Introduccion a Reinforcement Learning

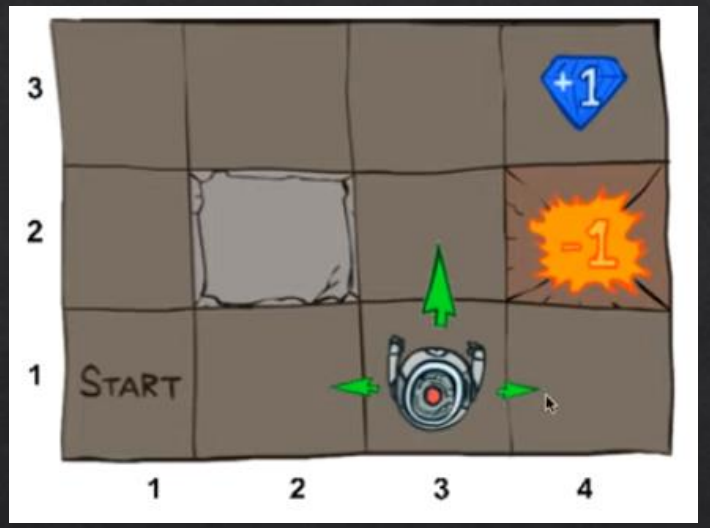
2.4 metodos de solución

Jorge Vasquez

Ejercicio de Grilla

Un MDP está definido por:

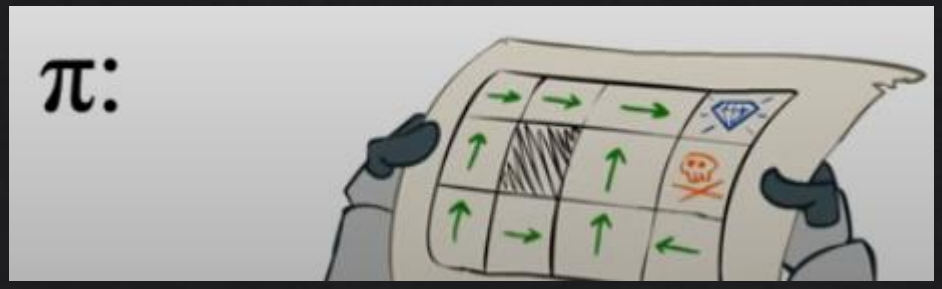
- Set de Estados S
- Set de acciones A
- Función de Transición P(s´ | s, a)
- Función de Refuerzo R (s, a, s')
- Estado inicial So
- Factor descuento γ
- Horizonte H



Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

Política:



Métodos de Soluciones Exactos

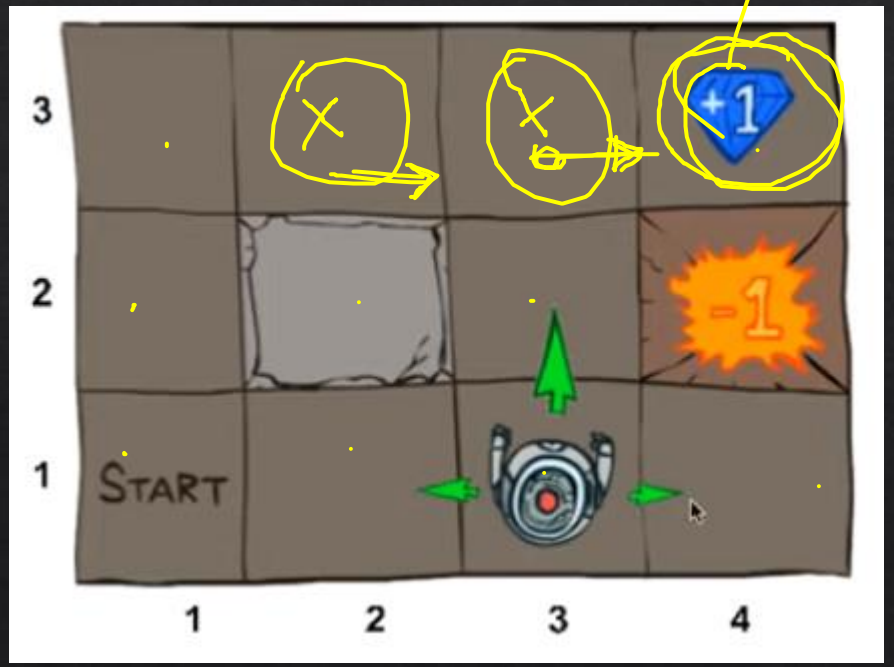
- Iteración de Valor

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

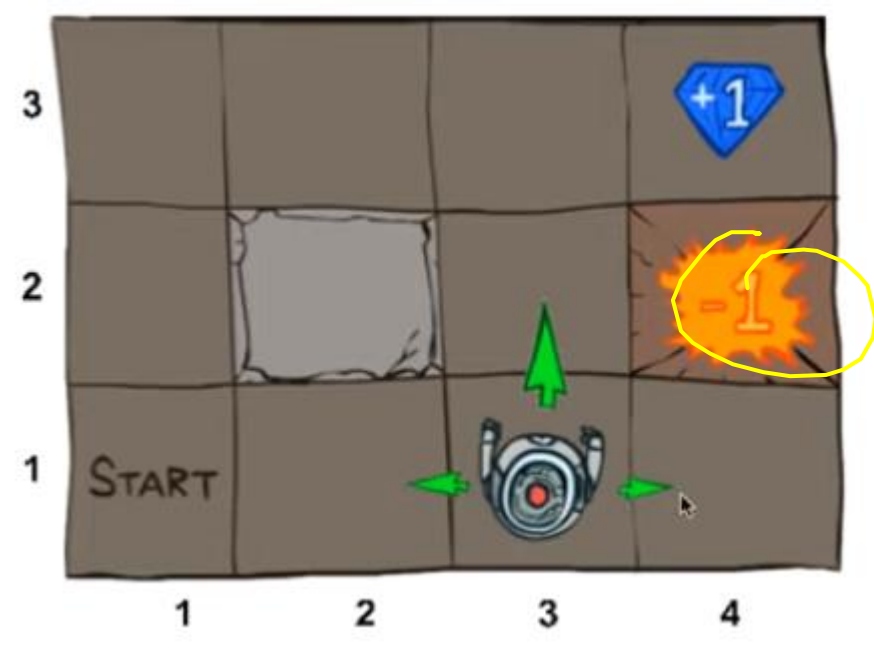


Asumamos:
Las acciones son exitosas en forma determinística, gamma = 1, y H = 100

- V* (4,3) = 1
- V* (3,3) =
- V* (2,3) =
- V* (1,1) =
- V* (4,2) =

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

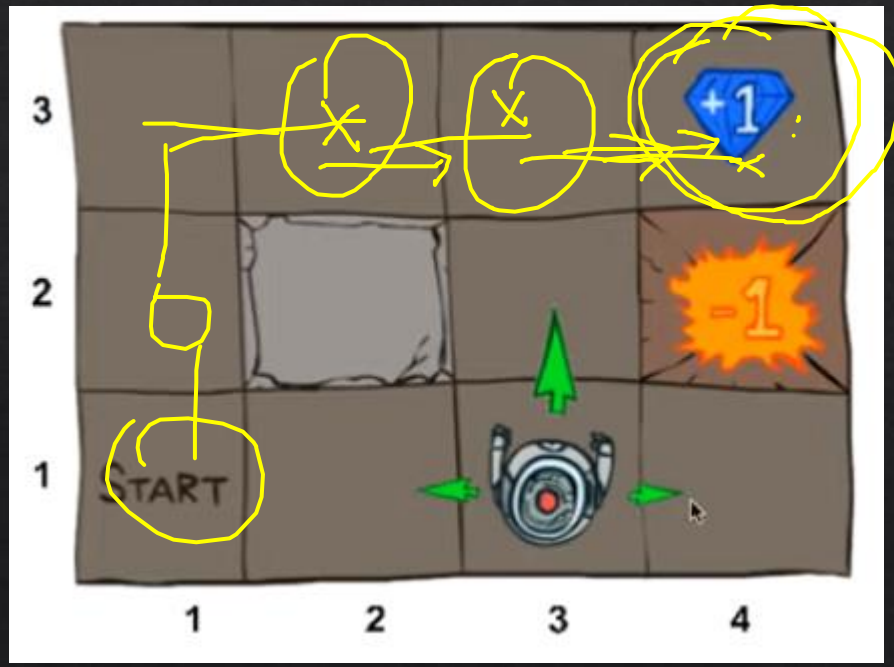


Asumamos:
Las acciones son exitosas en forma determinística, gamma = 1, y H = 100

