Reinforcement Learning

Presença

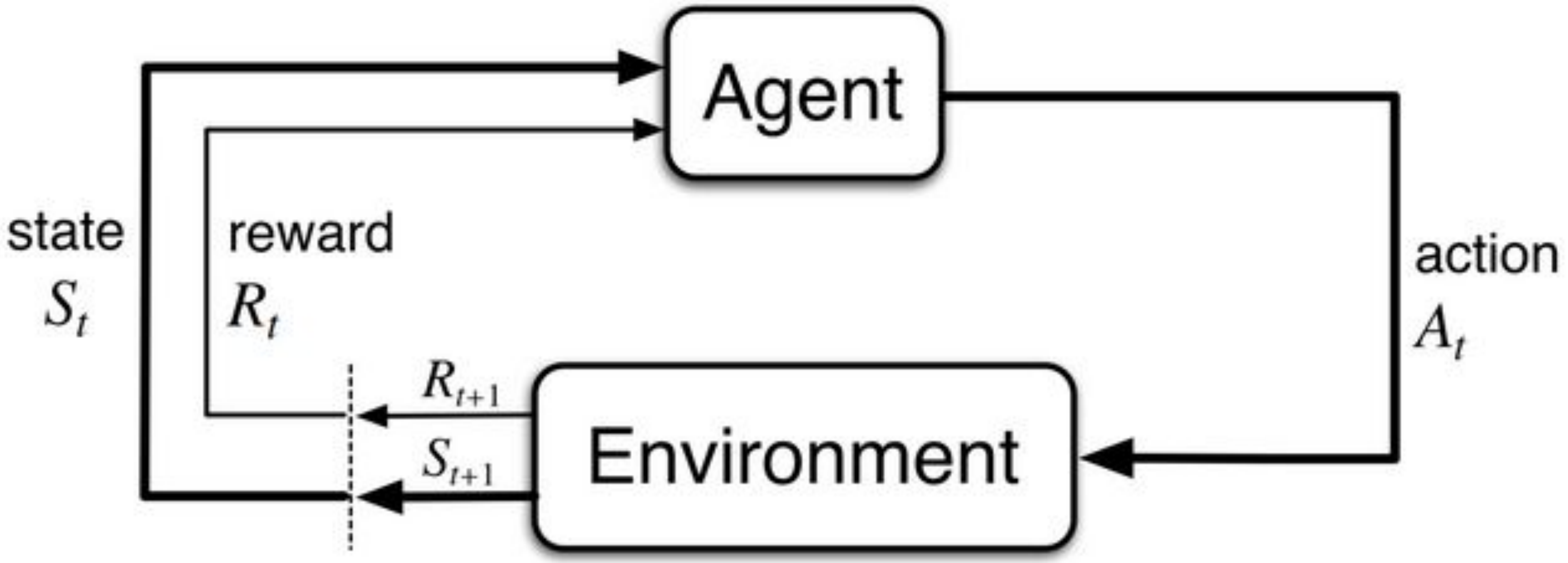
- Linktree: Presente na bio do nosso instagram
- Presença ficará disponível até 1 hora antes da próxima aula
- É necessário 70% de presença para obter o certificado

Presença

Revisão

Recapitulando

Diagrama básico do ciclo Agente-Ambiente:



- $t \in \{1, 2, 3, ..\}$
- $s \in S$
- $a \in A(s)$
- $r \in \mathbb{R}, r \neq \infty$

Recapitulando

Quatro principais elementos:

- Policy
- Recompensa
- Função valor
- Modelo

$$G_t = R_1 + R_2 + R_3 + ... + R_T = \sum_{i=1}^{T} R_i$$

$$G_t = \gamma^0 R_1 + \gamma^1 R_2 + \gamma^2 R_3 + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, 0 < \gamma < 1$$

