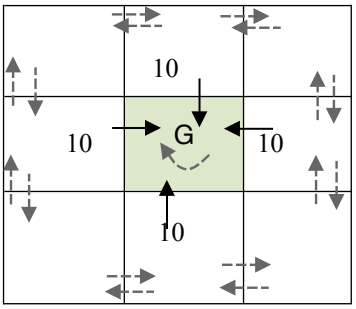
# Práctico 5
## Aprendizaje por Refuerzos

### Ejercicio 1
Sea un «mundo» en donde todos los movimientos tienen retorno nulo menos los cuatro indicados. Para aprender Q se realizan dos episodios, visitando los estados en el orden indicado (en gris la casilla de comienzo):

Mundo




Episodio 1

| | | |
|---|---|---|
| 2 | 3 | 4 |
| 1 | 6 | 5 |
| 0 | | |

Episodio 2

| | | |
|---|---|---|
| | 4 | 3 |
| | 5 | 2 |
| | 0 | 1 |

a) Calcule V* y Q utilizando un factor de descuento γ=0,8. Dé una política óptima.
b) Suponga que se aplica el algoritmo para aprender Q comenzando con todos los valores de la tabla en cero.
   i. Dé los valores de Q' al terminar el primer y el segundo episodio.
   ii. Dé los valores de Q' al terminar el primer y el segundo episodio si se guarda la secuencia de acciones seguida y se actualiza a Q' en orden inverso.
c) ¿Es necesariamente el primer episodio una secuencia de exploración? ¿Es el segundo episodio una secuencia de explotación? Justifique.

### Ejercicio 2
Sea el siguiente mundo redondo, en donde el agente se mueve horizontal o verticalmente, con γ=0,8. Se aplica el algoritmo Q, comenzando por la tabla $Q_0$. Todos los retornos son nulos salvo los indicados.

Mundo

| | | |
|---|---|---|
| $G_2$ | | 200 |
| 100 | | $G_1$ |
| A | | |

$Q_0$

| | | |
|---|---|---|
| | 200 80 | 100 |
| 20 | 100 | |
| 10 | 200 150 | |

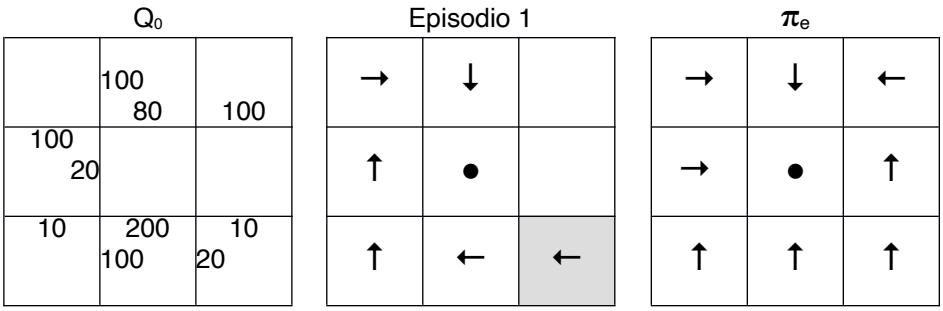
Episodio 1

| | | |
|---|---|---|
| | 3 | 4 |
| | 2 | 5 |
| | 1 | 0 |

i. Calcule V*, Q y dé dos políticas óptimas distintas para el mundo indicado.
ii. Dé una secuencia de explotación partiendo del cuadrado marcado con una A.
iii. Obtenga la tabla $Q_1$ resultante de realizar el episodio 1, guardando en memoria la secuencia y actualizando en orden inverso al recorrido.
iv. ¿Es la secuencia anterior de exploración o explotación? Justifique.

## Ejercicio 3

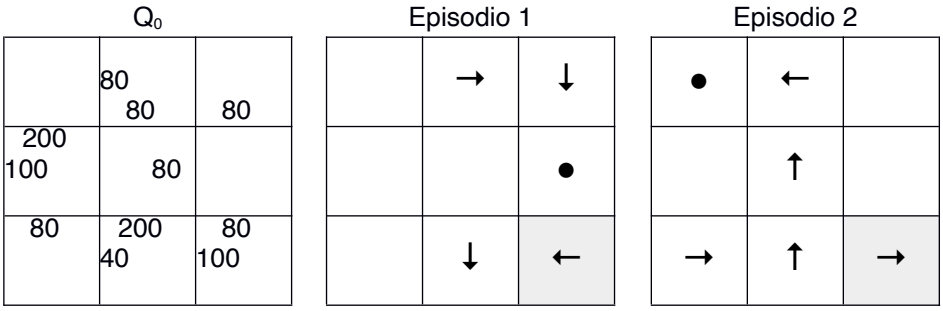
Sea un agente que se mueve horizontal o verticalmente con un factor de descuento de 0,9, en un mundo de 9 casilleros. El agente aplica el algoritmo Q, comenzando por la tabla $Q_0$, en donde todos los retornos son nulos salvo los indicados.




i. El primer episodio, ¿es una secuencia de exploración o explotación? Justifique.
ii. Dé un primer episodio alternativo que haga cambiar su respuesta de la parte anterior.
iii. Dé la tabla $Q_1$ sabiendo que la única recompensa no nula que recibe el agente es al realizar su última acción, con un retorno de +100.
iv. ¿Cuál sería el resultado de $Q_1$ si el agente guarda en memoria la secuencia y actualiza $Q_0$ en orden inverso al recorrido?
v. Sabiendo que solo tres acciones tienen retorno no nulo, establezca una función de recompensa de forma que una posible política óptima del mundo sea $\pi_e$. Tenga en cuenta las recompensas del episodio 1.
vi. Calcule V*y Q según la función de recompensa dada en la parte anterior.

## Ejercicio 4

Un agente se mueve horizontal o verticalmente en una cuadrícula circular, con $\gamma$=0,8. El agente aplica el algoritmo Q, realizando dos episodios de forma secuencial, empezando por la casilla sombreada. Se comienza con la tabla $Q_0$, en donde los valores son nulos salvo los indicados.




a) Dé la tablas $Q_1$ y $Q_2$ sabiendo que en ambos episodios la única recompensa no nula que recibe el agente es al realizar su última acción, con un retorno de +100.
b) ¿Puede afirmar que el algoritmo converge luego del episodio 2? ¿Puede afirmar lo contrario?
c) Dé una secuencia de exploración y otra de explotación para un posible episodio 3 partiendo del mismo punto sombreado.
d) Suponga que el agente guarda en memoria las acciones y retornos durante un episodio. ¿Cuál sería el resultado de la parte (a) si, al finalizar cada episodio, se actualiza Q en orden inverso al recorrido realizado?
e) Asumiendo que los únicos retornos no nulos son los de los episodios 1 y 2, dé V*, Q y $\pi$*.