

**Ejercicio 1**

Para un proceso de Markov donde  $p_{ss'} = P[S_{t+1} = s' \mid S_t = s]$  son las probabilidades de transición, demostrar que la probabilidad de un episodio se puede escribir como el producto de las probabilidades de transición y la condición inicial, es decir:

$$P[S_0 = s_0, S_1 = s_1, \dots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{(t+1)}} p_{s_{(t-1)} s_t} \cdots p_{s_1 s_2} p_{s_0 s_1} p_{s_0},$$

donde  $s_n \in S$  y  $p_{s_0} P[S_0 = s_0]$ .

**Solución**

Sea,

S: Secuencia de estados

Se aplica la regla de la cadena para la probabilidad de una secuencia de eventos, para un proceso de Markov,

$$\begin{aligned} P[S_0 = s_0, S_1 = s_1, \dots, S_{t+1} = s_{t+1}] &= P[S_{t+1} = s_{t+1}, S_t = s_t, \dots, S_0 = s_0] \\ &= P[S_{t+1} = s_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, S_0 = s_0] P[S_t = s_t, \dots, S_0 = s_0] \end{aligned}$$

Por la propiedad de Markov, la probabilidad de transición de un estado a otro depende solo del estado actual, por lo que la probabilidad de transición de un estado a otro depende solo del estado actual, matemáticamente se expresa como,

$$\begin{aligned} P[S_{t+1} = s_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, \dots, S_0 = s_0] &= P[S_{t+1} = s_{t+1} \mid S_t = s_t] \\ &= p_{s_t s_{t+1}} \end{aligned}$$

Por lo tanto, la probabilidad de una secuencia de eventos se puede escribir como,

$$P[S_0 = s_0, S_1 = s_1, \dots, S_t = s_t, S_{t+1} = s_{t+1}] = P[S_{t+1} \mid S_t = s_t] P[S_t = s_t, \dots, S_0 = s_0]$$

Se aplica recursivamente la regla de la cadena para obtener la probabilidad conjunta en términos de las probabilidades de transición,

$$P[S_0 = s_0, S_1 = s_1, \dots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}} P[S_t = s_t, \dots, S_0 = s_0]$$

Se aplica de nuevo la regla de la cadena para obtener la probabilidad conjunta en términos de las probabilidades de transición,

$$P[S_t = s_t, S_{t-1} = s_{t-1}, \dots, S_0 = s_0] = p_{s_{t-1} s_t} P[S_{t-1} = s_{t-1}, \dots, S_0 = s_0]$$

Tras aplicar la propiedad de Markov se llegó a que,

$$P[S_1 = s_1, S_0 = s_0] = p_{s_0 s_1} P[S_0 = s_0]$$

La probabilidad conjunta de la secuencia completa de estados es,

$$P[S_0 = s_0, S_1 = s_1, \dots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}} p_{s_{t-1} s_t} \cdots p_{s_1 s_2} p_{s_0 s_1} p_{s_0} P[S_0 = s_0]$$

Si se define  $p_{s_0} = P[S_0 = s_0]$ , la expresión se simplifica a,

$$P[S_0 = s_0, S_1 = s_1, \dots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}} p_{s_{t-1} s_t} \cdots p_{s_1 s_2} p_{s_0 s_1} p_{s_0}$$

Entonces, la probabilidad de un episodio se puede escribir como el producto de las probabilidades de transición y la condición inicial  $p_{s_0}$  [1].

## Ejercicio 2

Para el proceso de Markov, de la figura 1, calcular la probabilidad de cada uno de los siguientes episodios condicionados al estado inicial  $C1$ :

1. C1 C2 C3 Pass Sleep
2. C1 FB FB C1 C2 Sleep
3. C1 C2 C3 Pub C2 C3 Pass Sleep
4. C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

