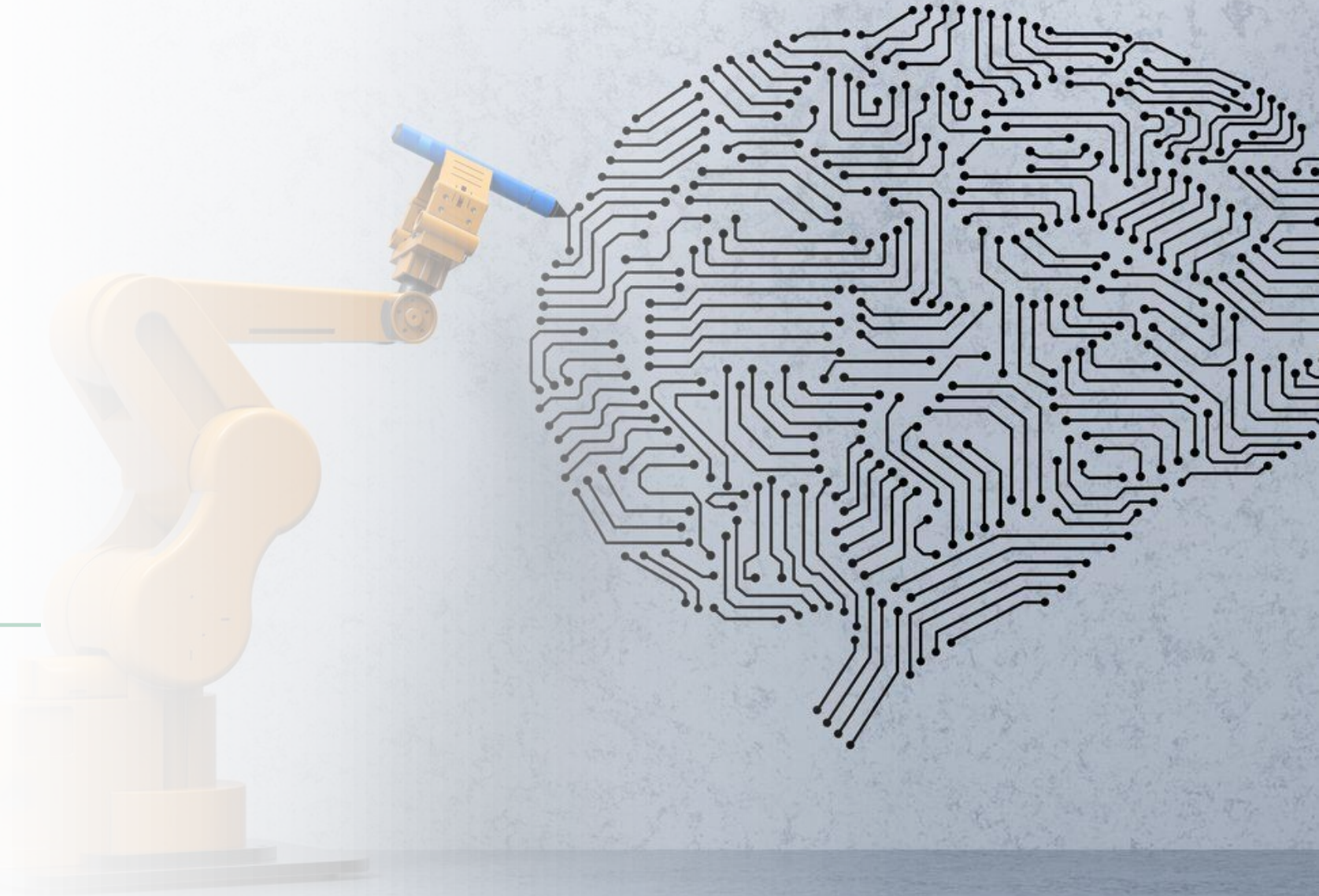


Aprendizaje por refuerzo

Clase 10: Exploración





Antes de empezar...

- Dudas de tarea 2
 - MC
 - Sarsa
 - Q-learning

El algoritmo SARSA

- Teorema: SARSA tabular converge a la función optima de acción valor, $q(s, a) \rightarrow q_*(s, a)$ si la política es GLIE

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Loop for each step of episode:

 Take action A , observe R, S'

 Choose A' from S' using policy derived from Q (e.g., ε -greedy)

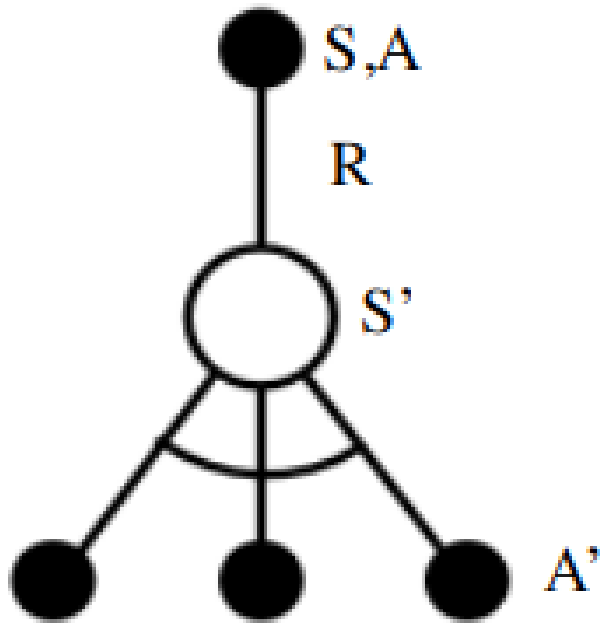
$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$

