Agentes Inteligentes

Franz Mayr
mayr@ort.edu.uy

Universidad ORT Uruguay
12 de marzo de 2023

Presentación de la materia

Presentación de la materia

- Dictada en conjunto con “Taller de Inteligencia Artificial”
- Docentes
  - Marcos Begerez
  - Franz Mayr
- Evaluaciones
  - Tareas: 35
    - En equipo de hasta 2 estudiantes
    - Entrega por aulas
  - Obligatorio: 30
    - En equipo de hasta 2 estudiantes
    - Entrega por gestión
  - Parcial: 35
    - Presencial

Agentes Inteligentes

Franz Mayr
mayr@ort.edu.uy

Universidad ORT Uruguay
12 de marzo de 2023

1. Introducción al Aprendizaje Reforzado

Inteligencia Artificial

- Computing Machinery and Intelligence.
  Alan Turing, Mind, 1950.
  `https://academic.oup.com/mind/article/LIX/236/433/986238`

Inteligencia Artificial

- Computing Machinery and Intelligence.
  Alan Turing, Mind, 1950.
  `https://academic.oup.com/mind/article/LIX/236/433/986238`
- Darthmouth Summer Research Project on Artificial Intelligence.
  J. McCarthy, M. Minsky, N. Rochester, C. Shannon, 1956.
  `https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth`

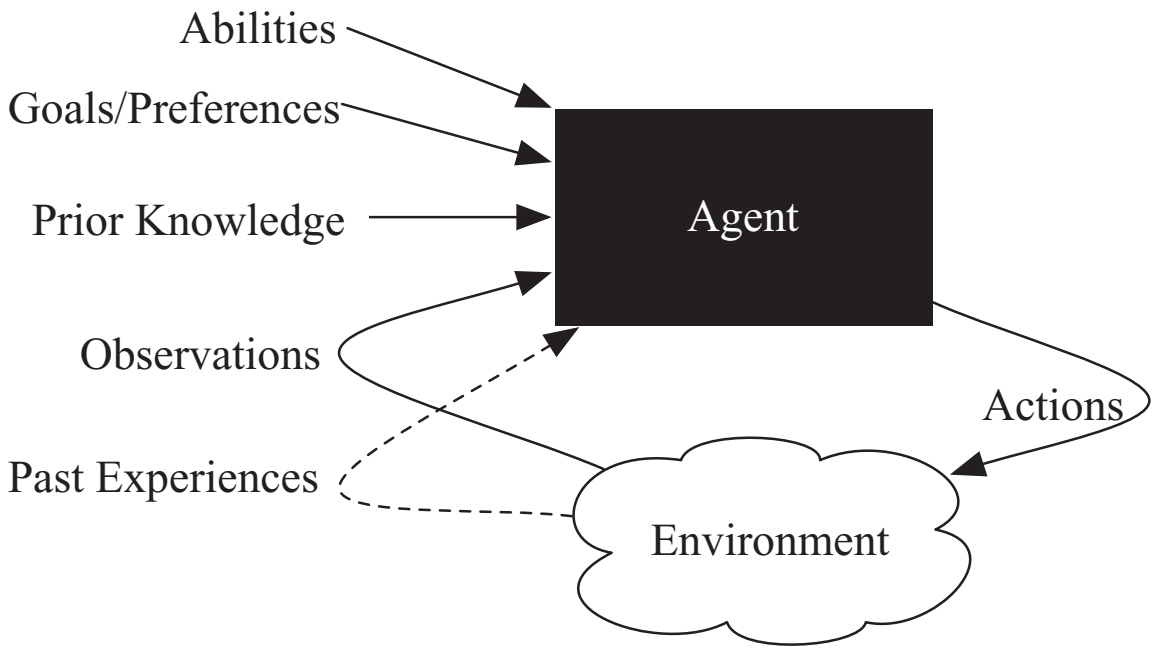
¿Qué es Inteligencia Artificial?

	Como humano	**Racionalmente**
Pensar	Enfoque cognitivo	Deducción lógica
Actuar	Test de Turing	Lograr el mejor resultado esperado

- S. Russel & P. Norvig (2010), Artificial Intelligence: A Modern Approach, 3$^{\text{ra}}$ ed., Prentice Hall, *capítulo 1*.

Agente

- Un **agente** es una entidad que interactúa con un **ambiente**



- Poole, D., Mackworth, A. (2010). Artificial Intelligence Foundations of Computational Agents. Cambridge University Press. Fig. 1.3.

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**
- Racionalidad: lograr el *mejor* resultado *esperado*

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**
- Racionalidad: lograr el *mejor* resultado *esperado*
  - Maximizar la *utilidad*

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**
- Racionalidad: lograr el *mejor* resultado *esperado*
  - Maximizar la *utilidad*
  - Minimizar la *pérdida*

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**
- Racionalidad: lograr el *mejor* resultado *esperado*
  - Maximizar la *utilidad*
  - Minimizar la *pérdida*
- El agente tiene que resolver un problema: saber qué hacer en cada situación

Agente Inteligente (o Racional)

- Un agente **inteligente** es el que **actúa racionalmente**
- Racionalidad: lograr el *mejor* resultado *esperado*
  - Maximizar la *utilidad*
  - Minimizar la *pérdida*
- El agente tiene que resolver un problema: saber qué hacer en cada situación
- El *aprendizaje* en general, y el *aprendizaje reforzado*, en particular, es una manera de hacerlo

Aprendizaje Reforzado: Ejemplos

- Helicóptero autónomo (Stanford University, 2008)
  `http://heli.stanford.edu/`

Aprendizaje Reforzado: Ejemplos

- Helicóptero autónomo (Stanford University, 2008)
  `http://heli.stanford.edu/`
- Juegos de Atari (Deepmind, 2013)
  `https://www.youtube.com/watch?v=V1eYniJ0Rnk&t=17s`

