1. Recap (MDP, V^*, Iteración de Valor)

2. Iteración de Política

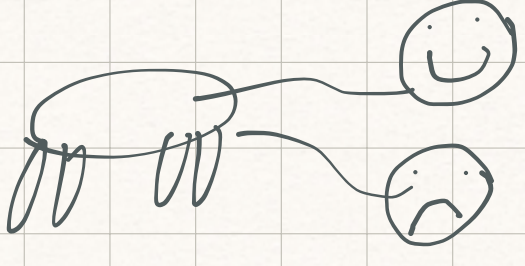
3. MDP Estados Continuos

* Tarea 1 (PDF)

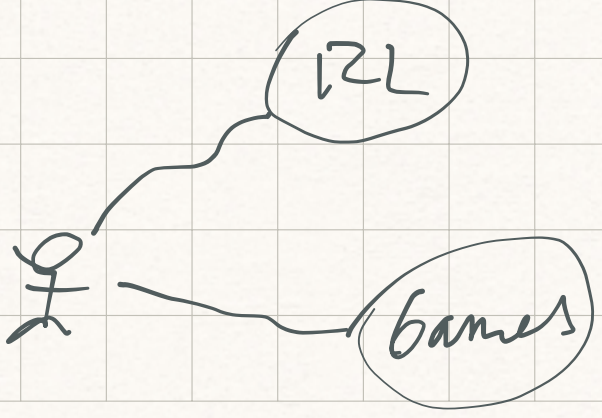
Recap

RL → AA, helicópteros, carritos con pole

Dinámica → Control RL



* alphaGo
* ajedrez
* alphaStar



Función de Recompensas

$$R(s) = +1 \quad \text{(para ganar)}$$
$$-1 \quad \text{(para perder)}$$
$$0 \quad \text{(el resto)}$$

recomp. (rewards)

estados

$R(3,2) = 0$

$R(4,3) = +1$

MDP

$$MDP\{S, A, P_{sa}, \gamma, R\}$$

$S \rightarrow$ set de estados

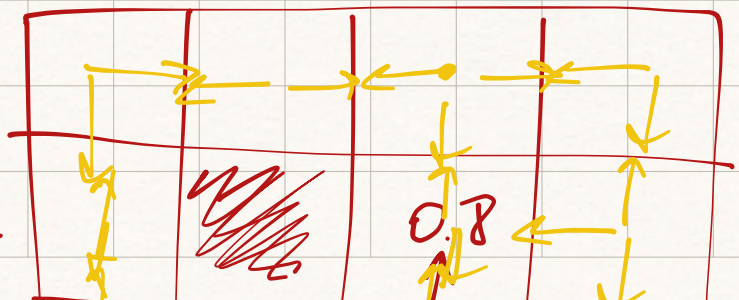
$A \rightarrow$ set de acciones

$P_{sa} \rightarrow$ Prob. de transición $\quad P_{sa}(s')$

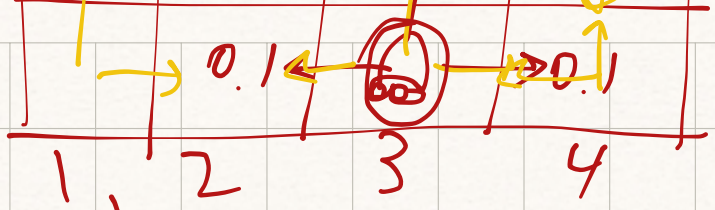
$\gamma \rightarrow$ factor de descuento $\quad \gamma \in [0,1)$

$\gamma = 0.9$, 0.99, 0.8

$R \rightarrow$ función de recompensas



recomp. inmediata vs recomp.



futuro

$R_0 + \gamma^1 R_1 + \gamma^2 R_2 + \dots$

S) a) 11 estados (pos.)

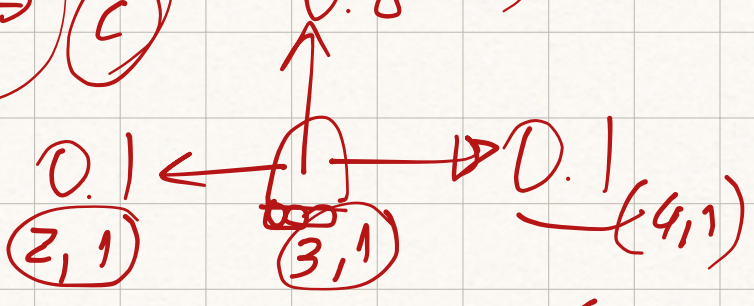
A) b) Acciones $\{S, N, E, O\}$ 4-dim

$S = \{-, -, -, -, \dots, -\}$

$S =$ "↓" $\quad O =$ "←"

$N =$ "↑" $\quad E =$ "→"

P) c)



$$P_{(3,1)N}((3,2)) = 0.8$$

$$P_{(3,1)E}((4,1)) = 0.1$$

$$P_{(3,1)O}((2,1)) = 0.1$$

$$P_{(3,1)}(3,3) = 0$$

d) $R((4,3)) = +1$

$R((4,2)) = 0.9$

$R(s) = 0 \quad \forall s$

Algoritmo!

s_0 (despertó)

$\downarrow$

a_0

$\downarrow$

$s_1 \sim P_{s_0 a_0}$

$\downarrow$

a_1

$\downarrow$

$s_2 \sim P_{s_1 a_1}$

Cuál es el pago total? (ganancia)

$$\Rightarrow \underline{R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots + \gamma^n R(s_n)}$$

$\gamma \cong 0.9$

Objetivo de un algo de RL

Arrojar la acción a través del tiempo para maximizar la sumatoria de recom. esperada

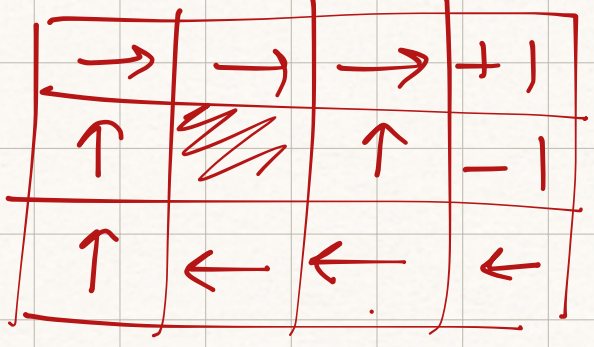
$$V = \max_{a}\left(E\left[R(S_0) + \gamma^1 R(S_1) + \gamma^2 R(S_2) + \dots\right]\Big|\pi\right)$$

Política (π) mapea acción para cada estado

$$\pi : S \rightarrow A$$

(Controlador)

Política dada π (π^*)



$$\pi(3,1) = \text{"}\leftarrow\text{"}$$

Estando en el estado s, tome acción $\pi(s)$

Obj. → Maximizar el total de recompensa esperada.

Conclusiones

MDP → formula el problema

Algo. de RL → Encontrar la política π que maximize de la ganancia esperada