

Agenda

Modelo / Simulador
fct. de R
MLI

1. Depuración de los algoritmos de RL
2. Algoritmos de Búsqueda de Política Directa (Policy Search)
3. *Multi-arm Bandits*
4. Deep Q Learning
5. Taller práctico (Tarea incluida)

Multi-Arm Bandits

Clase 5.3 Aprendiendo Acciones Simples, no secuenciales

Jorge Vasquez

Motivación

- ❖ Cuando las Acciones no tienen impacto en los siguientes estados

Motivación

- ❖ Acá acción resulta en una recompensa **inmediata**.
- ❖ Cuando queremos escoger acciones que maximizan nuestras recompensas inmediatas esperadas.
 - ❖ ¿Por que en modo **esperado**?
 - ❖ Porque las recompensas no son **determinísticas** (depende de una probabilidad)

Motivación

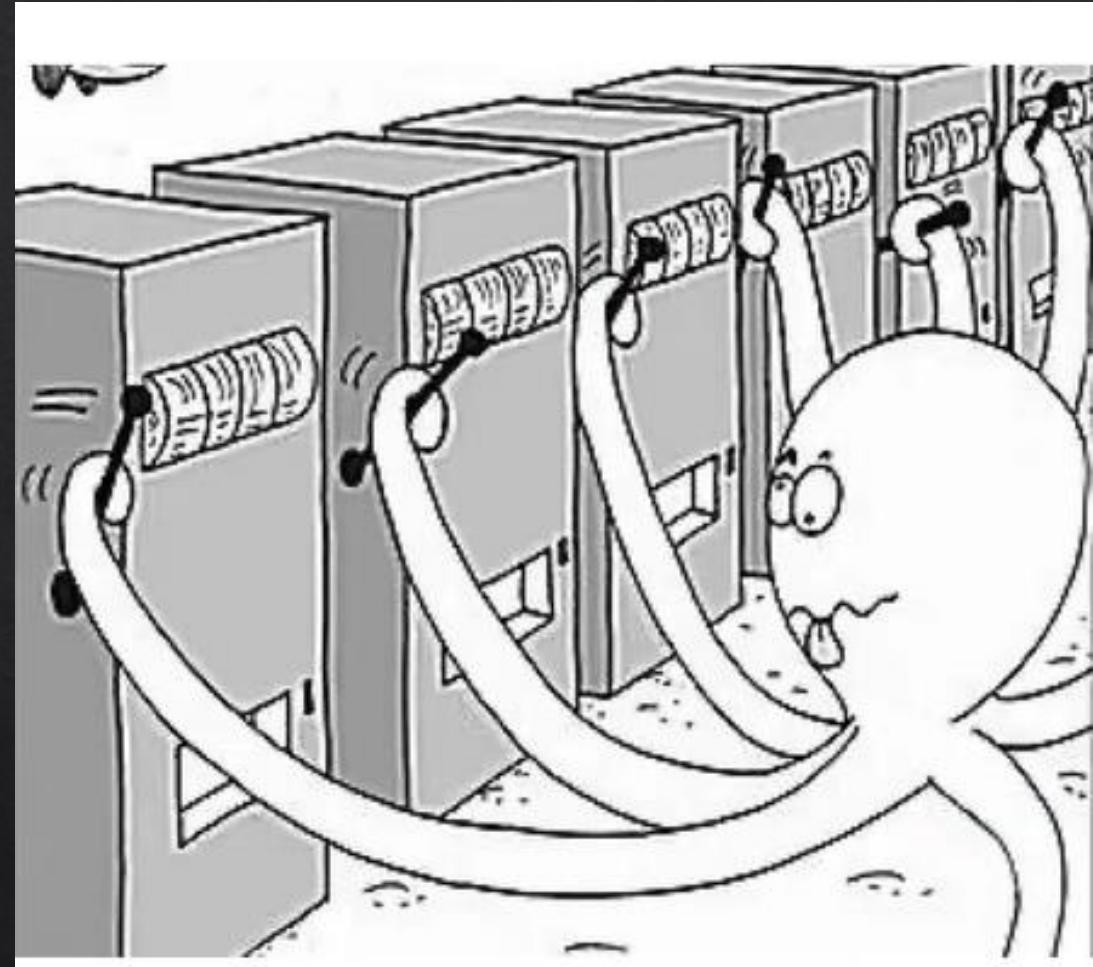
- ❖ Por ejemplo, mostrar un advertisements puede generar diferentes tasas de click en días diferentes
 - ❖ Acciones, es mostrar el advertisement
 - ❖ Recompensas, tasa de click
 - ❖ Entonces, queremos elegir el advertisement que maximize la tasa de click

