Aprendizaje Reforzado

Maestría en Data Mining, Universidad Austral

Javier Kreiner

Plan de la clase

Teoría:

- Exploración vs Explotación
- Bandidos Multi-Brazo
- 10-armed Test Bed
- Métodos: ε-greedy, optimista, cota superior de confianza
- Cadenas de Markov
- Procesos de Decisión de Markov
- Procesos de Recompensa de Markov
- Función de Valor
- Función de Acción-Valor
- Ecuación de Esperanza de Bellman
- Ecuación de Optimalidad de Bellman
- Evaluación de una política

Exploración vs Explotación

- **Explotación**: Tomar la acción que es **más conveniente** *en el momento*.
- **Exploración**: Tomar **decisiones sub-óptimas** con el propósito de *obtener más información.*

Ejemplos:

¿Qué publicidad muestro? ¿Dónde realizó pozos de petróleo? ¿Qué restorán elijo?

Bandidos Multi-brazo

Sólo hay acciones y recompensas.

$$A_t \longrightarrow R_t$$

$$q_*(a) = E\left[R_t \mid A_t = a\right]$$

$$a = 1, \ldots, k$$

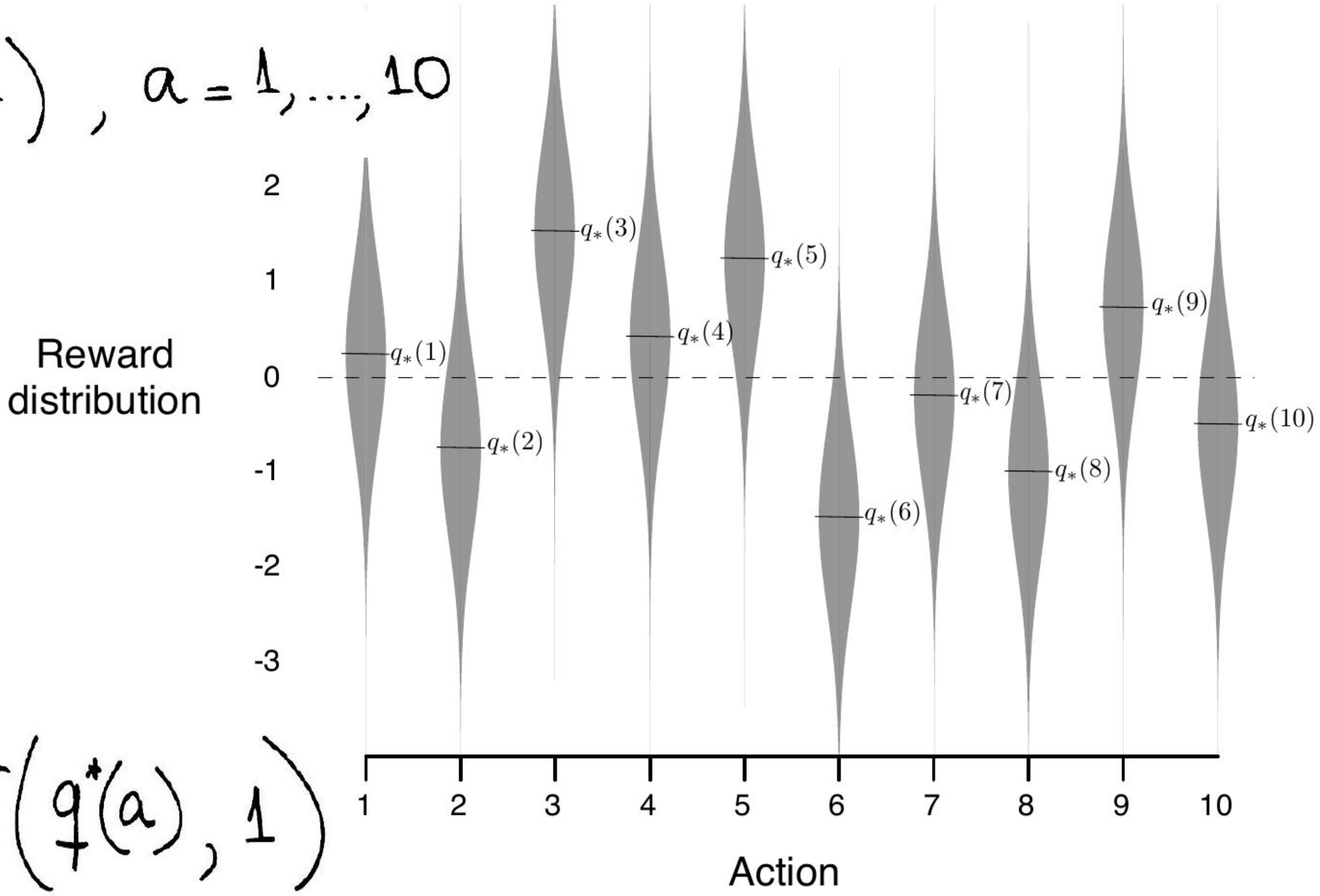
Métodos basados en la función de valor

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{N_t(a)}$$

$$A_t = \underset{a}{\operatorname{argmax}}\; Q_t(a)$$

10-armed Testbed

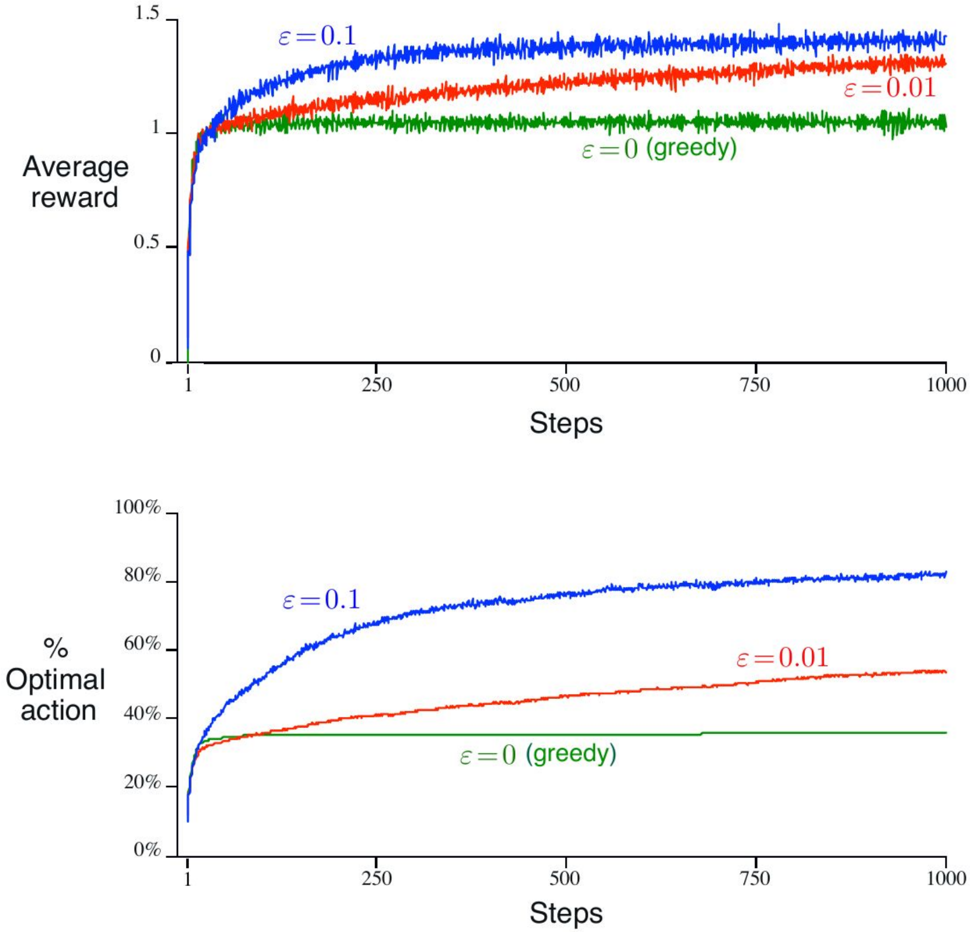
$$q^*(a) \sim \mathcal{N}(0, 1), \quad a = 1, \ldots, 10$$



$$R_t \mid A_t = a \sim \mathcal{N}(q^*(a), 1)$$

10-armed Testbed - e-greedy

- Dependiendo del ruído de la recompensa, se modifica el rendimiento del ε greedy.
- Inclusive, en casos donde la recompensa es determinística (en función de la acción) puede convenir ε-greedy (caso no estacionario).



