Entrega final Tarea 1: Programación dinámica

Código: EL7021-1

Nombre: José Luis Cádiz Sejas

Parte I:

Pregunta 1:

- **Espacio de estados:** Dado el espacio $(x, y) \in \Re^2$, definimos el espacio de estados:

$$S = \{(RewardGgrid(x, y) = -1)\ or\ (RewardGgrid(x, y) = 0)\}$$

Donde $RewardGgrid(x, y)$ es la función de recompensa que puede generar valores -1, 0 o NULL según si el estado es de transición, terminal o no factible respectivamente.

En particular si $(RewardGgrid(x, y) = 0)$, estamos hablando del estado terminal:

$$s_t \in S \mid (RewardGgrid(x, y) = 0)$$

- **Espacio de acciones**: $a \in \{0,1,2,3\}$ donde
{"0":"up","1":"down","2":"right","3":"left"}

- **Función de recompensa**: Función independiente de las acciones.

$$R(s, a) = \begin{cases} -1\ if\ s \neq s_t \\ 0\ if\ s = s_t \end{cases}\ donde\ s \in S$$

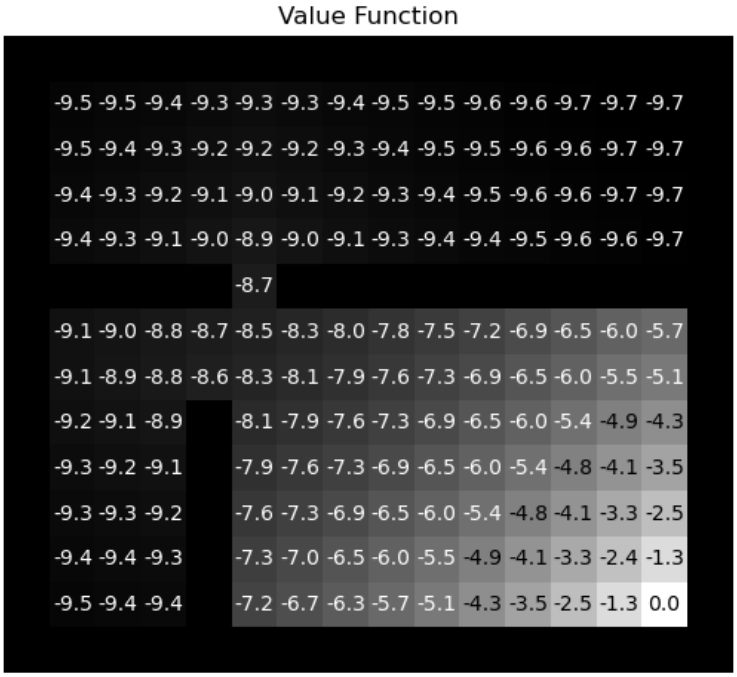
- **Función de transición de estados**: Esta función indica la probabilidad de transición del estado s al estado s'. Dada las acciones que define la política $\pi(a|s)$.

$$\mathrm{T}(s', s, a) = \begin{cases} p_{dir}\ \ if\ a = \pi(a|s) \\ \dfrac{1 - p_{dir}}{2}\ \ if\ a = \perp \pi(a|s)\ (Estados\ perpendiculares\ a\ la\ dirección\ de\ a) \\ 0\ if\ s \notin S\ (Restricción\ de\ paredes) \end{cases}$$