$S \leftarrow S'; A \leftarrow A';$

 until S is terminal

Algoritmo de control de aprendizaje Q

- $q_{t+1}(S_t, A_t) \leftarrow q_t(S_t, A_t) + \alpha_t(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(S_t, A_t))$
- Teorema
 - Control con aprendizaje Q converge a la función acción valor óptima $q \rightarrow q^*$ en el límite



Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

Initialize S

Loop for each step of episode:


Choose A from S using policy derived from Q (e.g., ε -greedy)

Take action A , observe R, S'

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_{a'} Q(S', a) - Q(S, A)]$

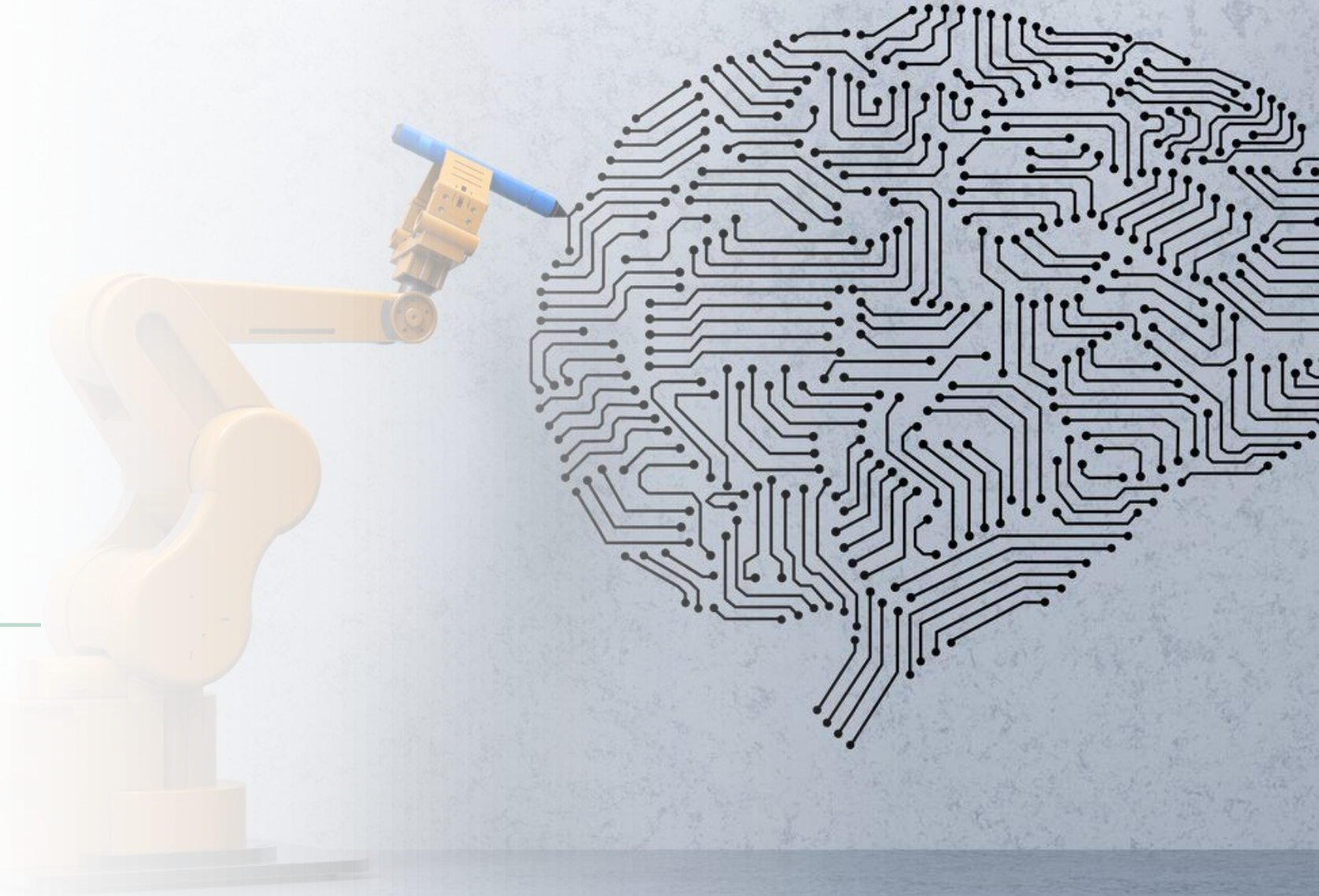
$S \leftarrow S'$

until S is terminal



Aprendizaje por refuerzo

Clase 10: Exploración



Para el día de hoy...

