Aprendizaje Reforzado
Clase 3

Julián Martínez

Evaluación de una política

$$v_\pi^{k+1}(s) = R(s) + \gamma \sum_{s'} v_\pi^{k}(s')\, P^{\pi}_{ss'}$$

Ecuaciones de Bellman

$$v_*(s) = \max_a \left[\sum_{s'} \left\{ r(s,a,s') + \gamma\, v_*(s') \right\} p^{a}_{s,s'} \right]$$

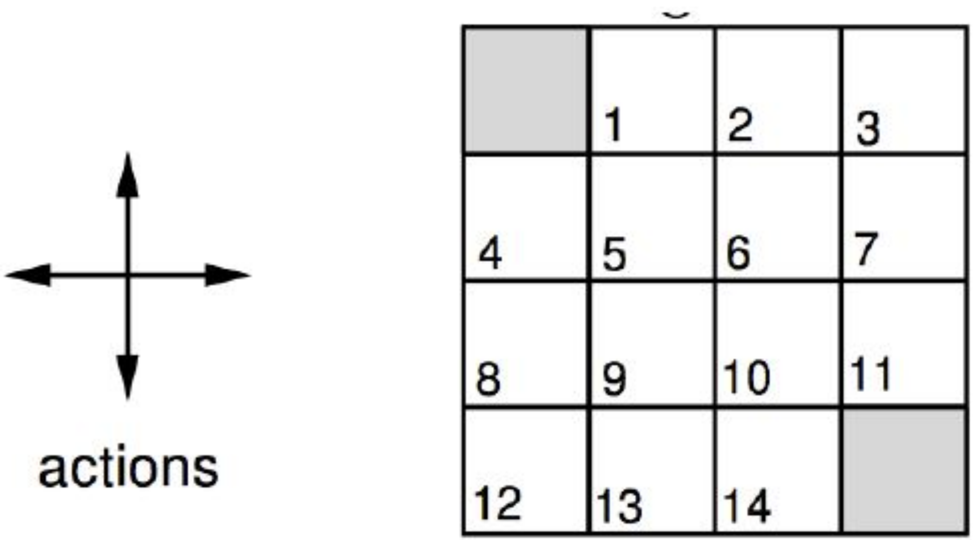
Mejora de una política

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

Evaluación y mejora

$$\pi_k \longrightarrow v_{\pi_k} \longrightarrow q_{\pi_k} \longrightarrow \pi_{k+1}$$

GridWorld (del sutton)



$r = -1$
on all transitions

Multi-armed bandits

Sólo hay acciones y recompensas.

$$A_t \longrightarrow R_t$$

$$A_t \in \{1, \dots, K\}$$

Objetivo

$$\max_{a_1, \dots, a_T} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

NO CONOZCO LA ALEATORIEDAD DEL REWARD!

$$q(a) := \mathbb{E}[R_t \mid A_t = a]$$

Action- Value function

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

¿Cómo elegir las acciones de manera óptima?

