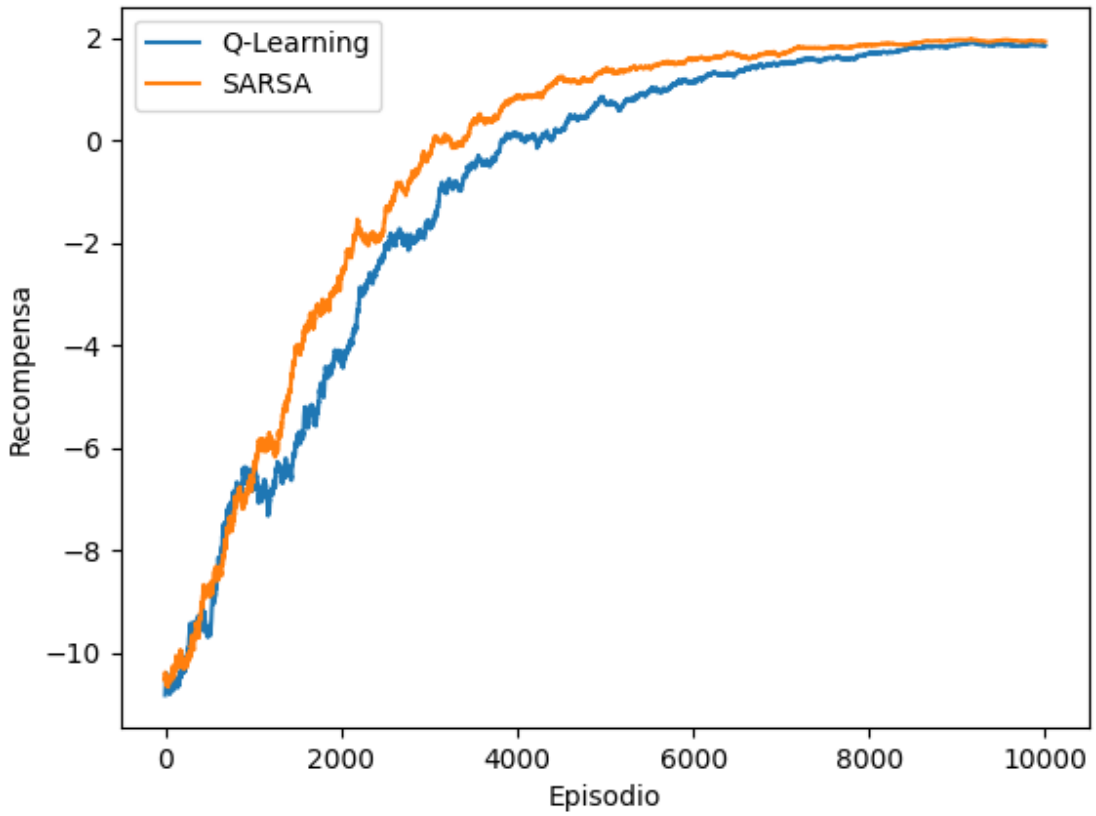
Pablo Zapata - Tarea 1 Aprendizaje por refuerzo

Todos los gráficos están suavizados para que no parezcan dientes de sierra y la recompensa cada episodio es la que obtuvo al final de este, independiente si llego a la meta o no.

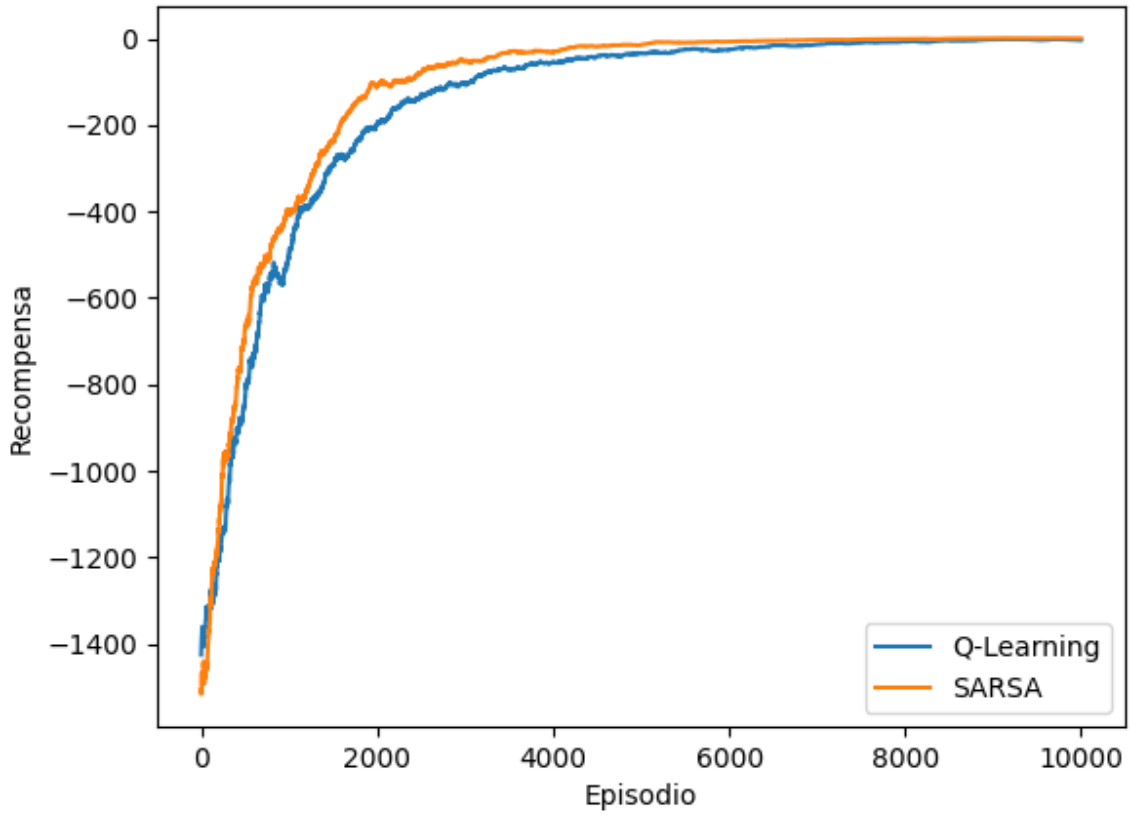
1. Curvas de aprendizaje para el mapa 1. Hiperparametros sin modificar.