- Exploración



Exploración vs explotación

- La toma de decisiones en línea tiene una elección fundamental
 - Explotar: realizar la mejor decisión dada la información
 - Explorar: obtener nueva información
- La mejor estrategia a largo plazo puede involucrar sacrificios en el corto plazo
- Obtener suficiente información para tomar las mejores decisiones

Ejemplos

Selección de restaurantes

- Explotación: ir al restaurante favorito
- Exploración: Probar un nuevo restaurante

Banners de anuncios

- Explotación: mostrar el anuncio más exitoso
- Exploración: mostrar un anuncio diferente

Juegos

- Explotación: elegir el movimiento que se cree el mejor
- Exploración: elegir un movimiento experimental



Entonces.

¿Cómo puede el agente descubrir estrategias con alta recompensa que requiera una secuencia larga de comportamientos que individualmente no son satisfactorias?

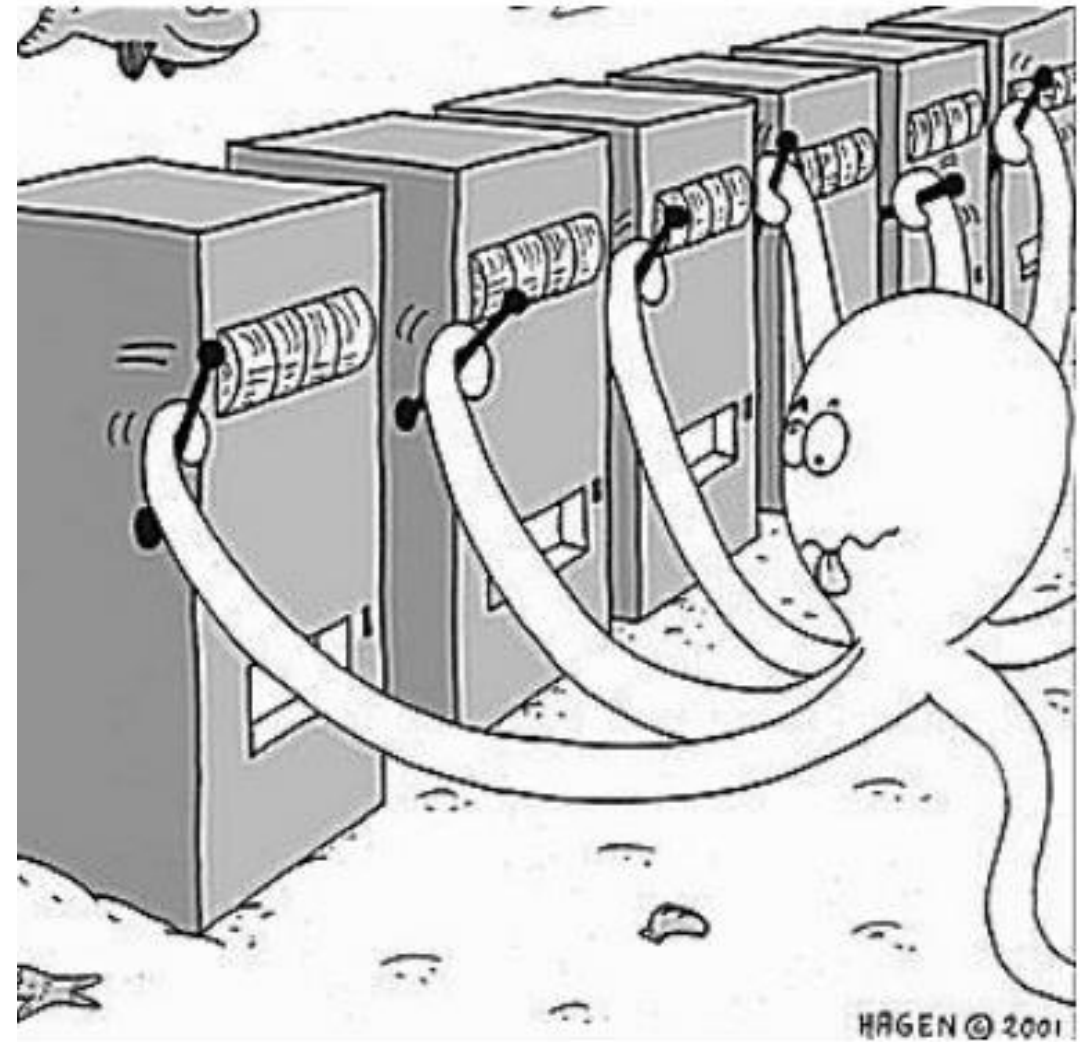
¿Cómo puede un agente decidir cuales comportamientos realizar o continuar haciendo lo que se sabe?