- V* (4,3) = 1
- V* (3,3) = 1
- V* (2,3) = 1
- V* (1,1) = 1
- V* (4,2) = -1

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$



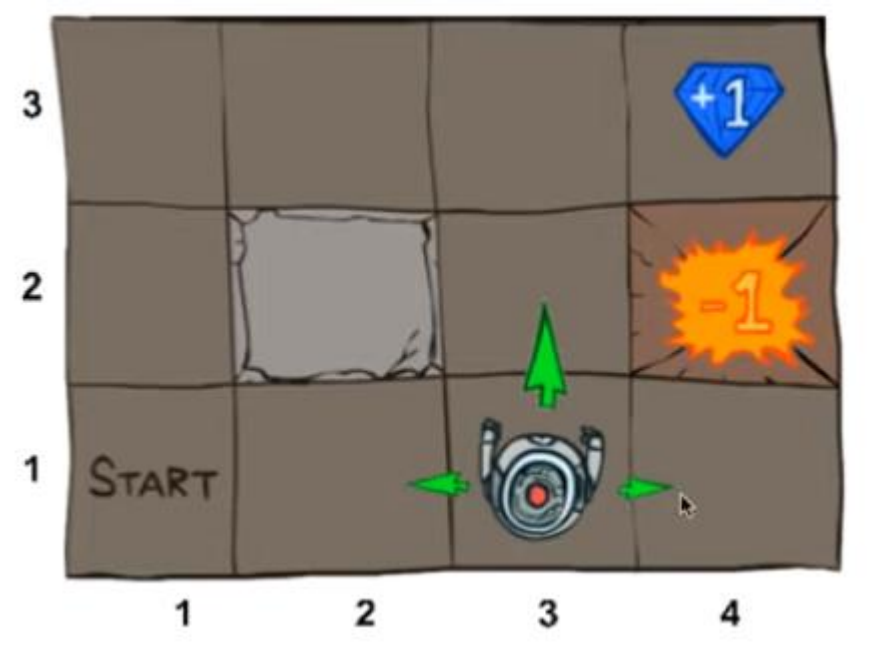
$\gamma = 0.9$

Asumamos:
Las acciones son existosas en forma determinística, gamma = 0.9, y H = 100

- V* (4,3) = 1
- V* (3,3) = $\gamma^0 r^0 + \gamma^1 r^1 = 0.9 \cdot 1 = 0.9$
- V* (2,3) = $\gamma^0 r^0 + \gamma^1 r^1 + \gamma^2 r^2 = 0.9^2 \cdot 1 = 0.81$
- V* (1,1) = $0.9^5 \cdot 1 =$
- V* (4,2) = -1

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

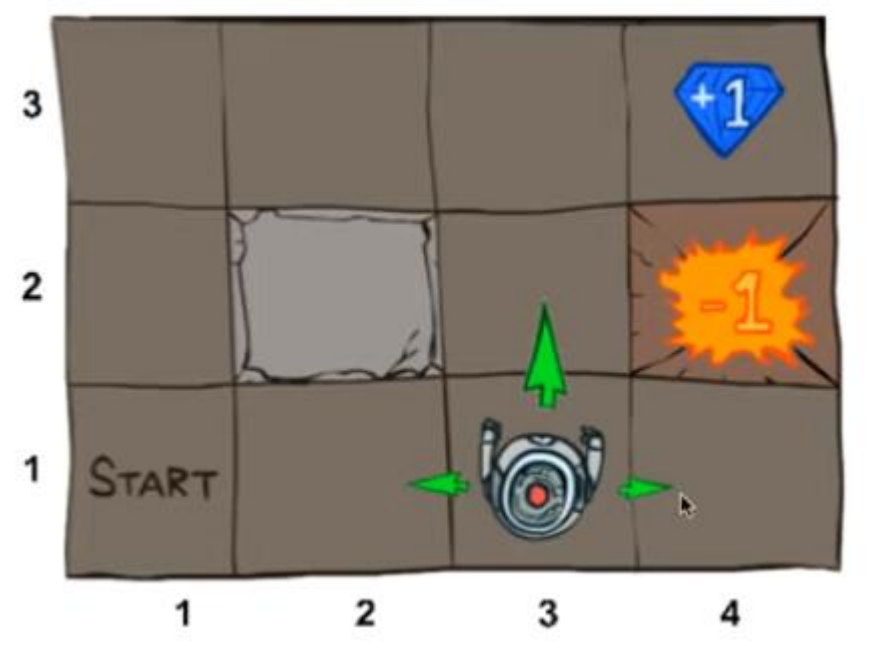


Asumamos:
Las acciones son exitosas en forma determinísticas, gamma = 0.9, y H = 100