Aprendizaje Reforzado: Ejemplos

- Helicóptero autónomo (Stanford University, 2008)
  `http://heli.stanford.edu/`
- Juegos de Atari (Deepmind, 2013)
  `https://www.youtube.com/watch?v=V1eYniJ0Rnk&t=17s`
- Juego del Go (AlphaGo, Deepmind, 2016)
  `https://deepmind.com/research/alphago/`

Aprendizaje Reforzado: Ejemplos

- Helicóptero autónomo (Stanford University, 2008)
  `http://heli.stanford.edu/`
- Juegos de Atari (Deepmind, 2013)
  `https://www.youtube.com/watch?v=V1eYniJ0Rnk&t=17s`
- Juego del Go (AlphaGo, Deepmind, 2016)
  `https://deepmind.com/research/alphago/`
- Juego del escondite multi-agente (OpenAI, 2019)
  `https://www.youtube.com/watch?v=kopoLzvh5jY`

Aprendizaje Reforzado: Ejemplos

- Helicóptero autónomo (Stanford University, 2008)
  `http://heli.stanford.edu/`
- Juegos de Atari (Deepmind, 2013)
  `https://www.youtube.com/watch?v=V1eYniJ0Rnk&t=17s`
- Juego del Go (AlphaGo, Deepmind, 2016)
  `https://deepmind.com/research/alphago/`
- Juego del escondite multi-agente (OpenAI, 2019)
  `https://www.youtube.com/watch?v=kopoLzvh5jY`
- Juego multi-agente Dota 2 (OpenAI, 2019)
  `https://openai.com/blog/openai-five/`

Aprendizaje Reforzado

Todos estos sistemas de Aprendizaje Reforzado tienen:

- Agente
- Ambiente
- Acción
- Recompensa

Aprendizaje Reforzado

Todos estos sistemas de Aprendizaje Reforzado tienen:

- Agente
- Ambiente
- Acción
- Recompensa

El Aprendizaje Reforzado consiste en aprender a mapear situaciones a acciones, con el objetivo de maximizar la recompensa a largo plazo.

Aprendizaje Reforzado

Todos estos sistemas de Aprendizaje Reforzado tienen:

- Agente
- Ambiente
- Acción
- Recompensa

El Aprendizaje Reforzado consiste en aprender a mapear situaciones a acciones, con el objetivo de maximizar la recompensa a largo plazo.

Características básicas del Aprendizaje Reforzado:

1. se aprende por prueba y error,
2. las recompensas pueden demorar en llegar,
3. debe buscarse un balance entre exploración y explotación.

Aprendizaje Reforzado

Imaginen jugar a un juego nuevo, del cual desconocen las reglas. Después de algo así como 100 movidas, tu oponente anuncia que perdiste. Eso es un buen resumen del aprendizaje reforzado.

Stuart Russel & Peter Norvig

Traducido de S. Russel & P. Norvig, "Artificial Intelligence: A Modern Approach", 3ra edición, Prentice Hall, 2010, pág. 831.

Aprendizaje Reforzado vs. Supervisado

Aprendizaje Supervisado:

Consiste en aprender a predecir un valor (ej.: la clase) de las instancias, a partir de un conjunto provisto de datos etiquetados (ej.: con la clase correcta de cada instancia).

Aprendizaje Reforzado vs. Supervisado

Aprendizaje Supervisado:

Consiste en aprender a predecir un valor (ej.: la clase) de las instancias, a partir de un conjunto provisto de datos etiquetados (ej.: con la clase correcta de cada instancia).

Aprendizaje Reforzado:

El agente va juntando datos a partir de su experiencia, en forma interactiva.

Las acciones suelen tener un efecto sobre el ambiente.

El éxito en una tarea suele ocurrir a largo plazo, no es inmediato.

Aprendizaje Reforzado vs. No Supervisado

Aprendizaje No Supervisado:

Consiste en buscar estructura o información útil en datos no etiquetados.

Por ejemplo: clustering y transformación dimensional.

Aprendizaje Reforzado vs. No Supervisado

Aprendizaje No Supervisado:

Consiste en buscar estructura o información útil en datos no etiquetados.

Por ejemplo: clustering y transformación dimensional.

Aprendizaje Reforzado:

El agente también realiza una búsqueda en la oscuridad, sin supervisión directa, pero sus objetivos son muy distintos.

No quiere entender los datos, sino aprender comportamientos en un mundo desconocido (parcial o totalmente).

Elementos del Aprendizaje Reforzado

Elementos del Aprendizaje Reforzado

Política

- Define el comportamiento del agente.
- Mapea estados del ambiente a acciones.
- En Psicología, se lo denomina “reglas estímulo-respuesta”.
- En general son estocásticas, asignando a cada acción una probabilidad.

Elementos del Aprendizaje Reforzado

Política

- Define el comportamiento del agente.
- Mapea estados del ambiente a acciones.
- En Psicología, se lo denomina “reglas estímulo-respuesta”.
- En general son estocásticas, asignando a cada acción una probabilidad.

Señal de recompensa

- En cada instante de tiempo, el ambiente envía al agente un valor numérico llamado recompensa.
- El objetivo del agente es maximizar la recompensa total acumulada, a largo plazo.
- Es una función estocástica dependiente del estado del ambiente y de las acciones tomadas.

Elementos del Aprendizaje Reforzado (cont.)

Funciones de valor

- Determinan cuán buenos resultan un estado y/o una acción, a largo plazo.
- Son la sumas de recompensas que el agente espera acumular en el futuro, a partir de cierto estado y/o cierta acción.
- No son directamente observables; el agente debe aprender a estimarlas.

Elementos del Aprendizaje Reforzado (cont.)

Funciones de valor

- Determinan cuán buenos resultan un estado y/o una acción, a largo plazo.
- Son la sumas de recompensas que el agente espera acumular en el futuro, a partir de cierto estado y/o cierta acción.
- No son directamente observables; el agente debe aprender a estimarlas.

