# Assignment 1 Bandidos de K-brazos

Luis Alejandro Medina

El presente trabajo presenta informe del resultado de la implementación de Bandido de K-brazos, en dos escenarios “estacionario” y “no estacionario”.

**Escenario Estacionario**

El primer escenario es con 10 máquinas traga monedas (Bandidos) donde cada una ofrece una recompensa entre uniform(-3,3) con media cero constante y desviación estándar constante (condición de estacionariedad).

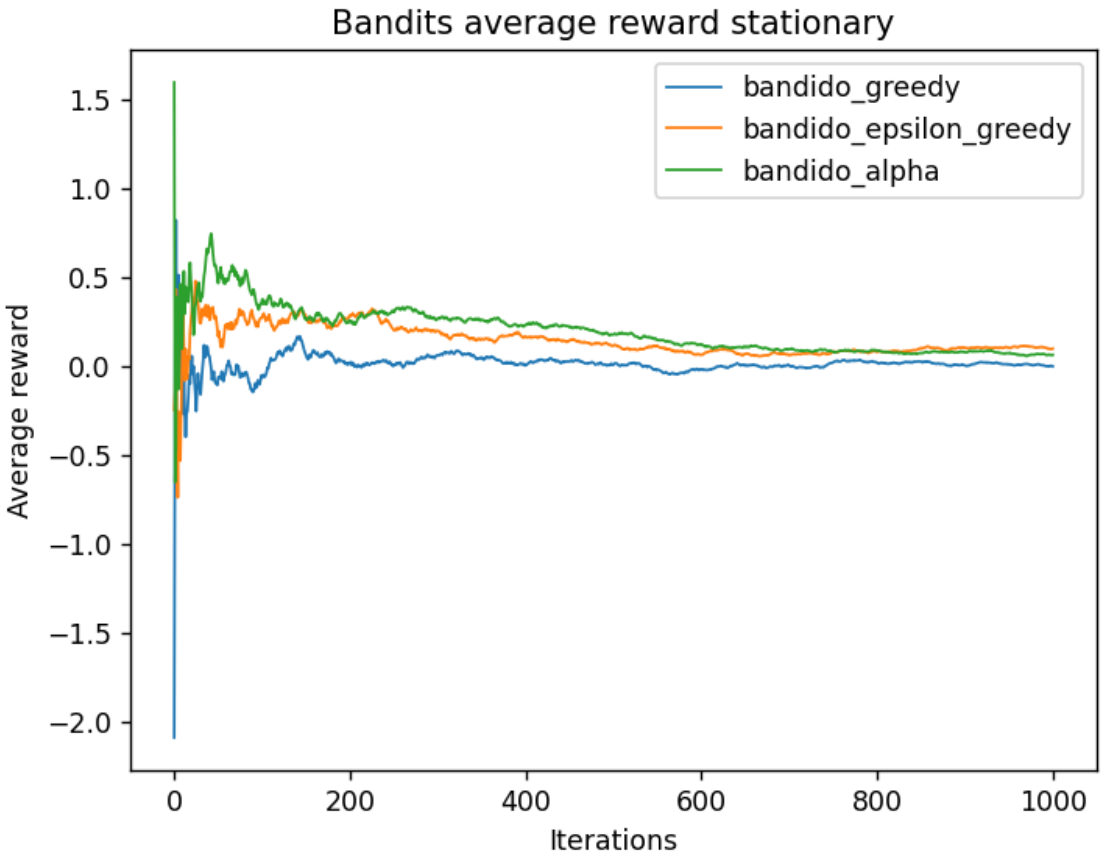
Existen 3 jugadores: Bandido_greedy, Bandido_epsilon_greedy y Bandido_alpha; El Bandido_greedy completamente codicioso (epsilon=0) solo elige la máquina con Q valor más alto y Bandido_epsilos_greedy (épsilon=0.1) que el 10% de las veces explora entre las máquinas con Q valor diferente al máximo, para ambos, el Q es definido $Q_k = Q_k + \left(\frac{1}{k}\right)[R_k - Q_k]$

Por último, el Bandido_alpha tiene un Q definido como $Q_k + alpha * [R_k - Q_k]$ con alpha=0.1 y con un grado de exploración de 10% (epsilon=0.1).

