Avance Tarea 2: Q-Learning

Código: EL7021-1

Nombre: José Luis Cádiz Sejas

Pregunta 1: Descripción del MDP.

- **Espacio de estados:** Sea $s \in S \mid s = (x, v)$, donde $x \in [-1.2, 0.6]$ representa la posición horizontal del vehículo y $v \in [-0.07, 0.07]$ representa el módulo de su velocidad.

- **Espacio de acciones**: El espacio de acciones A es tal que:

$$A = \{"0": "Acelerar\ hacia\ la\ izquierda", "1": "No\ acelerar", "2": "Acelerar\ hacia\ la\ derecha"\}$$

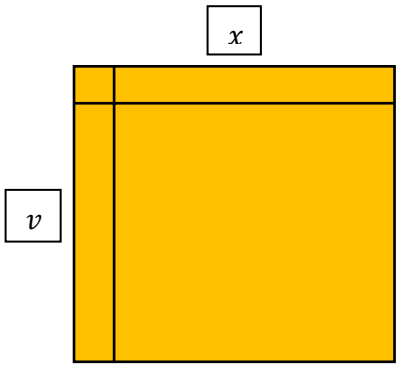
- **Función de recompensa**: La función de recompensa es tal que en cada paso de tiempo se otorga al agente una recompensa -1, si el agente logra llegar la cima de la montaña derecha, se otorga una recompensa de 0. Matemáticamente:

$$R(s) = \begin{cases} -1\ if\ x < 0.5 \\ 0\ if\ x \geq 0.5 \end{cases}$$

- **Función de transición de estados**: Sea el estado siguiente $s' = (x', v')$ dado el estado y la acción actual s, a respectivamente, la función de transición de estados estará definida por:

$$s' = \left(x', v'\right) = \mathrm{T}(s, a) \mid x' \in [-1.2, 0.6]\ , v' \in [-0.07, 0.07]$$

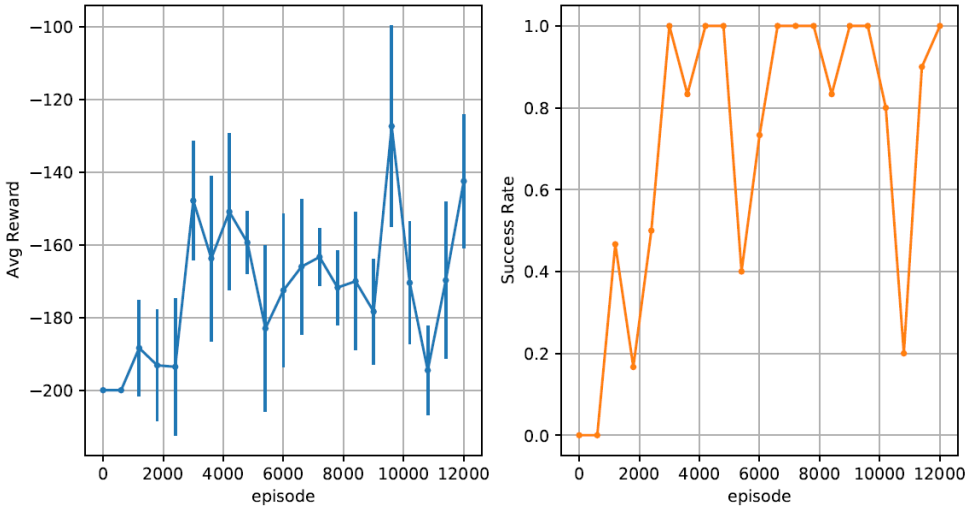
Pregunta 2: Discretización de los estados del ambiente. Dado el espacio 2D, se dividirá el espacio en 625 subespacios definidos por rangos. Esto de modo que cada variable será dividida en 25 subgrupos. Código adjunto.



Pregunta 3: Código adjunto.

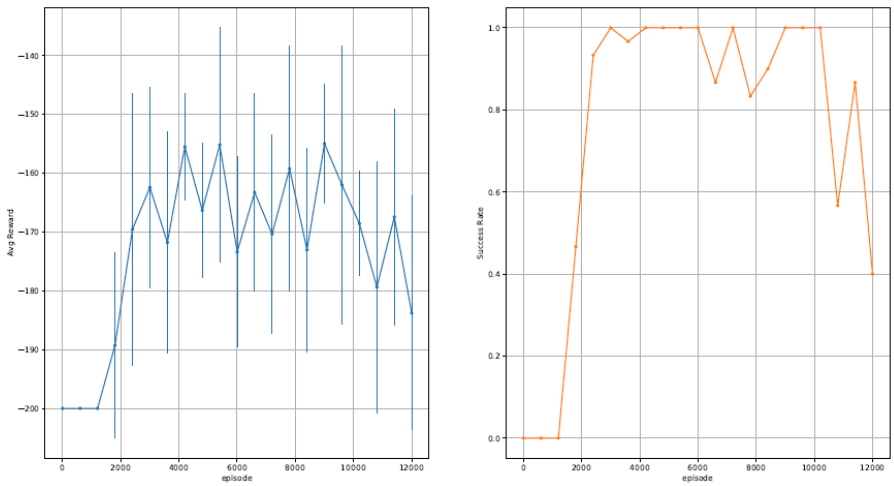
Pregunta 4: Resultados obtenidos:

Average Reward: -175.27, Reward Deviation: 21.68 | Average Steps: 175.27, Success Rate: 0.67



Pregunta 5: Implementación de decaimiento lineal de épsilon. Código adjunto. Resultados y comentarios:

Average Reward: -183.80, Reward Deviation: 19.87 | Average Steps: 183.80, Success Rate: 0.40



A partir de los resultados obtenidos, se observa que un épsilon mayor permite alcanzar buenos resultados en un número menor de episodios respecto de la parte anterior, esto se asocia a que se le permite al agente explorar nuevas posibilidades con mayor recurrencia. Sin embargo, casi al termino de los periodos, épsilon es casi el doble (0.2) respecto del épsilon igual a 0.1 de la parte anterior, lo que al termino del entrenamiento se observan resultados con menor desempeño producto de seguir explorando de manera más exhaustiva acciones no necesariamente optimas.