Multi Armed Bandits

Fernando Lozano

Universidad de los Andes

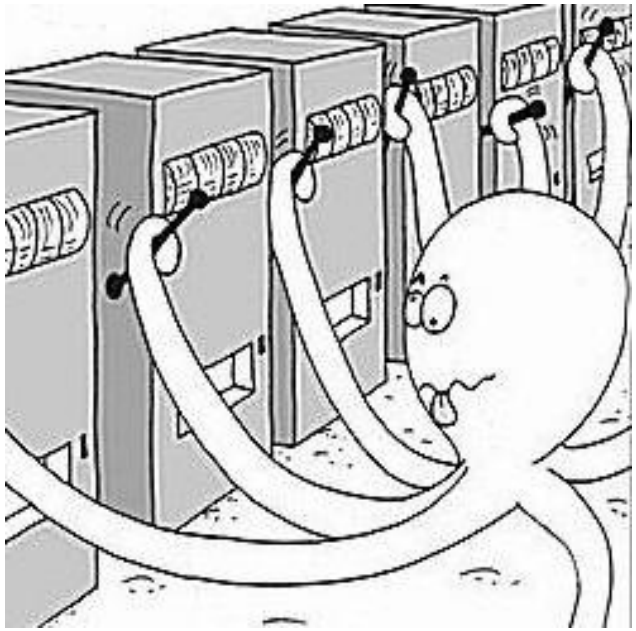
31 de enero de 2023

Evaluación vs. Instrucción

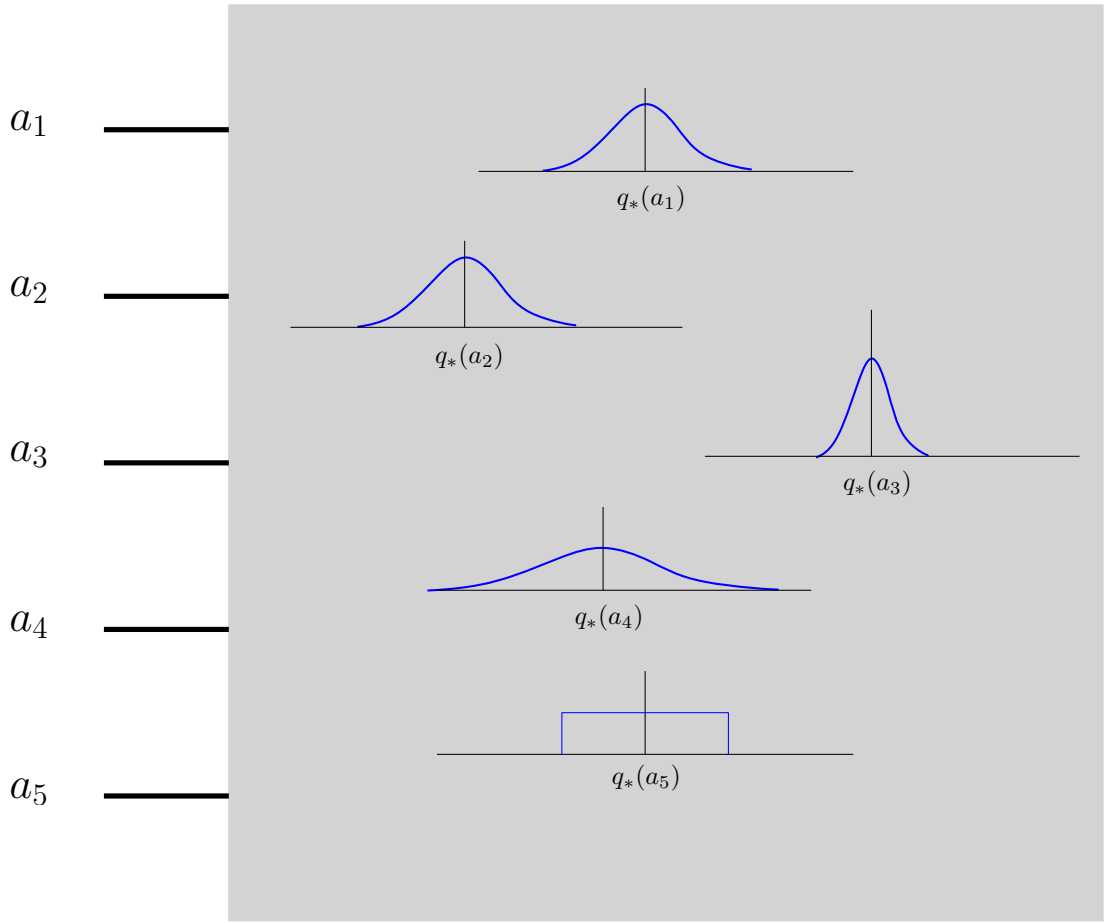
- Algoritmos de RL usan evaluación de acciones ejecutadas o a ejecutar.
- No hay supervisor/maestro que indique cuál es la mejor o peor acción en una situación dada
- Retroalimentación que recibe el agente depende de acción ejecutada.
- Dilema entre explotación y exploración.
- Estudiaremos escenario simplificado en que la situación en que se encuentra el agente no cambia.

