به نام خدا

Iran University
of Science and
Technology

یادگیری تقویتی در کنترل

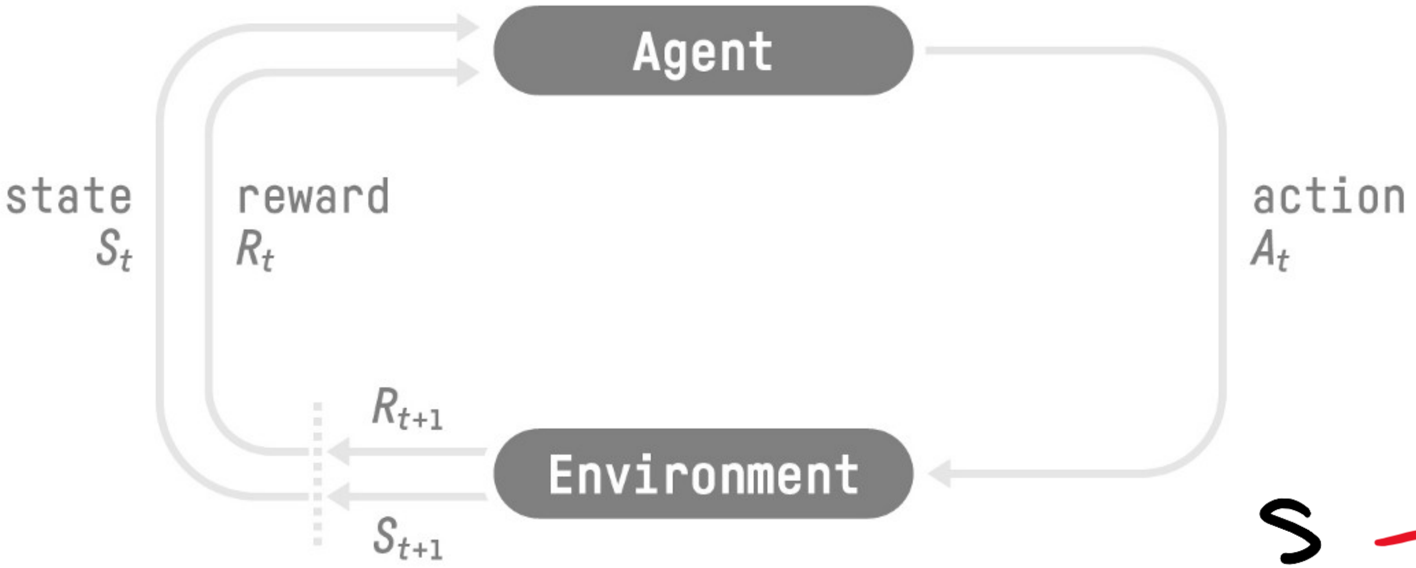
دکتر سعید شمقدری

دانشکده مهندسی برق
گروه کنترل

نیمسال اول ۱۴۰۵–۱۴۰۴

Monte Carlo Methods

دانش از محیط



عدم اطلاعات کامل؟

تجربه از محیط:

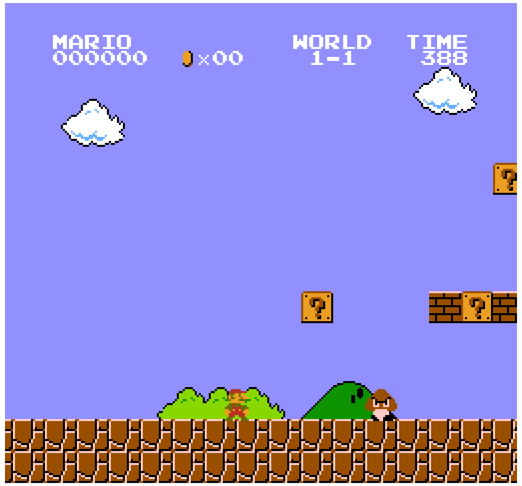
$S \rightarrow A \rightarrow R$

محیط واقعی/محیط شبیه سازی

راه حل؟

- تخمین دینامیک محیط ← حل با DP
- تخمین تابع Value از اندازه گیری (Learning)

فرض



Episodic Task

تخمین Value در انتهای هر اپیزود ؟؟

یک روش : متوسط گیری از Return های تجره شده از هر State

روش مونت کارلو

مشاهده Ret بیشتر ← تخمین بهتر .Exp. Ret

تعریف

First Visit:

در یک اپیزود، اولین بار عبور از حالت s

- روش Monte Carlo First Visit:
  تخمین V_π بر اساس متوسط Return تا First Visit

- روش Monte Carlo Every Visit:
  تخمین V_π بر اساس متوسط Return همه Visit ها تا s

الگوریتم

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

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

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

Loop forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

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

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

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

Monte Carlo Estimation of Action Values

Action Value or State Value?

$$q_\pi(s, a)$$

- ایده تخمین؟

  Visit شدن s و اعمال a برای آن (First Visit)
- چالش تخمین q؟

  عدم ویزیت s و a! (مخصوصا برای پالیسی قطعی)

راه حل ویزیت s و a های مختلف؟
Exploration دائم

Monte Carlo Estimation of Action Values

راه حل ویزیت s و a های مختلف؟
Exploration دائم

راه حل:

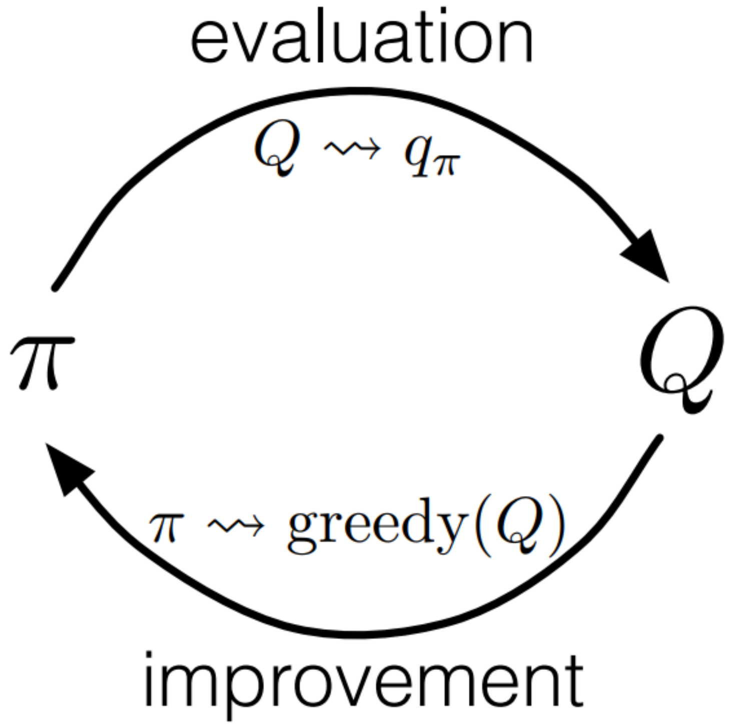
1. Exploring Starts:

شروع هر اپیزود با یک زوج s و a که احتمال انتخاب آن صفر نباشد.

سوال: تعداد ویزیت ها در بی نهایت تکرار اپیزود؟

2. Stochastic Policy

Monte Carlo Control



generalized policy iteration

سوال: مشکل پالیسی قطعی؟

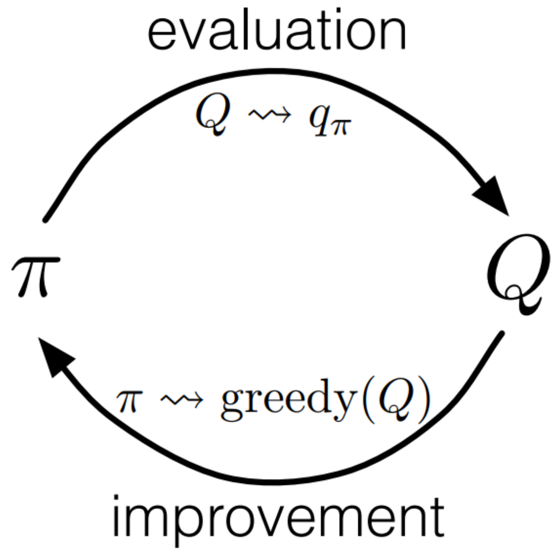
$$\pi_0 \xrightarrow{\mathrm{E}} q_{\pi_0} \xrightarrow{\mathrm{I}} \pi_1 \xrightarrow{\mathrm{E}} q_{\pi_1} \xrightarrow{\mathrm{I}} \pi_2 \xrightarrow{\mathrm{E}} \cdots \xrightarrow{\mathrm{I}} \pi_* \xrightarrow{\mathrm{E}} q_*$$

Monte Carlo Control

- انتهای هر اپیزود هم PE انجام میشود هم PI
- انتهای هر اپیزود بخشی از فضای s,a آپدیت میشود لذا GPI است

انتخاب گریدی پالیسی:

$$\pi(s) \doteq \arg\max_{a} q(s,a)$$



generalized policy iteration

Monte Carlo Control

قضیه بهبود سیاست

for all $s \in \mathcal{S}$

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s, a)) \\ &= \max_a q_{\pi_k}(s, a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \\ &\geq v_{\pi_k}(s). \end{aligned}$$

دو الزام:

- اپیزودهای با ES و بی نهایت اپیزود

Monte Carlo Control

دو الزام:

- اپیزودهای با ES و بی نهایت اپیزود

حل مشکل بی نهایت اپیزود:

Use **Value Iteration**

انتهای هر اپیزود:

- **Policy Evaluation**
- **Policy Improvement**

الگوریتم

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
- $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
- $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
- $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
- Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
- Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
    - Append G to $Returns(S_t, A_t)$
    - $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
    - $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Monte Carlo Control without Exploring Starts

راه حل:

- Stochastic Policy

تضمین انتخاب همه Action ها در بی نهایت تکرار

روش:

- On-Policy
- Off-Policy

سوال: روش MC-ES کدام است؟

Monte Carlo Control without Exploring Starts

Soft-Policy:

$$\pi(a|s) > 0$$

Epsilon Soft-Policy:

$$\pi(a|s) \geq \frac{\varepsilon}{|\mathcal{A}(s)|}$$

Epsilon Greedy Policy:

$$\text{probability of selection of} \begin{cases} \text{all nongreedy actions} & \frac{\varepsilon}{|\mathcal{A}(s)|} \\ \text{greedy actions} & 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|} \end{cases}$$

