Lecture 5: Monte Carlo Methods

Dr. Wen Fuxi

Motivating example: Monte Carlo estimation

▷ How can we estimate something without models?

- The simplest idea: Monte Carlo estimation.

Example: Flip a coin

The result (either head or tail) is denoted as a random variable x

▷ if the result is head, then $x = +1$

▷ if the result is tail, then $x = -1$

The aim is to compute $\mathbb{E}[x]$.

Motivating example: Monte Carlo estimation

Method 1: with model

- Suppose the probabilistic model is known as

$$p(X = 1) = 0.5, \quad p(X = -1) = 0.5$$

Then by Definition

$$\mathbb{E}[X] = \sum_x x\, p(x) = 1 \times 0.5 + (-1) \times 0.5 = 0$$

- Problem: It may be impossible to know the precise distribution!

Motivating example: Monte Carlo estimation

Method 2: without a model, or called model-free

- Idea: Flip the coin many times, and then calculate the average of the outcomes.
- Suppose we get a sample sequence: $\{x_1, x_2, \ldots, x_N\}$.

  Then, the mean can be approximated as

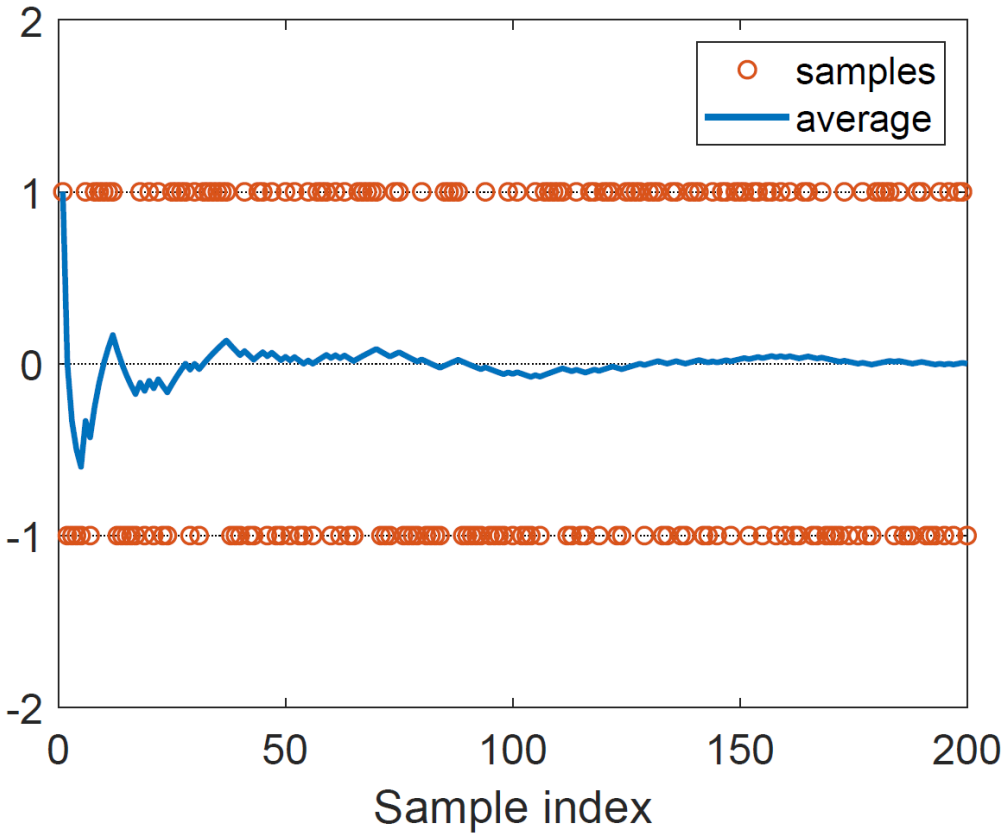
$$\mathbb{E}[X] \approx \bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j.$$

  This is the idea of Monte Carlo estimation!

Motivating example: Monte Carlo estimation

Question: Is the Monte Carlo estimation accurate?

- When N is small, the approximation is inaccurate.
- When N is large, the approximation is more accurate. (Law of Large Numbers)



Motivating example: Monte Carlo estimation

Law of Large Numbers

For a random variable X, suppose $\{x_j\}_{j=1}^N$ are some iid samples.
Let $\bar{x} = \frac{1}{N}\sum_{j=1}^N x_j$ be the average of the samples. Then,

$$\mathbb{E}[\bar{x}] = \mathbb{E}[X],$$

$$\mathrm{Var}[\bar{x}] = \frac{1}{N}\mathrm{Var}[X].$$

As a result, $\bar{x}$ is an unbiased estimate of $\mathbb{E}[X]$ and its variance decreases to zero as $N \to \infty$.

▷ The samples must be iid (independent and identically distributed)

Motivating example: Monte Carlo estimation

Stanisław Ulam – Magician of Mathematics

First developed by Stanislaw Ulam in the late 1940s (a Polish-American mathematician)

- Worked in the Manhattan Project to carry out hydro-dynamical calculations
- As it was secret, the MC method required a code name: Monte Carlo casino in Monaco was where Ulam's uncle liked to gamble.

Motivating example: Monte Carlo estimation

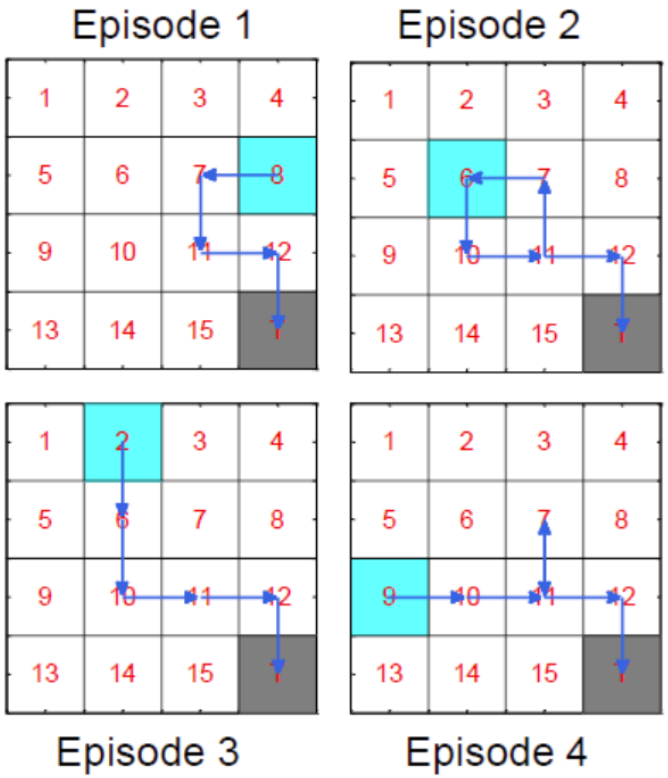
Why we care about mean estimation?

Because state value and action value are defined as expectations of random variables!

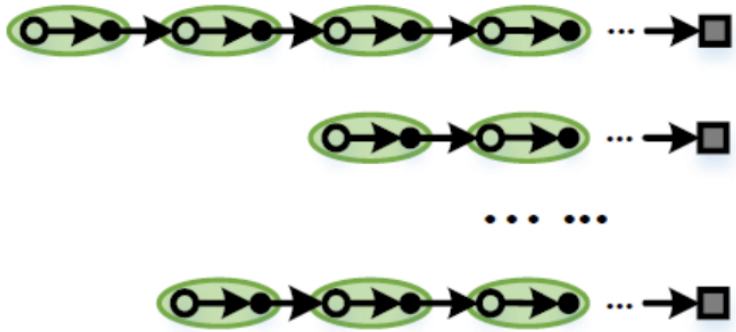
- Monte Carlo estimation refers to a broad class of techniques that rely on repeated random sampling to solve approximation problems. (simplest)
- Why do we care about Monte Carlo estimation? Because it does not require the model! (model-free)
- Use a complete trajectory and no **bootstrapping**.
- Only suits episodic tasks, i.e., tasks must terminate.