Para poder hacer la comparación entre los 3 jugadores es justo que todo tengan la misma “racha”, es decir, que las máquinas den los mismos resultados para los jugadores y sean las elecciones de los jugadores que originen las recompensas, para esto se entrega una lista de 10 listas cada una con longitud 1000 y cada valor la recompensa (R=uniform(-3,3))

$$maquinas = [maquina_1, \dots, maquina_{10}]$$

$$maquina_i = [r_1, \dots, r_{1000}] \quad r_i = uniform(-3,3)$$

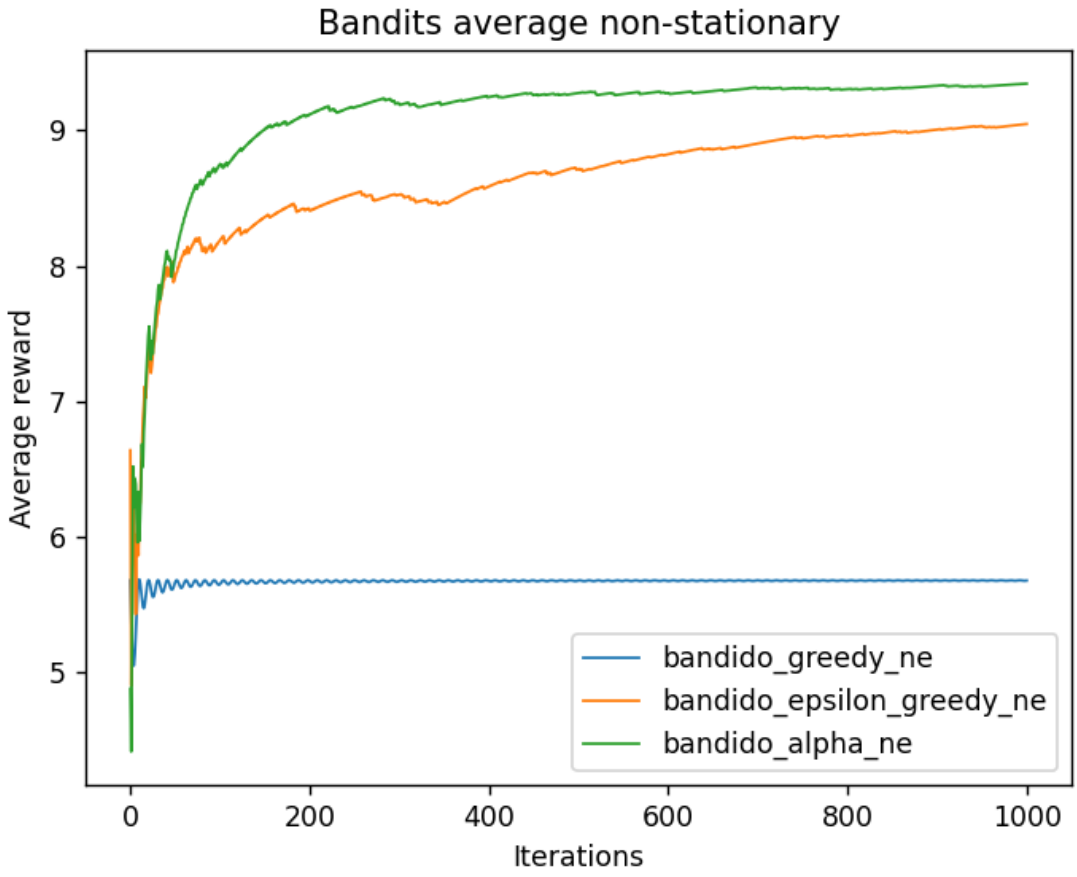



Como muestra la gráfica el bandido Alpha quien da más importancia a las recompensas más actuales comienza con un promedio superior a los otros, sin embargo, al aumentar el número de iteraciones el promedio converge a cero como los otros bandidos, recordemos que todas las máquinas tiene una media de cero pues entonces es de esperar que el promedio de las recompensas tienda a cero.

**Escenario No Estacionario**

Para el caso no estacionario se construyen 10 máquinas cada máquina con una recompensa definida $r = valorinicial + \sin(step * valorinicial)$, donde $valorinicial = uniform(1, numbrazos)$ .

Los 3 jugadores son bandido_greedy, bandido_epsilon_greedy, y bandido_alpha, el primero codicioso que siempre juega la máquina con mejor Q, el segundo que explora el 0.1 de las veces y el último que explora el 0.1 de las veces con un Alpha de 0.1 .




Como se muestra en la gráfica bandido epsilon greedy con exploración del 10% parecen converger al mismo rendimiento, mientras el bandido greedy que no explora se mantiene en un nivel de rendimiento, tal vez, el

comportamiento sinoidal de la recompensa sostiene en “equilibrio” la recompensa de la primera máquina que toma, es más, observe en la gráfica que la curva bandido_greedy_ne tiene comportamiento senoidal.