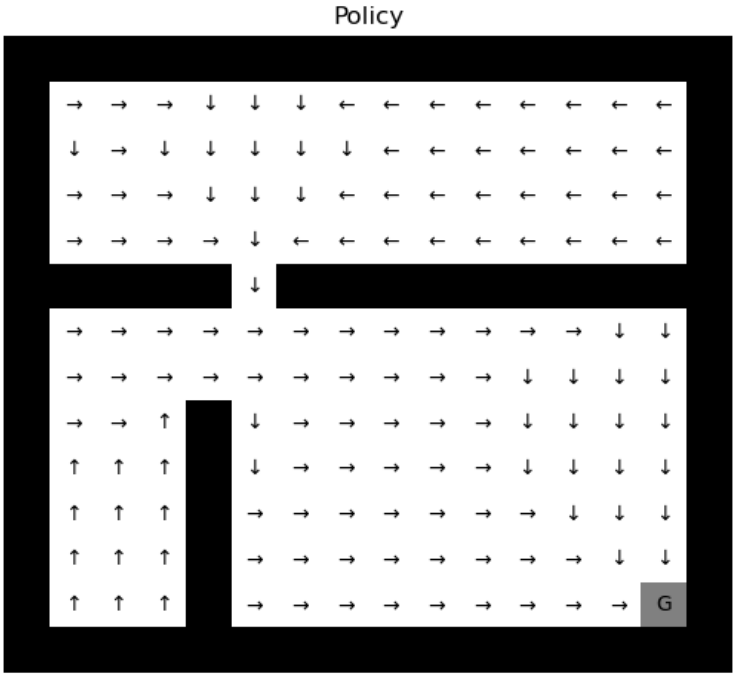
Pregunta 2: Código adjunto.

Pregunta 3:

- **Función de valor**:



- **Política aprendida:**



- **Número de iteraciones sobre la función de valor**: 253 iteraciones en los 11 llamados que se hizo a la función policy_evaluation.

Iteración Policy evaluation	Iteraciones dentro de Policy evaluation
1	84
2	16
3	16
4	41
5	20
6	19
7	34
8	15
9	6
10	1
11	1

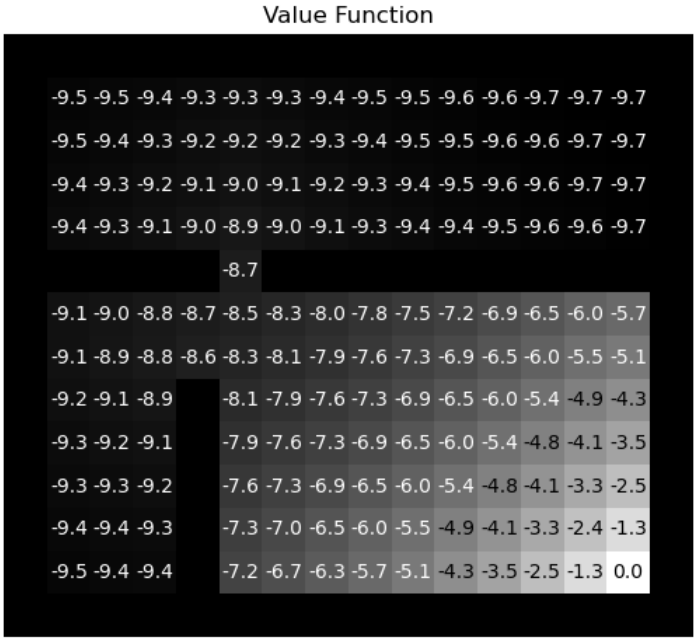
<u>Parte II</u>:

Pregunta 1: Código adjunto.

