Aprendizaje Reforzado Clase 3

Julián Martínez

Evaluación de una política

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

Ecuaciones de Bellman

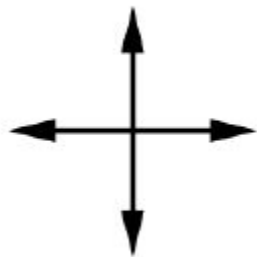$$v_*(s) = \max_a \left[\sum_{s'} \left\{ r(s,a,s') + \gamma \, v_*(s') \right\} P_{s,s'}^a \right]$$

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 (del sutton)

actions

$r = -1$
on all transitions

Multi-armed bandits

Sólo hay acciones y recompensas.

$$A_t \longrightarrow R_t$$

$$A_t \in \{1, ..., \kappa\}$$

Objetivo

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

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?

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

$$A^{\circ}_t := \text{argmax}\; Q_t(a)$$

¿Conviene esto?

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

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

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

t	A_t	R_t
1	(white lever)	(shock)
2	(black lever)	(cheese)
3	(black lever)	(shock)
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:

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

▶ When to stop being greedy?

Exploración vs Explotación

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

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

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

$$q^*(\cdot) \simeq \mathcal{O}_t(\cdot)$$

Menú 1 - Epsilon Greedy

$$A^o_t := \underset{a}{argmax}\ Q_t(a)$$

$$\Pi^{\varepsilon}_t(a) = \begin{cases} a^o_t & \text{with prob. } 1-\varepsilon \\ j & \text{with prob. } \varepsilon \cdot \frac{1}{k} \;,\; j=1,...,k \end{cases}$$

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

10-armed testbed

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

Reward distribution

$q_*(1)$, $q_*(2)$, $q_*(3)$, $q_*(4)$, $q_*(5)$, $q_*(6)$, $q_*(7)$, $q_*(8)$, $q_*(9)$, $q_*(10)$

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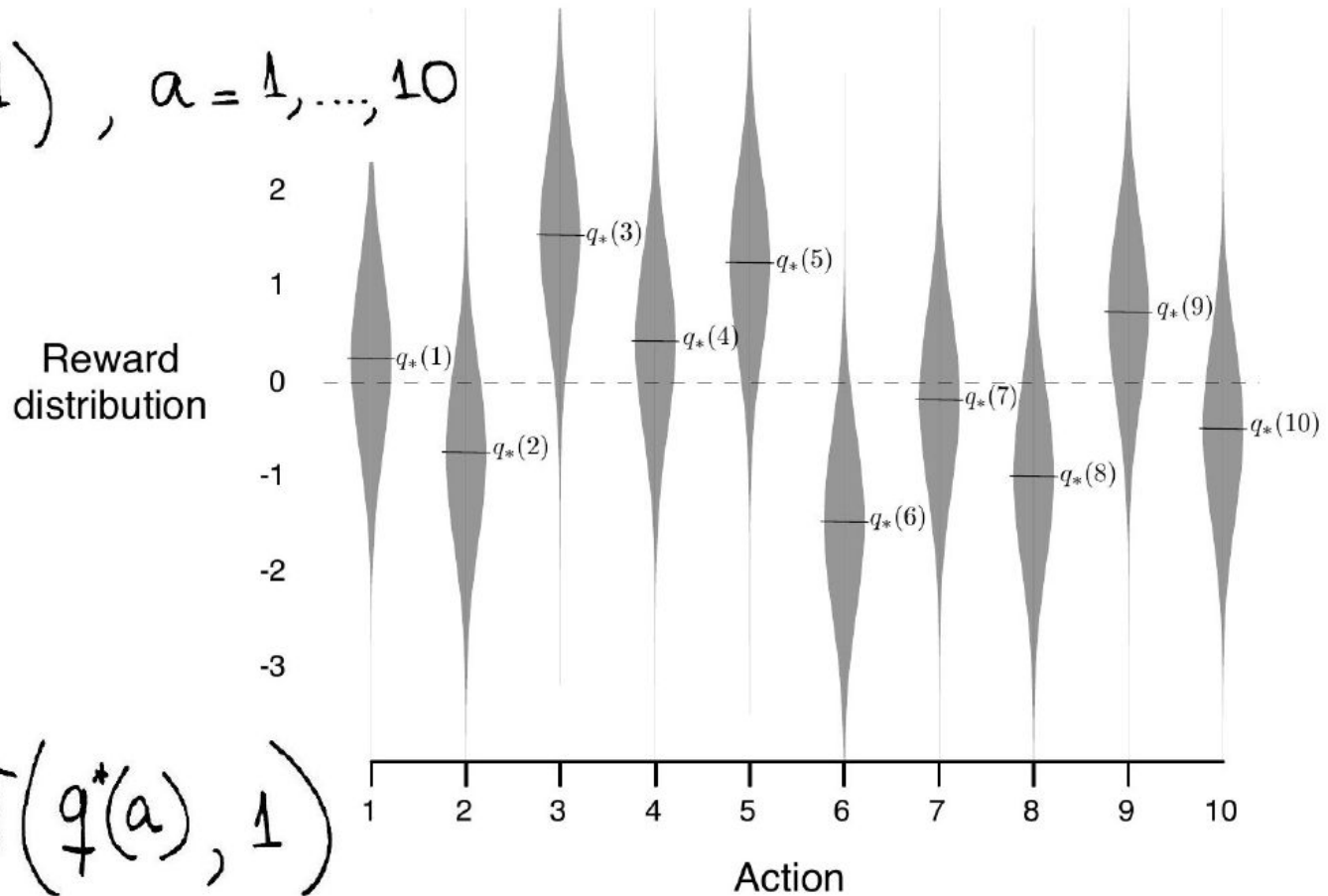
$$R_t \mid A_t = a \sim \mathcal{N}\left(q^*(a), 1\right)$$