Simples Acciones

- ❖ Este concepto en RL, de simple acción, se llaman multi-arm bandits, que son máquinas tragadoras de monedas
 - ❖ Queremos elegir una palanca por periodo de tiempo
 - ❖ Recompensas son dinero o no-dinero

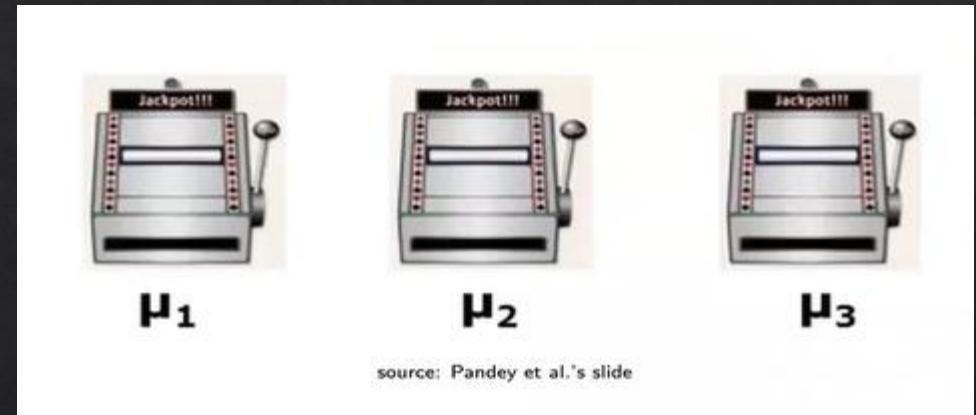


Bandits multi-brazos



Bandits Multi-Brazos

- ❖ Para cada timestep t , el agente escoge una las palancas k y juega.
 - ❖ La palanca k -esima produce una recompensa de $r_{k,t}$ cuando juega en timestep t
 - ❖ Las recompensas $r_{k,t}$ son diseñadas desde la distribución de probabilidades P_k , con una media de u_k .
 - ❖ El agente no sabe la distribución de probabilidades ni sus promedios
 - ❖ El objetivo del agente es maximizar las recompensas acumuladas



Bandidos Multi-brazos

- ◊ El agente no sabe la distribución de probabilidades ni sus promedios.
- ◊ El objetivo del agente es maximizar las recompensas acumuladas, dentro de un horizonte infinito de tiempo. O sea, encontrar la maquina o la palanca con el mas alto promedio de recompensas.



Bandidos Multi-Brazos

- ❖ Definición: La **valor de la acción** para la acción α (acá será el brazo k) es su recompensa promedio.



¿Estrategias?

Dilema de Exploración vs Explotación

- ◊ Supongamos que tenemos estimaciones:

$$Q_t(a) \approx q_*(a), \quad \forall a \quad \text{action-value estimates}$$

- ◊ No sabemos los q^*
- ◊ Q estimaciones
- ◊ q^* valores de evidencia (ground truth)

Dilema de Exploración vs Explotación

- ◊ Supongamos que tenemos estimaciones:

$$Q_t(a) \approx q_*(a), \quad \forall a \quad \text{action-value estimates}$$

- Definir la Acción más codiciosa (greedy) al tiempo t como:

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

Dilema de Exploración vs Explotación

- ◊ Supongamos que tenemos estimaciones:

$$Q_t(a) \approx q_*(a), \quad \forall a$$

- Definir la acción más codiciosa (greedy) al tiempo t como

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

If $A_t = A_t^*$ then you are *exploiting*
If $A_t \neq A_t^*$ then you are *exploring*

- ◊ No puedes hacer las dos, pero debes hacer las dos.
- ◊ No puedes parar de explorar , pero debes explorar menos a través del tiempo

Dilema de Exploración vs Explotación

- ❖ Tomar decisiones en línea envuelve esta decisión fundamental:
 - ❖ Explotación: Hacer la mejor decisión dada la información actual.
 - ❖ Exploración: Adquirir más información.
- ❖ La mejor estrategia a futura puede envolver sacrificios tempranos
- ❖ Adquirir suficiente información para tomas las mejores decisiones posibles

Dilema de Exploración vs Explotación

- ❖ Elegir un restaurant:
 - ❖ Explotación: ir a tu restaurant favorito.
 - ❖ Exploración: innovar con uno nuevo
- ❖ Buscar a los mineros:
 - ❖ Explotación: Perforar en la mejor ubicación conocida.
 - ❖ Exploración: Perforar una nueva ubicación.
- ❖ Jugar
 - ❖ Explotación: Jugar la movida que crees es la mejor.
 - ❖ Exploración: Experimentar una nueva movida.

