5. El mundo tiene el siguiente conjunto de estados S={s1, s2, s3, sF1, sF2} donde s1=estado inicial y, sF1 y sF2 son estados terminales:

sF1	s1	s2	s3	sF2

El mundo tiene el siguiente conjunto de acciones A={→, ←} donde:

→=Agente se mueve a la derecha una sola celda

←=Agente se mueve a la izquierda una sola celda

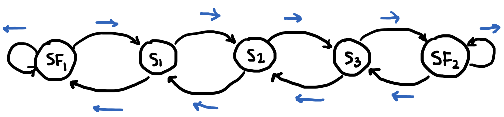
La función de recompensa $f_R(s, a, s_f) = f_R(s_f)$ solo depende del estado al que el Agente llega y esta definida como:

-10	0	-0.4	-0.4	10

Es decir, si el agente transiciona de s1 a s2 entonces recibe la recompensa -0.4 que esta definide en el estado s2. El agente tiene la siguiente función de acción $f_\pi(s)$:

$$f_\pi(s) = \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_{F1} \\ s_{F2} \end{array} \begin{bmatrix} \rightarrow \\ \rightarrow \\ \rightarrow \\ \leftarrow \\ \rightarrow \end{bmatrix}$$

(a) Build the graph of the world



(b) Write the transition function $f_{MT}(s,a)$

$$f_{MT}(s,a) = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ sF_1 \\ sF_2 \end{array} \begin{array}{c} \begin{array}{cc} \leftarrow & \rightarrow \end{array} \\ \begin{bmatrix} sF_1 & s_2 \\ s_1 & s_3 \\ s_2 & sF_2 \\ sF_1 & s_1 \\ s_3 & sF_2 \end{bmatrix} \end{array}$$

(c) Build all the possible trajectories from the initial state s_1 given the action function $f_\pi(s)$ that go to a final state, either sF_1 or sF_2

$$f_\pi(s) = \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ sF_1 \\ sF_2 \end{array} \begin{bmatrix} \rightarrow \\ \rightarrow \\ \rightarrow \\ \leftarrow \\ \rightarrow \end{bmatrix}$$

$\tau_1 = s_1, s_2, s_3, sF_2$

$\tau_2 = s_1, s_2, s_3, sF_2, s_3, sF_2$

(d) Build all the possible trajectories from the state s_2, given the action function $f_\pi(s)$ that go to a final state, either sF_1 or sF_2

$\tau_1 = s_2, s_3, sF_2$

$\tau_2 = s_2, s_3, sF_1, s_3, sF_2$

(e) Build all the possible trajectories from the state s_3, given the action function $f_\pi(s)$ that go to a final state, either sF_1 or sF_2

$\tau_1 = s_3, sF_2$

$\tau_2 = s_3, sF_2, s_3, sF_2$

(f) Calculate the accumulated reward of every possible trajectory in c, d, e using $\gamma = 0.7$.

sF_1	s_1	s_2	s_3	sF_2
-10	0	-0.4	-0.4	10

$s(0) = s_1$

$\tau_1 = s_1, s_2, s_3, sF_2$ $\quad f_{RA} = -0.4 + (0.7)(-0.4) + (0.7)^2(10) = 4.22$

$\tau_2 = s_1, s_2, s_3, sF_2, s_3, sF_2$ $\quad f_{RA} = -0.4 + (0.7)(-0.4) + (0.7)^2(10) + (0.7)^3(-0.4) + (0.7)^4(10) = 6.48$

$s(0) = s_2$

$\tau_1 = s_2, s_3, sF_2$ $\quad f_{RA} = -0.4 + (0.7)(10) = -0.4 + 7 = 6.6$

$\tau_2 = S_2, S_3, SF_2, S_3, SF_2$ $\quad f_{RA} = -0.4 + (0.7)(10) + (0.7)^2(0.4) + (0.7)^3(10) = 9.83$

$s(0) = S_3$

$\tau_1 = S_3, SF_2$ $\quad f_{RA} = 10$

$\tau_2 = S_3, SF_2, S_3, SF_2$ $\quad f_{RA} = 10 + (0.7)(-0.4) + (0.7)^2(10) = 14.62$

6. El mundo tiene el siguiente conjunto de estados S={s1, s2, s3, sF1, sF2} donde s1=estado inicial y, sF1 y sF2 son estados terminales:

sF1	s1	s2	s3	sF2

El mundo tiene el siguiente conjunto de acciones A={→, ←} donde:

→=Agente se mueve a la derecha una sola celda con probabilidad 0.8 y se mueve una sola celda a la izquierda con probabilidad 0.2

←=Agente se mueve a la izquierda una sola celda con probabilidad 0.8 y se mueve una sola celda a la derecha con probabilidad 0.2

La función de recompensa $f_R(s, a, s_f) = f_R(s_f)$ solo depende del estado al que el Agente llega y esta definida como:

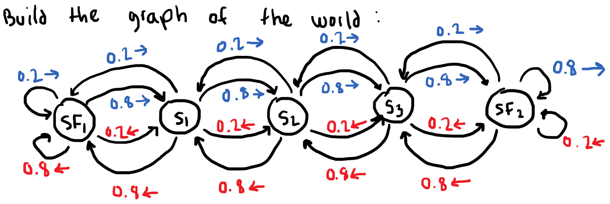
-10	0	-0.4	-0.4	10

Es decir, si el agente transiciona de s1 a s2 entonces recibe la recompensa -0.4 que esta definide en el estado s2.