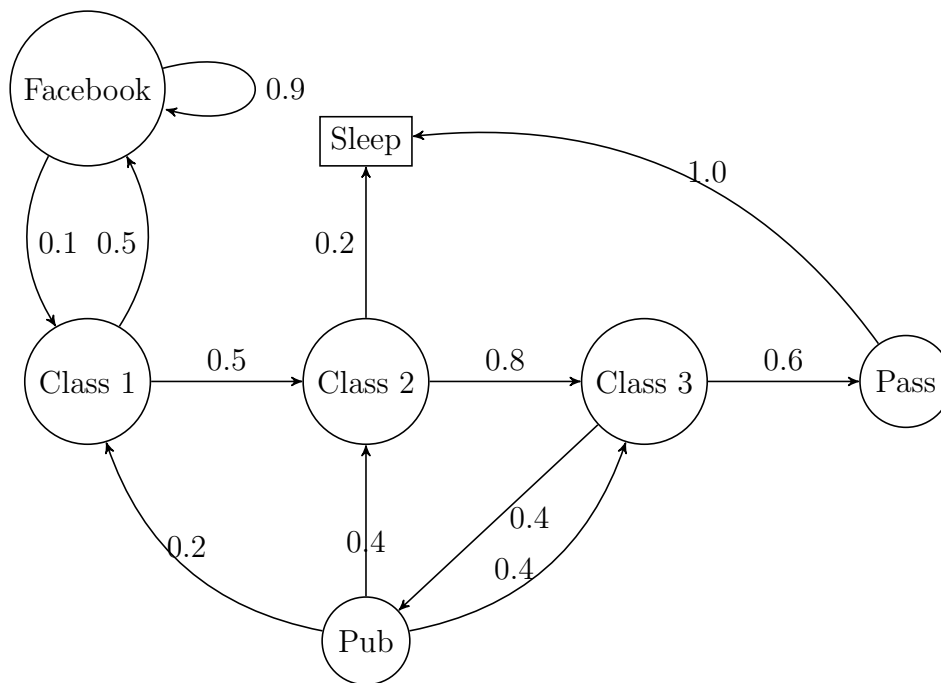


Figura 1: Grafo de transición de estados

## Solución

$C1 \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep$

Se tienen las probabilidades de transición de un estado a otro:

- $P(C1 \rightarrow C2) = 0.5$
- $P(C2 \rightarrow C3) = 0.8$
- $P(C3 \rightarrow Pass) = 0.6$
- $P(Pass \rightarrow Sleep) = 1.0$

Por lo tanto, la probabilidad de que el proceso de Markov pase por los estados  $C1 \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep$  es:

$$\begin{aligned} P(C1 \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep) &= P(C1 \rightarrow C2) \cdot P(C2 \rightarrow C3) \cdot P(C3 \rightarrow Pass) \cdot P(Pass \rightarrow Sleep) \\ &= 0.5 \cdot 0.8 \cdot 0.6 \cdot 1.0 \\ &= 0.24 \end{aligned}$$

$$C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow Sleep$$

Se tienen las probabilidades de transición de un estado a otro:

- $P(C1 \rightarrow FB) = 0.5$
- $P(FB \rightarrow FB) = 0.9$
- $P(FB \rightarrow C1) = 0.1$
- $P(C1 \rightarrow C2) = 0.5$
- $P(C2 \rightarrow Sleep) = 0.2$

Por lo tanto, la probabilidad de que el proceso de Markov pase por los estados  $C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow Sleep$  es:

$$\begin{aligned} P(C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow Sleep) &= P(C1 \rightarrow FB) \cdot P(FB \rightarrow FB) \cdot \dots \\ &= 0.5 \cdot 0.9 \cdot 0.1 \cdot 0.5 \cdot 0.2 \\ &= 0.0045 \end{aligned}$$

$$C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep$$

Se tienen las probabilidades de transición de un estado a otro:

- $P(C1 \rightarrow C2) = 0.5$
- $P(C2 \rightarrow C3) = 0.8$
- $P(C3 \rightarrow Pub) = 0.4$
- $P(Pub \rightarrow C2) = 0.4$
- $P(C2 \rightarrow C3) = 0.8$
- $P(C3 \rightarrow Pass) = 0.6$
- $P(Pass \rightarrow Sleep) = 1.0$

Por lo tanto, la probabilidad de que el proceso de Markov pase por los estados  $C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep$  es:

$$\begin{aligned} P(C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C2 \rightarrow C3 \rightarrow Pass \rightarrow Sleep) &= P(C1 \rightarrow C2) \cdot P(C2 \rightarrow C3) \cdot \dots \\ &= 0.5 \cdot 0.8 \cdot 0.4 \cdot 0.4 \cdot 0.8 \cdot 0.6 \cdot 1.0 \\ &= 0.03072 \end{aligned}$$

$C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C1 \rightarrow FB \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C2 \rightarrow Sleep$

Se tienen las probabilidades de transición de un estado a otro:

- $P(C1 \rightarrow FB) = 0.5$
- $P(FB \rightarrow FB) = 0.9$
- $P(FB \rightarrow C1) = 0.1$
- $P(C1 \rightarrow C2) = 0.5$
- $P(C2 \rightarrow C3) = 0.8$
- $P(C3 \rightarrow Pub) = 0.4$
- $P(Pub \rightarrow C1) = 0.2$
- $P(C1 \rightarrow FB) = 0.5$
- $P(FB \rightarrow FB) = 0.9$
- $P(FB \rightarrow FB) = 0.9$
- $P(FB \rightarrow C1) = 0.1$
- $P(C1 \rightarrow C2) = 0.5$
- $P(C2 \rightarrow C3) = 0.8$
- $P(C3 \rightarrow Pub) = 0.4$
- $P(Pub \rightarrow C2) = 0.4$
- $P(C2 \rightarrow Sleep) = 0.2$

Por lo tanto, la probabilidad de que el proceso de Markov pase por los estados  $C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C1 \rightarrow FB \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \rightarrow C3 \rightarrow Pub \rightarrow C2 \rightarrow Sleep$  es:

$$\begin{aligned}
 P(C1 \rightarrow FB \rightarrow FB \rightarrow C1 \rightarrow C2 \dots) &= P(C1 \rightarrow FB) \cdot P(FB \rightarrow FB) \cdot \dots \\
 &= 0.5 \cdot 0.9 \cdot \dots \cdot 0.2 \\
 &= (0.1)^2 \cdot (0.2)^2 \cdot (0.4)^3 \cdot (0.5)^4 \cdot (0.8)^2 \cdot (0.9)^3 \\
 &= 0.00000075 \\
 &= 7.5 \cdot 10^{-7}
 \end{aligned}$$

---

### **Ejercicio 3**

Demostrar que si la recompensa es constante  $R_t = R \forall t$  y el factor de descuento  $\gamma < 1$ , entonces,

$$G_t = \frac{R}{1 - \gamma}$$


---

**Solución**

El retorno  $G_t$  es la suma de las recompensas futuras descontadas a partir del tiempo  $t$ , es decir,

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

Dado que la recompensa es constante,  $R_t = R \forall t$ , entonces,  $R_{t+1} = R$ ,  $R_{t+2} = R$ ,  $\dots$ , por lo que,

$$G_t = R + \gamma R + \gamma^2 R + \gamma^3 R + \dots$$

Tras factorizar toma la forma,

$$G_t = R(1 + \gamma + \gamma^2 + \dots)$$

La serie geométrica  $1 + \gamma + \gamma^2 + \dots$  es infinita. Sin embargo, dado que  $\gamma < 1$ , la serie converge a,

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1 - \gamma}$$

Por lo tanto,

$$1 + \gamma + \gamma^2 + \dots = \frac{1}{1 - \gamma}$$

Que al remplazar en la expresión de  $G_t$  se obtiene,

$$G_t = R \left( \frac{1}{1 - \gamma} \right)$$

$$G_t = \frac{R}{1 - \gamma}$$

**Ejercicio 4**

Para el proceso de Recompensas Markoviano (MRP) de la figura 2, el retorno  $G_0$  de cada uno de los siguientes episodios con estado inicial C1 y  $\gamma = 0.5$ :

1. C1 C2 C3 Pass Sleep
2. C1 FB FB C1 C2 Sleep
3. C1 C2 C3 Pub C2 C3 Pass Sleep
4. C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

