Aprendizaje por refuerzo

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad de Sevilla

Planificación bajo incertidumbre

Asunciones en planificación clásica:

1. Conjuntos finitos de estados y acciones.
2. Estado en que se encuentra el sistema completamente observable.
3. Efectos deterministas de las acciones.
4. Sistema solo cambia por medio de las acciones.
5. Planes como secuencias de acciones.
6. Objetivo como conjuntos de estados.
7. Ejecución instantánea de acciones.
8. El problema no cambia mientras se planifica.

Cambios en planificación bajo incertidumbre:

1. Conjuntos finitos de estados y acciones.
2. Estado en que se encuentra el sistema completamente observable.
3. Efectos no deterministas de las acciones.
4. Sistema solo cambia por medio de las acciones.
5. Planes como políticas de acciones.
6. Objetivo como optimización de políticas.
7. Ejecución instantánea de acciones.
8. El problema no cambia mientras se planifica.

Efectos no deterministas de las acciones:

- Aplicar una misma acción a un mismo estado puede dar lugar a resultados distintos.
- Modelizado mediante probabilidades sobre los resultados de una acción.
- Comportamiento del sistema más realista.

Planes como políticas de acciones:

- Misma secuencia de acciones puede resultar en distintas secuencias de estados.
- Inviable solución del problema como secuencia de acciones.
- Una política establece qué acción aplicar en cada posible estado.

Objetivo como optimización de políticas:

- Objetivo no es alcanzar ciertos estados, sino elegir la mejor política.
- Modelizado mediante funciones de utilidad a maximizar.

Procesos de decisión de Markov

Proceso de decisión de Markov: (S, A, P)

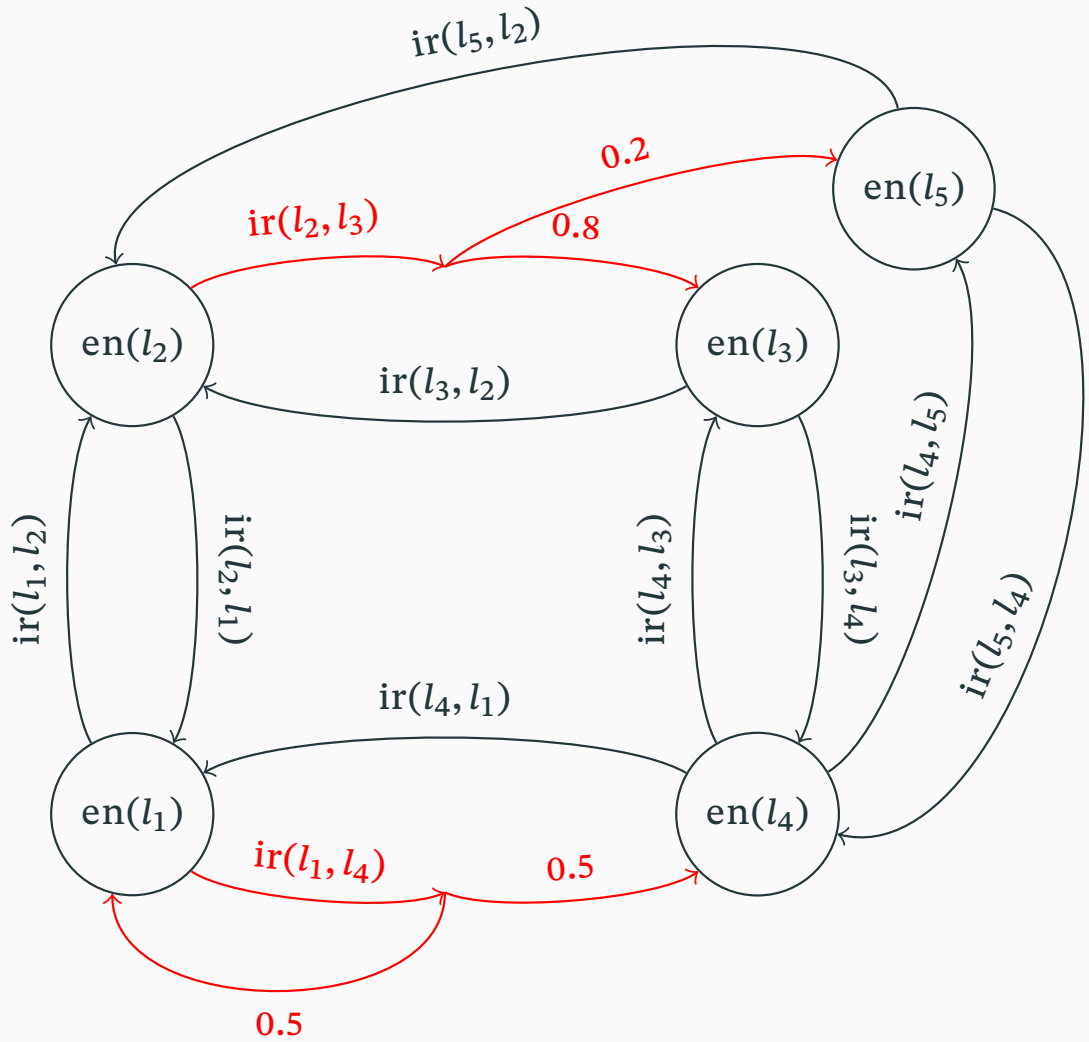
- S conjunto finito de estados.
- A conjunto finito de acciones.
- $P_a(s' \mid s)$ probabilidad de que la acción a lleve al sistema del estado s al estado s'.

