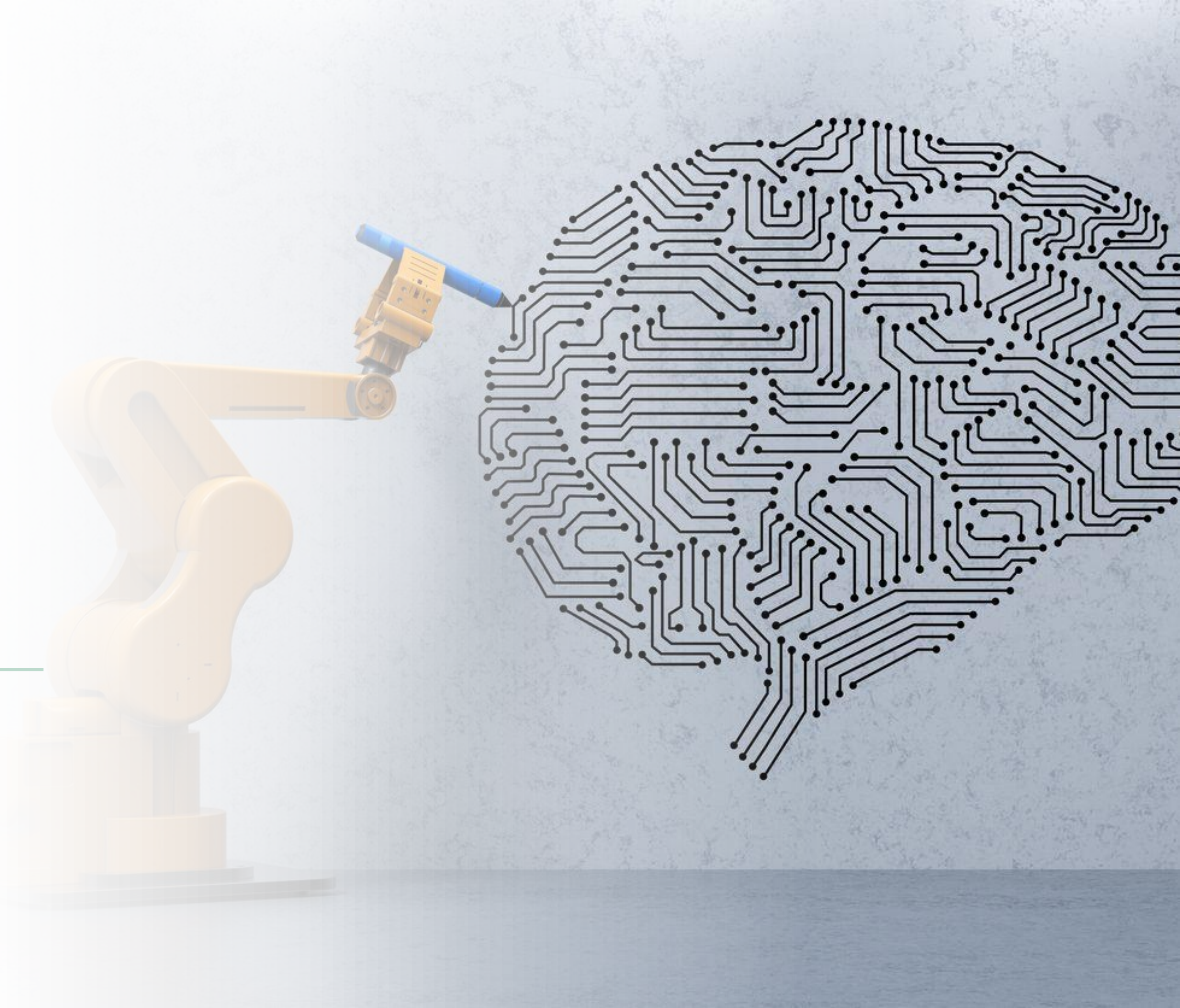
Aprendizaje por refuerzo

Clase 14: RL Bayesiano

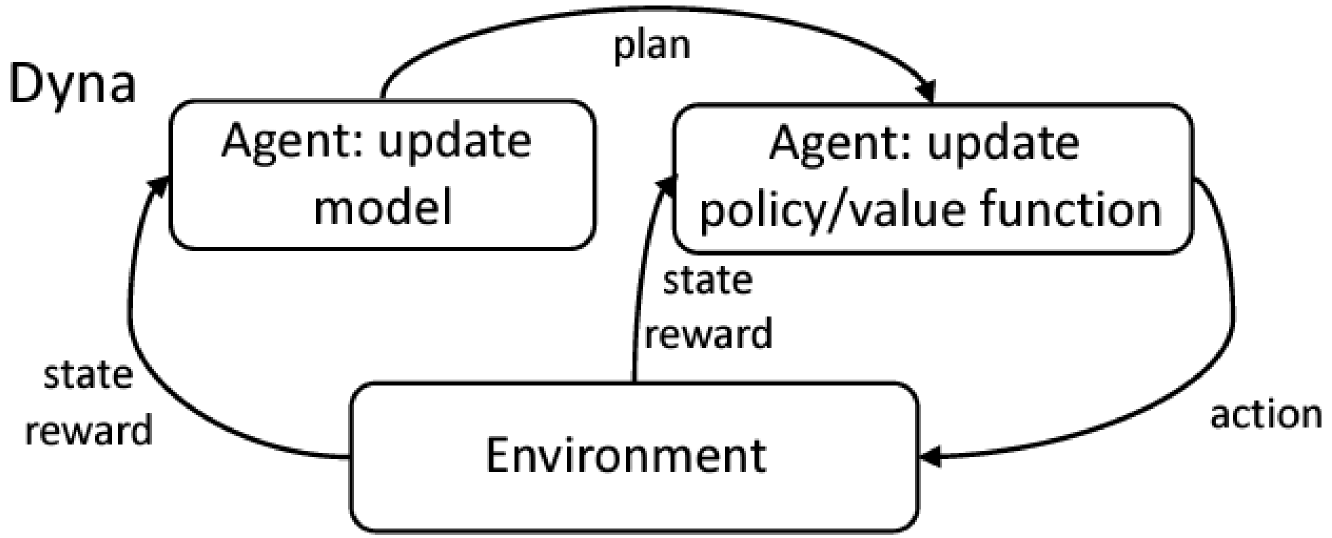


Para el día de hoy…

- RL Bayesiano

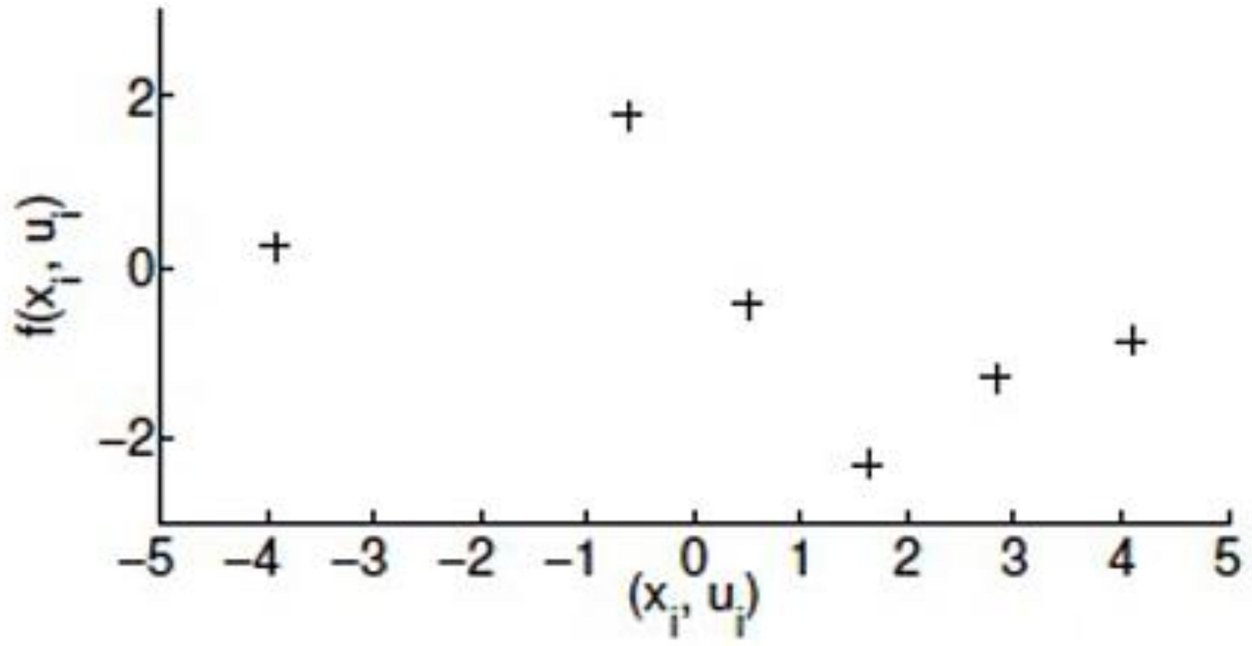
Recordando el pasado…

- RL libre de modelo: aprendizaje directo sin sesgo, necesita muchas iteraciones en el ambiente
- RL basada en modelo: aprendizaje indirecto con sesgo



Sesgo en el modelo

- Problemas
  - El modelo aprende de datos finitos
  - El modelo es imperfecto
