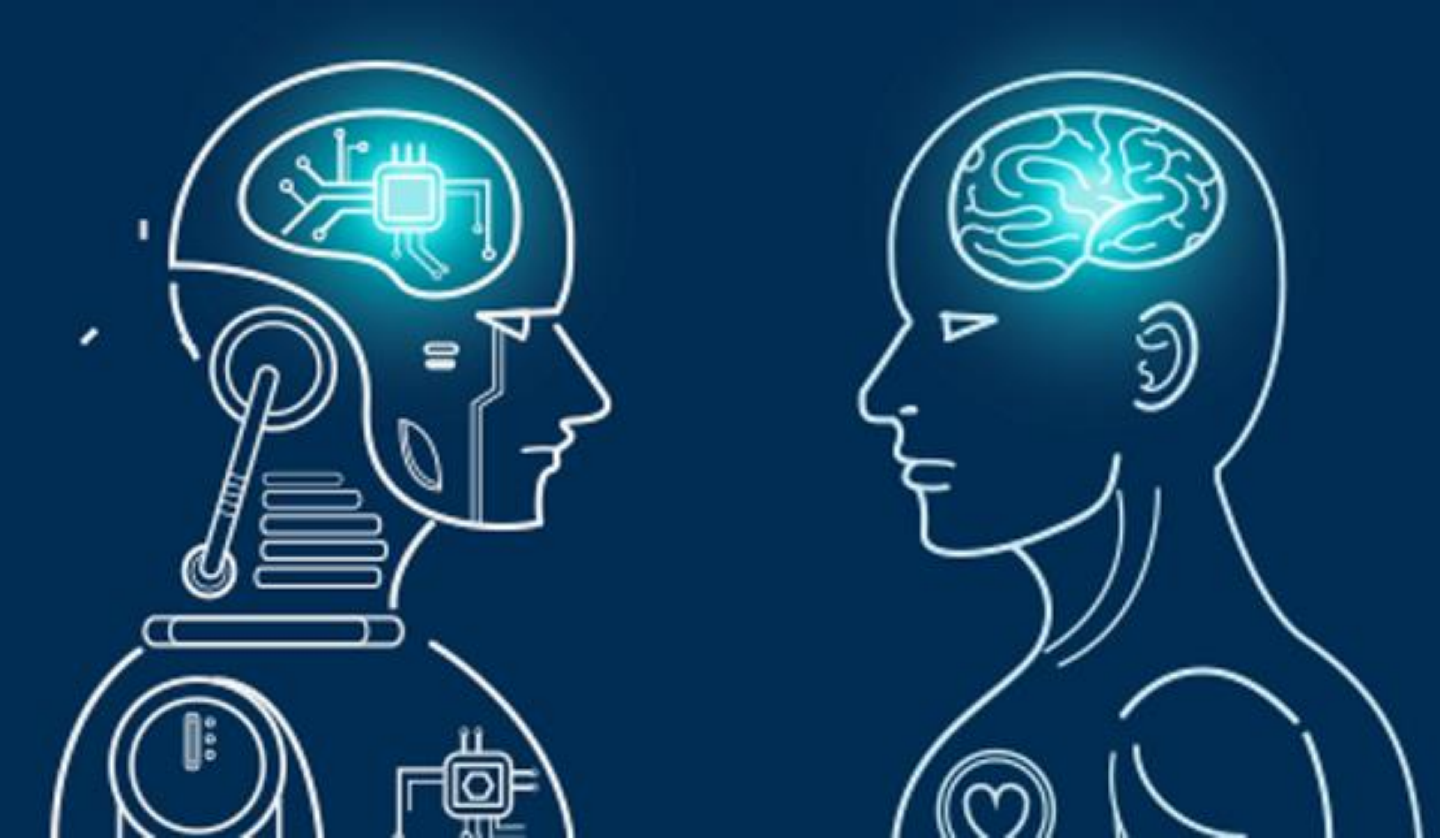
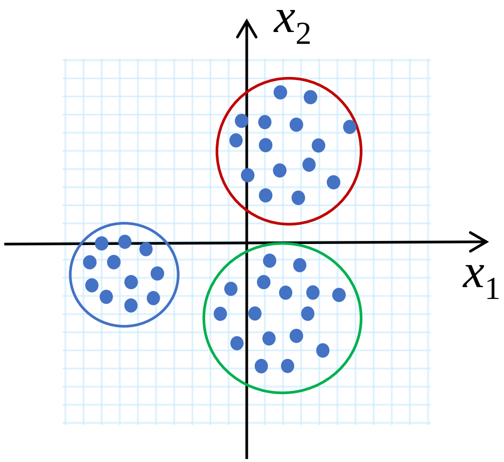


Aprendizaje Por Refuerzo

Aprendizaje Artificial

tipo

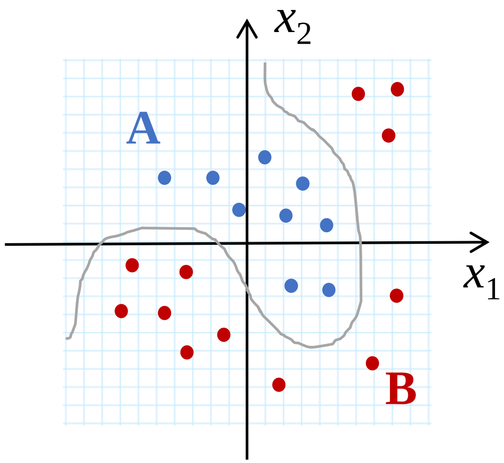
NO Supervisado



Datos	NO Etiquetados NO Estructurados	Retroalimentación	X
Función	Agrupación ↓Dimensionalidad		
Objetivo	Encontrar estructura en los Datos		