Acciones ejecutables en s: $A(s) = \{a \in A \mid \exists s' \in S(P_a(s' \mid s) \neq 0)\}$

Para todo estado $s \in S$ y toda acción ejecutable $a \in A(s)$ exigimos que $P_a(\cdot \mid s)$ sea una distribución de probabilidad:

$$0 \leq P_a(s' \mid s) \leq 1 \qquad \text{y} \qquad \sum_{s' \in S} P_a(s' \mid s) = 1$$

Ejemplo: robot que se mueve entre 5 posibles localizaciones.
Acciones: **ir**, **esperar**, con probabilidad **1** salvo donde se indica.
esperar deja al robot en la misma localización.



Un plan como secuencia de acciones no es conveniente:

$$\langle \mathrm{ir}(l_1, l_2), \mathrm{ir}(l_2, l_3), \mathrm{ir}(l_3, l_4) \rangle$$

Posición del robot: l_1

Un plan como secuencia de acciones no es conveniente:

$$\langle \mathrm{ir}(l_1, l_2), \mathrm{ir}(l_2, l_3), \mathrm{ir}(l_3, l_4) \rangle$$

Posición del robot: l_2

Un plan como secuencia de acciones no es conveniente:

$$\langle \mathrm{ir}(l_1, l_2), \mathrm{ir}(l_2, l_3), \mathrm{ir}(l_3, l_4) \rangle$$

Posición del robot: l_5

Un plan como secuencia de acciones no es conveniente:

$$\langle \mathrm{ir}(l_1, l_2), \mathrm{ir}(l_2, l_3), \mathrm{ir}(l_3, l_4) \rangle$$

Posición del robot: l_5

Un plan como secuencia de acciones no es conveniente:

$$\langle \mathrm{ir}(l_1, l_2), \mathrm{ir}(l_2, l_3), \mathrm{ir}(l_3, l_4) \rangle$$

Posición del robot: l_5

Política: asigna a cada estado una acción ejecutable.

$$\pi: S \to A, \quad \text{con } \pi(s) \in A(s), \forall s \in S$$

Cantidad de políticas posibles acotada por $|A|^{|S|}$.

Planificador: genera una política y se la proporciona al controlador.

Controlador: observa el estado del sistema y ejecuta la acción indicada por la política.

Ejemplos de políticas que tratan de llevar el robot a l_4:

$$\pi_1 = \Big\{\big(\text{en}(l_1), \text{ir}(l_1, l_2)\big), \big(\text{en}(l_2), \text{ir}(l_2, l_3)\big), \big(\text{en}(l_3), \text{ir}(l_3, l_4)\big), \big(\text{en}(l_4), \text{esperar}\big), \big(\text{en}(l_5), \text{esperar}\big)\Big\}$$

