

# Aprendizaje por refuerzo

Clase 21: RL multi-agente



Antes de  
empezar

- <https://cinvescomp.cinvestav.mx/CC2023/>



# Antes de empezar

---



Comentarios tarea 4



Comentarios de proyecto



# Proyecto

---

- (40 puntos) Implementación (idealmente notebook en colab):
- (30 puntos) Científico: resumen en extenso en inglés de 4 páginas  
(<https://neurips.cc/Conferences/2022/PaperInformation/StyleFiles> )
- (30 puntos) Presentación - 15 minutos y 5 de preguntas
- Entrega 24 de mayo





# Proyecto

---

- (40 puntos) Implementación (idealmente notebook en colab):
  - Código del método utilizado
  - Pruebas de pruebas generación de exactamente los mismos resultados que los presentados en las otras partes (fijar semillas y parámetros)

# Proyecto

- (30 puntos) Científico: resumen en extenso en inglés de 4 páginas (<https://neurips.cc/Conferences/2022/PaperInformation/StyleFiles> )
  - Abstract
    - El problema
    - ¿Por qué es interesante?
    - ¿Qué hicieron?
    - ¿Cuál fue el resultado?
  - Introducción
    - Describir en mayor detalle la parte del abstract
    - Trabajos relacionados
    - Metodología
      - Descripción de los métodos utilizados
      - Descripción de los datos/problema
      - Setup experimental
        - Equipo
        - Parámetros
        - Número de ejecuciones
  - Resultados y discusión
    - Gráfica de convergencia en los problemas
    - Tabla de resultados (calidad y tiempo)
    - Comparación de métodos
    - Discusión de observaciones relevantes de los resultados (¿funcionó?, ¿por qué?)
  - Conclusiones y trabajo futuro

# Proyecto

- (30 puntos) Presentación - 15 minutos y 5 de preguntas
  - Relevancia a ciencias de la computación
  - Relevancia del tema
  - Originalidad
  - Diapositivas
  - Presentación
  - Replicabilidad
  - Robustez
  - Preguntas
  - Confianza del revisor

## Para el día de hoy...

- Forma normal de juegos
- RL multi-agente
- Juegos de Markov





# Forma normal de un juego

- Un conjunto de jugadores/agentes  $\mathcal{I}$
- Un conjunto de acciones conjuntas  $a = (a_i), a_i \in \mathcal{A}$ , es la acción del agente  $i \in \mathcal{I}$
- Recompensa/pagos  $r_i(a)$  es la recompensa recibida por el agente  $i$  con la acción  $a$
- Cuando un juego en su forma normal se repite un número de veces (finito/infinito) se llama juego repetido

# Estrategias

- Estrategia/política:  $\pi_i \in \Delta(\mathcal{A}_i)$ :  $\pi_i(a_i)$  es la probabilidad que un agente  $i$  seleccione la acción  $a_i$ 
  - Pura (determinista): solo se juega una acción
  - Mixta (estocástica): una distribución sobre un conjunto de acciones
- Perfil: una estrategia para cada jugador  $\pi = (\pi_i)_i$
- Cada jugador desea maximizar su pago/recompensa
- El pago esperado de cada jugador  $i$  cuando se usa un perfil  $\pi$

$$r_i(\pi) = \sum_a r_i(a) \prod_{j \in \mathcal{I}} \pi_j(a_j)$$

## Un caso especial: juegos de dos jugadores

- El pago de juegos de dos jugadores puede ser representado con una matriz
- Dilema del prisionero: cada agente elige cooperar o acusar al otro

		Bob	
		cooperate	defect
Alex	cooperate	1, 1	-1, 2
	defect	2, -1	0, 0

# Estrategia dominante

- Una estrategia dominante  $\pi_i$  para un jugador  $i$  es una estrategia que es la mejor respuesta a todo  $\pi_{-i}$
- $r_i(\pi_i, \pi_{-i}) \geq r_i(\tilde{\pi}_i, \pi_{-i}), \forall \tilde{\pi}, \pi_{-i}$
- En un equilibrio, cada jugador adopta una estrategia dominante
- Es posible que no exista una estrategia dominante ni un equilibrio