tipo

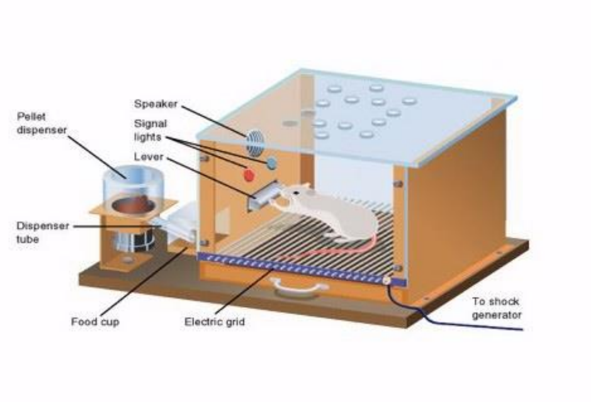
Supervisado



Datos	Etiquetados Estructurados	Retroalimentación	✓
Función	Clasificación Regresión		
Objetivo	Predicción		

tipo

Por Refuerzo



Datos	Estados Recompensa	Retroalimentación	✓
Función	Asociar acciones a estados		
Objetivo	Encontrar las mejores Acciones		

Aprendizaje Artificial

tipo

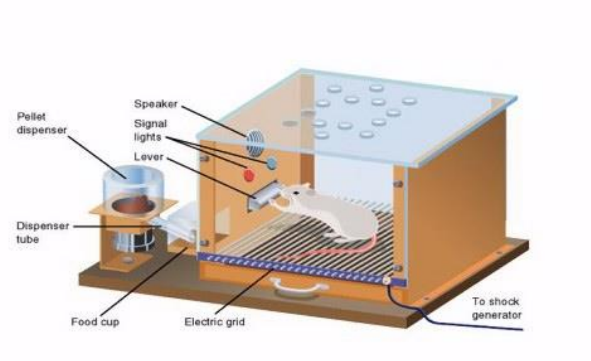
NO Supervisado

tipo

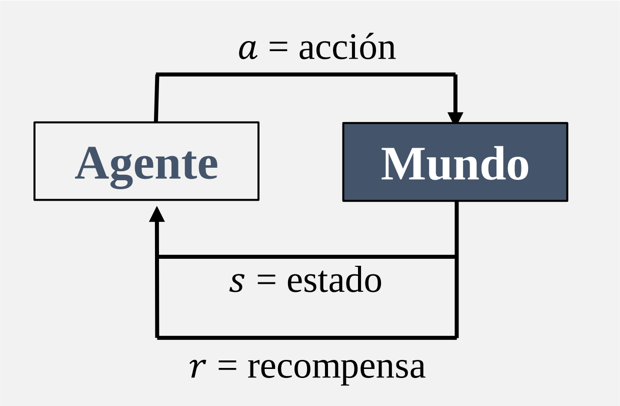
Supervisado

tipo

Por Refuerzo



estructura

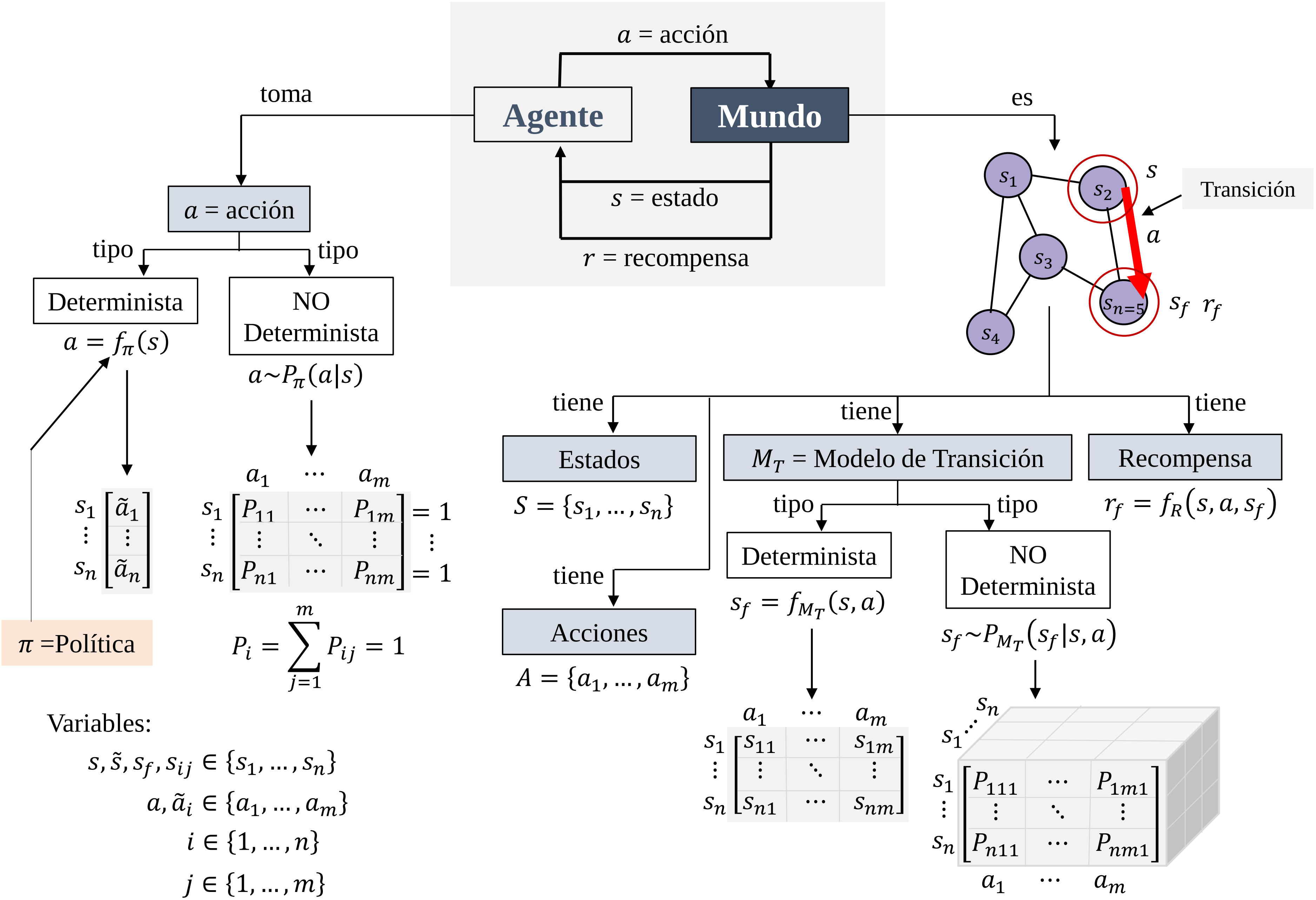

a = acción

Agente

Mundo

s = estado

r = recompensa



a = acción

Agente

Mundo

s = estado

r = recompensa

toma

a = acción

tipo

Determinista

$a = f_{\pi}(s)$

$$\begin{array}{c|c} s_1 & \tilde{a}_1 \\ \vdots & \vdots \\ s_n & \tilde{a}_n \end{array}$$

π =Política

tipo

NO Determinista

$a \sim P_{\pi}(a|s)$