$$a^{\circ} := \operatorname{argmax}_a \; q(a)$$

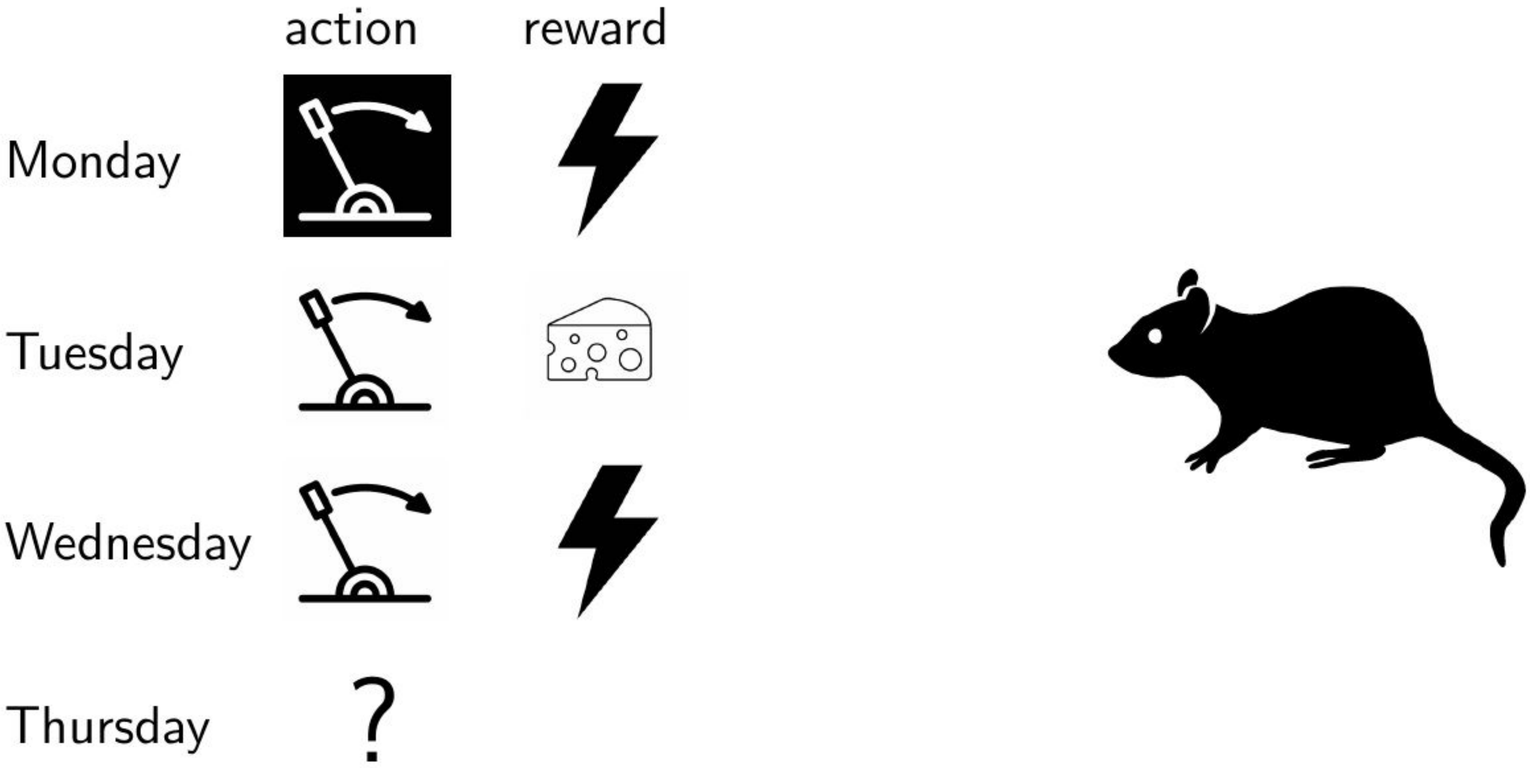
$$A_t^{\circ} := \operatorname{argmax}_a \; Q_t(a)$$

¿Conviene esto?

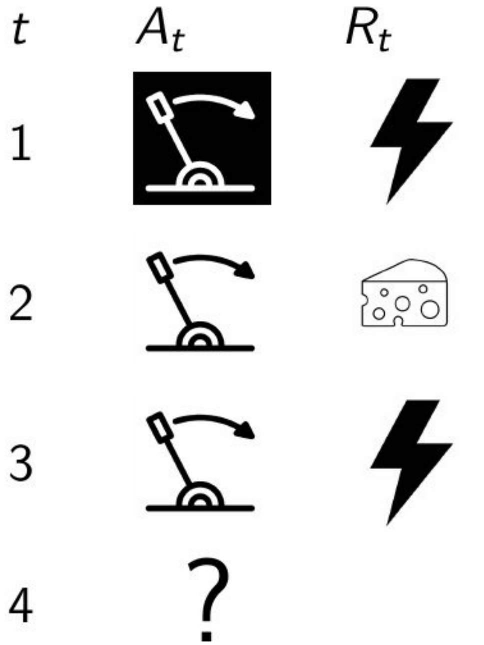
Más Ratas! (Tomado de los slides del curso de deepmind)

	action	reward
Monday		
Tuesday		
Wednesday	?	