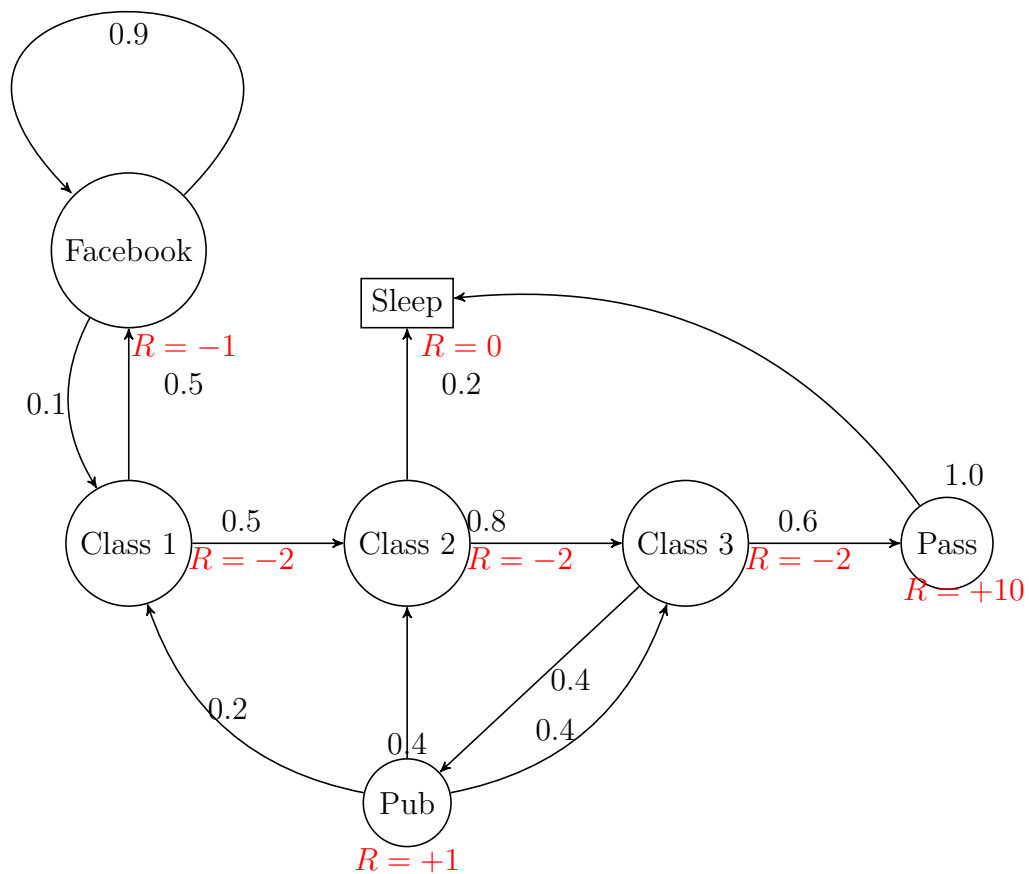


Figura 2: Grafo de transición de estados

**Solución**

Sea,

$\gamma \in [0, 1]$ : Factor de descuento.  $\gamma = 0.5$   $R_t \in \mathbb{R}$ : Recompensa.

$G_t$ : Retorno, es la acumulación de recompensas desde el instante  $t + 1$ .

El retorno se calcula como la suma de las recompensas futuras descontadas a partir del tiempo  $t$ , es decir,

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

**Episodio 1:** C1 C2 C3 Pass Sleep

Se calcula el retorno,

$$\begin{aligned} G_0 &= -2 + 0.5(-2) + 0.5^2(10) + 0.5^3(0) \\ &= -2 - 1 + 2.5 + 0 \\ &= -0.5 \end{aligned}$$

**Episodio 2:** C1 FB FB C1 C2 Sleep

Se calcula el retorno,

$$\begin{aligned}
G_0 &= -1 + 0.5(-1) + 0.5^2(-2) + 0.5^3(-2) + 0.5^4(0) \\
&= -1 - 0.5 - 0.5 - 0.25 + 0 \\
&= -2.25
\end{aligned}$$

**Episodio 3:** C1 C2 C3 Pub C2 C3 Pass Sleep

Se calcula el retorno,

$$\begin{aligned}
G_0 &= -2 + 0.5(-2) + 0.5^2(1) + 0.5^3(-2) + 0.5^4(-2) + 0.5^5(10) + 0.5^7(0) \\
&= -2 - 1 + 0.25 - 0.25 - 0.125 + 0.3125 + 0 \\
&= -2.8125
\end{aligned}$$

**Episodio 4:** C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

Se calcula el retorno,

$$\begin{aligned}
G_0 &= -1 + 0.5(-1) + 0.5^2(-2) + 0.5^3(-2) + 0.5^4(-2) + 0.5^5(1) + 0.5^6(-2) + 0.5^7(-1) + \\
&\quad 0.5^8(-1) + 0.5^9(-1) + 0.5^{10}(-2) + 0.5^{11}(-2) + 0.5^{12}(-2) + 0.5^{13}(1) + 0.5^{14}(-2) + 0.5^{15}(0) \\
&= -1 - 0.5 - 0.5 - 0.25 - 0.125 + 0.03125 - 0.03125 - 0.0078125 - 0.0039063 - 0.0019531 - \\
&\quad 0.0019531 - 0.0009766 - 0.0004883 + 0.0001221 - 0.0001221 + 0 \\
&= -2.3921
\end{aligned}$$

**Ejercicio 5**

Demostrar que si  $\gamma = 0$  y  $R_t = R(S_t)$  (la recompensa depende del estado), entonces

$$v(s) = \sum_{s' \in S} p(S_{t+1} = s' | S_t = s) R(S_{t+1})$$

**Solución**

Sea,

$v(s)$ : Valor esperado del retorno a partir del estado  $s$ .

$R(S_t)$ : Recompensa en el estado  $S_t$ .

$G_t$ : Retorno a partir del tiempo  $t$  y  $S_t = s$ .