$$\pi_2 = \Big\{\big(\text{en}(l_1), \text{ir}(l_1, l_2)\big), \big(\text{en}(l_2), \text{ir}(l_2, l_3)\big), \big(\text{en}(l_3), \text{ir}(l_3, l_4)\big), \big(\text{en}(l_4), \text{esperar}\big), \big(\text{en}(l_5), \text{ir}(l_5, l_4)\big)\Big\}$$

$$\pi_3 = \Big\{\big(\text{en}(l_1), \text{ir}(l_1, l_4)\big), \big(\text{en}(l_2), \text{ir}(l_2, l_1)\big), \big(\text{en}(l_3), \text{ir}(l_3, l_4)\big), \big(\text{en}(l_4), \text{esperar}\big), \big(\text{en}(l_5), \text{ir}(l_5, l_4)\big)\Big\}$$

- Con π_1 el robot puede quedarse esperando eternamente en l_5.
- Con π_2 el robot toma un camino largo pero seguro hasta l_4.
- Con π_3 el robot intenta tomar el camino directo de l_1 a l_4, pero puede quedarse atascado en l_1.

La ejecución de una política se corresponde con una historia: sucesión infinita de estados.

Cumplen la propiedad de Markov: la probabilidad de que el sistema se encuentre en un estado en un instante dado solo depende del estado en que se encontraba en el instante inmediatamente anterior.

Probabilidad de una historia $h = \{s_i\}_{i \geq 0}$ inducida por una política π:

$$\mathbb{P}(h \mid \pi) = \prod_{i \geq 0} P_{\pi(s_i)}(s_{i+1} \mid s_i)$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$
$$\mathbb{P}(h_3 \mid \pi_2) = 0.2$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$
$$\mathbb{P}(h_3 \mid \pi_2) = 0.2 \qquad \mathbb{P}(h_4 \mid \pi_3) = 0.5$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$
$$\mathbb{P}(h_3 \mid \pi_2) = 0.2 \qquad \mathbb{P}(h_4 \mid \pi_3) = 0.5 \qquad \mathbb{P}(h_5 \mid \pi_3) = 0.25$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$
$$\mathbb{P}(h_3 \mid \pi_2) = 0.2 \qquad \mathbb{P}(h_4 \mid \pi_3) = 0.5 \qquad \mathbb{P}(h_5 \mid \pi_3) = 0.25$$
$$\mathbb{P}(h_6 \mid \pi_3) = 0.125$$

Ejemplos de historias:

$$h_1 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_3), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_2 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_5), \ldots \rangle$$
$$h_3 = \langle \text{en}(l_1), \text{en}(l_2), \text{en}(l_5), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_4 = \langle \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_5 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_6 = \langle \text{en}(l_1), \text{en}(l_1), \text{en}(l_1), \text{en}(l_4), \text{en}(l_4), \ldots \rangle$$
$$h_7 = \langle \text{en}(l_1), \text{en}(l_1), \ldots \rangle$$

Probabilidades inducidas:

$$\mathbb{P}(h_1 \mid \pi_1) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_2) = 0.8 \qquad \mathbb{P}(h_1 \mid \pi_3) = 0$$
$$\mathbb{P}(h_2 \mid \pi_1) = 0.2 \qquad \mathbb{P}(h_2 \mid \pi_2) = 0 \qquad \mathbb{P}(h_2 \mid \pi_3) = 0$$
$$\mathbb{P}(h_3 \mid \pi_2) = 0.2 \qquad \mathbb{P}(h_4 \mid \pi_3) = 0.5 \qquad \mathbb{P}(h_5 \mid \pi_3) = 0.25$$
$$\mathbb{P}(h_6 \mid \pi_3) = 0.125 \qquad \mathbb{P}(h_7 \mid \pi_3) = 0$$

Función recompensa: $R: S \to \mathbb{R}$, establece preferencias entre los estados (asumimos R acotada).