		Bob	
		cooperate	defect
Alex	cooperate	1, 1	-1, 2
	defect	2, -1	0, 0



# Equilibrio de Nash

- En un equilibrio de Nash  $\pi^*$ , ningún jugador puede mejorar su recompensa esperada cambiando su política, si el resto mantiene la suya
- $\pi^*$  es la mejor respuesta para cada agente  $i$  si los otros agentes se quedan con  $\pi_{-1}^*$

- Para cada agente

$$r_i(\pi^*) \geq r_i(\pi_i, \pi_{-1}^*) \quad \forall \pi_i$$

- Toda estrategia dominante es un equilibrio de Nash

# Piedra-papel y tijera

- No existe una estrategia dominante
- Un equilibrio de Nash es cada jugador usar una estrategia mixta con  $\frac{1}{3}$  para cada opción
- Teorema: para un juego con jugadores y acciones finitas, existe un equilibrio de Nash con estrategia mixta

		Bob		
		rock	paper	scissor
Alex	rock	0, 0	-1, 1	1, -1
	paper	1, -1	0, 0	-1, 1
	scissor	-1, 1	1, -1	0, 0

## Un juego más...

- Cada jugador anota un número  $i \in [0,100]$
- El número que anoten es el número que consideren será  $\frac{2}{3}$  de la media de los valores que los jugadores adivinen
- El jugador que se encuentre más cerca, gana

# Juego de Markov

- Es una tupla  $G = (N, S, A, P, \{R_i\}_{i \in N}, \delta)$
- Donde
  - $N = \{1, \dots, n\}$  es un conjunto de jugadores
  - $S$  es el espacio de estados
  - $A = A_1 \times \dots \times A_n$  es el espacio de acciones donde  $A_i$  es el conjunto de acciones de  $i$
  - Para estados  $s \in S$  y  $a \in A$ ,  $P(\cdot, s, a)$  es la distribución de probabilidad  $P_{i,s \rightarrow s'}^a$
  - Para estados  $s \in S$  y  $a \in A$ ,  $R_i(s'|s, a)$  es la recompensa  $R_{i,s \rightarrow s'}^a$
  - $\delta \in (0,1)$  es un factor de descuento