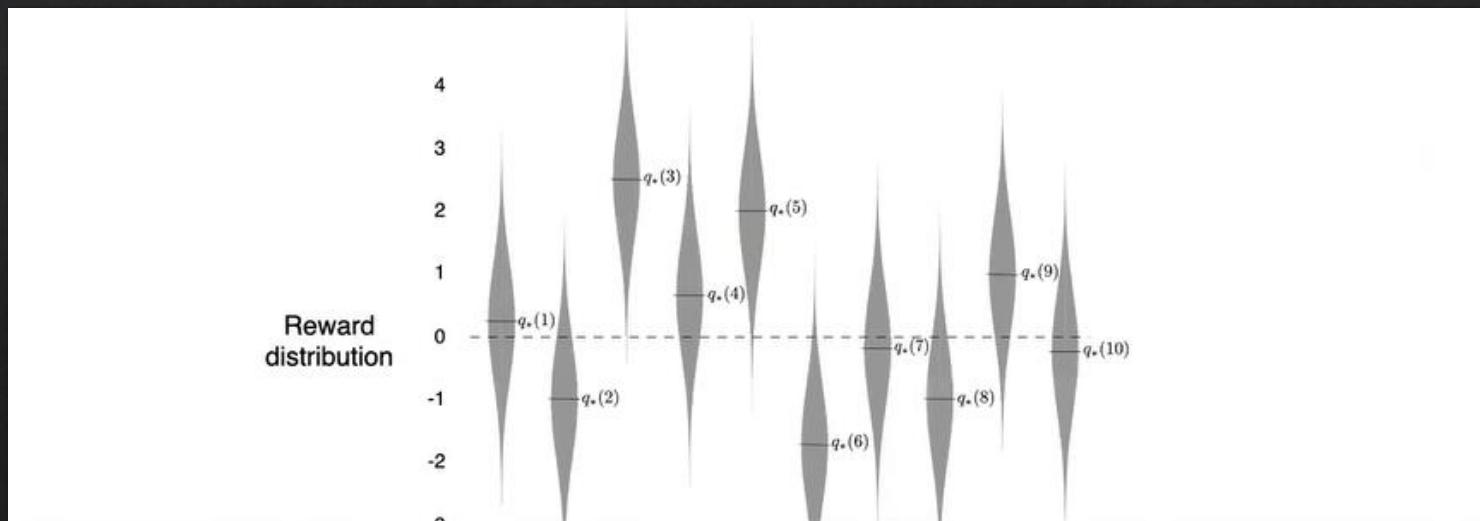
Tragadoras de Monedas con distribución de Bernoulli

- ◊ La **distribución de Bernoulli** es una distribución de probabilidad discreta de una variable *random*, toma el valor 1 con probabilidad p , y el valor 0 con probabilidad $q=1-p$, esto es la distribución de cualquier experimento simple que tiene la como respuesta si o no.
- ◊ Cada acción (tirar la palanca) resulta en éxito o fracaso (**las recompensas son binarias**).
- ◊ Promedio de cada recompensa para cada palanca representa la probabilidad de éxito.
- ◊ La acción de la palanca k pertenece a $1 \dots K$ produce un éxito con una probabilidad de $\theta_k = [0,1]$



Tragadoras de Monedas con distribución gaussiana

- ❖ Cada acción (tirar la palanca) resulta en un numero de los Reales
- ❖ Accion (brazo) k pertenece a $\{1 \dots K\}$ produce un reward promedio equivalente a la media de su distribución gaussiana.



Motivación del mundo real, Presentación de contenidos

- ◊ Tenemos variaciones para contenido web para invitaciones , A y B, y queremos decidir cual de los dos puede tratar mas usuarios.
- ◊ **Two arm bandits** : cada brazo corresponde a una variante de contenido mostrado a los usuarios.
- ◊ **Recompensas Medias** : el porcentaje total de usuarios que clicarían en cada invitación.



