Analisis Q-Learning - DoubleQ-Learning

Taxi-v3

Para este entorno a fin de poder medir la eficiencia de los distintos algoritmos, se han evaluado los siguientes parámetros en cada algoritmo: Medir la eficacia del aprendizaje, esto lo medimos con el número de episodios que son necesarios para que el valor se estabilice. Recompensa obtenida cuando se ha acabado el aprendizaje (la cual está relacionada con el número de pasos)

Para poder calcular el primer parámetro, se ha entrenado a dos (2) agentes distintos (uno por cada algoritmo) durante diez mil episodios. Una vez entrenado nuestro agente, su Tabla donde almacena los valores para cada par estado-acción ha tomado los valores óptimos para tomar las decisiones correctas. Tras esto se hace una media de los pasos necesarios y la puntuación obtenida.

Antes de pasar a comparar los algoritmos entre sí, es necesario encontrar los parámetros alfa y épsilon que mejor se ajustan a cada algoritmo. Para ello, se han llevado a cabo diferentes pruebas variando dichos parámetros, y he aquí los resultados para los 2 algoritmos.

Q-Learning.

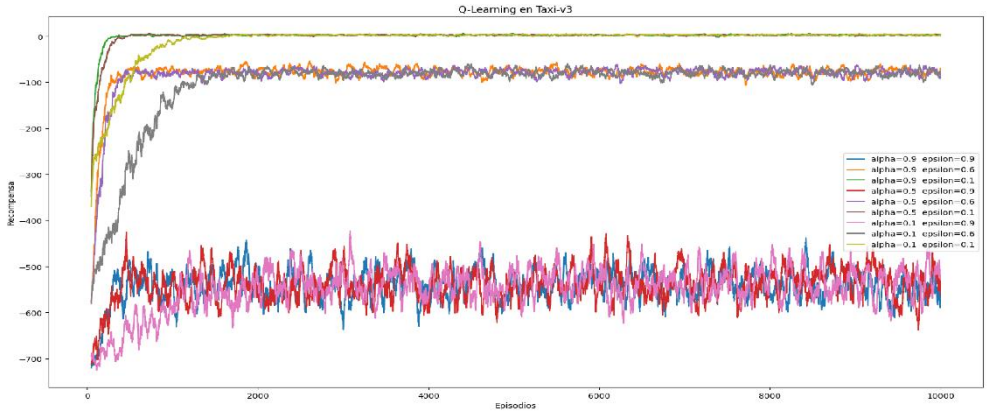

Figura 1. Alfa-épsilon. Q-Learning en el entorno Taxi-v3.

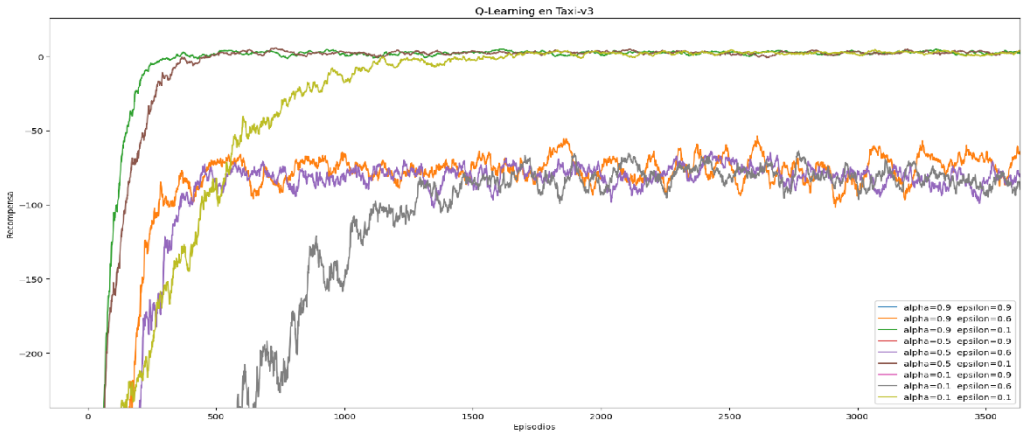

Figura 2. Alfa-épsilon. Q-Learning en el entorno Taxi-v3 Ampliada.

Para este algoritmo se puede observar en la figura 1 y en la figura 2 (que es la misma, pero ampliada) que la gráfica en color verde es la que nos da mejores resultados, esto teniendo un valor de **alfa = 0.9**, y **épsilon = 0.1**. Tras diversas pruebas realizadas con distintos valores de alfa y épsilon, es lo que mejor resultados ha dado para este entorno.