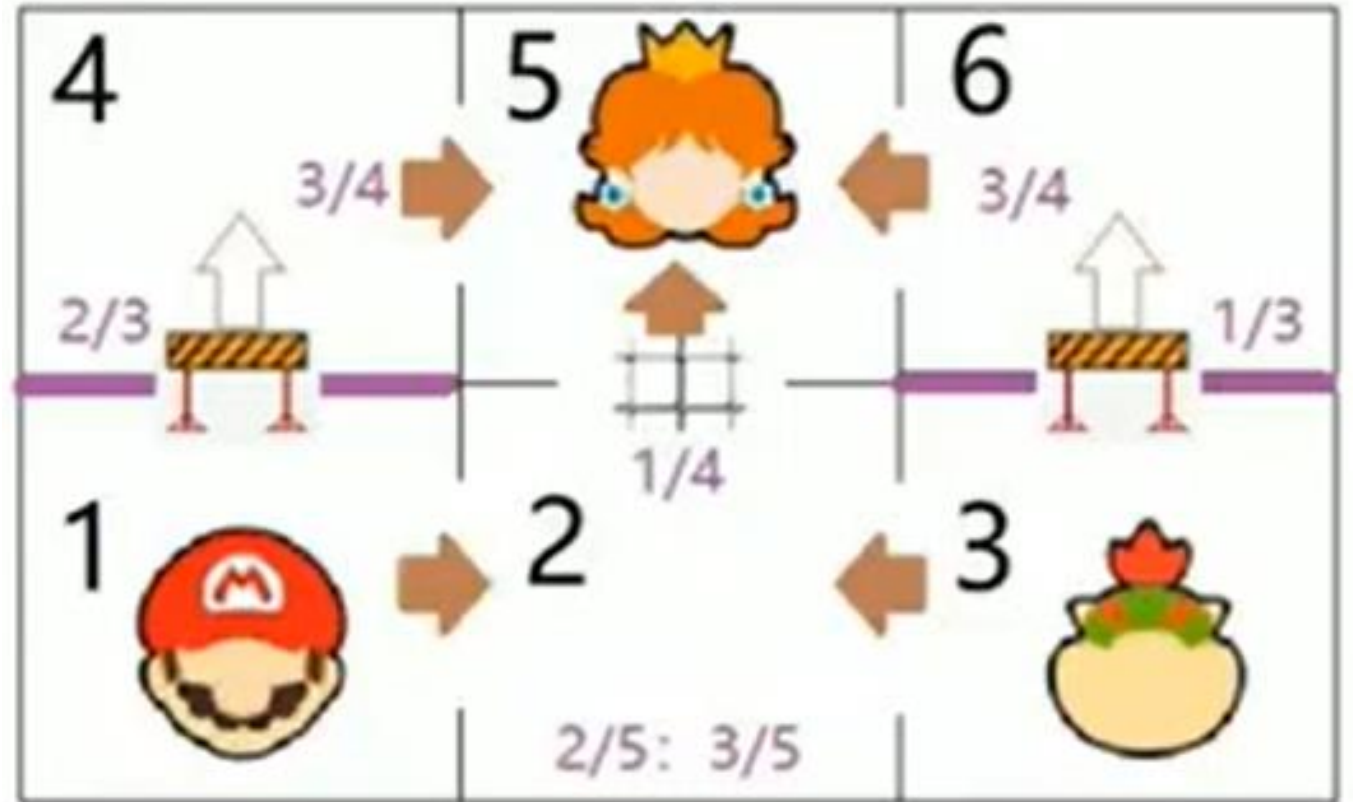


# El juego

- El juego inicia en  $s_1$
- En cualquier momento, el juego se encuentra en algún estado  $s_t$
- Cada jugador elige una acción simultáneamente  $a_t = (a_{t1}, \dots, a_{tN}) \in A$
- El siguiente estado es determinado por  $P(\cdot, s_t, a)$
- Los jugadores obtienen recompensas  $r_t = (r_{t1}, \dots, r_{tN})$  donde  $r_{ij} = R_i(s'|s, a)$  es el pago de  $i$
- En cada tiempo, todos los jugadores pueden observar la historia

