IIC3675 - RL: Tarea 1

Bruno Cerda Mardini

August 16, 2025

Pregunta a

Los resultados que obtuve si fueron los esperados. Se puede observar que el orden de los resultados es el mismo que en el gráfico del libro, de manera más especifica: El agente con $\epsilon = 0.1$ tiene la mayor recompensa promedio y el mejor rendimiento en elegir la acción óptima, luego en segundo lugar está el agente con $\epsilon = 0.01$ y finalmente el agente con $\epsilon = 0.0$ en último lugar.

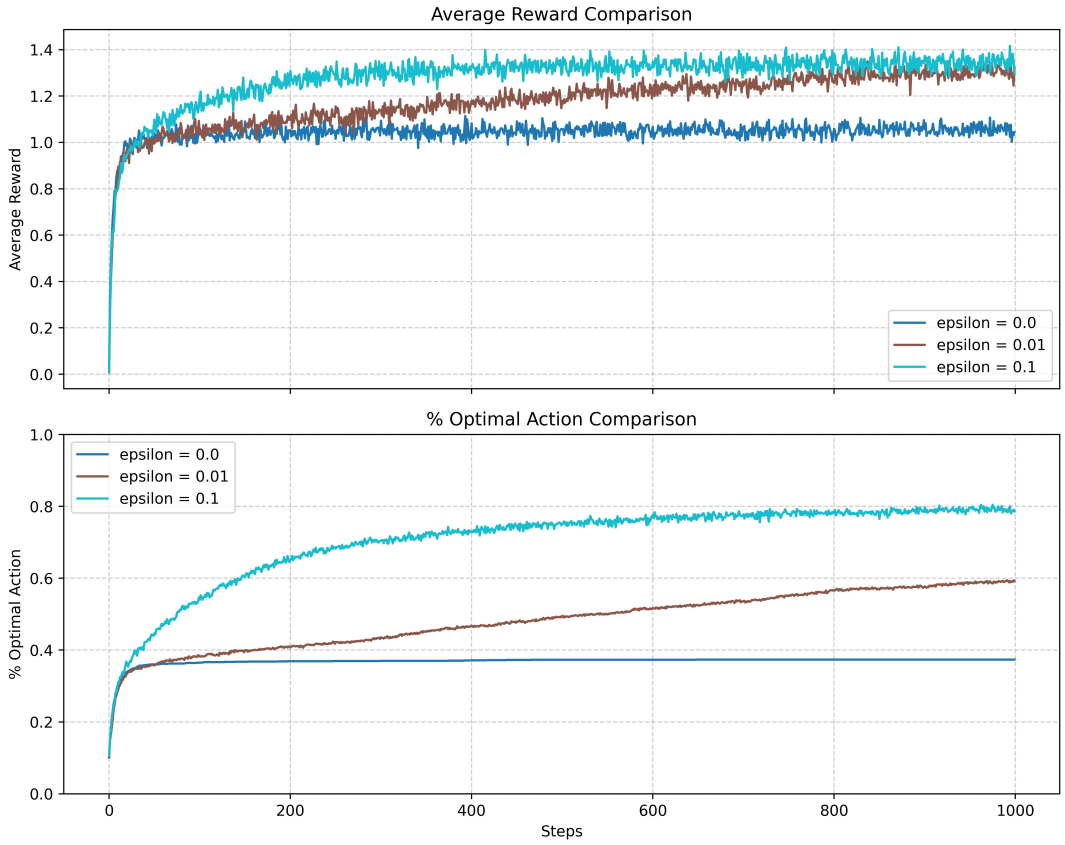

Figure 1: Gráficos para la comparación de recompensa promedio y el porcentaje de acción óptima, ambos con $\epsilon \in [0.0, 0.01, 0.1]$

.

Pregunta b

Que el rendimiento del agente no llegue a 90% con $\epsilon = 0.1$ tiene que ver con el conocimiento que ha aprendido y el número de pasos que ha efectuado.

Si suponemos que el agente ha aprendido totalmente como resolver el problema, y que tiene que elegir un 10% de las veces una acción random, entonces su rendimiento debería llegar al 91%, esto debido a que en ese 10% va a tender a elegir 1 de las acciones correctas (ya que son 10 arms en total).

Por lo tanto, el hecho de que llegue alrededor de 80% nos indica que el agente no ha aprendido con totalidad a resolver el problema. Esto se podría intentar solucionar aumentando el número de pasos posibles, pero de todas formas debido a la naturaleza estocástica de este problema, es probable que no llegue a ese 91%.

