

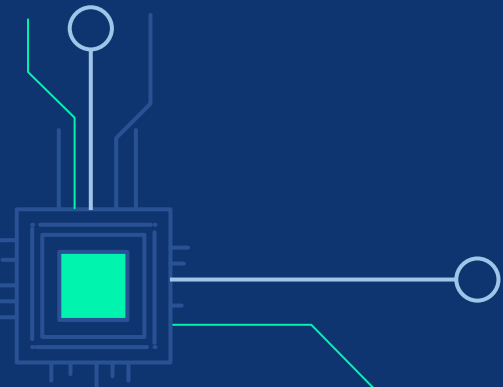


AYUDANTÍA 11

Reinforcement Learning



MARKOV DECISION PROCESS





MARKOV DECISION PROCESS

Supuestos:

- Agente activo
- Observabilidad Total
- Mundo Markoviano



MARKOV DECISION PROCESS





¿CÓMO SE DEFINE UN MDP?





MARKOV DECISION PROCESS

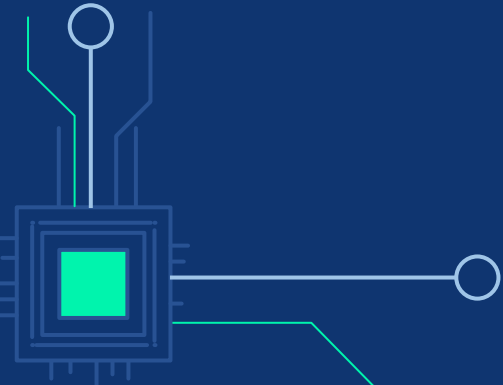
MDP = (S, R, A, P)

- S : Conjunto de estados posibles
- R : Conjunto de Rewards
- A : Conjunto de acciones posibles
- P : Conjunto de probabilidades entre estados.

$$P_{ij}^k = P(s_j | s_i, a_k)$$



REINFORCEMENT LEARNING





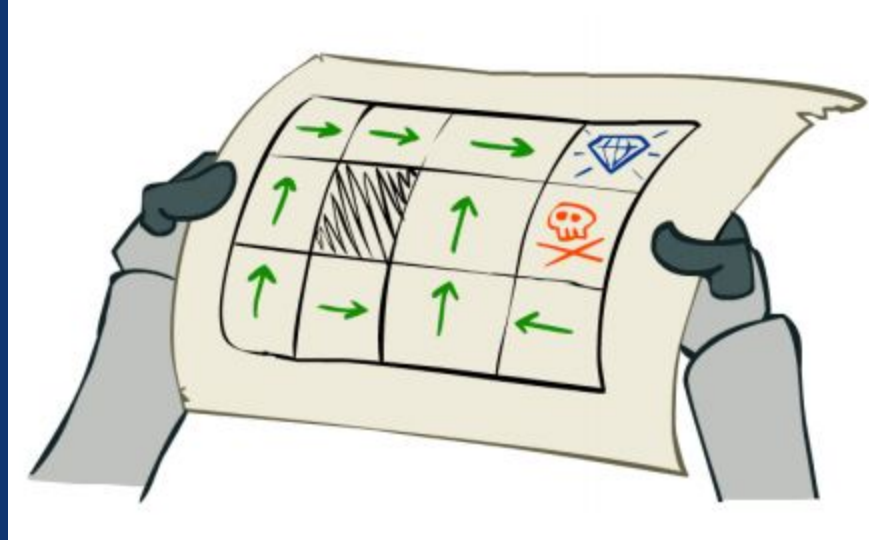
REINFORCEMENT LEARNING

Actuación del agente:

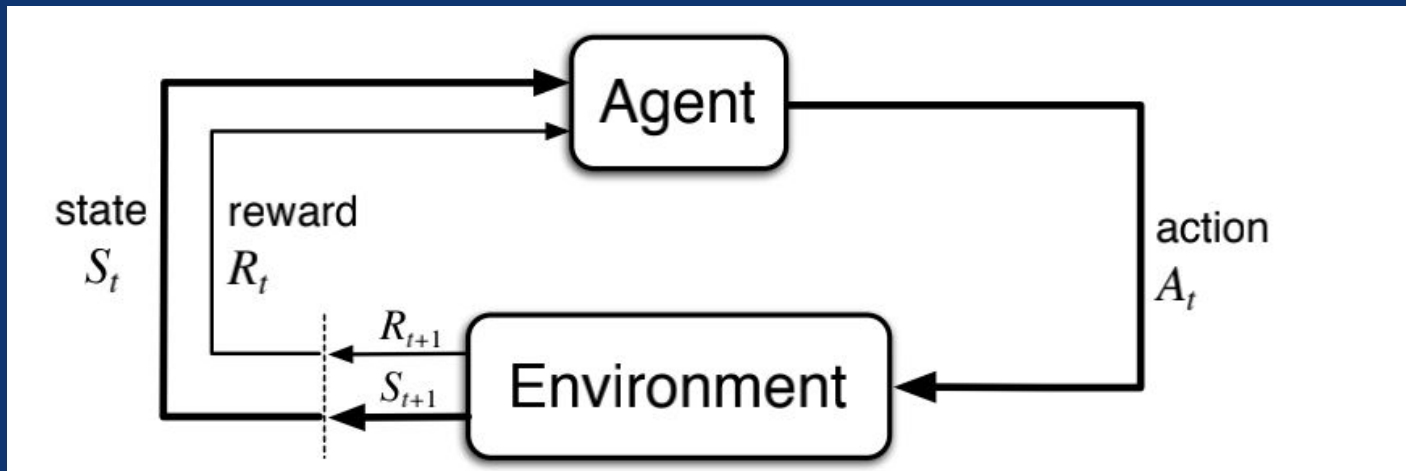
- Maximizar recompensa esperada
- Generar política de acción



REINFORCEMENT LEARNING



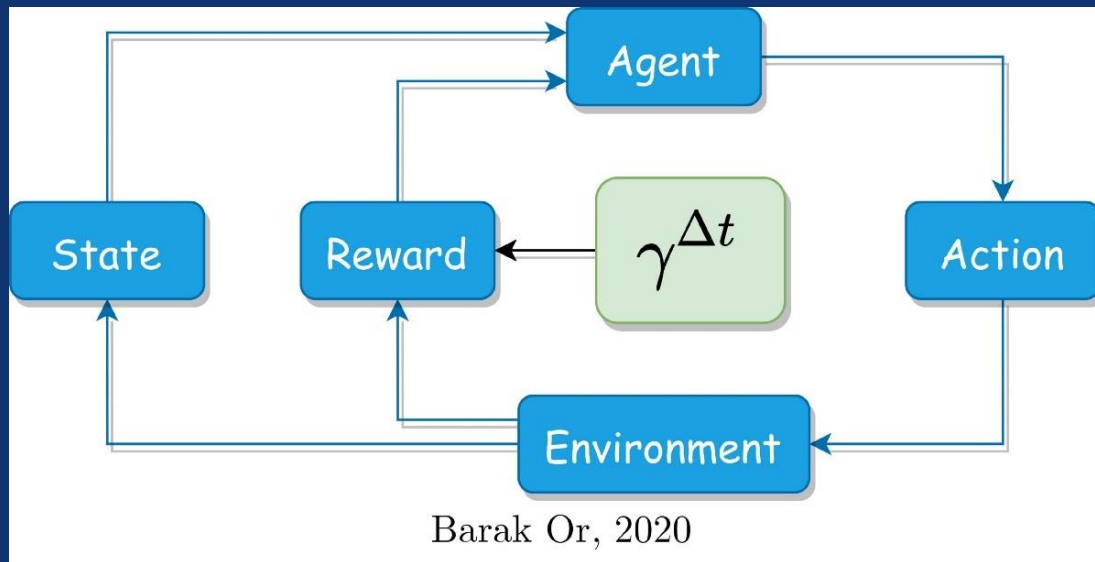
REINFORCEMENT LEARNING



Experiencias de entrenamiento



REINFORCEMENT LEARNING



Factor de descuento



REINFORCEMENT LEARNING

$$v_{\pi}(s) = \mathbb{E}_{\pi} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s]$$

Expected

Reward
discounted

Given that state

Función Valor





REINFORCEMENT LEARNING

$$V(s) = \max_a \left(R(s, a) + \gamma \sum_{s'} P(s, a, s') V(s') \right)$$

Ecuación Bellman



REINFORCEMENT LEARNING

Iteración de valores

```
Initialize  $V(s)$  arbitrarily
loop until policy good enough
  loop for  $s \in \mathcal{S}$ 
    loop for  $a \in A$ 
       $Q(s, a) := R(s) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') \hat{V}(s')$ 
    end loop
     $\hat{V}(s) := \max_a Q(s, a)$ 
  end loop
```



