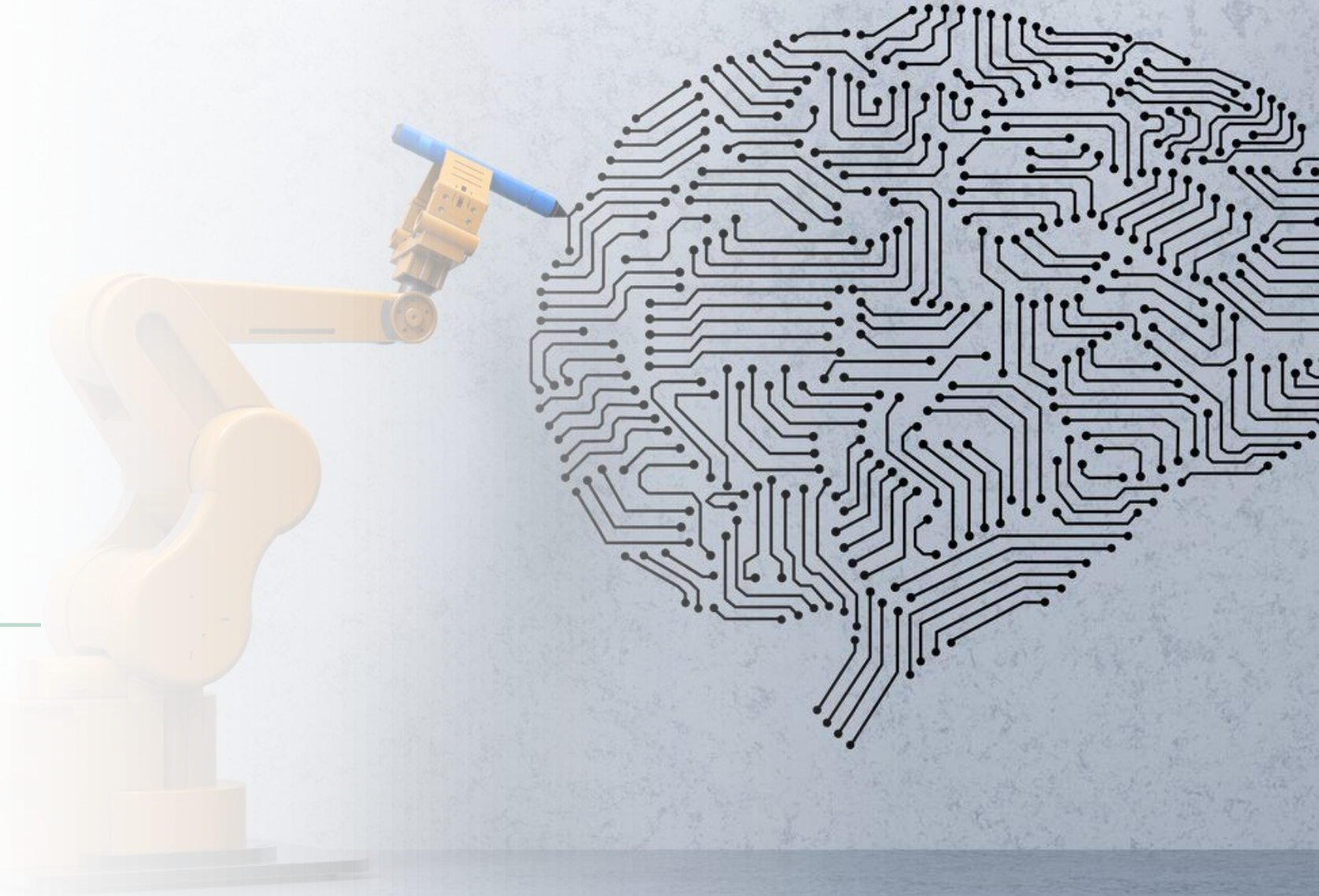


Aprendizaje por refuerzo

Clase 24: meta RL



Para el día de hoy...

- Meta RL



Meta aprendizaje

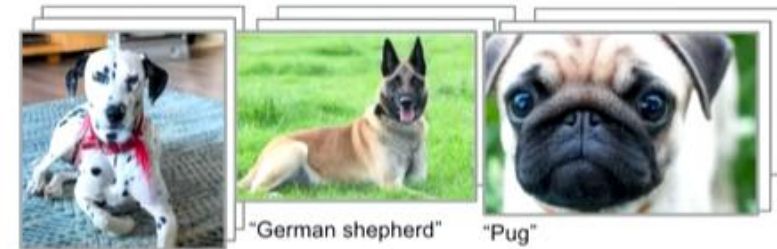
- Los métodos que hemos revisados se hiperespecializan

reinforcement learning



Robot art by Matt Spangler, mattspangler.com

supervised learning



"Dalmation"

"German shepherd"

"Pug"



corgi



???

El problema

- Aprender una regla de adaptación
- $\theta^* = \arg \max_{\theta} \sum_{i=1}^n \mathbb{E}_{\pi_{\phi_i}(\tau)} [R(\tau)]$
- Donde: $\phi_i = f_{\theta}(\mathcal{M}_i)$



\mathcal{M}_1



\mathcal{M}_2



\mathcal{M}_3

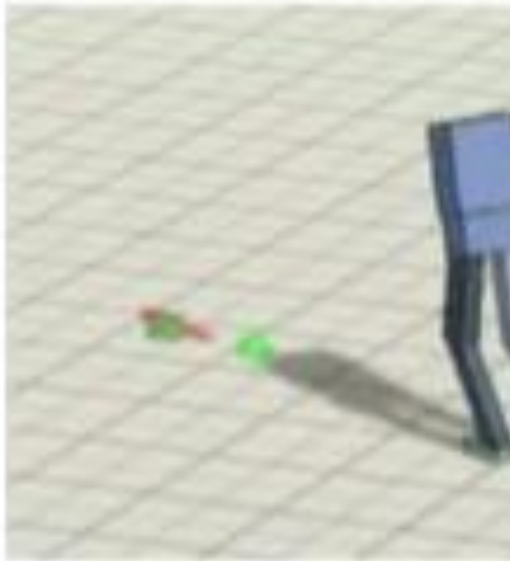


\mathcal{M}_{test}

Relación con contextos



ω : stack location



ω : walking d

- Meta RL puede ser visto como política contextual donde la información de la tarea se infiere de la experiencia
- La información de la tarea puede ser acerca de la dinámica o la función de recompensa
- Las recompensas son una generalización de las metas

Adaptación



Explorar: recolectar tanta información como sea posible



Adaptar: utilizar los datos para obtener la política óptima

Algoritmo general

- En entrenamiento
 - Muestrar de una tarea i , recolectar datos D_i
 - Adaptar la política calculando $\phi_i = f(\theta, D_i)$
 - Recolectar datos D'_i con la política adaptada π_{ϕ_i}
 - Actualizar θ de acuerdo a $\mathcal{L}(D'_i, \phi_i)$



