Q - Learning

Configuración

Se ha implementado el algoritmo de Q -learning según el guión, donde:

- Se han realizado 50 bucles para obtener la media de las utilidades y la moda de las políticas.
- Se realizan 1000 *trials* dentro de cada bucle.
- Dentro de cada trial se obtiene un estado aleatorio no obstáculo y se va avanzando hasta encontrar el estado final.
- El valor epsilon del Q-learning esde 0.92, es decir, en el 92% d las veces realizará una explotación (se avanza hacia donde es más probable encontrar el estado final) y en el 8% una exploración.
- El valor gamma de actualización es de 0.95 (factor de descuento), y el alfa se calcula dependiendo de la iteración en la que se encuentra y del número de veces qe se a explorado el estado en el que se encuentre.

Todos estos parametros son editables y se pueden realizar diferentes pruebas donde, siempre que sean valores lógicos y se converja en cada bucle, funcionará.

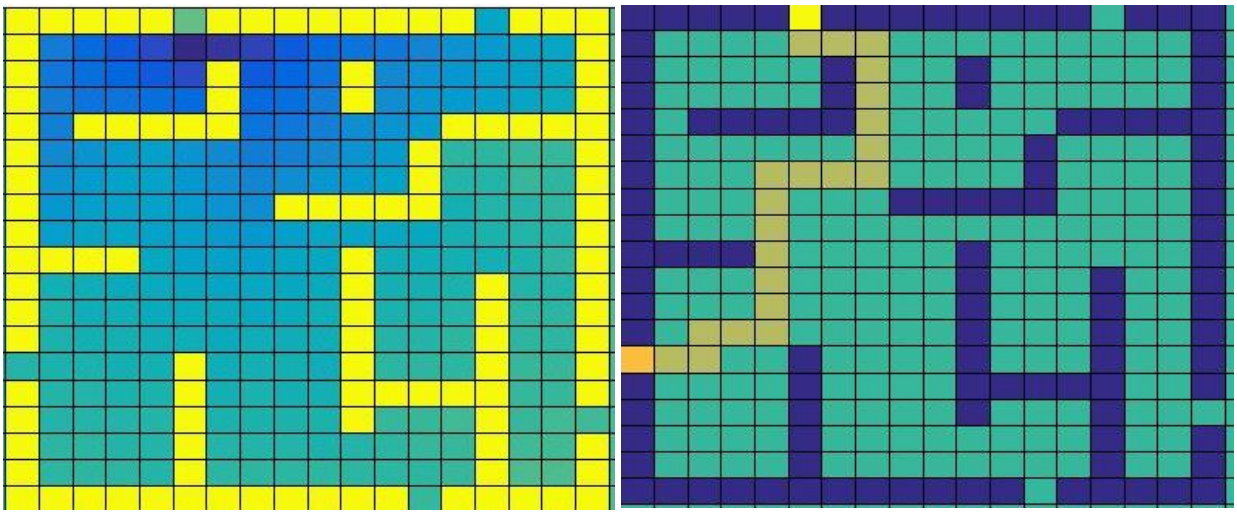
En comparación con MDP, los resultados no son idénticos, pero tan solo en los estados donde las utildiades de los posibles estados siguientes son casi idénticas. Lo que si es idéntico es el coste total.

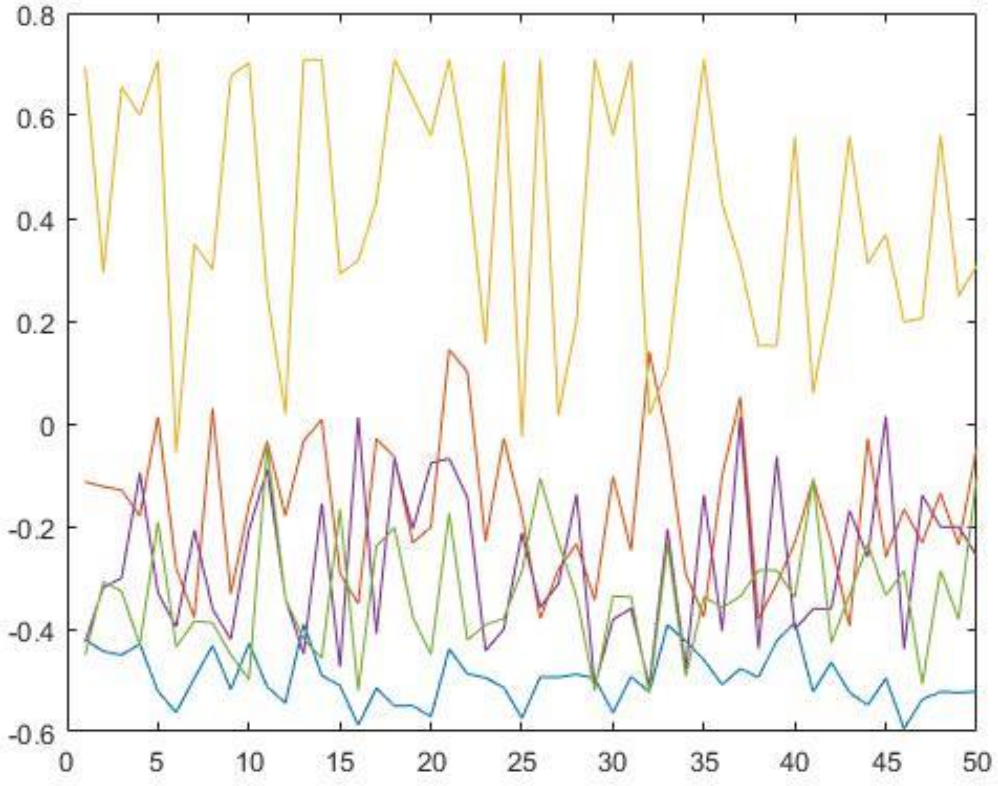
A continuación se muestran 3 ejecuciones:

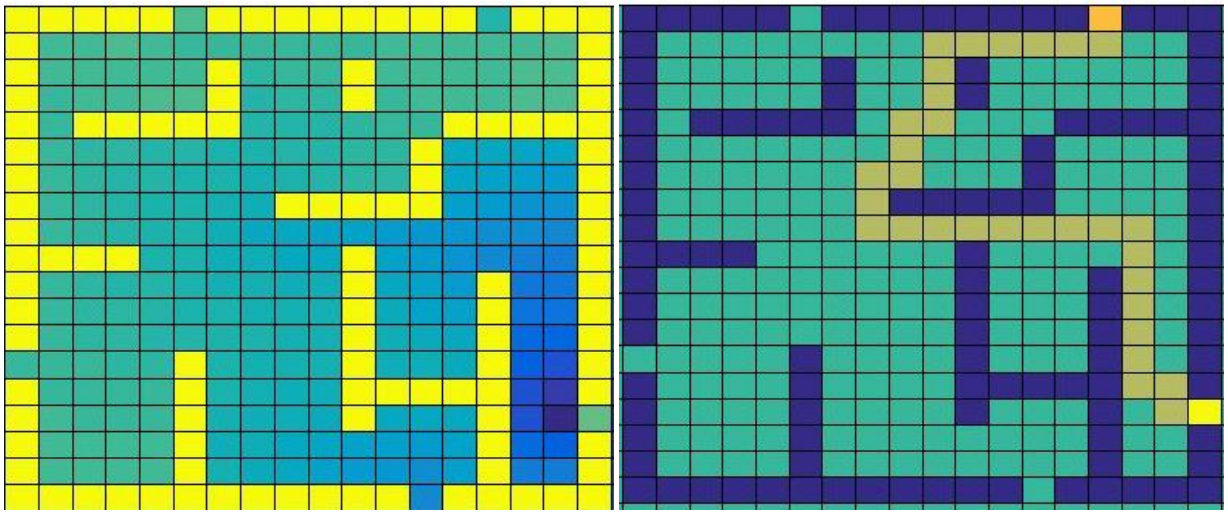
En la primera imagen se muestran las utilidades finales en cada estado del laberinto.

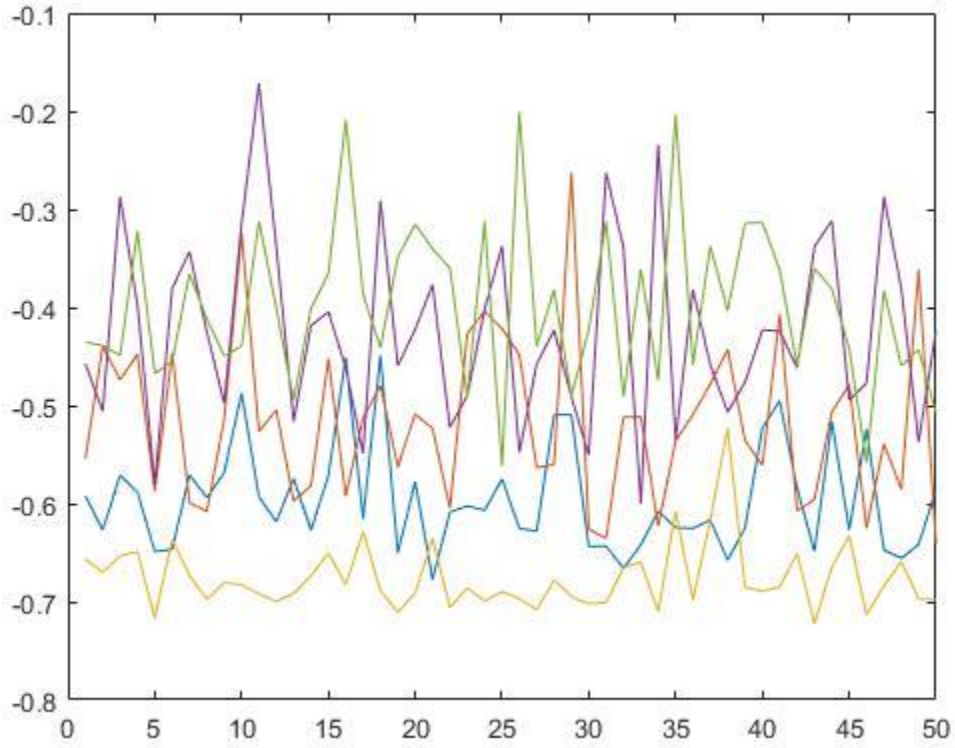
En la segunda, se muestra el camino óptimo