Modelo del ambiente (opcional)

- Es algo que el agente puede usar para predecir cómo reaccionará el ambiente ante una acción determinada.
- Cuando contamos con un modelo del ambiente, podemos usar métodos *model-based*; de lo contrario, debemos usar métodos *model-free*.

Breve historia del Aprendizaje Reforzado

Dos corrientes independientes:

- (1950s-) Teoría del control óptimo. Cómo diseñar un controlador capaz de maximizar/minimizar alguna métrica de un sistema dinámico en el tiempo. Principalmente offline; ej: programación dinámica.

Breve historia del Aprendizaje Reforzado

Dos corrientes independientes:

- (1950s-) Teoría del control óptimo. Cómo diseñar un controlador capaz de maximizar/minimizar alguna métrica de un sistema dinámico en el tiempo. Principalmente offline; ej: programación dinámica.
- (1850s-) Aprendizaje por prueba y error. Estudio del comportamiento animal. Psicología experimental: estímulos, recompensas, patrones de conducta.

Breve historia del Aprendizaje Reforzado

Dos corrientes independientes:

- (1950s-) Teoría del control óptimo. Cómo diseñar un controlador capaz de maximizar/minimizar alguna métrica de un sistema dinámico en el tiempo. Principalmente offline; ej: programación dinámica.
- (1850s-) Aprendizaje por prueba y error. Estudio del comportamiento animal. Psicología experimental: estímulos, recompensas, patrones de conducta.

En los 1980s comenzaron a juntarse ámbas corrientes, en cierta forma unificadas con la aparición de los métodos de aprendizaje por diferencias temporales.

La sección 1.7 del libro de S&B tiene una descripción detallada y muy recomendable de la historia del área.

Bibliografía

- **R.S. Sutton & A.G. Barto (2018), Reinforcement Learning. An Introduction, MIT Press, 2nd ed.**
- Poole, D., Mackworth, A. (2010). Artificial Intelligence Foundations of Computational Agents. Cambridge University Press.
- Szepesvári, C. (2010). Algorithms for Reinforcement Learning. Morgan & Claypool.
- S. Russel & P. Norvig (2010), Artificial Intelligence: A Modern Approach, 3$^{\text{ra}}$ ed., Prentice Hall, *capítulo 21.*
- T.M. Mitchell (1997), Machine Learning, McGraw-Hill, *capítulo 13.*

Temas de las clases teóricas

1. Introducción; bandidos de k brazos.
2. Procesos de decisión de Markov; programación dinámica.
3. Métodos Monte Carlo.
4. Métodos de diferencias temporales (Q-learning).
5. Métodos de aproximación de función de valor.
6. Bootstrapping de n pasos.
7. Aprendizaje y planificación (Dyna-Q, Monte Carlo Tree Search).
8. Métodos de gradientes; Reinforce; Actor-critic.

Agentes Inteligentes

Franz Mayr
mayr@ort.edu.uy

Universidad ORT Uruguay
12 de marzo de 2023

2. Bandidos de k brazos

Máquina tragamonedas *(bandido)*

Al accionar la palanca, una máquina tragamonedas paga una recompensa numérica (ej.: \$0, \$5) según una función estocástica que podemos suponer estacionaria (que no varía en el tiempo).

Bandidos de k brazos

Bandidos de k brazos

- Acción: elegir una palanca (o *brazo*) entre k posibles.

Bandidos de k brazos

- Acción: elegir una palanca (o *brazo*) entre k posibles.
- Cada palanca otorga una recompensa numérica, según su propia distribución de probabilidad.
- Todas las distribuciones son estacionarias (no varían en el tiempo) y desconocidas.
- Cada acción es independiente de las anteriores.

Bandidos de k brazos

- Acción: elegir una palanca (o *brazo*) entre k posibles.
- Cada palanca otorga una recompensa numérica, según su propia distribución de probabilidad.
- Todas las distribuciones son estacionarias (no varían en el tiempo) y desconocidas.
- Cada acción es independiente de las anteriores.
- **Objetivo:** Maximizar las recompensas acumuladas sobre cierto período de tiempo.

Bandidos de k brazos

Sea A_t la acción elegida en el tiempo t, con recompensa R_t.

Bandidos de k brazos

Sea A_t la acción elegida en el tiempo t, con recompensa R_t.

El valor de una acción $a \in \mathcal{A}$ se define como:

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

Bandidos de k brazos

Sea A_t la acción elegida en el tiempo t, con recompensa R_t.

El valor de una acción $a \in \mathcal{A}$ se define como:

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

$Q_t(a)$ es una estimación de $q_*(a)$ en el instante de tiempo t.

Bandidos de k brazos

Sea A_t la acción elegida en el tiempo t, con recompensa R_t.

El valor de una acción $a \in \mathcal{A}$ se define como:

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

$Q_t(a)$ es una estimación de $q_*(a)$ en el instante de tiempo t.

En el instante t, llamamos acción greedy a una acción a que tiene máximo valor estimado $Q_t(a)$. (Observación: Puede haber más de una acción greedy.)

Bandidos de k brazos

Sea A_t la acción elegida en el tiempo t, con recompensa R_t.

El valor de una acción $a \in \mathcal{A}$ se define como:

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

$Q_t(a)$ es una estimación de $q_*(a)$ en el instante de tiempo t.

En el instante t, llamamos acción greedy a una acción a que tiene máximo valor estimado $Q_t(a)$. (Observación: Puede haber más de una acción greedy.)

- Explotación: Elegir una acción greedy.
- Exploración: Elegir otra acción.

En el Aprendizaje Reforzado es importante buscar un balance entre exploración y explotación. Con la metáfora de los bandidos de k brazos podemos estudiarlo de manera simple.

Métodos action-value

Los métodos action-value estiman $Q_t(a)$ y usan esas estimaciones para seleccionar las acciones a ejecutar.