Algoritmo y caso no estacionario

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

Caso no estacionario

$$Q_{n+1} \doteq Q_n + \alpha\Big[R_n - Q_n\Big] = \quad (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i.$$

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty$$

Controla la fluctuaciones del comienzo

$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty.$$

Garantiza la convergencia

Método “Optimístico”

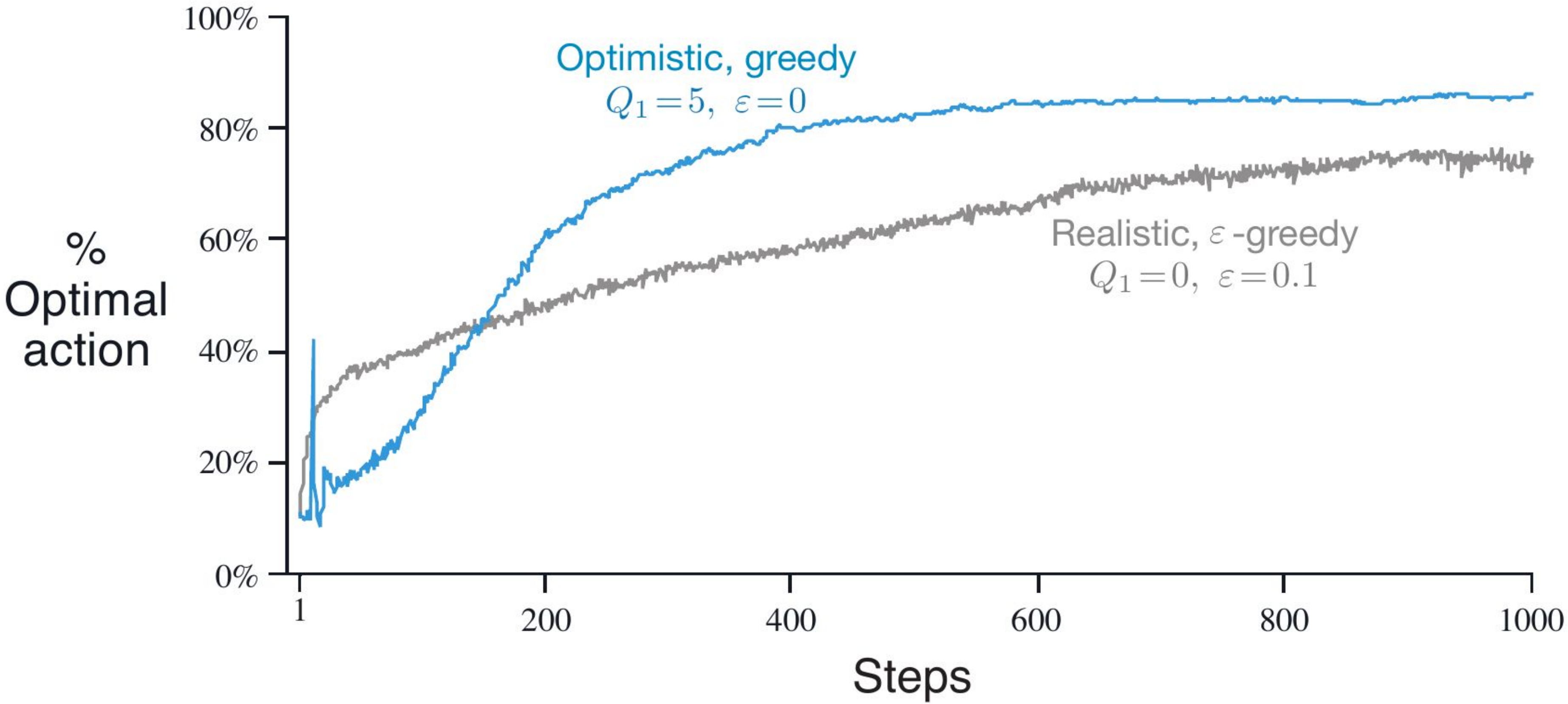
El método anterior es muy dependiente de las *condiciones iniciales*,

$$Q_1(a)$$

$$Q_2(a) = Q_1(a) + \frac{1}{N_1(a)}\left[R_1 - Q_1(a)\right]$$

Si defino todas las condiciones iniciales de manera ***optimista***, la recompensa no va a cumplir con las expectativas por lo que *muchas veces voy a cambiar de acción.*