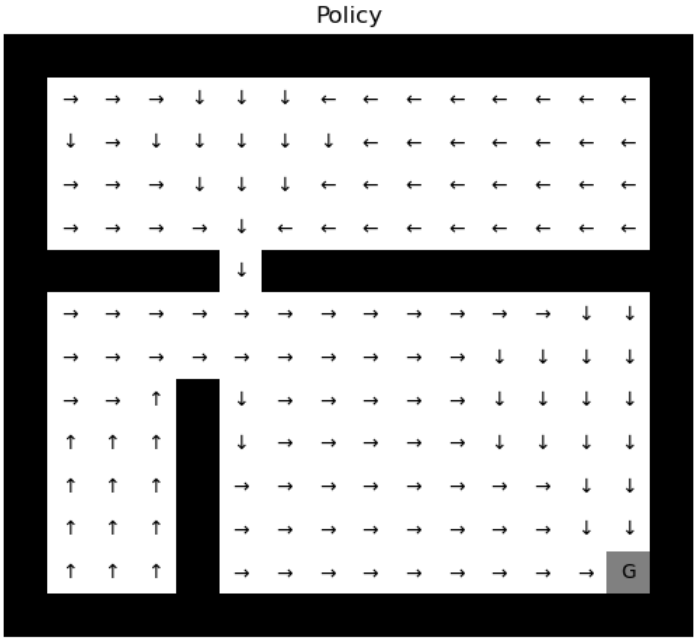
Pregunta 2: $1 - p = 0.2$

- **# de iteraciones:** 36

- **Función de valor**:



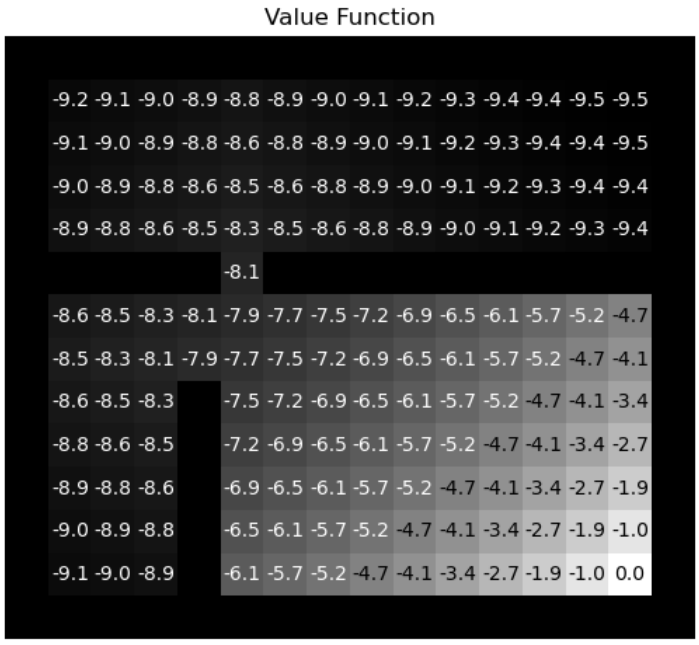
- **Política aprendida:**



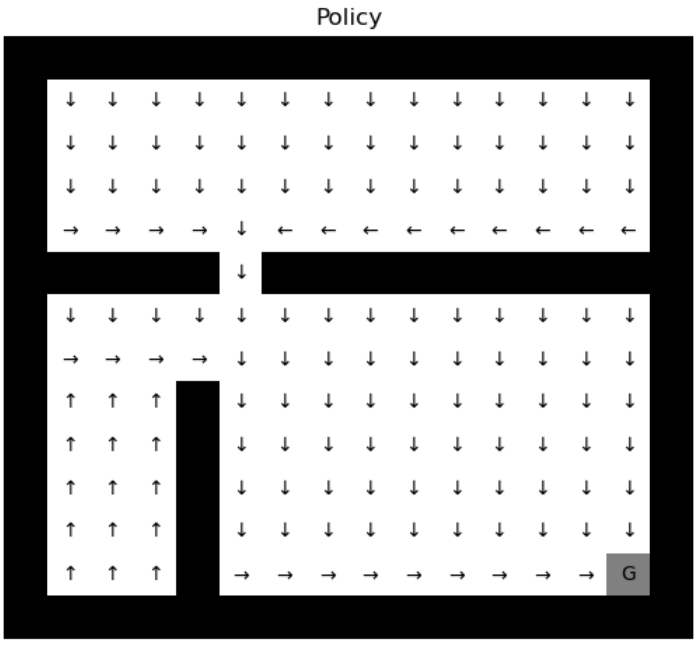
- **Comentarios**: Se obtiene la misma función de valor y política que con policy_evaluation pero con un número menor de iteraciones sobre la función de valor (253 vs 36).