Métodos action-value

Los métodos action-value estiman $Q_t(a)$ y usan esas estimaciones para seleccionar las acciones a ejecutar.

La forma más sencilla de computar la estimación $Q_t(a)$ es calcular el promedio de recompensas obtenidas al seguir la acción a:

Métodos action-value

Los métodos action-value estiman $Q_t(a)$ y usan esas estimaciones para seleccionar las acciones a ejecutar.

La forma más sencilla de computar la estimación $Q_t(a)$ es calcular el promedio de recompensas obtenidas al seguir la acción a:

$$Q_t(a) \doteq \frac{\text{suma de recompensas al tomar } a \text{ antes de } t}{\text{cantidad de veces que se tomó } a \text{ antes de } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

(o bien 0 si el denominador es 0)

Métodos action-value

Los métodos action-value estiman $Q_t(a)$ y usan esas estimaciones para seleccionar las acciones a ejecutar.

La forma más sencilla de computar la estimación $Q_t(a)$ es calcular el promedio de recompensas obtenidas al seguir la acción a:

$$Q_t(a) \doteq \frac{\text{suma de recompensas al tomar } a \text{ antes de } t}{\text{cantidad de veces que se tomó } a \text{ antes de } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

(o bien 0 si el denominador es 0)

Observación: A medida que crece el denominador, $Q_t(a)$ converge a $q_*(a)$, gracias a la ley de los grandes números.

Métodos action-value

Si ya computamos $Q_t(a)$ para todas las acciones disponibles, podemos usarla para seleccionar una acción greedy:

$$A_t \doteq \arg\max_a Q_t(a)$$

(con desempate aleatorio)

Métodos action-value

Si ya computamos $Q_t(a)$ para todas las acciones disponibles, podemos usarla para seleccionar una acción greedy:

$$A_t \doteq \arg\max_a Q_t(a)$$

(con desempate aleatorio)

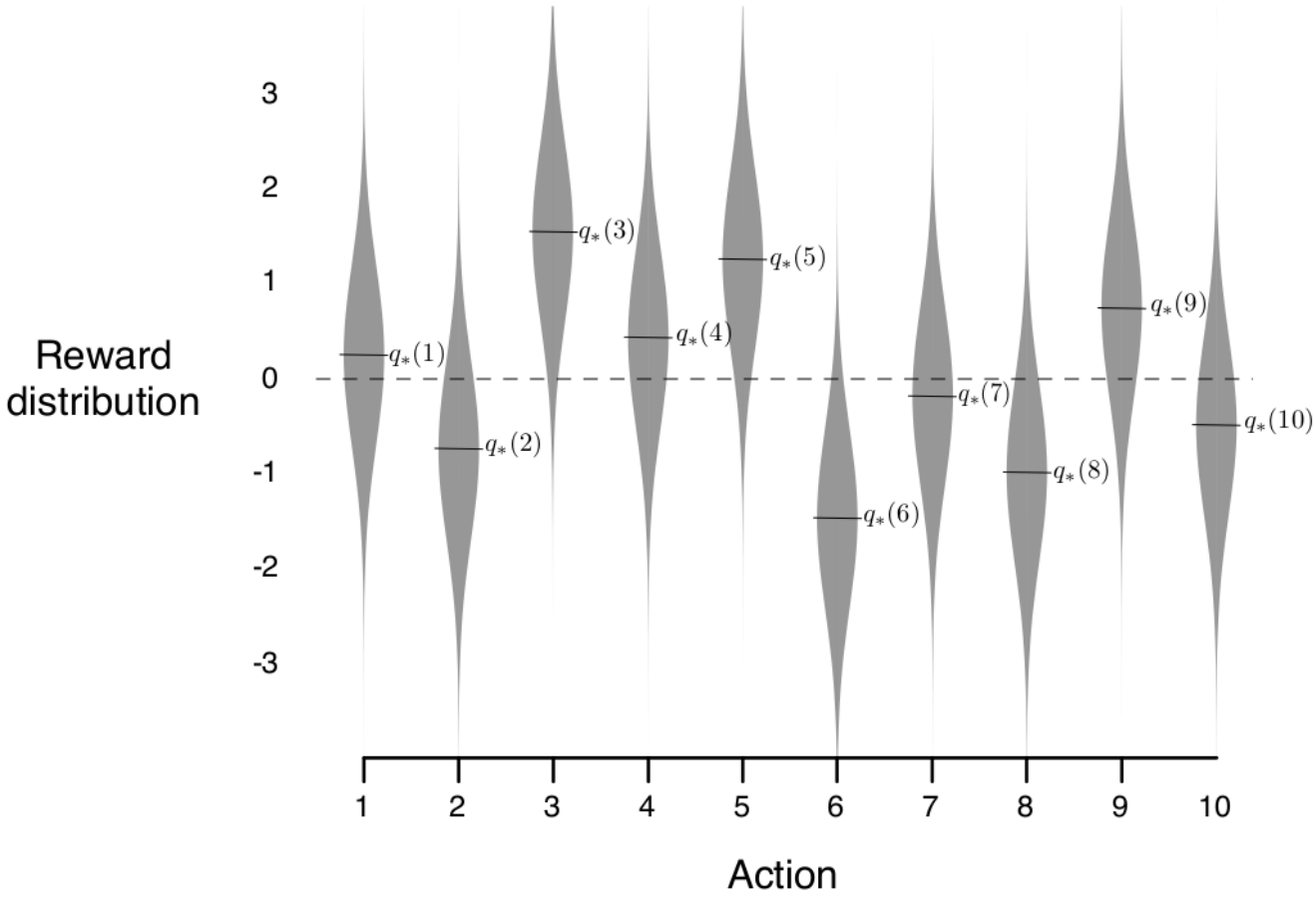
Otra forma de usar $Q_t(a)$ para elegir una acción:

- Con probabilidad ε (baja), se elige una acción al azar.
- Con probabilidad $1 - \varepsilon$, se elige una acción greedy.

