Procesos Markovianos de Decisión II

Reinforcement Learning & HMM

Instituto Tecnologico Autónomo de México

Primavera 2017

I. Tareas de aprendizaje por reforzamiento (recordatorio)

Una **tarea de aprendizaje por reforzamiento** (en tiempo discreto) está compuesta de los siguientes elementos

- Un **proceso de estados** $(S_t)_{t=0}^{\infty}$ con valores es un conjunto $\mathcal{S}$
- Un **proceso de acciones** $(A_t)_{t=0}^{\infty}$ que toma el **agente** con valores en un conjunto $\mathcal{A}(\mathcal{S})$ de acciones disponibles según el estado.
- Un **proceso de pagos** $(R_t)_{t=0}^{\infty}$ que se reciben tras tomar una acción según el estado
- Una **familia de probabilidades de transición** dictadas por el **ambiente/naturaleza** que gobiernan la probabilidad de estar en el estado s' y recibir un pago r en el tiempo $t+1$ si se está en el estado s al tiempo t y se toma la acción a

$$\mathbb{P}(S_{t+1}=s', R_{t+1}=r_t | A_t=a, S_t=s) \rightsquigarrow \rho_{t,a,s}(s', r)$$

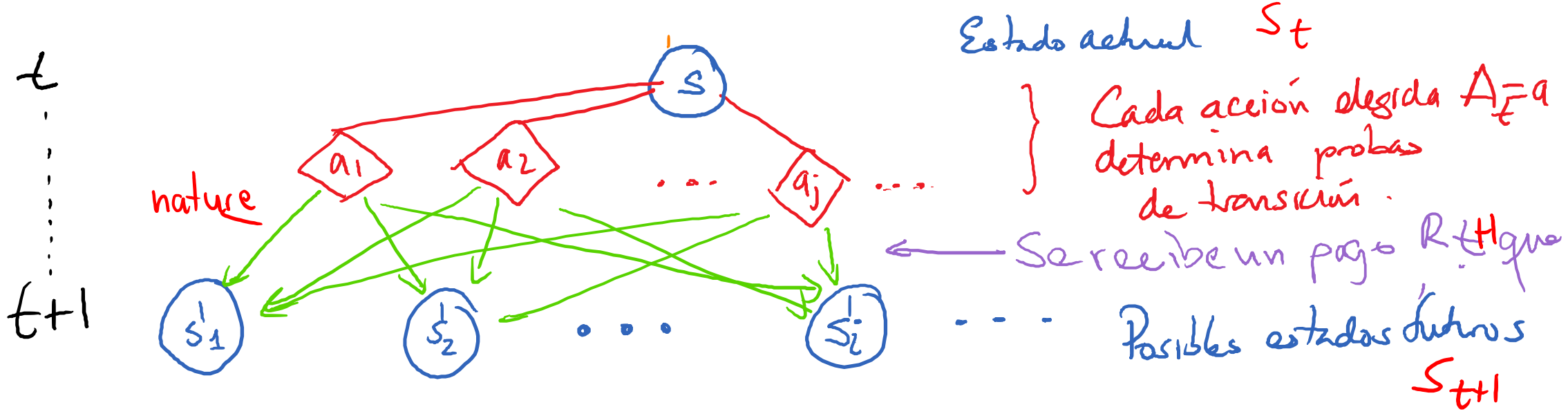
I. Procesos Markovianos de Decisión

- Una tarea de aprendizaje por reforzamiento es una **Proceso Markoviano de Decisión** (MDP) si cumple la propiedad markoviana, i.e., las probabilidades de transición dependen únicamente del estado y acción actual y no del pasado

$$\mathbb{P}(S_{t+1} = s_{t+1}, R_{t+1} = r_{t+1} \,|S_0 = s_0, R_0 = r_0, A_0 = a_0, \dots, S_t = s_t, R_t = r_t, A_t = a_t)$$
$$= \mathbb{P}(S_{t+1} = s', R_{t+1} = r_{t+1} | S_t = s_t, A_t = a_t)$$

Árbol de Decisión

- En cada momento del tiempo t, estado $S_t = s$, tenemos un árbol de decisión



Probas de transición entre estados dado acción.

$$\mathbb{P}(S_{t+1}=s' \mid S_t=s, A_t=a)$$

$$= \sum_r \mathbb{P}(S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a) = \sum_r p_{t,s,a}(s',r)$$

a pasos

$$\mathbb{P}(R_{t+1}=r \mid S_t=s, A_t=a) = \sum_{s'} \mathbb{P}(S_{t+1}=s', R_{t+1}=r \mid A_t=a, S_t=a)$$

$$= \sum_{s'} p_{t,s,a}(s',r)$$

Pagos esperado según estado y acción.