Basic idea of Monte Carlo estimation

Evaluate the value function by calculating the average return from a collection of episodes.



Generate a few episodes by interacting with environment (state-action pairs)



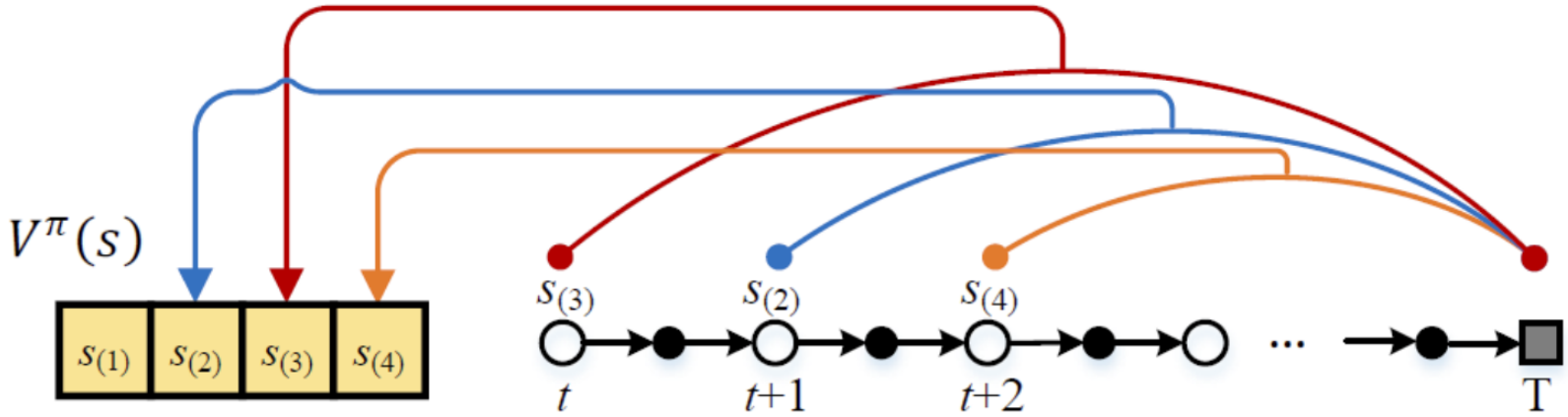
Basic idea of Monte Carlo estimation

Estimate of state-value function

$$v_\pi(s) = \text{Avg}\{G_{t:T} \mid S_t = s\}$$

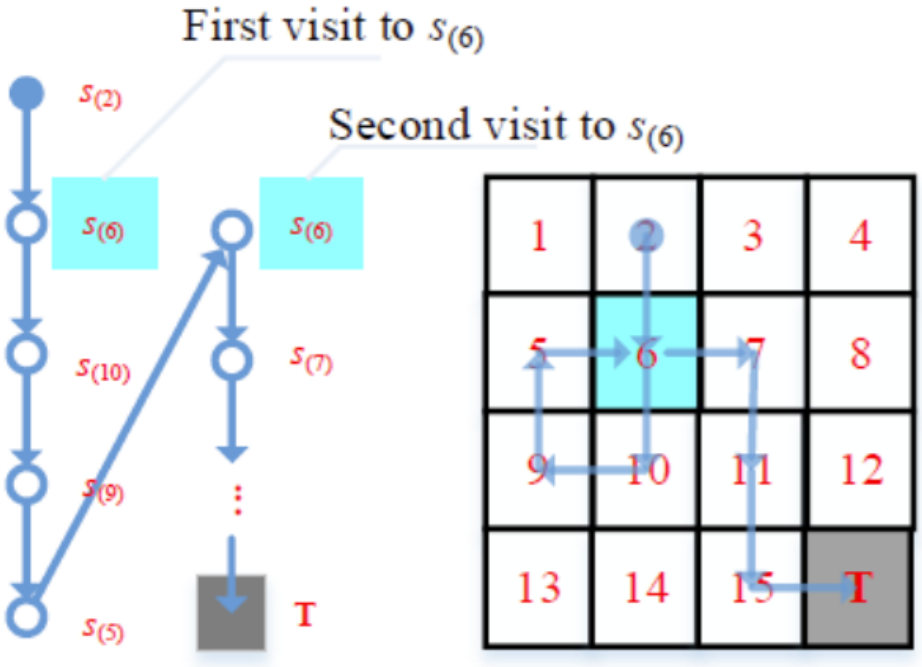
$$G_{t:T} = \sum_{i=0}^{T-t} \gamma^i r_{t+i}$$

Average all the returns following a particular state s



(a) Estimation of state-value function

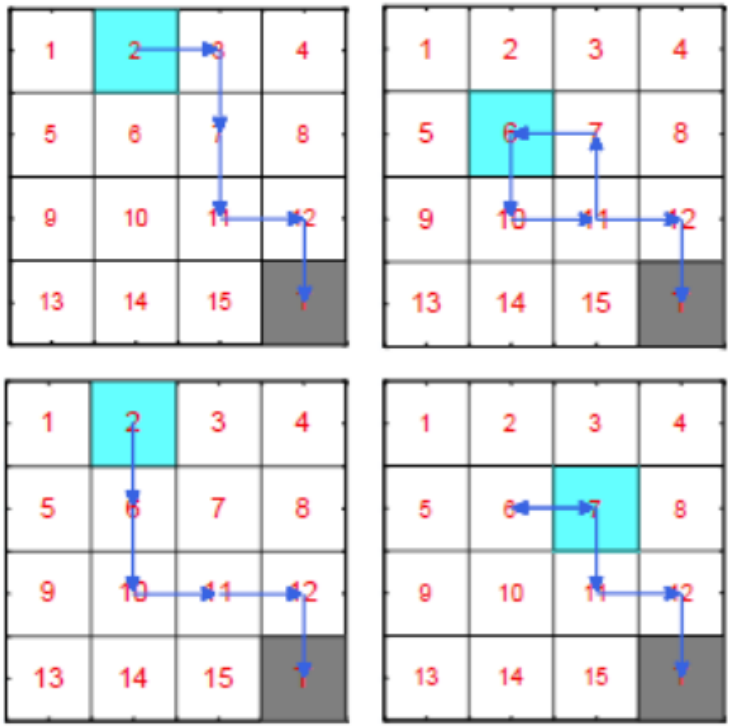
First Visit MC VS Every Visit MC

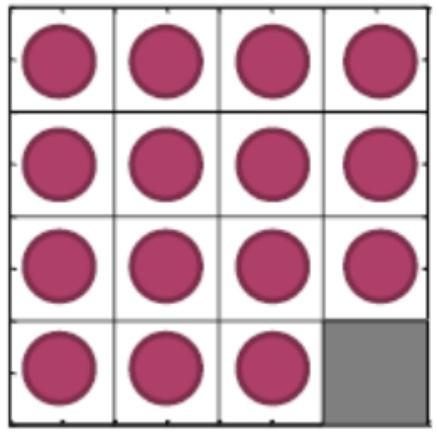


How many states can be updated with MC estimation?

