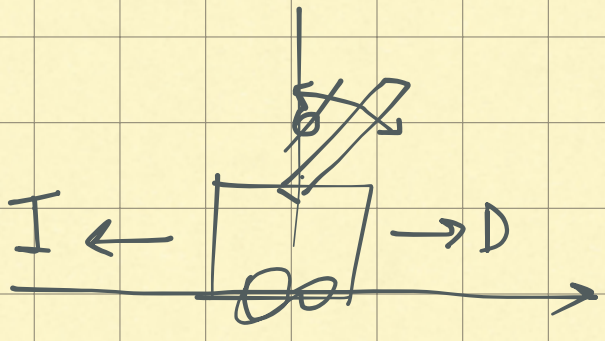
Clase 5.1 "Policy Search", POMDP

$$V^* \longrightarrow \pi^*$$

Búsqueda de Política Directa

Caso de uso "CartPole"



$$A = \{L, R\}$$

$$S = \begin{pmatrix} 1 \\ x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{pmatrix}$$

Paso 1: Encontrar el conjunto de funciones para obtener una π aproximada.

Reg. Log.

$$y \in \{0, 1\} \qquad h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Buscar la π Directamente

a) $\pi_\theta(s)$ — depende de un conjunto de estados

dado un conj. de parámetros

$$\pi_\theta(s) = \frac{1}{1 + e^{-\theta^T s}}$$

y $s = \begin{pmatrix} 1 \\ x \\ x \\ \emptyset \\ \emptyset \end{pmatrix}$

Paso: Escoger una Política (Estocástica)

función $\pi : S \times A \to \mathbb{R}$

$\pi(s, a)$

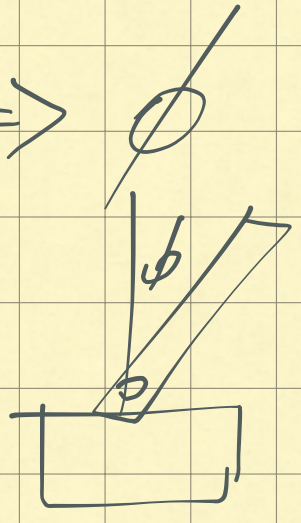
Ejemplo $\pi_\theta(s, \text{"D"}) = \dfrac{1}{1 + e^{-\theta^T s}}$

$$\pi_\theta(s, \text{"I"}) = 1 - \frac{1}{1+e^{-\theta^T s}}$$

$\varepsilon = 0.3$

ε-greedy $\begin{cases} 30\% \text{ aleatorio} \\ 70\% \text{ greedy} \end{cases}$

$$s = \begin{pmatrix} x \\ \dot{x} \\ \phi \\ \dot{\phi} \end{pmatrix}, \quad \theta = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \phi$$



En este caso, $\theta^T = \phi$

$$\pi_\theta(s, \text{"D"}) = \frac{1}{1+e^{-\phi}}$$

CartPole

$\phi \neq 10$

Un poco más complejo:

$$\dot{x} \quad \theta = \begin{bmatrix} 0 \\ -0.5 \\ 0 \end{bmatrix}$$

$$\phi \qquad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Cuanto debe acelerar a la derecha, dependiendo de estos 2 parámetros

(pos, vel. lineal, ángulo, vel. ángulo)

Policy Search

→ El Objetivo es encontrar este conjunto de parámetros θ

→ Para ejecutar $\pi_\theta(s, a)$

→ Maximicemos la ganancia

$$\max_{\theta} E\left[R(s_0, a_0) + \dots + R(s_t, a_t) \mid \pi_\theta\right]$$

Encontrar θ:

- Horizonte Finito T
- Estado Inicial S_0

$$\max_{\theta} E\left[R(S_0,a_0)+R(S_1,a_1)+\dots+R(S_t,a_t)\,|\,\pi_\theta\right]$$

$T=1$

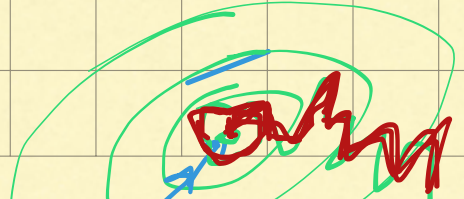
$$=\sum P(S_0,a_0,S_1,a_1)\left[R(S_0,a_0)+R(S_1,a_1)\right]$$

ganancia

$$=\sum P(S_0)\,\pi_\theta(S_0,a_0)\,P_{S_0,a_0}(S_1)\,\pi_\theta(S_1,a_1)\,[g]$$

↳ Derivar en fn de θ, usando algo. de gradientes ascendentes (estocásticos)

$0/0$



Algoritmo Reinforce (Reinforce Algorithm)