Algunos algoritmos

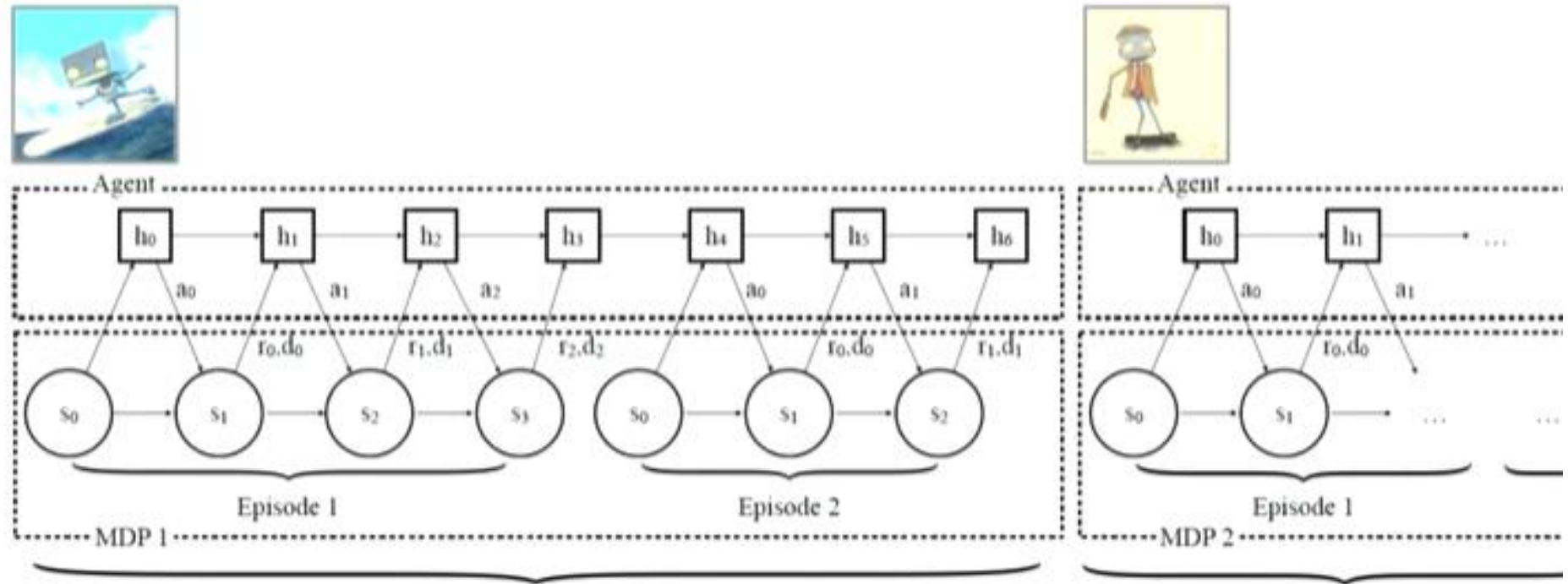
Solución 1: recurrencia

- Duan et al. 2016, Wang et al. 2016, Heess et al. 2016
- Implementar una política como una red recurrente
- Entrenar en un conjunto de tareas

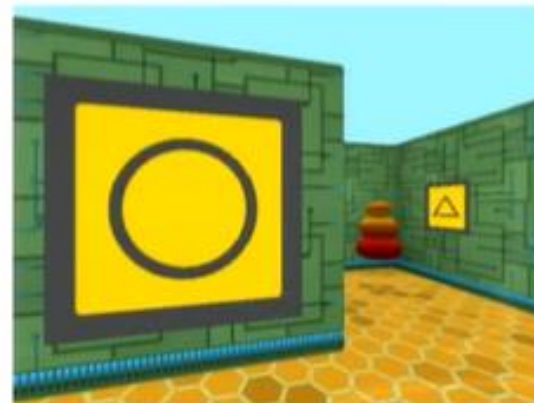
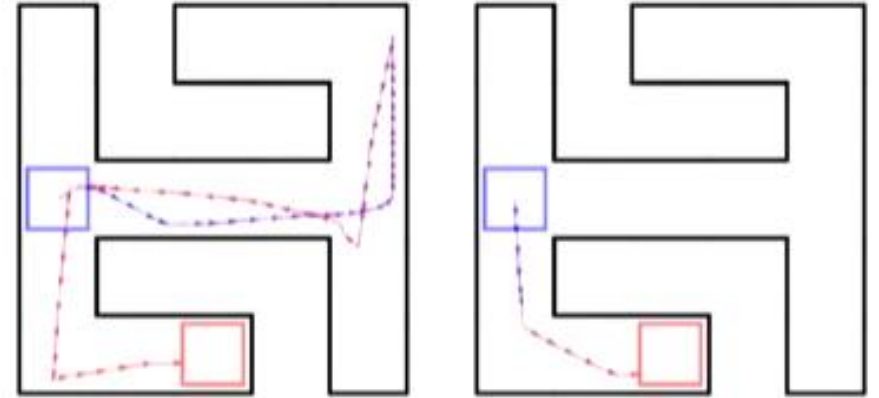
$\theta^* = \arg \max_{\theta} \sum_{i=1}^n E_{\pi_{\phi_i}(\tau)} [R(\tau)]$