One armed bandit

Multi-armed Bandits

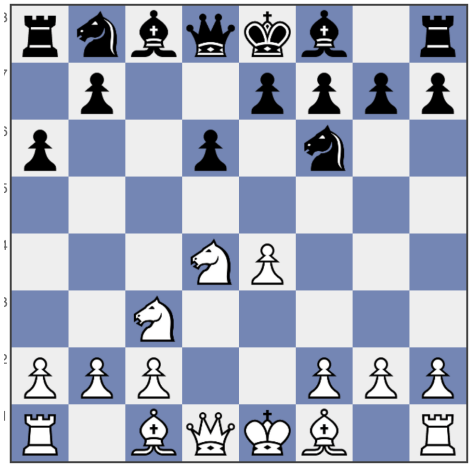


- Seleccionar repetidamente entre una de k posibles opciones (acciones).
- Al seleccionar una opción se recibe una recompensa de una distribución de probabilidad que depende de la acción ejecutada.
- Objetivo: Maximizar la recompensa total esperada en un horizonte de tiempo (p.ej. 1000 acciones).



$$q_*(a_i) = \mathbb{E}\left[R_t \mid A_t = a_i\right]$$

Ejemplo



- Acción: jugada posible en este tablero.
  - Seguir jugando contra el mismo oponente, siguendo la misma estrategia.
  - Terminar la partida: ganar, perder o empatar.
- Repetir 1000 veces.

Dilema exploración/explotación

- Si conocieramos $q_*(a_i)$, problema resuelto!
- Al seleccionar a_i recibimos información sobre $q_*(a_i)$.
- $Q_t(a_i)$: estimativo de $q_*(a_i)$ en la iteración t, basado en experiencia anterior.
- Selección greedy:

$$a = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, Q_t(a_i)$$

  - Explotar conocimiento actual.
  - Mejora estimativo de acción greedy.
  - Seleccionar acción diferente: exploración.
- Balancear explotación y exploración.
- Extensa bibliografía sobre balance óptimo.

Métodos de valor de acción (Action-value methods)

- Estimar $q_*(a)$: promedio de recompensas observadas cuando se selecciona a:

$$Q_t(a) = \begin{cases} \frac{\sum_{i=1}^{t-1} R_i \times \mathbb{I}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{I}_{A_i=a}} & \sum_{i=1}^{t-1} \mathbb{I}_{A_i=a} \neq 0 \\ v_0 & \sum_{i=1}^{t-1} \mathbb{I}_{A_i=a} = 0 \end{cases}$$

- Selección greedy:

$$a = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, Q_t(a_i)$$

- Selección ϵ-greedy:

$$\mathbf{P}\left[A_t = a\right] = \begin{cases} 1 - \epsilon + \frac{\epsilon}{k} & a = \arg\max_{i \in \{1,\ldots,k\}} Q_t(a_i) \\ \frac{\epsilon}{k} & \text{demás acciones.} \end{cases}$$