A este método se lo conoce como ε-greedy.

Ejemplo: 10-armed bandits

Simulación de bandidos de k brazos ($k = 10$). Los valores $q_*(a)$ de las acciones $a = 1, ..., 10$ se muestran a continuación:



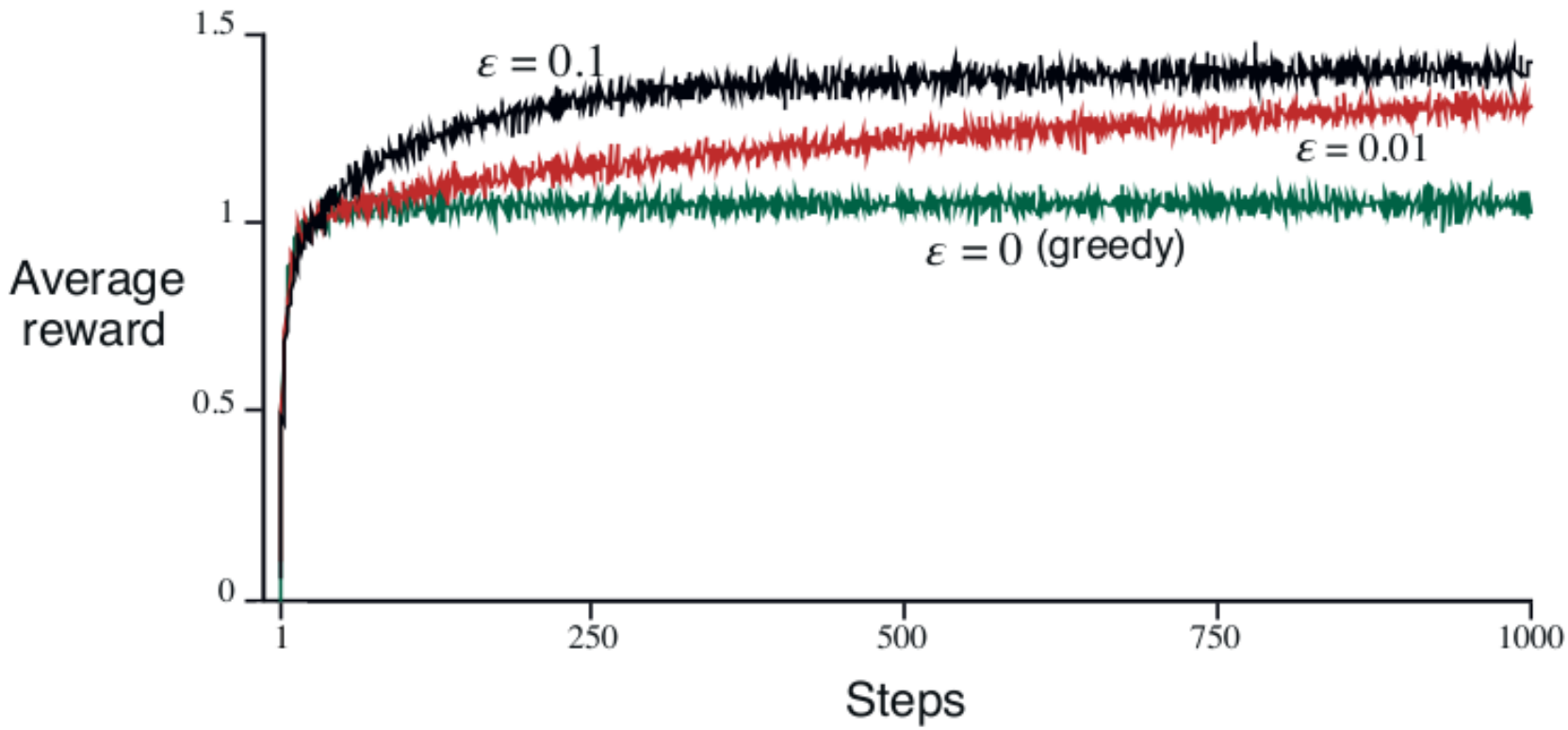
Ejemplo: 10-armed bandits

Una ejecución = Aplicación durante 1000 instantes de tiempo de un método de aprendizaje ε-greedy.

Ejemplo: 10-armed bandits

Una ejecución = Aplicación durante 1000 instantes de tiempo de un método de aprendizaje ε-greedy.