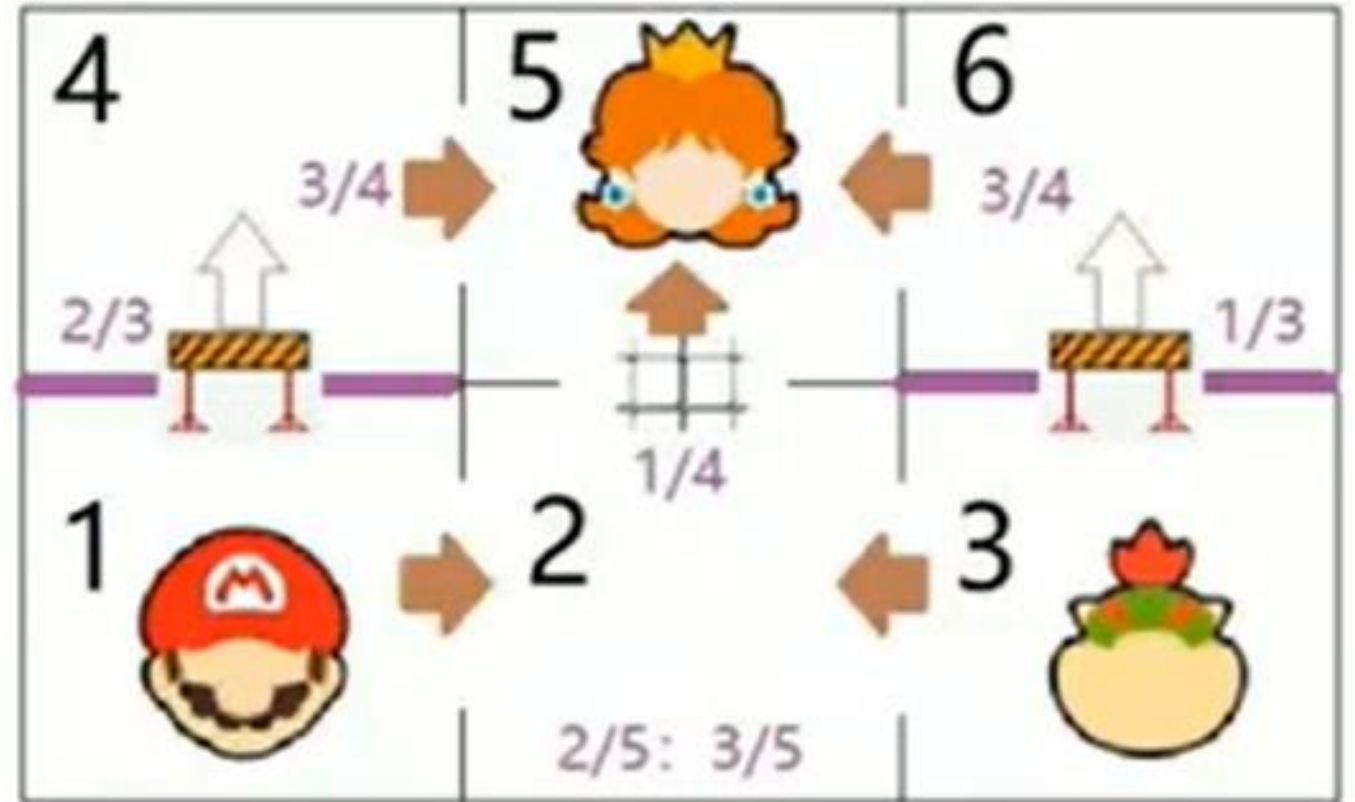
# Ejemplo: grid world

- Tenemos dos jugadores: Mario (M) y Bowser (B)
- Estados
  - M: {1,2,4,5}
  - B: {2,3,5,6}
- Cada celda solo puede tener un agente
- Quien llegue a la celda 5 gana
- Existen movimientos probabilistas
- Si ambos agentes desean entrar al mismo lugar necesitan pelear donde M tiene  $\frac{2}{5}$  de probabilidad de ganar



# Modelado como juego de Markov

- Estados:  $(\{1,2,4\} \times \{2,3,6\} \cup \{q_1, q_2\}) \setminus \{(2,2)\}$
- Acciones:  $A_1 = \{r, u\}, A_2 = \{l, u\}$



# Transiciones y recompensas

From state	Action	To state	Probability	payoff
$(1,3)$	$(u,u)$	$(1,3)$	$2/9$	$(0,0)$
$(1,3)$	$(u,u)$	$(4,3)$	$4/9$	$(0,0)$
$(1,3)$	$(u,u)$	$(1,6)$	$1/9$	$(0,0)$
$(1,3)$	$(u,u)$	$(4,6)$	$2/9$	$(0,0)$
$(1,3)$	$(r,l)$	$(1,2)$	$3/5$	$(-c,c)$
$(1,3)$	$(r,l)$	$(2,3)$	$2/5$	$(c,-c)$
$(1,3)$	$(u,l)$	$(1,2)$	$1/3$	$(0,0)$
$(1,3)$	$(u,l)$	$(4,2)$	$2/3$	$(0,0)$
$(1,3)$	$(r,u)$	$(2,3)$	$2/3$	$(0,0)$
$(1,3)$	$(r,u)$	$(2,6)$	$1/3$	$(0,0)$
$(4,6)$	$\cdot$	$q_2$	$3/5$	$(-1-c, 1+c)$
$(4,6)$	$\cdot$	$q_1$	$2/5$	$(1+c, -1-c)$
$(4,2)$	$\cdot$	$q_2$	$1/4$	$(-1, 1)$
$(4,2)$	$\cdot$	$q_1$	$3/4$	$(1, -1)$
$(2,6)$	$\cdot$	$q_2$	$3/4$	$(-1, 1)$
$(2,6)$	$\cdot$	$q_1$	$1/4$	$(1, -1)$
$(1,2), (1,6)$	$\cdot$	$q_2$	$1$	$(-1, 1)$
$(2,3), (4,3)$	$\cdot$	$q_1$	$1$	$(1, -1)$



# Utilidad

- Cada jugador recibe recompensa descontada de sus estrategias
- Para un jugador  $i$ , dada el estado inicial  $s \in S$ ,  $V_i^\pi(s)$  es la recompensa esperada para el jugador  $i$  para cualquier jugada iniciando en  $s$  y consistente con  $\pi$

$$V_i^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_{t+1,i} | s_0 = s \right]$$

# Proposición

- En un juego de Markov, supongamos que todos los oponentes del jugador  $i$  juegan una estrategia de Markov. Entonces, la mejor respuesta del jugador  $i$  es jugar una estrategia de Markov
- El valor de una estrategia se calcular de la misma forma que en MDPs

