

# APRENDIZAJE REFORZADO

## CLASE 3

**Julían Martínez**

Evaluación de una política

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

Ecuaciones de Bellman

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

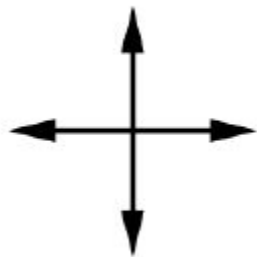
Mejora de una política

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

Evaluación y mejora

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

# GRIDWORLD (DELTUTTON)



actions

~

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

$r = -1$   
on all transitions

# MULTI-ARMED BANDITS

Sólo hay acciones y recompensas.

$$A_t \rightarrow R_t$$



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

Objetivo

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

NO CONOZCO LA ALEATORIEDAD DEL REWARD!

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

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)

	action	reward
Monday		
Tuesday		
Wednesday		
Thursday	?	





# MÁS RATAS! (TOMADO DE LOS SLIDES DEL CURSO DE DEEPMIND)

$t$	$A_t$	$R_t$
1		
2		
3		
4	?	















- ▶ 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)

$t$	$A_t$	$R_t$
1		
2		
3		
4		
5		
6		



► Cheese:  $R = +1$

► Shock:  $R = -1$

► Then:

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

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

► When to stop being greedy?

# EXPLORACIÓN VS EXPLOTACIÓN

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

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

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

$$q^*(\cdot) \approx Q_t(\cdot)$$

# MENÚ 1 - EPSILON GREEDY

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

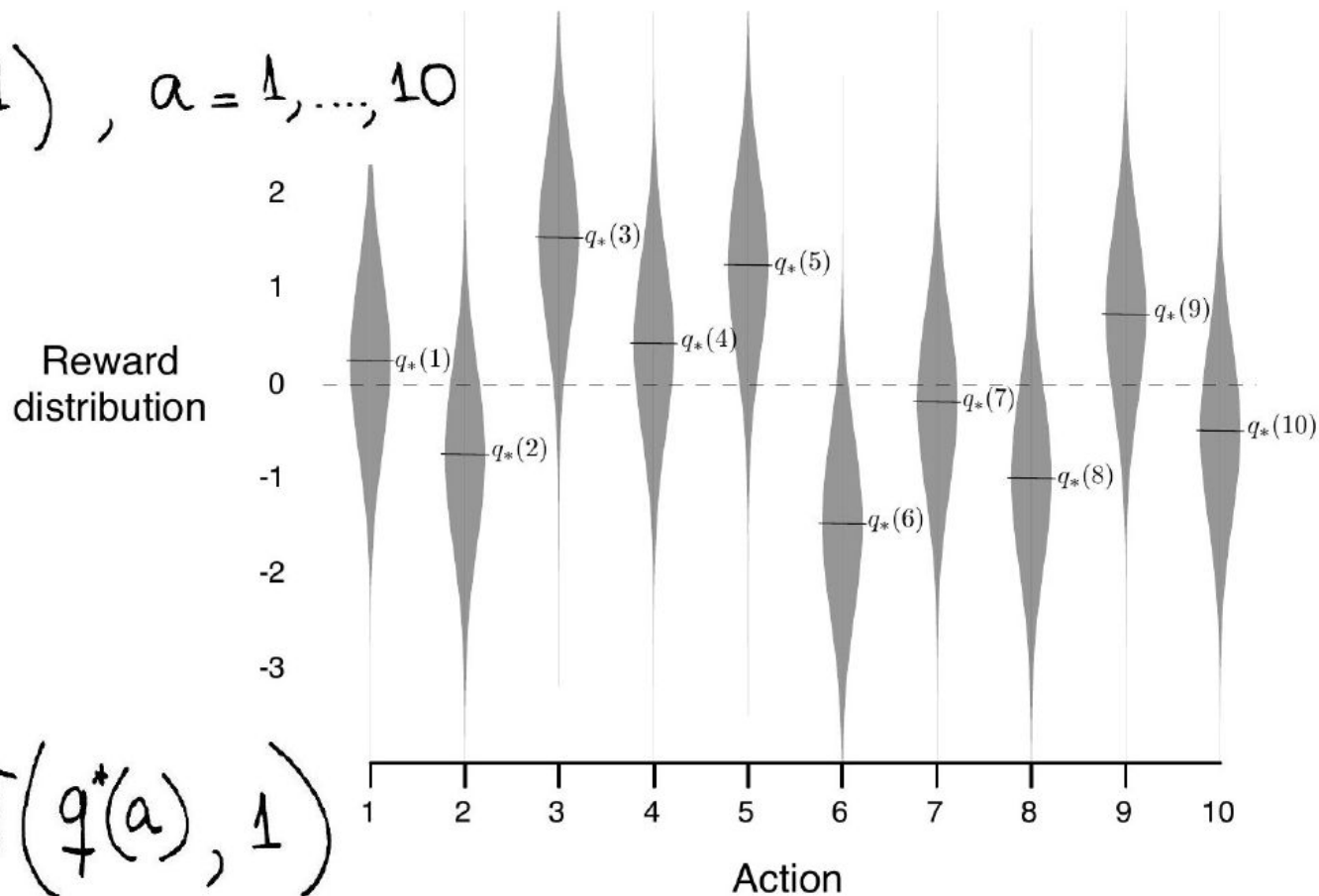
$$\pi_t^\varepsilon(a) = \begin{cases} a_t^o & \text{with prob. } 1-\varepsilon \\ j & \text{with prob. } \varepsilon \cdot \frac{1}{k} ; j=1, \dots, k \end{cases}$$

