**Entrega final Tarea 1:** Programación dinámica

**Código:** EL7021-1

**Nombre:** José Luis Cádiz Sejas

**Parte I**:

**Pregunta 1:**

* **Espacio de estados:** Dado el espacio , definimos el espacio de estados:

Donde 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.

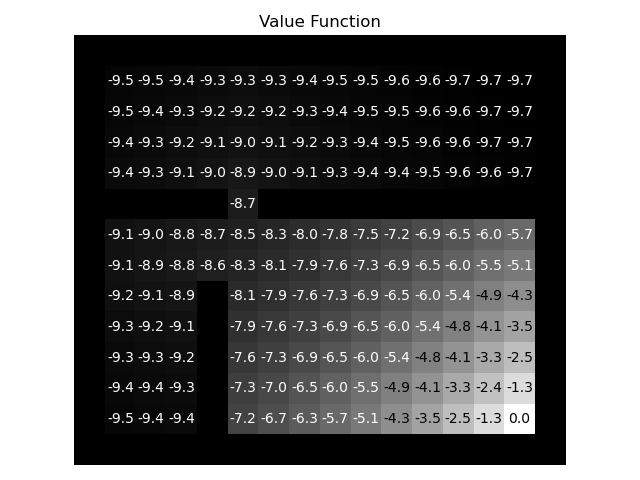
En particular si , estamos hablando del estado terminal:

* **Espacio de acciones**: donde {"0":"up","1":"down","2":"right","3":"left"}
* **Función de recompensa**: Función independiente de las acciones.
* **Función de transición de estados**: Esta función indica la probabilidad de transición del estado al estado . Dada las acciones que define la política

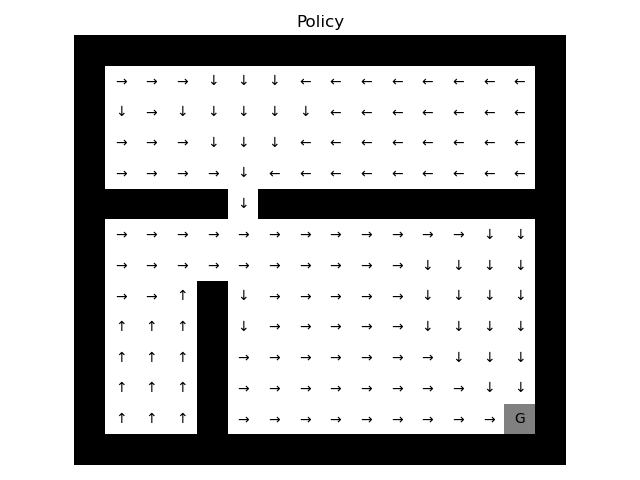
**Pregunta 2:** Código adjunto.

**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.

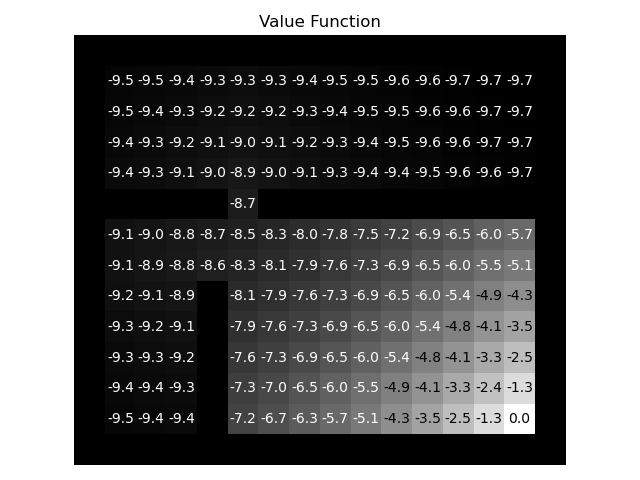
|  |  |
| --- | --- |
| **Iteración Policy evaluation** | **Iteraciones dentro de Policy evaluation** |
| 1 | 84 |
| 2 | 16 |
| 3 | 16 |
| 4 | 41 |
| 5 | 20 |
| 6 | 19 |
| 7 | 34 |
| 8 | 15 |
| 9 | 6 |
| 10 | 1 |
| 11 | 1 |

**Parte II**:

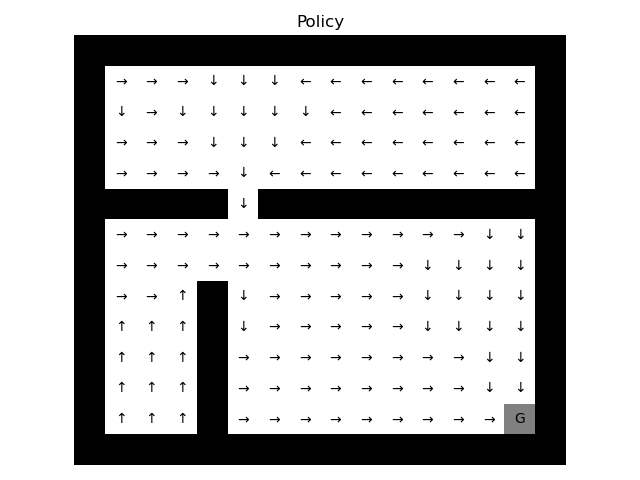
**Pregunta 1:** Código adjunto.