Más Ratas! (Tomado de los slides del curso de deepmind)



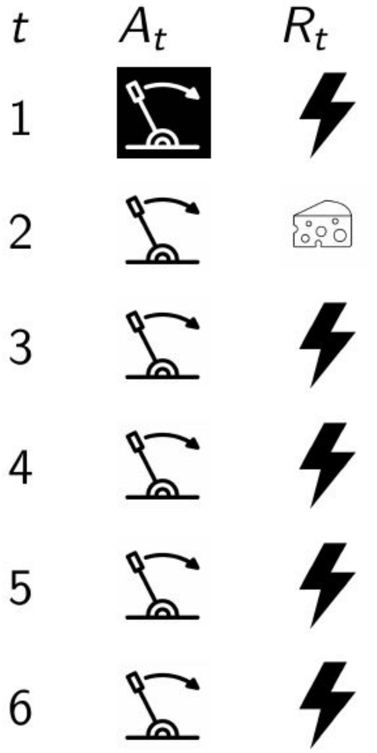
Más Ratas! (Tomado de los slides del curso de deepmind)



- Cheese: $R = +1$
- Shock: $R = -1$
- Then:

$$Q_3(\text{white lever}) = 0$$
$$Q_3(\text{black lever}) = -1$$

Más Ratas! (Tomado de los slides del curso de deepmind)



- Cheese: $R = +1$
- Shock: $R = -1$
- Then:

$$Q_6(\text{white lever}) = -\mathbf{0.6}$$
$$Q_6(\text{black lever}) = -1$$

- When to stop being greedy?

Exploración vs Explotación

- **Explotación**: Tomar la acción que es **más conveniente** *en el momento*.

$$a^{o} := \operatorname{argmax} \; q(a)$$

- **Exploración**: Tomar **decisiones sub-óptimas** con el propósito de *obtener más información.*

$$q^{*}(\cdot) \simeq Q_t(\cdot)$$

Menú 1 - Epsilon Greedy

$$A_t^{\circ} := \arg\max_a Q_t(a)$$

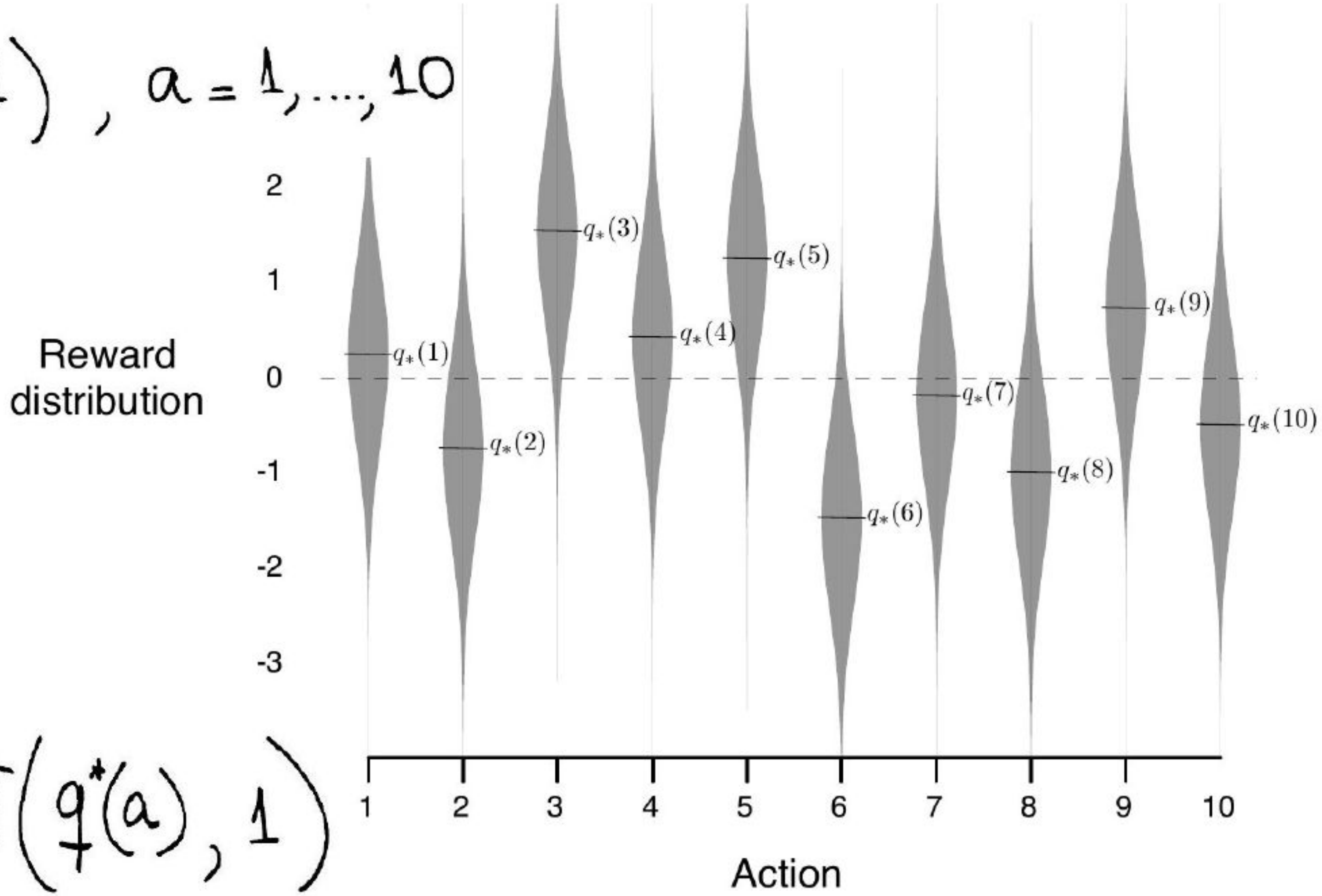
$$\pi_t^{\varepsilon}(a) = \begin{cases} a_t^{\circ} & \text{with prob. } 1-\varepsilon \\ j & \text{with prob. } \varepsilon \cdot \frac{1}{k} \ ; \ j = 1, \ldots, k \end{cases}$$