PG where $\phi_i = f_{\theta}(\mathcal{M}_i)$

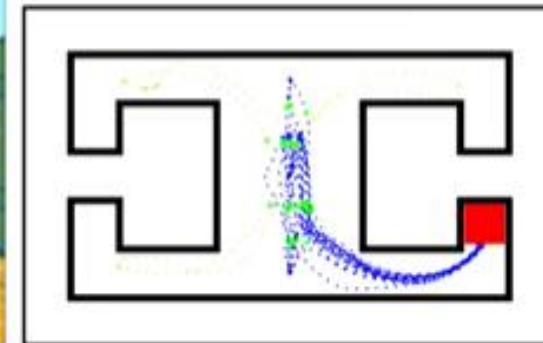
RNN



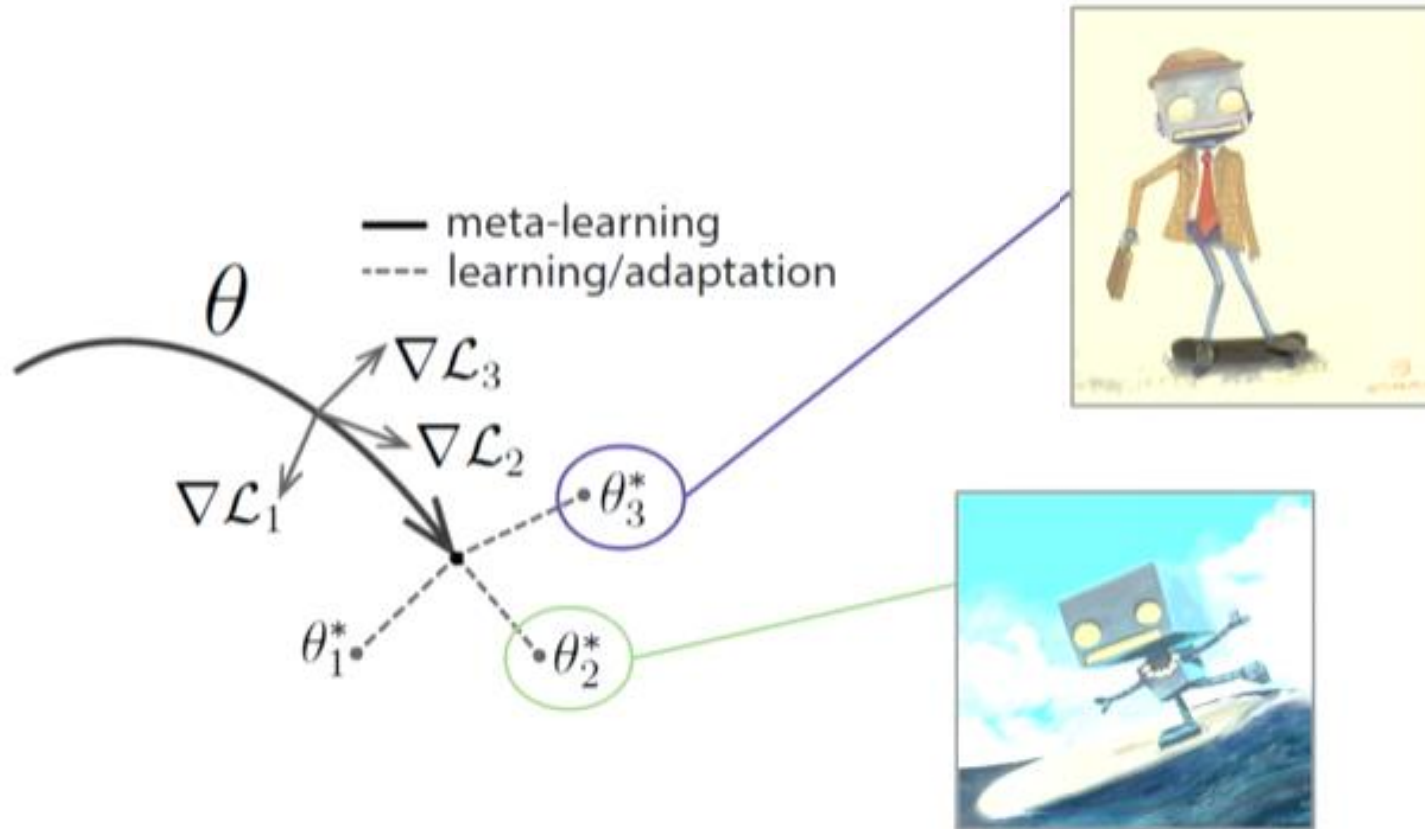
Un ejemplo



(a) Labryinth I-maze



(b) Illustrative Episode



Solución 2: optimización

Aprender una inicialización de parámetros para la cual un ajuste fino funcione

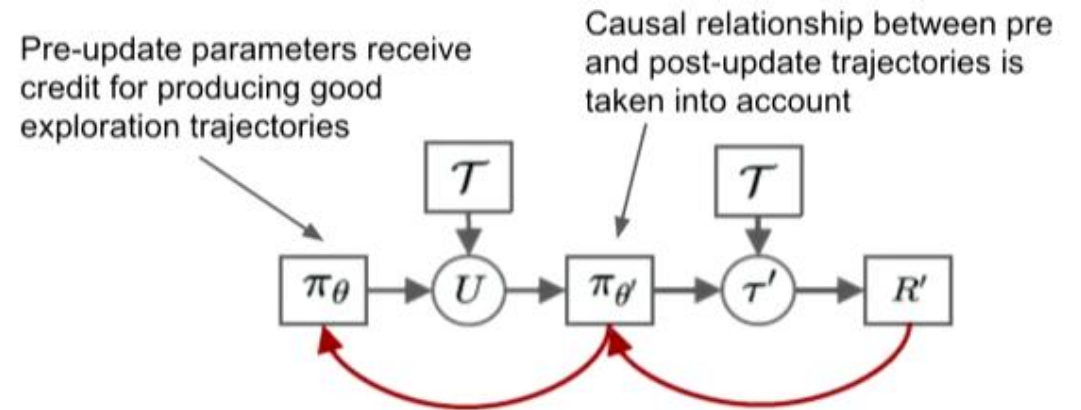
Algoritmo

while training:

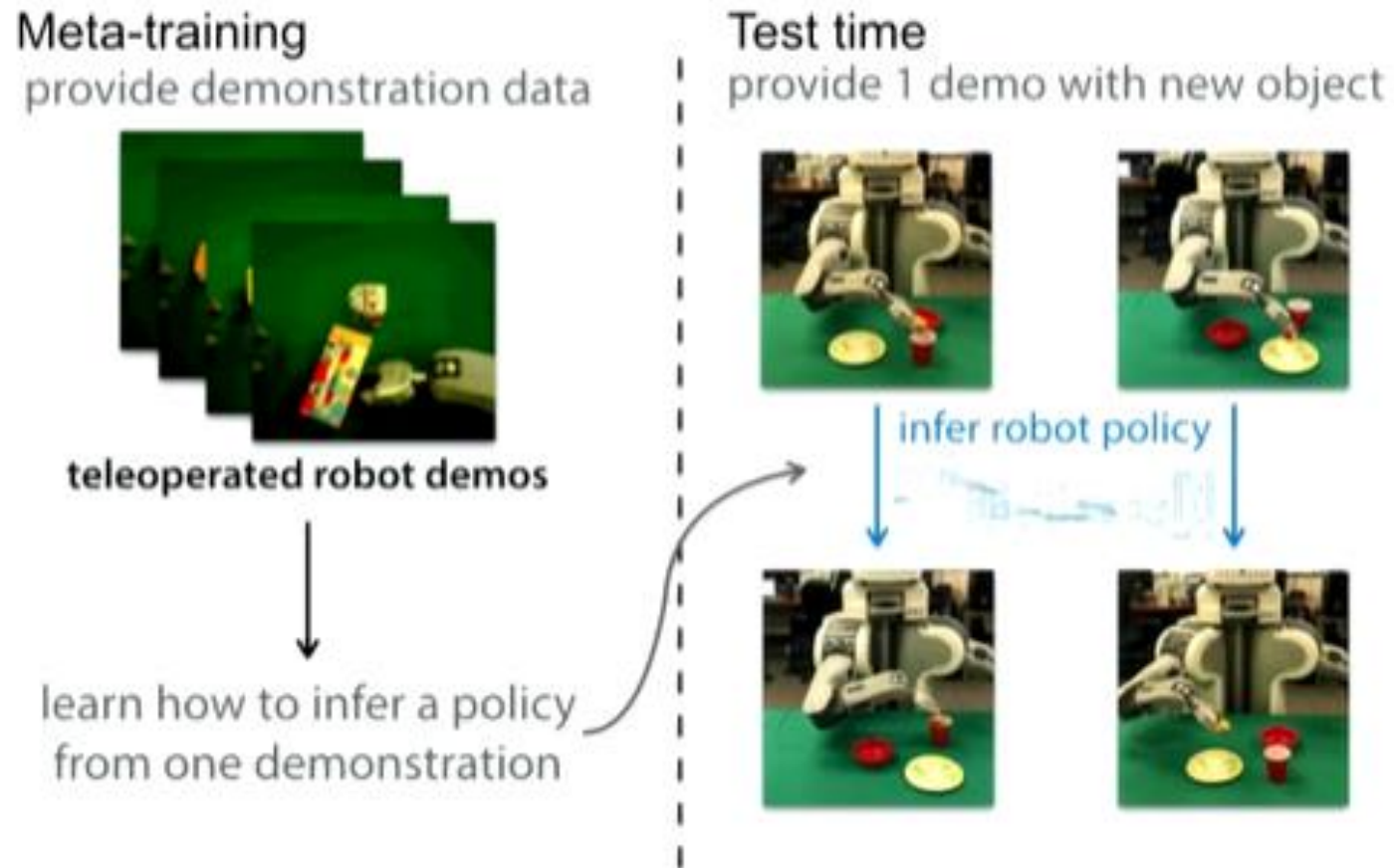
 for i in tasks:

1. sample k episodes $\mathcal{D}_i = \{(s, a, s', r)\}_{1:k}$ from π_θ
2. compute adapted parameters $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}_i(\pi_\theta, \mathcal{D}_i)$
3. sample k episodes $\mathcal{D}'_i = \{(s, a, s', r)_{1:k}\}$ from π_ϕ

update policy parameters $\theta \leftarrow \theta - \nabla_\theta \sum_i \mathcal{L}_i(\mathcal{D}'_i, \pi_{\phi_i})$