Es decir, **voy a explorar**!



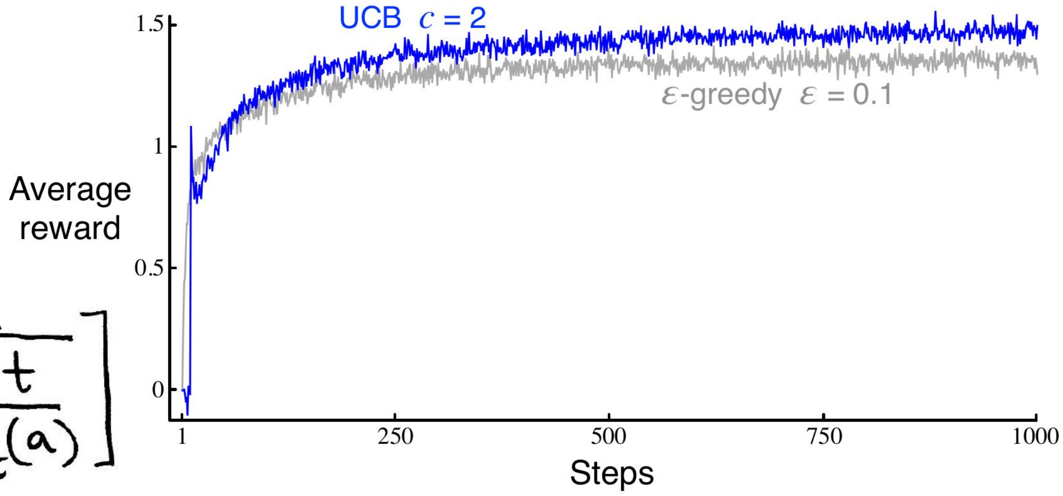
100%
80%
60%
40%
20%
0%
% Optimal action
Optimistic, greedy
$Q_1 = 5, \ \varepsilon = 0$
Realistic, ε-greedy
$Q_1 = 0, \ \varepsilon = 0.1$
1
200
400
600
800
1000
Steps

Selección de acción basada en Cota Superior de Confianza (UCB)

e-greedy selecciona las acciones no óptimas *sin utilizar **nada** de información* sobre las mismas.

Cota de la varianza del Q

$$A_t = \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$



Programación

Propiedad de Markov

$$\mathbb{P}\left[S_{t+1} \mid S_t\right] = \mathbb{P}\left[S_{t+1} \mid S_1, \ldots, S_t\right]$$

Proceso o Cadena de Markov

A Markov process is a memoryless random process, i.e. a sequence of random states $S_1, S_2, ...$ with the Markov property.

Definition

A *Markov Process* (or *Markov Chain*) is a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$

- $\mathcal{S}$ is a (finite) set of states
- $\mathcal{P}$ is a state transition probability matrix,
  $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$

Proceso de Recompensa de Markov

A Markov reward process is a Markov chain with values.

