Aprendizaje Reforzado

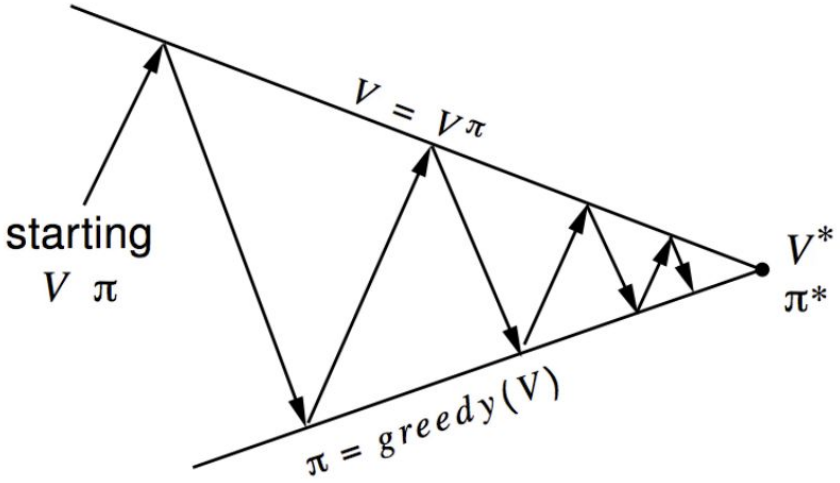
Maestría en Ciencia de Datos, DC - UBA

Julián Martínez
Javier Kreiner

Monte Carlo y Diferencias Temporales (programación)

- Ejemplo de Blackjack
- Predicción Monte Carlo
- Predicción TD
- Control Monte Carlo on-policy con políticas epsilon greedy

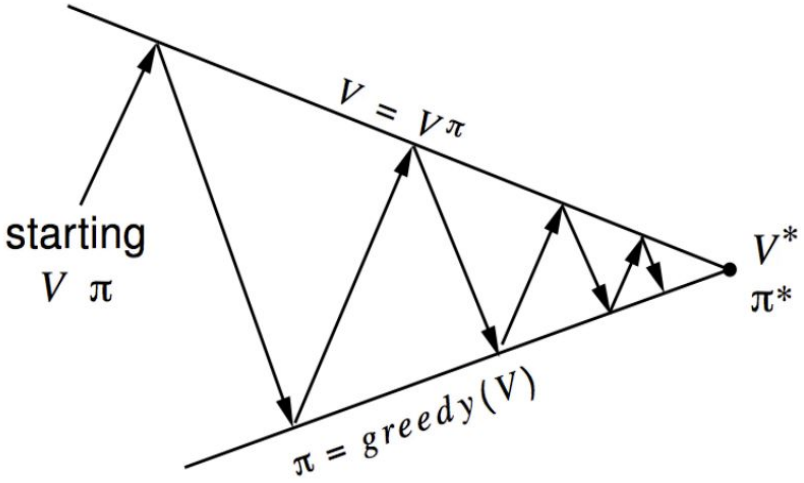
Control (Improvement) - Monte Carlo



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

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

Control (Improvement) - Monte Carlo

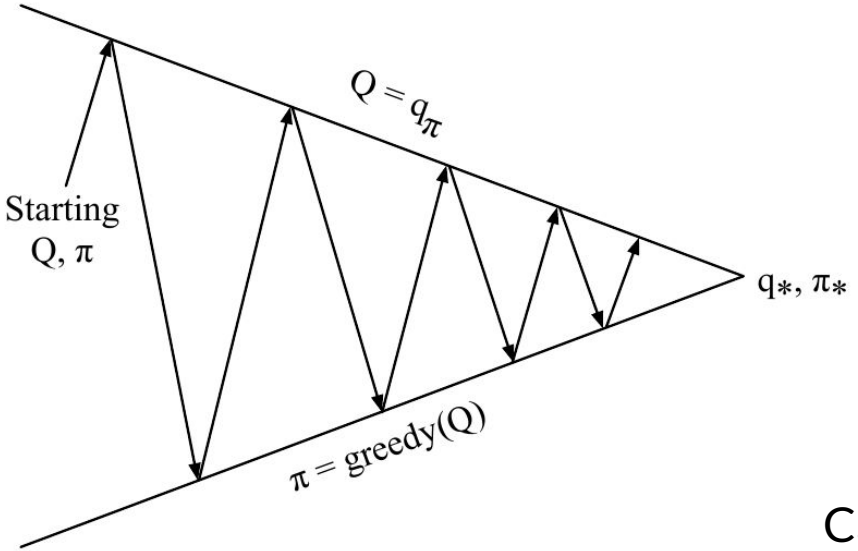


- No tengo como dato la matriz de transición. (*Model Free*)
- No tengo la esperanza *exacta*, donde están *todos* los posibles escenarios. (*Exploration vs. Explotation*)

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

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

Los parches - Dependiente del modelo (p)



Hacer evaluación MC de

$$q_{\pi_k}(s,a) =: Q_k(s,a)$$

Cambio la esperanza que aproximo.