  - Existe el riesgo que la planeación sobreajuste el modelo
  - Riesgo de malas políticas
- Solución
  - Representar la incertidumbre en el modelo



RL Bayesiano

- Representa explícitamente la incertidumbre
- Beneficios
  - Balance entre exploración y explotación
  - Mitiga el sesgo en el modelo
  - Reduce la cantidad de datos necesarios
- Desventajas
  - Complejidad de cómputo

Definición

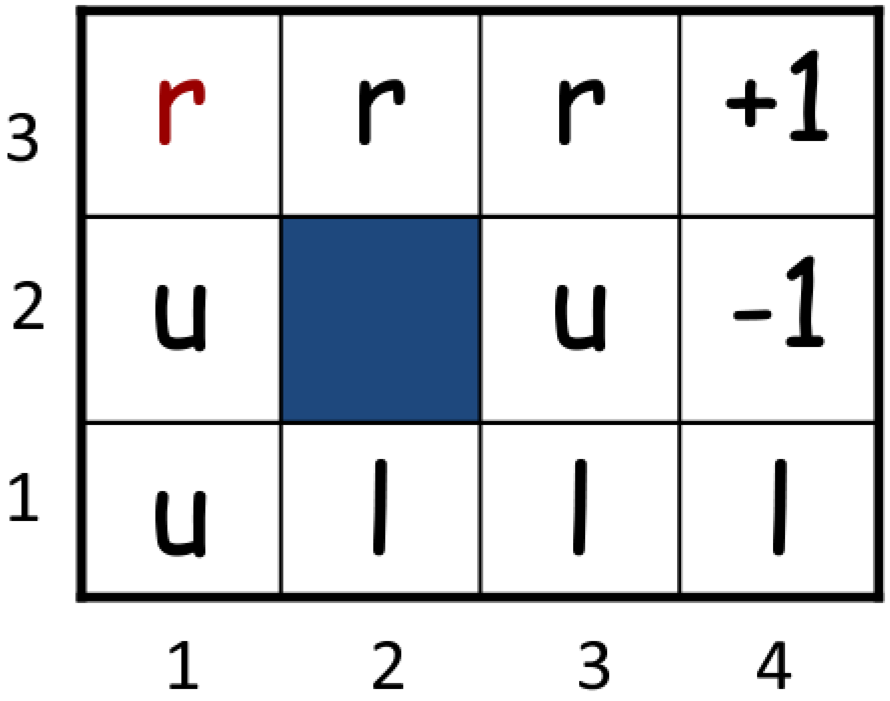
- Idea: aumentar el conjunto de estados con la distribución sobre parámetros desconocidos
- Elementos
  - Estados de información $(s, b) \in \mathcal{S}, \mathcal{B}$
    - Estados físicos $s \in \mathcal{S}$
    - Estados de creencias $b \in \mathcal{B}$ donde $b(\theta) = \mathbb{P}(\theta)$
  - Acciones $a \in \mathcal{A}$
  - Recompensas $r \in \mathbb{R}$
  - Modelo $p(r, s', b' | s, b, a)$
- Objetivo: encontrar una política $\pi: \mathcal{S} \times \mathcal{B} \to \mathcal{A}$ que maximice las recompensas esperadas