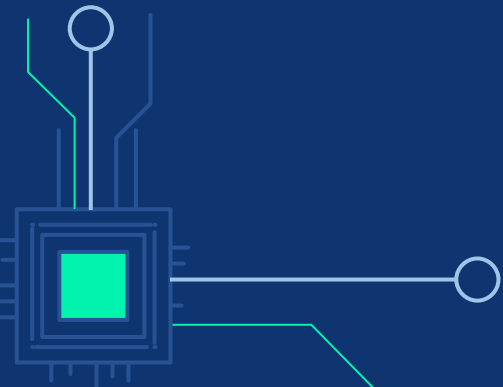
REINFORCEMENT LEARNING

Iteración de políticas

```
Choose an arbitrary policy  $\pi'$ 
Loop
     $\pi := \pi'$ 
    Compute value function of policy  $\pi$ :
    #solve linear equations over  $V_\pi(s)$ 
     $V_\pi(s) := R(s) + \gamma \sum_{s' \in S} T(s, \pi(s), s') V_\pi(s')$ 
    Improve the policy at each state
     $\pi'(s) := \arg \max_a (R(s) + \gamma \sum_{s' \in S} T(s, a, s') V_\pi(s'))$ 
until  $\pi = \pi'$ 
```



Q-LEARNING





Q-LEARNING

$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a')$$



Q-LEARNING

$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a')$$

↑
Valor de la
acción tomada
en el estado
actual

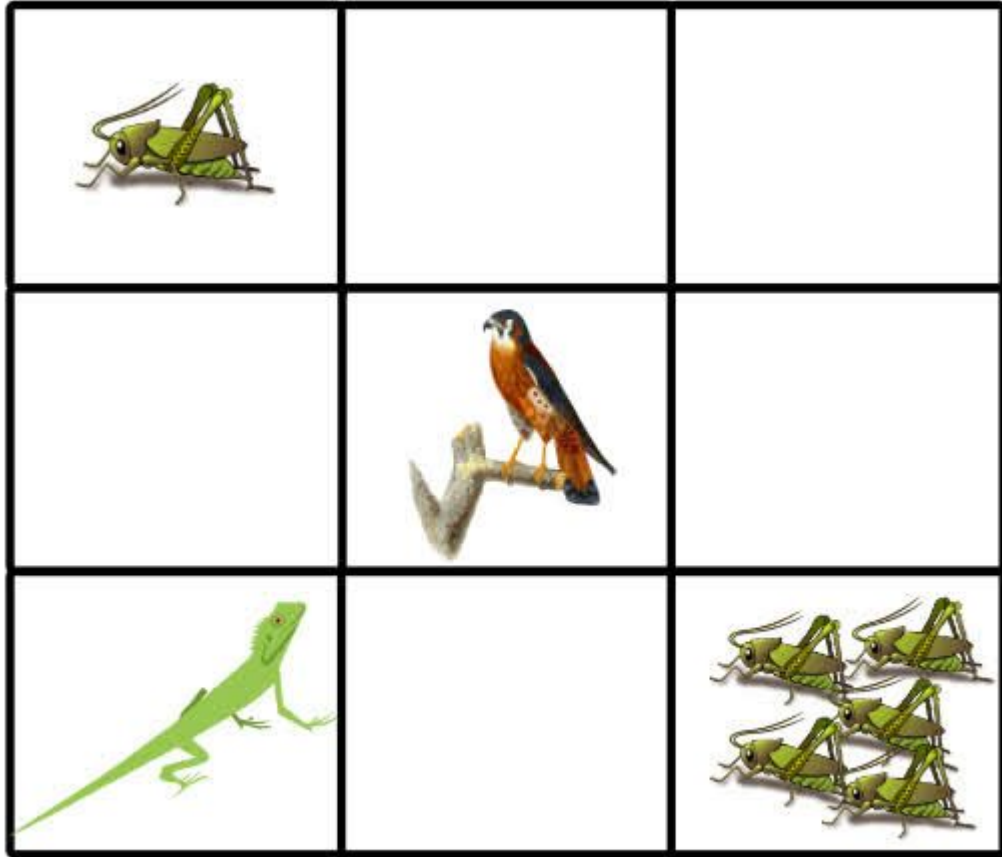
↗
Recompensa
entregada por
tomar la acción
en el estado
actual