Motivación del mundo real, Presentación de contenidos

- ❖ Para una película particular, queremos decidir que imagen mostrar a todos los usuarios de Netflix
- ❖ Acciones: subir una de las K imágenes a la pantalla
- ❖ Recompensas Medias: el promedio de tasa de click observado.



Regret

- ◆ La **valor de la acción** es la **media de la recompensa** por la acción a

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a], \quad \forall a \in \{1, \dots, k\} \quad (\text{expected return})$$

- ◆ El valor **optimal** es:

$$v_* = q(a^*) = \max_{a \in \mathcal{A}} q_*(a)$$

- ◆ El **regret** es la oportunidad perdida para un paso de tiempo

$$I_t = \mathbb{E}[v_* - q_*(a_t)]$$

- ◆ EL **regret total** es la oportunidad total perdida

$$L_t = \mathbb{E} \left[\sum_{t=1}^T v_* - q_*(a_t) \right]$$

Regret

- ◆ Count $N_t(a)$ es el Numero de veces que la acción a ha sido seleccionada antes que el tiempo t
- ◆ Gap Δ_a es la diferencia en valor entre la acción a y la acción optimal a^* : $\Delta_a = v^* - q^*(a)$
- ◆ Regret es una función de Gaps y Counts:

$$\begin{aligned} L_t &= \mathbb{E} \left[\sum_{t=1}^T v_* - q_*(a_t) \right] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)](v_* - q_*(a)) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)]\Delta_a \end{aligned}$$

Formación de Estimaciones de Acción-Valor

- ❖ Si los enfocamos en una acción, consideramos solo recompensas y sus estimaciones después de recompensas $n+1$

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

- ❖ Esto lo podemos hacer también incremental, hacemos una suma corriente y contamos:

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

- ❖ Esto es una forma estándar de la regla de actualización (aprendizaje) :

$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize} \begin{matrix} \text{error} \\ \boxed{\text{Target} - \text{OldEstimate}} \end{matrix}$$

Actualización incremental

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

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

Bandits no estacionarias

Bandits no estacionarias

- ❖ Las funciones de recompensas estaban congeladas
- ❖ Supongamos los valores de acciones reales cambian lentamente a través del tiempo
- ❖ En este caso, promedios en muestras no son buena idea

Bandits no estacionarias

- ◆ Acá, es mejor es usar “*weighted average*”, exponencial

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

where $\alpha \in (0,1]$ and constant

- Mientras más pequeño es i , más pequeño es $(1 - \alpha)^{n-i}$ → es olvidar recompensas inmediatas

Selección de Acciones en bandits multi-arms

I. Periodo de Exploración fijo , Greedy

1. Localizar un período fijo de tiempo para explorar mientras testeas *bandits* uniformemente en forma aleatoria
2. Estimar recompensas medias para todas las acciones
3. Seleccionar la acción que es óptima para recompensas medias estimadas
4. Volver al 2

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} r_i \mathbf{1}(A_i = a)$$

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}} Q_t(a)$$

I. Periodo de Exploración fijo, Greedy

- Después de un periodo de exploración fijo ,
se forman recompensas estimadas:



$$Q_t(a_1) = 0.3$$



$$Q_t(a_2) = 0.5$$



$$Q_t(a_3) = 0.1$$

II. Selección de Acciones con ϵ -Greedy

- ❖ Selección acción tipo greedy , siempre estas explotando
- ❖ Con ϵ -greedy eres codicioso la mayoría del tiempo, pero con una probabilidad ϵ , tomas una acción aleatoria
- ❖ Esta es la forma mas simple de balancear el dilema de explotación- exploración

II. Selección de Acciones con ε -Greedy

A simple bandit algorithm

Initialize, for $a = 1$ to k :

$$\begin{aligned} Q(a) &\leftarrow 0 \\ N(a) &\leftarrow 0 \end{aligned}$$

Repeat forever:

$$A \leftarrow \begin{cases} \arg \max_a Q(a) & \text{with probability } 1 - \varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases} \quad (\text{breaking ties randomly})$$