Cabe destacar que, aunque en las gráficas anteriores no están todos los pares posibles de alfa-épsilon, ya que son infinitos, y el incluir demasiados pares dificultaría la visualización y comprensión de dichas gráficas, se ha optado por diferentes combinaciones manteniendo bien alfa = 0.9 o alfa = 0.1 ya que en pruebas previas se ha observado como para valores intermedios de alfa, se obtienen peores resultados.

A continuación, la figuras 3 nos muestran la evolución de aprendizaje del agente con los valores de alfa = 0.9, y épsilon = 0.1, se puede observar como parece converger a 0, pero realmente converge a un valor cercano a 2, que es la recompensa obtenida, tal como se puede observar en la figura 4, que es la misma gráfica, pero ampliada.

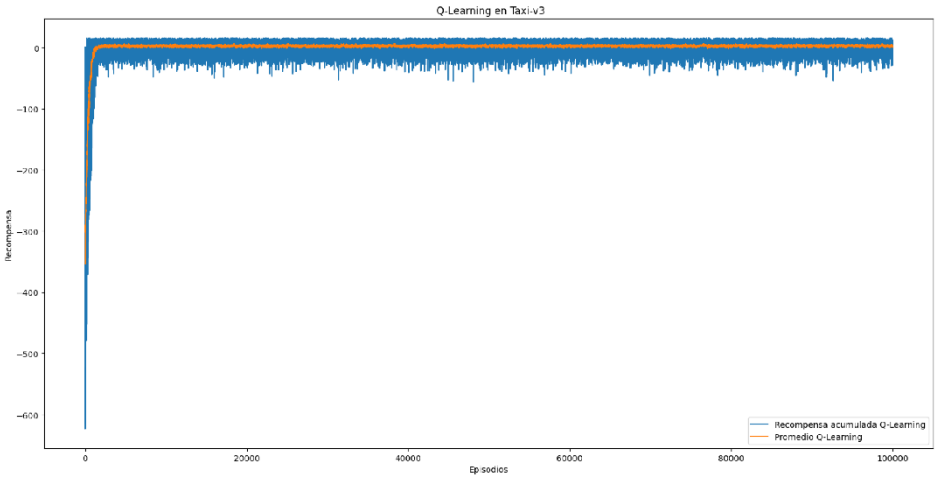

Figura 3. Qlearning en Taxi-v3, Evolución del aprendizaje

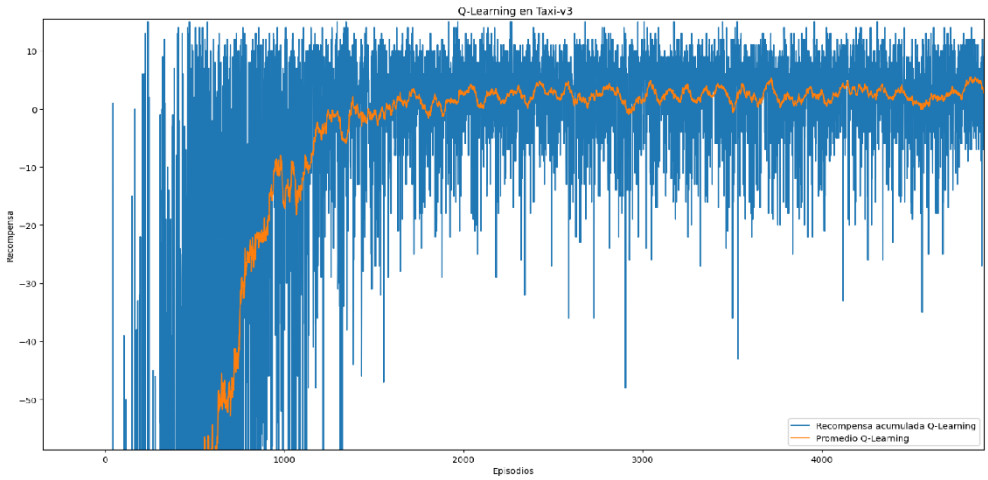

Figura 4. Qlearning en Taxi-v3 Evolución del aprendizaje Ampliada.

Double Q-Learning.

Para este algoritmo hay que tener en cuenta dos cosas. Tal como se observa en la figura 5, la primera la apreciamos en la gráfica con un color marrón, que es la correspondiente a los valores de alfa = 0.5 y épsilon = 0.1. La segunda es la coloreada de verde claro, correspondiente a los valores de **alfa = 0.1** y **épsilon = 0.1**.