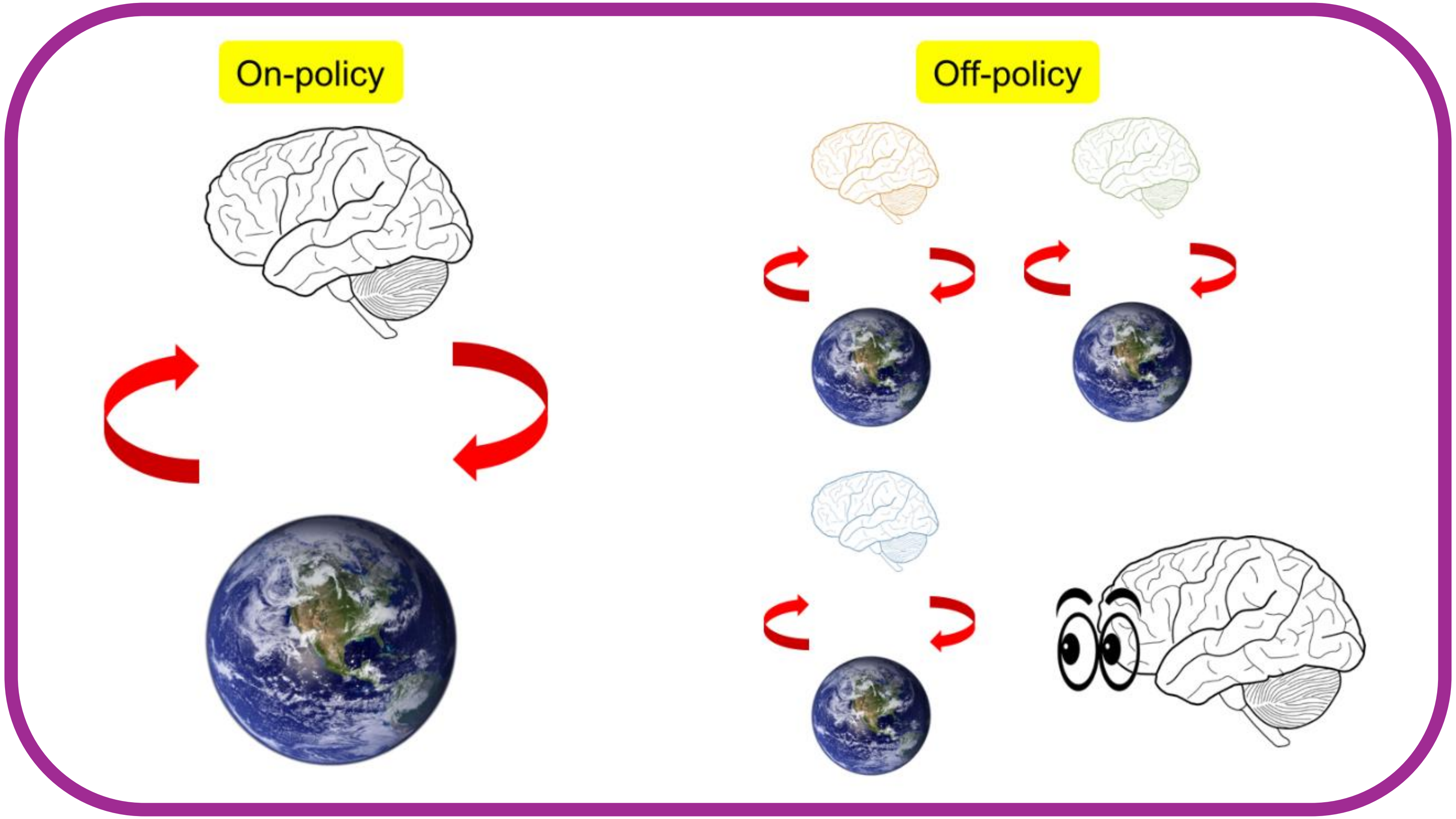
Monte Carlo Control without Exploring Starts

اثبات بهتر بودن سیاست اپسیلون گریدی π' نسبت به سیاست های سافت π :

$$\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a|s) q_\pi(s,a) \\
&= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;+\; (1-\varepsilon) \max_a q_\pi(s,a) \\
&\geq \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;+\; (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{1-\varepsilon} q_\pi(s,a) \\
&= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;-\; \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;+\; \sum_a \pi(a|s) q_\pi(s,a) \\
&= v_\pi(s).
\end{aligned}$$

On-policy RL vs Off-policy RL

دو مفهوم مهم در یادگیری تقویتی!



الگوریتم

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

Algorithm parameter: small $\varepsilon > 0$

Initialize:

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

Repeat forever (for each episode):

- Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
    - Append G to $Returns(S_t, A_t)$
    - $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
    - $A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)
    - For all $a \in \mathcal{A}(S_t)$:

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

Monte Carlo Control without Exploring Starts

رویکرد دوم برای اثبات:
انتقال اتفاقی بودن پالیسی به محیط

Action *a*:

- $P = 1 - \varepsilon \rightarrow$ مشابه محیط اصلی
- $P = \varepsilon \rightarrow$ مشابه واکنش محیط به یک عمل تصادفی با توزیع یکنواخت

Monte Carlo Control without Exploring Starts

رویکرد دوم برای اثبات:
انتقال اتفاقی بودن پالیسی به محیط
$\widetilde{V}_*$: تابع ارزش برای محیط جدید

$$\begin{aligned}\widetilde{v}_*(s) &= (1-\varepsilon)\max_a \widetilde{q}_*(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a \widetilde{q}_*(s,a) \\ &= (1-\varepsilon)\max_a \sum_{s',r} p(s',r|s,a)\Big[r+\gamma\widetilde{v}_*(s')\Big] \\ &+ \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a\sum_{s',r} p(s',r|s,a)\Big[r+\gamma\widetilde{v}_*(s')\Big]\end{aligned}$$