- V* (4,3) = 1
- V* (3,3) = 0.9
- V* (2,3) =
- V* (1,1) =
- V* (4,2) =

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

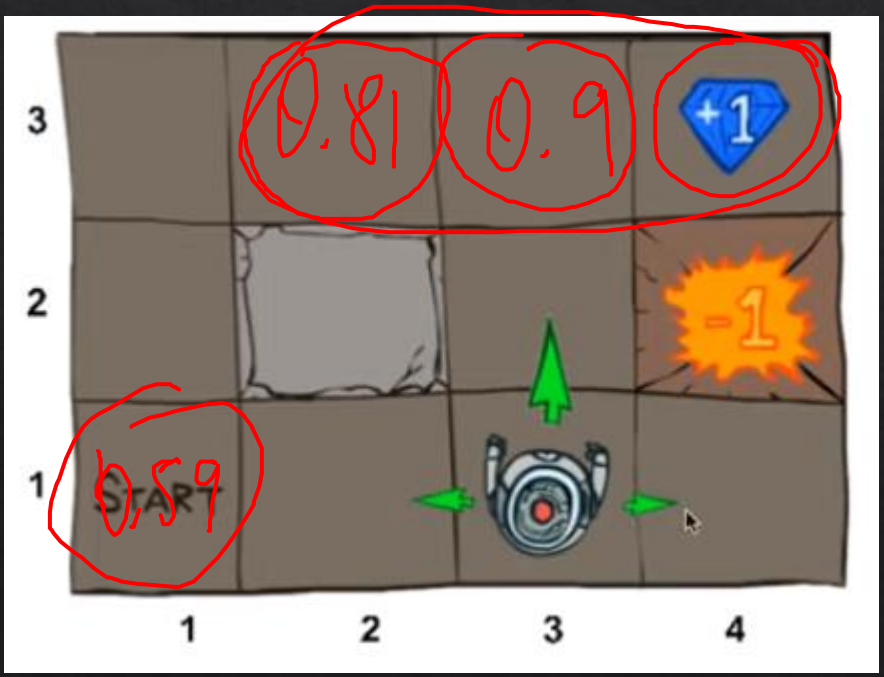


Asumamos:
Las acciones son son exitosas en forma determinística,
gamma = 0.9, y H = 100

- V* (4,3) = 1
- V* (3,3) = 0.9
- V* (2,3) = 0.9*0.9=0.81
- V* (1,1) =
- V* (4,2) =

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

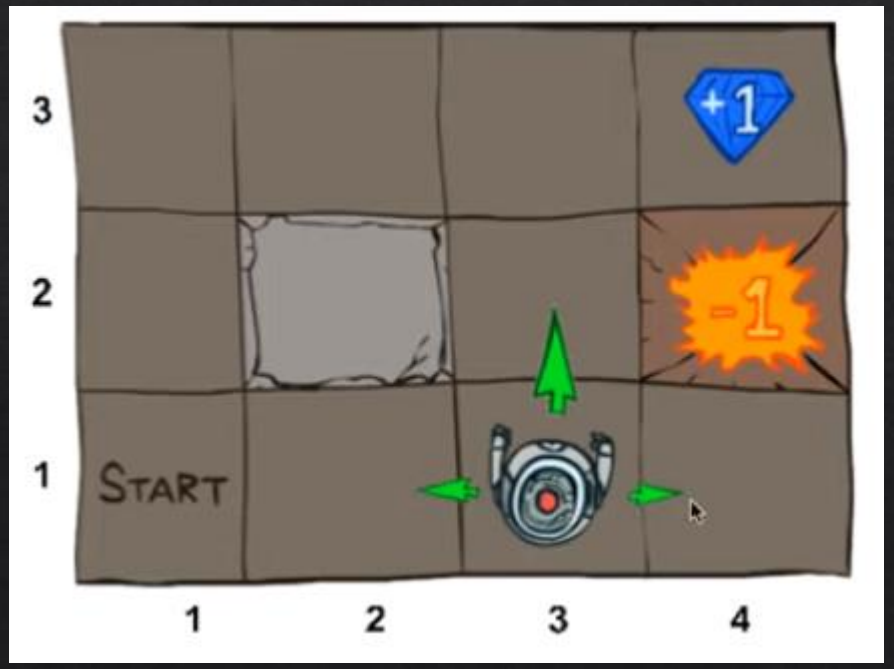


Asumamos:
Las acciones son son exitosas en forma determinística, gamma = 0.9, y H = 100