→ Monte Carlo Policy Gradient

{

Sample $S_0, a_0, S_1, a_1, \ldots$ (T = 1)

Compute $R(S_0) + R(S_1)$

Update

$$\theta := \theta + \alpha \left[\frac{\nabla_\theta \pi_\theta(S_0, a_0)}{\pi_\theta(S_0, a_0)} + \frac{\nabla_\theta \pi_\theta(S_1, a_1)}{\pi_\theta(S_1, a_1)} \right] g$$

(α: tasa de aprendizaje)

}

En cada iteración, actualizamos θ

Entonces, El algoritmo Reinforce

- las actualizaciones tienen aleatoriedad
- depende de la secuencia de estados
- Hasta que logre el Target

$$\nabla_\theta \left[\sum_{s_0, a_0, s_1, a_1} P(s_0)\, \pi_\theta(s_0, a_0)\, P_{s_0 a_0}(s_1)\, \pi_\theta(s_1, a_1) \right] (g)$$

Aplicamos "Regla del Producto"

$$\frac{d}{d\theta} f(\theta) g(\theta) h(\theta)$$

$$= f'(\theta) g(\theta) h(\theta) + f(\theta) g'(\theta) h(\theta) + f(\theta) g(\theta) h'(\theta)$$

$$\sum \left[P(s_0)\, \pi_\theta(s_0, a_0)\, \nabla_\theta \pi_\theta(s_0, a_0)\, P(s_1)\, \pi_\theta(s_1, a_1) \right.$$

$$\frac{\nabla_\theta \pi_\theta(s_0, a_0)}{\pi_\theta(s_0, a_0)}$$

$$+ P(s_0)\pi_\theta(s_0, a_0) P(s_1)\pi_\theta(s_1, a_1) \frac{\nabla_\theta \pi_\theta(s_1, a_1)}{\pi_\theta(s_1, a_\cdot)} \Big]$$

factorizando

$$= \sum P(s_0, a_0, s_1, a_1) \left[\frac{\nabla_\theta \pi_\theta(s_0, a_0)}{\pi_\theta(s_0, a_0)} + \frac{\nabla_\theta \pi_\theta(s_1, a_1)}{\pi_\theta(s_1, a_\cdot)} \right] \cdot g$$

Gradient Update

Conclusiones:

POMDP

$$S = \begin{pmatrix} \not{x} \\ x \\ \not{\phi} \\ \not{\phi} \end{pmatrix} \begin{matrix} \leftarrow \text{sensor} \\ \\ \leftarrow \text{sensor} \\ \end{matrix}$$

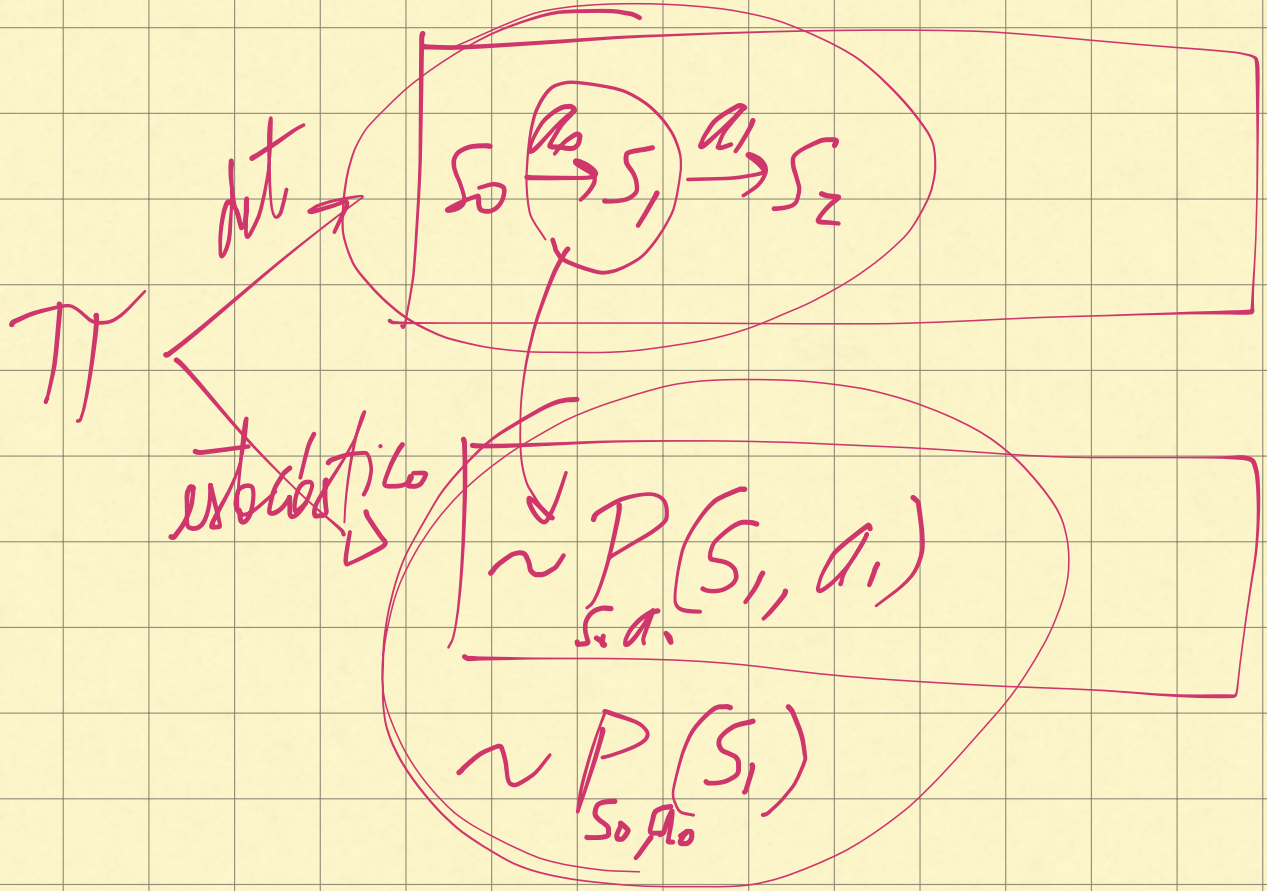
- Siempre tengo una medición / observación parcial, ruidosa del estado

