Introducci?n al Aprendizaje por Refuerzos

June 3, 2020

1 Facultad Regional Villa María

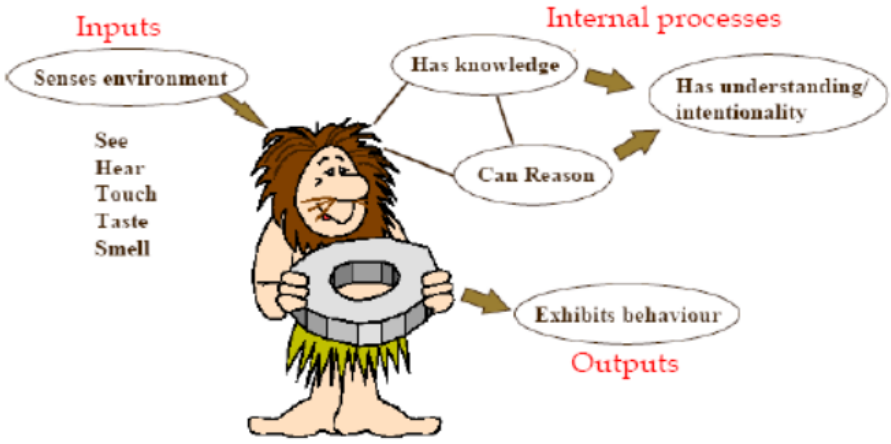
1.1 5to año - Ingeniería en Sistemas de Información

1.1.1 Introducción al Aprendizaje por Refuerzos

- Introducción. Modelo Agente Entorno. Agente Situado. Arquitectura Actor-Crítico.
- Aprendizaje por Refuerzos. Elementos. Ciclo del Aprendizaje por Refuerzos. Definición Formal.
- Procesos de Decisión de Markov. Función de Valor. Ecuación de Bellman. Optimalidad.
- Aproximaciones al Aprendizaje. Model Free y Model Based.
  - Iteración de Política.
  - Iteración de Valor.

1.2 Introducción: Entidad Inteligente -> Agente Situado

- El desarrollo de la inteligencia requiere que la entidad o el agente esté situada/o en un entorno **(Measuring universal intelligence: Towards an anytime intelligence test, Hernandez-Orallo & Dowe, Artificial Intelligence, 2010).** Agent.bb



1.3 Agent-Environment Framework

- El agente y su entorno interactúan a través de la ejecución de acciones, observación de estados y señales de reward (recompensa). La inteligencia tendrá efecto sólo si el agente tiene

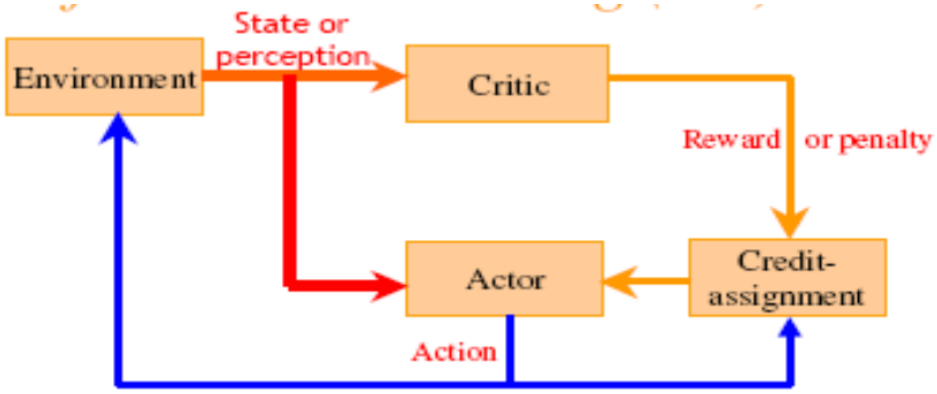


Diagram.bb

claramente definidos objetivos o metas que persigue activamente mientras ocurre la interac-