Monte Carlo Control without Exploring Starts

رویکرد دوم برای اثبات:

انتقال اتفاقی بودن پالیسی به محیط

$\tilde{V}_*$: تابع ارزش برای محیط جدید

در شرایط همگرایی:

$$\begin{aligned} v_\pi(s) &= (1-\varepsilon)\max_a q_\pi(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a q_\pi(s,a) \\ &= (1-\varepsilon)\max_a \sum_{s',r} p(s',r|s,a)\Big[r+\gamma v_\pi(s')\Big] \\ &\quad + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a \sum_{s',r} p(s',r|s,a)\Big[r+\gamma v_\pi(s')\Big] \end{aligned}$$

نتیجه:

$$v_\pi = \widetilde{v}_*$$

Off-Policy Prediction via Importance Sampling

الزام ضمنی:
Stochastic μ
مثال:
Epsilon Greedy

تولید اپیزودها با پالیسی μ
تخمین پالیسی π
که:

$$\mu \neq \pi$$

Target Policy: π
Behavior Policy: μ

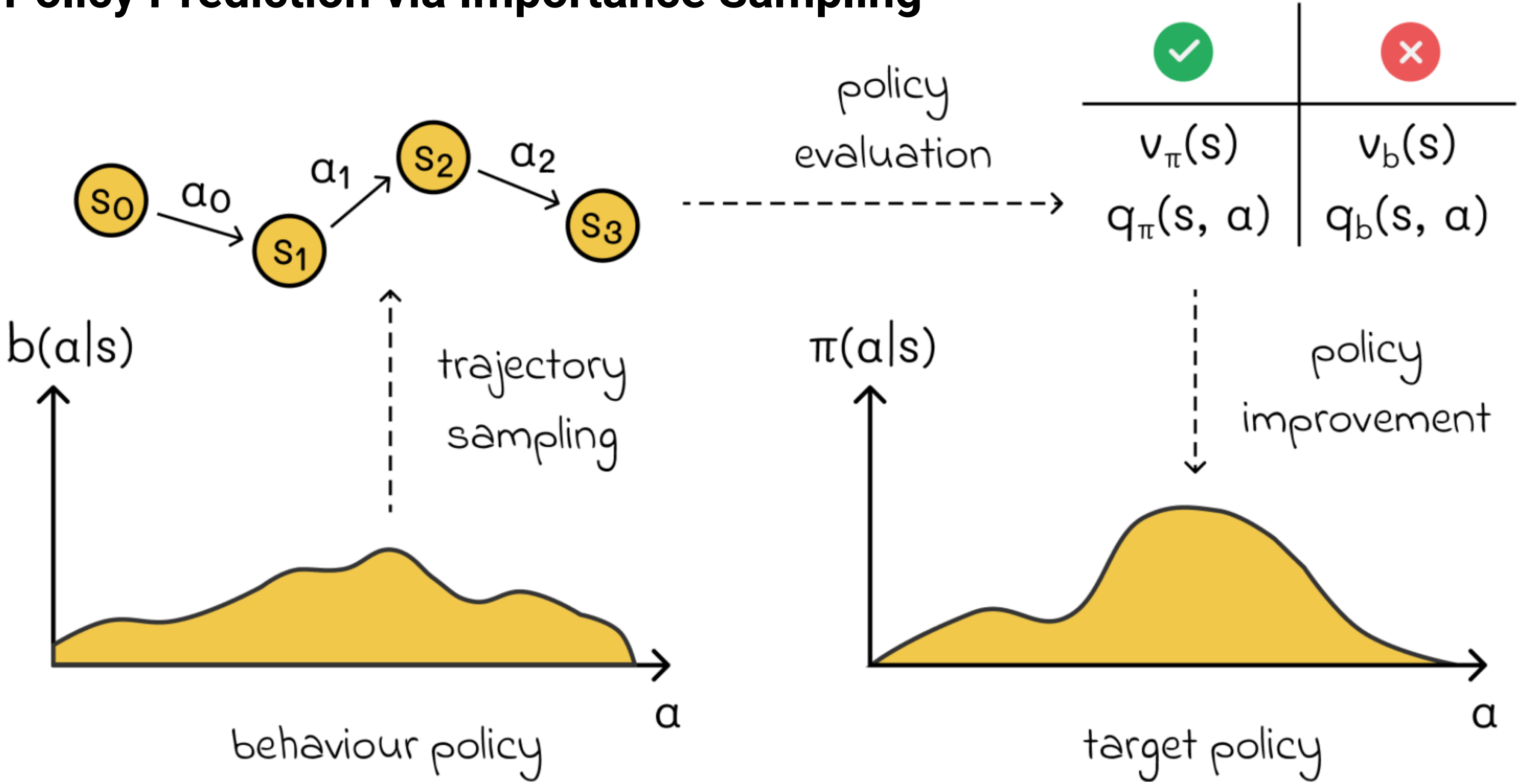
Off-policy

الزام تخمین V_π با اپیزودهای μ؟

فرض Coverage

$$\mu(a|s) > 0 \ \rightarrow \ \pi\,(a|s) > 0$$

Off-Policy Prediction via Importance Sampling



Importance Sampling

تخمین V_π با اپیزودهای μ

وزن دهی Return هایی که تحت μ به دست آمده اند، با نسبت (ρ) احتمال رخداد تراژکتوری ها تحت μ و π

Starting state S_t

احتمال وقوع s-a ها تحت پالیسی π:

$$\text{trajectory, } A_t, S_{t+1}, A_{t+1}, \ldots, S_T$$

$$\begin{aligned}
\Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T-1} \sim \pi\} \\
&= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),
\end{aligned}$$

Importance Sampling Ratio

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

First Visit – Every Visit

$$\{\rho_t^{T(t)}\}_{t\in\mathcal{T}(s)}$$

توجه: عدم وابستگی به p

$\mathcal{T}(s)$: set of all time steps in which state s is visited

Also, let T denote the first time of termination following time t.