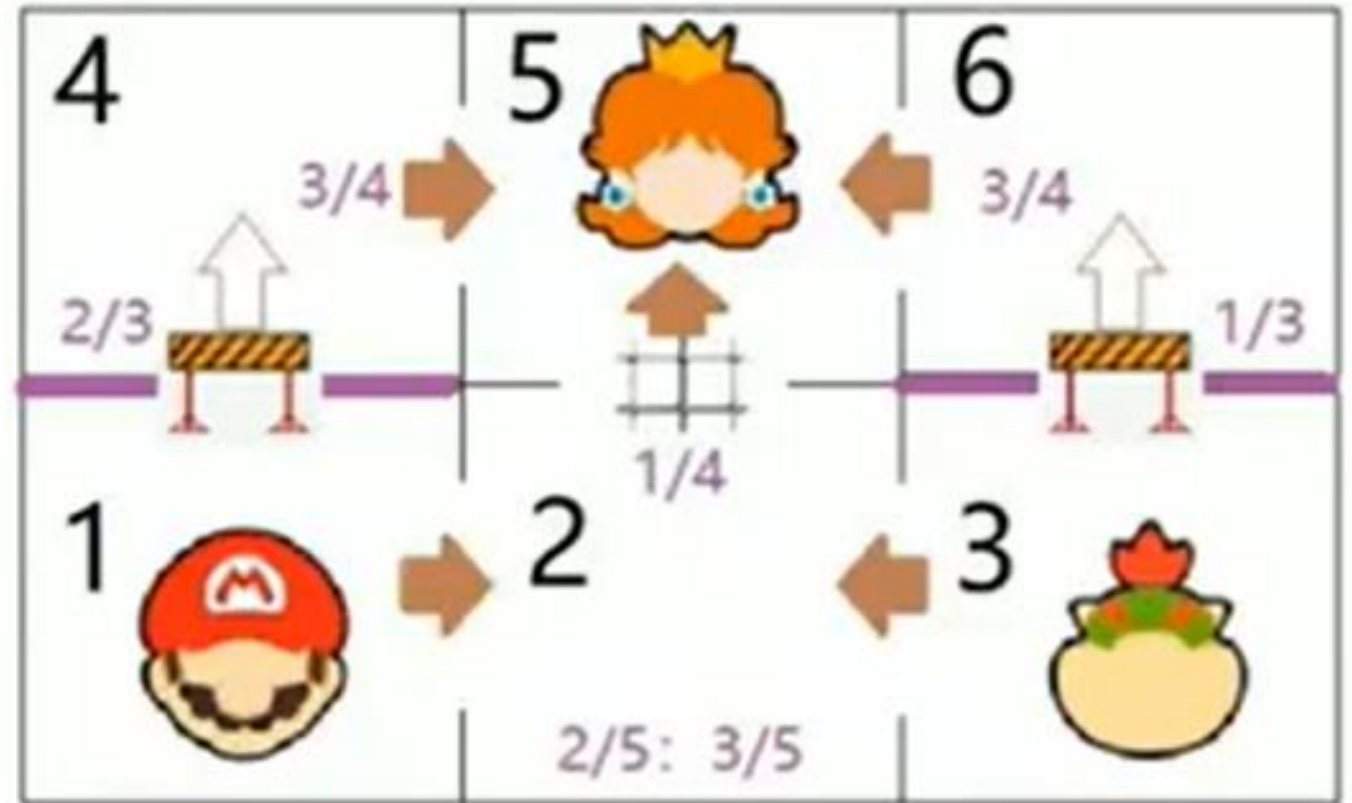
# Estrategia óptima para un solo jugador

- Idea: identificar la respuesta óptima del jugador  $i$  a la estrategia de Markov  $\bar{\pi}$  de los oponentes

$$V_i^*(s) = \max_{a \in A_i} \sum_{s' \in S} P_{s \rightarrow s'}^{a, \bar{\pi}(s)} \left( R_{s \rightarrow s'}^{a, \bar{\pi}(s)} + \gamma V_i^*(s') \right)$$