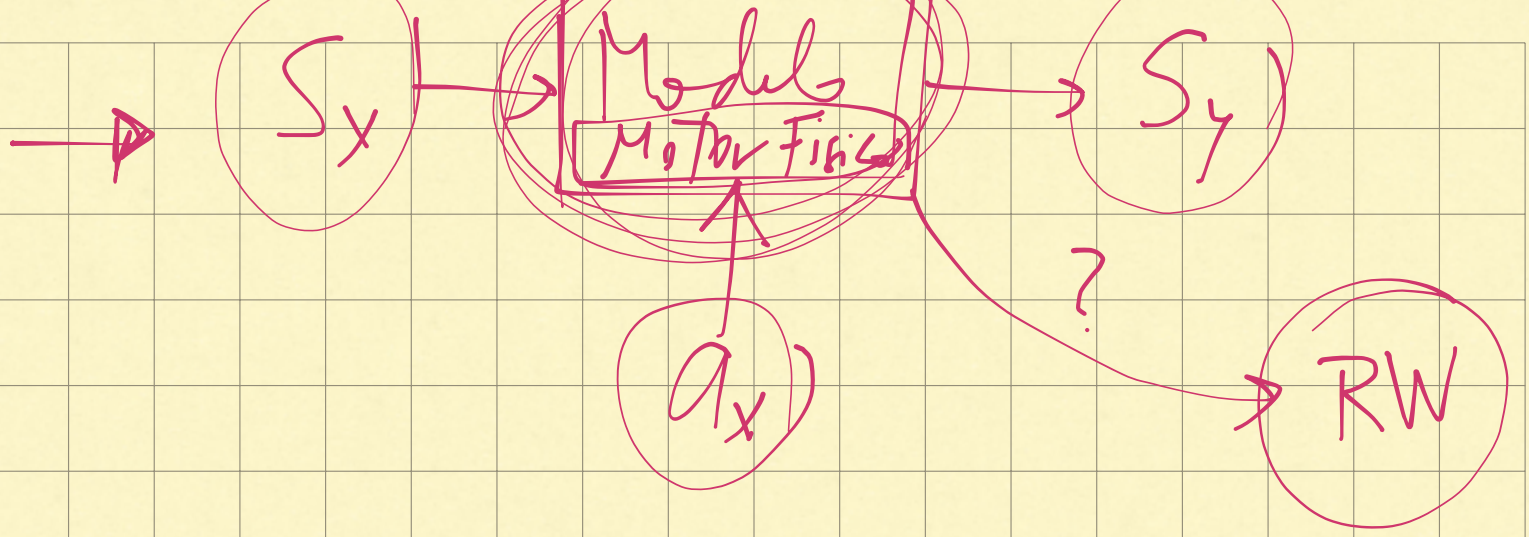
- Generalización

- Funciona muy bien POMDP

$$y = \begin{pmatrix} x \\ \phi \end{pmatrix} + \text{ruido}$$

$$\pi_\theta(y, \text{"1"}) = \frac{1}{1 + e^{-\theta^T y}}$$




S_x
Module
Matériel
S_y
q_x
?
RW