Definition

A *Markov Reward Process* is a tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{P}$ is a state transition probability matrix,
  $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}_s = \mathbb{E}[R_{t+1} \mid S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1]$

Proceso de recompensa de Markov

Función de recompensa

$$\mathcal{R}_s := E[R_{t+1} | S_t = s]$$

Retorno

$$G_t := R_{t+1} + \gamma R_{t+2} + \cdots$$

Función de Valor

$$v(s) = E[G_t | S_t = s]$$

Ecuación de Bellman

Podemos descomponer la función de valor en dos partes:

- La recompensa inmediata
- La función de valor en el próximo paso descontada

$$
\begin{aligned}
v(s) &= \mathbb{E}\left[G_t \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \dots\right) \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s\right]
\end{aligned}
$$

Ecuación de Bellman

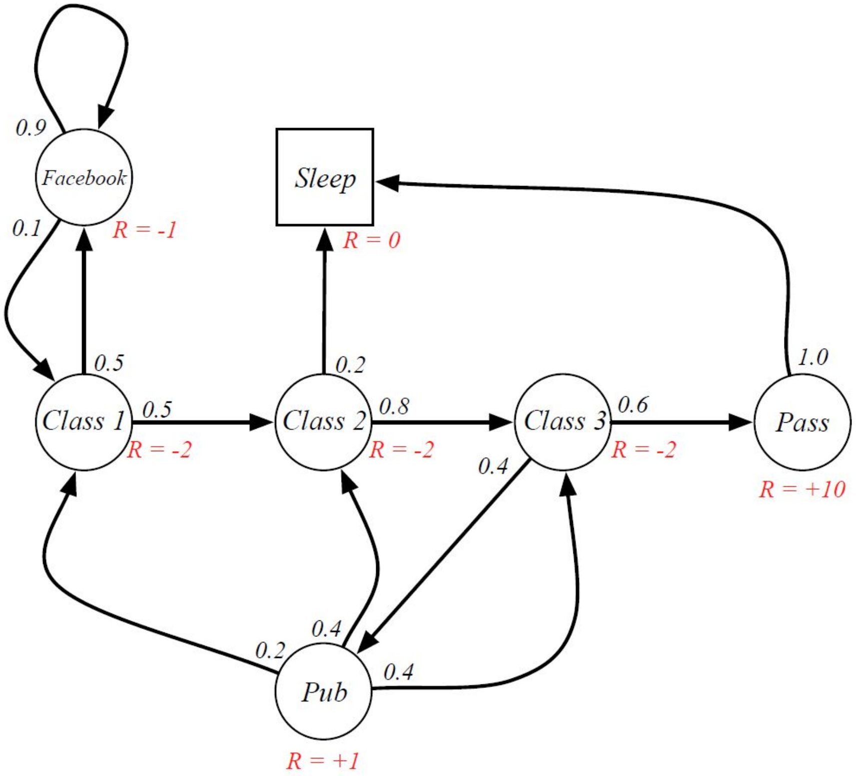
$$v(s) = \mathcal{R}_s + \gamma \sum_{s'} p_{s,s'} v(s')$$

$$v = (v(s_1), \ldots, v(s_K))$$

Un sistema de K ecuaciones!

- Con matrices: $\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \ldots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{11} & \ldots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$
- Escrito en una línea: $v = \mathcal{R} + \gamma p v$

Un ejemplo



0.9
Facebook
0.1
R = -1
Sleep
R = 0
0.5
0.2
1.0
Class 1
0.5
Class 2
0.8
Class 3
0.6
Pass
R = -2
R = -2
R = -2
0.4
R = +10
0.4
0.2
0.4
Pub
R = +1

Proceso de Decisión de Markov

Se agrega un *acción* la cual modifica el ambiente.