Utilidad de una historia $h = \{s_i\}_{i \geq 0}$ inducida por una política π:

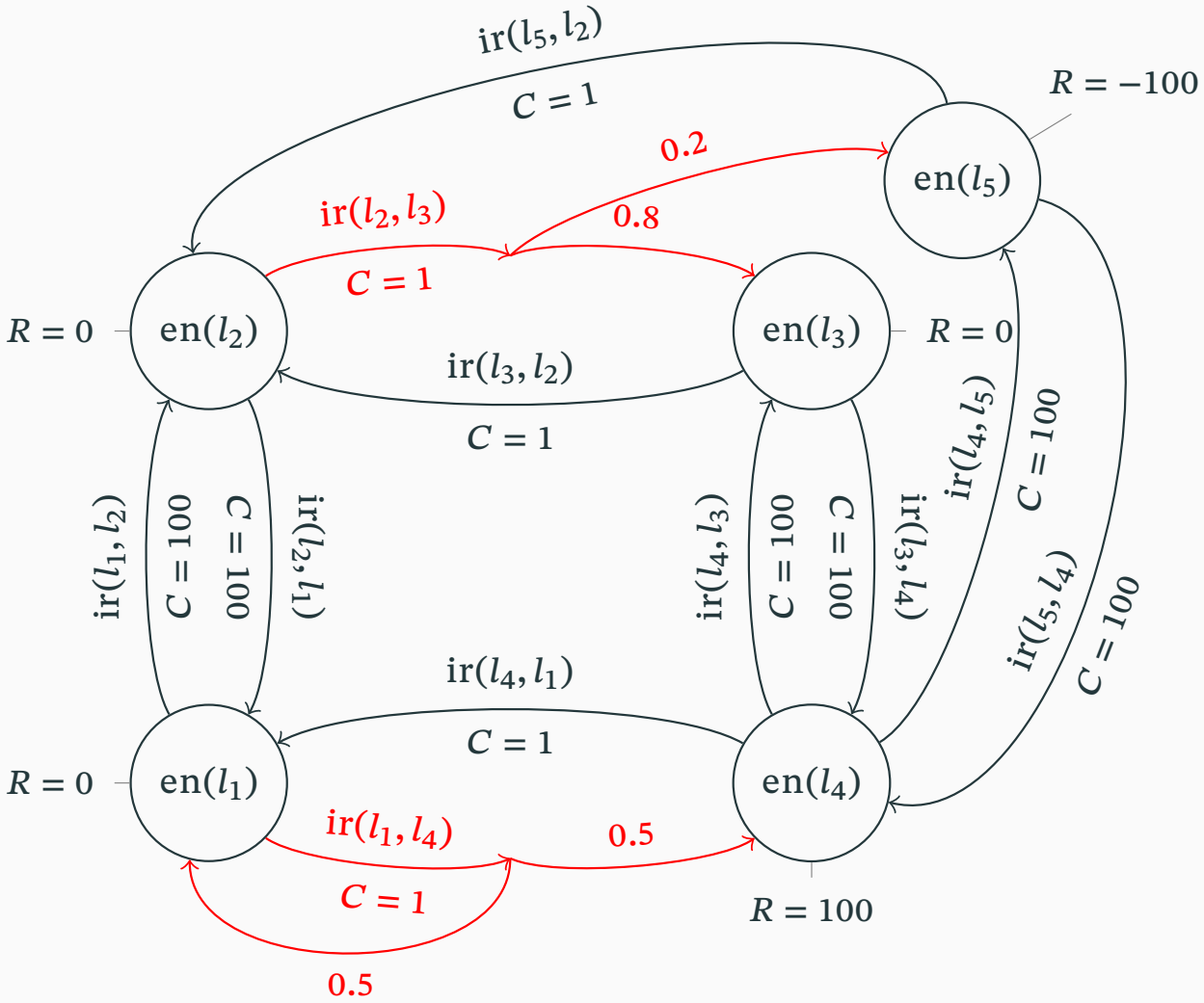
$$U(h \mid \pi) = \sum_{i \geq 0} R(s_i)$$

Puede también considerarse un coste (acotado), $C: S \times A \to \mathbb{R}$, para las acciones, en cuyo caso:

$$U(h \mid \pi) = \sum_{i \geq 0} R\big(s_i, \pi(s_i)\big)$$

con $R(s, a) = R(s) - C(s, a)$, para cada estado s y cada acción a.

El coste de **esperar** en cada estado es **0**.



Utilidad de las historias puede tomar valores infinitos, lo que impide su comparación.

Asegurar valores acotados: usar factor de descuento γ, con $0 < \gamma < 1$:

$$U(h \mid \pi) = \sum_{i \geq 0} \gamma^i R\big(s_i, \pi(s_i)\big)$$

Justificación del factor de descuento: reduce la contribución de recompensas y costes futuros distantes del estado actual.