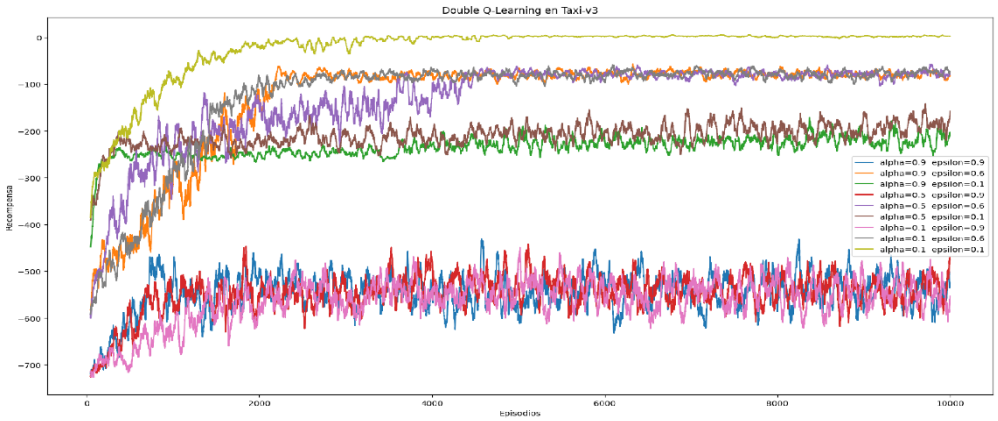

Figura 5. Alfa-épsilon. Doble Q-Learning en el entorno Taxi-v3.

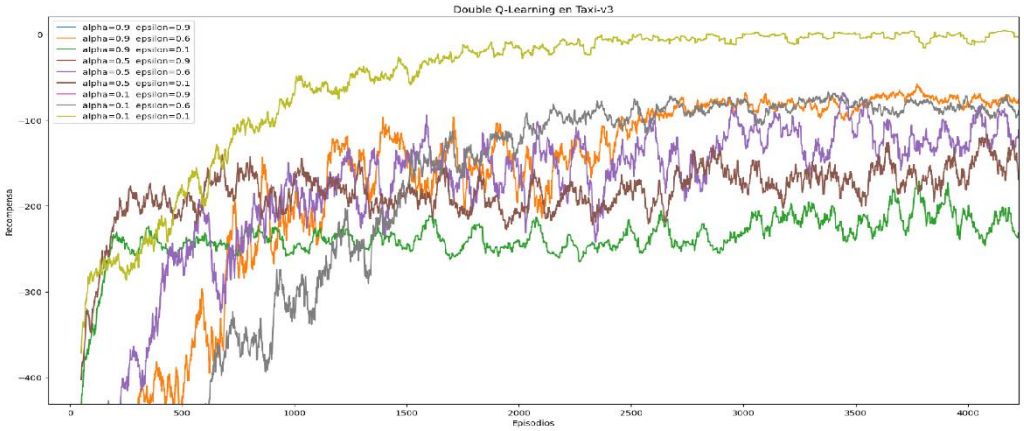

Figura 6. Alfa-épsilon. Doble Q-Learning en el entorno Taxi-v3 Ampliada.

Como se observa en la figura 6, la primera línea descrita anteriormente (color marrón) empieza a obtener mejor recompensa de forma más temprana, sin embargo, llega un momento en que apenas mejora, es por ello se ha optado por tomar los valores de la segunda línea (color verde claro) ya que es la que maximiza la recompensa y la que no ofrece los mejor resultados.

A continuación, la figuras 7 nos muestran la evolución de aprendizaje del agente con los valores de alfa = 0.1, y épsilon = 0.1, se puede observar como parece converger a 0, pero realmente converge a un valor cercano a 2, que es la recompensa obtenida, tal como se puede observar en la figura 8.