Testbed

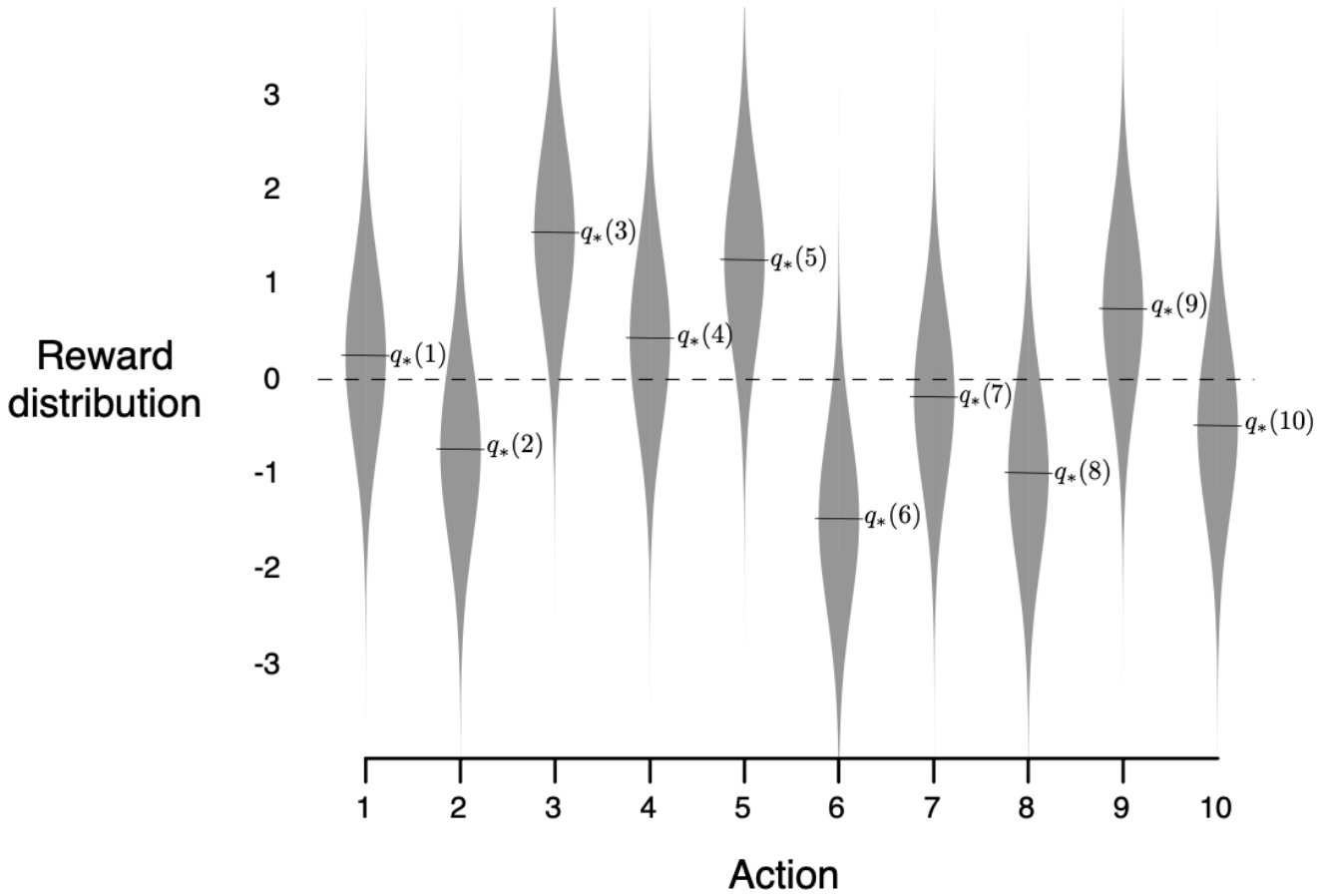


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$, unit-variance normal distribution, as suggested by these gray distributions.