Pregunta 3: $1 - p = 0$

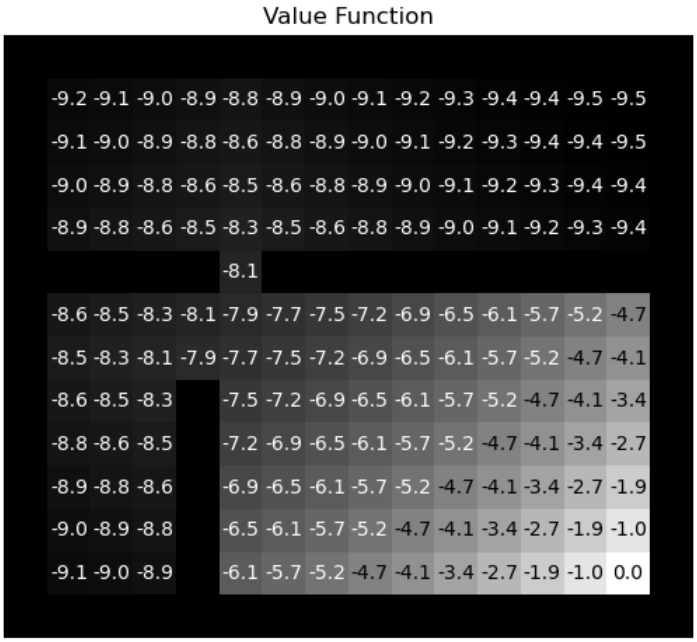
- **Policy_iteration:**
  - **# de iteraciones:** 117
  - **Función de valor:**