Por definición, el valor esperado del retorno a partir del estado  $s$  es,

$$v(s) = E[G_t | S_t = s]$$

Donde,  $G_t$  está dado por,

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

Si  $\gamma = 0$  y  $R_t = R(S_t)$ , entonces el agente no considera el futuro y la recompensa depende del estado actual. Por lo tanto, el retorno  $G_t$  se reduce a,

$$G_t = R_t = R(S_t)$$

Por otro lado,  $v(s)$  se puede expresar en términos del siguiente estado  $S_{t+1}$  y la recompensa  $R(S_{t+1})$  como,

$$v(s) = E[G_t | S_t = s] = E[R(S_{t+1}) | S_t = s]$$

Por lo tanto, el valor esperado del retorno a partir del estado  $s$  es la suma de las recompensas esperadas en el siguiente estado ponderadas por la probabilidad de transición al siguiente estado, es decir,

$$v(s) = \sum_{s' \in S} p(S_{t+1} = s' | S_t = s) R(S_{t+1})$$

En otras palabras, el valor de un estado es la expectativa de la recompensa inmediata en el siguiente estado ponderada por la probabilidad de transición al siguiente estado [1].

### **Ejercicio 6**

Demostrar la ecuación de *Bellman* para MRPs con  $N$  estados

$$v = r + \gamma P v,$$

donde  $v \in \mathbb{R}^N$  es el vector de valores de estados,  $r \in \mathbb{R}^N$  es el vector de recompensas medias partiendo de cada estado,  $P \in \mathbb{R}^{N \times N}$  es la matriz de transiciones y  $\gamma$  es el factor de descuento.

### **Solución**

Sea,

$v \in \mathbb{R}^N$ : Vector de valores de estados.

$r \in \mathbb{R}^N$ : Vector de recompensas medias partiendo de cada estado.

$P \in \mathbb{R}^{N \times N}$ : Matriz de transiciones.

$\gamma \in [0, 1)$ : Factor de descuento.

$i$ : Índice del estado.

Partiendo de la ecuación de *Bellman* para un estado  $i$ ,

$$v_i = r_i + \gamma \sum_{j=1}^N P_{ij} v_j$$

Es decir, el valor de un estado  $i$  es la recompensa inmediata  $r_i$  más el valor esperado de los estados siguientes ponderados por la probabilidad de transición desde  $i$  a  $j$ ,  $P_{ij}$ , siendo la matriz de transiciones, y descontados por el factor  $\gamma$ . El primer término se puede expresar en forma matricial como,

$$r$$

El segundo término es la suma ponderada de los valores de los estados siguientes.

$$\gamma \sum_{j=1}^N P_{ij} v_j$$

Es decir, el factor de descuento  $\gamma$  resta valor a los estados futuros [1]. Este término se puede expresar en forma matricial como,

$$\gamma P v$$

Al sumar ambos términos se obtiene la ecuación de *Bellman* para MRPs con  $N$  estados,

$$v = r + \gamma P v$$



**Ejercicio 7**

[Programación] Resuelva la ecuación de Bellman para el MRP del Ejercicio 4 por los siguientes métodos para  $\gamma \in \{0, 0.5, 0.9\}$ :

1. Iterando  $v_{n+1} = r + \gamma P v_n$
2. Usando algún método iterativo para resolver sistemas lineales, por ejemplo usando eliminación Gaussiana (vía `numpy.linalg.solve()`)
3. Calculando la inversa de la matriz  $I - \gamma P$

**Solución**

Se tiene la matriz de transición  $P$ , mostrada en la Tabla 1.

Tabla 1: Matriz de transición

	FB	C1	C2	C3	Pub	Pass	Sleep
FB	0.9	0.1	0	0	0	0	0
C1	0.5	0	0.5	0	0	0	0
C2	0	0	0	0.8	0	0	0.2
C3	0	0	0	0	0.4	0.6	0
Pub	0	0.2	0.4	0.4	0	0	0
Pass	0	0	0	0	0	0	1
Sleep	0	0	0	0	0	0	0

El vector de recompensas  $r$ , se muestra en la Tabla 2.

Tabla 2: Vector de recompensas

	$r_i$
FB	-1
C1	-2
C2	-2
C3	-2
Pub	1
Pass	10
Sleep	0

**Método 1: Iterando**  $v_{n+1} = r + \gamma P v_n$  Se iteró el cálculo de  $v_{n+1}$  hasta que la diferencia entre  $v_{n+1}$  y  $v_n$  sea menor a  $10^{-6}$ . A continuación se muestran los valores de  $v_{n+1}$  para  $\gamma \in 0, 0.5, 0.9$  en la Tabla 3<sup>1</sup>.