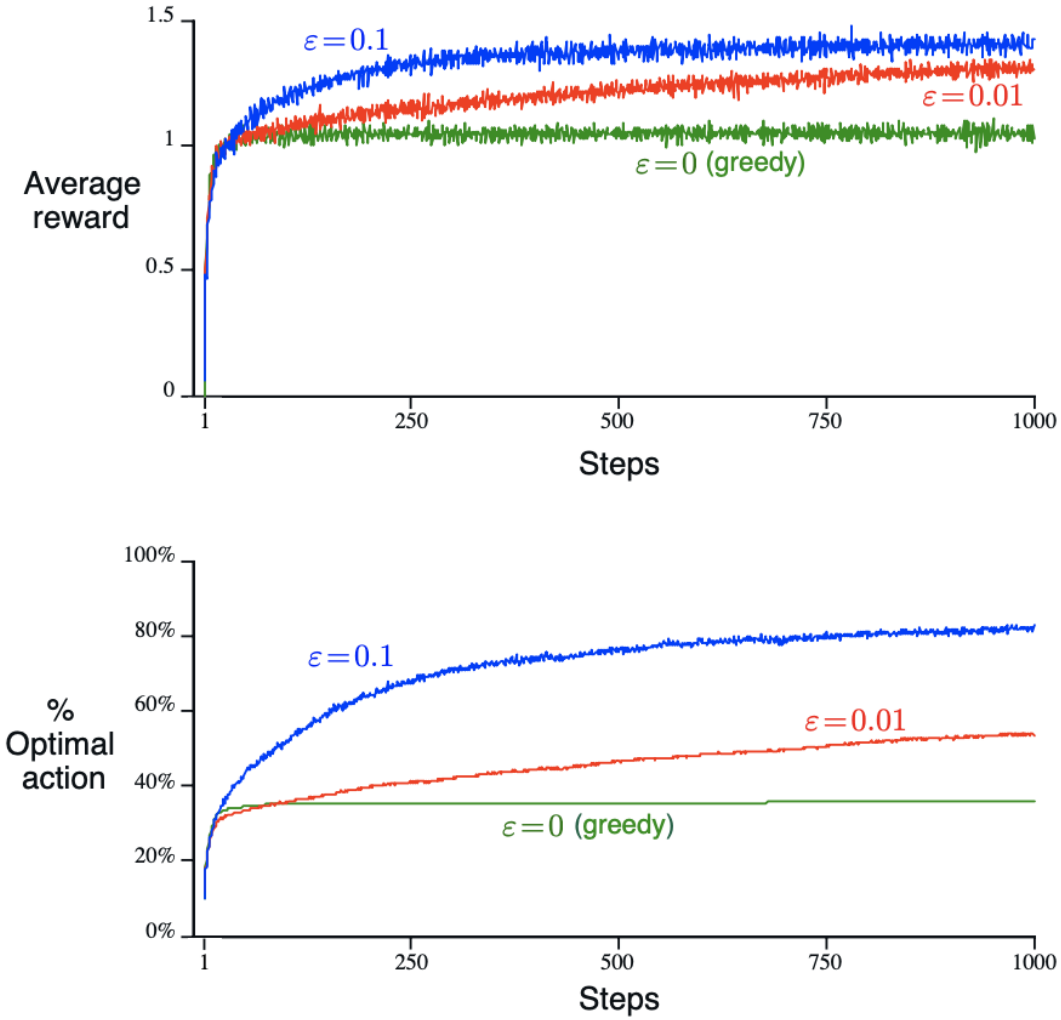


Figure 2.2: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.

Preguntas

- Qué pasa si acciones tienen varianza muy grande?
- Qué pasa si acciones tienen varianza muy pequeña?
- Qué pasa si los $q_*(a_i)$ cambian con el tiempo?

Implementación eficiente

- Sea Q_n el estimativo del valor de a después de haberla seleccionado $n-1$ veces:

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

- Cálculo incremental:

$$\begin{aligned} Q_{n+1} &= (Q_n(n-1) + R_n)\,\frac{1}{n} \\ &= \underbrace{Q_n}_{\text{Valor viejo}} + \underbrace{\frac{1}{n}}_{\text{step size}} \underbrace{[R_n - Q_n]}_{\text{Actualización}} \end{aligned}$$

- Actualización en la dirección del objetivo.
- $[R_n - Q_n]$: error en el estimativo.

Caso no estacionario

- Probabilidades cambian con el tiempo.
- Dar más peso a observaciones recientes que a observaciones pasadas.
- Tamaño de paso $\alpha \in (0, 1]$:

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha \left[R_n - Q_n \right] \\
Q_{n+1} &= \alpha R_n + (1 - \alpha) Q_n \\
&= \alpha R_n + (1 - \alpha) \left[\alpha R_{n-1} + (1 - \alpha) Q_{n-1} \right] \\
&= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \left[\alpha R_{n-2} + (1 - \alpha) Q_{n-2} \right] \\
&\vdots \\
&= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \\
&\quad \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\
&= (1 - \alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1 - \alpha)^{n-i} R_i
\end{aligned}
$$

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

- Promedio pesado:

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

- Peso de R_i más antiguos es menor.
- Decaimiento exponencial con exponente $(1-\alpha)$.