Recapitulando

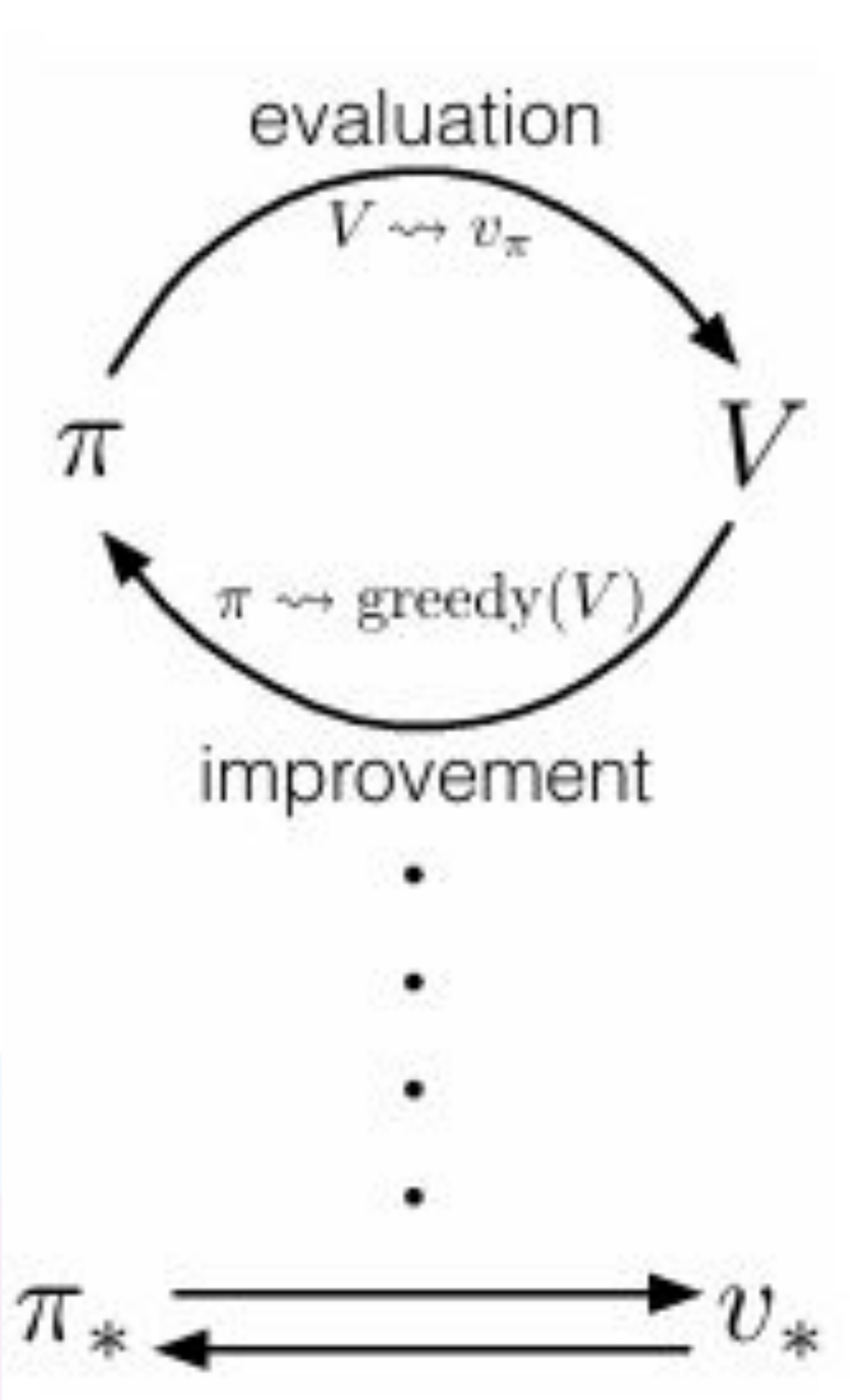
1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   - $\Delta \leftarrow 0$
   - Loop for each $s \in \mathcal{S}$:
     - $v \leftarrow V(s)$
     - $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s)) \big[r + \gamma V(s') \big]$
     - $\Delta \leftarrow \max(\Delta, |v - V(s)|)$