$$R \leftarrow \text{bandit}(A)$$

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

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

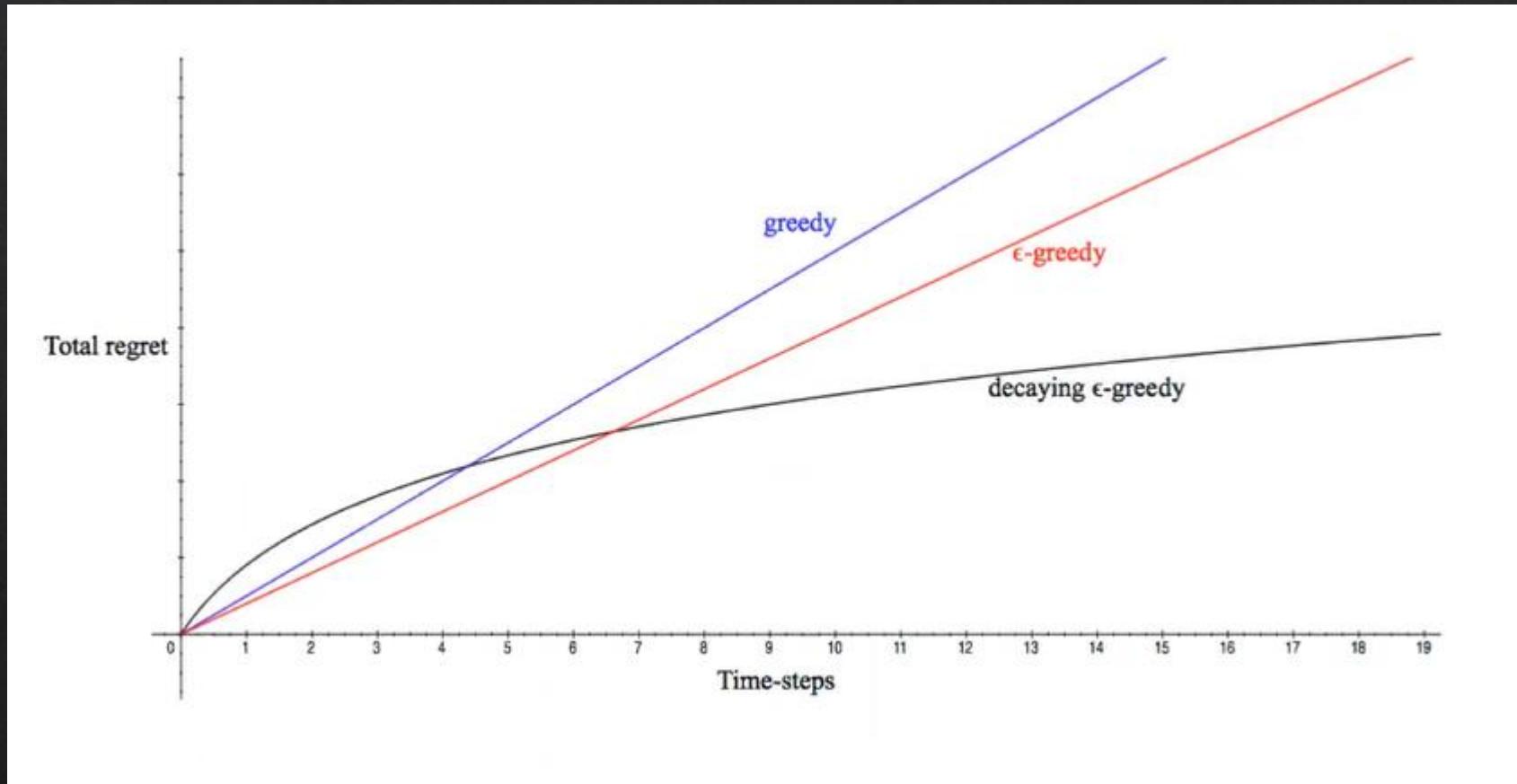
II. Selección de Acciones con ε -Greedy

- ❖ Este algoritmo continua para siempre
 - ❖ Con probabilidad $1 - \varepsilon$ seleccionar
 - ❖ Con probabilidad ε seleccionar acción aleatoria
- ❖ ε constante asegura regret mínimo
 - ❖ ε -greedy tiene regret total del tipo lineal

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}} Q_t(a)$$

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

II. Selección de Acciones con ϵ -Greedy



II. Selección de Acciones con ε -Greedy

Ejemplo con una muestra de bandits de 10 brazos

$$q_*(a) \sim \mathcal{N}(0, 1)$$
$$R_t \sim \mathcal{N}(q_*(a), 1)$$