$$p^a_{s,s'} := P(S_{t+1} = s' | S_t = s, A_t = a)$$

$$\mathcal{R}^a_s := E[R_{t+1} | S_t = s, A_t = a]$$

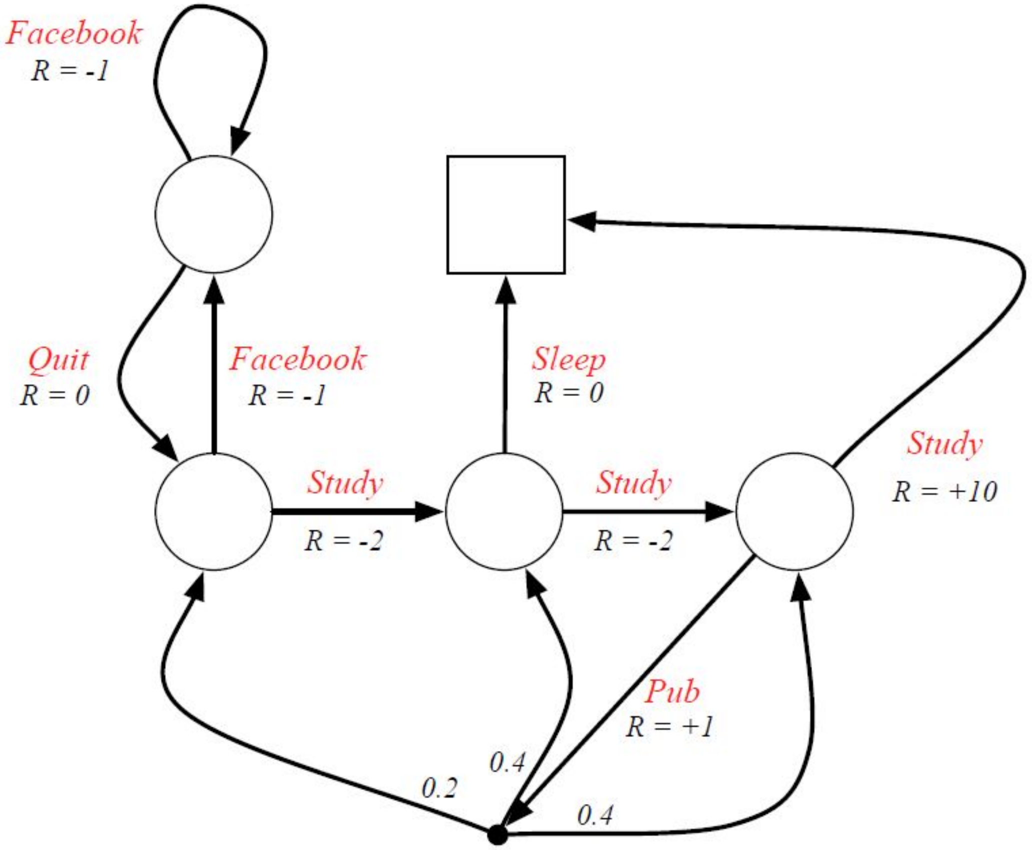
Proceso de Decisión de Markov

Definition

A *Markov Decision Process* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
- $\mathcal{P}$ is a state transition probability matrix,
  $\mathcal{P}_{ss'}^{a} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}_s^a = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$
- γ is a discount factor $\gamma \in [0, 1]$.

Ejemplo



Facebook
R = -1
Quit
R = 0
Facebook
R = -1
Sleep
R = 0
Study
R = +10
Study
R = -2
Study
R = -2
Pub
R = +1
0.4
0.2
0.4

Política

$$\pi(a|s) = \mathbb{P}\left[A_t = a \mid S_t = s\right]$$

- Es una distribución de probabilidad sobre las acciones dado un estado
- Define el comportamiento del agente
- Dado un proceso de decisión de Markov y una política fija, la secuencia

$S_1, R_2, S_2, \ldots$ es un proceso de recompensa de Markov $\langle \mathcal{S}, \mathcal{P}^{\pi}, \mathcal{R}^{\pi}, \gamma \rangle$

Donde:

$$\mathcal{P}^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^{a}_{ss'}$$

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

Funciones de Valor:

- Función de valor de estado:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

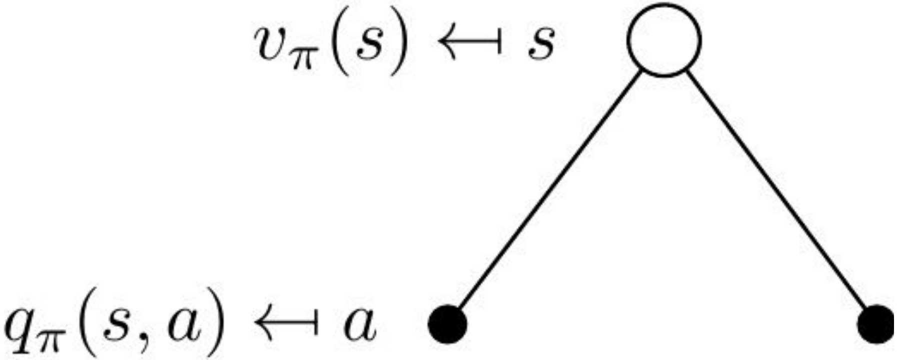
$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