- V* (4,3) = 1
- V* (3,3) = 0.9
- V* (2,3) = 0.9*0.9=0.81
- V* (1,1) = 0.9^5=0.59
- V* (4,2) = -1

Función de Valor Optimal V*

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$



Asumamos:
Las acciones son exitosas con probabilidades de 0.8, gamma = 0.9, y H = 100

- V* (4,3) = 1
- V* (3,3) = 0.8 * 0.9 * V*(4,3) + 0.1 * 0.9 * V*(3,2) + 0.8 * 0.9 * V*(3,2)
- V* (2,3) =
- V* (1,1) =
- V* (4,2) =

Iteración de Valor

- $V_0^*(s)$ = valor óptimo para estado s cuando H=0
  - $V_0^*(s) = 0 \; \forall s$
- $V_1^*(s)$ = valor óptimo para estado s cuando H=1
  - $V_1^*(s) = max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_0^*(s'))$
- $V_2^*(s)$ = valor óptimo para estado s cuando H=2
  - $V_1^*(s) = max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_0^*(s'))$
- $V_k^*(s)$ = valor óptimo para estado s cuando H=k
  - $V_k^*(s) = max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$

Eq. Bellman

Iteración de Valor

algoritmo:

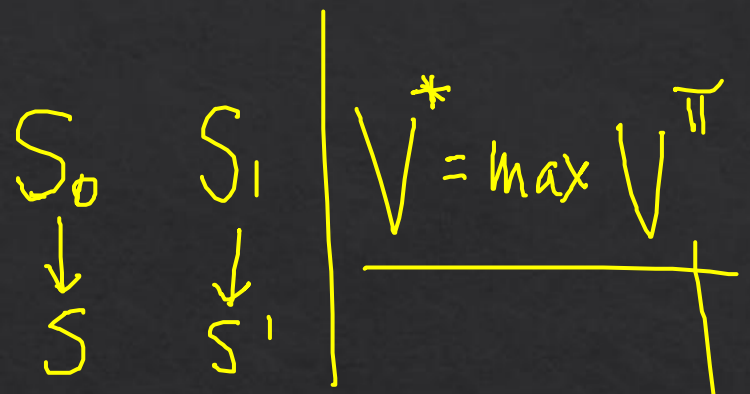
Comienza con $V_0^*(s) = 0 \; \forall s$

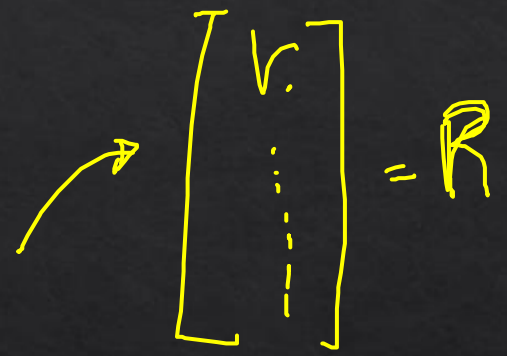
Para k=1, …., H

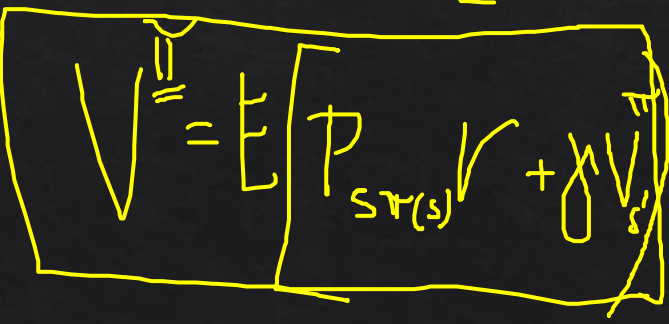
Para todos los estados s en S:

$$V_k^*(s) = max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$

$$\pi_k^*(s) = argmax_a \sum_{s} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$