ción. Interaction.bb

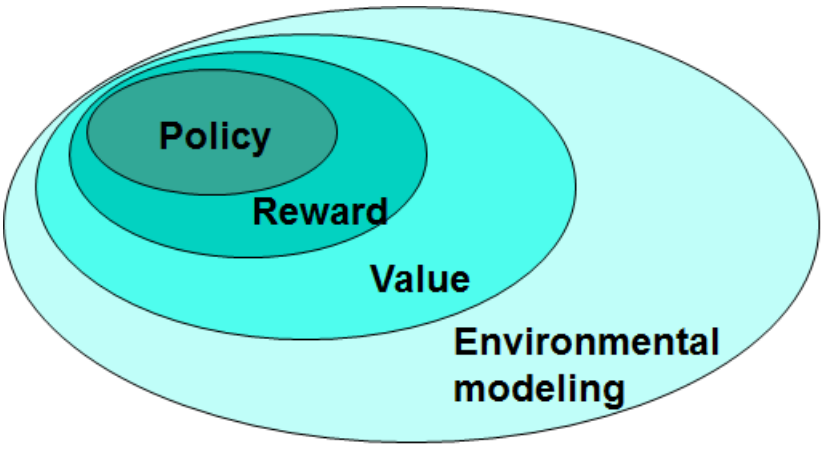
1.4 Arquitectura Actor-Crítico

1.5 Aprendizaje por Refuerzos

- La toma de decisiones secuencial involucra aprender sobre nuestro entorno y elegir acciones que maximizan el retorno esperado. El RL computacional, inspirado por estas ideas, las formalizo y produjo un impacto importante en robótica, machine learning y neurociencias.
- El Aprendizaje por Refuerzos (RL) consiste en un agente que se encuentra en algún estado $s \in S$ inmerso en un entorno E y toma acciones $a \in A$ en busca de una meta. El agente puede ser modelado formalmente como una función f, que toma un historial de interacción como entrada, y devuelve una acción a tomar. Una manera conveniente para representar el agente es una medida de probabilidad sobre el set A de acciones, en base a un historial de interacción:

  $$f(a_n|s_1a_1r_2s_2a_2...r_ns_n)$$

  que representa la probabilidad de la acción a en el ciclo n dado un historial de interacción.
- Problema RL: £Cómo el agente produce la distribución de probabilidad sobre las acciones?
- Dilema de exploración - explotación: debido a que el Agente no recibe ejemplos de entrenamiento, debe probar alternativas, procesar los resultados de sus acciones y modificar su



Elements.bb

comportamiento en algún sentido. £Cuándo explotar este conocimiento vs. cuándo probar nuevas estrategias?

1.6 Elementos del Aprendizaje por Refuerzos

- **Policy (Política):**

Una política define la manera de comportarse de un agente, en cualquier momento de tiempo dado. Basicamente, es un mapeo de un estado o percepción s a una acción a, pudiendo ser estocásticas.

- **Reward Function (Función de Recompensa)**

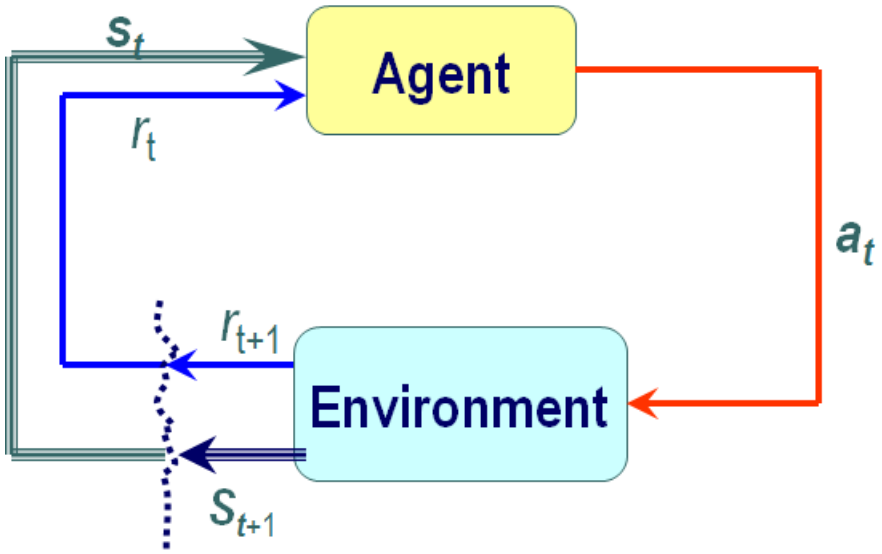
Define cuantitativamente el objetivo del agente. Es un mapeo de un par estado-acción a un número real que indica "cuán deseable" es ejecutar dicha acción en ese estado. Asimismo, el único objetivo del agente es maximizar la recompensa total que recibe a lo large del tiempo. Cabe mencionar que, si bien la función de reward no puede ser alterada por el agente, provee las bases para cambiar la política del mismo.