Principios

- Exploración ingenua: añadir ruido a una política voraz (ϵ -voraz)
- Inicialización optimista: suponer lo mejor hasta ser probado lo contrario
- Optimismo bajo incertidumbre: preferir acciones con valores inciertos
- Pareo de probabilidades: seleccionar acciones de acuerdo a la probabilidad de que tan buenas son
- Búsqueda de estado de información: búsqueda incorporando el valor de la información

El bandido multi-brazo

- Es una tupla $(\mathcal{A}, \mathcal{R})$
- \mathcal{A} es un conjunto de m acciones
- $\mathcal{R}^a(r) = \mathbb{P}[r|a]$ es una distribución de probabilidad desconocida sobre recompensas
- En cada paso t el agente selecciona una acción $a_t \in \mathcal{A}$
- El ambiente genera una recompensa $r_t \sim \mathcal{R}^{a_t}$
- El objetivo es maximizar la recompensa acumulativa $\sum_{\tau=1}^t r_{\tau}$



Costo de oportunidad (Regret)

- El acción-valor es la media de recompensa para una acción a

$$Q(a) = \mathbb{E}[r|a]$$

- El valor óptimo V^* es

$$V^* = Q(a^*) = \max_{a \in \mathcal{A}} Q(a)$$

- El regret es la oportunidad pérdida en un paso

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^t V^* - Q(a_\tau)\right]$$

- Maximizar la recompensa cumulativa \equiv minimizar regret total

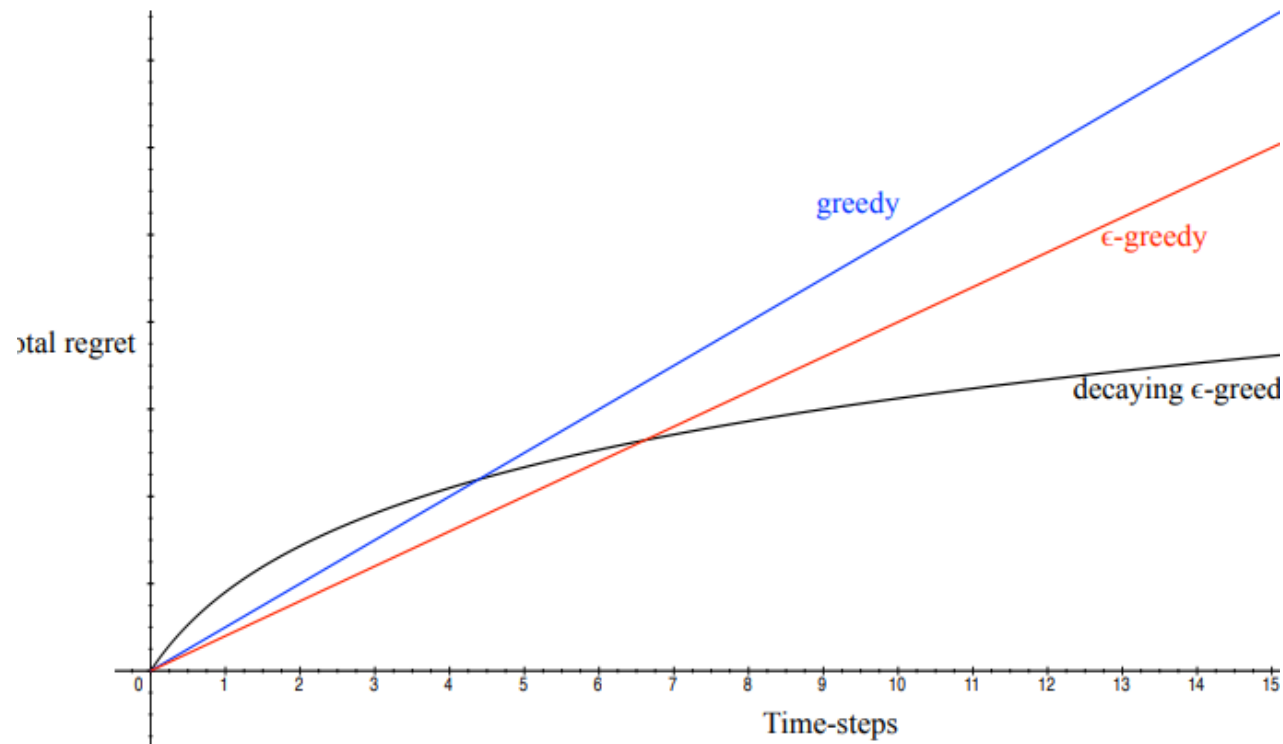
Contando el regret

- El conteo $N_t(a)$ es el número de selecciones de la acción a
- La brecha Δ_a es la diferencia entre el valor de la acción a y la acción óptima a^* , $\Delta_a = V^* - Q(a)$
- El regret es la función de brechas y conteos

$$\begin{aligned} L_t &= \mathbb{E} \left[\sum_{\tau=1}^t V^* - Q(a_\tau) \right] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)](V^* - Q(a)) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)]\Delta_a \end{aligned}$$

- Un buen algoritmo se asegura que se tiene un conteo pequeño para brechas grandes
- Problema: las brechas son desconocidas

Regret lineal o sublineal



- Si un algoritmo explora por siempre, tendrá regret total lineal
- Si un algoritmo nunca explora, tendrá regret total lineal
- ¿Es posible hacer algo mejor?