- First-visit MC: estimate as the average of returns only following first visits to s
- Every-visit MC: estimate as the average of returns following all visits to s

An example: Given a policy, 4 episodes are collected





? $V^{\pi}(s_{(7)})$

$V^{\pi}(s_{(10)})$

$V^{\pi}(s_{(5)})$

Starting state

$$r(s,a,s') = \begin{cases} -1 & \text{if } s' \neq \text{T} \\ +9\,' & \text{if } s' = \text{T} \end{cases}$$

Discount factor: $\gamma = 0.9$

Model-free RL Problem

$$\max_{\pi} J(\pi) = \mathbb{E}_{s\sim d_{\text{init}}(s)}\{v^{\pi}(s)\}$$

Subject to:

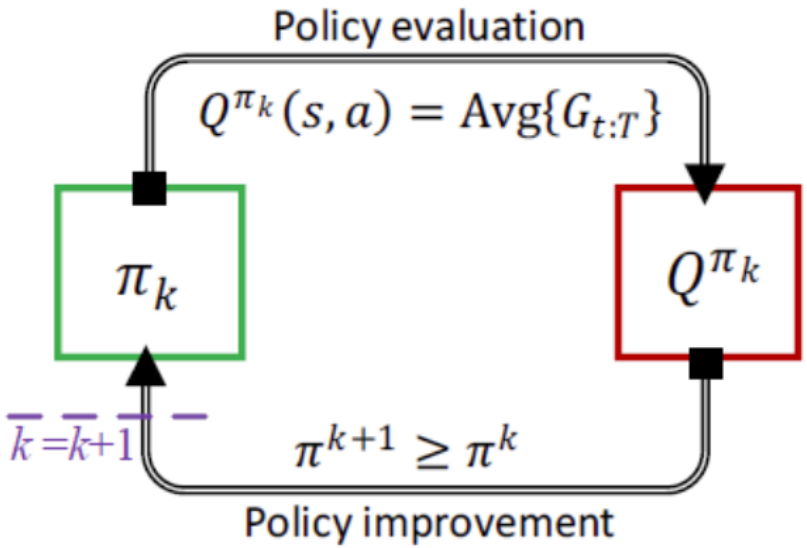
$$\mathcal{D} = \left\{ \begin{matrix} s_0, a_0, \\ s_1, a_1, \\ s_2, a_2, \\ \cdots\cdots, \\ s_t, a_t, \\ s_{t+1}, a_{t+1}, \\ \cdots \end{matrix} \right\}$$

(1) Environment model
(2) State-action samples
(3) Policy
(4) Reward (value function)

Optimal policy: π^* ?

Monte Carlo learning algorithm

(1) Policy EValuation (PEV) and (2) Policy IMprovement (PIM)



- PEV: Apply Monte Carlo estimation to approximate the true action-value function of the current policy.
- PIM: Find a new better policy based on the estimate of the action value function.

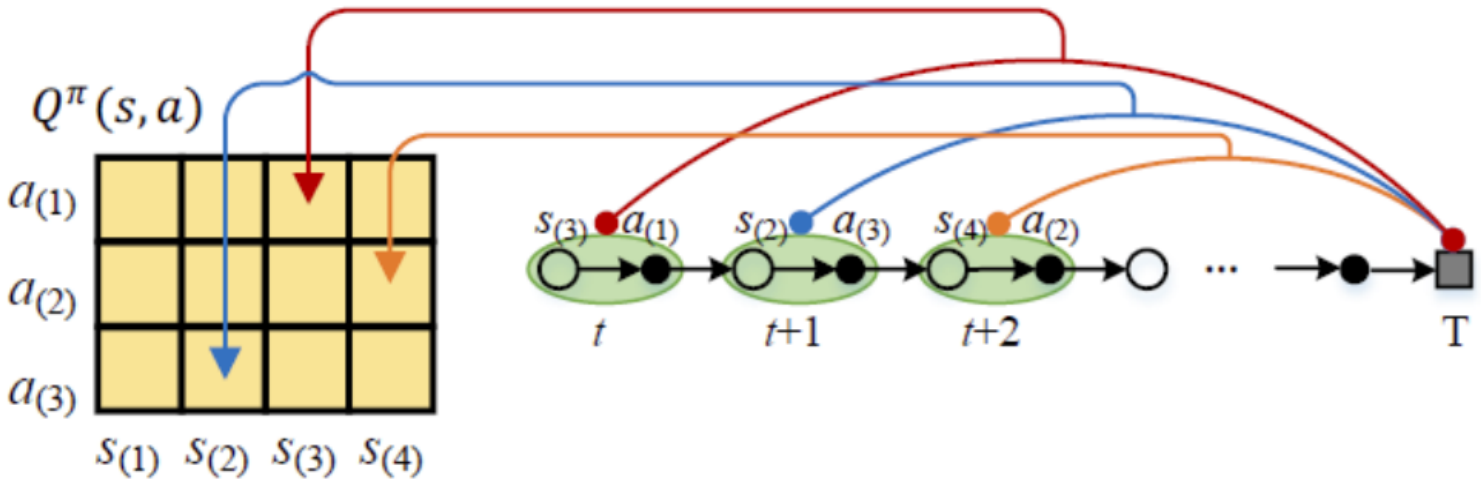
Policy EValuation (PEV)

MC estimate of action-value function $Q_\pi(s, a)$

$$Q_\pi(s, a) = \text{Avg}\{G_{t:T} \mid S_t = s, A_t = a\}$$

$$G_{t:T} = \sum_{i=0}^{T-t} \gamma^i r_{t+i}$$

Average all the returns followed after a particular state-action pair (s, a)



(b) Estimation of action-value function

Convert policy iteration to be model-free

The key idea is to convert the policy iteration algorithm to be model-free.

It is simple to understand if you

- Understand the policy iteration algorithm well, and
- Understand the idea of Monte Carlo mean estimation.

Convert policy iteration to be model-free

Policy iteration has two steps in each iteration:

$$\begin{cases} \text{Policy evaluation: } v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k} \\ \text{Policy improvement: } \pi_{k+1} = \arg\max_{\pi} \left(r_\pi + \gamma P_\pi v_{\pi_k} \right) \end{cases}$$

The elementwise form of the policy improvement step is:

$$\begin{aligned} \pi_{k+1}(s) &= \arg\max_{\pi} \sum_a \pi(a \mid s) \left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_{\pi_k}\left(s'\right) \right] \\ &= \arg\max_{\pi} \sum_a \pi(a \mid s) q_{\pi_k}(s, a), \quad s \in \mathcal{S} \end{aligned}$$

- The key is to calculate $q_{\pi_k}(s, a)$!

Convert policy iteration to be model-free

Two expressions of action value:

- Expression 1: requires the model:

$$q_{\pi_k}(s,a) = \sum_r p(r \mid s,a) r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v_{\pi_k}\left(s'\right)$$

- Expression 2: does not require the model:

$$q_{\pi_k}(s,a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

Idea to achieve model-free RL

We can use expression 2 to obtain $q_{\pi_k}(s,a)$ based on data (samples or experiences)!

Convert policy iteration to be model-free

How to obtain $q_{\pi_k}(s, a)$ based on data?