Utilidad esperada de estado s bajo política π:

$$U_\pi(s) = \mathbb{E}\big[U(h \mid \pi)\big] = \sum_{h \in H(s)} \mathbb{P}(h \mid \pi) U(h \mid \pi)$$

donde $H(s)$ es el conjunto de todas las historias con estado inicial s.

Caracterización de U_π: es la única solución del sistema de ecuaciones lineales

$$U(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s' \in S} P_{\pi(s)}(s' \mid s) U(s'), \quad \forall s \in S$$

Máxima utilidad esperada de un estado s:

$$U^*(s) = \max_\pi U_\pi(s)$$

Ecuaciones de Bellman: U^* es la única solución de

$$U(s) = \max_{a \in A(s)} \left(R(s,a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U(s') \right), \quad \forall s \in S$$

Política óptima π^*: cualquier política π tal que $U_\pi = U^*$.

La política óptima puede obtenerse eligiendo para cada estado una acción que maximice su utilidad esperada:

$$\pi^*(s) \in \arg\max_{a \in A(s)} \left(R(s,a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U^*(s') \right), \quad \forall s \in S$$

Algoritmos de cálculo de política óptima

- Ecuaciones de Bellman sistema de ecuaciones no lineales. Resolución directa complicada en general.
- Algoritmo de iteración de valores:
  1. Basándose en ecuaciones de Bellman, genera secuencia de funciones de utilidad que tiende a U^*.
  2. Tras cumplirse criterio de parada, deriva política a partir de la función de utilidad.
  3. Garantiza obtener política óptima en número finito de pasos.
- Algoritmo de iteración de políticas:
  1. Genera una secuencia de políticas que converge a π^* en un número finito de pasos.
  2. Cada política se deriva a partir de la función de utilidad de la política anterior.
  3. Cada función de utilidad se calcula mediante sistema de ecuaciones lineales.

Algoritmo de iteración de valores:

1 Inicializar arbitrariamente U_0
2 $i \leftarrow 0$
3 **repetir**
4 **para cada** $s \in S$ **hacer**
5 $U_{i+1}(s) = \max_{a \in A(s)} \Big(R(s,a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U_i(s') \Big)$
6 $i \leftarrow i + 1$
7 **hasta que** se cumple criterio de parada
8 **para cada** $s \in S$ **hacer**
9 $\pi(s) \in \arg\max_{a \in A(s)} \Big(R(s,a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U_i(s') \Big)$
10 **devolver** π

Criterio de parada: $\|U_i - U_{i-1}\| \stackrel{\text{def}}{=} \max_{s \in S} |U_i(s) - U_{i-1}(s)| < \varepsilon$

Garantiza $\|U_i - U^*\| < \frac{2\gamma}{1-\gamma}\varepsilon$

Algoritmo de iteración de políticas:

1 Inicializar arbitrariamente π_0
2 $i \leftarrow 0$
3 **repetir**
4 resolver el sistema de ecuaciones lineales
5 $U_i(s) = R(s, \pi_i(s)) + \gamma \sum_{s' \in S} P_{\pi_i(s)}(s' \mid s) U_i(s'), \quad s \in S$
6 **para cada** $s \in S$ **hacer**
7 $\pi_{i+1}(s) \in \arg\max_{a \in A(s)} \left(R(s, a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U_i(s') \right)$
8 $i \leftarrow i + 1$
9 **hasta que** $\pi_i = \pi_{i-1}$
10 **devolver** π_i

Aprendizaje por refuerzo

Iteración de valores e iteración de políticas requieren de un conocimiento completo de la dinámica del sistema.

En un problema real las funciones $P_a(s' \mid s)$ y $R(s, a)$ son, en general, desconocidas.

Algoritmos de aprendizaje por refuerzo buscan política óptima mediante *ensayo y error*, aprendiendo a partir de la interacción del sistema con el entorno.

Entorno: todo lo que no puede ser controlado por el sistema.

Interacción: real o simulada.

Equilibrio necesario entre explotación y exploración:

- Explotación: repetir las mejores acciones encontradas para maximizar la recompensa acumulada.
- Exploración: considerar todas las posibles acciones para asegurar encontrar la mejor.