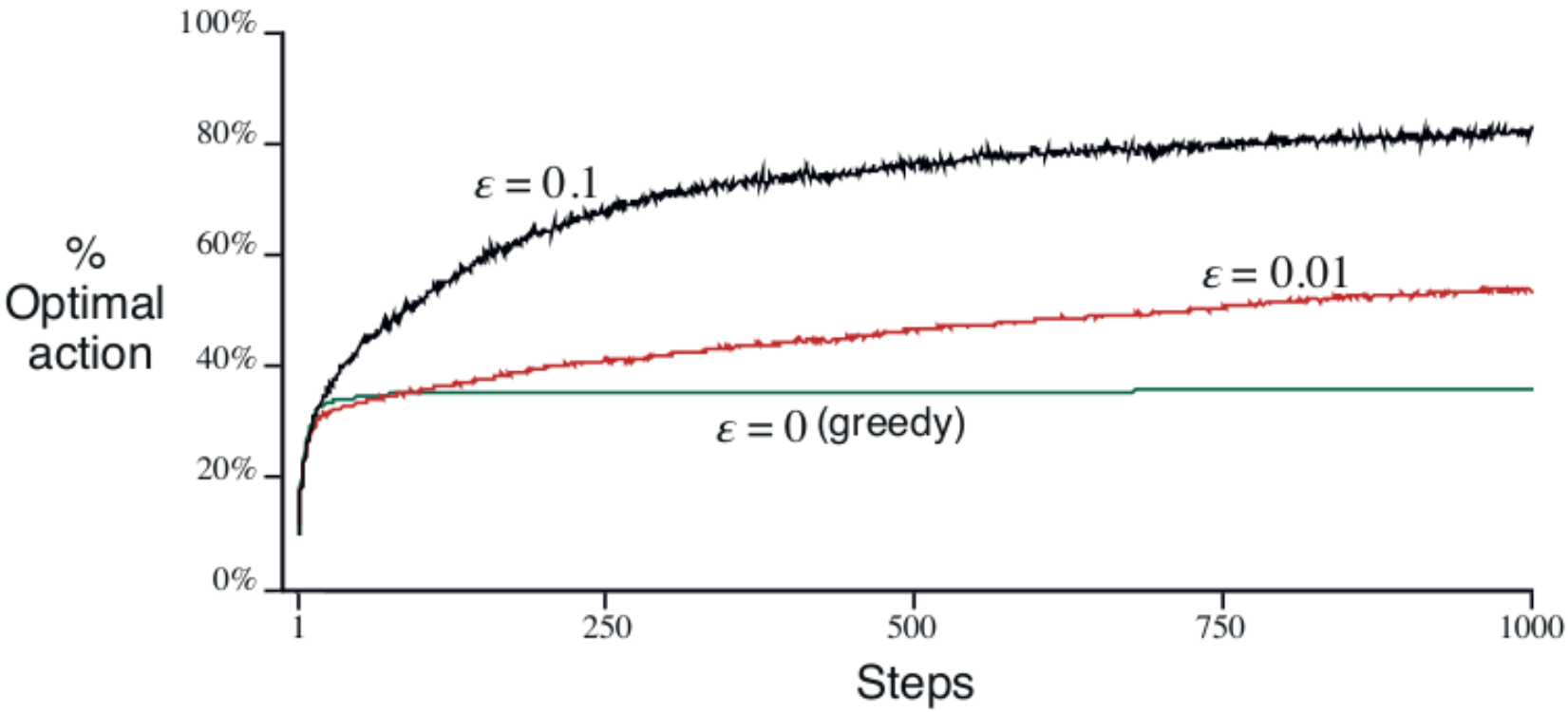
Evolución de la recompensa obtenida en cada instante, promediada sobre 2000 ejecuciones:



Ejemplo: 10-armed bandits

Una ejecución = Aplicación durante 1000 instantes de tiempo de un método de aprendizaje ε-greedy.

Evolución del porcentaje de elecciones óptimas, promediado sobre 2000 ejecuciones:



Implementación incremental

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\Big(R_n + \sum_{i=1}^{n-1} R_i\Big) \\
&= \frac{1}{n}\Big(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\Big) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]
\end{aligned}
$$

(para $n = 1$: $Q_2 = R_1$ con Q_1 arbitrario)

Esto requiere almacenar sólo Q_n y n en memoria, y la actualización luego de tomar cada acción es $O(1)$. Mucho mejor que almacenar todas las recompensas.

Bandidos de k brazos: Algoritmo

Inicializar, para cada $a = 1 \,..\, k$:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Repetir:

$$A \leftarrow \begin{cases} \text{una acción al azar} & \text{con probabilidad } \varepsilon \\ \arg\max_a Q(a) & \text{con probabilidad } 1 - \varepsilon \end{cases}$$

$R \leftarrow \mathit{bandido}(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)} \big[R - Q(A)\big]$

(Si hay empate en el caso greedy $(1 - \varepsilon)$, se desempata al azar.)

Actualización de estimaciones

$$Q_{n+1} = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]$$

Actualización de estimaciones

$$Q_{n+1} = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]$$

Esta forma de actualizar estimaciones es frecuente en el Aprendizaje Reforzado:

$$\boxed{NuevaEst = ViejaEst + TasaAct\Big[Objetivo - ViejaEst\Big]}$$

Actualización de estimaciones

$$Q_{n+1} = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]$$

Esta forma de actualizar estimaciones es frecuente en el Aprendizaje Reforzado:

$$\boxed{\mathit{NuevaEst} \leftarrow \mathit{ViejaEst} + \mathit{TasaAct}\,\Big[\mathit{Objetivo} - \mathit{ViejaEst}\Big]}$$

donde:

- *ViejaEst* y *NuevaEst* son las estimaciones vieja y nueva.
- *TasaAct* es una tasa de actualización; suele denotarse α.
- $\Big[\mathit{Objetivo} - \mathit{ViejaEst}\Big]$ se denomina error de estimación.

Problemas no estacionarios

El método visto es efectivo para problemas estacionarios, en los cuales los valores de las acciones no varían en el tiempo.

Para problemas no estacionarios (el caso más realista), tiene sentido dar mayor peso a recompensas recientes al actualizar nuestras estimaciones.

Problemas no estacionarios

El método visto es efectivo para problemas estacionarios, en los cuales los valores de las acciones no varían en el tiempo.

Para problemas no estacionarios (el caso más realista), tiene sentido dar mayor peso a recompensas recientes al actualizar nuestras estimaciones.

Por ejemplo, puede usarse una tasa $\alpha \in (0, 1]$ constante:

$$\begin{aligned} Q_{n+1} &\doteq Q_n + \alpha\Big[R_n - Q_n\Big] \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \end{aligned}$$

(ver próxima página)

Q_{n+1} es el promedio ponderado de recompensas pasadas.

Problemas no estacionarios

El método visto es efectivo para problemas estacionarios, en los cuales los valores de las acciones no varían en el tiempo.

Para problemas no estacionarios (el caso más realista), tiene sentido dar mayor peso a recompensas recientes al actualizar nuestras estimaciones.

Por ejemplo, puede usarse una tasa $\alpha \in (0, 1]$ constante:

$$\begin{aligned} Q_{n+1} &\doteq Q_n + \alpha\Big[R_n - Q_n\Big] \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \end{aligned}$$

(ver próxima página)

Q_{n+1} es el promedio ponderado de recompensas pasadas.