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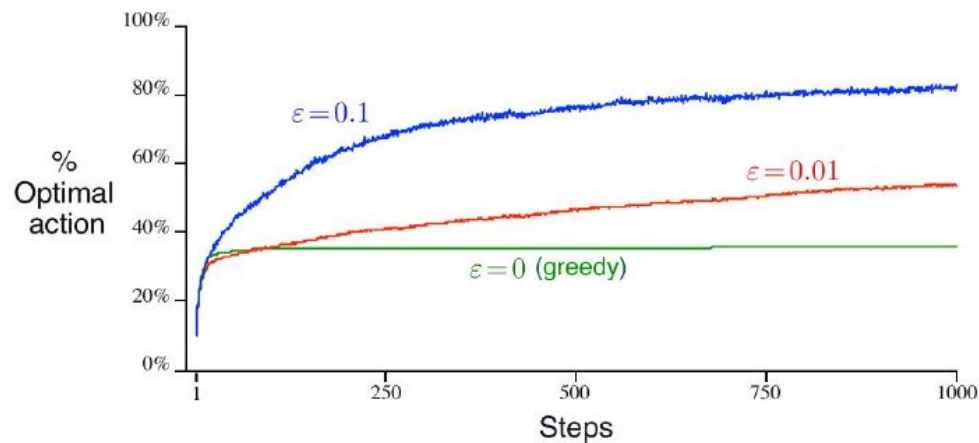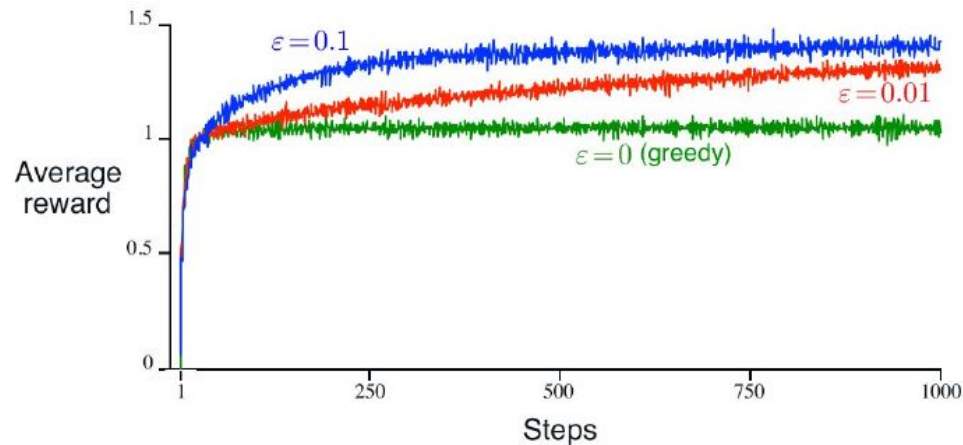
¿Cómo explorar de una manera "Inteligente"? (Regret)

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

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

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

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: $U_t(a)$ ⟋

$$P\left(q(a) \leq \emptyset_t(a) + U_t(a) \; ; \; \forall a \right) \geq s$$

EXPORATION	vs	EXPLOTATION
$\tilde{\triangle}\, q$	vs	$\underset{\triangle}{\overset{q}{\vert}}$
$N_t(a)$	vs	$\Delta_a\left(\emptyset_t(a)\right)$

Visitas ~ Certidumbre

$\cdot \; N_t(a) \downarrow \;\Rightarrow\; Var(\emptyset_t(a)) \uparrow \;\Rightarrow\; U_t(a) \uparrow$

$\cdot \; N_t(a) \uparrow \;\Rightarrow\; Var(\emptyset_t(a)) \downarrow \;\Rightarrow\; U_t(a) \downarrow$

PROPUESTA

$$a_t = \underset{a}{\text{argmax}} \; \emptyset_t(a) + U_t(a)$$

¿Matemática estas ahí?

Hoeffding's Inequality $\qquad X_i$ iid in $[0,1]$. Then

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

Si $\quad R_t \in [0,1]$

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

Cierre

Ejercicio