Política ε-voraz ($\varepsilon \in (0, 1)$): en cada paso elegir con probabilidad $1 - \varepsilon$ la mejor acción y con probabilidad ε una acción aleatoria.

Método de Montecarlo: estima $U_\pi(s) = \mathbb{E}\big[U(h \mid \pi)\big]$ mediante promedio de recompensas (obtenidas interaccionando con el entorno) acumuladas a lo largo de historias iniciadas en s e inducidas por π.

Asumimos existencia de estado terminal absorbente: siempre se alcanza bajo π para cualquier estado inicial; solo transiciona a sí mismo, generando recompensa 0.

$h_1, \dots, h_n$ historias (finitas, acaban en terminal) iniciadas en s y $U_1, \dots, U_n$ utilidades correspondientes inducidas por π (sumas finitas, con descuento, de las recompensas proporcionadas por el entorno). Ley fuerte de los grandes números asegura:

$$U_\pi(s) = \lim_{n \to +\infty} \frac{U_1 + \cdots + U_n}{n}$$

En realidad, cada historia proporciona valores empíricos de $U_\pi(s)$ para cada estado s que aparece en ella.

Montecarlo de primera visita: para cada s se estima $U_\pi(s)$ mediante el promedio de las utilidades a partir de la primera vez que aparece s en cada historia.

Montecarlo de cada visita: para cada s se estima $U_\pi(s)$ mediante el promedio de las utilidades a partir de cada vez que aparece s en cada historia.

Estimar U_π explota política π, pero esta puede no ser óptima.

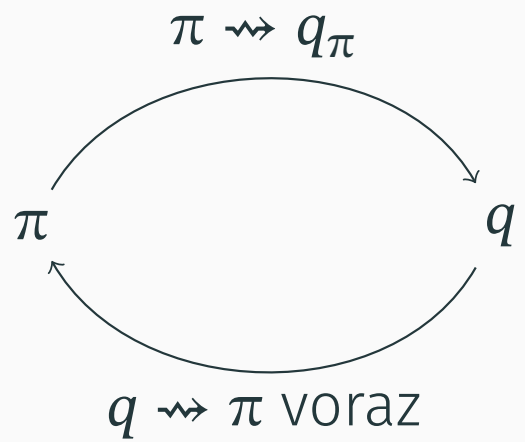
Iteración de políticas deriva a partir de U_π nueva política mejor o igual que π, pero requiere conocer $P_a(s' \mid s)$.

Igual que para estimar $U_\pi(s)$, puede usarse Montecarlo para estimar utilidad de estado s y acción a: valor esperado de recompensas acumuladas al aplicar a a s y política π a continuación.

$$q_\pi(s, a) \stackrel{\text{def}}{=} R(s, a) + \gamma \sum_{s' \in S} P_a(s' \mid s) U_\pi(s')$$

Inicios exploratorios: forma de considerar políticas distintas de π. Se elige aleatoriamente estado inicial y primera acción a ejecutar y se sigue política π a partir de ahí.

Análogo a iteración de políticas, Montecarlo intercala estimación de utilidad de pares estado-acción con mejora de política mediante criterio voraz.



En cada iteración, estimación de q_π actualizada a partir de única historia generada bajo π.

Montecarlo de primera visita con inicios exploratorios:

1 Inicializar arbitrariamente $\pi(s) \in A(s), \quad \forall s \in S$
2 Inicializar arbitrariamente $q(s,a) \in \mathbb{R}, \quad \forall s \in S, a \in A(s)$
3 $Racum(s,a) \leftarrow$ lista vacía, $\quad \forall s \in S, a \in A(s)$
4 **repetir**
5 elegir aleatoriamente s_0 no terminal y $a_0 \in \mathrm{A}(s_0)$
6 generar $s_0, a_0, R_0, s_1, a_1, R_1, \ldots, s_T, a_T, R_T, s_{T+1}$
(s_{T+1} terminal, $a_i = \pi(s_i)$ para $i > 0$, $R_i = R(s_i, a_i)$ para $i \geq 0$)
7 **para cada** $t = 0, \ldots, T$ **hacer**
8 **si** s_t, a_t es la primera vez que ocurre en la secuencia **entonces**
9 $U \leftarrow \sum_{i=t}^{T} \gamma^{i-t} R_i$
10 añadir U a $Racum(s_t, a_t)$
11 $q(s_t, a_t) \leftarrow$ media de los valores de $Racum(s_t, a_t)$
12 $\pi(s_t) \leftarrow \arg\max_{a \in A(s_t)} q(s_t, a)$
13 **hasta que** se cumple la condición de parada
14 **devolver** π

Montecarlo de cada visita: eliminar el condicional en línea 8.