El modelo en RL Bayesiano

- $p(r, s', b' | s, b, a) = p(r, s' | s, b, a) p(b' | r, s', s, b, a)$
  - Modelo físico: $p(r, s' | s, b, a)$
  - Modelo de creencias: $p(b' | r, s', s, b, a)$

Ejemplo: gridworld

- $\gamma = 1$
- Recompensa: -0.04 para estados no terminales



$$\Pr(i', j' | i, j, right, \theta) = \begin{cases} \theta & i' = i + 1 \text{ and } j' = j \\ \frac{1-\theta}{2} & i' = i \text{ and } (j' = j + 1 \text{ } or \text{ } j' = j - 1) \\ 0 & \text{otherwise} \end{cases} .$$

$$\Pr(i', j' | i, j, up, \theta) = \begin{cases} \theta & i' = i \text{ and } j' = j + 1 \\ \frac{1-\theta}{2} & (i' = i + 1 \text{ } or \text{ } i' = i - 1) \text{ and } j' = j. \\ 0 & \text{otherwise} \end{cases} .$$

Creencias

- Modelemos la incertidumbre con respecto a θ con una distribución Beta

$$b(\theta) = k\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Actualización de creencias: teorema de Bayes

$$b'(\theta) = b^{s,a,s'}(\theta) = b(\theta|s,a,s') \propto b(\theta)p(s'|s,a,\theta)$$

Ejemplo de actualización de creencias

- A priori
- $b(\theta) = Beta(\theta; \alpha, \beta) = k\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Posteriori para $i, j, up \rightarrow i', j'$ donde $i' = i$ y $j' = j + 1$

3	r	r	r	+1
2	u		u	-1
1	u	l	l	l
	1	2	3	4

$$
\begin{aligned}
b'(\theta) = b^{s,a,s'}(\theta) = b(\theta|s,a,s') &= b(\theta|i,j,up,i',j') \\
&\propto b(\theta)Pr(i',j'|i,j,up,\theta) \\
&= k\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta \\
&= k\theta^{\alpha}(1-\theta)^{\beta-1} \propto Beta(\theta; \alpha+1, \beta)
\end{aligned}
$$

Modelo físico

- Considere $s = (i, j), a = right, s' = (i', j')$
- donde $i' = i$ y $j' = j - 1$
- Distribución predictiva
  - $p(s'|s, b, a) = \int_\theta p(s'|s, a, \theta) b(\theta) d\theta$
  - $= \int_\theta p(i', j'|i, j, right, \theta) Beta(\theta; \alpha, \theta) d\theta$
  - $= \int_\theta \frac{1-\theta}{2} k\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\beta}{2(\alpha+\beta)}$

$$\Pr(i', j'|i, j, right, \theta) = \begin{cases} \theta & i' = i + 1 \text{ and } j' = j \\ \frac{1-\theta}{2} & i' = i \text{ and } (j' = j + 1 \text{ } or \text{ } j' = j - 1) \\ 0 & \text{otherwise} \end{cases}.$$

$$\Pr(i', j'|i, j, up, \theta) = \begin{cases} \theta & i' = i \text{ and } j' = j + 1 \\ \frac{1-\theta}{2} & (i' = i + 1 \text{ } or \text{ } i' = i - 1) \text{ and } j' = j. \\ 0 & \text{otherwise} \end{cases}.$$