Tamaño de paso variante

$$Q_{n+1} = Q_n + \alpha_n \left[R_n - Q_n \right]$$

- Condiciones de convergencia (con probabilidad 1):

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty$$

- $\alpha_n = \frac{1}{n}$ satisface estas condiciones (caso estacionario).
- $\alpha_n = \alpha$ no satisface estas condiciones (caso no estacionario).

Inicialización Optimista

- $Q_1(a) = v_0$ introduce sesgo en los algoritmos vistos:
  - Desaparece en promedio simple, cuando se han usado todas las acciones.
  - Decrece exponencialmente en el promedio pesado.
- $v_0 \gg$ puede fomentar exploración inicial en métodos $\epsilon-$greedy.
- Caso no estacionario.

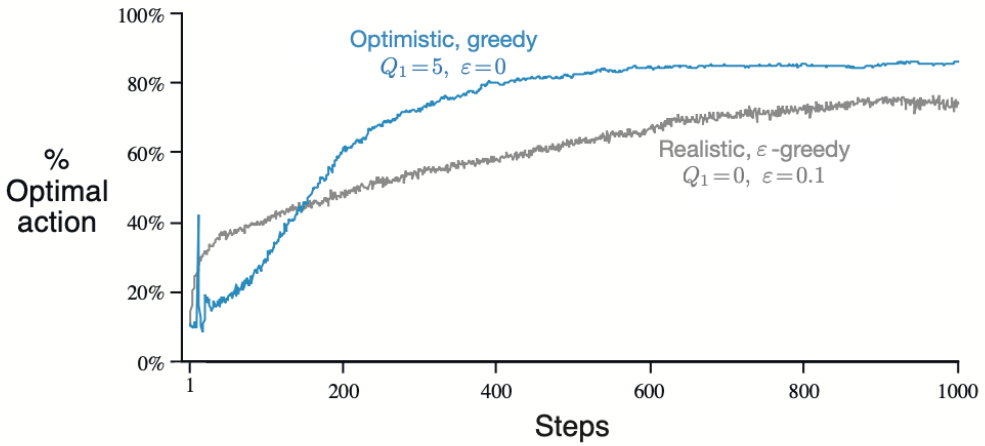


Figure 2.3: The effect of optimistic initial action-value estimates on the 10-armed testbed. Both methods used a constant step-size parameter, $\alpha = 0.1$.

Selección con intervalo de confianza (UCB)

- Al explorar $\epsilon-$greedy escoge acción de manera aleatoria.
- Preferible seleccionar acciones:
  1. Con estimativo de valor cercano al de la acción greedy.
  2. Preferir acciones con estimativos inciertos.

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

- $N_t(a)$: Número de veces que se ha seleccionado a.
- c: balance entre valor del estimativo e incertidumbre (intervalo de confianza).
- Cota superior en intervalo de confianza.
- Al crecer t:
  - ★ Se seleccionan todas las acciones.
  - ★ No se seleccionan acciones con estimativos bajos o que se han seleccionado muchas veces.