   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   - *old-action* $\leftarrow \pi(s)$
   - $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \mid s, a) \big[r + \gamma V(s') \big]$
   - If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*

   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2



Métodos Monte Carlo

- Estimar valores teóricos através de simulações/experiência
- Em Reinforcement Learning:
  - Simulação do Agente no Ambiente seguindo uma política
  - Papel: policy-evaluation (no contexto do GPI)
- Alternativa ao método DP

Métodos Monte Carlo

- Três principais vantagens sobre DP:
  - Aprendizado por experiência/sem modelo
  - Foco em estados frequentes
  - Facilidade em simular o episódio

Monte Carlo Prediction

- Exemplo de como funciona na prática uma estimativa com MC
  - estimaremos $v_\pi(s)$
  - na prática, é melhor estimar $q_\pi(s, a)$ (s/modelo)
- Lembrando que um modelo nada mais é que termos disponível $p(s', r|a, s)$

Monte Carlo Prediction

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 Prediction

Exemplo:

Episódio 1 = [(A, w, 10), (B, x, -5), (C, y, 15), (D ,z, 20)]

$G_3 = 20$	Returns(A) = [40]	V(A) = 40
$G_2 = 15 + 20$	Returns(B) = [30]	V(B) = 30
$G_1 = -5 + 15 + 20$	Returns(C) = [35]	V(C) = 35
$G_0 = 10 - 5 + 15 + 20$	Returns(D) = [20]	V(D) = 20

Monte Carlo Prediction

Exemplo:

Episódio 2 = [(B, y, -25), (D, y, 20), (A, x, -10), (C ,w, 5)]

$G_3 = 5$	Returns(A) = [40, -5]	V(A) = 17,5
$G_2 = 5 - 10$	Returns(B) = [30, -10]	V(B) = 10
$G_1 = 20 + 5 - 10$	Returns(C) = [35, 5]	V(C) = 20
$G_0 = -25 + 20 + 5 - 10$	Returns(D) = [20, 15]	V(D) = 17,5

Estimando valores de ação

- Apenas $v^\pi(s)$ não é suficiente para escolher ações
- Definição de valor de ação:

$$q^\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

- Política escolhe melhor ação em cada estado por $Q(s,a)$

Estimando valores de ação

- Analogamente aos valores de estado, mas usando pares (s,a)
- Métodos principais:
  - First-visit: média dos retornos após a primeira ocorrência
  - Every-visit: média dos retornos após todas as ocorrências

- Ambas convergem se cada (s,a) for visitado infinitas vezes

Exploração

- Problema: políticas determinísticas exploram apenas uma ação por estado.
- Duas soluções:
  - Exploring starts: iniciar episódios em todo (s,a) com probabilidade > 0 (teórico).
  - Políticas estocásticas *soft*/ϵ-soft: todas as ações têm probabilidade positiva (prático).

Estimando valores de ação

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 $Returns(S_t)$

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

Controle com Métodos Monte Carlo (MC)

Monte Carlo Control

Objetivo: Usar a amostragem de Monte Carlo para encontrar a **política ótima (π∗)**

Estrutura: Segue o padrão da **Iteração de Política Generalizada (GPI)**

- **Avaliação (E):** Estimar $q_\pi(s,a)$ rodando episódios e calculando a média dos retornos
- **Melhoria (I):** Tornar a política *greedy* em relação a $q_\pi(s,a)$

Grande Vantagem: É um método **livre de modelo** (*model-free*). Não precisamos conhecer a dinâmica do ambiente, pois *q(s,a)* já nos diz o valor de cada ação

Monte Carlo Control - algoritmo

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 $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Desafios Teóricos do Modelo

A abordagem "clássica" do Controle MC depende de duas suposições pouco realistas:

- **Infinitos Episódios**
  - A avaliação da política exigiria um número **infinito** de episódios para calcular o valor exato de q_π
- **Inícios Exploratórios (*Exploring Starts*)**
  - O método assume que podemos **iniciar** um episódio **a partir de qualquer par estado-ação** para garantir que tudo seja explorado

Soluções Práticas

Para o Problema Infinitos Episódios:

Não esperar a convergência! Intercalar avaliação e melhoria a cada episódio

- **Ciclo prático:**
  1. Roda-se um episódio completo
  2. Usa-se os retornos para dar um pequeno passo na **avaliação** (atualiza Q)
  3. Melhora-se a política para os estados visitados (**melhoria**)

Para o Problema *Exploring Starts*:

Uso de políticas que **garantem a exploração contínua**

Resolvendo o Problema da Exploração

Problema: Como garantir a exploração contínua sem a suposição de *inícios exploratórios*?

Abordagem On-Policy: Avaliar e melhorar a **mesma política** que o agente usa para agir e coletar experiência

Solução Prática: Política *ϵ-greedy*

- É uma política *soft* (nunca para de explorar)
- **Com probabilidade $1-\epsilon$ (Exploit):** Age de forma *greedy*, escolhendo a melhor ação conhecida
- **Com probabilidade ϵ (Explore):** Ignora o que sabe e escolhe uma ação **aleatória**

GPI On-Policy e o Trade-off

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

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

GPI On-Policy e o Trade-off

Adaptação do GPI:

- **Avaliação:** Continua igual (estima-se *q(s,a)* para a política *ϵ-greedy*)
- **Melhoria (A Mudança):** A nova política não se torna 100% *greedy*. Em vez disso, ela se torna *ϵ-greedy* em relação aos novos valores de *q*

O Trade-off Final:

- **Vantagem:** Elimina a necessidade de *inícios exploratórios*
- **Desvantagem:** O algoritmo **não converge** para a política ótima absoluta π_*
- **Convergência:** Encontra a melhor política possível **que ainda é ϵ-soft**

Predição Off-Policy

Até agora usávamos a abordagem de aprender com base em uma política que se aproximava da ótima. Na predição off-policy vamos utilizar **duas políticas distintas**:

- **Política alvo (π):** política que está sendo aprendida pelo agente (ganancioso/*greedy*, sem exploração).
- **Política de Comportamento (*b*)**: política que o agente usa para explorar o mundo, ela é estocástica e exploratória (ex: *ϵ-greedy).*

Objetivo: estimar a função valor $v_\pi(s)$ ou $q_\pi(s,a)$ da política alvo π, usando retornos G_t que foram obtidos utilizando a política ***b***.

On-policy é um caso especial da off-policy, onde π e ***b*** são as mesmas!

Predição Off-Policy

Condição Essencial: Cobertura (Coverage)

- Para que possamos aprender sobre a política alvo π a partir de **b**, precisamos garantir que **b** explore o suficiente!
- A política de comportamento **b não pode evitar completamente** ações que a política alvo π tomaria.

Suposição de Cobertura: Toda ação que π poderia tomar em um estado **s** deve ter uma probabilidade de ser escolhida por **b** naquele mesmo estado.

Importance Sampling

É uma técnica para estimar valores de uma distribuição utilizando amostras de outras.

Importance-Sampling Ratio (ρ): é a razão entre a probabilidade de uma trajetória (A_t, S_{t+1}, A_{t+1},..., S_T) ocorrer sob a política alvo $\boldsymbol{\pi}$ e a probabilidade dessa mesma trajetória ocorrer sob a política de comportamento ***b***.

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

Importance Sampling

O que significa o valor de ρ?

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

•ρ *> 1: **Mais provável** de acontecer com a política alvo. A experiência é relevante.*

•ρ *< 1: **Menos provável** de acontecer com a política alvo.*

•ρ *= 0:Política alvo **nunca** teria feito essa trajetória. Experiência irrelevante para* π.