Algoritmos voraces

- Consideramos algoritmos que estimen $\hat{Q}_t(a) \approx Q(a)$
- Estimar el valor de cada acción por evaluación de Monte-Carlo

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{t=1}^T r_t 1(a_t = a)$$

- El algoritmo voraz selecciona acción con el valor más alto

$$a_t^* = \arg \max_{a \in \mathcal{A}} \hat{Q}_t(a)$$

- La estrategia voraz puede quedarse con una acción subóptima por siempre
- La estrategia voraz tiene regret lineal

Algoritmo ϵ -Voraz

- El algoritmo ϵ -voraz continua explorando por siempre
 - Con probabilidad $1 - \epsilon$ selecciona $a = \arg \max_{a \in \mathcal{A}} \hat{Q}(a)$
 - Con probabilidad ϵ selecciona una acción aleatoria
- ϵ constante minimiza el regret

$$l_t \geq \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \Delta_a$$

- ϵ -voraz tiene regret total lineal

Inicialización Optimista

- Idea: inicializar $Q(a)$ a un valor alto
- Actualizar el valor de acción por evaluación de Monte-Carlo
- Iniciar con $N(a) > 0$

$$\hat{Q}_t(a_t) = \hat{Q}_{t-1} + \frac{1}{N_t(a_t)} (r_t - \hat{Q}_{t-1})$$

- Motiva la exploración sistemática
- Puede quedarse con una acción subóptima
- Voraz + inicialización optimista tiene regret total lineal
- ϵ -voraz + inicialización optimista tiene regret total lineal

Algoritmo ϵ_t -voraz con disminución

- Elegir un calendario de disminución para $\epsilon_1, \epsilon_2, \dots$
- Considerar el siguiente calendario

$$c > 0$$
$$d = \min_{(a | \Delta_a > 0)} \Delta_i$$
$$\epsilon_t = \min \left\{ 1, \frac{c|\mathcal{A}|}{d^2 t} \right\}$$

- Este algoritmo tiene regret total logarítmico
- Sin embargo, el calendario requiere conocimiento de las brechas
- Objetivo: encontrar un algoritmo sublineal para cualquier bandido multi-brazo (sin conocimiento de \mathcal{R})

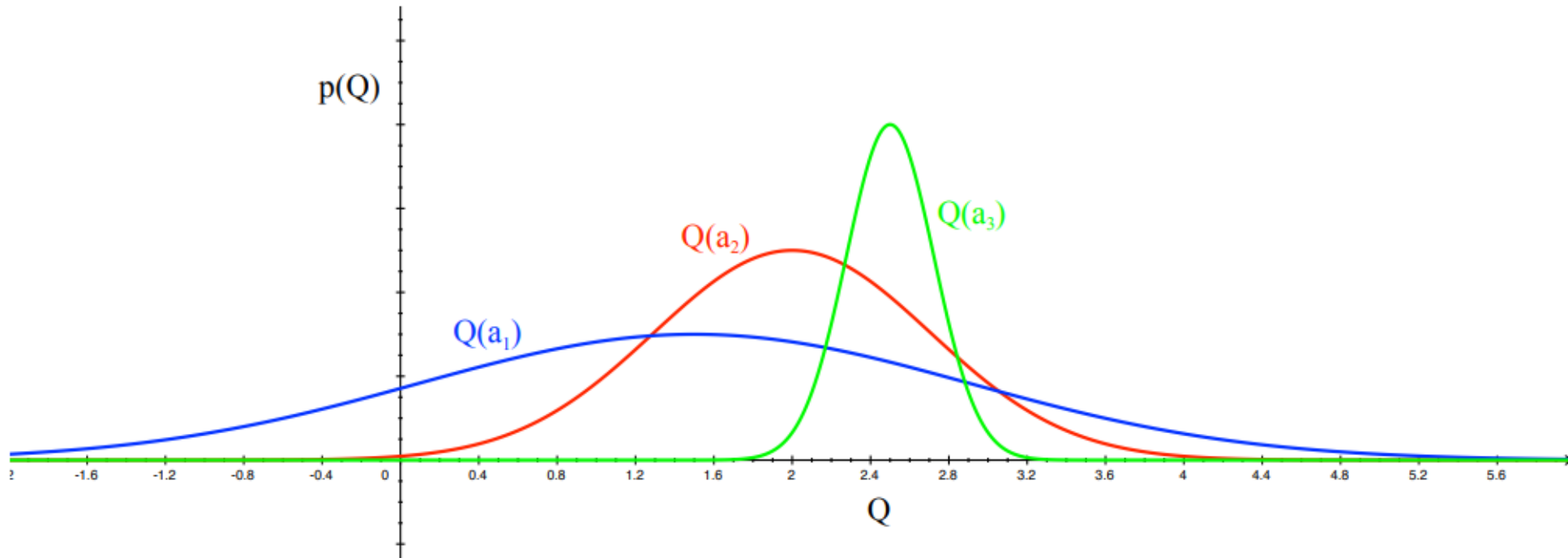
Limite inferior

- El desempeño de cualquier algoritmo está determinado por la similitud entre el brazo óptimo y el resto
- Los problemas complejos tienen brazos que lucen similares con diferentes medias
- Esto se describe con la brecha Δ_a y la similitud en las distribuciones $KL(\mathcal{R}^a || \mathcal{R}^{a*})$
- Teorema Lai y Robbins: El regret total asintótico es al menos logarítmico en el número de pasos

$$\lim_{t \rightarrow \infty} L_t \geq \log t \sum_{(a | \Delta_a > 0)} \frac{\Delta_a}{KL(\mathcal{R}^a || \mathcal{R}^{a*})}$$