**Pregunta 2:**

* **# de iteraciones:** 36
* **Función de valor**:



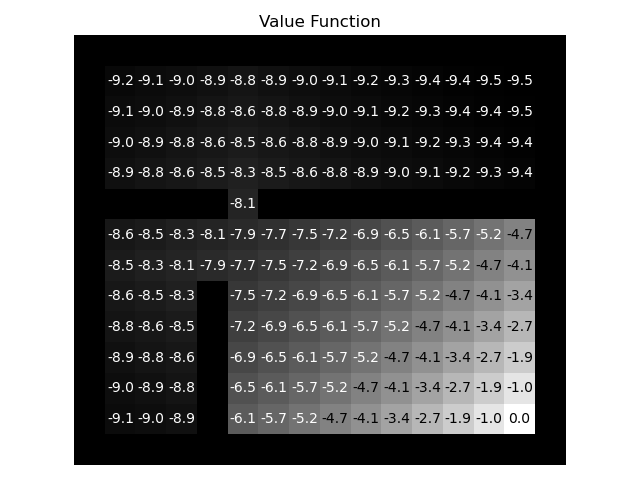
* **Política aprendida:**



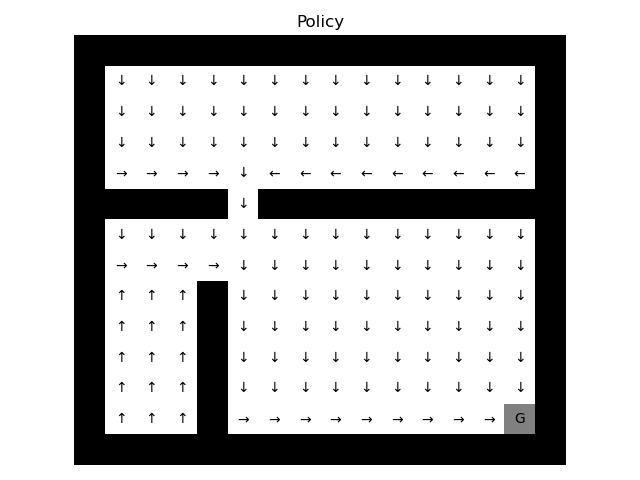
* **Comentarios**: Se obtiene la misma función de valor y política que con policy\_evaluation pero con un número menor de iteraciones sobre la función de valor (253 vs 36).

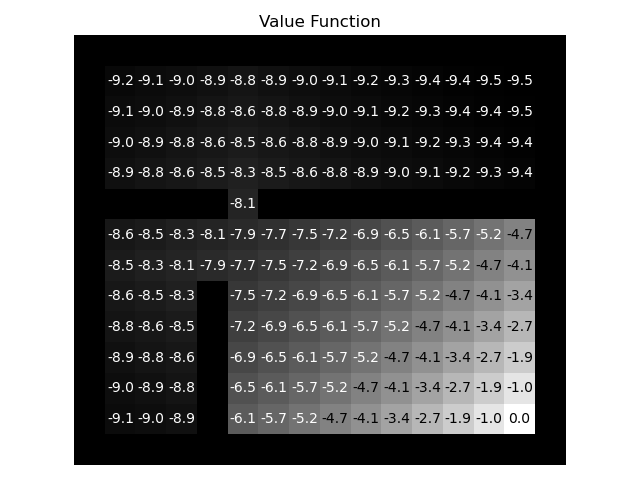
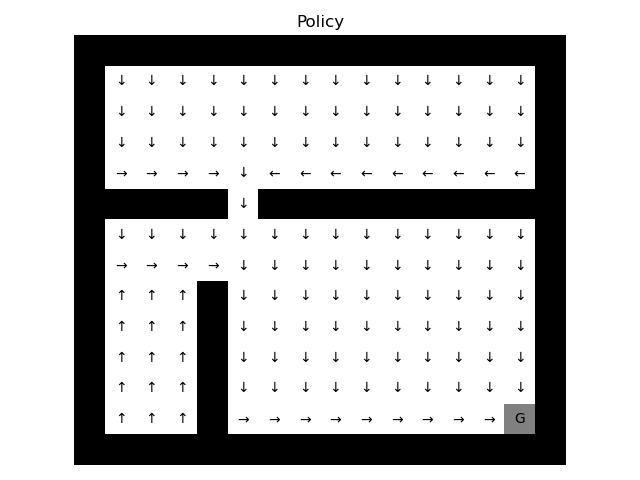
**Pregunta 3:**