- Starting from (s, a), following policy π_k, generate an episode.
- The return of this episode is $g(s, a)$, which is a sample of G_t in

$$q_{\pi_k}(s, a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

- Suppose we have a set of episodes and hence $\left\{g^{(j)}(s, a)\right\}$. Then,

$$q_{\pi_k}(s, a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right] \approx \frac{1}{N}\sum_{i=1}^{N} g^{(i)}(s, a)$$

Fundamental idea: When a model is unavailable, we can use data.

The MC Basic algorithm

Given an initial policy π_0, there are two steps in the k th iteration.

Step 1: Policy Evaluation

This step aims to estimate $q_{\pi_k}(s,a)$ for all (s,a). Specifically, for each (s,a), run sufficiently many episodes. The average of their returns, denoted as $q_k(s,a)$, is used to approximate $q_{\pi_k}(s,a)$.

- The first step of the **policy iteration algorithm** calculates $q_{\pi_k}(s,a)$ by firstly solving the state values v_{π_k} from the Bellman equation. This requires the model.
- The first step of the **MC Basic algorithm** is to directly estimate $q_k(s,a)$ from experience samples. This does not require the model.

Step 2: Policy Improvement

This step aims to solve $\pi_{k+1}(s) = \arg\max_\pi \sum_a \pi(a \mid s) q_k(s,a)$ for all $s \in \mathcal{S}$.
The greedy optimal policy is $\pi_{k+1}\left(a_k^* \mid s\right) = 1$ where $a_k^* = \arg\max_a q_k(s,a)$.

- This step is exactly the same as the second step of the policy iteration algorithm.

The MC Basic algorithm

Pseudocode: MC Basic algorithm (a model-free variant of policy iteration)

Initialization: Initial guess π_0.

Aim: Search for an optimal policy.

For the k th iteration ($k = 0, 1, 2, \ldots$), do

For every state $s \in \mathcal{S}$, do

For every action $a \in \mathcal{A}(s)$, do

Collect sufficiently many episodes starting from (s, a) *following* π_k

Policy evaluation:

$q_{\pi_k}(s, a) \approx q_k(s, a) =$ average return of all the episodes starting from (s, a)

Policy improvement:

$a_k^*(s) = \arg\max_a q_k(s, a)$

$\pi_{k+1}(a \mid s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a \mid s) = 0$ otherwise

$k = k + 1$

The MC Basic algorithm

MC Basic is a variant of the policy iteration algorithm

The model-free algorithms are built upon model-based ones. It is, therefore, necessary to understand model-based algorithms first before studying model-free algorithms.

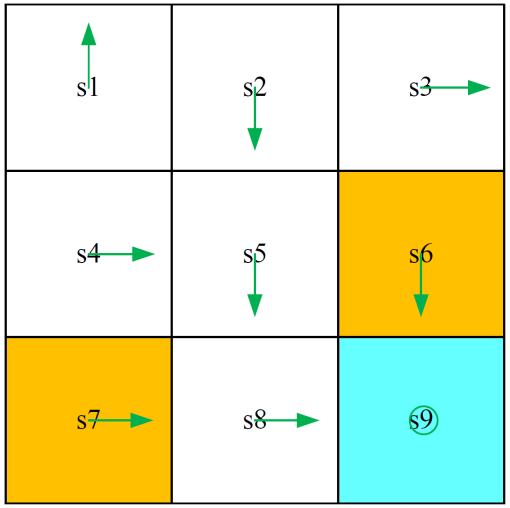
- MC Basic is useful for revealing the core idea of MC-based model-free RL, but it is not practical due to its low efficiency.

Why does MC Basic estimate action values instead of state values?

That is because state values cannot be used directly to improve policies. When models are unavailable, we should estimate action values directly.

- Since policy iteration is convergent, the convergence of MC Basic is also guaranteed to be convergent, provided there are sufficient episodes.

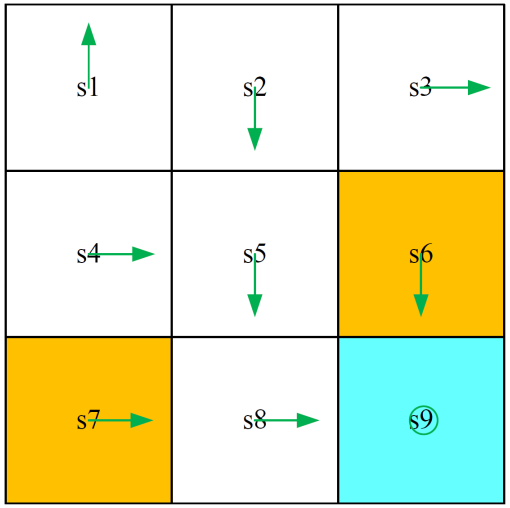
Illustrative example 1: step by step



Task:

- An initial policy is shown in the figure.
- Use MC Basic to find the optimal policy.
- $r_{\text{boundary}} = -1, r_{\text{forbidden}} = -1, r_{\text{target}} = 1, \gamma = 0.9.$

Illustrative example 1: step by step



Outline: Given the current policy π_k

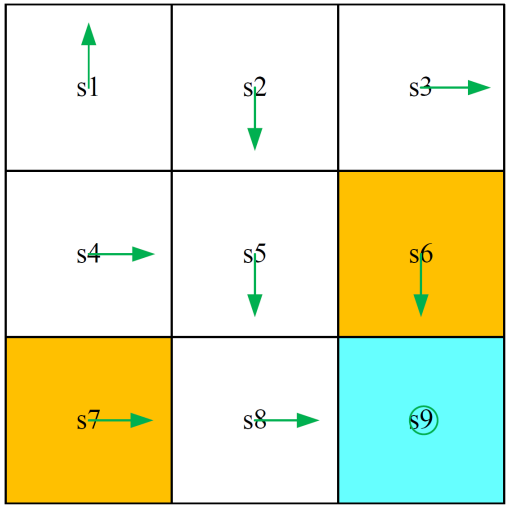
- Step 1 - Policy evaluation: calculate $q_{\pi_k}(s, a)$

  How many state-action pairs? 9 states $\times 5$ actions $= 45$ state-action pairs!

- Step 2 - Policy improvement: select the greedy action

  $a^*(s) = \arg\max_a q_{\pi_k}(s, a)$

Illustrative example 1: step by step

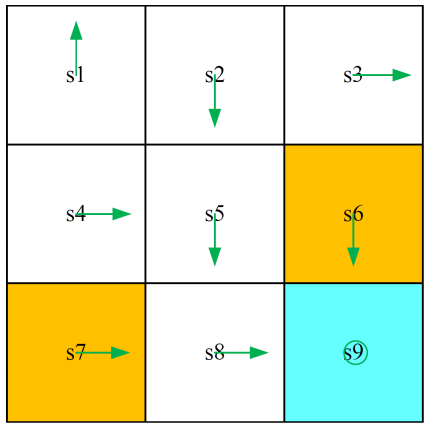


Due to space limitation, we only show $q_{\pi_k}(s_1, a)$

▷ Step 1 - Policy evaluation:

- Since the current policy is deterministic, one episode would be sufficient to get the action value! Otherwise, if the policy or model is stochastic, many episodes are required!