Pregunta c

Los resultados obtenidos si son los esperados, ya que se puede observar el mismo comportamiento representado en el gráfico del libro: Se demuestra un mejor rendimiento al utilizar los parámetros optimistas ($Q_1 = 5, \epsilon = 0$), llegando cerca del 80%. Mientras que con los parámetros realistas ($Q_1 = 0, \epsilon = 0.1$) se llega aproximadamente al 75% de rendimiento. Cabe destacar que durante los primeros pasos, principalmente debido a la exploración inicial por consecuencia de la recompensa inicial, la estrategia con parámetros realistas tiene un mejor desempeño que la estrategia con parámetros optimistas, pero al largo plazo esto se invierte.

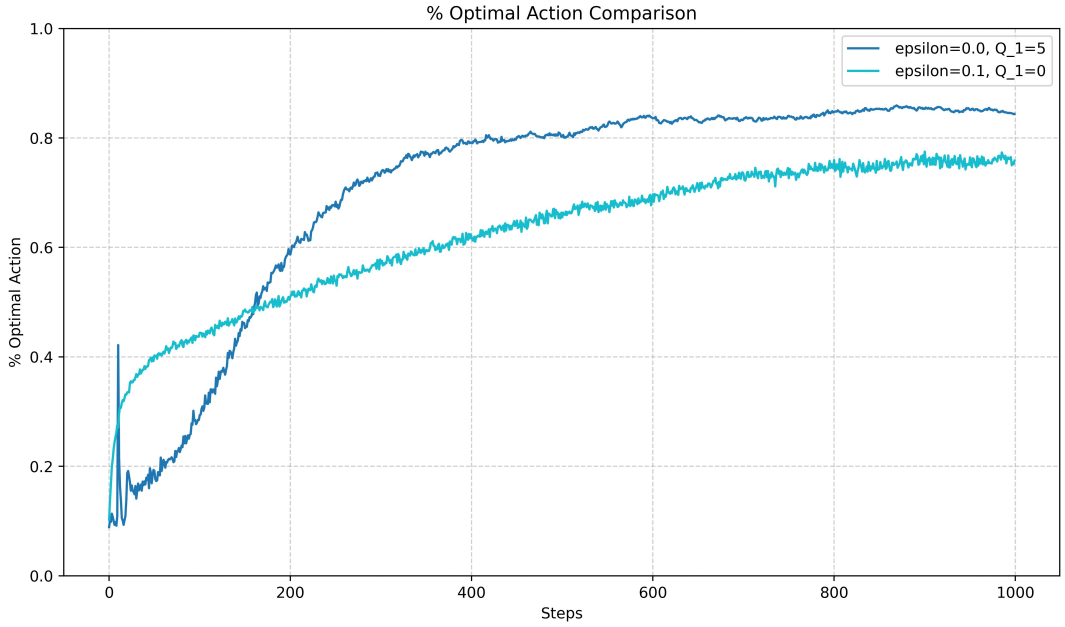

Figure 2: Grafico de porcentaje de accion optima, se muestran dos sets de parametros distintos, uno optimista ($\epsilon = 0.0$, $Q_1 = 5$), y otro realista ($\epsilon = 0.1$, $Q_1 = 0$).

.

Pregunta d

La subida y bajada abrupta que ocurre es consecuencia de la combinación de un agente greedy ($\epsilon = 0$) y valores optimistas iniciales para las acciones ($Q_1 = 5$).

Al inicio el agente va a elegir de manera random uniforme cualquiera de las acciones, ya que todas van a tener el mismo valor. Luego, las acciones no óptimas van a ser reducidas de mayor manera, debido a la fórmula de la actualizacion de la creencia de los valores de las acciones por parte del agente. Después de esto en un paso determinado el agente va a elegir siempre la acción con el mayor valor, las cuales van a ser las acciones óptimas verdaderas, y luego estas van a ser degradadas, teniendo como resultados nuevamente un descenso en el rendimiento del agente.

Pregunta e

Hay múltiples razones que aportan para explicar por qué el agente selecciona la acción óptima él 85% de las veces y no él 100%.

Dado que el agente tiene $\epsilon = 0$, este siempre va a elegir la opción que él cree que es óptima. El rendimiento de 85% nos indica que el agente no ha aprendido bien como resolver el problema en su totalidad, ya que él 15% restante elige acciones que él cree que son las más óptimas, pero en realidad no lo son.

La naturaleza estocástica de este problema definitivamente tiene relación con lo anterior, ya que las recompensas provienen de una distribución, por lo que el agente en ocasiones es "engañado" por recompensas de acciones no óptimas.