Observación: Como $1 - \alpha < 1$, entonces $(1-\alpha)^{n-i}$ decrece exponencialmente cuando $i \to 0$ (Ejercicio: probarlo).

Problemas no estacionarios

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

Valores iniciales optimistas

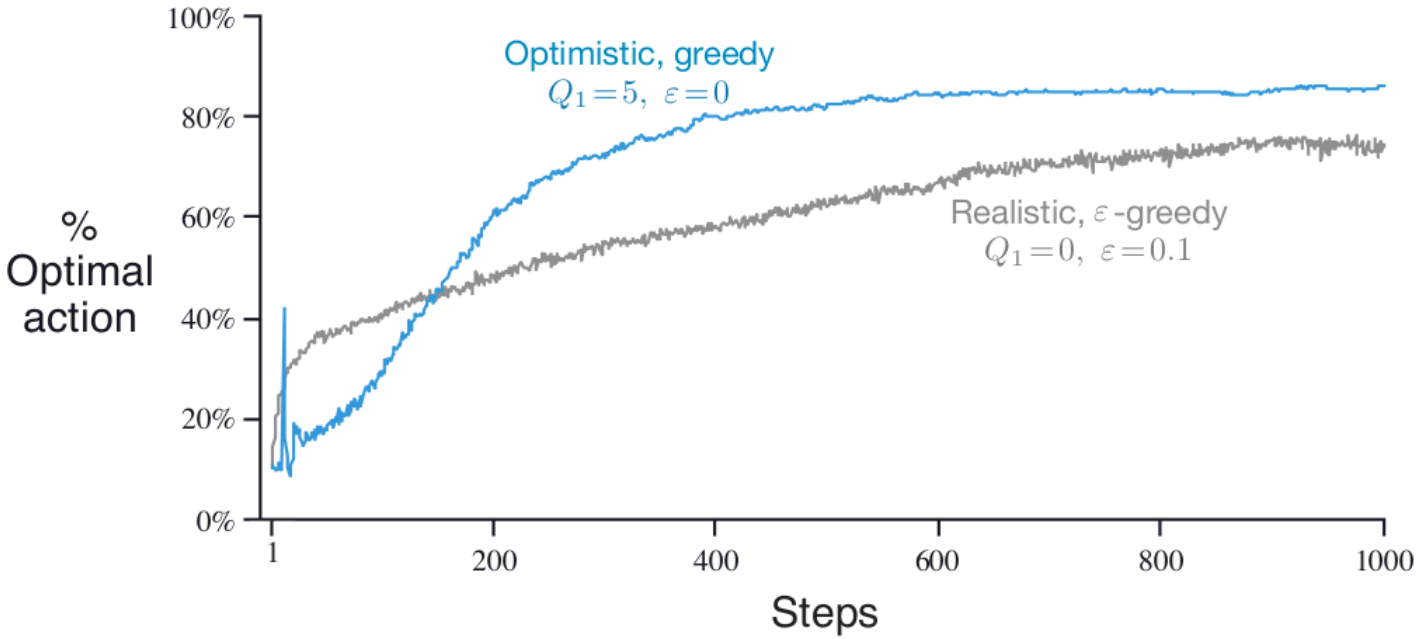
Los valores iniciales de las acciones (es decir, $Q_1(a)$) pueden usarse para guiar o alentar la exploración inicial.

Volviendo al ejemplo, con Q_1 muy alto para todas las acciones, forzamos la exploración de todas las acciones al menos una vez.

Valores iniciales optimistas

Los valores iniciales de las acciones (es decir, $Q_1(a)$) pueden usarse para guiar o alentar la exploración inicial.

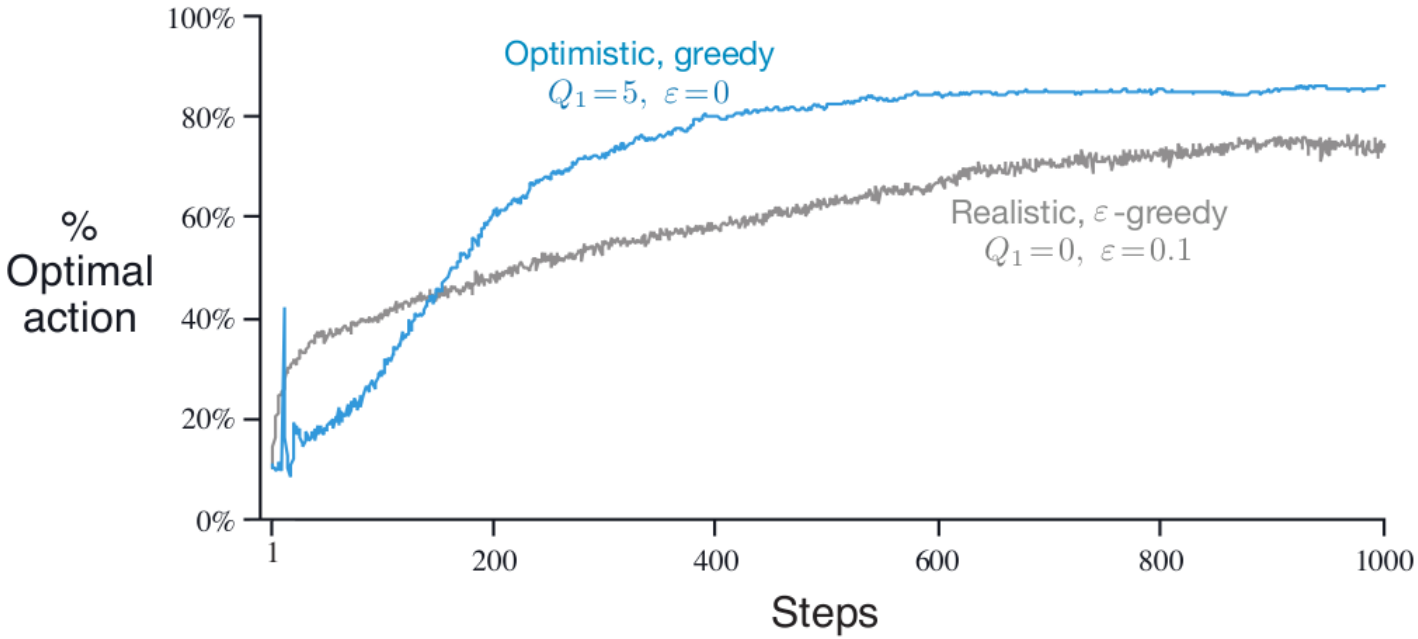
Volviendo al ejemplo, con Q_1 muy alto para todas las acciones, forzamos la exploración de todas las acciones al menos una vez.