Vantagem: As probabilidades do ambiente se cancelam. Método livre de modelo (model-free).

Importance Sampling (Analogia)

- **Problema:** saber o valor da nossa política alvo ($\boldsymbol{\pi}$), mas só temos dados de uma política de comportamento (b).
- **Analogia:** Como descobrir a altura média dos brasileiros (alvo) usando apenas uma amostra de jogadores de basquete (comportamento)?
- **Solução:** Dar um peso menor aos jogadores mais altos (mais comuns na amostra do que no alvo) e peso maior aos mais baixos.
- **Importance-Sampling Ratio (ρ)**: “fator de correção” para dizer quão mais provável (ou improvável) era uma experiência ter acontecido sob a política alvo.

Importance Sampling (Exemplo)

- Agente no estado ***s***
- Apenas duas ações a serem tomadas: $\boldsymbol{a_1}$ e $\boldsymbol{a_2}$
- **Política $\boldsymbol{\pi}$:** determinística e *greedy* escolhe de acordo com as probabilidades:
  - $\boldsymbol{\pi(a_1)} = 1$
  - $\boldsymbol{\pi(a_2)} = 0$
- Política *b*: *ϵ-greedy* permitindo exploração de acordo com as probabilidades
  - $b(a_1) = 0.8$
  - $b(a_2) = 0.2$

Para ação a_1:

$$\rho = \boldsymbol{\pi(a_1)}/b(a_1) = 1.0/0.8 = 1.25$$

Ou seja, a ação é relevante, pois a política alvo era mais provável de ter escolhido ela, e seu peso aumentado em 25%.

Para ação a_2:

$$\rho = \boldsymbol{\pi(a_2)}/b(a_2) = 0.0/0.2 = 0.0$$

A política alvo jamais escolheria a ação, então ela se torna irrelevante e seu peso sendo 0.

Corrigindo o valor esperado

Com o Importance Sampling Ratio corrigimos o retorno esperado da política ***b*** para o retorno esperado da política $\boldsymbol{\pi}$

$$\mathbb{E}[G_t|S_t = s] = v_b(s) \longrightarrow \mathbb{E}[\rho_{t:T-1}G_t \mid S_t = s] = v_\pi(s)$$

Estimando a função valor de estado

Existem duas formas de estimarmos $v_\pi(s)$.

- ***Ordinary Importance Sampling***: escala cada retorno com seu ρ.

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$$

$\mathcal{T}(s)$ é o conjunto de visitas ao estado s e $|\mathcal{T}(s)|$ é o número de vezes que ele foi visitado

Vantagem: é **não enviesado**. Em média o valor estimado converge para $v_\pi(s)$

Desvantagem: possui **alta variância.** Caso ρ seja muito alto, pode levar a estimativas muito altas.

