




Aprendizaje Reforzado

Maestría en Ciencia de Datos, DC - UBA

Julián Martínez
Javier Kreiner

Evaluación de Política


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

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

$\exists \pi_* / \pi_* \geq \pi \forall \pi$ such that

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

$$\pi_*(s) = \arg \max_a q_*(s, a)$$

$$v_*(s) = \max_a q_*(s, a)$$

Optimalidad de Bellman

T_{ij} = Costo de viajar de i a j

O_{ij} = Costo del viaje ÓPTIMO de i a j

$$O_{ij} = \min_k [T_{ik} + O_{kj}]$$

Ecuaciones de optimalidad para MDP



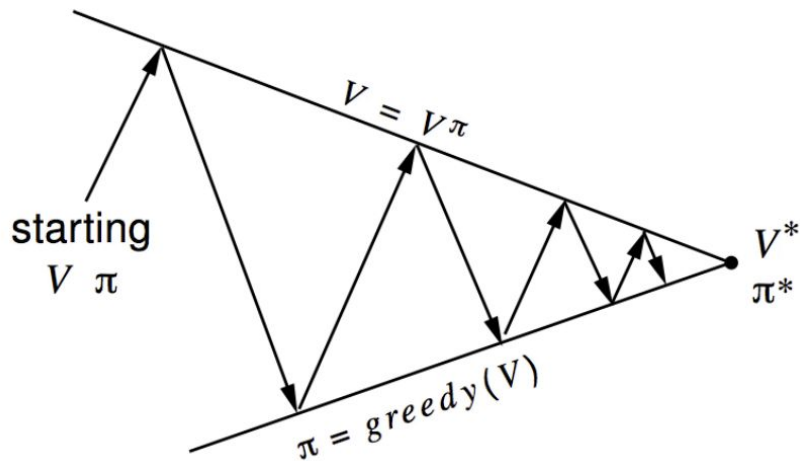
$$v_*(s) = \max_a [\mathcal{R}_s^a + \sum_{s'} p_{s,s'}^a v_*(s')]$$

Son ecuaciones NO lineales!


¿Cómo obtener v_* y π_* ?