$$\begin{array}{cccc} & a_1 & \cdots & a_m \\ s_1 & P_{11} & \cdots & P_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & P_{n1} & \cdots & P_{nm} \end{array} \begin{array}{c} \\ = 1 \\ \vdots \\ = 1 \end{array}$$

$$P_i = \sum_{j=1}^{m} P_{ij} = 1$$

es

Transición

tiene

Estados

$S = \{s_1, \ldots, s_n\}$

tiene

Acciones

$A = \{a_1, \ldots, a_m\}$

tiene

M_T = Modelo de Transición

tipo

Determinista

$s_f = f_{M_T}(s, a)$

$$\begin{array}{cccc} & a_1 & \cdots & a_m \\ s_1 & s_{11} & \cdots & s_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & s_{n1} & \cdots & s_{nm} \end{array}$$

tipo

NO Determinista

$s_f \sim P_{M_T}(s_f | s, a)$

$$\begin{array}{cccc} & & & \\ s_1 & P_{111} & \cdots & P_{1m1} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & P_{n11} & \cdots & P_{nm1} \\ & a_1 & \cdots & a_m \end{array}$$

tiene

Recompensa

$r_f = f_R(s, a, s_f)$

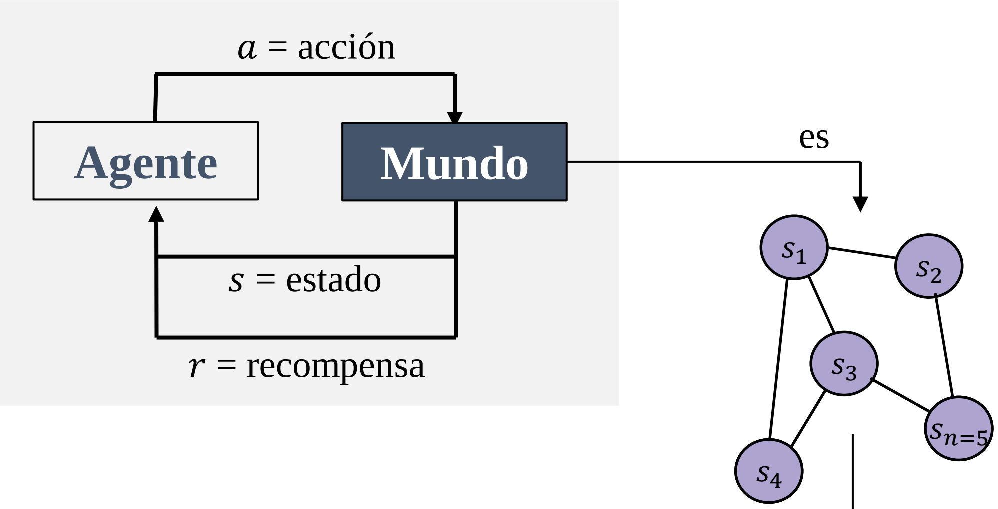
Variables:

$$s, \tilde{s}, s_f, s_{ij} \in \{s_1, \ldots, s_n\}$$

$$a, \tilde{a}_i \in \{a_1, \ldots, a_m\}$$

$$i \in \{1, \ldots, n\}$$

$$j \in \{1, \ldots, m\}$$



a = acción

Agente

Mundo

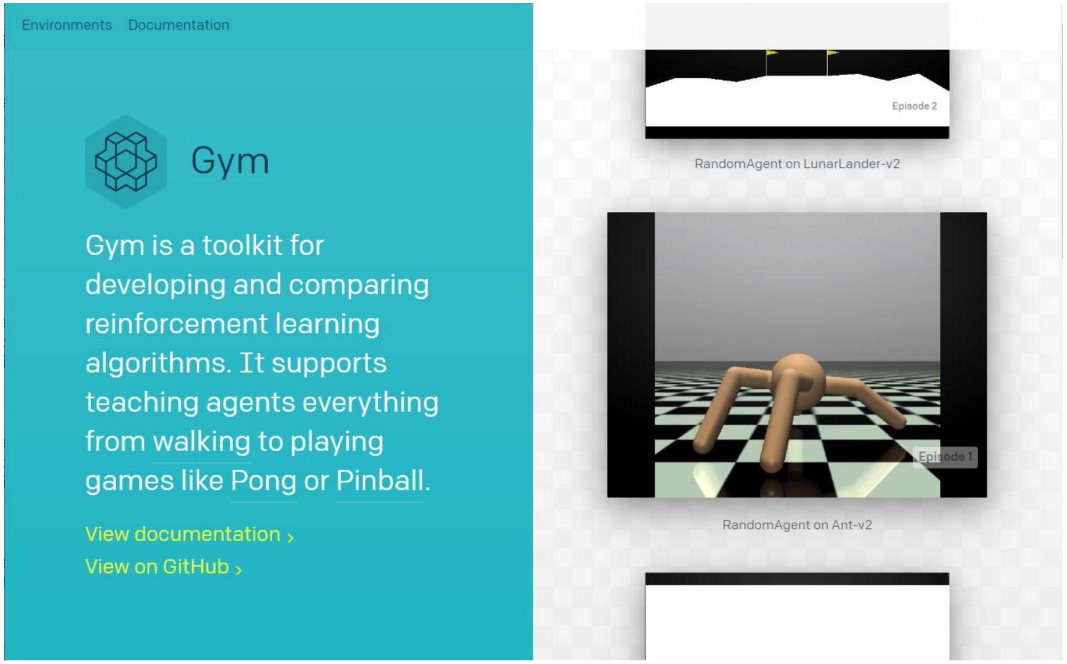
s = estado

r = recompensa

es

s_1 s_2 s_3 s_4 $s_{n=5}$

OpenIA Gym: https://gym.openai.com



Gym

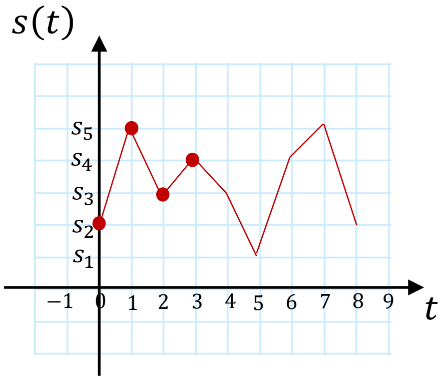
Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents everything from walking to playing games like Pong or Pinball.

RandomAgent on LunarLander-v2

RandomAgent on Ant-v2

tiene

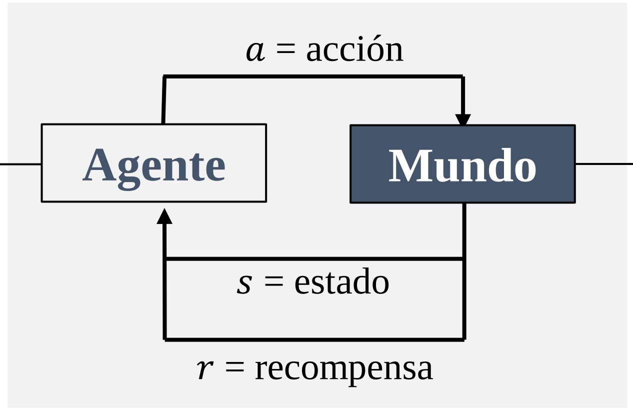
Trayectoria



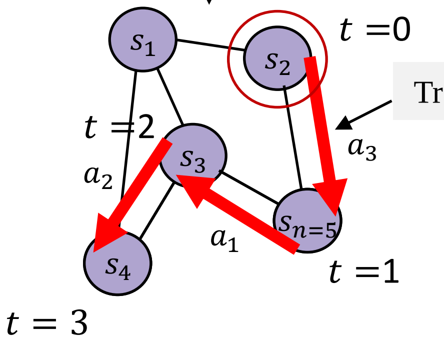
$$\tau = s(0), s(1), s(2), s(3), \dots, s(T)$$
$$= s_2, s_5, s_3, s_4, \dots, s_T$$

Trayectoria

con $t \in \{0, \dots, T\}$



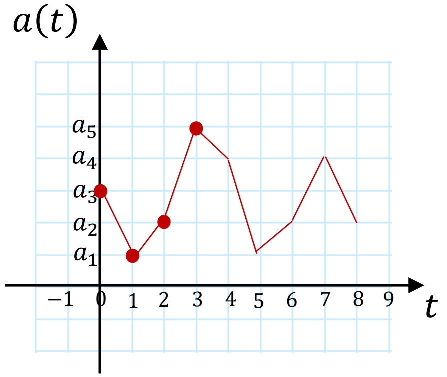
es



Trayectoria

tiene

Recompensa Acumulada



$$\tau_a = a(0), a(1), a(2), \dots, a(T-1)$$
$$= a_3, a_1, a_2, \dots, a_{(T-1)}$$

$$f_{RA}(\tau) = \sum_{t=0}^{(T-1)} \gamma^t f_R(s(t), a(t), s(t+1))$$

con $r_{(t+1)} = f_R(s(t), a(t), s(t+1))$

$\gamma \in [0{,}1]$

$$= r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{(T-1)} r_T$$

NOTA: Convergencia

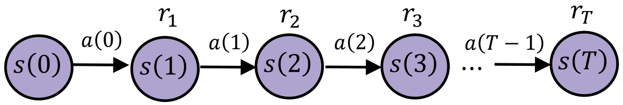
$$\sum_{n=0}^{N} a x^n = a\left[\frac{1 - x^{(N+1)}}{1 - x}\right], x \neq 1$$