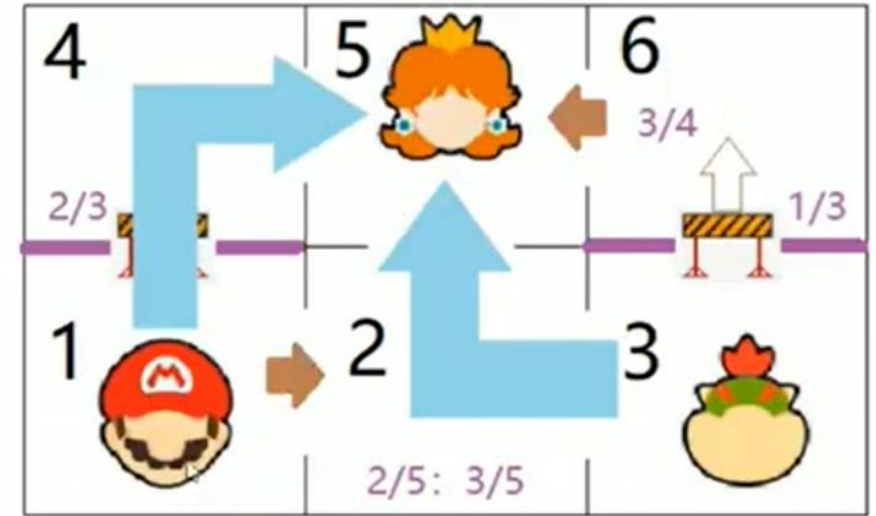
# Regresamos al ejemplo

- Solo importa la decisión en (1,3)
- Existen 4 estrategias de Markov
  - $\pi((1,3)) = (r, l)$
  - $\pi((1,3)) = (r, u)$
  - $\pi((1,3)) = (u, l)$
  - $\pi((1,3)) = (u, u)$





# El resultado




- Solo importa la decisión en (1,3)
- Existen 4 estrategias de Markov
  - $\pi((1,3)) = (r, l)$
  - $\pi((1,3)) = (r, u)$
  - $\pi((1,3)) = (u, l)$
  - $\pi((1,3)) = (u, u)$

Mario,Bowser	$l$	$u$
$r$	$\left(-\frac{\gamma+c}{5}, \frac{\gamma+c}{5}\right)$	$\left(\frac{\gamma}{2}, -\frac{\gamma}{2}\right)$
$u$	$(0, 0)$	$\left(\frac{13\gamma-2c\gamma}{45-10\gamma}, -\frac{13\gamma-2c\gamma}{45-10\gamma}\right)$

# Equilibrio perfecto de Markov

- Dado un juego de Markov, una estrategia  $\pi = (\pi_1, \dots, \pi_n)$  es un equilibrio perfecto de Markov si
  - Cada  $\pi_i$  es un estrategia de Markov
  - Cada  $s \in S, \pi(s) = (\pi_1(s), \dots, \pi_n(s))$  es un equilibrio de Nash para le juego que inicia en  $s$
- Dadas las misma condiciones del equilibrio de Nash se puede garantizar la existencia de este equilibrio



Un algoritmo  
para resolver el  
problema para  
juegos de suma  
cero de dos  
jugadores

---

**ShapleyValueIteration**( $S, A, P, R, \gamma$ )

INPUT: A 2-player zero-sum Markov game  $(S, A, P, R, \gamma)$

OUTPUT: Optimal strategy profile  $\pi^* = (\pi_1^*, \pi_2^*)$

$\forall s \in S$ : Set  $V(s) \leftarrow 0$

**repeat**

**for**  $s \in S$  **do**

$T(s) \leftarrow \sum_{s' \in S} P_{s \rightarrow s'}^{(a_1, a_2)} (R_{1, s \rightarrow s'}^{(a_1, a_2)} + \gamma V(s'))$

$V'(s) \leftarrow \text{val}(T(s))$

**end for**

  Set  $V \leftarrow V'$

**until**  $V$  converges

**for**  $s \in S$  **do**

  Set  $T(s) \leftarrow \sum_{s' \in S} P_{s \rightarrow s'}^{(a_1, a_2)} (R_{1, s \rightarrow s'}^{(a_1, a_2)} + \gamma V(s'))$

  Find  $\alpha_1 \in \Delta(A_1), \alpha_2 \in \Delta(A_2)$  such that  $\alpha_1^T T(s) \alpha_2 = \text{val}(T(s))$

  Set  $\pi_1^*(s) = \alpha_1$  and  $\pi_2^*(s) = \alpha_2$

**end for**

**return**  $\pi^*$



# Para la otra vez...

- RL multi-agente II



The End.



iimas