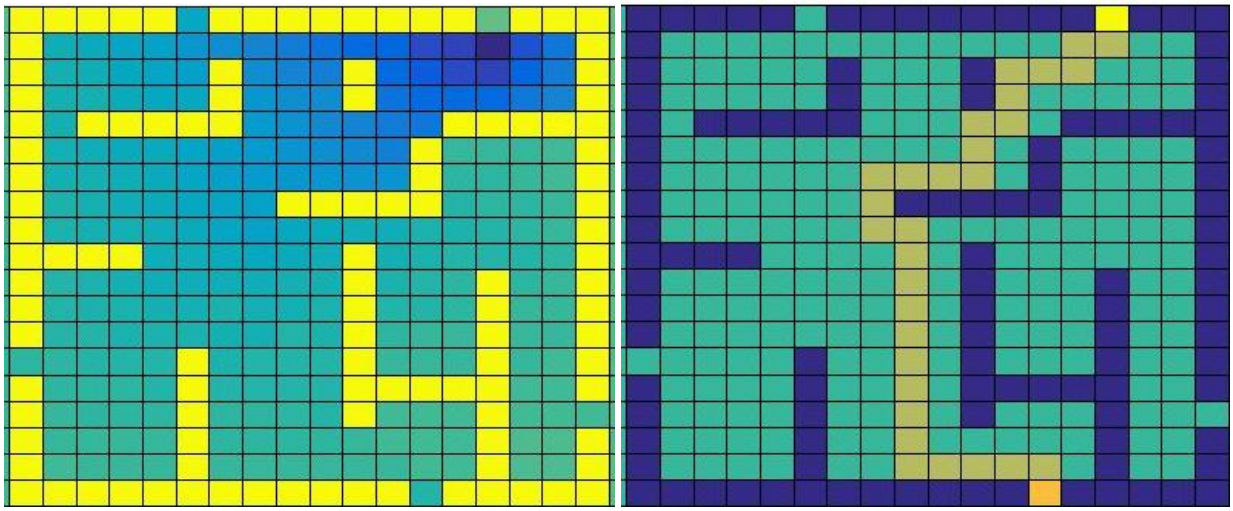
Y en la tercera se muestra la variación en la finalización de cada bucle de los valores de la uilidad de cinco estados al azar.

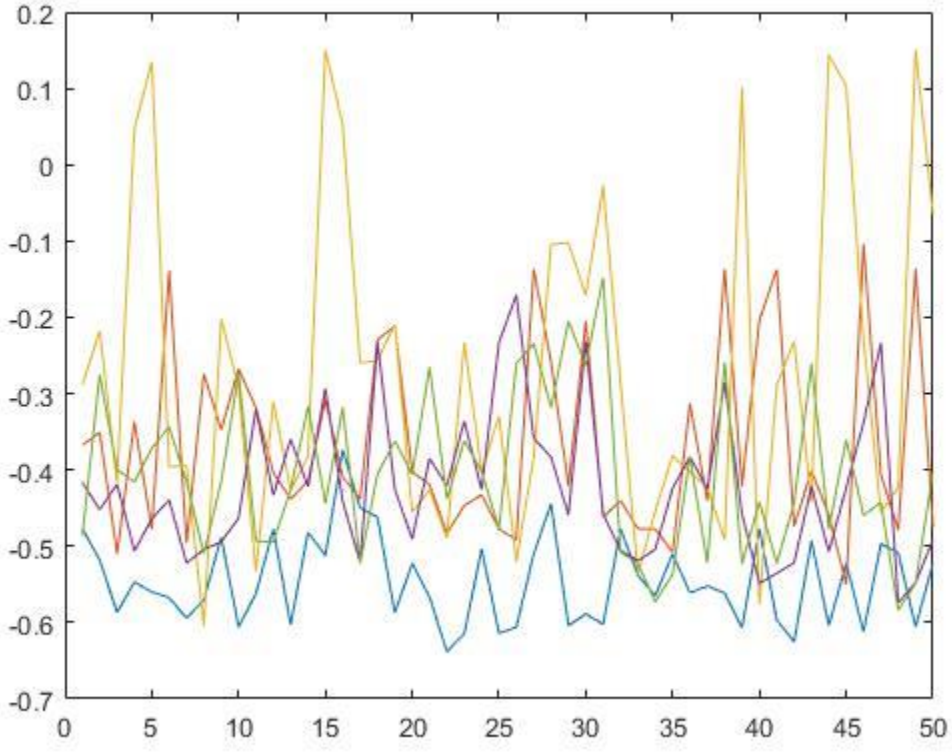



0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
5
10
15
20
25
30
35
40
45
50




-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
0
5
10
15
20
25
30
35
40
45
50




0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
0
5
10
15
20
25
30
35
40
45
50