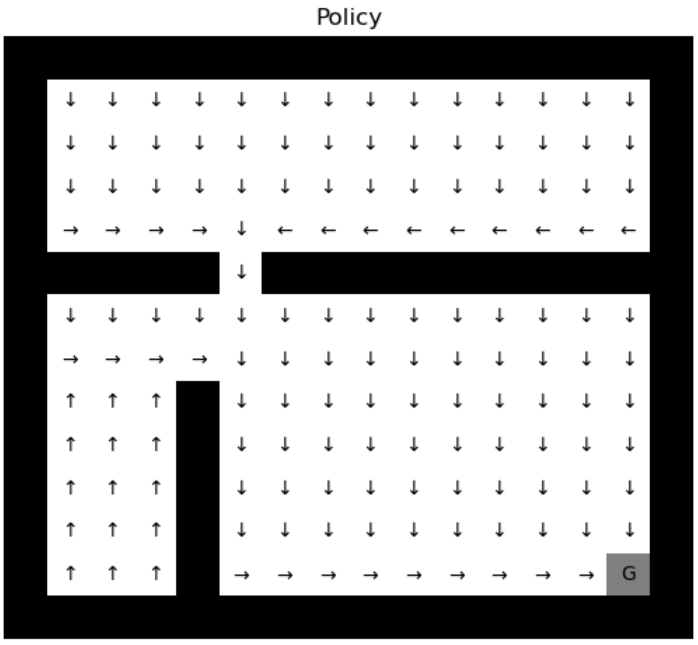
- **Política aprendida:**



- **Value_iteration:**
  - **# de iteraciones:** 30
  - **Función de valor:**



- **Política aprendida:**

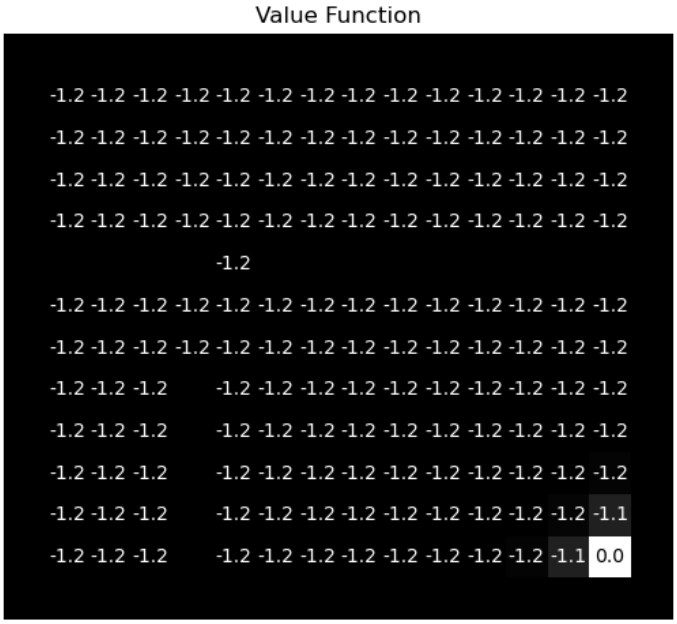


- **Análisis:** Para ambos casos se obtiene la misma función de valor y política aprendida. Por otro lado, se observa a partir de la función de valor, que el efecto que tiene el hecho de que el ambiente sea determinista ($p_{dir} = 1$), disminuye el costo de llegar a la meta final. Además, observando la política aprendida, se aprecia que las direcciones aprendidas son más directas en comparación a cuando el ambiente tiene cierto grado de incertidumbre.

  Adicionalmente también se observa que el número de iteraciones para aprender la política optima disminuye en un ambiente determinista.

Pregunta 4: $1 - p = 0.4$

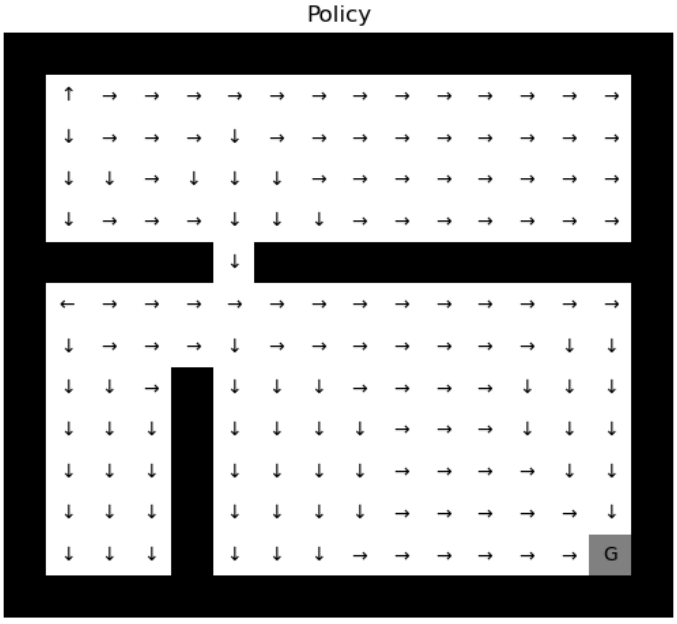
- **Gamma=0.2:**
    - **# de iteraciones:** 7
    - **Función de valor:**



Value Function

-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
				-1.2									
-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2		-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2		-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2		-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
-1.2	-1.2	-1.2		-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.1
-1.2	-1.2	-1.2		-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.1	0.0