Planeación

- Dado que el modelo es conocido, tratarlo como un MDP
- Beneficios
  - Resolver el problema con iteración de política/valor
  - Exploración/explotación óptima (de acuerdo a creencias)
- Desventajas
  - Cómputo complicado
- Ecuación de Bellman

$$V^*(s,b) = \max_a \mathbb{E}[r|s,b,a] + \gamma \sum_{s'} p(s'|s,a,b) V^*(s', b^{s,a,s'}) \;\; \forall s$$

- Donde $\mathbb{E}[r|s,b,a] = \int_\theta b(\theta) \int_r pdf(r|s,a,\theta) r dr d\theta$

Iteración de valor

valueIteration(BayesianRL)

$V_0^*(s,b) \leftarrow \max_a E[r|s,b,a] \quad \forall s$

For $t = 1$ to h do

$\quad V_t^*(s,b) \leftarrow \max_a E[r|s,b,a] + \gamma \sum_{s'} \Pr(s'|s,a,b)\, V_{t-1}^*(s', b^{s,a,s'}) \ \forall s$

Return V^*

valueIteration(MDP)

$V_0^*(s) \leftarrow \max_a E[r|s,a] \ \forall s$

For $t = 1$ to h do

$\quad V_t^*(s) \leftarrow \max_a E[r|s,a] + \gamma \sum_{s'} \Pr(s'|s,a)\, V_{t-1}^*(s') \ \forall s$

Return V^*

Exploración/explotación

- Ya no es necesario preguntarnos si explorar o explotar
- Todo se contiene un solo objetivo: maximizar recompensa total esperada
  - $V^{\pi}(s,b) = \sum_t \gamma^t \mathbb{E}[r_t|s_t,b_t]$
  - Política óptima π^*: $V^{\pi^*}(s,b) \geq V^{\pi}(w,b)\ \forall s,b$
- Dado el conocimiento dado a priori

Algoritmo para RL Bayesiano

- Fuera de línea: planeación (sin el ambiente)
  - Encontrar π^* y/o V^* por medio de algún algoritmo (iteración de política/valor, etc.)
- En línea (con el ambiente)
  - Inicializar $s_0, b_0, n \leftarrow 0$
  - Repetir
    - Ejecutar la política $a_n \leftarrow \pi(s_n, b_n)$
    - Obtener s_{n+1} y r_n del ambiente
    - Actualizar las creencias: $b_{n+1}(\theta) = b_n^{s_n, a_n, r_n, s_{n+1}}(\theta) = b_n(\theta | s_n, a_n, r_n, s_{n+1})$
    - $n \leftarrow n + 1$

Retos de RL Bayesiana

- La fase fuera de línea es bastante complicada
  - Utiliza funciones de aproximación
  - El espacio de creencias es continuo
  - Un buen plan debe tomar en cuenta todos los posibles estados, lo cual es intratable
- Alternativa: planeación parcial
  - Muestreo de Thompson
  - PILCO

Muestro de Thompson en RL Bayesiana

Idea: muestrar modelos θ_i en cada paso y planear para esos MDPs

ThompsonSamplingInBayesianRL(s,b)

Repeat

Sample $\theta_1, \dots, \theta_k \sim \Pr(\theta)$

$Q^*_{\theta_i} \leftarrow solve(MDP_{\theta_i})\ \forall i$

$\hat{Q}(s,a) \leftarrow \frac{1}{k}\sum_{i=1}^{k} Q^*_{\theta_i}(s,a)\ \forall a$

$a^* \leftarrow \text{argmax}_a\, \hat{Q}(s,a)$

Execute a^* and receive r, s'