- **Value function (Función de Valor)**

La función de valor se diferencia de la función de reward en el sentido de que indica "cuán deseable" es, a largo plazo, ejecutar una acción en un determinado estado. Así, el valor de un estado s es la cantidad total de reward que el agente espera obtener a futuro comenzando la interacción en el estado s.

- **Environment (Entorno)**

El entorno se encuentra constitutido por todo aquel elemento (real o simulado) que el agente no puede controlar. Es con quién el agente interactúa a partir de la ejecución de acciones de control.



Cycle.bb

1.7 Ciclo del Aprendizaje por Refuerzos

1.7.1 Definición formal

- Si el problema de RL dado tiene un conjunto finito de estados y acciones y satisface la propiedad de Markov entonces puede definirse como un Proceso de Decisión de Markov

$$MDPFinito = (S, A, P(.), R(.),) \tag{1}$$

donde

$$S = s_1, s_2, ..., s_n$$

es un conjunto finito de estados.

$$A = a_1, a_2, ..., a_m$$

es un conjunto finito de acciones.

$$P_a(s, s') = P(s_{t+1} = s | s_t = s, a_t = a)$$

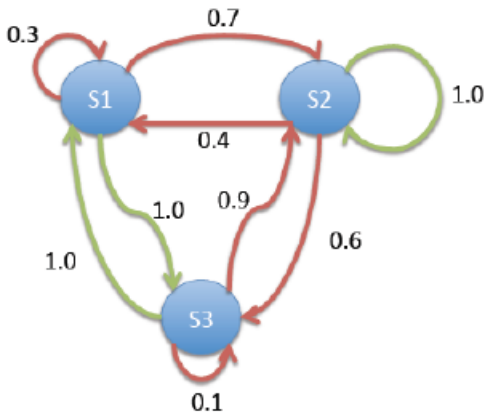
es la probabilidad de que la acción a tomada en tiempo t y en estado s lleve al agente al estado s' en tiempo t+1

$$R_a(s, s')$$

es la recompensa inmediata recibido tras transicionar, luego de tomar la acción a, desde el estado s al estado s'

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

es el factor de descuento, representando la diferencia en la importancia de la recompensa a corto plazo vs la recompensa a largo plazo.



R(s1) = +1
R(s2) = 0
R(s3) = -1

Función de transición T:

S	A	S'	p
S1	R	S1	0.3
	R	S2	0.7
	V	S3	1.0
S2	R	S1	0.4
	R	S2	0.6
	V	S2	1.0
S3	V	S1	1.0
	R	S3	0.1
	R	S2	0.9

- Un episodio (instancia) de este MDP forma una secuencia finita

$$s_0, a_0, r_1, s_1, a_1, r_2, s_2, ..., s_{n-1}, a_{n-1}, r_n, s_n$$

donde

$$s_n$$

es un estado final (o n es el tiempo de corte).

- La recompensa total del episodio está dado por

$$R = r_1 + r_2 + ... + r_n$$

- En consecuencia, la recompensa a futuro partiendo del tiempo t está dado por

$$R_t = r_t + r_{t+1} + ...$$

- Hay que considerar que el ambiente es estocástico en la mayor parte de los entornos reales y, por tanto, la recompensa suele diverger mientras más alejado se encuentre el instante de tiempo considerado. Es por esto que se utiliza un parámetro llamado *factor de descuento*, para descontar el valor de las recompensas futuras. De esta manera,