مثال از Importance Sampling: محاسبه متوسط درآمد خانوارها

Importance Sampling Ratio

روش تخمین V_π (ordinary):

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{|\mathcal{T}(s)|}$$

روش دیگر تخمین V_π (Weighted):

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)}}$$

Q Box!

بررسی دو روش تخمین برای یک بار مشاهده؟

بررسی دو روش تخمین برای $\rho = 10$؟

بررسی واریانس:

- روش معمولی بدون حد است بخاطر بدون حد بودن واریانس ρ
- با فرض Ret محدود، واریانس تخمین محدود و 0←

Importance Sampling Ratio

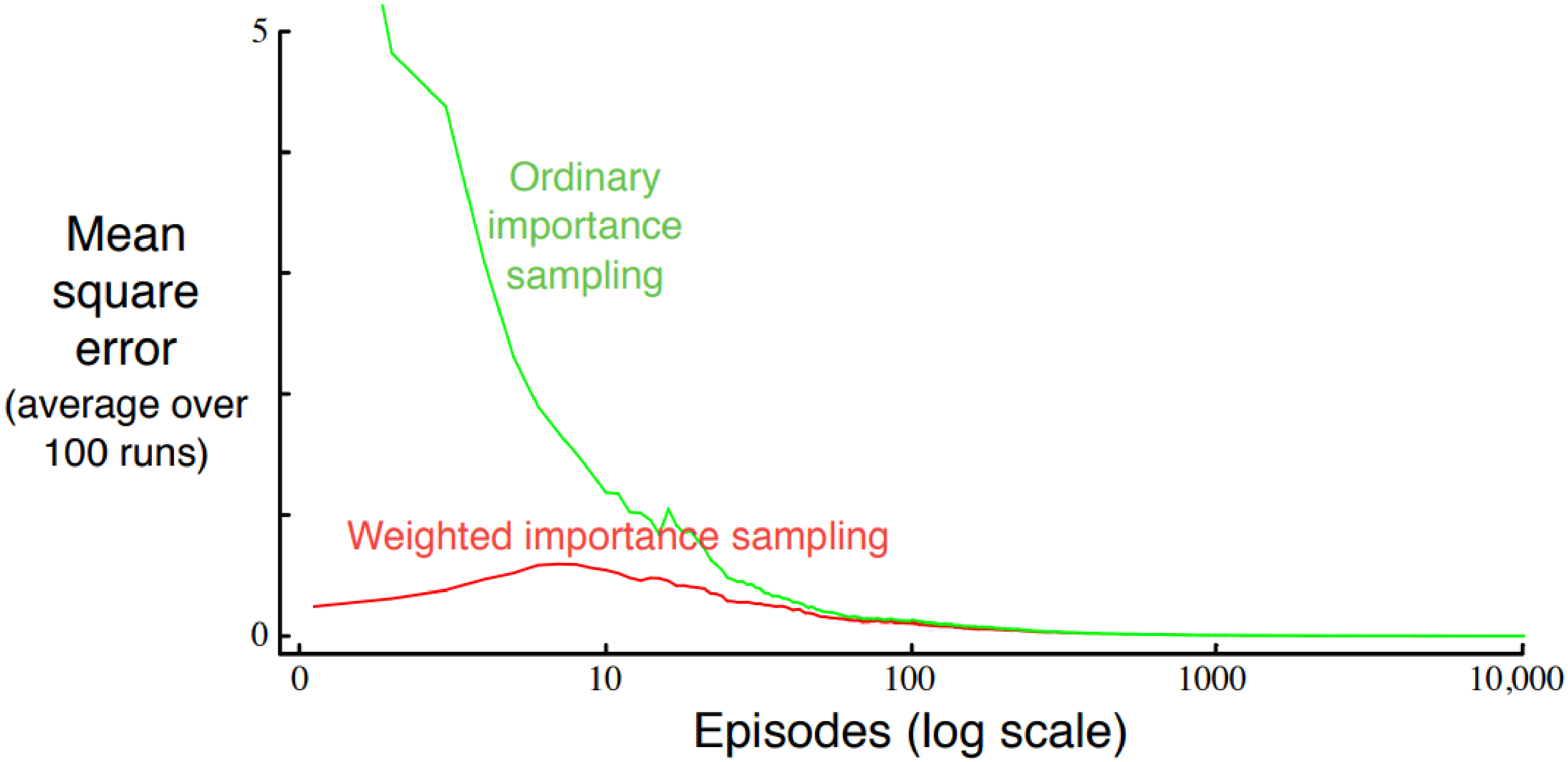
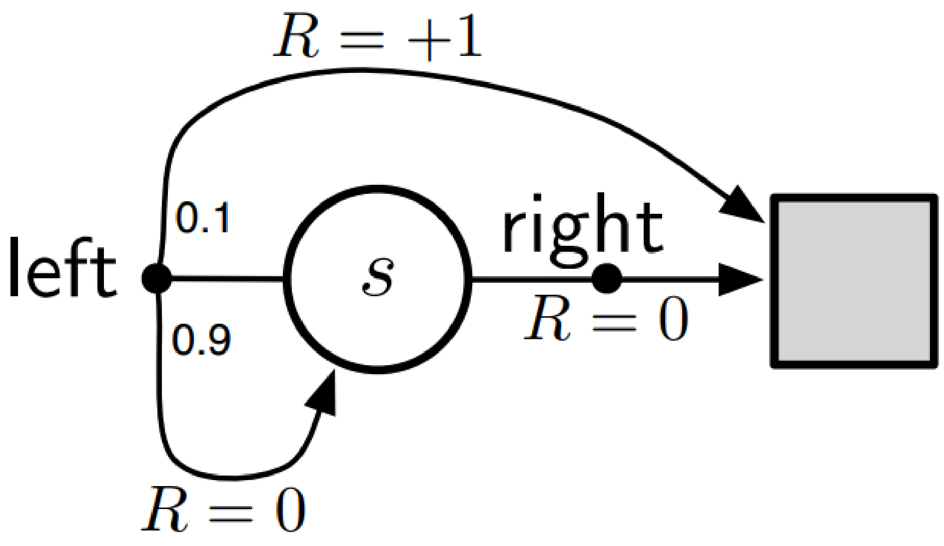