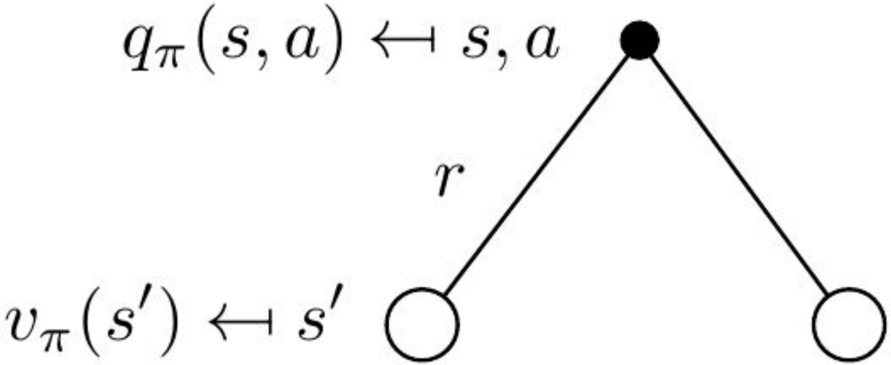
- Función de valor de estado-acción:

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

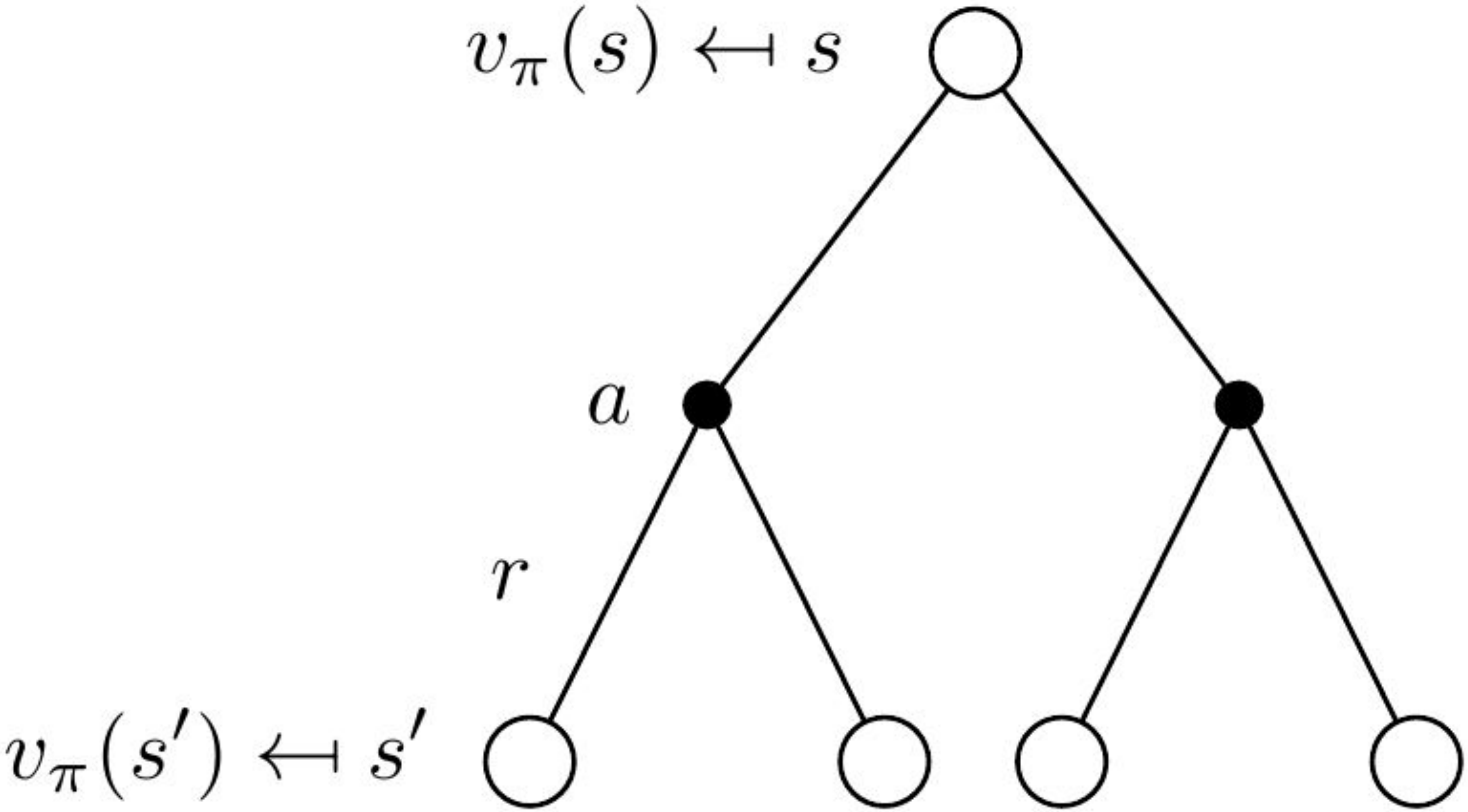
$$q_\pi(s, a) = \mathbb{E}_\pi \left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a \right]$$