* **Policy\_iteration:**
  + **# de iteraciones:** 117
  + **Función de valor:**



* + **Política aprendida:**

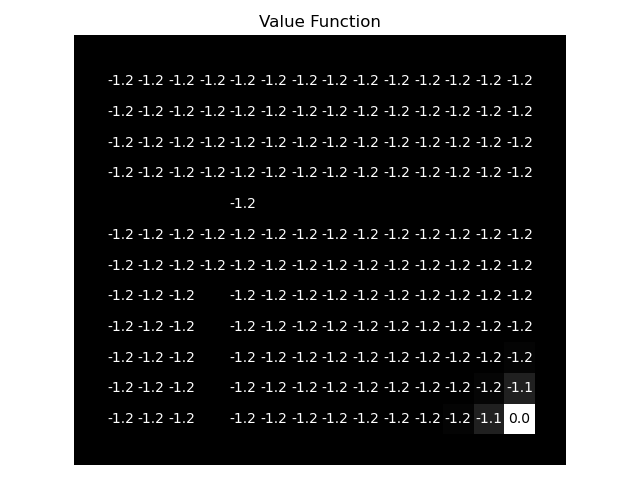


* **Value\_iteration:**
  + **# de iteraciones:** 30
  + **Función de valor:** 
  + **Política aprendida: **
  + **Análisis:** Para ambos casos se obtiene la misma función de valor y política aprendida. Por otro lado, se observa a partir de la función de valor, que el efecto que tiene el hecho de que el ambiente sea determinista , 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.

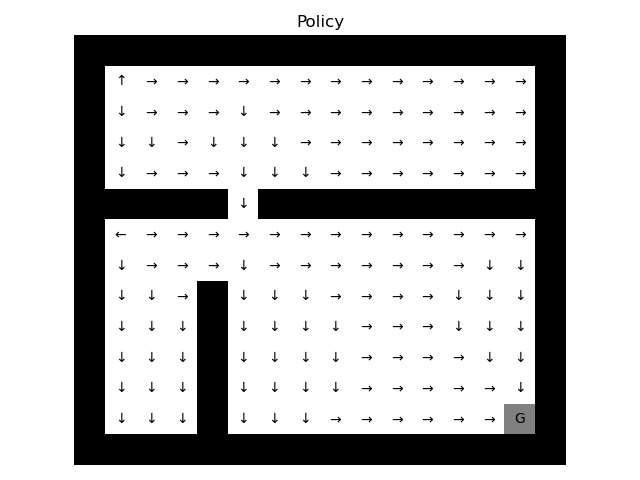
Adicionalmente también se observa que el número de iteraciones para aprender la política optima disminuye en un ambiente determinista.

**Pregunta 4**:

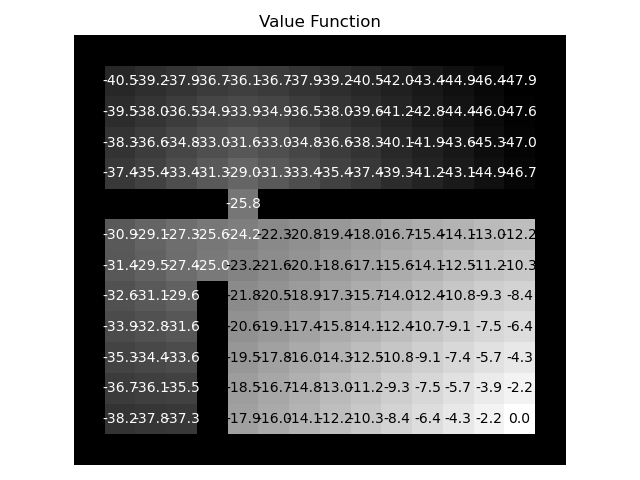
* **Gamma=0.2:**
  + **# de iteraciones:** 7
  + **Función de valor:**



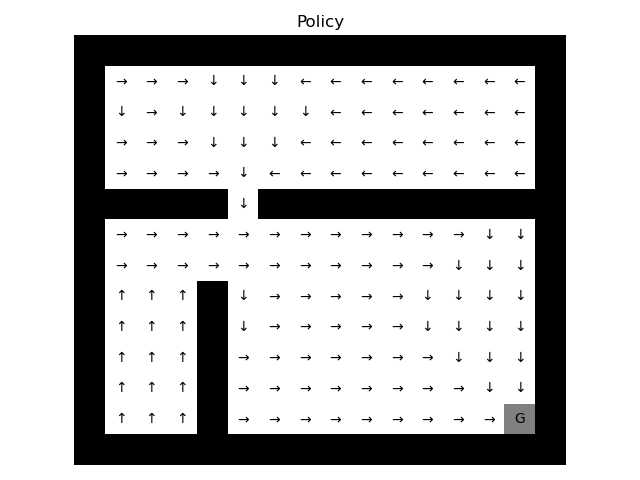
* + **Política aprendida:**

****

* **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.