Obs:

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

# 10-ARMED TESTBED

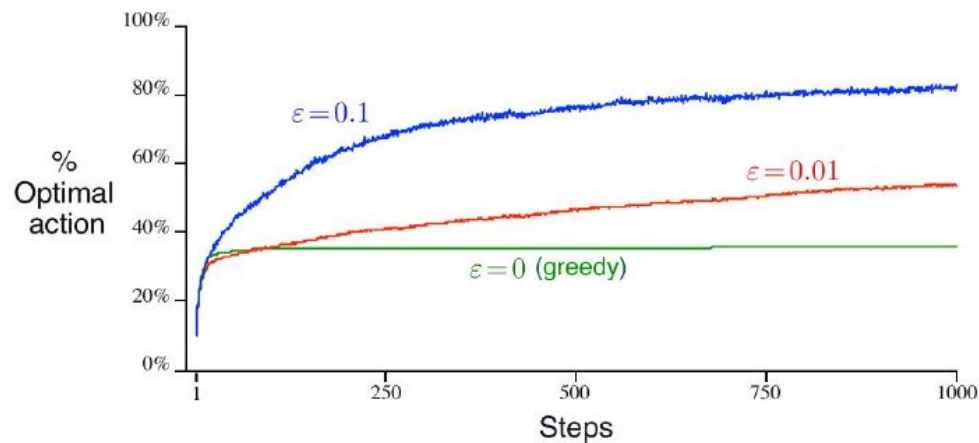
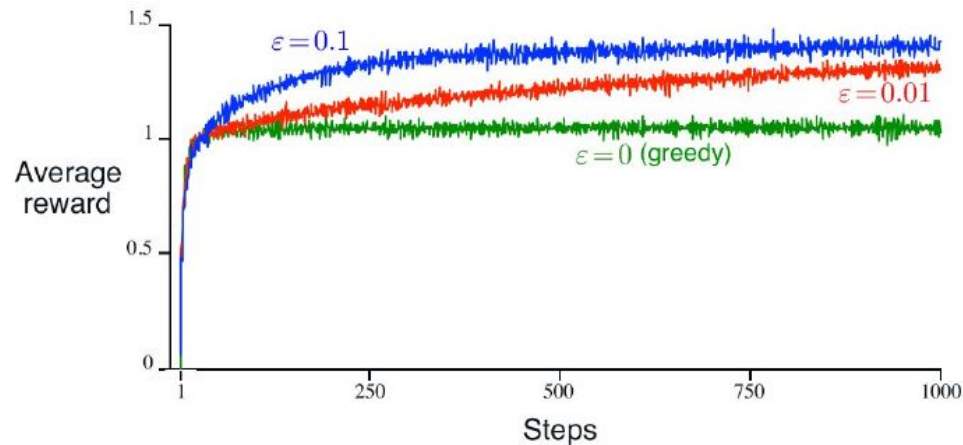
$$q^*(a) \sim \mathcal{N}(0, 1), \quad a = 1, \dots, 10$$



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



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



¿CÓMO EXPLORAR DE UNA MANERA "INTELIGENTE"? (REGRET)

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

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

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

# REGRET ANALYSIS

$$\Delta_a = v_* - q(a)$$

ACTION REGRET

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

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



PARANDO LA PELOTA...

UPPER CONFIDENCE BOUND:  $U_t(a)$  /

$$P(\underbrace{q(a)}_{\text{EXPORATION}} \leq \underbrace{Q_t(a)}_{\text{EXPLOTATION}} + \underbrace{U_t(a)}_{\text{EXPLOTATION}}; \forall a) \geq \rho$$

EXPORATION

vs

EXPLOTATION



vs



$N_t(a)$

vs

$\Delta_a (Q_t(a))$

VISITAS  $\sim$  CERTIDUMBRE

$$\cdot N_t(a) \downarrow \Rightarrow V_{2r}(Q_t(a)) \uparrow \Rightarrow U_t(a) \uparrow$$

$$\cdot N_t(a) \uparrow \Rightarrow V_{2r}(Q_t(a)) \downarrow \Rightarrow U_t(a) \downarrow$$

PROPUESTA

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

¿MATEMÁTICA ESTAS AHÍ?

Hoeffding's Inequality

$X_i$  iid in  $[0,1]$ . Then

$$P(\mathbb{E}[X] \geq \bar{X}_t + \mu) \leq e^{-2n\mu^2}$$

S.  $R_t \in [0,1]$

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

# ALGORITMO UCB

$$S_1 \quad p = \frac{1}{t}$$

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

$a_{t+1}$

$$a_t = \operatorname{argmax}_a$$

$$Q_t(a) + c \sqrt{\frac{\log t}{N_t(a)}}$$

$$\Delta a \uparrow \Rightarrow$$

$$q(a) < v^*$$

$$\Rightarrow$$

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

$$\Rightarrow N_t(a) \downarrow$$

### 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>