Ecuación de Bellman (bis)





$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a) \qquad q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$


$v_\pi(s) \leftarrow s$
a
r
$v_\pi(s') \leftarrow s'$

Evaluación de Política

$$v_\pi(s) = \sum_a [\mathcal{R}_s^a + \sum_{s'} v_\pi(s') p_{s,s'}^a] \pi(a|s)$$

$$= \mathcal{R}_s^\pi + \sum_{s'} v_\pi(s') p_{s,s'}^\pi$$

Método Iterativo

$$v_\pi^{k+1}(s) = \mathcal{R}_s^\pi + \sum_{s'} v_\pi^{k}(s') p_{s,s'}^\pi$$

Función de Valor Óptima

$$v_*(s) = \max_\pi v_\pi(s)$$

$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

Optimalidad de MDP

$$\exists\ \pi_* \ / \pi_* \geq \pi\ \forall\ \pi \qquad \text{such that}$$

$$v_*(s) = v_{\pi_*}(s), \qquad q_*(s,a) = q_{\pi_*}(s,a) \qquad \forall\ s, a$$

$$\pi_*(s) = arg\,\mathrm{max}_a\, q_*(s,a)$$

$$v_*(s) = \mathrm{max}_a\, q_*(s,a)$$

Ecuaciones de optimalidad para MDP

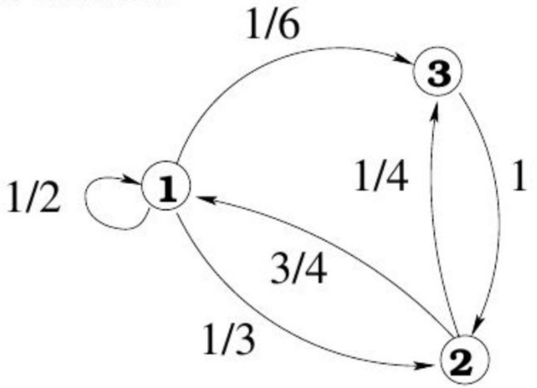
$$v_*(s) = \max_a [\mathcal{R}_s^a + \sum_{s'} p_{s,s'}^a v_*(s')]$$

Son ecuaciones NO lineales!

¿Cómo obtener v_* y π_*?

Ejercicios para entregar

1. Simular el proceso de recompensa del slide 20 y obtener el retorno esperado con γ=½
2.



$$R(1) = -2, \qquad R(2) = 3, \qquad R(3) = 5$$

$$G(S_1, S_2) = R(S_1) + R(S_2)$$

Calcular $E_{\mathbf{e1}}[G(S_1, S_2)]$ de manera analítica y por monte carlo.

Lectura recomendada:

- Introducción a Contextual Bandits: https://towardsdatascience.com/contextual-bandits-and-reinforcement-learning-6bdfeaece72a
- Capítulos 2 y 3 de Sutton and Barto
- Desafíos de Reinforcement Learning: https://www.alexirpan.com/2018/02/14/rl-hard.html