$b(\theta) \leftarrow b(\theta)\Pr(r, s' | s, a^*, \theta)$

$s \leftarrow s'$

Actor critico Bayesiano

- PILCO: Deisenroth, Rasmussen (2011)
  - $b(\theta)$: modelo de transición con proceso Gaussiano
- Deep PILCO: Gal, McCallister, Rasmussen (2016)
  - $b(\theta)$: modelo de transición con redes neuronales Bayesianas

PILCO(s, b, π)

Repeat

Repeat

Critic: $V_b^\pi \leftarrow policyEvaluation(b, \pi)$

Actor: $\pi \leftarrow \pi + \alpha\, \partial V_b^\pi / \partial \pi$

$a \leftarrow \pi(s, b)$

Execute a and receive r, s'

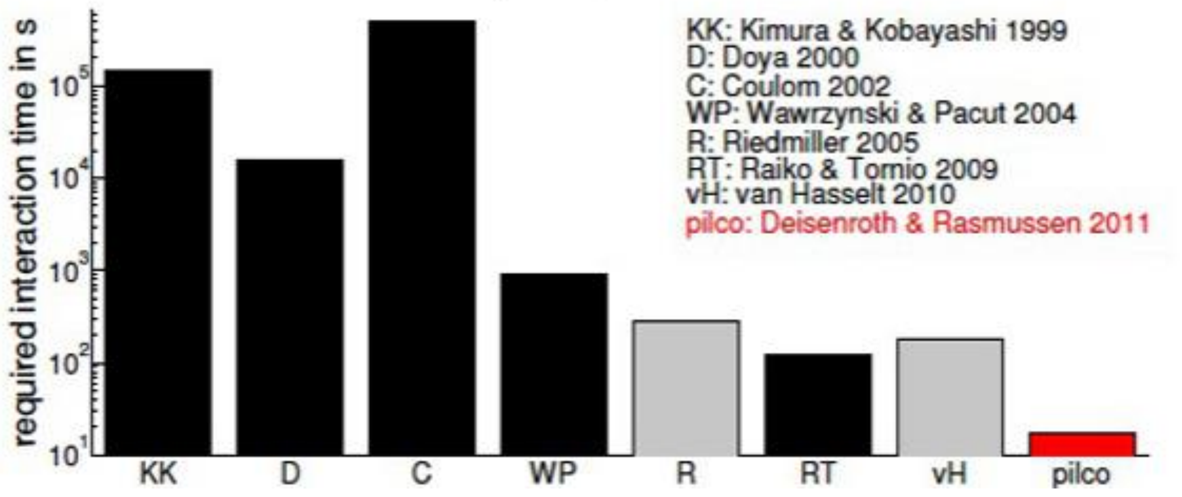
$b \leftarrow b^{s,a,r,s'}$ and $s \leftarrow s'$

Resultados

Table 1. PILCO's data efficiency scales to high dimensions.