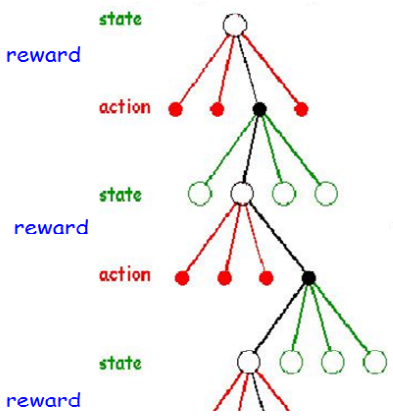
$$R_t = r_t + r_{t+1} + {}^2r_{t+2} + {}^3r_{t+3} + ... = r_t + (r_{t+1} + r_{t+2} + {}^2r_{t+3}...) = r_t + R_{t+1} \quad (2)$$

- Si utilizamos $= 0$, el agente priorizará sólo la recompensa inmediata, mientras que $\gamma = 1$ hará que considere todas los recompensas de la misma manera, independientemente del momento en donde las reciba. Problem Statement.bb

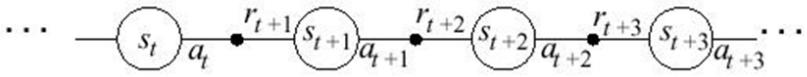
de Estado Valor.bb

Función de Estado - Valor para la Política π :

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\}$$



Predicción y control en secuencias *estado-acción-recompensa* , con un objetivo a largo plazo.



Definition.bb

Given an MDP $\langle S, A, T, R \rangle$, a policy is a computable function that outputs for each state $s \in S$ an action $a \in A$ (or $a \in A(s)$). Formally, a *deterministic* policy π is a function defined as $\pi : S \to A$. It is also possible to define a *stochastic* policy as $\pi : S \times A \to [0,1]$ such that for each state $s \in S$, it holds that $\pi(s,a) \geq 0$ and $\sum_{a \in A} \pi(s,a) = 1$

1.8 Procesos de Decisión de Markov

1.8.1 Función de Valor

- El valor de un estado es el retorno esperado por el agente, comenzando la interacción en dicho estado, dependiendo de la política ejecutada por el agente.

- El valor de la ejecución de una acción en un estado es el retorno esperado por el agente, comenzando la interacción en dicho estado a partir de la ejecución de dicha acción, dependiendo de la política ejecutada por el agente.

Una propiedad fundamental de las funciones de valor es que satisfacen ciertas propiedades recursivas. Para cualquier política y cualquier estado s, V(s) y Q(s,a) pueden ser definidas recursivamente en términos de la denominada *Ecuación de Bellman* ** (Bellman, 1957) **

1.8.2 Ecuación de Bellman

- La idea básica es:

de Accion Valor.bb

Función de Acción - Valor para la Política π :

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\}$$

$$\begin{aligned} R_t &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} \cdots \\ &= r_{t+1} + \gamma\left(r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} \cdots\right) \\ &= r_{t+1} + \gamma R_{t+1} \end{aligned}$$

- Entonces,

- O, sin el operador de valor esperado:

La ecuación anterior refleja el hecho de que el valor de un estado se encuentra definido en términos de la recompensa inmediata y los valores de los estados siguientes ponderados en función de las probabilidades de transición, y adicionalmente un factor de descuento.

1.8.3 Ecuación de Optimalidad de Bellman

La Ecuación de Optimalidad de Bellman refleja el hecho de que el Valor de un estado bajo la política óptima debe ser igual al retorno esperado para la mejor acción en dicho estado:

Al mismo tiempo, la acción óptima para un estado s dada la función de valor, puede obtenerse mediante:

La política anterior se denomina **Política Greedy**, dado que selecciona la mejor acción para cada estado, teniendo en cuenta la función de valor V(s). De manera análoga, la función de acción-valor óptima puede expresarse como:

- Valor.bb

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\left\{R_t \middle| s_t = s\right\} \\ &= E_{\pi}\left\{r_{t+1} + \gamma V^{\pi}(s_{t+1}) \middle| s_t = s\right\} \end{aligned}$$

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{\pi}(s') \right]$$

Equation.bb

$$V^{*}(s) = \max_{a \in A} \sum_{s' \in S} T(s,a,s') \left(R(s,a,s') + \gamma V^{*}(s') \right)$$

de Optimalidad Valor.bb

$$\pi^{*}(s) = \arg\max_{a} \sum_{s' \in S} T(s,a,s') \left(R(s,a,s') + \gamma V^{*}(s') \right)$$