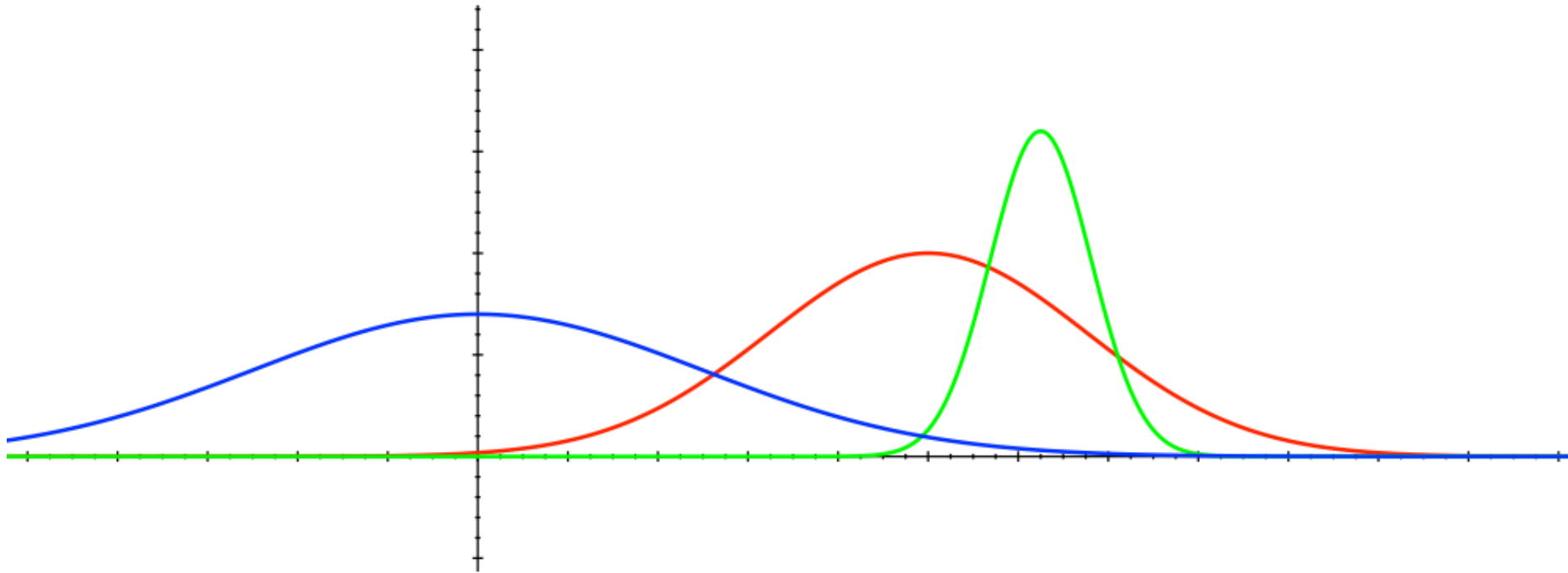
Optimismo bajo incertidumbre I

- ¿Cuál acción debemos elegir?
- Entre más incertidumbre, más importante la exploración
- Puede resultar en la mejor acción



Optimismo bajo incertidumbre II

- Después de elegir la acción azul hay menos incertidumbre
- Podemos elegir otro camino
- Hasta llegar a la mejor acción



Limite de superior confianza (UCB)

- Estima la confianza superior $\hat{U}_t(a)$ para cada valor acción
- Tal que $Q(a) \leq \hat{Q}_t(a) + \hat{U}_t(a)$ con alta probabilidad
- Esto depende del número de veces $N(a)$ que ha sido seleccionado
 - $N_t(a)$ pequeño $\Rightarrow \hat{U}_t(a)$ grande (el valor estimado es incierto)
 - $N_t(a)$ grande $\Rightarrow \hat{U}_t(a)$ pequeño (el valor estimado es cierto)
- Seleccionar la acción que maximiza UCB

$$a_t = \arg \max_{a \in \mathcal{A}} \hat{Q}_t(a) + \hat{U}_t(a)$$

Desigualdad de Hoeffding

- Teorema: sean X_1, \dots, X_t variables aleatorias iid en $[0,1]$ y sea $\bar{X}_t = \frac{1}{t} \sum_{\tau=1}^t X_\tau$ sea la media empírica, entonces

$$\mathbb{P}[\mathbb{E}[X] > \bar{X}_t + u] \leq e^{-2tu^2}$$

- Cuando utilizamos la desigualdad a las recompensas del bandido condicionado a la acción a

$$\mathbb{P}[Q(a) > \hat{Q}_t(a) + U_t(a)] \leq e^{-2N_t(a)U_t(a)^2}$$

Calculando UCB

- Elegimos una probabilidad p que su valor exceda UCB
- Ahora resolvemos para $U_t(a)$

$$e^{-2N_t(a)U_t(a)^2} = p$$

$$U_t(a) = \sqrt{-\frac{\log(p)}{2N_t(a)}}$$

- Reducir p al observar más recompensas, por ej. $p = t^{-4}$
- Asegura que seleccionemos la acción óptima para $t \rightarrow \infty$