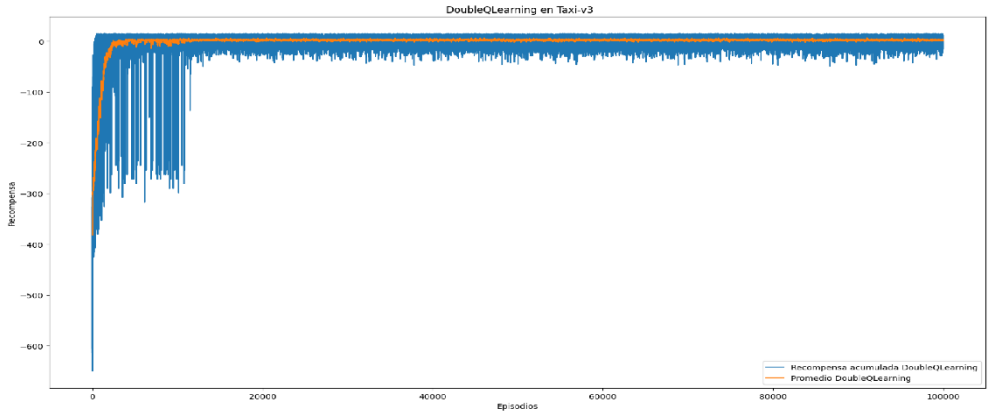

Figura 7. Alfa-épsilon. Doble Q-Learning en el entorno Taxi-v3.

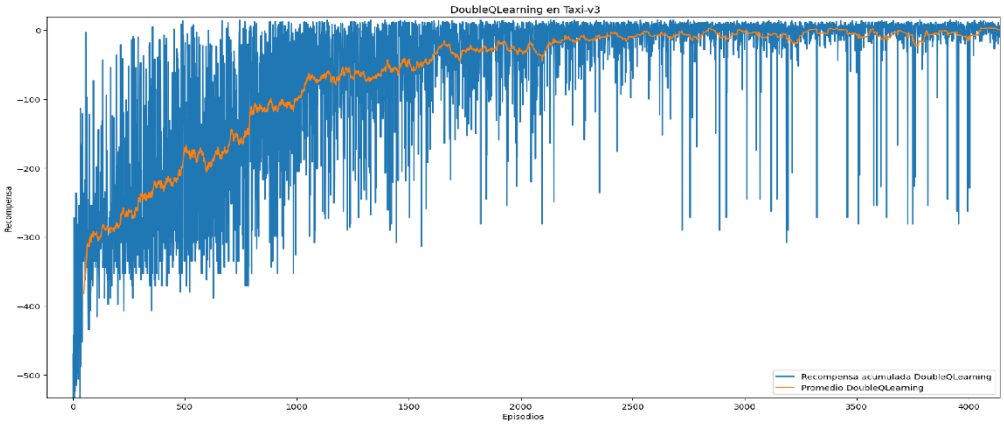

Figura 8. Alfa-épsilon. Doble Q-Learning en el entorno Taxi-v3 Ampliada.

Por ultimo tenemos la comparación de aprendizaje de ambos algoritmos, tanto en el caso de Q-Learning como en el de Doble Q-Learning, ambos consiguen resolver el problema en muy pocos episodios. Siendo Q-Learning el que obtiene mejores resultados, ya que consigue resolver el problema en menos episodios y a pesar que ambos convergen en torno a un valor similar Q-Learning se observa más estable para este entorno. Esto lo podemos apreciaren las Figuras 9 y 10 mostradas a continuación:

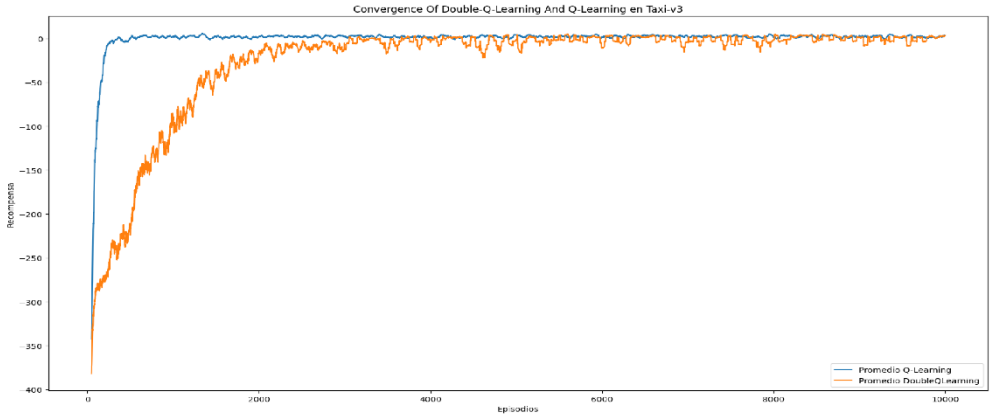

Figura 9. Q-Learning and Doble Q-Learning en el entorno Taxi-v3 Comparativa.