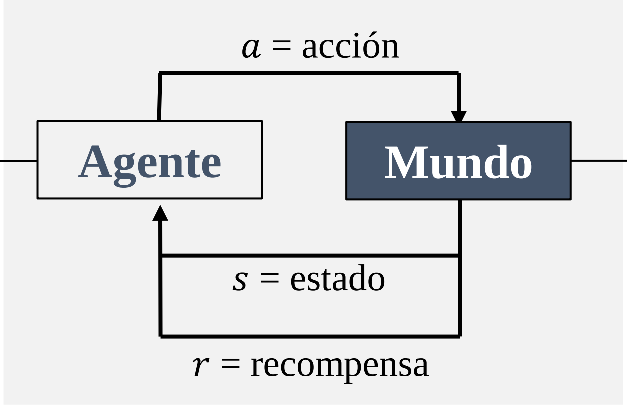
$$= r_1 + \gamma\left[r_2 + \gamma r_3 + \cdots + \gamma^{(T-2)} r_T\right]$$

$$= r_1 + \gamma\left[\sum_{t=1}^{(T-1)} \gamma^{(t-1)} f_R(s(t), a(t), s(t+1))\right]$$

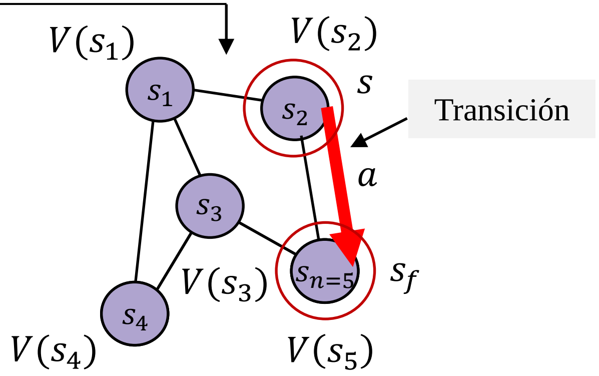
$$= r_1 + \gamma\left[f_{RA}(\tilde{\tau})\right]$$

con $\tilde{\tau} = s(1), s(2), s(3), \dots, s(T)$





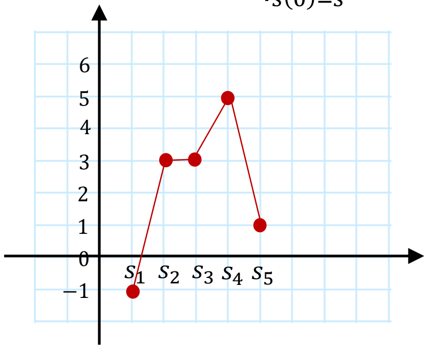
es



tiene

Recompensa Acumulada Promedio

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{s(0)=s}$$



tiene

Ecuación de Bellman

tipo

M_T Determinista

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{s(0)=s}$$

$$= \overline{f_R(s,a,s_f) + \gamma V(s_f)}$$

$$= f_R(s,a,s_f) + \gamma V(s_f)$$

con: $a = f_\pi(s)$

$$= f_R(s,f_\pi(s),s_f) + \gamma V(s_f)$$

tipo

M_T NO determinista

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{s(0)=s}$$

$$= \overline{f_R(s,a,s_f) + \gamma V(s_f)}$$

$$= \sum_{s_f \in S} P_{M_T}(s_f|s,a)\left[f_R(s,a,s_f) + \gamma V(s_f)\right]$$

con: $a = f_\pi(s)$

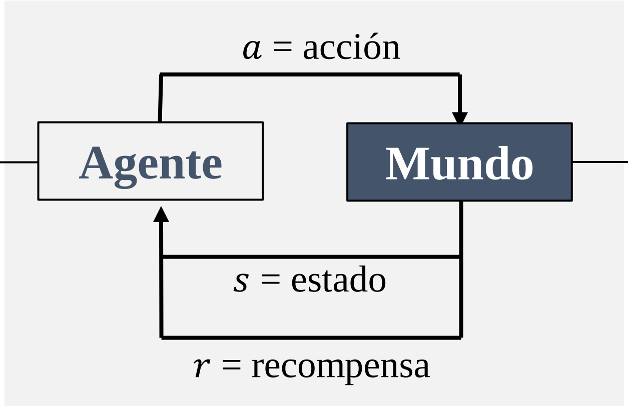
$$= \sum_{s_f \in S} P_{M_T}\left(s_f|s,f_\pi(s)\right)\left[f_R(s,f_\pi(s),s_f) + \gamma V(s_f)\right]$$

Solución: Iteración de Valor

$$V(s) \leftarrow f_R(s,f_\pi(s),s_f) + \gamma V(s_f)$$

$$V(s) \leftarrow \sum_{s_f \in S} P_{M_T}\left(s_f|s,f_\pi(s)\right)\left[f_R(s,f_\pi(s),s_f) + \gamma V(s_f)\right]$$