Optima.bb

$$Q^{*}(s,a) = \sum_{s'} T(s,a,s') \left(R(s,a,s') + \gamma \max_{a'} Q^{*}(s',a') \right)$$

Valor Optima.bb

Comparación de políticas de actuación y política óptima

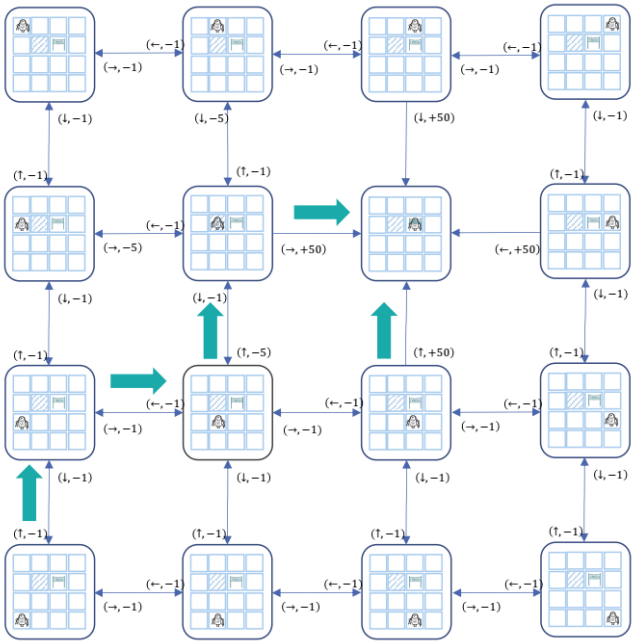
$\pi \geq \pi' \Leftrightarrow v_{\pi}(s) \geq v'_{\pi}(s)$

$\pi_{*} \geq \pi' \ \forall \pi'$ *(garantizada)*

Política π_1 $(\gamma = 1)$:

Estado	Acción
$(3,0)$	↑
$(2,0)$	→
$(2,1)$	↑
$(1,1)$	→
$(2,2)$	↑
...	

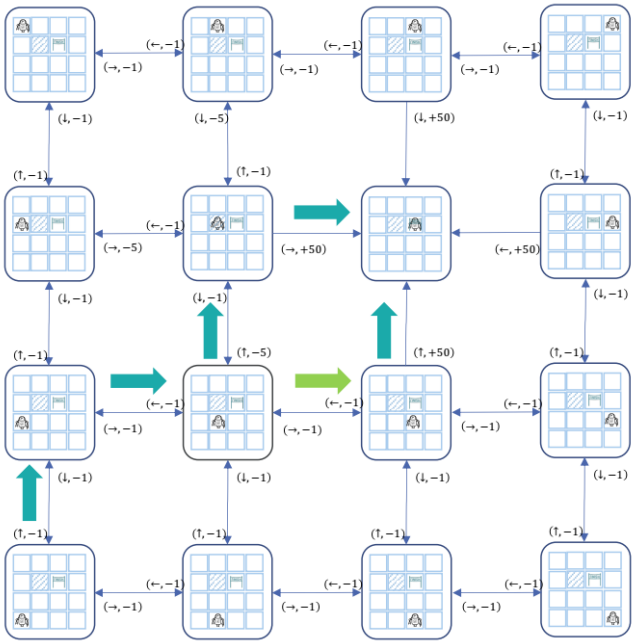
$v_{\pi_1}(\langle 2,1 \rangle) = -5 + 50 = 45$
$v_{\pi_1}(\langle 2,2 \rangle) = 50$
$v_{\pi_1}(\langle 3,0 \rangle) = -1 - 1 - 5 + 50 = 43$
$v_{\pi_1}(\langle 2,0 \rangle) = -1 - 5 + 50 = 42$



Política π_2 $(\gamma = 1)$:

Estado	Acción
$(3,0)$	↑
$(2,0)$	→
$(2,1)$	→
$(1,1)$	→
$(2,2)$	↑
...	