Figure 5.3: Weighted importance sampling produces lower error estimates of the value of a single blackjack state from off-policy episodes. ■

Ten Independent Runs of the First Visit MC Algorithm using ordinary importance sampling



$$\pi(\text{left}|s) = 1$$

$$b(\text{left}|s) = \frac{1}{2}$$

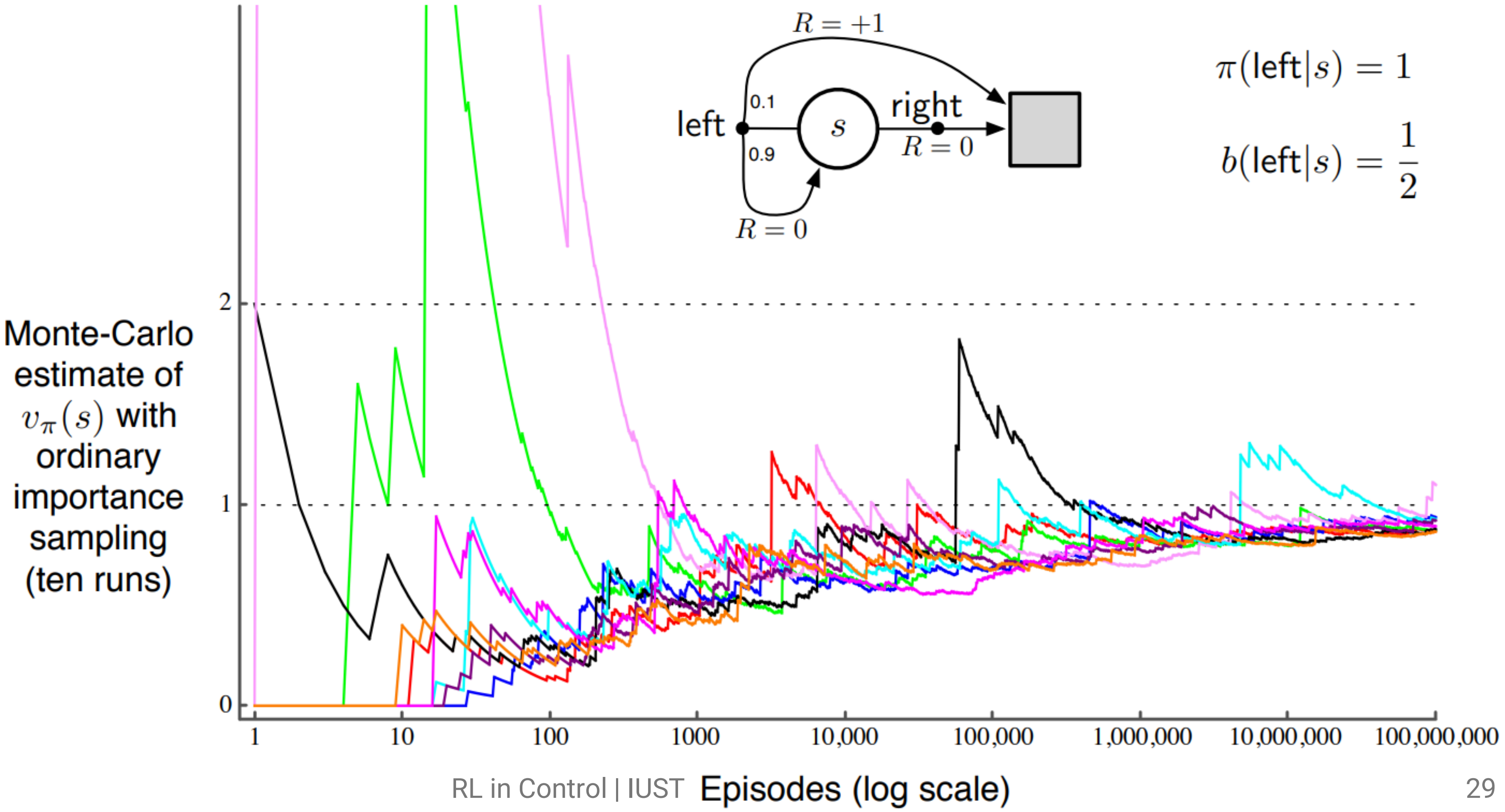
the target policy that always selects left.

behavior policy that selects right and left with equal probability

Ten Independent Runs of the First Visit MC Algorithm

variance of the importance-sampling-scaled returns is infinite

عدم همگرایی پس از 10^6 اپیزود!