EJEMPLO:



toma

a = acción

tipo

determinista

$a = f_{\pi}(s)$

3	→	→	→	+1
2	↑	🟦	↑	−1
1	↑	←	←↑	←
	1	2	3	4

da

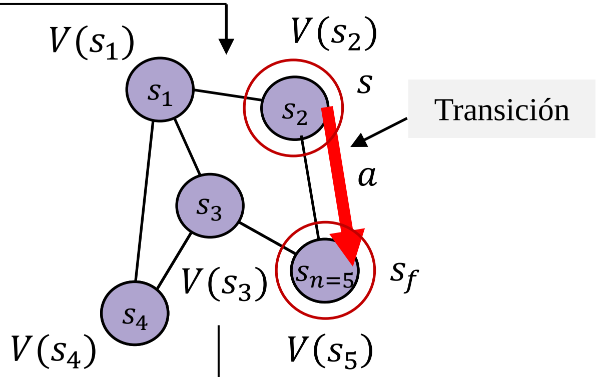
Recompensa Acumulada Promedio

$$V(s) = \sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)]$$

$\gamma = 1$

3	0.8516	0.9078	0.9578	+1
2	0.8016	🟦	0.7003	−1
1	0.7453	0.6953	0.6514	0.4279
	1	2	3	4

es



tiene

Estados

$S = \{s_{\mathrm{I}}, s_{21}, \dots, s_{\mathrm{F1}}, s_{\mathrm{F2}}\}$

3	s_{13}	s_{23}	s_{33}	s_{F1} +1
2	s_{12}	🟦	s_{32}	s_{F2} −1
1	s_{I} START	s_{21}	s_{31}	s_{41}
	1	2	3	4

tiene

Acciones

$A = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

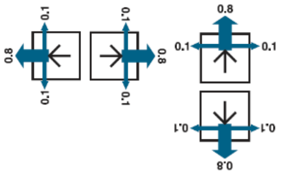
tiene

M_T = Modelo de Transición

tipo

NO determinista

$s_f \sim P_{M_T}(s_f|s,a)$

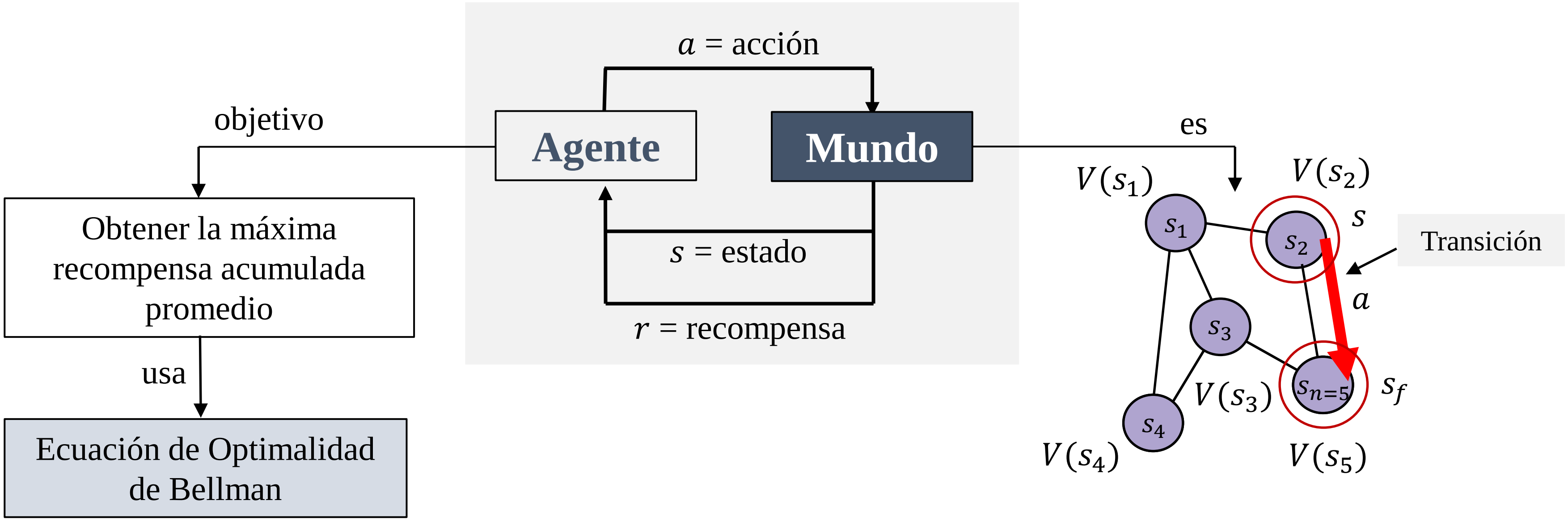


tiene

Recompensa

$r_f = f_R(s_f)$

3	-0.04	-0.04	-0.04	+1
2	-0.04	🟦	-0.04	−1
1	START	-0.04	-0.04	-0.04
	1	2	3	4


objetivo

Obtener la máxima recompensa acumulada promedio

usa

Ecuación de Optimalidad de Bellman

a = acción

Agente

Mundo

s = estado

r = recompensa

es

$$V(s) = \max_a \left[\overline{f_R(s,a,s_f) + \gamma V(s_f)}\right] = \max_a \left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)]\right]$$

definir

$$Q(s,a) = \left[\overline{f_R(s,a,s_f) + \gamma V(s_f)}\right]$$

$$= \left[\overline{f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)]}\right] = \sum_{s_f \in S} P_{M_T}(s_f|s,a)\left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)]\right]$$

$$V(s) = \max_a[Q(s,a)]$$

	V(s)
s1	V(s1)
s2	V(s2)
s3	V(s3)
s4	V(s4)
s(n=5)	V(s5)