$v_{\pi_2}(\langle 2,1 \rangle) = -1 + 50 = 49 \geq 45$
$v_{\pi_2}(\langle 2,2 \rangle) = 50 \geq 50$
$v_{\pi_2}(\langle 3,0 \rangle) = -1 - 1 - 1 + 50 = 47 \geq 43$
$v_{\pi_2}(\langle 2,0 \rangle) = -1 - 1 + 50 = 48 \geq 42$



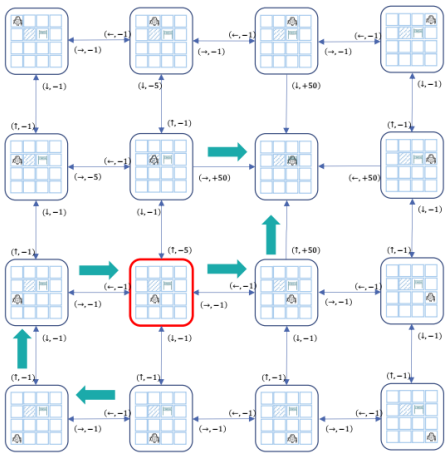
$$q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Política π_2 $(\gamma = 1)$:

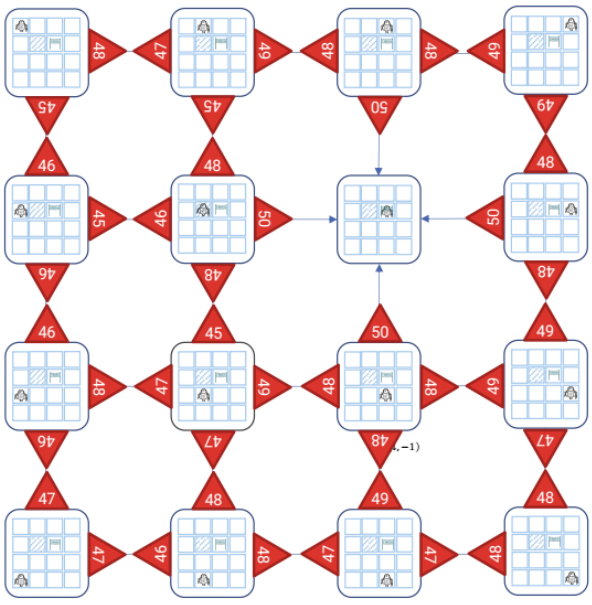
Estado	Acción
$(3,0)$	$\uparrow$
$(2,0)$	$\rightarrow$
$(2,1)$	$\rightarrow$
$(1,1)$	$\rightarrow$
$(2,2)$	$\uparrow$
$(3,1)$	$\leftarrow$
...	

$$v_\pi(s) = q_\pi(s, \pi(a|s))$$

$q_{\pi_2}(\langle 2,1\rangle, \uparrow) = -5 + 50 = 45$
$q_{\pi_2}(\langle 2,1\rangle, \rightarrow) = -1 + 50 = \mathbf{49}$
$q_{\pi_2}(\langle 2,1\rangle, \leftarrow) = -1 - 1 - 1 + 50 = 47$
$q_{\pi_2}(\langle 2,1\rangle, \downarrow) = -1 - 1 - 1 - 1 - 1 + 50 = 45$



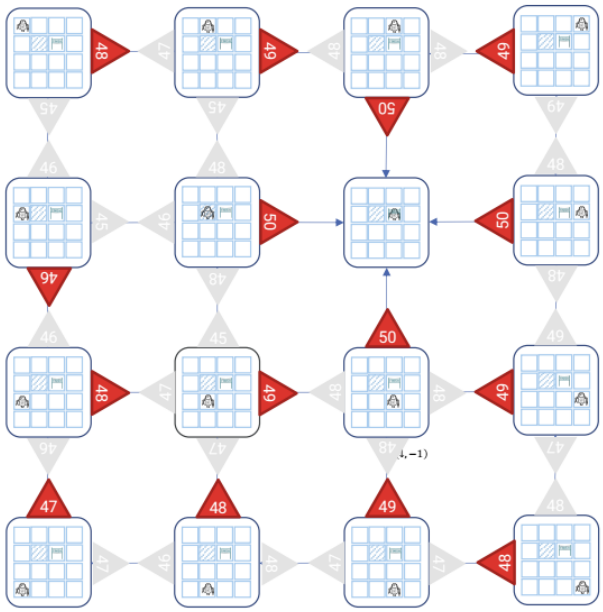
$$q_{\pi_*}(s,a)$$



$q_{\pi_*}(s,a)$

Resaltar valor máximo de cada estado:

$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a)$



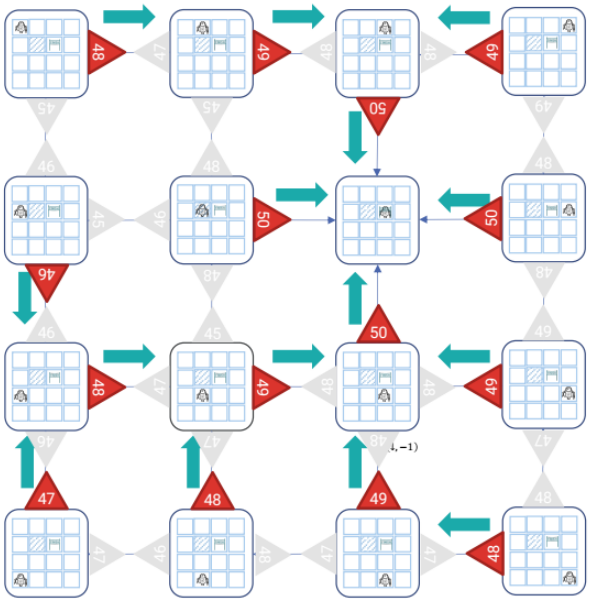
$q_{\pi_*}(s,a)$

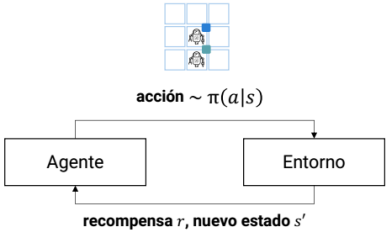
Resaltar valor máximo de cada estado:

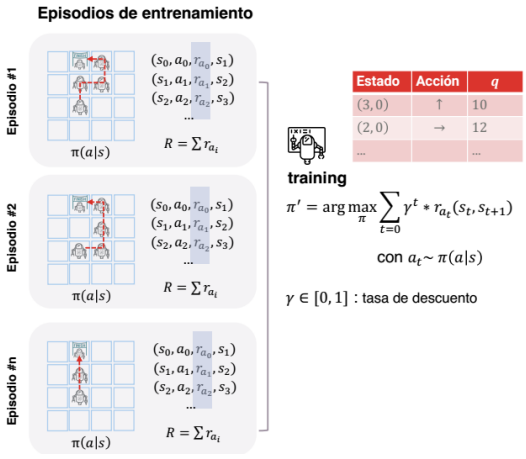
$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a)$

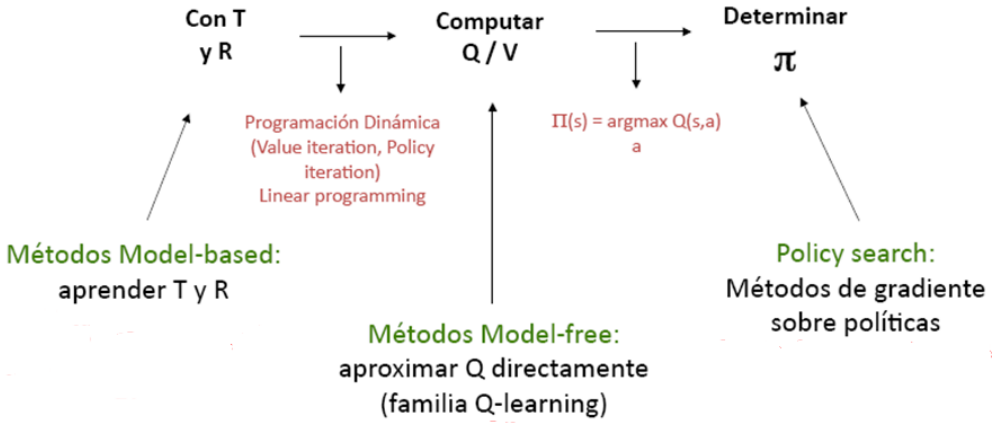
Reconstruir π_*

Estado	Acción
$(3,0)$	→
$(2,0)$	↑
...	