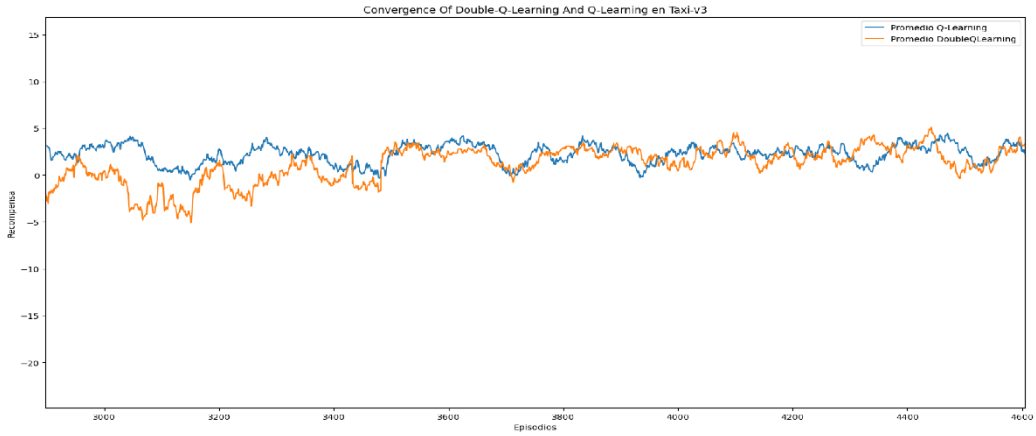

Figura 10. Q-Learning and Doble Q-Learning en el entorno Taxi-v3 Comparativa Ampliada.

CliffWalking-v0

Para este entorno, igual que para el anterior, se han evaluado los mismos parámetros en cada algoritmo a fin de poder medir la eficiencia de los distintos algoritmos. De igual manera que en el entorno anterior, antes de comparar los diferentes algoritmos entre sí, es necesario ver que parámetros alfa y épsilon se ajustan mejor a cada algoritmo.

Q-Learning. Tal y como se muestra en la figura 11, el agente obtiene mejores resultados durante el entrenamiento para los valores de alpha = 0.9 y épsilon = 0.1, además en la figura 12 observamos cómo son las gráficas con épsilon 0.1 son las que de mejor rendimiento.

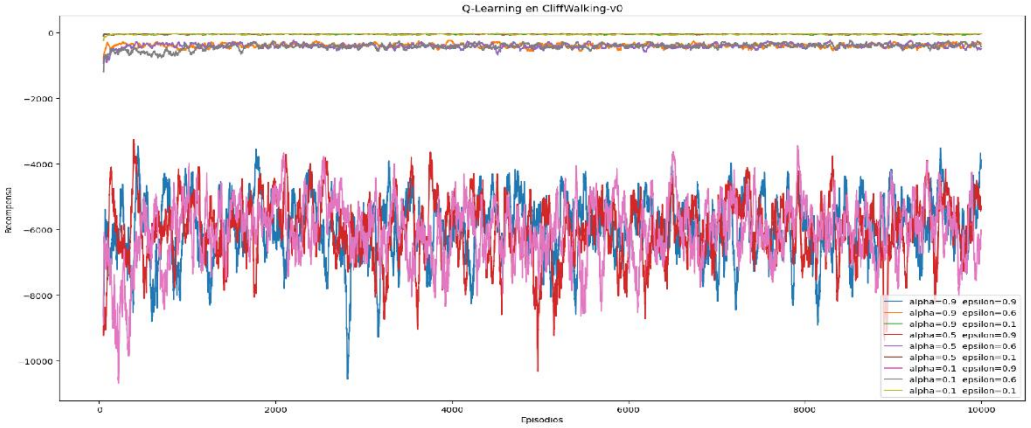

Figura 11. Alfa-épsilon. Q-Learning en el entorno CliffWalking-v0.