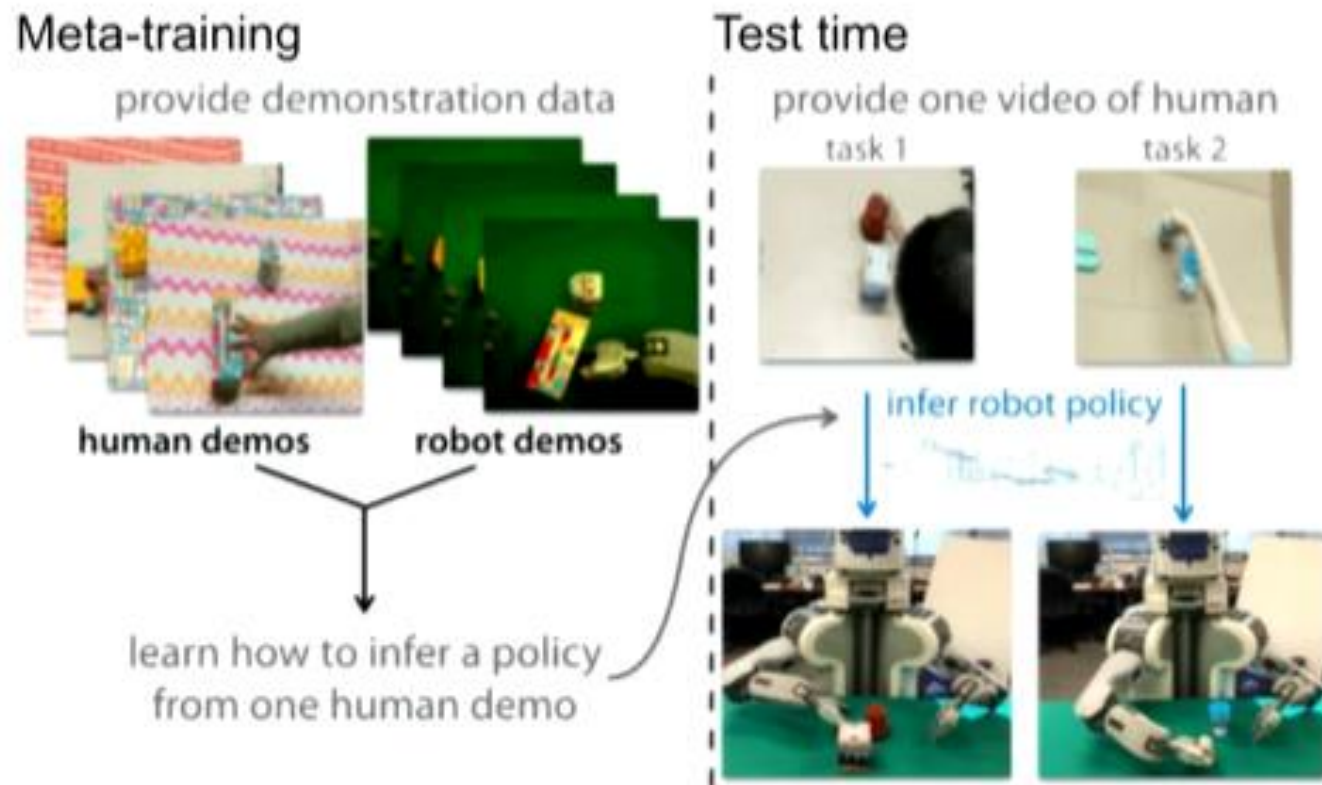
Meta RL en robótica con demo de robot



- Tarea: realizar una tarea dada una sola demostración al robot
- Entrenamiento: ejecutar clonación del comportamiento para adaptación

$$\phi_i = \theta - \alpha \nabla_{\theta} \sum_t ||\pi_{\theta}(o_t) - a_t^*||^2$$

Meta RL en robótica con demo de humano



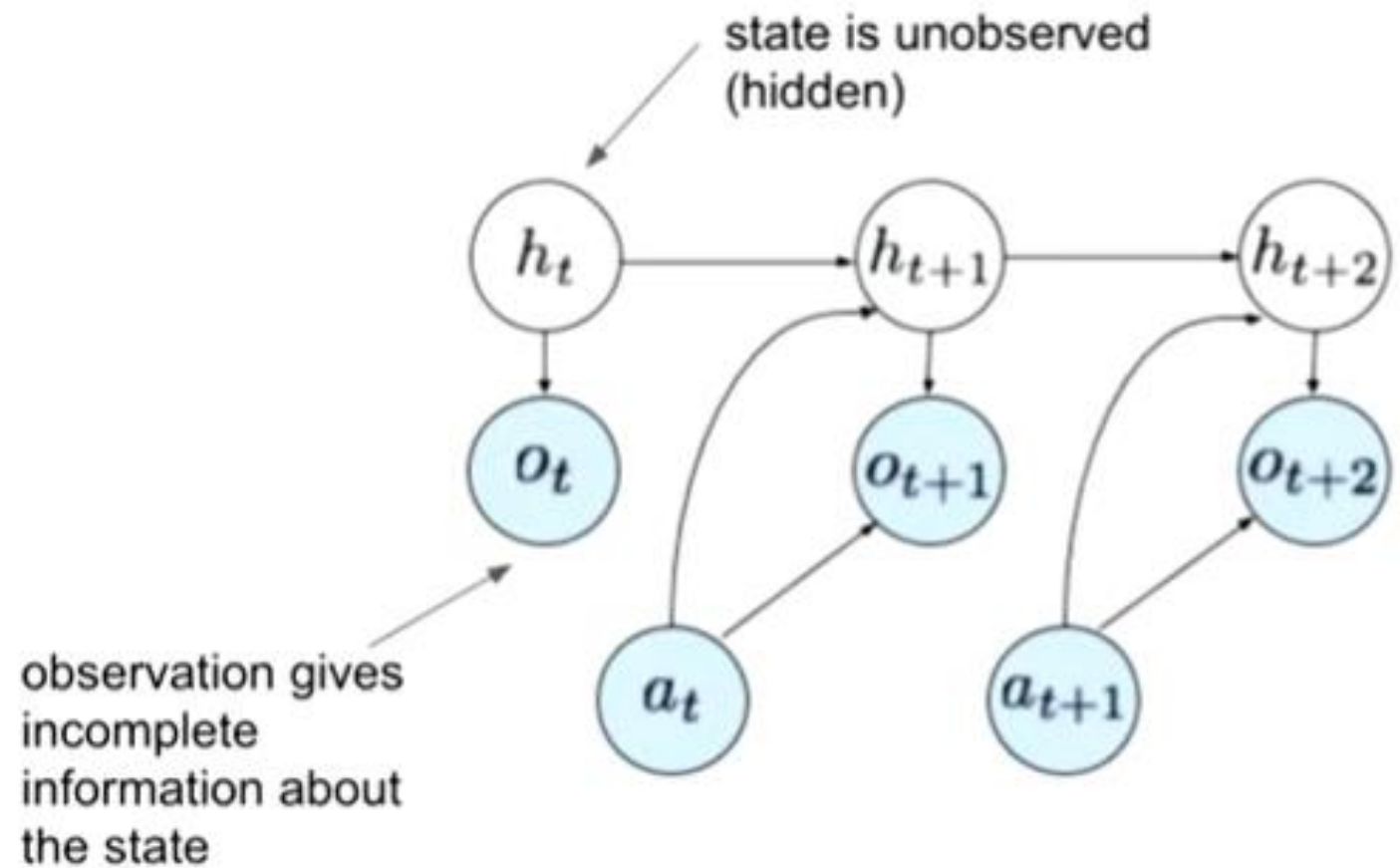
- Tarea: realizar una tarea dada una sola demostración de un humano

$$\phi = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\psi}(\theta, \mathbf{d}^h)$$

- Entrenamiento: aprender una función de pérdida que adapte la política

$$\phi = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\psi}(\theta, \mathbf{d}^h)$$