El agente tiene la siguiente función de acción $f_\pi(s)$:

$$f_\pi(s) = \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_{F1} \\ s_{F2} \end{array} \begin{bmatrix} \rightarrow \\ \rightarrow \\ \rightarrow \\ \leftarrow \\ \rightarrow \end{bmatrix}$$

(a) Build the graph of the world:



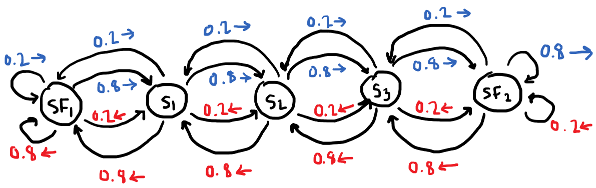
(b) Write the transition function $P_{MT}(s_f | s, a)$

$$P_{MT}(s_f | s, a) = \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ sF_1 \\ sF_2 \end{array} \begin{array}{cc} \leftarrow & \rightarrow \\ \end{array} \left[\begin{array}{cc} 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} 0.2 & 0.8 \\ 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \\ 0 & 0 \\ 0.8 & 0.2 \end{array}\right] \left[\begin{array}{cc} 0.8 & 0.2 \\ 0 & 0 \\ 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \\ 0.2 & 0.8 \end{array}\right]$$

c. Construya todas las trayectorias posibles a partir del estado inicial s1 dada la función de acción $f_\pi(s)$ que lleven a un estado final ya sea sF1 o sF2

(NOTA: Dado que es un número infinito de trayectorias solo escriba 10)

$$f_\pi(s) = \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_{F1} \\ s_{F2} \end{array} \begin{bmatrix} \rightarrow \\ \rightarrow \\ \rightarrow \\ \leftarrow \\ \rightarrow \end{bmatrix}$$



$\tau_1 = S_1, S_2, S_3, SF_2$

$\tau_2 = S_1, SF_1$

$\tau_3 = S_1, S_2, S_1, SF_1$

$\tau_4 = S_1, S_2, S_1, S_2, S_3, SF_2$

$\tau_5 = S_1, S_2, S_3, S_2, S_3, SF_2$

$\tau_6 = S_1, S_2, S_3, S_2, S_1, SF_1$

$\tau_7 = S_1, SF_1, S_1$

$\tau_8 = S_1, S_2, S_3, SF_2, S_3$

$\tau_9 = S_1, S_2, S_1, S_2, S_3, S_2, S_3, SF_2$

$\tau_{10} = S_1, S_2, S_1, S_2, S_3, S_2, S_1, SF_1$

d. Construya todas las trayectorias posibles a partir del estado s2 dada la función de acción $f_\pi(s)$ que lleven a un estado final ya sea sF1 o sF2
(NOTA: Dado que es un número infinito de trayectorias solo escriba 10)

τ_1 = S_2, S_3, SF_2
τ_2 = S_2, S_1, SF_1
τ_3 = S_2, S_3, S_2, S_1, SF_1
τ_4 = S_2, S_1, S_2, S_3, SF_2
τ_5 = $S_2, S_1, S_2, S_3, S_2, S_1, SF_1$
τ_6 = S_2, S_1, SF_1, S_1
τ_7 = S_2, S_3, SF_2, S_3
τ_8 = S_2, S_3, S_2, S_3, SF_2
τ_9 = $S_2, S_3, S_2, S_1, S_2, S_3, SF_2$
τ_{10} = S_2, S_1, S_2, S_1, SF_1

e. Construya todas las trayectorias posibles a partir del estado s3 dada la función de acción $f_\pi(s)$ que lleven a un estado final ya sea sF1 o sF2
(NOTA: Dado que es un número infinito de trayectorias solo escriba 10)

τ_1 = S_3, SF_2
τ_2 = S_3, S_2, S_1, SF_1
τ_3 = S_3, S_2, S_3, SF_2
τ_4 = $S_3, S_2, S_1, S_2, S_3, SF_2$
τ_5 = $S_3, S_2, S_3, S_2, S_1, SF_1$
τ_6 = S_3, SF_2, S_3
τ_7 = S_3, S_2, S_1, SF_1, S_1
τ_8 = $S_3, S_2, S_3, S_2, S_3, SF_2$
τ_9 = $S_3, S_2, S_1, S_2, S_1, SF_1$
τ_{10} = $S_3, S_2, S_3, S_2, S_3, S_2, S_3, SF_2$

d. Calcule la recompensa acumulada de cada posible trayectoria en los incisos c, d, e usando γ=0.7.

$\rightarrow s(0) = s_1$

τ_1 = 4.22
τ_2 = −10
τ_3 = −5.3
τ_4 = 1.67
τ_5 = 1.39
τ_6 = −3.28
τ_7 = −10
τ_8 = 4.08
τ_9 = 0.28
τ_{10} = −2

$\rightarrow s(0) = s_2$

τ_{11} = 6.6
τ_{12} = −7
τ_{13} = −4.11
τ_{14} = 2.95
τ_{15} = −2.29
τ_{16} = −7
τ_{17} = 6.4
τ_{18} = 2.55
τ_{19} = 0.77
τ_{20} = −3.71

$\rightarrow s(0) = s_3$

τ_{21} = 10
τ_{22} = −5.3
τ_{23} = 4.22
τ_{24} = 1.67
τ_{25} = −3.28
τ_{26} = 9.72
τ_{27} = −5.3
τ_{28} = 1.39
τ_{29} = −2.99
τ_{30} = 2.20

All these numbers were calculated using:
$r = r_1 + \gamma[f_{AR}(\tau)]$
Programmed in python

The code

```
r = [-10,0,-0.4,-0.4,10]

g = 0.7

for t in ts:
    acc = 0
    for i in range(1,len(t)):
        s = t[i]
        acc += (g**(i-1))*(r[s])
    print(f"T{ts.index(t)+1}:{acc}")
```