Algo más que hay que considerar en esta pregunta es el parámetro α. En este caso, el parámetro tiene un valor de 0.1, lo cual no es un valor bajo. Este valor alto permite que se le dé un peso más grande a los valores de las recompensas más recientes, pero en específico esto va a causar que sea afectado de peor manera por la varianza presente, lo cual va a hacer más difícil en general que converja en un rendimiento más óptimo. Un valor de α más chico podría ayudar a llegar a un rendimiento más alto, sin embargo, se necesitarían más pasos para poder lograr esto.

Pregunta f

Los resultados observados en el gráfico nos indica que si son los esperados, ya que son los mismos que en el gráfico del libro. En específico, los resultados que tienen Baseline obtienen un mejor rendimiento con respecto a los que no tienen Baseline. Cabe destacar también que un valor de α más chico es relevante, principalmente porque resulta en un mejor rendimiento (Con Baseline True/False respectivamente).

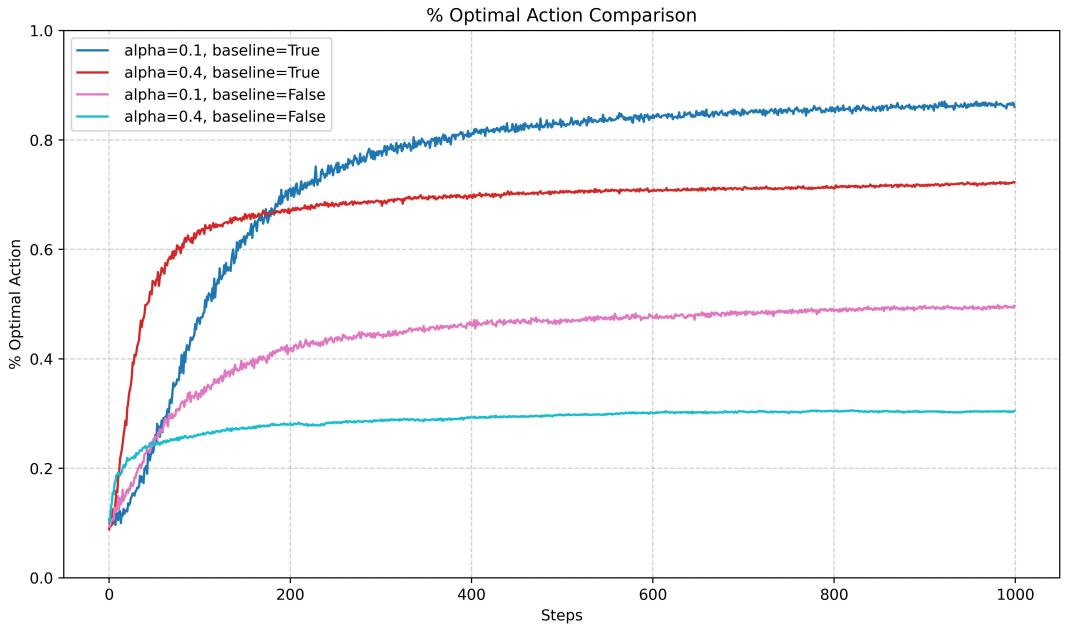

Figure 3: Gráfico de porcentaje de acción óptima, se muestran cuatro sets de parámetros distintos, en donde se varían los α y si es que hay o no Baseline.

Pregunta g

Si utilizáramos $\mu = 0$ eventualmente el promedio de las recompensas utilizadas como baseline en la fórmula de ascenso de gradiente van a tender a 0.

Esto implica que la decisión de tener o no una baseline es irrelevante, ya que en la fórmula de actualización se utilizaría 0 como valor de recompensa promedio, que tendría como consecuencia un igual rendimiento en ambos casos. La figura 4 muestra este efecto.

También se puede observar en la figura 4 que el rendimiento se centra aproximadamente en él 10% para todos los parámetros, esto quiere decir que el agente está eligiendo de manera random la acción que él cree que es óptima, y esto ocurre debido a que en la fórmula del ascenso del gradiente, al actualizar los valores de las acciones, el agente no va a tener información relevante sobre cuál acción es mejor, ya que el promedio pasado de recompensas va a tender a 0, por lo que a lo largo del tiempo el agente va a creer que todas las acciones son igual de buenas.

Figure 4: Gráfico de porcentaje de acción óptima, se muestran cuatro sets de parámetros distintos, en donde se varían los α y si es que hay o no Baseline. Se puede observar que a diferencia del resultado anterior, aquí los rendimientos son iguales debido a $\mu = 0$

,