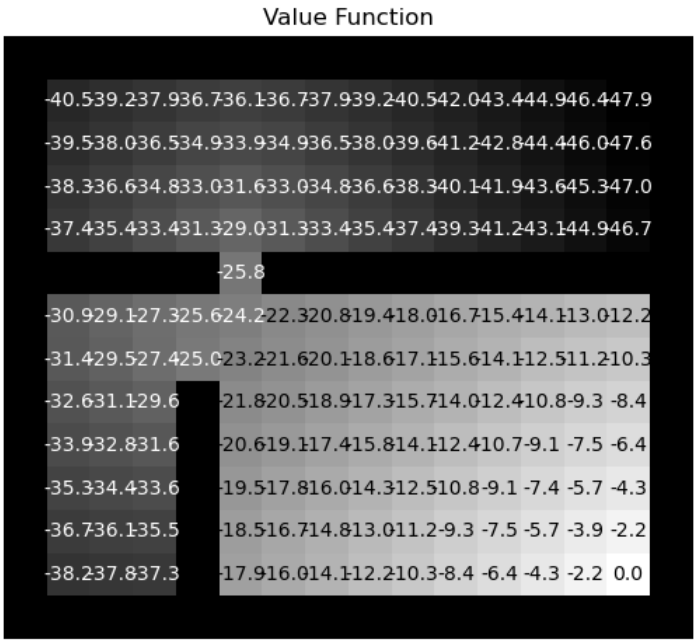
-
    - **Política aprendida:**



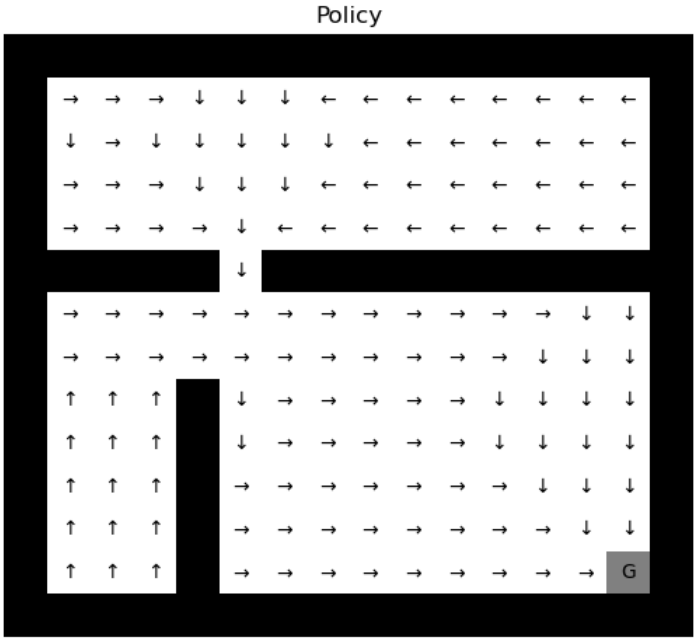
Policy

↑	→	→	→	→	→	→	→	→	→	→	→	→	→
↓	→	→	→	↓	→	→	→	→	→	→	→	→	→
↓	↓	→	↓	↓	↓	→	→	→	→	→	→	→	→
↓	→	→	→	↓	↓	↓	→	→	→	→	→	→	→
				↓									
←	→	→	→	→	→	→	→	→	→	→	→	→	→
↓	→	→	→	↓	→	→	→	→	→	→	→	↓	↓
↓	↓	→		↓	↓	↓	→	→	→	→	↓	↓	↓
↓	↓	↓		↓	↓	↓	↓	→	→	→	↓	↓	↓
↓	↓	↓		↓	↓	↓	↓	→	→	→	→	↓	↓
↓	↓	↓		↓	↓	↓	↓	→	→	→	→	→	↓
↓	↓	↓		↓	↓	↓	→	→	→	→	→	→	G

- **Gamma=1:**
  - **# de iteraciones:** 64
  - **Función de valor:**



  - **Política aprendida:**



- **Interpretación:** La diferencia entre ambas políticas aprendidas radica en el grado de importancia que se les dan a las recompensas futuras, para el caso en que gamma=0.2, se le está dando gran importancia las recompensar inmediatas, por otro lado, para el caso gamma=1 se le está dando la mayor importancia posible a las recompensas futuras.

  Para el caso gamma=0.2, no se alcanza a aprender una política óptima en los casos en que los estados iniciales están muy alejados de la meta, lo cual tiene sentido debido a que el contexto del problema amerita en darle relevancia a las recompensas en el largo plazo. Esto también se aprecia en la función de valor, en donde para la mayoría de los estados se obtiene un valor -1.2.

  Para el caso gamma=1, se logra obtener una política optima para todos los estados, pero se observa un notable aumento del número de iteraciones para obtener la política.

Pregunta 5: Gamma=1 representa que el agente le da exactamente la misma importancia a cada una de las recompensas inmediatas y futuras, lo cual en el contexto del problema es útil, sin embargo, puede haber un aumento del tiempo para encontrar la política optima en comparación con un gamma=0.9.