Los algoritmos de programación dinámica (iteración de valores/políticas) son locales: la estimación de U y π para cada estado solo depende de las estimaciones para los estados accesibles desde él. Pero requieren conocer la dinámica del sistema.

El método de Montecarlo no es local: deben generarse secuencias completas para poder estimar q y π. Pero puede aprender directamente de la experiencia, sin conocer la dinámica del sistema.

El método de las diferencias temporales (DT) combina ideas de ambos para realizar estimaciones locales a partir de la experiencia.

Idea básica del método de Montecarlo para estimar U_π:

$$U_\pi(s) \simeq U_\pi^n(s) \stackrel{\text{def}}{=} \frac{U_1 + \cdots + U_n}{n}$$

con U_i recompensas acumuladas de la historia h_i.

Las medias anteriores se pueden calcular de forma incremental al aumentar la cantidad n de historias generadas:

$$\begin{aligned}
U_\pi^n(s) &= \frac{1}{n}\sum_{i=1}^{n} U_i \\
&= \frac{1}{n}\left(U_n + \sum_{i=1}^{n-1} U_i\right) \\
&= \frac{1}{n}\left(U_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} U_i\right) \\
&= \frac{1}{n}\left(U_n + (n-1)U_\pi^{n-1}(s)\right) \\
&= U_\pi^{n-1}(s) + \frac{1}{n}\left(U_n - U_\pi^{n-1}(s)\right)
\end{aligned}$$

En lugar de usar la recompensa acumulada de toda la historia, diferencias temporales suma la actual con la estimación de la utilidad del siguiente estado.

Generada la secuencia inducida por π

$$s_0, a_0, R_0, s_1, a_1, R_1, \dots, s_T, a_T, R_T, s_{T+1}$$

actualiza las estimaciones como sigue:

$$U(s_t) \leftarrow U(s_t) + \alpha(s_t)\big(R_t + \gamma U(s_{t+1}) - U(s_t)\big), \quad t = 0, \dots, T$$

$\delta_t \stackrel{\text{def}}{=} R_t + \gamma U(s_{t+1}) - U(s_t)$ es el error DT.

Las estimaciones convergen a U_π con probabilidad 1 si el factor de aprendizaje $\alpha(s_t)$ tiende a 0 adecuadamente. En la práctica se suele usar un valor constante $\alpha \in (0, 1]$.

Para q función de utilidad de pares estado-acción

$$\delta_t \stackrel{\text{def}}{=} R_t + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t)$$

y la actualización de las estimaciones es

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha(s_t, a_t)\big(R_t + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t)\big)$$

El algoritmo *Q-learning* aproxima q^*, función de utilidad óptima de pares estado-acción, con independencia de la política π seguida:

$$\delta_t \stackrel{\text{def}}{=} R_t + \gamma \operatorname{máx}_{a \in A(s_{t+1})} q(s_{t+1}, a) - q(s_t, a_t)$$

Algoritmo *Q-learning*:

1 Inicializar arbitrariamente $q(s, a) \in \mathbb{R}, \quad \forall s \in S, a \in A(s)$
2 $q(\text{terminal}, a) \leftarrow 0, \quad \forall a \in A(\text{terminal})$
3 **repetir**
4 elegir aleatoriamente s no terminal
5 **repetir**
6 elegir $a \in A(s)$ según política ε-voraz derivada de q
7 realizar acción a y observar R y s'
8 $q(s, a) \leftarrow q(s, a) + \alpha\big(R + \gamma \operatorname{máx}_{a' \in A(s')} q(s', a') - q(s, a)\big)$
9 $s \leftarrow s'$
10 **hasta que** s es terminal
11 **hasta que** se cumple la condición de parada
12 **devolver** política voraz derivada de q

Parámetros del algoritmo: $\alpha \in (0, 1], \varepsilon \in (0, 1)$.