<sup>1</sup>El cálculo se hizo con Excel, la hoja de cálculo utilizada se encuentra disponible en el repositorio de GitHub.

Tabla 3: Método iterativo - Valores de  $v_{n+1}$  para  $\gamma \in 0, 0.5, 0.9$ 

	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 0.9$
FB	-1.000	-2.083	-7.6438
C1	-2.000	-2.908	-5.013
C2	-2.000	-1.550	0.943
C3	-2.000	1.125	4.087
Pub	1.000	0.624	1.908
Pass	10.000	10.000	10.000
Sleep	0.000	0.000	0.000

**Método 2: Usando eliminación Gaussiana** Se realizó el cálculo de  $v$  para  $\gamma \in 0, 0.5, 0.9$  usando eliminación Gaussiana, con ayuda de la función `numpy.linalg.solve()`. A continuación se muestran los valores de  $v$  en la Tabla 4<sup>2</sup>.

Tabla 4: Método de eliminación Gaussiana - Valores de  $v$  para  $\gamma \in 0, 0.5, 0.9$ 

	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 0.9$
FB	-1.000	-2.082	-7.637
C1	-2.000	-2.908	-5.013
C2	-2.000	-1.550	0.943
C3	-2.000	1.125	4.087
Pub	1.000	0.624	1.908
Pass	10.000	10.000	10.000
Sleep	0.000	0.000	0.000

**Método 3: Calculando la inversa de la matriz  $I - \gamma P$**  Para calcular la inversa de la matriz  $I - \gamma P$ , se utilizó la función `numpy.linalg.inv()`. A continuación se muestran los valores de  $v$  en la Tabla 5<sup>3</sup>.

Tabla 5: Método de inversa - Valores de  $v$  para  $\gamma \in 0, 0.5, 0.9$ 

	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 0.9$
FB	-1.000	-2.082	-7.6438
C1	-2.000	-2.908	-5.013
C2	-2.000	-1.550	0.943
C3	-2.000	1.125	4.087
Pub	1.000	0.624	1.908
Pass	10.000	10.000	10.000
Sleep	0.000	0.000	0.000

### Conclusión

No se observan diferencias significativas en los resultados obtenidos por los tres métodos, por lo que se puede concluir que los tres métodos son válidos para resolver la ecuación de ‘Bellman’.

Por otro lado, se compararon los tiempos de cómputo para los métodos de Eliminación Gaussiana (Solver) e Inversa, los resultados se muestran en la Figura 3.

<sup>2</sup>El cálculo se hizo con NumPy, la función `solve_belman(...)` se encuentra disponible en el repositorio de GitHub.

<sup>3</sup>El cálculo se hizo con NumPy, la función `solve_belman_inverse(...)` se encuentra disponible en el repositorio de GitHub.

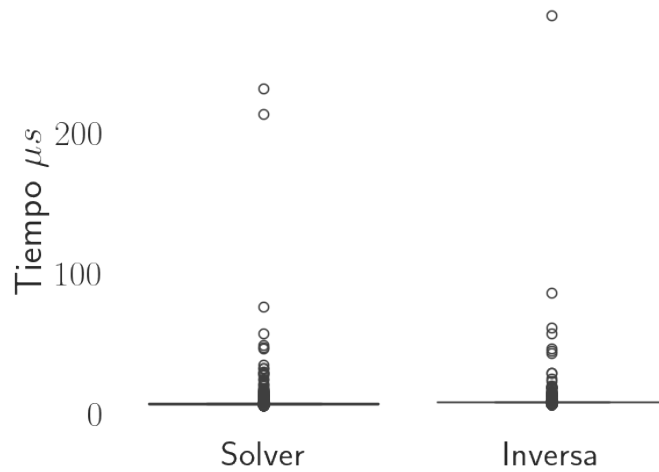


Figura 3: Comparación de tiempos de cómputo entre Solver e Inversa

Los resultados se muestran en la Tabla 6.

Tabla 6: Tiempo de ejecución de los métodos de solución

	Solver	Inversa
Media	7.827	8.340
Desviación estándar	7.467	6.894
Mínimo	5.722	6.676
Percentil 25	6.914	7.868
Percentil 50	6.914	7.868
Percentil 75	7.153	8.106
Máximo	230.789	283.241

De la Tabla 6, se observa que el método de eliminación Gaussiana (Solver) es el más rápido en términos de tiempo de ejecución, en comparación con el método de inversa. Este resultado es esperado, ya que el método de inversa requiere calcular la inversa de la matriz  $I - \gamma P$ , lo cual es computacionalmente costoso.