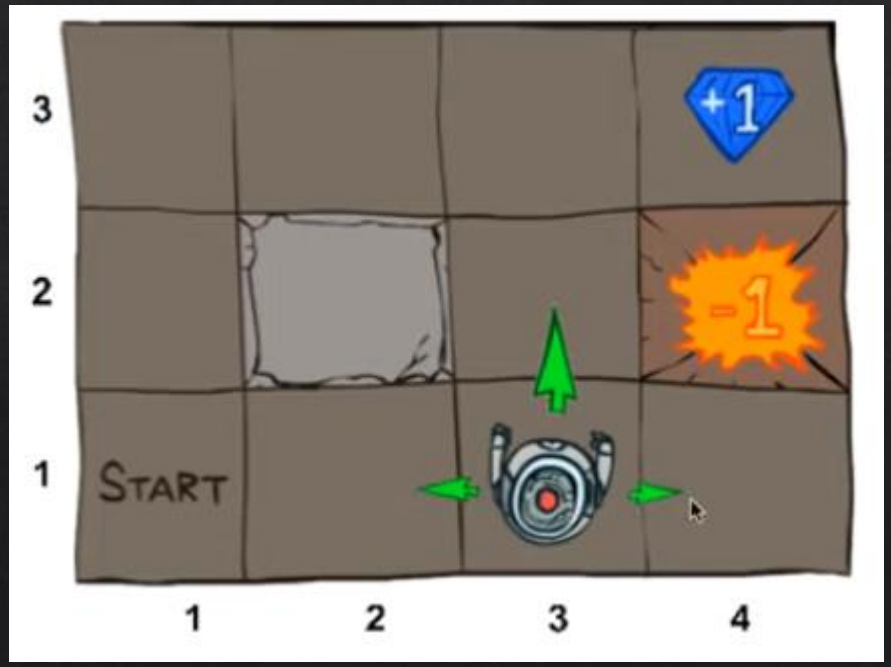




Iteración de Valor

$$V_0^*(s) \leftarrow 0$$

k=0



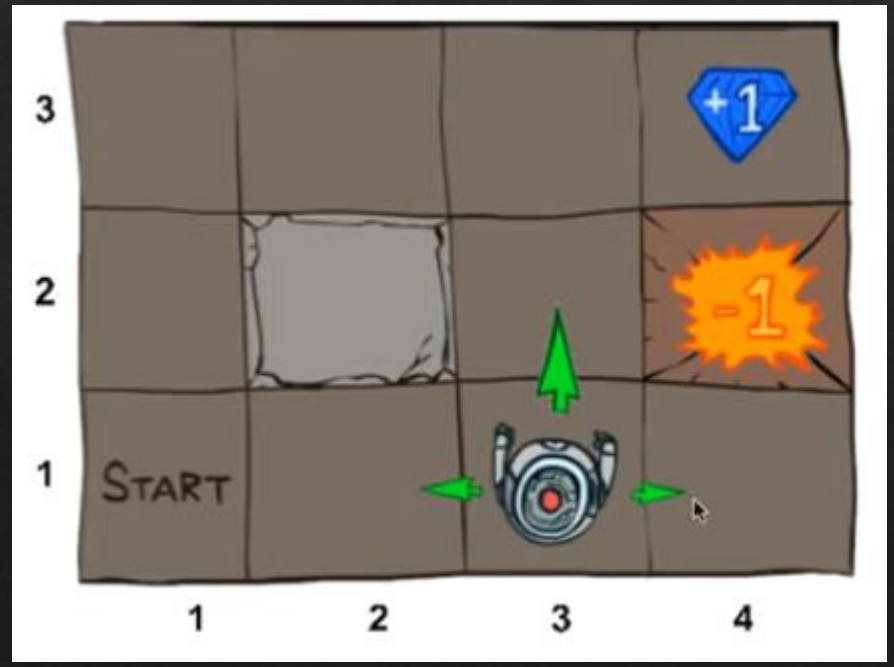
0	**0**	**0**	**0**
0		0	0
0	0	0	0

Valores después de 0 iteración

Iteración de Valor

$$V_2^*(s) \leftarrow max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_1^*(s'))$$