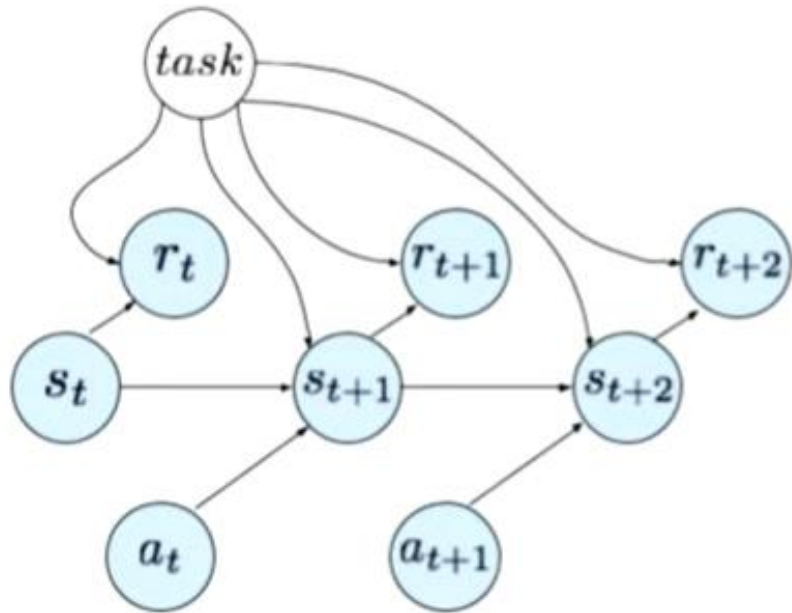
POMDPs



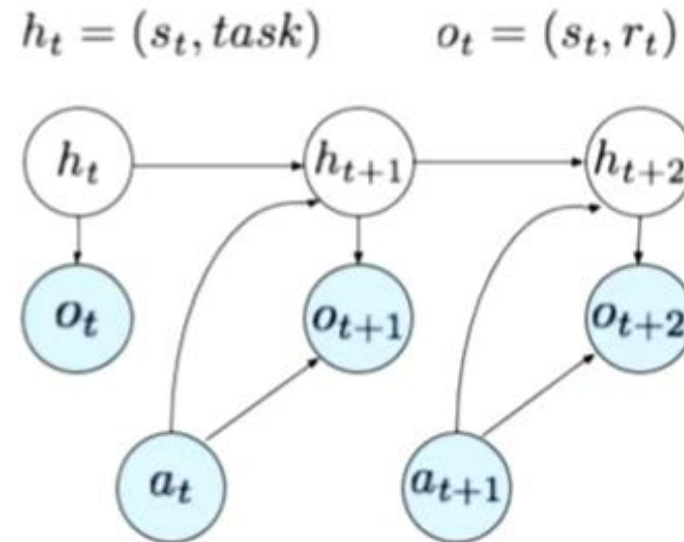
POMDP como meta RL

- Enfoques
 - Política con memoria
 - Estimación explícita del estado

meta-RL...