Por otro lado, se recomienda tener en cuenta este resultado al momento de seleccionar el método de solución a utilizar, ya que el tiempo de cómputo puede ser un factor importante en la elección del método. Si bien el caso en particular no presenta diferencias significativas, en problemas más grandes, la diferencia en tiempo de cómputo puede ser considerable.

### **Ejercicio 8**

Demostrar que, dada una política estocástica  $\pi(a \mid s)$ , la función de valor de estado puede escribirse como

$$v_{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) q_{\pi}(s, a)$$

### **Solución**

Sea,

$v_{\pi}(s)$ : Valor esperado del retorno a partir del estado  $s$ .

$q_{\pi}(s, a)$ : Función de valor de acción bajo una política  $\pi$ , es el valor esperado del retorno a partir del estado  $s$  y la acción  $a$ .

$\pi(a | s)$ : Política estocástica, asigna la probabilidad de seleccionar la acción  $a$  en el estado  $s$ .

$R_t$ : Recompensa total acumulada desde  $t$ .

Partiendo del valor esperado de estado  $s$ , bajo la política  $\pi$ ,

$$v_\pi(s) = E_\pi[R_t | S_t = s]$$

El valor esperado, se puede expresar considerando todas las posibles acciones que se pueden tomar en el estado  $s$ , ponderadas por la probabilidad de seleccionar cada acción, es decir,

$$v_\pi(s) = \sum_{a \in \mathcal{A}} P(A_t = a | S_t = s) E[R_t | S_t = s, A_t = a]$$

Se incorpora la política estocástica,

$$P(A_t = a | S_t = s) = \pi(a | s)$$

Se incorpora la función de valor de acción,

$$E[R_t | S_t = s, A_t = a] = q_\pi(s, a)$$

Al remplazar ambas en la expresión inicial, se obtiene,

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a | s) q_\pi(s, a)$$

Entonces, se tiene una expresión que relaciona la función de estado y acción [1].

### **Ejercicio 9**

Demostrar que, dada una política estocástica  $\pi(a | s)$ , la función de valor de estado puede escribirse como

$$q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma v_\pi(s')]$$

### **Solución**

Sea,

$\pi(a | s)$ : Política estocástica, es una función que asigna a cada estado  $s$  una distribución de probabilidad sobre el conjunto de acciones  $a$ , es decir,  $\pi(a | s) = P[A_t = a | S_t = s]$ .

$q_\pi(s, a)$ : Función de valor de acción bajo una política  $\pi$ , es el valor esperado del retorno a partir del estado  $s$  y la acción  $a$ .

$p(s', r | s, a)$ : Función de transición conjunta de estado y recompensas, es la probabilidad de que se obtenga el estado  $s'$  y la recompensa  $r$  al tomar la acción  $a$  en el estado  $s$ .

$v_\pi(s)$ : Valor de estado bajo la política  $\pi$ , es el valor esperado del retorno a partir del estado  $s$ .

La función de valor esperado de acción  $q_\pi(s, a)$ , se puede expresar como,

$$q_\pi(s, a) = E_\pi[R_t | S_t = s, A_t = a]$$

Es decir,  $q_\pi(s, a)$  es la suma de la recompensa inmediata  $r$  y el valor esperado del retorno a partir del estado  $s'$ . Al tomar la acción  $a$  en el estado  $s$  se obtiene la recompensa inmediata  $r$  y se transiciona al estado  $s'$ . Si se tiene en cuenta las recompensas futuras y todas las posibles transiciones y recompensas, la función de valor de acción se puede expresar como,

$$q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) E[R_t | S_t = s', A_t]$$

Las recompensas futuras se puede dividir en dos partes, que incluyen la recompensa inmediata  $r$  y el valor esperado del retorno a partir del estado  $s'$ , ponderado por el factor de descuento  $\gamma$ ,

$$E[R_t | S_t = s', A_t] = r + \gamma v_\pi(s')$$

Al remplazar la expresión anterior en la función de valor de acción, se obtiene,

$$q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma v_\pi(s')]$$

Entonces, el valor esperado total es igual al valor esperado de las recompensas inmediatas y futuras ponderadas por el factor de descuento  $\gamma$  [1].

### **Ejercicio 10**

Demostrar que la función de valor de estado óptima es

$$v_*(s) = \max_a q_*(s, a) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v_*(s')]$$

### **Solución**

Sea,

$v_*(s)$ : Valor de estado óptimo, es el valor esperado del retorno a partir del estado  $s$  bajo una política óptima.

$q_*(s, a)$ : Función de valor de acción óptima, es el valor esperado del retorno a partir del estado  $s$  y la acción  $a$  bajo una política óptima.

