Aprendizaje Reforzado

Maestría en Ciencia de Datos, DC - UBA

Julián Martínez
Javier Kreiner

Plan de la clase

Teoría:

- Cadenas de Markov
- Procesos de Decision de Markov
- Procesos de Recompensa de Markov
- Función de Valor
- Funcion de Acción-Valor
- Ecuación de Esperanza de Bellman
- Ecuación de Optimalidad de Bellman
- Evaluación de una política
- Iteración de política
- Iteración de (función de) valor

Markov Reward Processes

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

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

Un sistema de K ecuaciones!

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

$$v = \mathcal{R} + \gamma p v$$

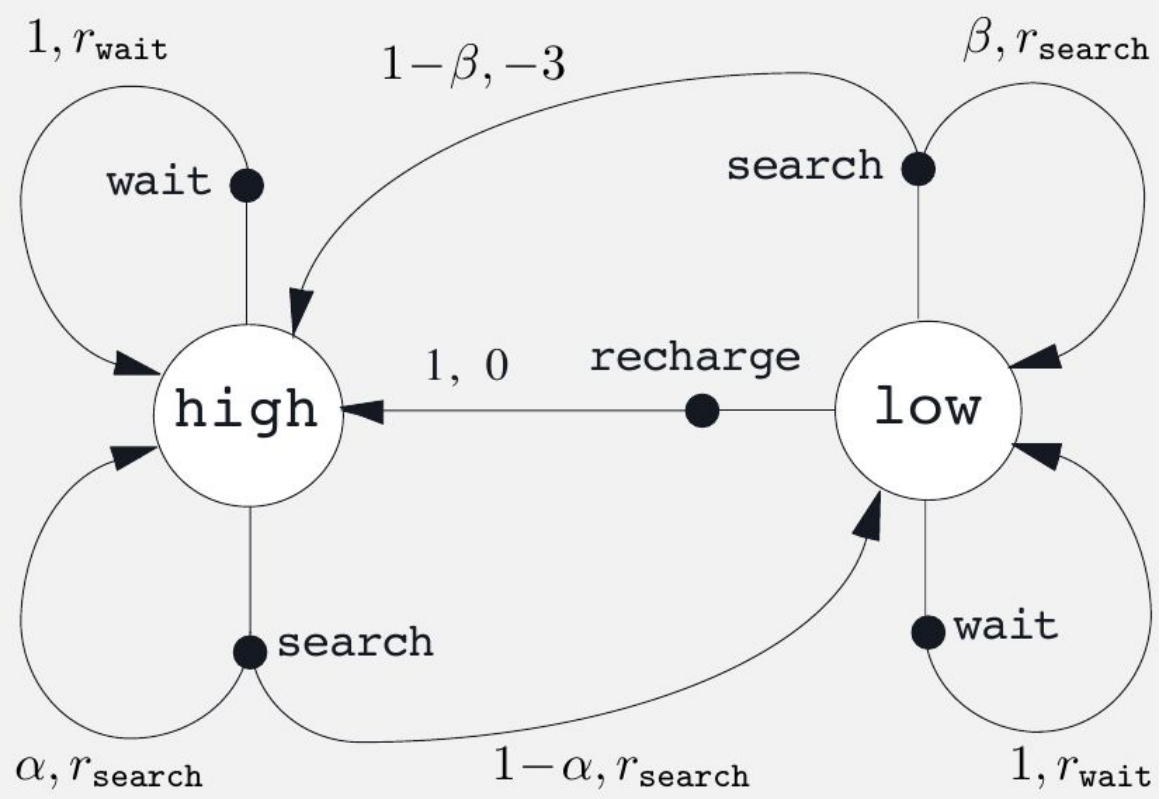
Proceso Markovianos de Decisión

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

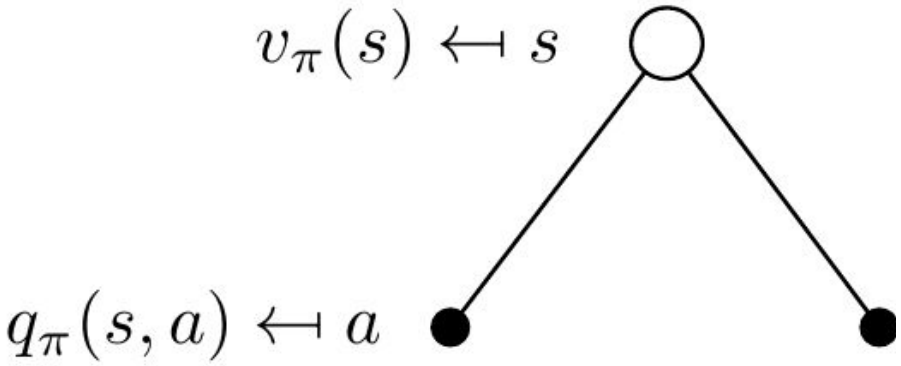
$$p^a_{s,s'} := P(S_{t+1} = s' | S_t = s, A_t = a)$$