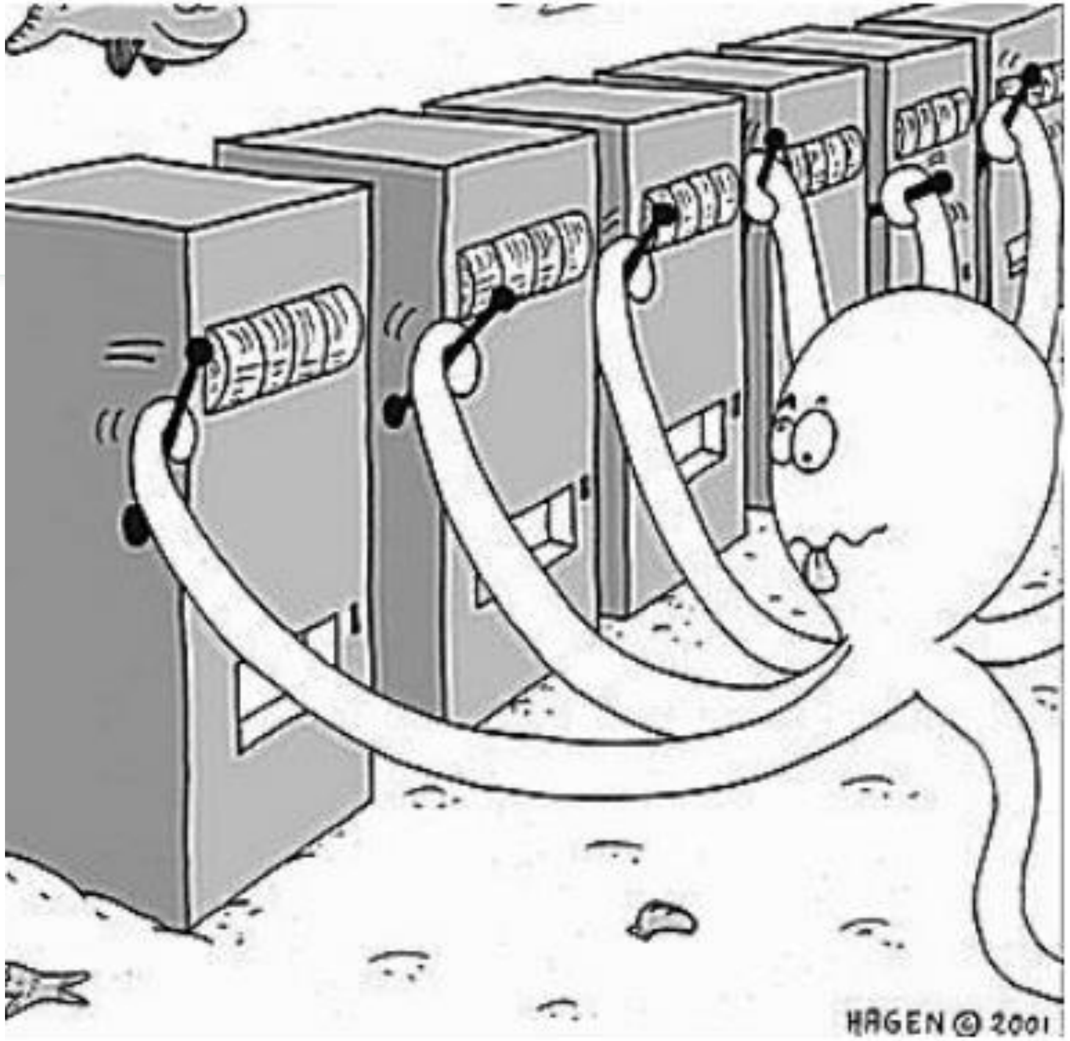
	cart-pole	cart-double-pole	unicycle
state space	$\mathbb{R}^4$	$\mathbb{R}^6$	$\mathbb{R}^{12}$
# trials	≤ 10	20–30	≈ 20
experience	$\approx 20\,\mathrm{s}$	$\approx 60\,\mathrm{s}$–$90\,\mathrm{s}$	$\approx 20\,\mathrm{s}$–$30\,\mathrm{s}$
parameter space	$\mathbb{R}^{305}$	$\mathbb{R}^{1816}$	$\mathbb{R}^{28}$