Illustrative example 1: step by step



- Starting from (s_1, a_1), the episode is $s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \ldots$. Hence, the action value is

$$q_{\pi_0}(s_1, a_1) = -1 + \gamma(-1) + \gamma^2(-1) + \ldots$$

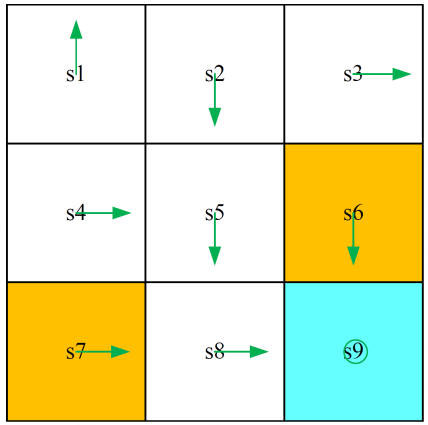
- Starting from (s_1, a_2), the episode is $s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_3} \ldots$. Hence, the action value is

$$q_{\pi_0}(s_1, a_2) = 0 + \gamma 0 + \gamma^2 0 + \gamma^3(1) + \gamma^4(1) + \ldots$$

- Starting from (s_1, a_3), the episode is $s_1 \xrightarrow{a_3} s_4 \xrightarrow{a_2} s_5 \xrightarrow{a_3} \ldots$. Hence, the action value is

$$q_{\pi_0}(s_1, a_3) = 0 + \gamma 0 + \gamma^2 0 + \gamma^3(1) + \gamma^4(1) + \ldots$$

Illustrative example 1: step by step



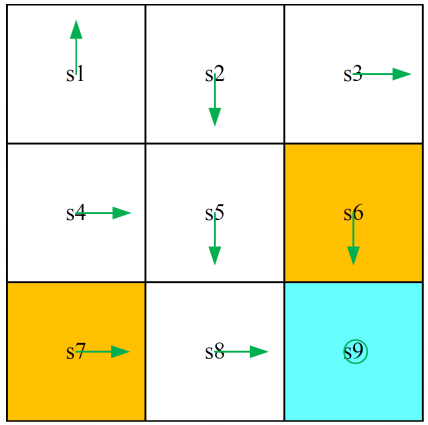
- Starting from (s_1, a_4), the episode is $s_1 \xrightarrow{a_4} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \ldots$. Hence, the action value is

$$q_{\pi_0}(s_1, a_4) = -1 + \gamma(-1) + \gamma^2(-1) + \ldots$$

- Starting from (s_1, a_5), the episode is $s_1 \xrightarrow{a_5} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \ldots$. Hence, the action value is

$$q_{\pi_0}(s_1, a_5) = 0 + \gamma(-1) + \gamma^2(-1) + \ldots$$

Illustrative example 1: step by step



▷ Step 2 - Policy improvement:

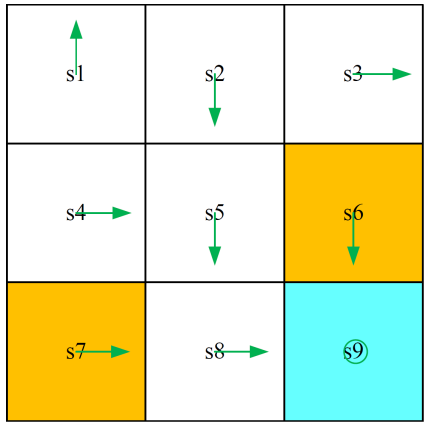
- By observing the action values, we see that $q_{\pi_0}(s_1, a_2) = q_{\pi_0}(s_1, a_3)$ are the maximum.
- As a result, the policy can be improved as

$$\pi_1(a_2 \mid s_1) = 1 \text{ or } \pi_1(a_3 \mid s_1) = 1.$$

In either way, the new policy for s_1 becomes optimal.

One iteration is sufficient for this simple example!

Illustrative example 1: step by step



Exercise: Now update the policy for s_3 using MC Basic!

Use data more efficiently

The MC Basic algorithm:

- Advantage: Reveal the core idea clearly!
- Disadvantage: Too simple to be practical.

However, MC Basic can be further enhanced to increase efficiency.

Use data more efficiently

▷ Consider a grid-world example, following a policy π, we can get an episode such as

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots$$

Visit: Every time a state-action pair appears in the episode, it is called a visit of that state-action pair.

▷ Methods to use the data: Initial-visit method

- Just calculate the return and approximate $q_\pi(s_1, a_2)$.
- This is what the MC Basic algorithm does.
- Disadvantage: Not fully utilize the data.

Use data more efficiently

▷ The episode also visits other state-action pairs.

$$
\begin{aligned}
s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots \quad & [\text{ original episode }] \\
s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots \quad & [\text{ episode starting from } (s_2, a_4)] \\
s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots \quad & [\text{ episode starting from } (s_1, a_2)] \\
s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots \quad & [\text{ episode starting from } (s_2, a_3)] \\
s_5 \xrightarrow{a_1} \dots \quad & [\text{ episode starting from } (s_5, a_1)]
\end{aligned}
$$

Can estimate $q_\pi(s_1, a_2), q_\pi(s_2, a_4), q_\pi(s_2, a_3), q_\pi(s_5, a_1), \dots$

Data-efficient methods:

- First-visit method
- Every-visit method

Update value estimate more efficiently

Another aspect in MC-based RL is when to update the policy.

There are two methods.

The first method

in the policy evaluation step involves collecting all episodes starting from a state-action pair and then using the average return to approximate the action value.

- This is the one adopted by the MC Basic algorithm.
- The problem with this method is that the agent has to wait until all episodes have been collected.

The second method

Uses the return of a single episode to approximate the action value.

In this way, we can improve the policy episode-by-episode.

Update value estimate more efficiently

▷ Will the second method cause problems?

- One may say that the return of a single episode cannot accurately approximate the corresponding action value.
- In fact, we have done that in the truncated policy iteration algorithm introduced in the last chapter!

▷ Generalized Policy Iteration:

- Not a specific algorithm.
- It refers to the general idea or framework of switching between policy-evaluation and policy-improvement processes.
- Many RL algorithms fall into this framework.

MC Exploring Starts

If we use data and update the estimate more efficiently, we get a new algorithm called MC Exploring Starts:

Algorithm (an efficient variant of MC Basic)

Initialization: Initial policy $\pi_0(a \mid s)$ and initial value $q(s, a)$ for all (s, a).

Returns $(s, a) = 0$ and $\text{Num}(s, a) = 0$ for all (s, a).

Goal: Search for an optimal policy.

For each episode, do

Episode generation: Select a starting state-action pair (s_0, a_0) and ensure that all pairs can be possibly selected (this is the exploring-starts condition). Following the current policy, generate an episode of length T : $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$.

Initialization for each episode: $g \leftarrow 0$

For each step of the episode, $t = T-1, T-2, \ldots, 0$, do

$g \leftarrow \gamma g + r_{t+1}$

$\text{Returns}(s_t, a_t) \leftarrow \text{Returns}(s_t, a_t) + g$

$\text{Num}(s_t, a_t) \leftarrow \text{Num}(s_t, a_t) + 1$

Policy evaluation:

$q(s_t, a_t) \leftarrow \text{Returns}(s_t, a_t) / \text{Num}(s_t, a_t)$

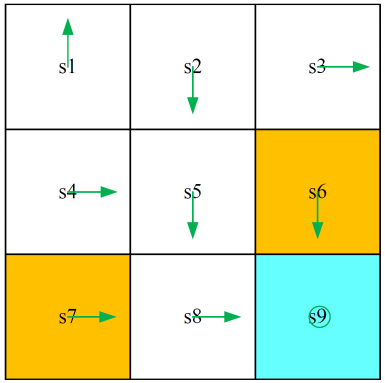
Policy improvement:

$\pi(a \mid s_t) = 1$ if $a = \arg\max_a q(s_t, a)$ and $\pi(a \mid s_t) = 0$ otherwise

MC Exploring Starts

▷ What is Exploring Starts?