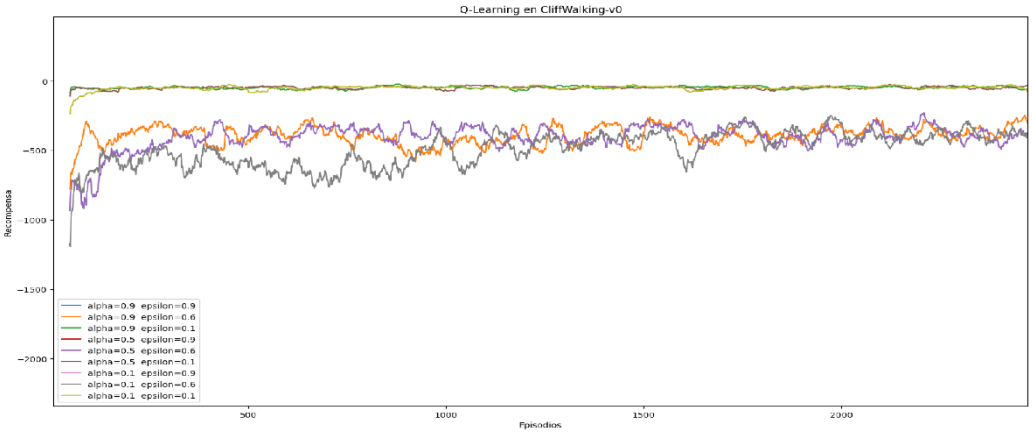

Figura 12. Alfa-épsilon. Q-Learning en el entorno CliffWalking-v0 Ampliada.

A continuación, la figuras 13 nos muestran la evolución de aprendizaje del agente con los valores de alfa = 0.1, y épsilon = 0.1, se puede observar como parece converger a 0, pero realmente converge a un valor cercano a -20, que es la recompensa obtenida, tal como se puede observar en la figura 14 (que es la misma que la anterior pero ampliada).

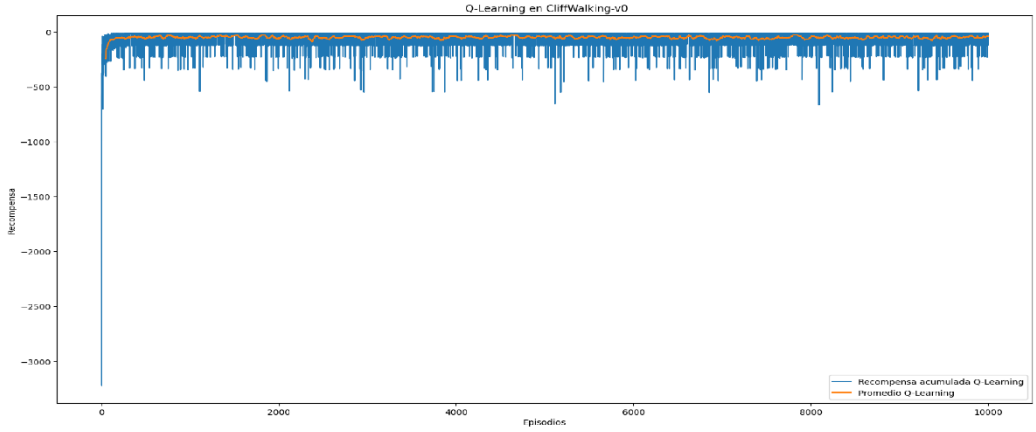

Figura 13. Qlearning en CliffWalking-v0, Evolución del aprendizaje.

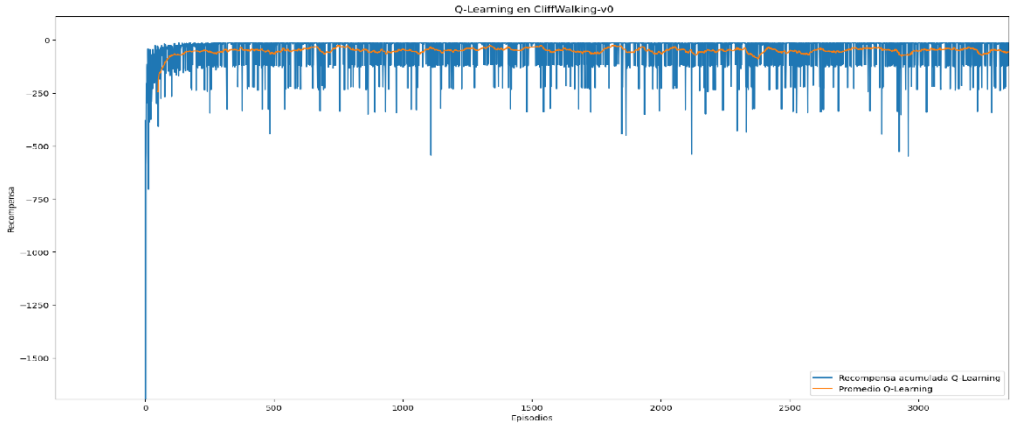

Figura 14. Qlearning en CliffWalking-v0, Evolución del aprendizaje Ampliada.

DoubleQ-Learning.

Como se puede observar en la figura 15, apreciamos que la gráfica de color verde claro, que es la correspondiente a los valores de **alfa = 0.1** y **épsilon = 0.1** es la que llega más temprano a la que maximiza la recompensa y la que no ofrece los mejor resultados. Esto lo podemos apreciar con un poco más de detalle en la Figura 16.