Valores iniciales optimistas

Los valores iniciales de las acciones (es decir, $Q_1(a)$) pueden usarse para guiar o alentar la exploración inicial.

Volviendo al ejemplo, con Q_1 muy alto para todas las acciones, forzamos la exploración de todas las acciones al menos una vez.



Ejercicio: El pico al principio de la curva azul no es ruido del azar (recordar que cada curva promedia 2000 ejecuciones). ¿A qué se debe?

Métodos Upper-Confidence-Bound (UCB)

En los métodos ε-greedy, la elección aleatoria de acciones se hace sin ningún criterio. Sería mejor orientar la exploración para reducir la incertidumbre de las acciones.

Un ejemplo es el método upper-confidence-bound (UCB, o cota superior de confianza) para selección de acciones:

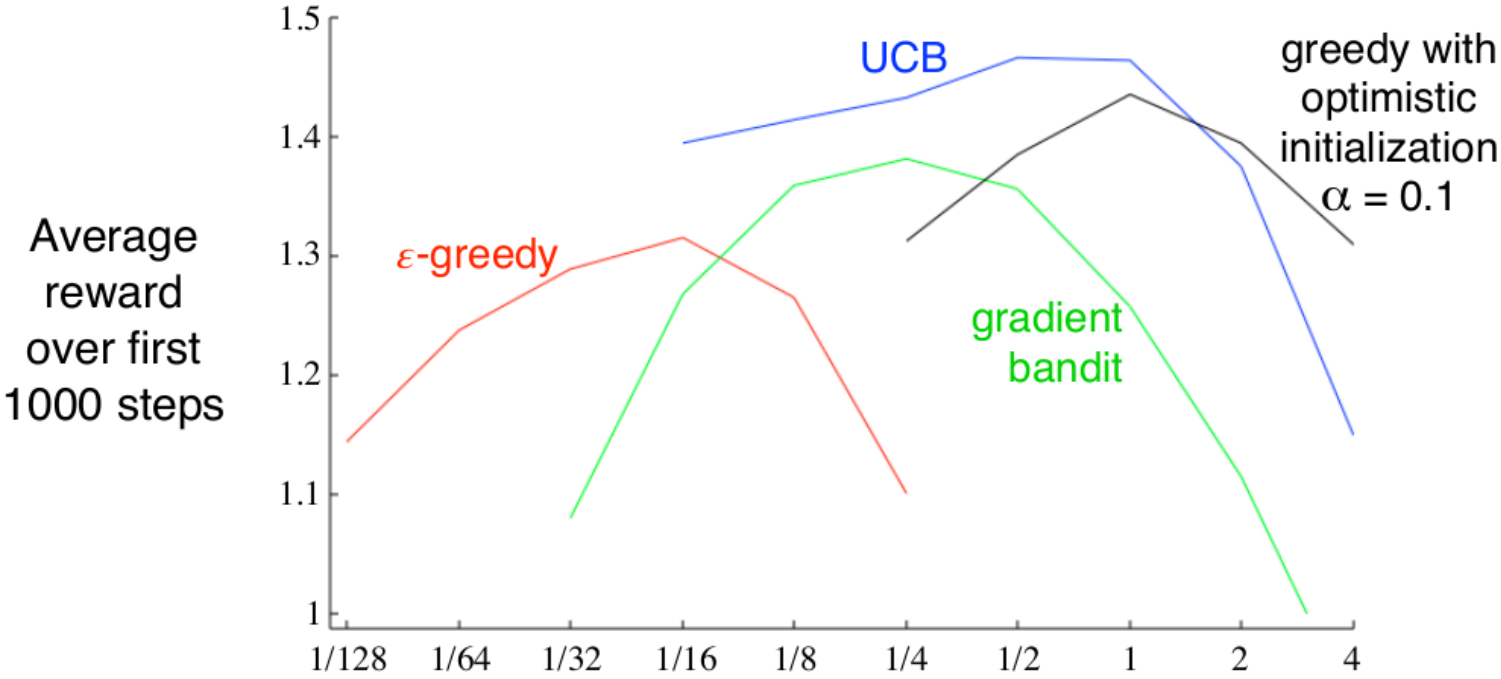
$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

El parámetro c controla el grado de exploración.

Si $N_t(a) = 0$, a es una acción seleccionable.

Comparación de métodos

Estudio comparativo de los parámetros de los algoritmos vistos:



Cada punto es la recompensa promedio obtenida tras 1000 instantes de tiempo con un algoritmo dado y con el valor especificado para el parámetro correspondiente.

Resumen - Bandidos de k brazos

- Acciones, recompensas estocásticas con distribución estacionaria y desconocida.
- Valor de una acción: $q_*(a)$; y su estimación: $Q_t(a)$.
- Dilema de exploración vs. explotación.
- Métodos action-value y selección ε-greedy de acciones.
- Fórmula general de actualización de estimaciones en Aprendizaje Reforzado.
- Problemas no estacionarios: ponderación de recompensas más recientes.
- Variantes: valores iniciales optimistas, UCB.