Ten Independent Runs of the First Visit MC Algorithm

$$\mathrm{Var}[X] \doteq \mathbb{E}\Big[\big(X - \bar{X}\big)^2\Big] = \mathbb{E}\big[X^2 - 2X\bar{X} + \bar{X}^2\big] = \mathbb{E}\big[X^2\big] - \bar{X}^2$$

$$\mathbb{E}_b\left[\left(\prod_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{b(A_t|S_t)} G_0\right)^2\right]$$

$$\begin{aligned}
&= \frac{1}{2} \cdot 0.1 \left(\frac{1}{0.5}\right)^2 && \text{(the length 1 episode)} \\
&\quad + \frac{1}{2} \cdot 0.9 \cdot \frac{1}{2} \cdot 0.1 \left(\frac{1}{0.5}\frac{1}{0.5}\right)^2 && \text{(the length 2 episode)} \\
&\quad + \frac{1}{2} \cdot 0.9 \cdot \frac{1}{2} \cdot 0.9 \cdot \frac{1}{2} \cdot 0.1 \left(\frac{1}{0.5}\frac{1}{0.5}\frac{1}{0.5}\right)^2 && \text{(the length 3 episode)} \\
&\quad + \cdots \\
&= 0.1 \sum_{k=0}^{\infty} 0.9^k \cdot 2^k \cdot 2 = 0.2 \sum_{k=0}^{\infty} 1.8^k = \infty. && \blacksquare
\end{aligned}$$

Incremental Implementation

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \qquad n \geq 2,$$

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big], \qquad n \geq 1,$$

and

$$C_{n+1} \doteq C_n + W_{n+1},$$

الگوریتم

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$

Loop forever (for each episode):
    $b \leftarrow$ any policy with coverage of π
    Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