- Exploring starts means we need to generate sufficiently many episodes $\underbrace{\text{starting}}_{\text{starts}}$ from $\underbrace{\text{every}}_{\text{exploring}}$ state-action pair.
- For example, we need episodes starting from $\{(s_1, a_j)\}_{j=1}^{5}$, $\{(s_2, a_j)\}_{j=1}^{5}$, $\ldots$, $\{(s_9, a_j)\}_{j=1}^{5}$.
  Otherwise, if there are no episodes starting from (s_i, a_j), then (s_i, a_j) may not be visited by any episodes, and hence $q_\pi(s_i, a_j)$ cannot be estimated.



- Both MC Basic and MC Exploring Starts need this assumption.

MC Exploring Starts

Why do we need to consider Exploring Starts?

- In theory, only if every action value for every state is well explored, can we select the optimal actions correctly.

  Otherwise, if an action is not explored, it may turn out to be the optimal one and thus be missed.

- In practice, exploring starts is difficult to achieve.

  For many applications, especially those involving physical interactions with the environment, it is challenging to collect episodes starting from every state-action pair.

Can we remove the requirement of Exploring Starts? We next demonstrate that this can be achieved by employing soft policies.

Soft policies

What is a soft policy?

- A policy is soft if the probability of taking any action is positive.

  Deterministic policy: For example, greedy policy

  Stochastic policy: For example, soft policy

Why introduce soft policies?

- A few sufficiently long episodes can visit every state-action pair for soft policy.
- Then, we do not need to have a large number of episodes starting from every state-action pair.

  Hence, the requirement to explore starts can be removed.

ε-greedy policies

What soft policies will we use? Answer: ε-greedy policies

What is the ε-greedy policy?

$$\pi(a \mid s) = \begin{cases} 1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1), & \text{for the greedy action} \\ \frac{\varepsilon}{|\mathcal{A}(s)|}, & \text{for the other } |\mathcal{A}(s)| - 1 \text{ actions} \end{cases}$$

where $\varepsilon \in [0, 1]$ and $|\mathcal{A}(s)|$ is the number of actions for s.

- Example: if $\epsilon = 0.2$, then

$$\frac{\varepsilon}{|\mathcal{A}(s)|} = \frac{0.2}{5} = 0.04, \quad 1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = 1 - 0.04 \times 4 = 0.84$$

- The chance to choose the greedy action is always greater than other actions, because

$$1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|} \geq \frac{\varepsilon}{|\mathcal{A}(s)|}$$

ε-greedy policies

ε-greedy policies can balance Exploitation and Exploration.

- When $\varepsilon \to 0$, it becomes greedy!

$$\pi(a \mid s) = \begin{cases} 1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = 1, & \text{for the greedy action} \\ \frac{\varepsilon}{|\mathcal{A}(s)|} = 0, & \text{for the other } |\mathcal{A}(s)| - 1 \text{ actions} \end{cases}$$

More exploitation but less exploration.

- When $\varepsilon \to 1$, it becomes a uniform distribution.

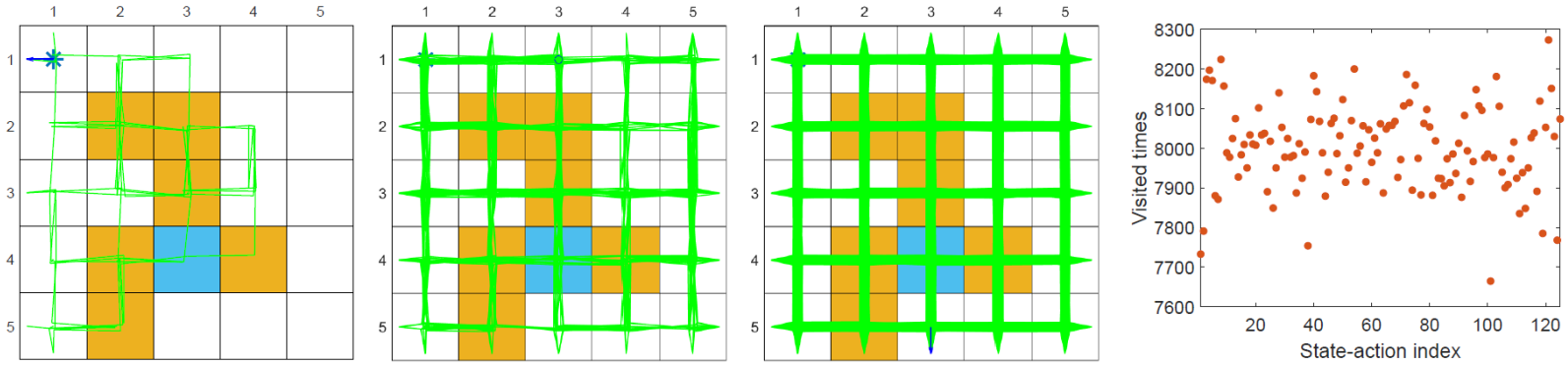
$$\pi(a \mid s) = \begin{cases} 1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = \frac{1}{|\mathcal{A}(s)|}, & \text{for the greedy action} \\ \frac{\varepsilon}{|\mathcal{A}(s)|} = \frac{1}{|\mathcal{A}(s)|}, & \text{for the other } |\mathcal{A}(s)| - 1 \text{ actions} \end{cases}$$

More exploration but less exploitation.

Examples to demonstrate the exploration ability

Can a single episode visit all state-action pairs?

When $\varepsilon = 1$, the policy (uniform distribution) has the strongest exploration ability.

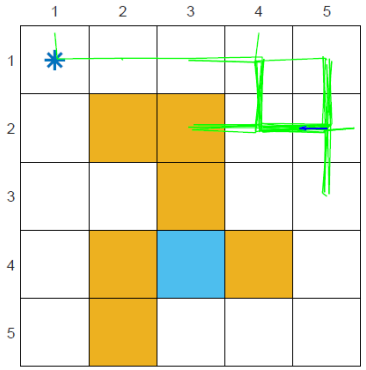


(a) 100 steps (b) 1000 steps (c) 10000 steps (d)

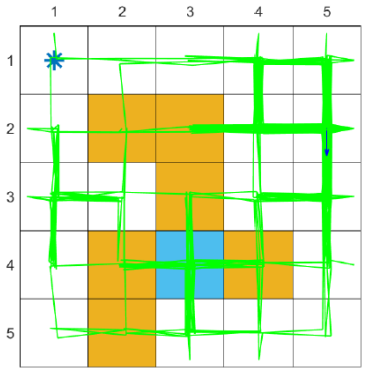
Examples to demonstrate the exploration ability

Can a single episode visit all state-action pairs?

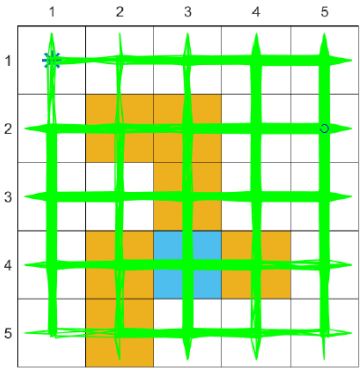
When ε is small, the exploration ability of the policy is also small.