Evaluación y Mejora

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



Evaluation / Improvement


$$\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_*,$$

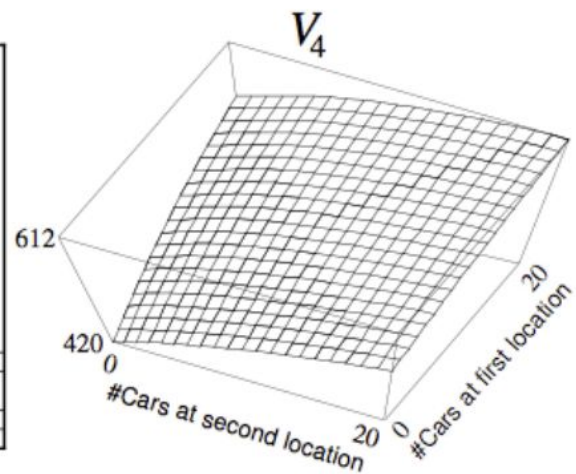
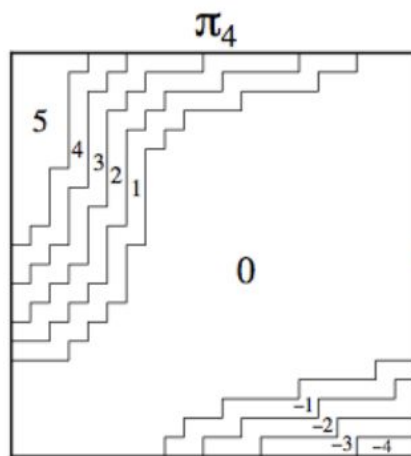
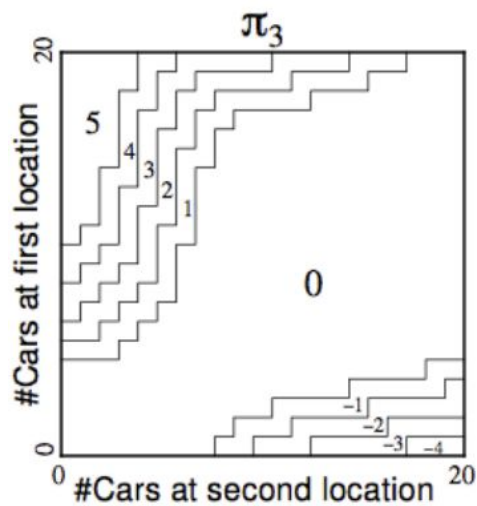
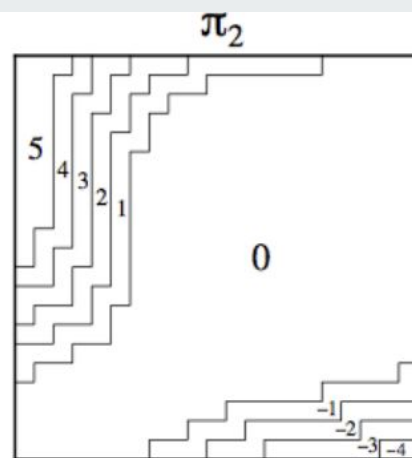
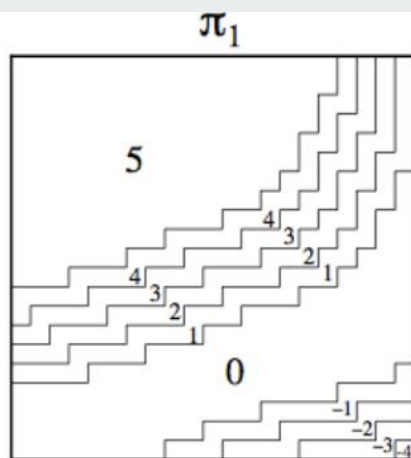
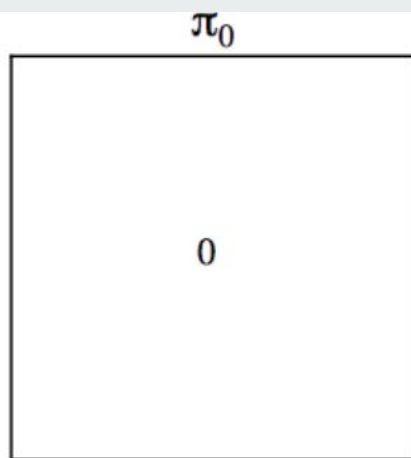
$$q_{\pi_k}(s, a) = \mathcal{R}_s^a + \gamma \sum_{s'} v_{\pi_k}(s') p_{s, s'}^a$$

$$\pi_{k+1}(s) = \arg \max_a q_{\pi_k}(s, a)$$

Alquiler de autos



- Dos terminales, A y B.
- Un máximo de 20 autos por terminal.
- Puedo mover máximo en cada noche 5 autos de una terminal a otra. Cada auto cuesta 2\$ moverlo.
- La cantidad de autos *demandados* en cada una de las terminales sigue una distribución de Poisson de medias 3 y 4 respectivamente.
- La cantidad de autos *retornados* en cada una de las terminales sigue una distribución de Poisson de medias 2 y 3 respectivamente.
- Cada auto alquilado da una ganancia de 10\$.
- *Si alguna de las dos terminales se queda sin autos se acaba el negocio.*



Dos pasos en uno



$$v_*^{k+1}(s) = \max_a [\mathcal{R}_s^a + \sum_{s'} p_{s,s'}^a v_*^k(s')]$$

Otras variantes:

- Actualizar un sólo estado en cada iteración evaluation / improvement.
- Actualizar algunos estados en evaluation y otros en improvement.
- No actualizar los estados que sean poco probables.



Plan de la case

Práctica: (python)

- Evaluación de una política
- Iteración de política
- Iteración de Valor



Ejercicio

- Problema del apostador/Calcular la política óptima para el problema de la batería
- Ejercicio matemática



Lectura recomendada

- AlphaStar de deepmind le gana a profesionales del Starcraft 2:
<https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>
- hilo de twitter con aplicaciones de RL:
<https://twitter.com/jackclarkSF/status/919584404472602624>