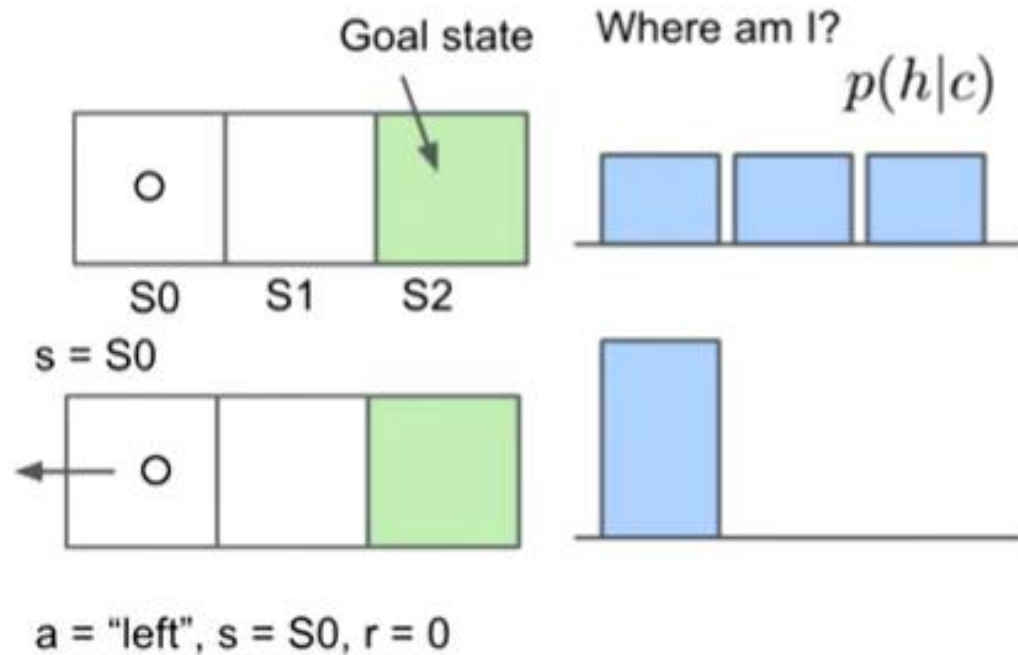


...as a POMDP

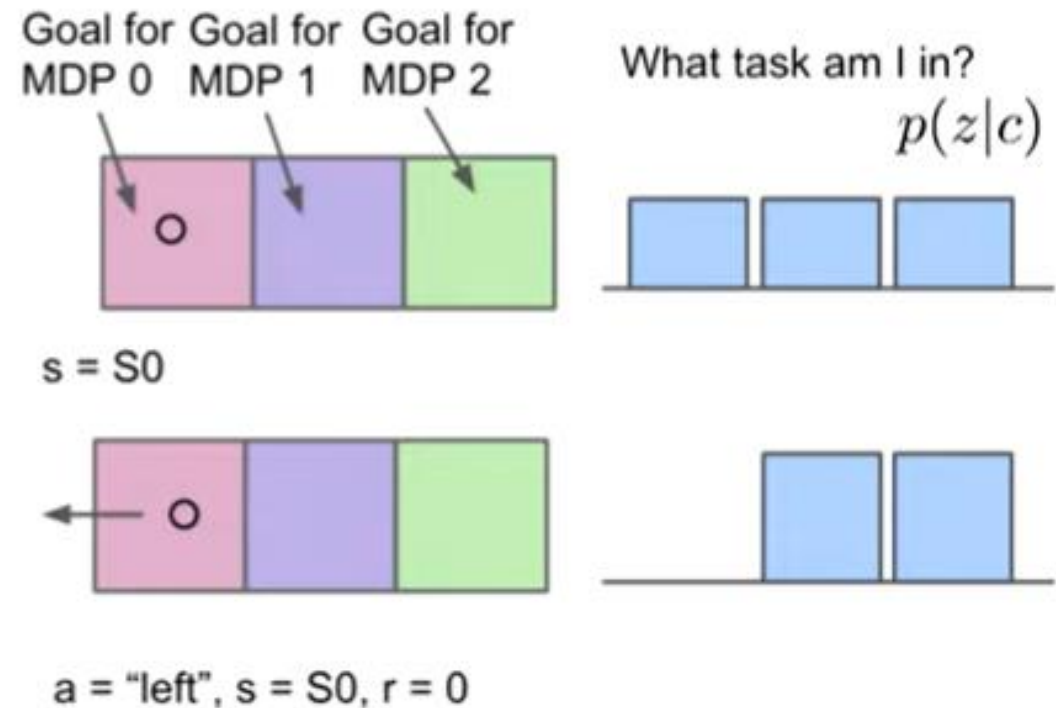


Modelo de creencias

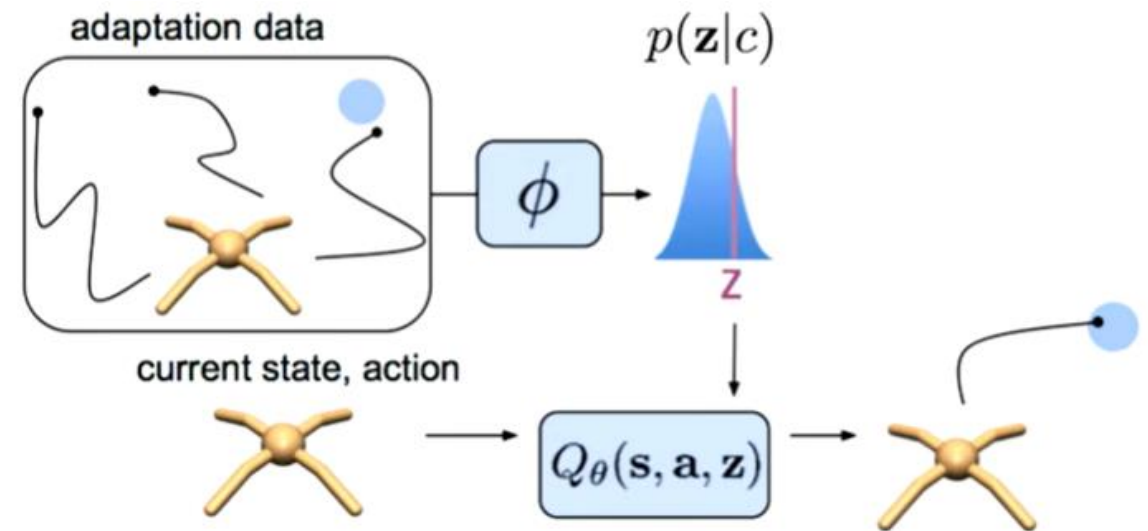
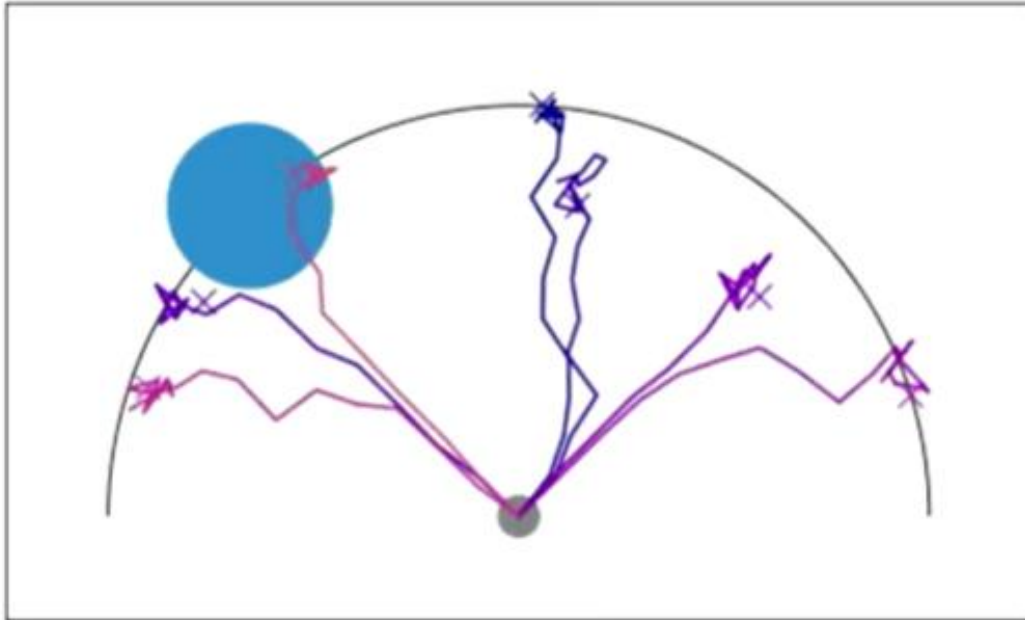
POMDP for unobserved state



POMDP for unobserved task



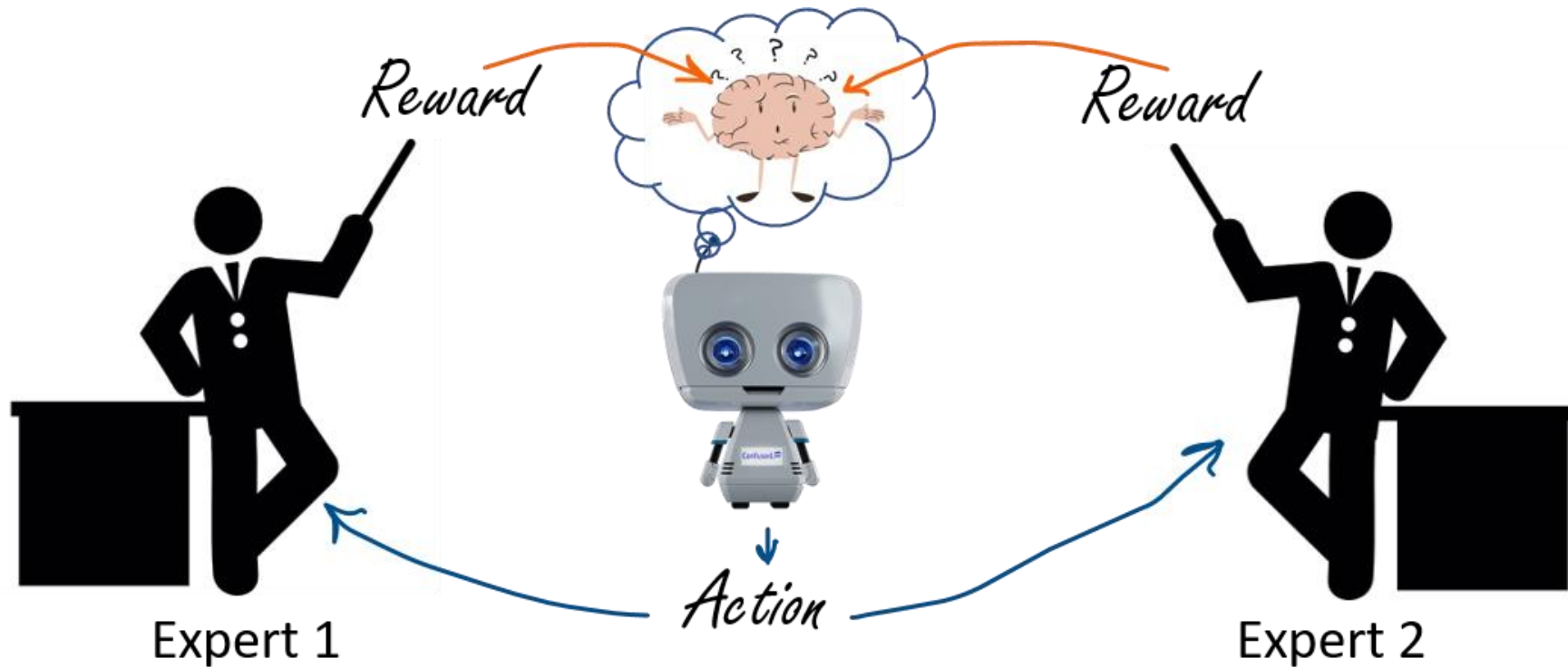
Estados de creencia de tareas




Comentarios

- Meta RL encuentra un procedimiento de adaptación que puede adaptar rápidamente la política a una nueva tarea
- Existen tres clases de soluciones: RNN, optimización y creencia de tareas.
- Existe conexión entre contextos y POMDPs
- Problemas abiertos: mejor exploración, definir distribución de tareas, meta aprendizaje en línea

RL multi-recompensa





“Toda meta puede ser descrita
como la maximización de
recompensas esperadas”