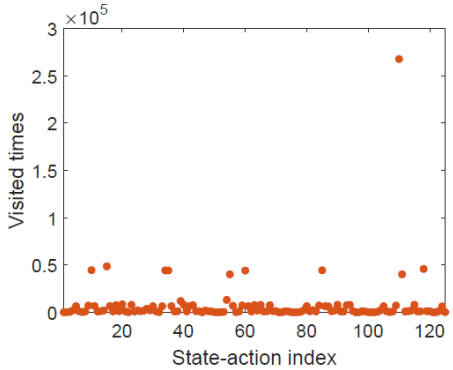
(a) 100 steps



(b) 1000 steps



(c) 10000 steps



(d)

MC ε-Greedy algorithm

▷ How to embed ε-greedy into the MC-based RL algorithms?

Originally, the policy improvement step in MC Basic and MC Exploring Starts is to solve

$$\pi_{k+1}(s) = \arg\max_{\pi \in \Pi} \sum_a \pi(a \mid s) q_{\pi_k}(s, a)$$

where Π denotes the set of all possible policies.

The optimal policy here is

$$\pi_{k+1}(a \mid s) = \begin{cases} 1, & a = a_k^* \\ 0, & a \neq a_k^* \end{cases}$$

where $a_k^* = \arg\max_a q_{\pi_k}(s, a)$.

MC ε-Greedy algorithm

▷ How to embed ε-greedy into the MC-based RL algorithms?

Now, the policy improvement step is changed to solve

$$\pi_{k+1}(s) = \arg\max_{\pi \in \Pi_\varepsilon} \sum_a \pi(a \mid s) q_{\pi_k}(s, a)$$

where Π_ε denotes the set of all ε-greedy policies with a fixed value of ε.

The optimal policy here is

$$\pi_{k+1}(a \mid s) = \begin{cases} 1 - \frac{|\mathcal{A}(s)|-1}{|\mathcal{A}(s)|}\varepsilon, & a = a_k^*, \\ \frac{1}{|\mathcal{A}(s)|}\varepsilon, & a \neq a_k^*. \end{cases}$$

Summary:

- MC ε-Greedy is the same as that of MC Exploring Starts except that the former uses ε-greedy policies.
- It does not require exploring starts, but still requires visiting all state-action pairs in a different form.

MC ε-Greedy algorithm

Algorithm: MC ϵ-Greedy (a variant of MC Exploring Starts)

Initialization: Initial policy $\pi_0(a \mid s)$ and initial value $q(s, a)$ for all (s, a). Returns $(s, a) = 0$ and $\mathrm{Num}(s, a) = 0$ for all (s, a).$\epsilon \in (0, 1]$

Goal: Search for an optimal policy.

For each episode, do

Episode generation: Select a starting state-action pair (s_0, a_0) (the exploring starts condition is not required).

Following the current policy, generate an episode of length T: $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$.

Initialization for each episode: $g \leftarrow 0$

For each step of the episode, $t = T-1, T-2, \ldots, 0$, do

$g \leftarrow \gamma g + r_{t+1}$

$\mathrm{Returns}(s_t, a_t) \leftarrow \mathrm{Returns}(s_t, a_t) + g$

$\mathrm{Num}(s_t, a_t) \leftarrow \mathrm{Num}(s_t, a_t) + 1$

Policy evaluation:

$q(s_t, a_t) \leftarrow \mathrm{Returns}(s_t, a_t) \,/\, \mathrm{Num}(s_t, a_t)$

Policy improvement:

Let $a^* = \arg\max_a q(s_t, a)$ and

$$\pi(a \mid s_t) = \begin{cases} 1 - \frac{|\mathcal{A}(s_t)|-1}{|\mathcal{A}(s_t)|}\epsilon, & a = a^* \\ \frac{1}{\mathcal{A}(s_t)|}\epsilon, & a \neq a^* \end{cases}$$

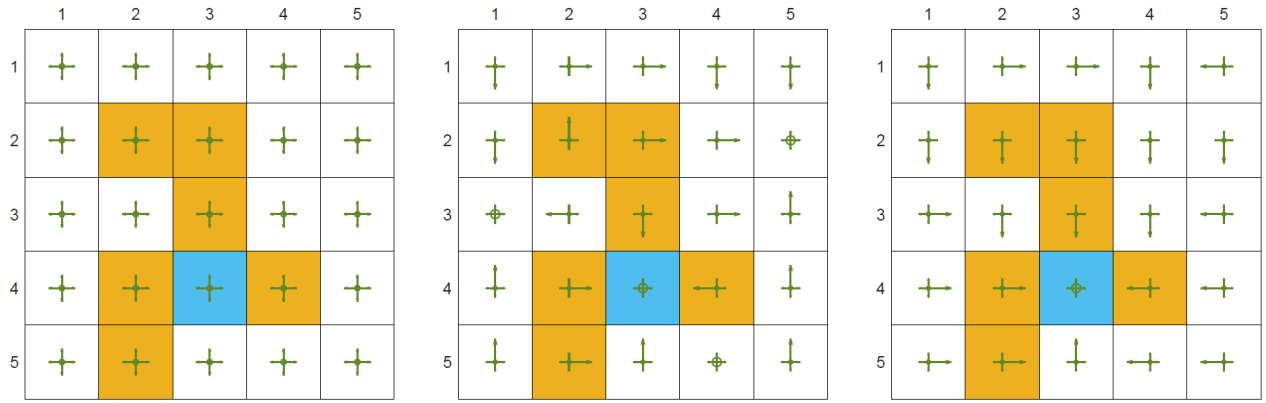
Example

▷ Demonstrate the MC ε-Greedy algorithm.

In every iteration, do the following:

- In the episode generation step, use the previous policy generates a single episode of 1 million steps!
- In the rest steps, use the single episode to update the policy.
- Two iterations can lead to the optimal ε-greedy policy.

Here, $r_{\text{boundary}} = -1, r_{\text{forbidden}} = -10, r_{\text{target}} = 1, \gamma = 0.9$.



(a) Initial policy (b) After the first iteration (c) After the second iteration

Optimality vs Exploration

▷ Compared to greedy policies,

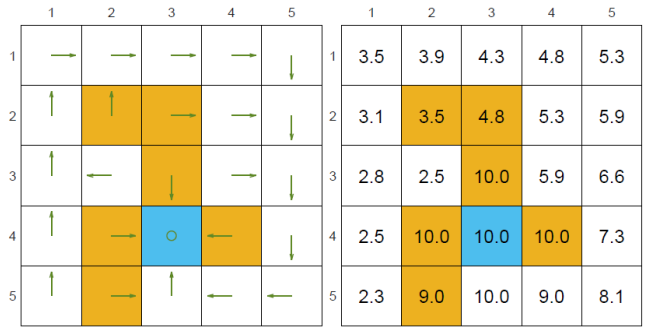
- The advantage of ε-greedy policies is that they have strong exploration ability when ε is large.
  - Then, the exploring starts condition is not required.

- The disadvantage is that ε-greedy polices are not optimal in general.
  - It is only optimal in the set Π_ε of all ε-greedy policies.
  - ε cannot be too large. We can also use a decaying ε.

▷ Next, we use examples to illustrate the concept. The setup is $r_{\text{boundary}} = -1$, $r_{\text{forbidden}} = -10, r_{\text{target}} = 1, \gamma = 0.9$

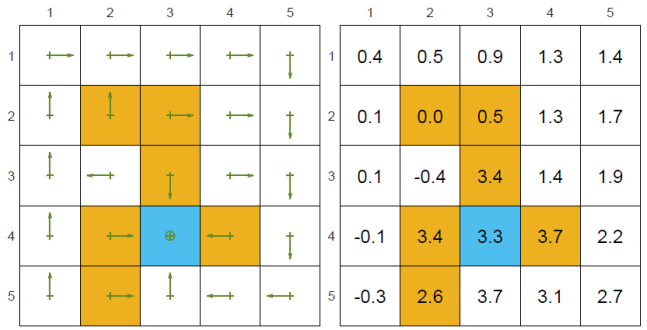
Optimality

▷ Given an ε-greedy policy, what is its state value?