Figura 15. Alfa-épsilon. Doble Q-Learning en el entorno CliffWalking-v0.

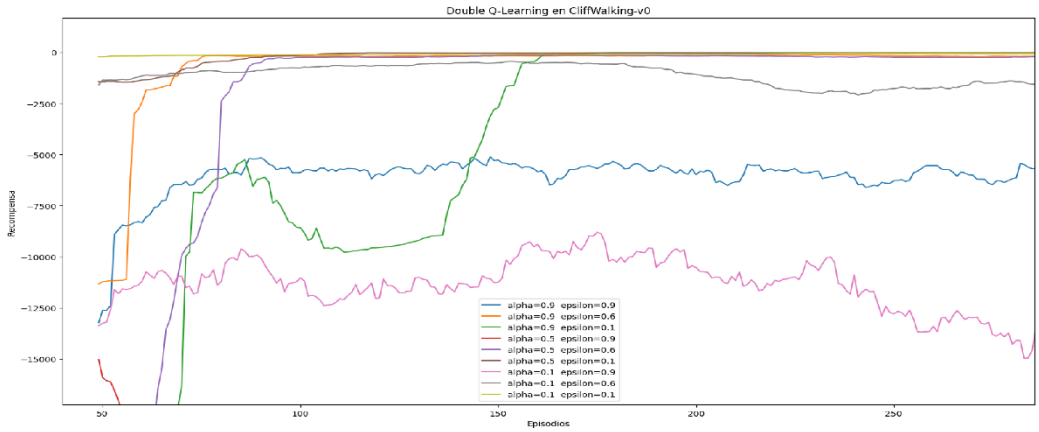

Figura 16. Alfa-épsilon. Doble Q-Learning en el entorno CliffWalking-v0 Ampliada.

Ahora en la figura 17 podemos ver la evolución de aprendizaje del agente con los valores de alfa = 0.1, y épsilon = 0.1 en el entorno CliffWalking-V0, se puede observar como parece converger a 0, pero realmente converge a un valor cercano a 2, que es la recompensa obtenida, tal como se puede observar en la figura 18.

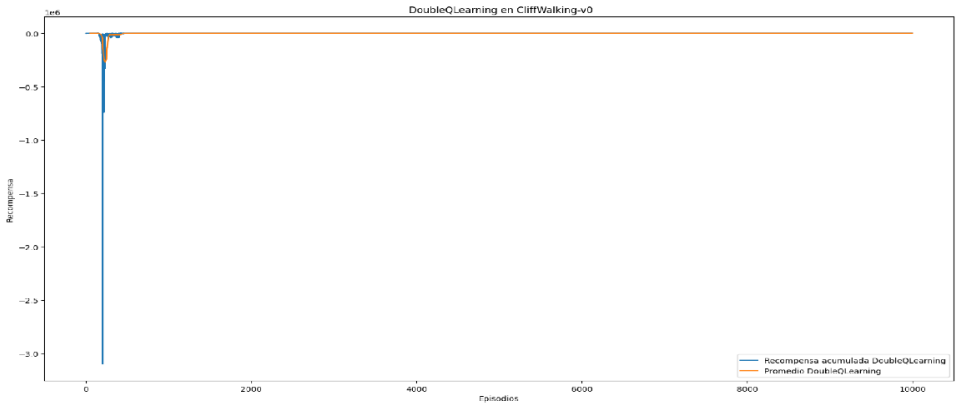

Figura 17.Doble Qlearning en CliffWalking-v0, Evolución del aprendizaje.

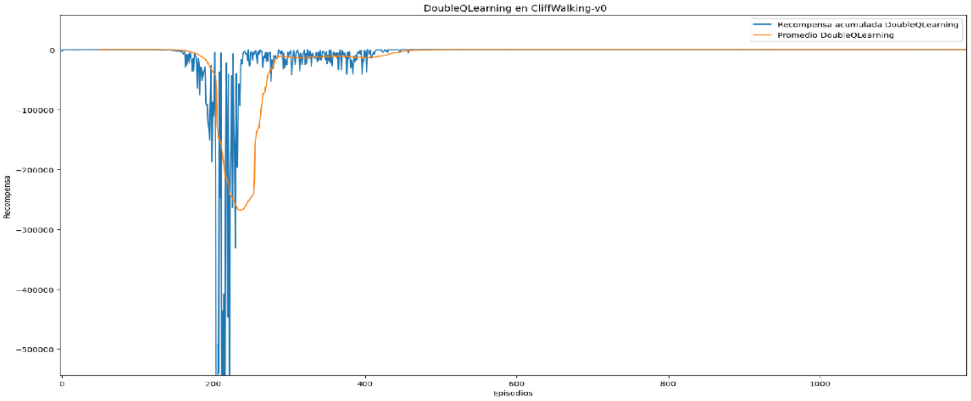

Figura 18.Doble Qlearning en CliffWalking-v0, Evolución del aprendizaje, Ampliada.