Obs:

$$Q_t(a) \xrightarrow[t \to \infty]{} q(a) \quad \text{c.s.} \quad \forall a$$

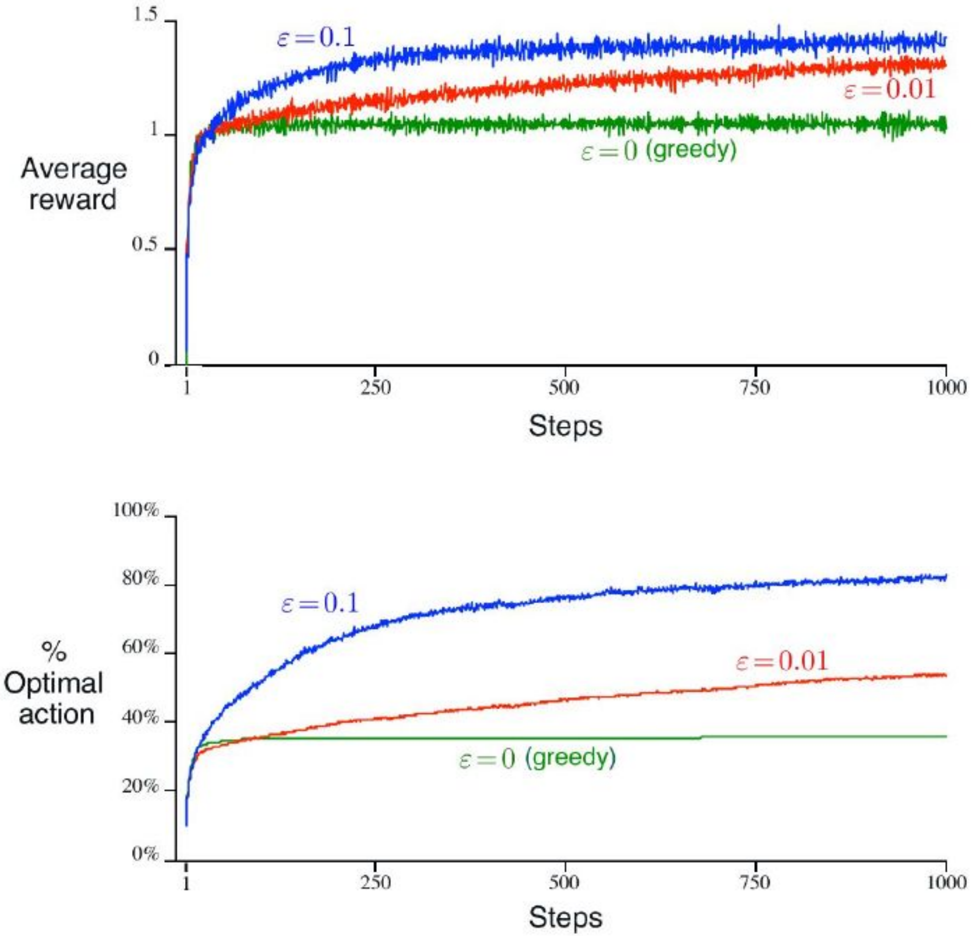
10-armed testbed

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



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

- 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).



¿Cómo explorar de una manera "Inteligente"? (Regret)

$$v_* := \max_a q(a) = \max_a \mathbb{E}[R_t \mid A_t = a]$$

$$L_t = \sum_{i=1}^{t} (v_* - q(a_i)) = \mathbb{E}\left[\sum_{i=1}^{t} (v_* - R_i)\right]$$

$$\text{Minimize } L_t = \max_{a_1, \ldots, a_T} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

Regret Analysis

$$\Delta_a = v_* - q(a) \qquad \text{ACTION REGRET}$$

$$L_t = \sum_{i=1}^{t} v_* - q(a_i) = \sum_a N_t(a)\left(v_* - q(a)\right)$$

$$= \sum_a N_t(a)\,\Delta_a$$

Parando la pelota...

UPPER CONFIDENCE BOUND: $U_t(a)$ /

$$P\left(q(a) \leq Q_t(a) + U_t(a) \;;\; \forall a\right) \geqslant \rho$$

EXPORATION vs EXPLOTATION

$\overset{\simeq}{⚠}\ q$ vs $\dfrac{q}{\underline{⚠}}$

$N_t(a)$ vs $\Delta_a\left(Q_t(a)\right)$

Visitas ~ Certidumbre

- $N_t(a) \downarrow \Rightarrow Var(Q_t(a)) \uparrow \Rightarrow U_t(a) \uparrow$