Probar un poco todo (epsilon - greedy policy)

$$\pi^{\varepsilon}(a|s) = \begin{cases} (1-\varepsilon) + \varepsilon \frac{1}{|\mathcal{A}|} & \text{si } a = arg\max_a q_\pi(s,a) \\ \varepsilon \frac{1}{|\mathcal{A}|} & \text{caso contrario} \end{cases}$$

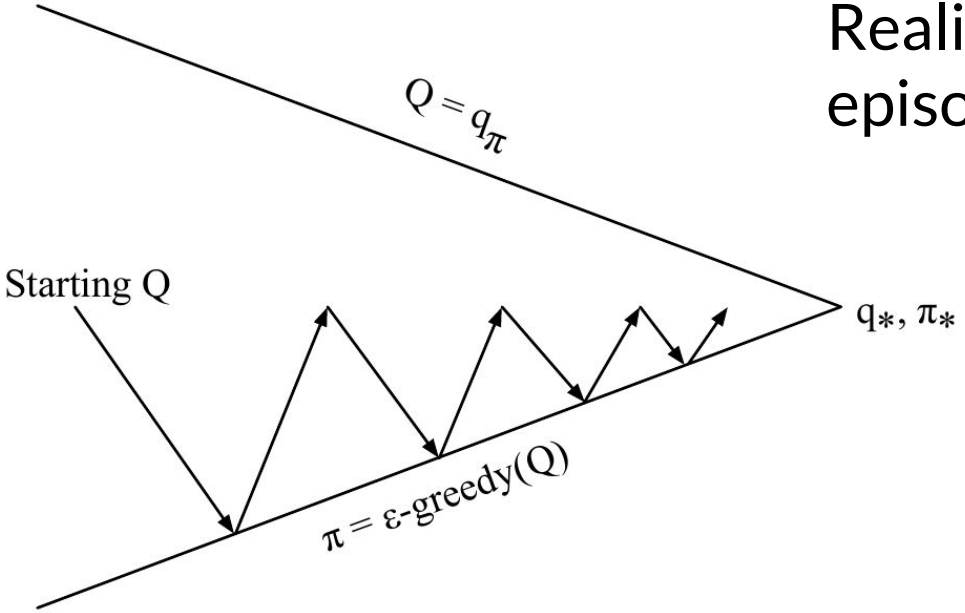
Teorema:

Si π^{ε} una política ε-greedy, $\pi'(s) := arg\max_a q_{\pi^{\varepsilon}}(s,a)$.

Entonces:

$$v_{\pi^{\varepsilon}}(s) \leq v_{\pi'^{\varepsilon}}(s)$$

Algunas mejoras



Realizar la actualización en cada episodio.

GLIE Monte Carlo

- Simular el episodio k utilizando la política π_k^{ε}: $\{S_1^k, A_1^k, R_2^k, \ldots, S_T^k\}$.
- Para cada par (s, a) del episodio

$$N^{k+1}(s,a) = N^k(s,a) + 1$$

$$Q^{k+1}(s,a) = Q^k(s,a) + \frac{1}{N^{k+1}(s,a)}(G^{k+1}(s,a) - Q^k(s,a))$$

- $$\varepsilon = \frac{1}{k}, \qquad \pi_{k+1}^{\varepsilon} = \varepsilon - \text{greedy}(Q(s,a))$$

Ejercicio - Leer:

That any ε-greedy policy with respect to q_π is an improvement over any ε-soft policy π is assured by the policy improvement theorem. Let π' be the ε-greedy policy. The conditions of the policy improvement theorem apply because for any $s \in \mathcal{S}$:

$$\begin{aligned} q_\pi(s, \pi'(s)) &= \sum_a \pi'(a|s) q_\pi(s,a) \\ &= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;+\; (1-\varepsilon) \max_a q_\pi(s,a) \qquad (5.2) \\ &\geq \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \;+\; (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{1-\varepsilon} q_\pi(s,a) \end{aligned}$$

(the sum is a weighted average with nonnegative weights summing to 1, and as such it must be less than or equal to the largest number averaged)

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

Thus, by the policy improvement theorem, $\pi' \geq \pi$ (i.e., $v_{\pi'}(s) \geq v_\pi(s)$, for all $s \in \mathcal{S}$).

Ejercicio (programación)

- Obtener la política óptima y la función de valor para esa política para el ambiente Gridworld (visto en la parte de programación dinámica) utilizando control Monte Carlo, ¿cuántos episodios simulados se necesitan para obtener un resultado con 2 dígitos de precisión para todos los estados en la función de valor óptima?

Lectura recomendada

- OpenAI entrenando manipulación robótica con simulaciones y como pasar lo aprendido en simulaciones a problemas reales: https://blog.openai.com/learning-dexterity/