↗
Tasa de
descuento.
Cuánta
importancia le
doy a las
recompensas
futuras

↑
El valor que me
entrega la mejor
acción en el
próximo estado

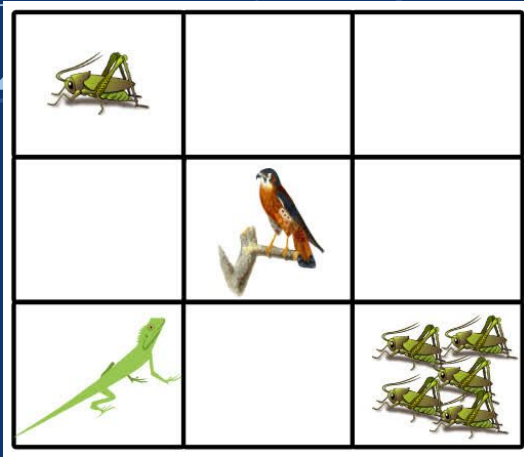






Estado	Recompensa
Un grillo	+1
Vacío	- 1
5 Grillos	+10 (Fin del juego)
Pájaro	- 10 (Fin del juego)

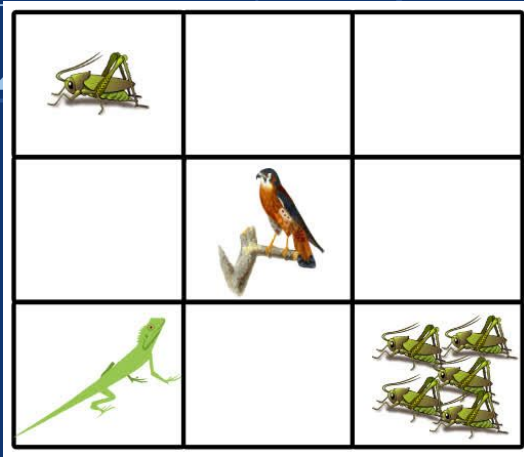




Estado	Recompensa
Un grillo	+1
Vacío	- 1
5 Grillos	+10 (Fin del juego)
Pájaro	- 10 (Fin del juego)

	Left	Right	Up	Down
1 Grillo	0	0	0	0
Vacío 1	0	0	0	0
Vacío 2	0	0	0	0
Vacío 3	0	0	0	0
Pájaro	0	0	0	0
Vacío 4	0	0	0	0
Vacío 5	0	0	0	0
Vacío 6	0	0	0	0
5 Grillos	0	0	0	0





$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a')$$

Estado	Recompensa
Un grillo	+1
Vacío	- 1
5 Grillos	+10 (Fin del juego)
Pájaro	- 10 (Fin del juego)

	Left	Right	Up	Down
1 Grillo	0	0	0	0
Vacío 1	0	0	0	0
Vacío 2	0	0	0	0
Vacío 3	0	0	0	0
Pájaro	0	0	0	0
Vacío 4	0	0	0	0
Vacío 5	0	0	0	0
Vacío 6	0	0	0	0
5 Grillos	0	0	0	0



¿EXPLORAR O EXPLOTAR?





Q-LEARNING

$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a')$$

↑
Valor de la
acción tomada
en el estado
actual

↗
Recompensa
entregada por
tomar la acción
en el estado
actual

↗
Tasa de
descuento.
Cuánta
importancia le
doy a las
recompensas
futuras

↑
El valor que me
entrega la mejor
acción en el
próximo estado










-1 (90%)

-100 (10%)





		
		
	<p>-1 (90%)</p> <p>-100 (10%)</p>	

$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a') \gggggg$$

Q-LEARNING

$$Q(s, a) = r(s, a) + \gamma \arg \max_{a'} Q(s', a')$$

↑
Valor de la
acción tomada
en el estado
actual

↗
Recompensa
entregada por
tomar la acción
en el estado
actual

↗
Tasa de
descuento.
Cuánta
importancia le
doy a las
recompensas
futuras

↑
El valor que me
entrega la mejor
acción en el
próximo estado



Q-LEARNING

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r(s, a) + \gamma \arg \max_{a'} Q(s', a'))$$

Valor de la acción tomada en el estado actual

Valor antiguo en la tabla

Cuánta importancia le doy a mi experiencia nueva

Recompensa entregada por tomar la acción en el estado actual

Tasa de descuento. Cuánta importancia le doy a las recompensas futuras

El valor que me entrega la mejor acción en el próximo estado