Por ultimo tenemos la comparación de ambos algoritmos para este entorno, esto lo tenemos en las figuras 19 y 20, en ellas podemos observar que el agente Qlearning empieza a obtener mejor recompensa de forma más temprana, sin embargo, el doble Qlearning es el que nos ofrece la maximiza la recompensa ya que converge un valor cercano a 2 mientras que QLearning lo hace en un valor cercano a menos 20. Por esto concluimos que para este entorno es mejor el agente DobleQ-Learning.

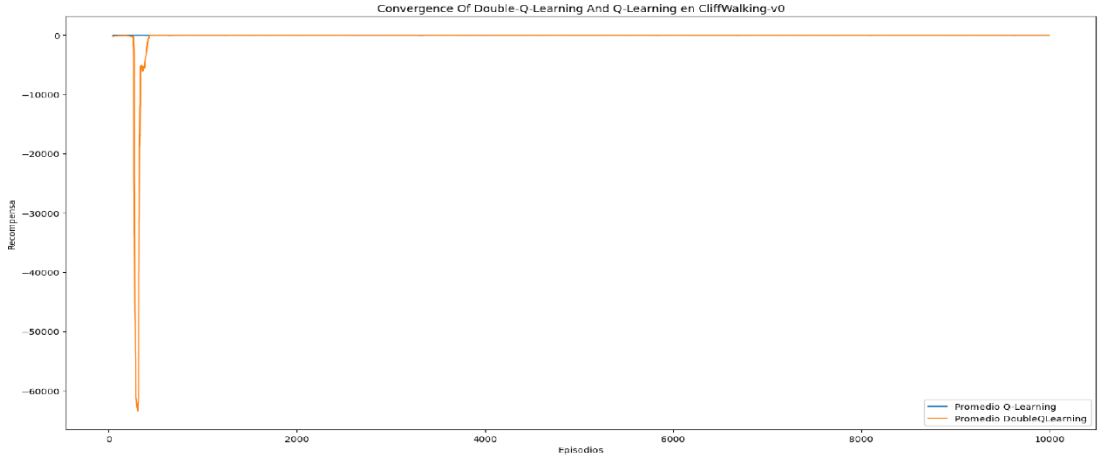

Figura 19. Q-Learning and Doble Q-Learning en el entorno CliffWalking-v0 Comparativa.

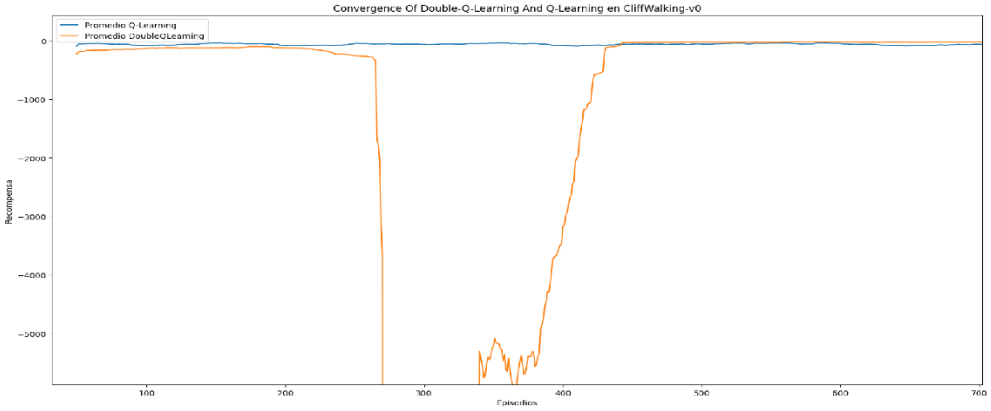

Figura 20. Q-Learning and Doble Q-Learning en el entorno CliffWalking-v0 Comparativa Ampliada.

Autores: Natanael Rojo

C.I: 26.488.388

Heberto Gutiérrez

C.I: 24.752.816

Github: https://github.com/NatanaelRojo/computer-systems-activities/tree/main/homework-9