Racionalidad

- Teorema [Ramsey, 1931; von Neumann & Morgenstern, 1944]
- Dadas preferencias que satisfagan los axiomas, existe una función real U tal que:

$$U(A) \geq U(B) \leftrightarrow A \succsim B$$

$$U([p_1, S_1; \dots; p_n, S_n]) = \sum_i p_i U(S_i)$$

The Axioms of Rationality

Orderability

$$(A \succ B) \vee (B \succ A) \vee (A \sim B)$$

Transitivity

$$(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$$

Continuity

$$A \succ B \succ C \Rightarrow \exists p [p, A; 1 - p, C] \sim B$$

Substitutability

$$A \sim B \Rightarrow [p, A; 1 - p, C] \sim [p, B; 1 - p, C]$$

Monotonicity

$$A \succ B \Rightarrow$$

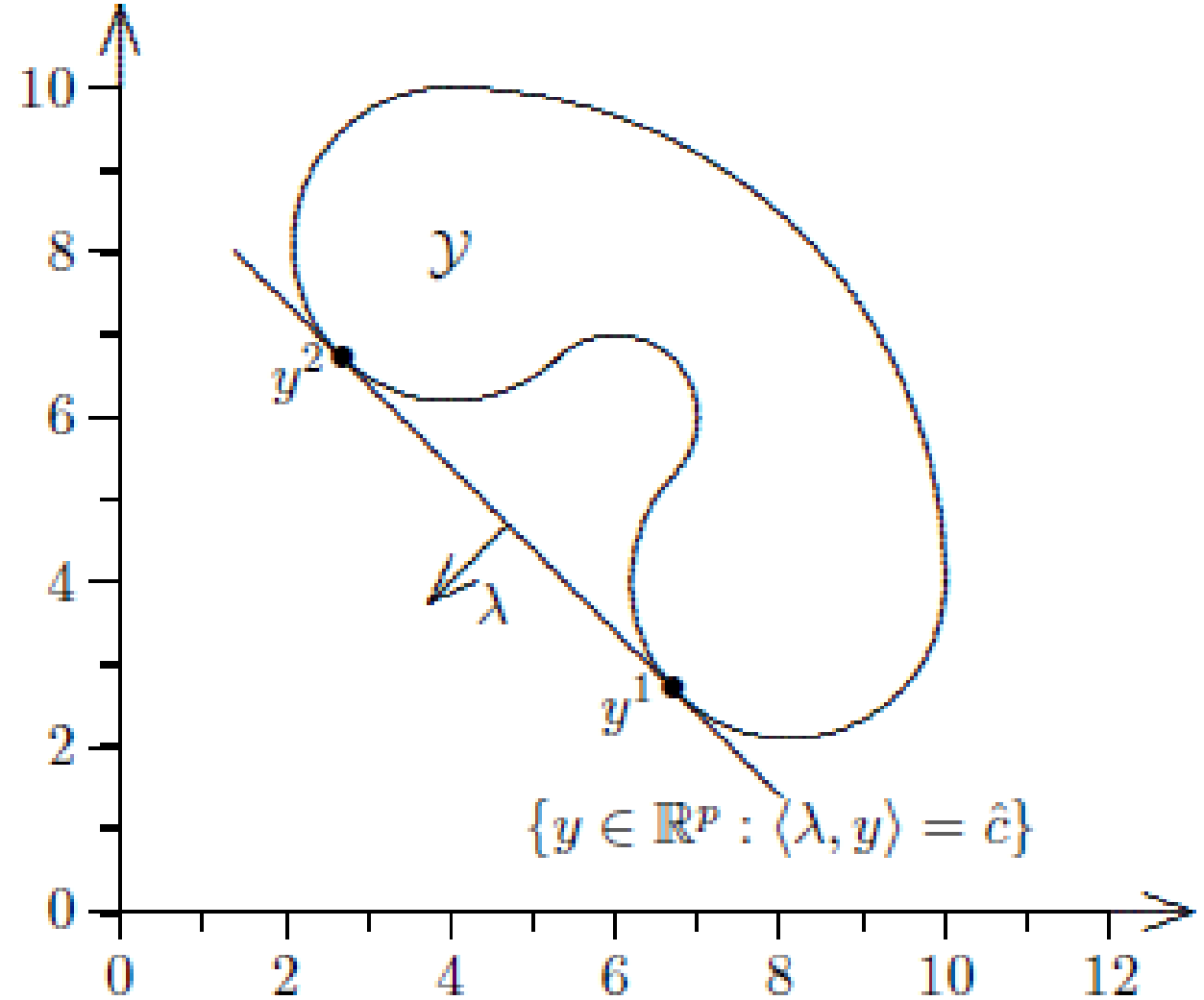
$$(p \geq q \Leftrightarrow [p, A; 1 - p, B] \succeq [q, A; 1 - q, B])$$



Método de suma ponderada

$$\begin{aligned} \min_{x \in Q} \quad & \sum_{i=1}^k w_i f_i(x) \\ \text{s.t.} \quad & \sum_{i=1}^k w_i = 1 \end{aligned}$$

$w_i \geq 0$ para $i = \{1, \dots, k\}$



Los problemas avanzados...

Multi-agente

Multi-tarea

Meta-aprendizaje

Ensamblajes

Multi-recompensa



¿Pero... somos racionales?

Optimización multi-objetivo

- La vida es acerca de decidir
 - Individualmente
 - En grupo
- Típicamente involucran algún conflicto
- “Queremos todo”



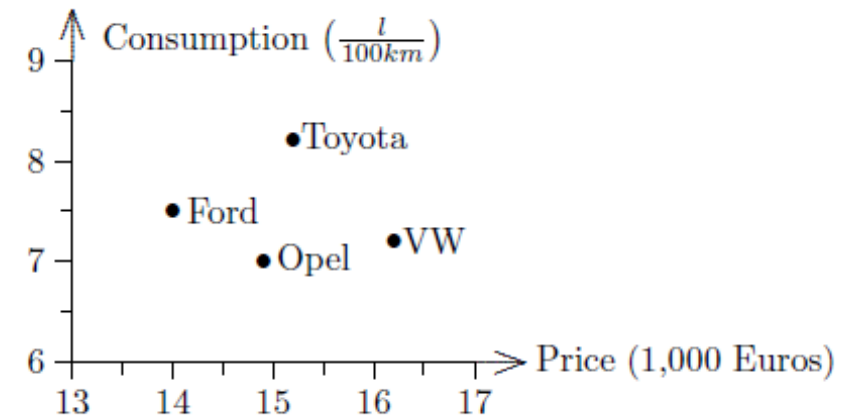
Las 3 B



Un mejor ejemplo...

- Queremos comprar un nuevo auto y hemos identificado cuatro modelos que nos agradan: VW Golf, Opel Astra, Ford Focus, Toyota Corolla
- La decisión la será tomada de acuerdo a:
 - Precio
 - Consumo de combustible
 - Potencia
- ¿Cuál es la mejor alternativa?

		Alternatives			
		VW	Opel	Ford	Toyota
Criteria	Price (1,000 Euros)	16.2	14.9	14.0	15.2
	Consumption ($\frac{l}{100km}$)	7.2	7.0	7.5	8.2
	Power (kW)	66.0	62.0	55.0	71.0



Un poco de historia...

- La primera referencia a este tipo de situaciones se atribuye a Pareto (1896), quien escribió:

"We will say that the members of a collectivity enjoy maximum ophelimity in a certain position when it is impossible to find a way of moving from that position very slightly in such a manner that the ophelimity enjoyed by each of the individuals of that collectivity increases or decreases. That is to say, any small displacement in departing from that position necessarily has the effect of increasing the ophelimity which certain individuals enjoy, and decreasing that which others enjoy, of being agreeable to some and disagreeable to others."