k=1



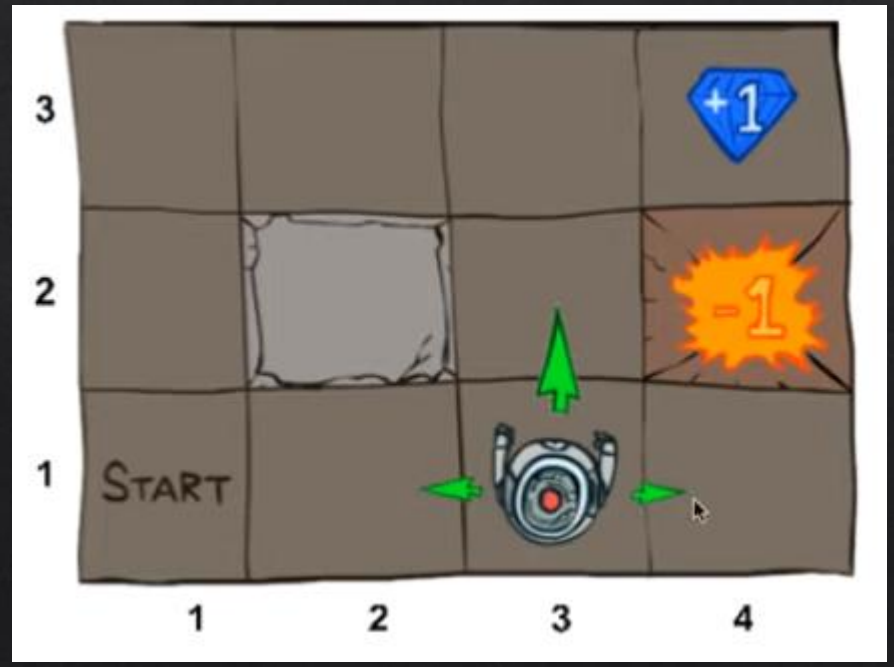
0	**0**	**0**	1
0		0	-1
0	0	0	0

Valores después de 1 iteración

Iteración de Valor

$$V_3^*(s) \leftarrow max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_1^*(s'))$$

k=2

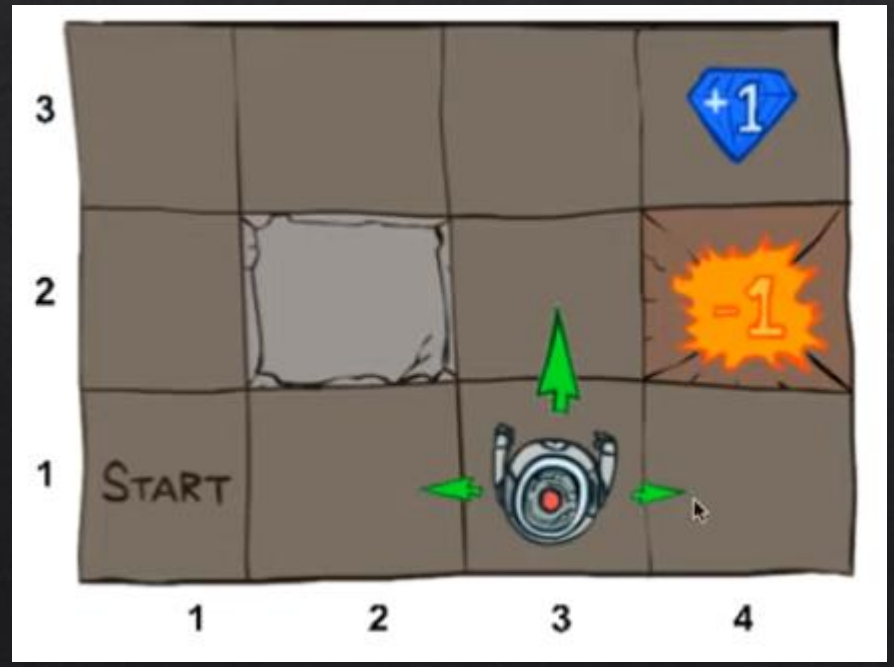


0	**0**	**0.72**	**1**
0		0	-1
0	0	0	0

Valores después de 2 iteraciones

Iteración de Valor

$$V^*_{k+1}(s) \leftarrow max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V^*_k(s'))$$