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

UCB1

- $a_t = \arg \max_{a \in \mathcal{A}} Q(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$
- Teorema: el algoritmo UCB obtiene regret asintótico total logarítmico

$$\lim_{t \rightarrow \infty} L_t \leq 8 \log t \sum_{a | \Delta_a > 0} \Delta_a$$

Ejemplo

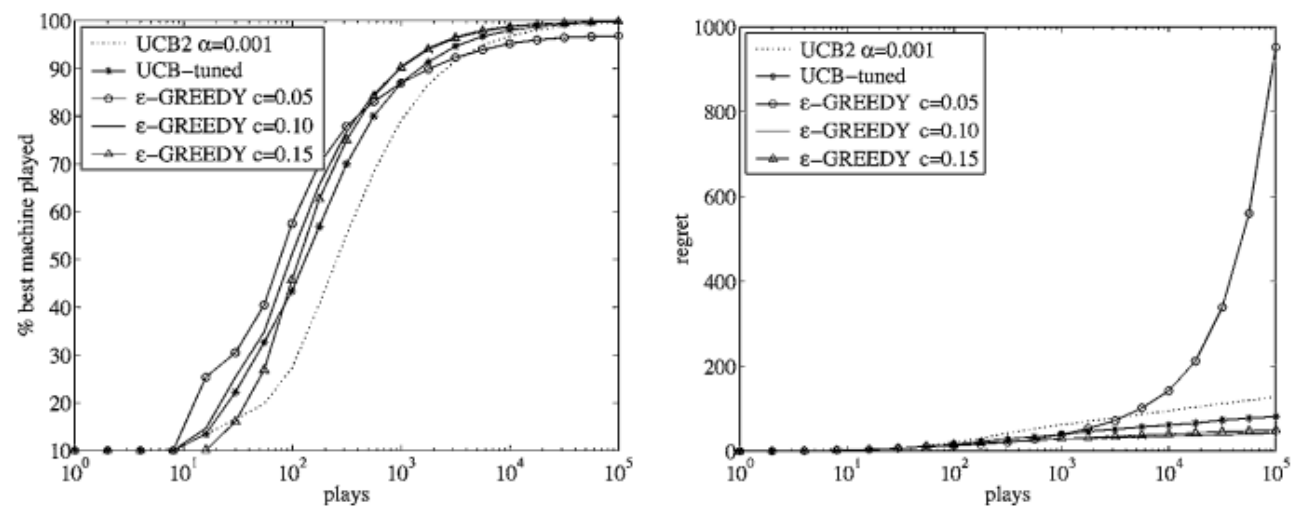
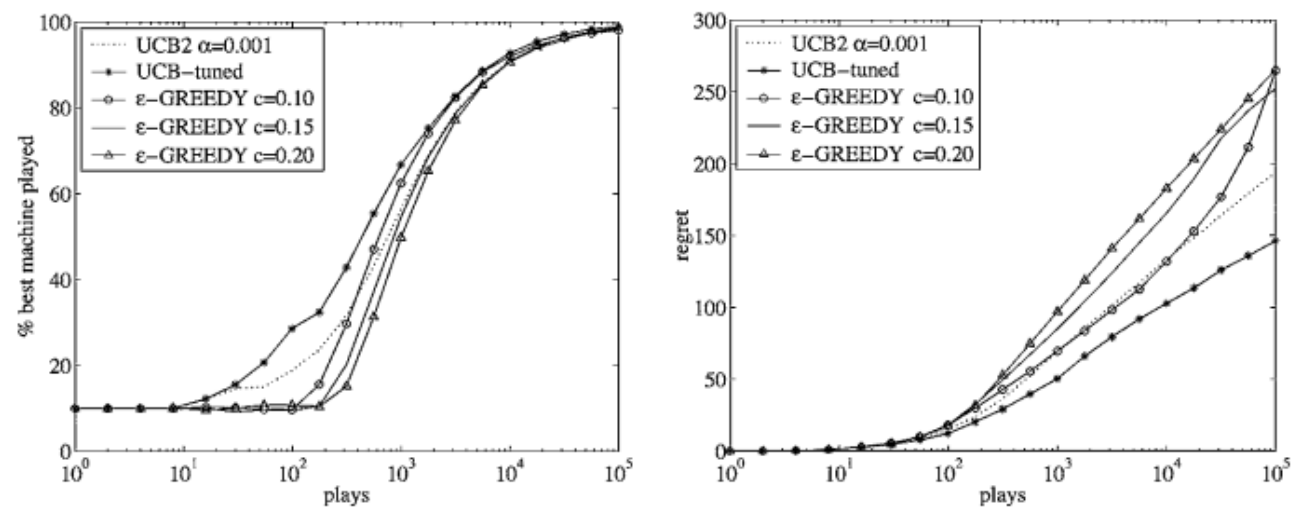


Figure 9. Comparison on distribution 11 (10 machines with parameters 0.9, 0.6, ..., 0.6).