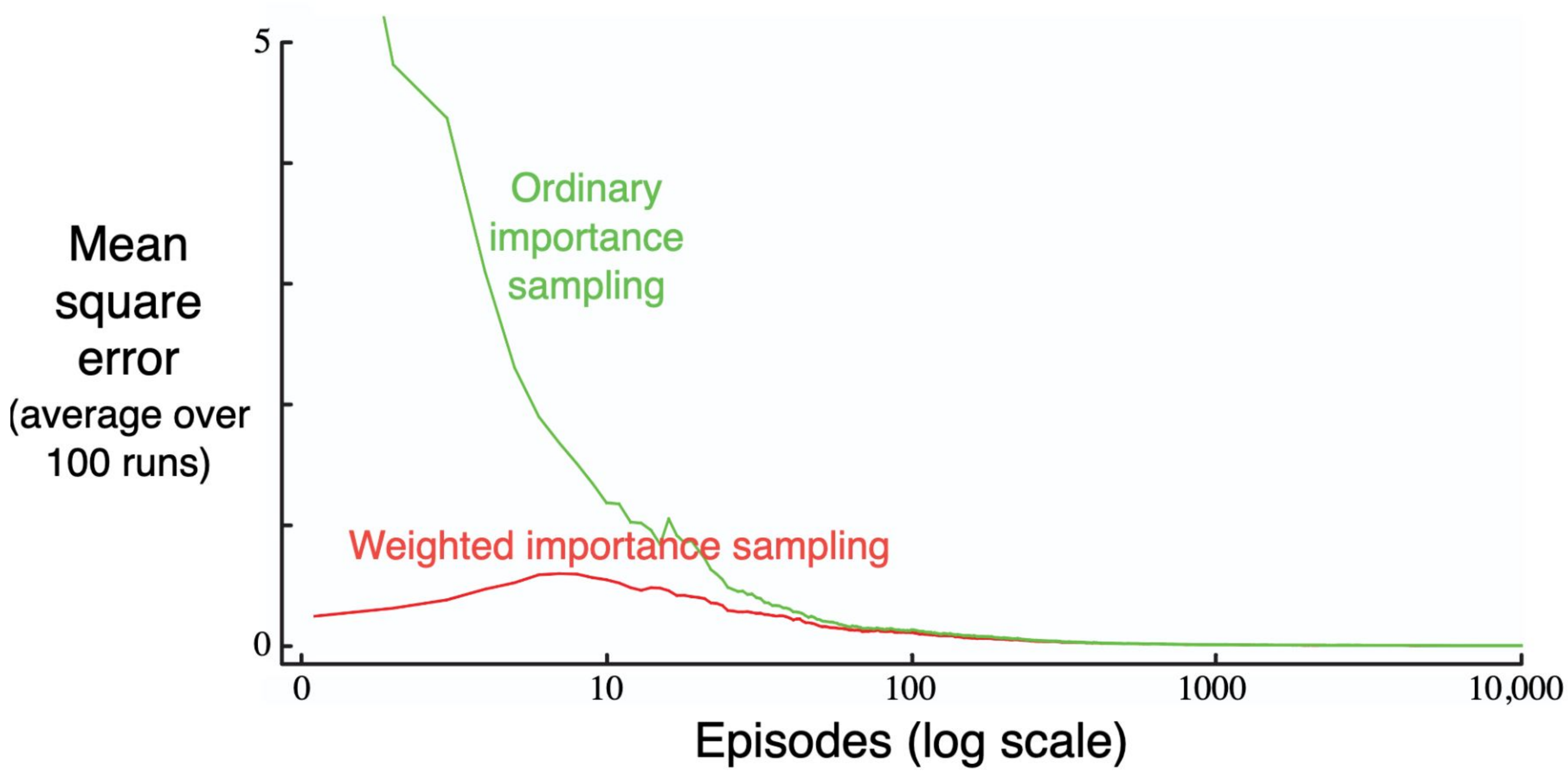
- ***Weighted Importance Sampling***: realiza uma média ponderada dos retornos.

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Vantagem: Possui uma variância muito menor que a anterior. Na prática, converge muito mais rápido e de forma confiável, sendo escolha padrão.

Desvantagem: é enviesado. Entretanto o viés desaparece a medida que mais dados são coletados.

Estimando a função valor de estado



Abordagem incremental

Como atualizar médias de forma eficiente?

- **Problema:** Ineficiência da média completa. A cada etapa, recalcular todos os retornos é muito caro computacionalmente.
- **Solução proposta:** Ajustar a estimativa antiga com base na nova informação obtida.
- **Como funciona:** Mantemos a soma acumulada dos pesos (C) e atualizamos a média (V) a cada passo.

Abordagem incremental

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

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

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

Estimando a função valor de ação

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 $W \neq 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

Controle MC off policy

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

data@icmc.usp.br

@data.icmc

/c/DataICMC

/icmc-data

data.icmc.usp.br

obrigado!