k=9

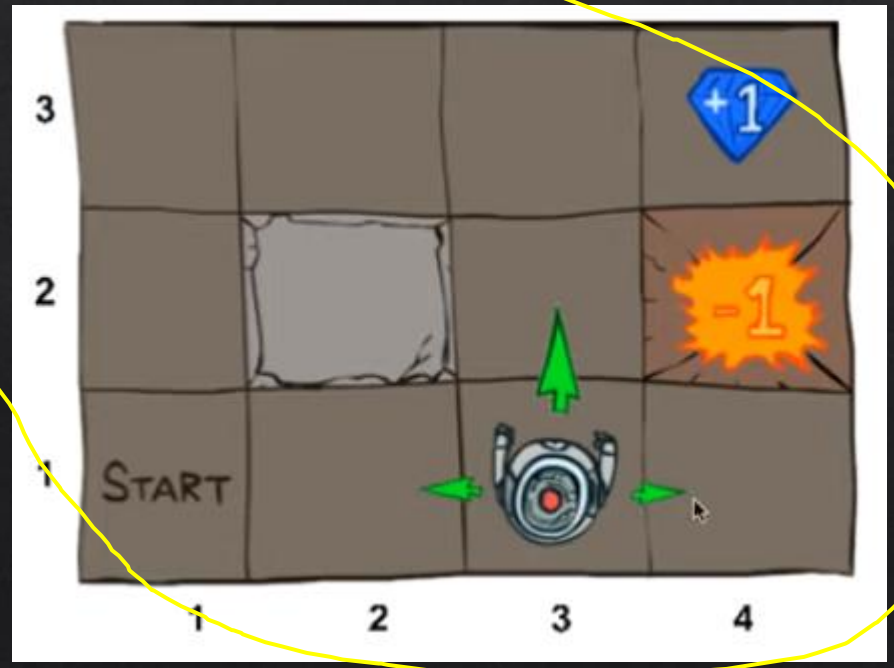
0.64	**0.74**	**0.85**	**1.00**
0.55		0.57	-1.00
0.46	0.40	0.47	0.27

Valores después de 9 iteraciones

Iteración de Valor

$$V^*_{k+1}(s) \leftarrow max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V^*_k(s'))$$

k=100



0.64	**0.74**	**0.85**	**1.00**
0.55		0.57	-1.00
0.49	0.43	0.48	0.28

Valores después de 100 iteraciones

Convergencia de la iteración de Valor

Teorema:
El valor iterado converge. En convergencia, encontramos que la función de valor optimal V^* para el problema de horizontes infinitos descontados , lo que satisface a las ecuaciones de Bellman.

$$\forall S \in S:$$

$$V^*(s) \leftarrow max_A \sum_{s'} T(s,a,s') \left[(R(s,a,s') + \gamma V_k^*(s')\right]$$

$$\pi \leftarrow \text{arg max}$$

Convergencia de la iteración de Valor

Ecuaciones de Bellman.

$$V^*(s) \leftarrow max_A \sum_{s'} T(s, a, s') \left[(R(s, a, s') + \gamma V_k^*(s') \right]$$

Convergencia de la iteración de Valor

Ecuaciones de Bellman.

$$V^*(s) \leftarrow max_A \sum_{s'} T(s, a, s') \left[(R(s, a, s') + \gamma V_k^*(s') \right]$$

Convergencia de la iteración de Valor

Ecuaciones de Bellman (Update Rule)

$$V^*(s) \leftarrow max_A \sum_{s'} P(s,a,s') [(R(s,a,s') + \gamma V_k^*(s')]$$

- Def → Regla de Actualización
- Recursiva → programación dinámica

Convergencia de la iteración de Valor

Ahora sabemos como actuar para horizontes infinitos con recompensas descontadas.

1. Hacer correr la iteración del valor hasta su convergencia
2. Esto genera V*, lo que nos dice como actuar, y se escribe así:

$$\pi^*(s) = argmax_{a \in A} \sum_{s'} T(s, a, s') \left[(R(s, a, s') + \gamma V^*(s') \right]$$

Notar que la política optimal de horizontes infinitos es estacionaria, esto quiere decir la acción optima para el estado s es la misma acción siempre

Convergencia de la iteración de Valor

- V* (s) suma de recompensas esperadas acumuladas desde el estado s, actuando óptimamente para pasos infinitos.
- V_H^{**} (s) suma de recompensas esperadas acumuladas desde el estado s, actuando óptimamente para H numero pasos.
- Adicionalmente, recompensas coleccionadas sobre tiempo H+1, H*2

$$\gamma^{H+1}R(s_{H+1}) + \gamma^{H+2}R(s_{H+2}) + \ldots \leq \gamma^{H+1}R_{max} + \gamma^{H+2}R_{max} + \ldots = \frac{\gamma^{H+1}}{1-\gamma}R_{max}$$

- Tiende a cero cuando H va a infinito
- Entonces,

$$V_H^* \xrightarrow{H\to\infty} V*$$