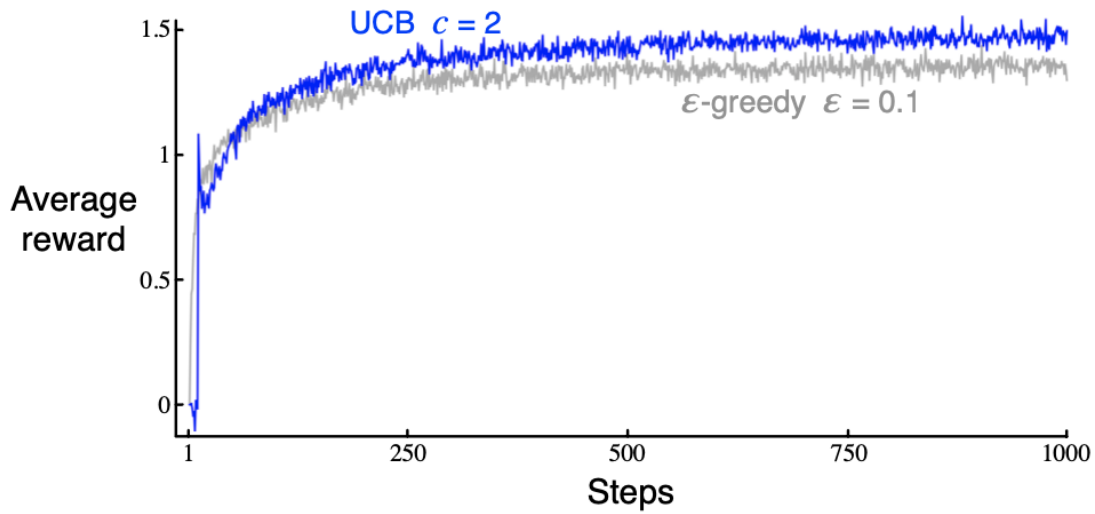


Figure 2.4: Average performance of UCB action selection on the 10-armed testbed. As shown, UCB generally performs better than ε-greedy action selection, except in the first k steps, when it selects randomly among the as-yet-untried actions.

Gradient Bandit

- No usa estimativos directamente para seleccionar acción.
- Usa medida de preferencia relativa entre acciones $H_t(a)$
- Distribución soft-max:

$$\mathbf{P}\left[A_t = a_i\right] \doteq \pi_t(a) \doteq \frac{e^{H_t(a_i)}}{\sum_{j=1}^{k} e^{H_t(a_j)}} = \frac{1}{1 + \sum_{\substack{j=1 \\ j \neq i}}^{k} \frac{e^{H_t(a_j)}}{e^{H_t(a_i)}}}$$

- Algoritmo:
  - $H_1(a_i) = 0 \; i = 1, \ldots, k$
  - Selecciona acción A_t, recibe R_t,

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t))$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \qquad \forall a \neq A_t$$

donde el baseline $\bar{R}_t$ es el promedio de recompensas recibidas hasta t.

- Ascenso de gradiente estocástico sobre $\mathbb{E}[R_t] = \sum_a \pi_t(a) q_*(a)$:

$$H_{t+1}(a) = H_t(a) + \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

- No conocemos $q_*(a)$.
- Se puede mostrar que:

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\left[H_t(A_t) + \alpha(R_t - \bar{R}_t)(\mathbb{I}_{a=A_t} - \pi_t(A_t))\right]$$

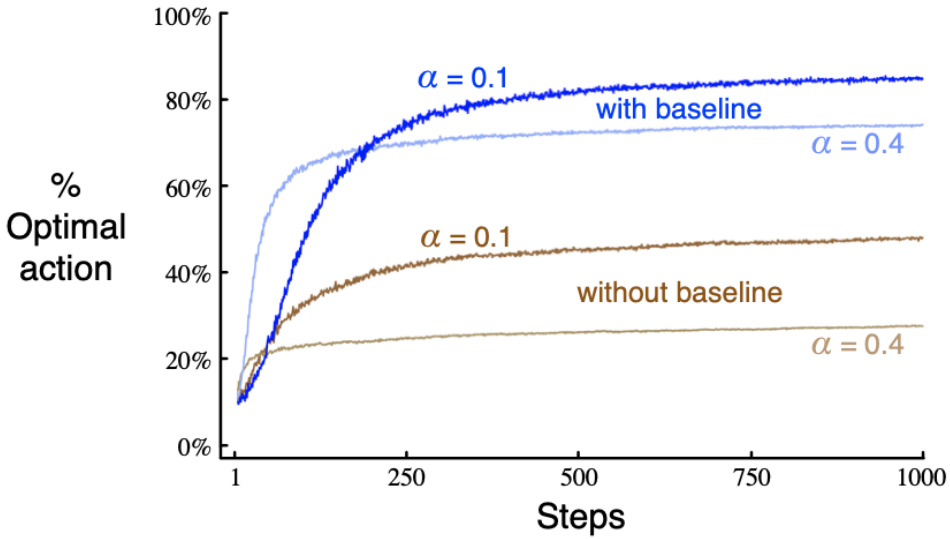


Figure 2.5: Average performance of the gradient bandit algorithm with and without a reward baseline on the 10-armed testbed when the $q_*(a)$ are chosen to be near $+4$ rather than near zero.

Hacia el problema de RL

- Asociar diferentes acciones a diferentes situaciones $\longrightarrow$ búsqueda asociativa (contextual bandits).
- Acciones afectan también la siguiente situación $\longrightarrow$ Aprendizaje por refuerzo.