		Q(s,a)	
	a1	a2	a(m=3)
s1	Q(s1,a1)	Q(s1,a2)	Q(s1,a3)
s2	Q(s2,a1)	Q(s2,a2)	Q(s2,a3)
s3	Q(s3,a1)	Q(s3,a2)	Q(s3,a3)
s4	Q(s4,a1)	Q(s4,a2)	Q(s4,a3)
s(n=5)	Q(s5,a1)	Q(s5,a2)	Q(s5,a3)

a = acción

Agente

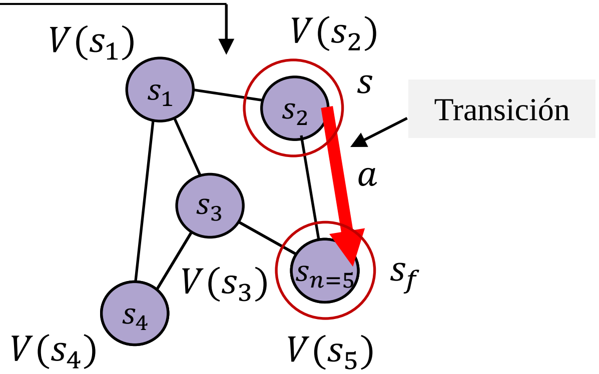
Mundo

s = estado

r = recompensa

objetivo

es



Obtener la máxima recompensa acumulada promedio

usa

Ecuación de Optimalidad de Bellman

Solución

Aproximación

Iteración de Valor

$$V(s) \leftarrow \max_a \left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)] \right]$$

Desconocidos

$$Q(s,a) \leftarrow \sum_{s_f \in S} P_{M_T}(s_f|s,a) \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right]$$

Desconocidos

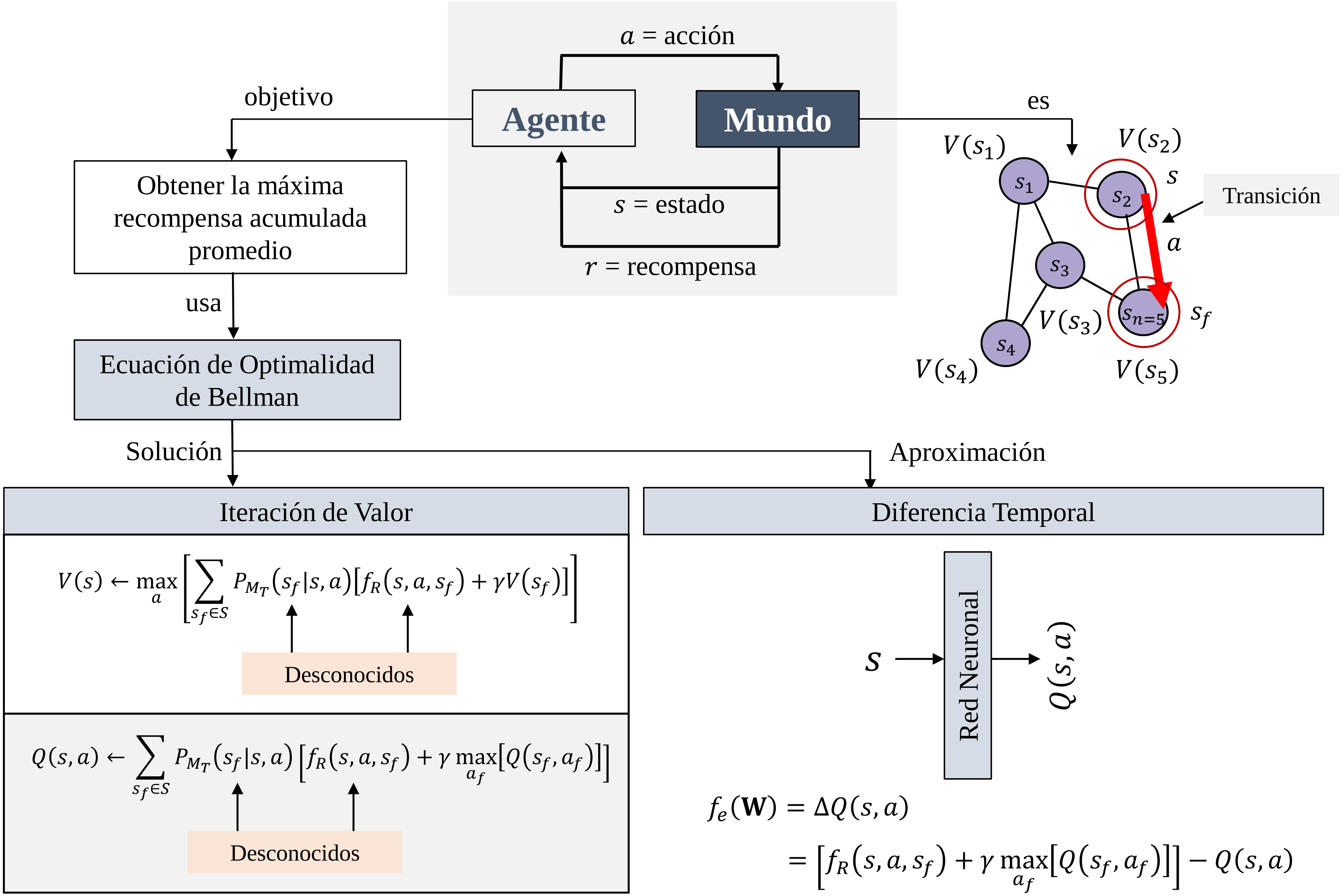
Diferencia Temporal

$$V(s) \leftarrow V(s) + \alpha \Delta V(s)$$

$$\Delta V(s) = [f_R(s,a,s_f) + \gamma V(s_f)] - V(s)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha \Delta Q(s,a)$$

$$\Delta Q(s,a) = \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right] - Q(s,a)$$



a = acción

objetivo

Agente

Mundo

es

s = estado

r = recompensa

Obtener la máxima recompensa acumulada promedio

usa

Ecuación de Optimalidad de Bellman

Solución

Aproximación

Iteración de Valor

$$V(s) \leftarrow \max_{a} \left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)] \right]$$