$p(s', r | s, a)$ : Función de transición conjunta de estado y recompensas, es la probabilidad de que se obtenga el estado  $s'$  y la recompensa  $r$  al tomar la acción  $a$  en el estado  $s$ .

La función de valor de estado óptima se puede expresar como,

$$v_*(s) = \max_a q_*(s, a)$$

Es decir, si se conoce la función de valor de acción óptima, la función de valor de estado óptima es el máximo valor de la función de valor de acción óptima. La función de valor de acción óptima se puede expresar como,

$$\begin{aligned} q_*(s, a) &= E[R_t | S_t = s, A_t = a] \\ &= \sum_{s', r} p(s', r | s, a) E[R_t | S_t = s', A_t = a] \\ &= \sum_{s', r} p(s', r | s, a) [r + \gamma v_*(s')] \end{aligned}$$

La función de valor óptima se puede escribir como la suma de la recompensa inmediata  $r$  y el valor esperado del retorno a partir del estado  $s'$ , ponderado por el factor de descuento  $\gamma$  [1]. Al remplazar la expresión de la función de valor de acción óptima en la función de valor de estado óptima, se obtiene,

$$v_*(s) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v_*(s')]$$

### **Ejercicio 11**

Demostrar que la función de valor estado-acción verifica la siguiente ecuación:

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) \left[ r + \gamma \sum_{a'} q_{\pi}(s', a') \pi(a' | s') \right]$$

### Solución

Sea,

$q_{\pi}(s, a)$ : Función de valor de acción bajo una política  $\pi$ , es el valor esperado del retorno a partir del estado  $s$  y la acción  $a$ .

$p(s', r | s, a)$ : Función de transición conjunta de estado y recompensas, es la probabilidad de que se obtenga el estado  $s'$  y la recompensa  $r$  al tomar la acción  $a$  en el estado  $s$ .

$\pi(a' | s')$ : Política estocástica, asigna la probabilidad de seleccionar la acción  $a'$  en el estado  $s'$ .

La función de valor de acción  $q_{\pi}(s, a)$ , se puede expresar como,

$$\begin{aligned} q_{\pi}(s, a) &= E[R_t | S_t = s, A_t = a] \\ &= \sum_{s', r} p(s', r | s, a) E[R_t | S_t = s', A_t = a] \end{aligned}$$

Se incluye la expectativa total, que considera la recompensa inmediata  $r$  y el valor esperado del retorno a partir del estado  $s'$ , ponderado por el factor de descuento  $\gamma$  [1],

$$E[R_t | S_t = s', A_t = a] = r + \gamma \sum_{a'} q_{\pi}(s', a') \pi(a' | s')$$

Al sustituir en la función de valor de acción, se obtiene,

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) \left[ r + \gamma \sum_{a'} q_{\pi}(s', a') \pi(a' | s') \right]$$

### Ejercicio 12

Demostrar que la función de valor estado-acción óptima es:

$$q_*(s, a) = \sum_{s', r} p(s', r | s, a) \left[ r + \gamma \max_{a'} q_*(s', a') \right]$$

### Solución

Sea,

$q_*(s, a)$ : Función de valor de acción óptima, es el valor esperado del retorno a partir del estado  $s$  y la acción  $a$  bajo una política óptima.

$p(s', r | s, a)$ : Función de transición conjunta de estado y recompensas, es la probabilidad de que se obtenga el estado  $s'$  y la recompensa  $r$  al tomar la acción  $a$  en el estado  $s$ .

$v_*(s')$ : Valor de estado óptimo, es el valor esperado del retorno a partir del estado  $s'$  bajo una política óptima.

La función de valor de estado acción óptima  $q_*(s, a)$ , se puede expresar como,

$$q_*(s, a) = E[R_t | S_t = s, A_t = a]$$

Es decir,  $q_*(s, a)$  es la suma de la recompensa inmediata  $r$  y el valor esperado del retorno a partir del estado  $s'$  [1]. El valor óptimo de la función de valor de acción se puede expresar como,

$$q_*(s, a) = \sum_{s', r} p(s', r | s, a) E[R_t | S_t = s', A_t = a]$$

Donde,

$$E[R_t | S_t = s', A_t = a] = r + \gamma v_*(s') = \max_{a'} q_*(s', a')$$

Al sustituir en la función de valor de acción óptima, se obtiene,

$$q_*(s, a) = \sum_{s', r} p(s', r | s, a) \left[ r + \gamma \max_{a'} q_*(s', a') \right]$$

---

## Referencias

- [1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.