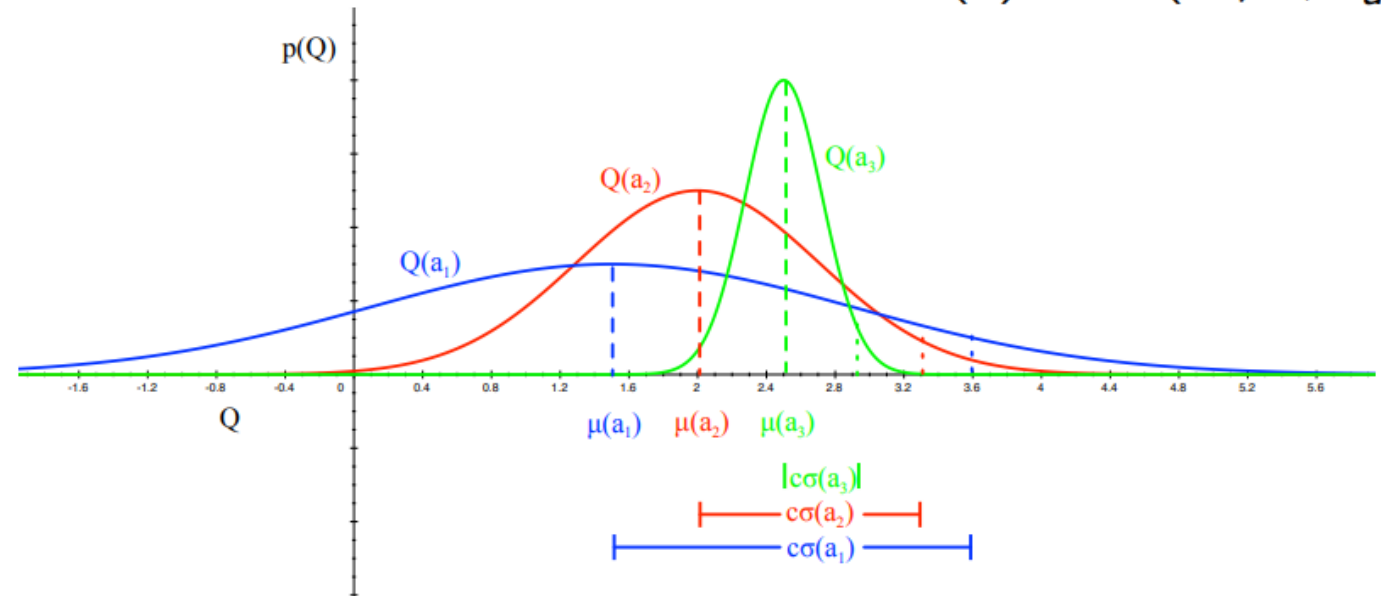


Ahora... Bandidos Bayesianos

- Hasta ahora no hemos hecho suposiciones acerca de la distribución de la recompensa \mathcal{R}
- Los bandidos Bayesianos explotan la información a priori de recompensas $p[\mathcal{R}]$
- Calculan la distribución de recompensas a posteriori $p[\mathcal{R}|h_t]$ donde $h_t = a_1, r_1, \dots, a_{t-1}, r_{t-1}$ es la historia
- Usamos la distribución a posteriori para guiar la exploración
 - Límites de confianza superiores (UCB Bayesianos)
 - Pareo de probabilidades (muestreo de Thompson)
- Mejor desempeño si el conocimiento a priori es preciso

Ejemplo de UCB Bayesianos

- Supongamos que la distribución de recompensa es Gaussiana, $\mathcal{R}_a(r) = \mathcal{N}(r; \mu_a, \sigma_a^2)$
- Calcular la posteriori Gaussiana sobre μ_a y σ_a^2 (por medio del teorema de Bayes)
- $p[\mu_a, \sigma_a^2 | h_t] \propto p[\mu_a, \sigma_a^2] \prod_{(t|a_t = a)} \mathcal{N}(r_t; \mu_a, \sigma_a^2)$
- Elegir la acción que maximice la desviación estándar de $Q(a)$
$$a_t = \arg \max \mu_a + \frac{c\sigma_a}{\sqrt{N(a)}}$$





Para la otra vez...

- Exploración II



iimas

A close-up photograph of a vintage typewriter. The focus is on the carriage and the typebars, which are arranged in a semi-circle. The words "The End." are printed in a classic typewriter font on a piece of paper emerging from the machine. The background is a light pink wall.

The End.