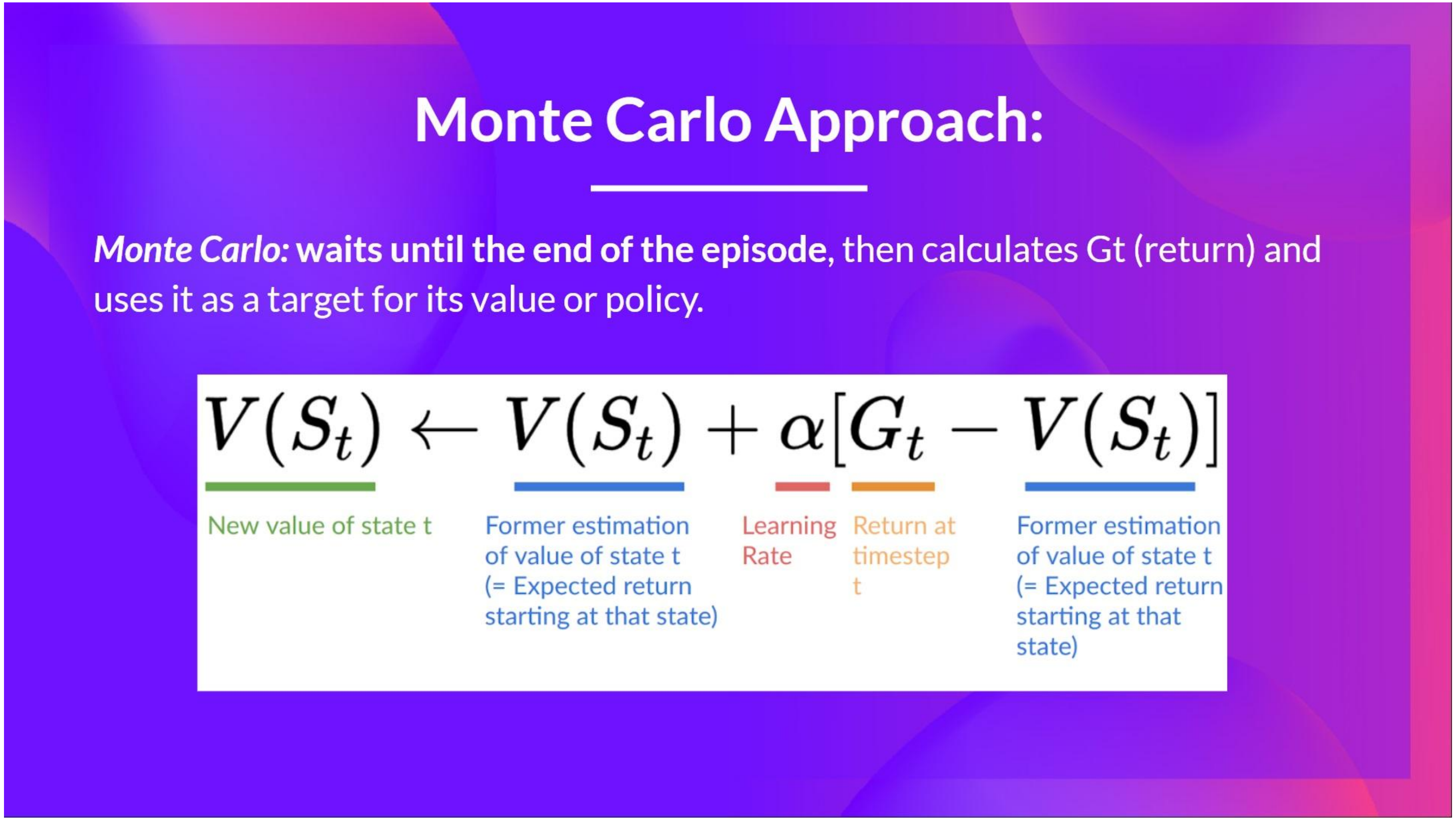
الگوریتم

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

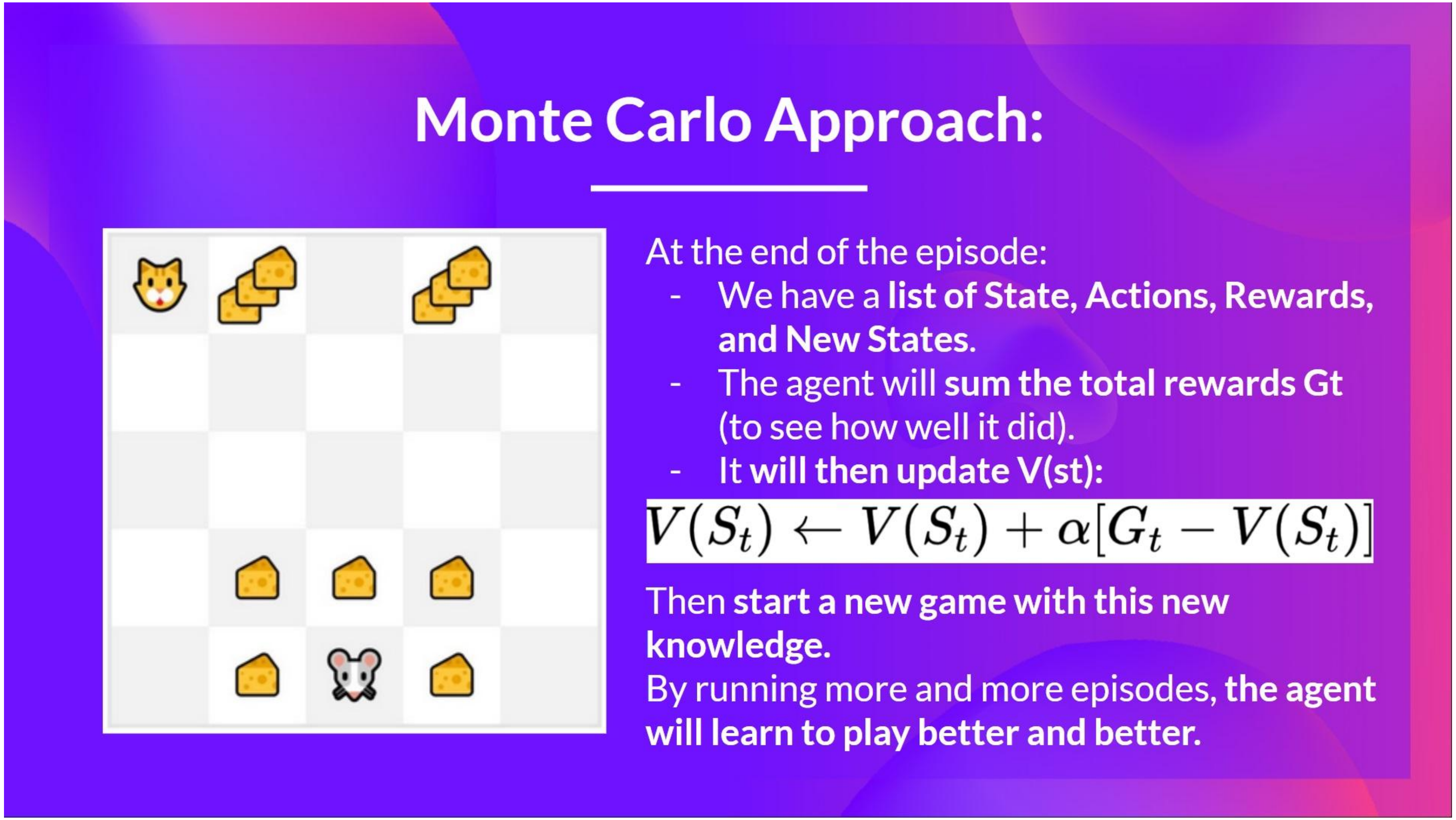
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
- $Q(s,a) \in \mathbb{R}$ (arbitrarily)
- $C(s,a) \leftarrow 0$
- $\pi(s) \leftarrow \arg\max_a Q(s,a)$ (with ties broken consistently)

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

جمع‌بندی ...


Monte Carlo Approach:
Monte Carlo: waits until the end of the episode, then calculates Gt (return) and uses it as a target for its value or policy.
$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$
New value of state t
Former estimation of value of state t (= Expected return starting at that state)
Learning Rate
Return at timestep t
Former estimation of value of state t (= Expected return starting at that state)

جمع‌بندی ...


Monte Carlo Approach:
At the end of the episode:
- We have a list of State, Actions, Rewards, and New States.
- The agent will sum the total rewards Gt (to see how well it did).
- It will then update V(st):
$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$
Then start a new game with this new knowledge.
By running more and more episodes, the agent will learn to play better and better.