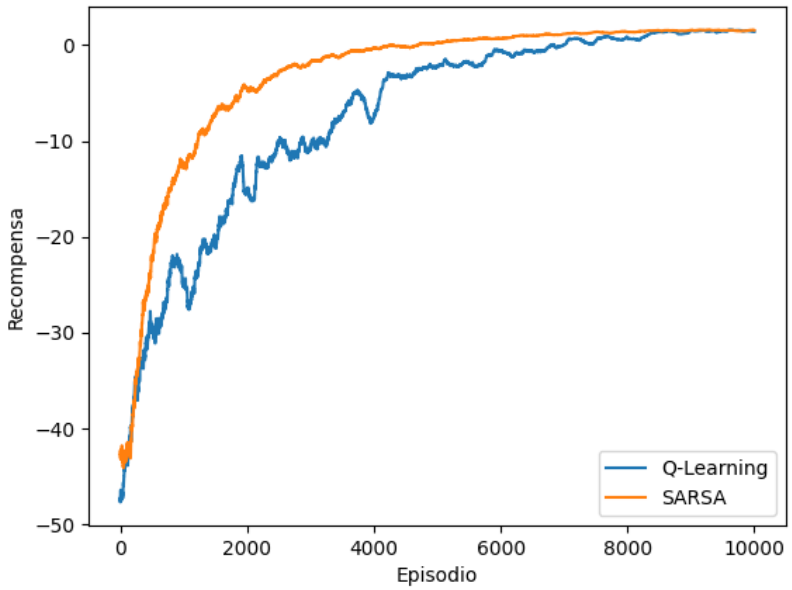
Aumente el gamma a 1, esto debido a que el agente debe recorrer más o menos harto para llegar a la recompensa final, por lo que con un gamma mayor hace que tome con más consideración las recompensas futuras. También reduje a 0.05 el learning rate para que no descarte tanto conocimiento por episodio y aumente los steps máximos a 1000, debido a que 40 son necesarios para el camino más largo y con la gran aleatoriedad inicial, se necesitan de varios para lograr llegar a la meta. Ambos llegan a una política óptima de recompensa final 2.

SARSA encontrará la política óptima cuando su política es greedy y le permite explorar infinitamente cada estado y acción. Q-Learning encuentra la política óptima cuando explora cada estado y acción infinitamente.

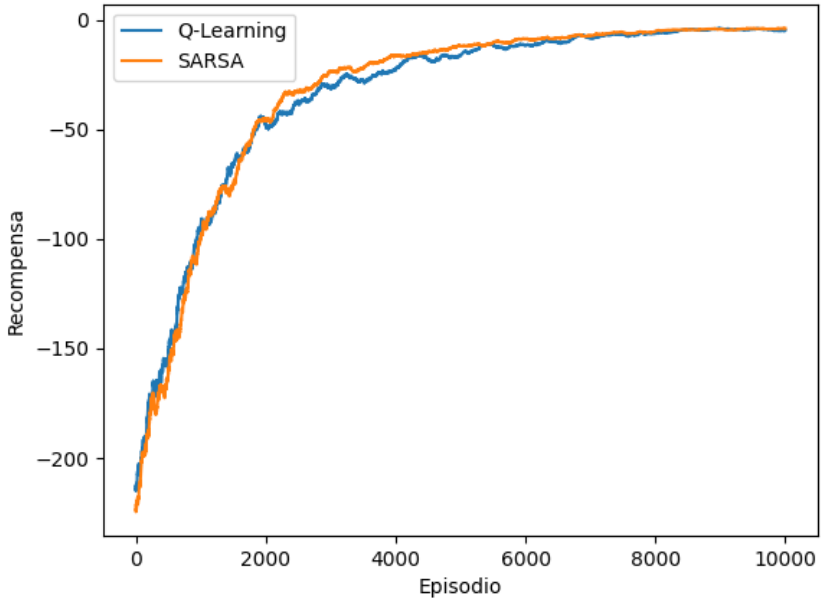
Curvas de aprendizaje para el mapa 2 con gamma = 1, learning_rate = 0.05, epsilon= 1 y max_steps = 1000:



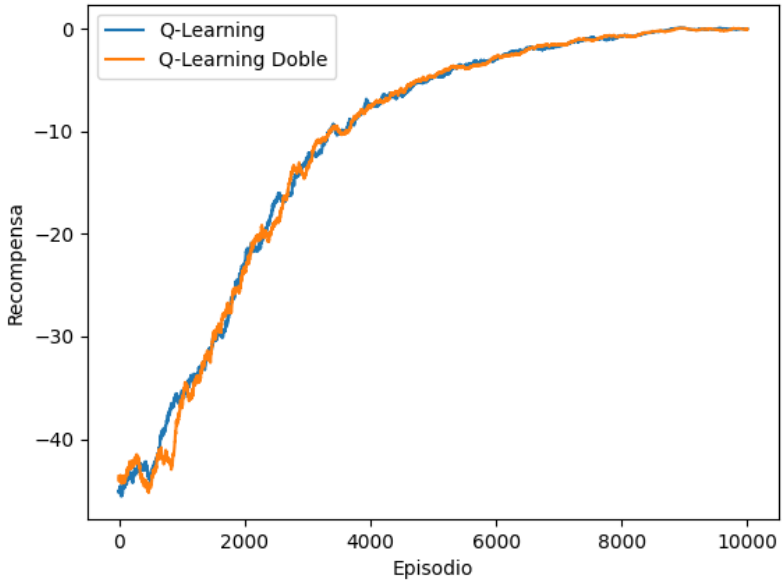
2. Se usaron los hiperparametros modificados. En el mapa 1, SARSA converge de mejor manera que Q-Learning debido a que intentará tomar un camino más seguro. Ambos algoritmos llegan a recompensas finales de 2 la mayoría del tiempo, aunque Q-Learning lo hace más consistentemente.



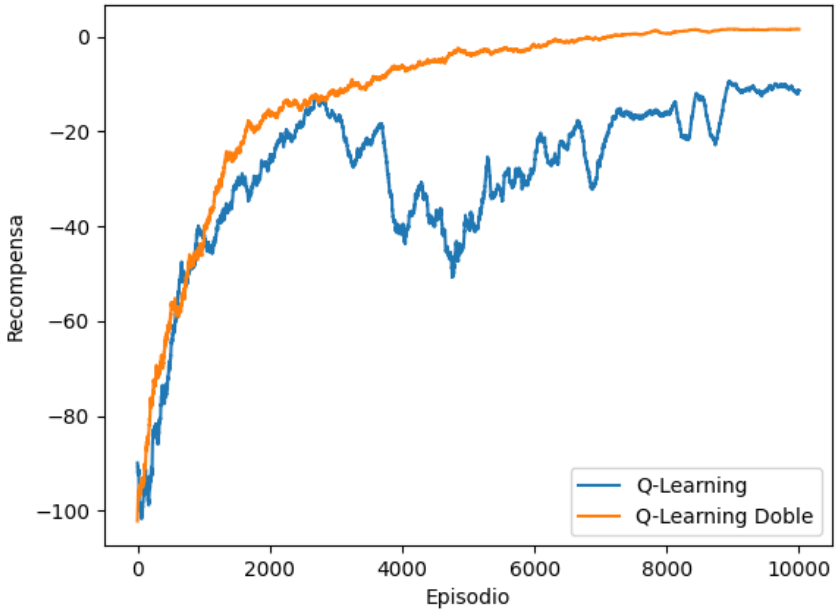
En el mapa 2, SARSA y Q-Learning toman el camino más corto para llegar a la meta, pasando por una bomba y chocando contra paredes ocasionalmente, resultando en recompensas negativas la mayor parte del tiempo. Toman este camino debido a que un corto implicaría chocar con menos paredes.



3. Para el mapa 2, estocástico con hiperparametros modificados, ambos métodos se comportan casi igual, esto es debido a que la mayoría del ambiente es casi lo mismo. La ventaja de doble Q-Learning es que aprende a no elegir un camino donde el promedio de las acciones de un estado da una recompensa negativa, el problema es que la mayoría de los estados son iguales, dos acciones dan 0 de recompensa y las otras dos es chocar.



Sin embargo, si se eliminase la penalización de chocar contra una pared, Q-Learning doble gana bastante ventaja:



4. En el mapa 2, estocástico con hiperparametros modificados, SARSA y Q-Learning tomaban el camino corto con la bomba. SARSA Lambda y Q Lambda también toman este camino. No hay mucha diferencia entre sus curvas de aprendizaje (El azul apenas se ve pero si está).