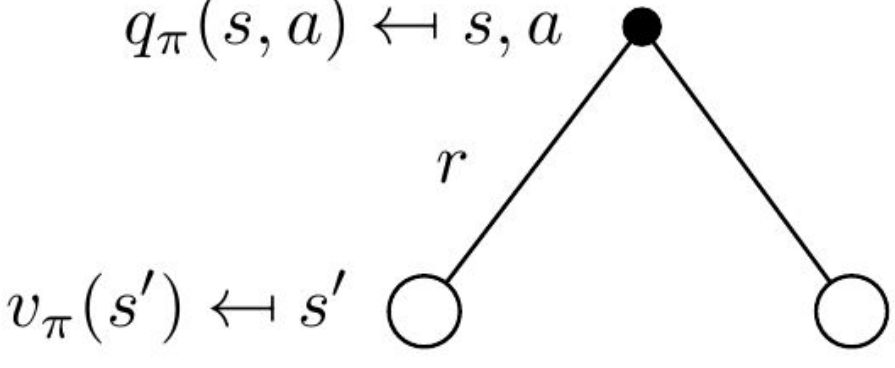
$$\mathcal{R}^a_s := E[R_{t+1} | S_t = s, A_t = a]$$

Un ejemplo


$1, r_{\texttt{wait}}$
$1-\beta, -3$
$\beta, r_{\texttt{search}}$
wait
search
high
$1,\ 0$
recharge
low
search
wait
$\alpha, r_{\texttt{search}}$
$1-\alpha, r_{\texttt{search}}$
$1, r_{\texttt{wait}}$

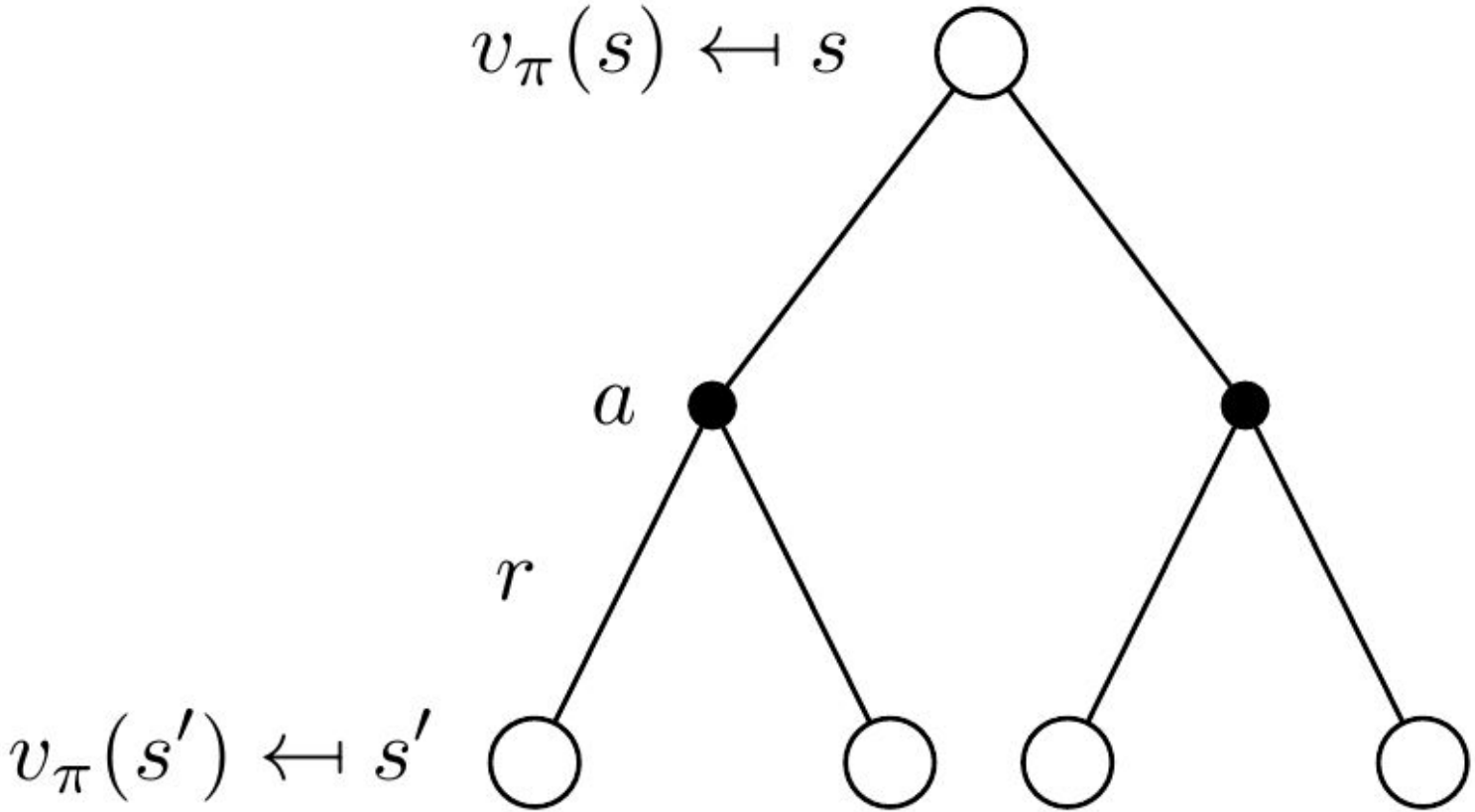
Ecuación de Bellman (bis)





$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$

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


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

Introducción a OpenAI Gym

Lectura recomendada:

- Desafíos de Reinforcement Learning: https://www.alexirpan.com/2018/02/14/rl-hard.html

Competición MIT

- Competición en MIT: https://selfdrivingcars.mit.edu/deeptraffic/