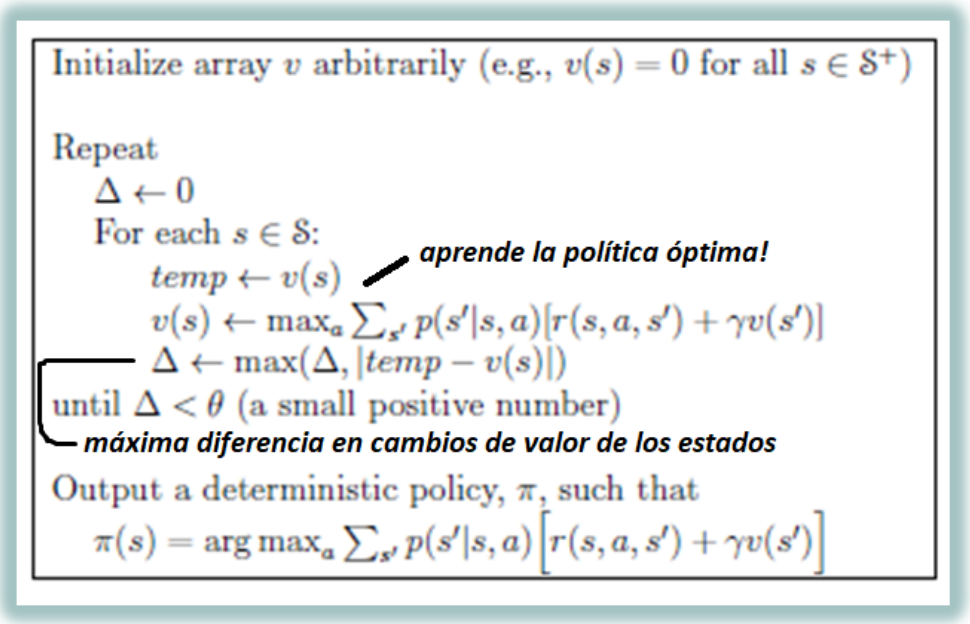








al aprendizaje.bb



de Valor.bb

1.9 Ejemplo MDP y Políticas (*)

1.10 Aproximaciones para el aprendizaje de V y Q

1.10.1 Model Based vs. Model Free

- Model-free aprende Q/V directamente y presenta muy baja complejidad computacional.
- Model-based aprende T y R y usa un algoritmo de planning para encontrar la política. Uso eficiente de los datos/experiencia. Alto costo computacional.

1.10.2 Programación Dinámica: Iteración de Valor e Iteración de Política (Model Based)

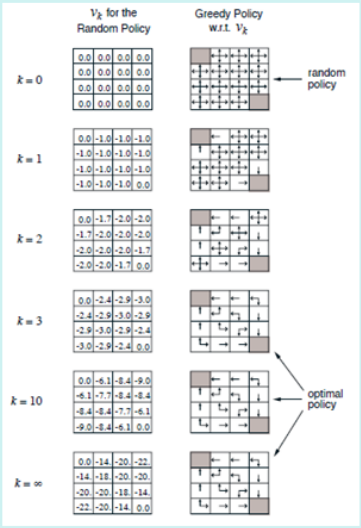
Iteración de Valor

Iteración de Política

1. Initialization
   $v(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Repeat
      $\Delta \leftarrow 0$
      For each $s \in \mathcal{S}$:
         $temp \leftarrow v(s)$
         $v(s) \leftarrow \sum_{s'} p(s'|s, \pi(s)) \Big[r(s, \pi(s), s') + \gamma v(s') \Big]$
         $\Delta \leftarrow \max(\Delta, |temp - v(s)|)$
   until $\Delta < \theta$ (a small positive number)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
      $temp \leftarrow \pi(s)$
      $\pi(s) \leftarrow \arg\max_a \sum_{s'} p(s'|s, a) \Big[r(s, a, s') + \gamma v(s') \Big]$
      If $temp \neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return v and π; else go to 2



de Politica.bb ***Secuencia de evaluación y mejora*** $\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_{*}.$