Cartpole problem

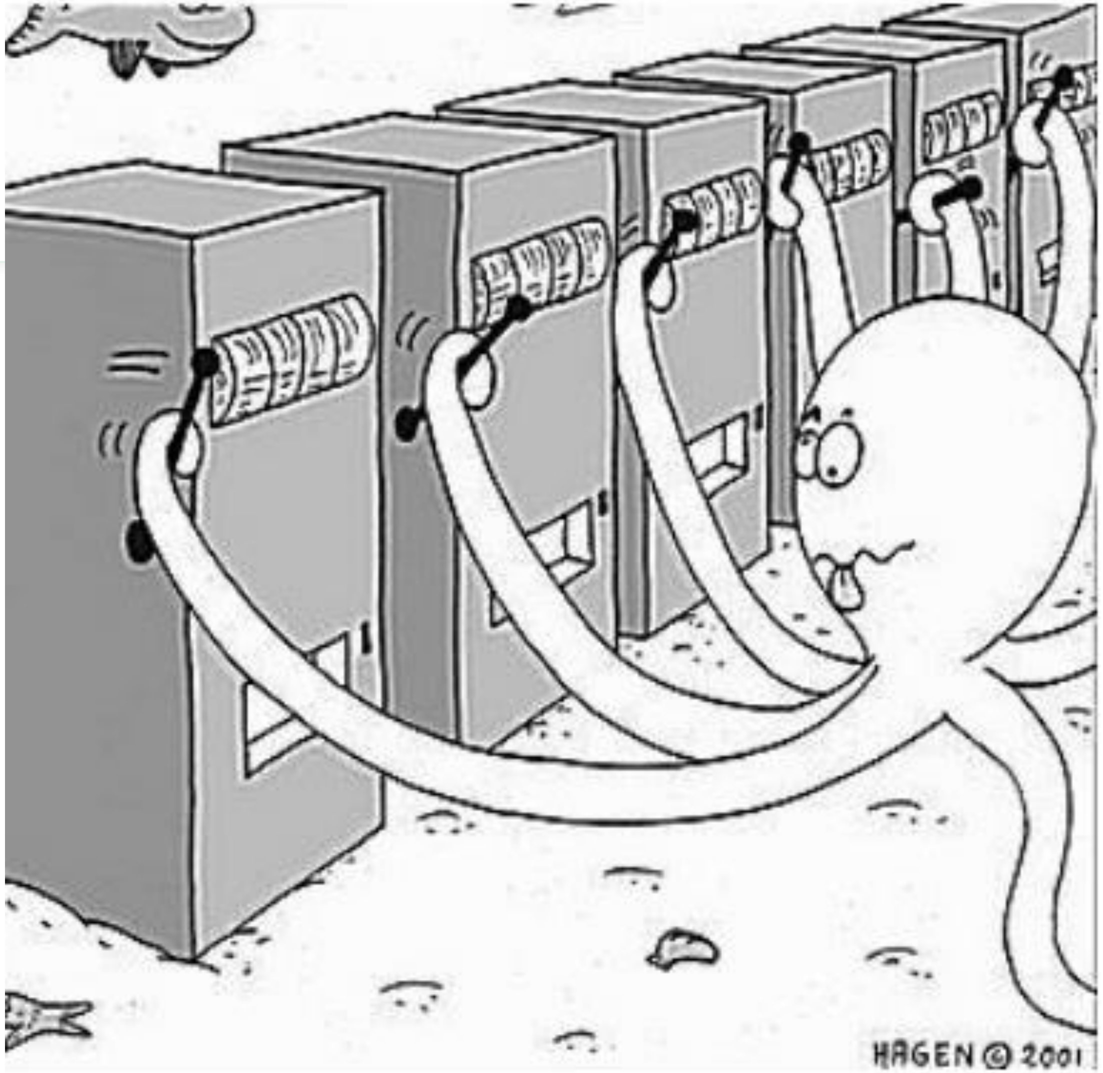


Recordando el bandido multi-brazo

- Es una tupla $(\mathcal{A}, \mathcal{R})$
- $\mathcal{A}$ es un conjunto de m acciones
- $\mathcal{R}^a(r) = \mathbb{P}[r|a]$ es una distribución de probabilidad desconocida sobre recompensas
- En cada paso t el agente selecciona una acción $a_t \in \mathcal{A}$
- El ambiente genera una recompensa $r_t \sim \mathcal{R}^{a_t}$
- El objetivo es maximizar la recompensa cumulativa $\sum_{\tau=1}^{t} r_\tau$



¿Qué pasa si tenemos un número infinito de brazos?



Para la otra vez…

• Implementación I