Desconocidos

$$Q(s,a) \leftarrow \sum_{s_f \in S} P_{M_T}(s_f|s,a) \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right]$$

Desconocidos

Diferencia Temporal

s

Red Neuronal

$Q(s,a)$

$$f_e(\mathbf{W}) = \Delta Q(s,a)$$
$$= \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right] - Q(s,a)$$

APENDICE: Ecuaciones

Recompensa Acumulada Promedio

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{\tilde{s}(0)=s}$$

$$= \frac{1}{N}\sum_{k=1}^{N} f_{RA}(\tau_k)\Big|_{\widetilde{s_k}(0)=s}$$

$$= \frac{1}{N}\left[f_{RA}(\tau_1)\Big|_{\widetilde{s_1}(0)=s} + \cdots + f_{RA}(\tau_N)\Big|_{\widetilde{s_N}(0)=s}\right]$$

con $f_{RA}(\tau_k)|_{\widetilde{s_k}(0)=s} = r_0^k + \gamma r_1^k + \gamma^2 r_2^k + \cdots + \gamma^T\, r_{T_k}^k$

$$\begin{aligned} = \;& \frac{1}{N}\left[r_0^1 + \gamma r_1^1 + \gamma^2 r_2^1 + \gamma^3 r_3^1 + \cdots + \gamma^{T_1}\, r_{T_1}^1\right] + \\ & \frac{1}{N}\left[r_0^2 + \gamma r_1^2 + \gamma^2 r_2^2 + \gamma^3 r_3^2 + \cdots + \gamma^{T_2}\, r_{T_2}^2\right] + \\ & \vdots \\ & \frac{1}{N}\left[r_0^N + \gamma r_1^N + \gamma^2 r_2^N + \gamma^3 r_3^N + \cdots + \gamma^{T_N}\, r_{T_N}^N\right] \end{aligned}$$

$$\begin{aligned} = \;& \frac{1}{N}[r_0^1] + \frac{1}{N}\gamma\left[r_1^1 + \gamma^1 r_2^1 + \gamma^2 r_3^1 + \cdots + \gamma^{(T_1-1)}\, r_{T_1}^1\right] + \\ & \frac{1}{N}[r_0^2] + \frac{1}{N}\gamma\left[r_1^2 + \gamma^1 r_2^2 + \gamma^2 r_3^2 + \cdots + \gamma^{(T_2-1)}\, r_{T_2}^2\right] + \\ & \vdots \\ & \frac{1}{N}[r_0^N] + \frac{1}{N}\gamma\left[r_1^N + \gamma^1 r_2^N + \gamma^2 r_3^N + \cdots + \gamma^{(T_N-1)}\, r_{T_N}^N\right] \end{aligned}$$

$$\begin{aligned} \tau_1 &= \widetilde{s_1}(0), \widetilde{s_1}(1), \widetilde{s_1}(2), \widetilde{s_1}(3), \dots, \widetilde{s_1}(T_1) \\ \tau_2 &= \widetilde{s_2}(0), \widetilde{s_2}(1), \widetilde{s_2}(2), \widetilde{s_2}(3), \dots, \widetilde{s_2}(T_2) \\ & \vdots \\ \tau_N &= \widetilde{s_N}(0), \widetilde{s_N}(1), \widetilde{s_N}(2), \widetilde{s_N}(3), \dots, \widetilde{s_N}(T_N) \end{aligned}$$

$$\begin{aligned} \tau_1 &= \widetilde{s_1}(1), \widetilde{s_1}(2), \widetilde{s_1}(3), \dots, \widetilde{s_1}(T_1) \\ \tau_2 &= \widetilde{s_2}(1), \widetilde{s_2}(2), \widetilde{s_2}(3), \dots, \widetilde{s_2}(T_2) \\ & \vdots \\ \tau_N &= \widetilde{s_N}(1), \widetilde{s_N}(2), \widetilde{s_N}(3), \dots, \widetilde{s_N}(T_N) \end{aligned}$$

$$\begin{aligned} = \;& \frac{1}{N}[r_0^1] + \frac{1}{N}\gamma\left[f_{RA}(\tau_1)\Big|_{\widetilde{s_1}(0)=\widetilde{s_1}(1)} = r_1^1 + \gamma^1 r_2^1 + \gamma^2 r_3^1 + \cdots + \gamma^{(T_1-1)}\, r_{T_1}^1\right] + \\ & \frac{1}{N}[r_0^2] + \frac{1}{N}\gamma\left[r_1^2 + \gamma^1 r_2^2 + \gamma^2 r_3^2 + \cdots + \gamma^{(T_2-1)}\, r_{T_2}^2\right] + \\ & \vdots \\ & \frac{1}{N}[r_0^N] + \frac{1}{N}\gamma\left[r_1^N + \gamma^1 r_2^N + \gamma^2 r_3^N + \cdots + \gamma^{(T_N-1)}\, r_{T_N}^N\right] \end{aligned}$$

$$= s_1, s_2, s_3, s_4, s_5$$

$$= a_1, a_2, a_3, a_4, a_5$$

$$= 0.8a_1, 0.2a_1, 1a_2, 0.8a_2, 0.2a_2, 0.3a_3, 0.4a_3, a_4, a_5$$