El problema

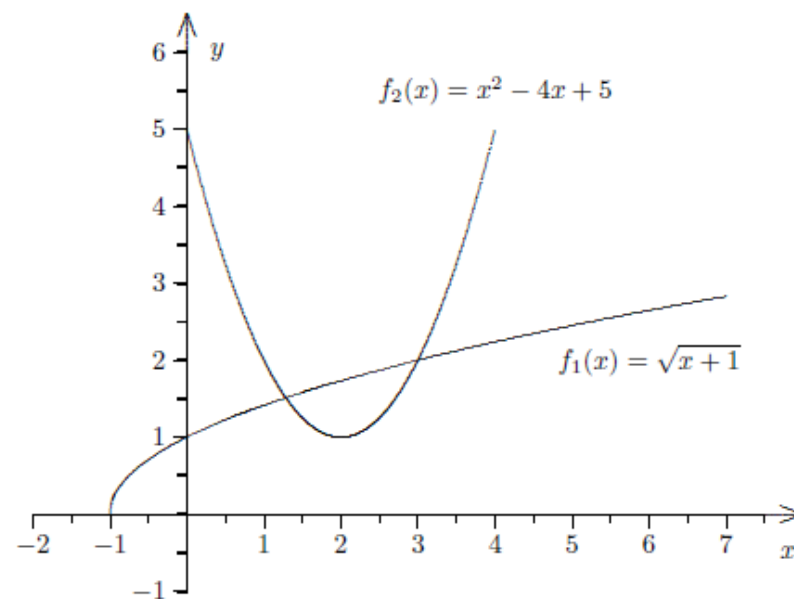
$$\min_{x \in Q} F(x)$$

Donde:

$$Q = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, i = 1, \dots, I \text{ and}$$

$$h_j(x) = 0, j = 1, \dots, m\}$$

$$F: Q \rightarrow \mathbb{R}^k$$

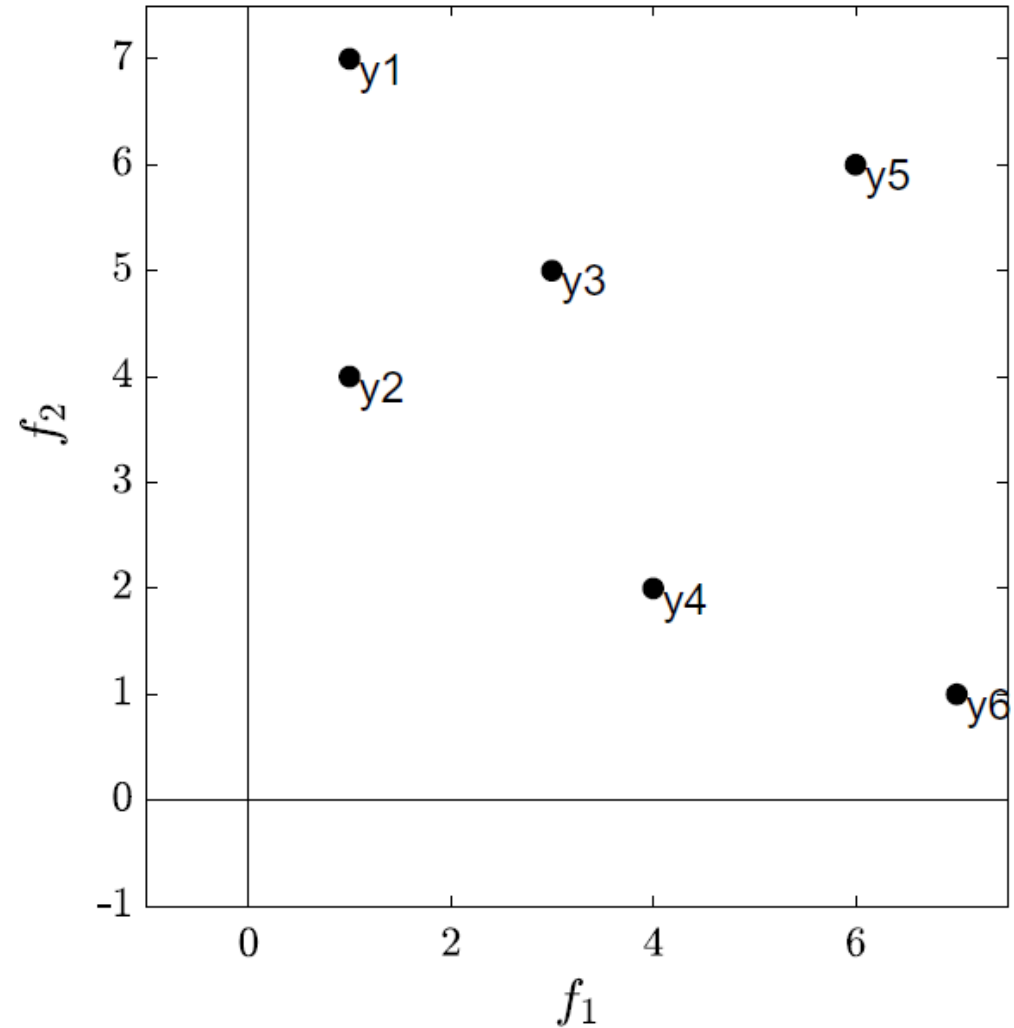


Dominancia de Pareto

- Sean $v, w \in R^k$. Entonces el vector v es menos que w ($v <_p w$), si $v_i < w_i$ para todo $i \in \{1, \dots, k\}$. La relación \leq_p se define de forma análoga
- Un vector $y \in Q$ se dice que es dominado por un vector $x \in Q$ ($x < y$) con respecto al problema multi-objetivo si

$$F(x) \leq_p F(y) \text{ y } F(x) \neq F(y)$$

Si no, y se llama no dominado por x .



¿Y cómo
encontramos
esas
soluciones de
Pareto?

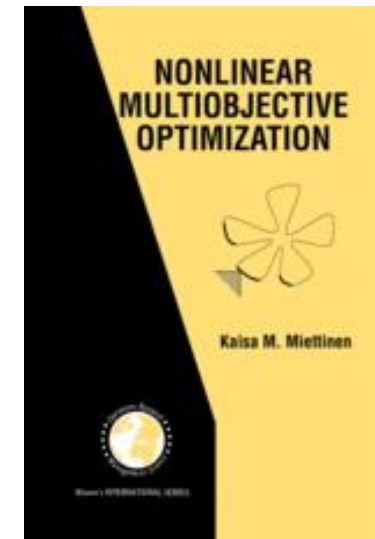
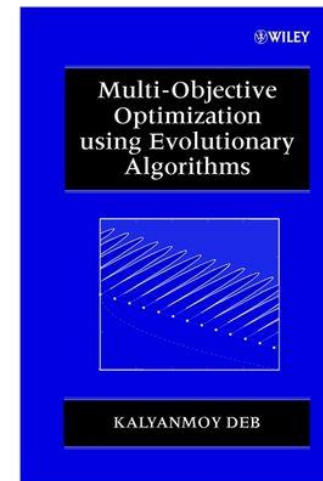
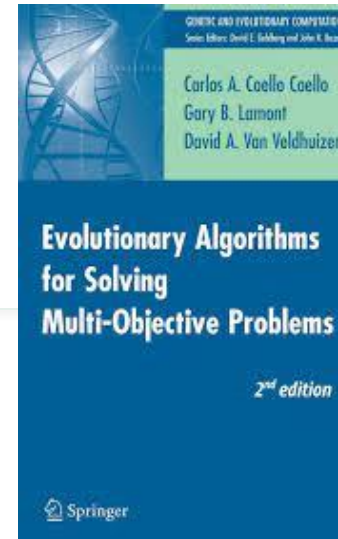


Se nos acabó
el tiempo...



Para saber más

- Libros
 - Matthias Ehrgott: *Multicriteria Optimization*, Springer 2005
 - KaizaMiettinen: *Nonlinear Multiobjective Optimization*, Kluwer, 1999
 - Carlos Coello et al.: *Evolutionary Algorithms for Solving Multi-objective Problems*
 - Kalyanmoy Deb: *Multi-objective Optimization using Evolutionary Algorithms*
 - Curso de optimización multi-objetivo en PCIC ;)





Para la otra vez...

- Cierre del curso

The End.



iimas