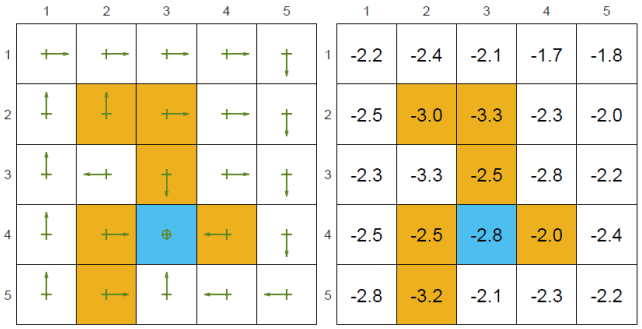
	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

$\varepsilon = 0$



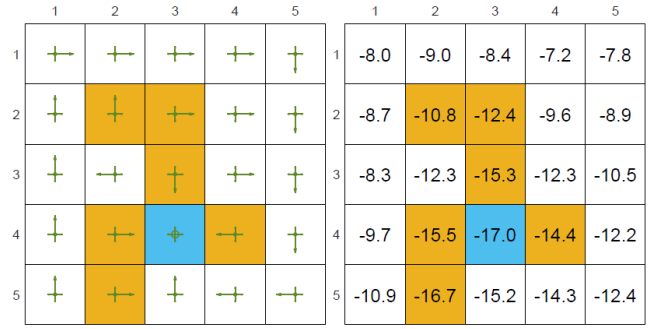
	1	2	3	4	5
1	0.4	0.5	0.9	1.3	1.4
2	0.1	0.0	0.5	1.3	1.7
3	0.1	-0.4	3.4	1.4	1.9
4	-0.1	3.4	3.3	3.7	2.2
5	-0.3	2.6	3.7	3.1	2.7

$\varepsilon = 0.1$



	1	2	3	4	5
1	-2.2	-2.4	-2.1	-1.7	-1.8
2	-2.5	-3.0	-3.3	-2.3	-2.0
3	-2.3	-3.3	-2.5	-2.8	-2.2
4	-2.5	-2.5	-2.8	-2.0	-2.4
5	-2.8	-3.2	-2.1	-2.3	-2.2

$\varepsilon = 0.2$



	1	2	3	4	5
1	-8.0	-9.0	-8.4	-7.2	-7.8
2	-8.7	-10.8	-12.4	-9.6	-8.9
3	-8.3	-12.3	-15.3	-12.3	-10.5
4	-9.7	-15.5	-17.0	-14.4	-12.2
5	-10.9	-16.7	-15.2	-14.3	-12.4

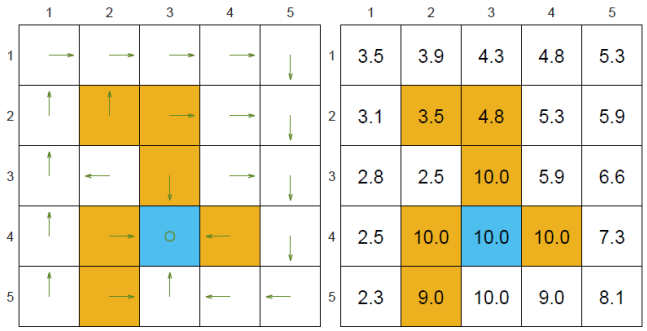
$\varepsilon = 0.5$

▷ When ε increases, the optimality of the policy becomes worse!

▷ Why is the state value of the target state negative?

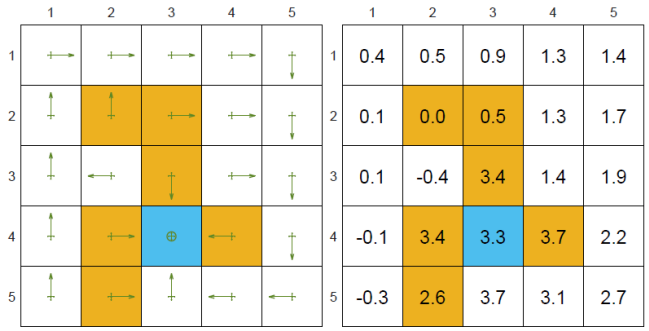
Consistency

▷ Find the optimal ε-greedy policies and their state values?



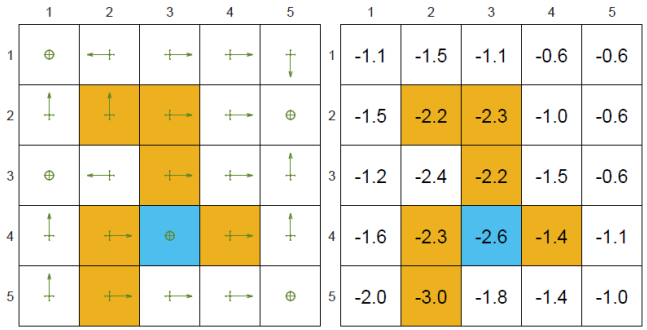
	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

$\varepsilon = 0$



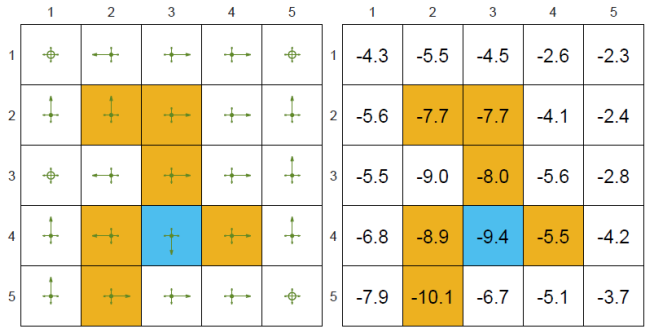
	1	2	3	4	5
1	0.4	0.5	0.9	1.3	1.4
2	0.1	0.0	0.5	1.3	1.7
3	0.1	-0.4	3.4	1.4	1.9
4	-0.1	3.4	3.3	3.7	2.2
5	-0.3	2.6	3.7	3.1	2.7

$\varepsilon = 0.1$



	1	2	3	4	5
1	-1.1	-1.5	-1.1	-0.6	-0.6
2	-1.5	-2.2	-2.3	-1.0	-0.6
3	-1.2	-2.4	-2.2	-1.5	-0.6
4	-1.6	-2.3	-2.6	-1.4	-1.1
5	-2.0	-3.0	-1.8	-1.4	-1.0

$\varepsilon = 0.2$



	1	2	3	4	5
1	-4.3	-5.5	-4.5	-2.6	-2.3
2	-5.6	-7.7	-7.7	-4.1	-2.4
3	-5.5	-9.0	-8.0	-5.6	-2.8
4	-6.8	-8.9	-9.4	-5.5	-4.2
5	-7.9	-10.1	-6.7	-5.1	-3.7

$\varepsilon = 0.5$

▷ The optimal ε-greedy policies are not consistent with the greedy optimal one! Why is that? Consider the target for example.

Summary

Key Points:

- Mean estimation by the Monte Carlo methods
- Three algorithms:
  1. MC Basic
  2. MC Exploring Starts
  3. MC ε-Greedy
- Relationship among the three algorithms
- Optimality vs exploration of ε-greedy policies