- $N_t(a) \uparrow \Rightarrow Var(Q_t(a)) \downarrow \Rightarrow U_t(a) \downarrow$

Propuesta

$$a_t = \underset{a}{\operatorname{argmax}} \; Q_t(a) + U_t(a)$$

¿Matemática estas ahí?

<u>Hoeffding's Inequality</u> X_i iid in $[0,1]$. Then

$$P\left(\mathbb{E}[X] \geq \bar{X}_t + u\right) \leq e^{-2nu^2}$$

Si $R_t \in [0,1]$

$$P\left(q(a) \geq Q_t(a) + U_t(a)\right) \leq e^{-2N_t(a)U_t(a)^2}$$

ALGORITMO UCB

Si $\quad p = \frac{1}{t}$

$$\Rightarrow \quad U_t(a) = \sqrt{\frac{\log t}{2N_t(a)}}$$

$$a_t = \underset{a}{\operatorname{argmax}} \; Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}}$$

$$\Delta_a \uparrow \implies q(a) << v^*$$

$$\implies Q_t(a) + U_t(a) < v^*$$

$$\implies N_t(a) \downarrow$$

Ejercicio 3.2 (Opcional)

$$\Delta_a N_t(a) \leq O(\log t) \quad \forall a$$

Sugerencia: Pensar en los incrementos de $N_t(a)$

El lado oscuro...

- Excelente survey sobre regret analysis y Bandit problems: http://sbubeck.com/SurveyBCB12.pdf
- Un capo en las concentration inequalities: https://www.youtube.com/watch?v=SnNj6cMeDq0


TENGO HASTA AHÍ...
GENERADORMEMES.COM