$$r_{t+1}(s,a) = \mathbb{E}(R_{t+1} \mid S_t = s, A_t = a)$$

$$= \sum_r r \underbrace{\mathbb{P}(R_{t+1} = r \mid S_t = s, A_t = a)}_{\text{marginal de } R_{t+1} \text{ en probas de transición}}$$

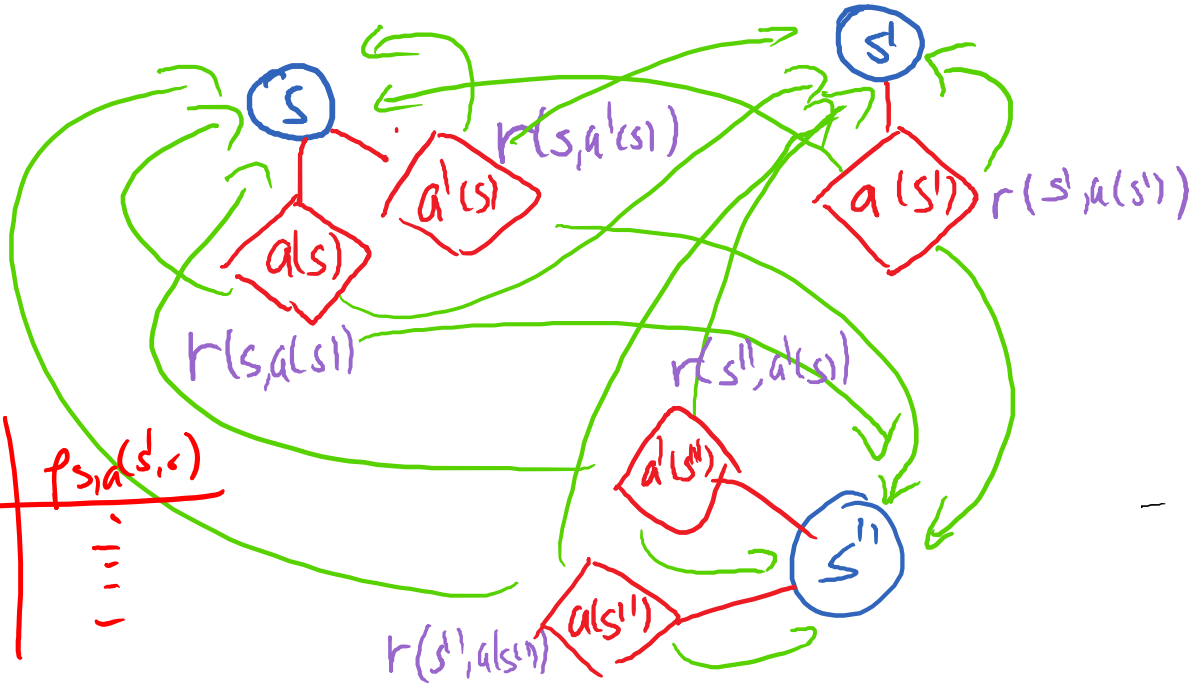
$$= \sum_r r \left[\sum_{s'} \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \right]$$

$$= \sum_r \sum_{s'} r \, p_{t,s,a}(s', r)$$

Diagrama de transición

$$p_{t,s,a}(s',r) = p_{s,a}(s,r)$$
$$\Rightarrow r_{t+1}(s,a) = r(s,a)$$

- Cuando el MDP es **homogéneo**,i.e., las probabilidades de transición no dependen del tiempo, entonces podemos englobar todos los árboles de decisión en un **diagrama de transición y decisión**



tabla

s	a	s'	r	$p_{s,a}(s',r)$
⋮	⋮	⋮	⋮	⋮

Estados
Acciones
Probabilidades
Pagos

Ejemplos

- Dominó
- Robot que recicla (libro)
- Trading USD

desde la perspectiva del jugador.

$S = \{ \text{tablero} + \text{mano} + \dots \}$ $A(s) = \{ \text{fichas a tirar} \}$

$Pagos = \{ 0 \text{ en cada jugada}, 1 \text{ cuando ganas} \}$

Pruebas de transacción dependen de los otros

Opciones de a millón de USD

Estados $S = \{ \text{tener}, \text{no tener}, \text{info mercado} \}$

$A(\text{tener}) = \{ \text{mantener}, \text{vender} \}$

$A(\text{no tener}) = \{ \text{comprar}, \text{no comprar} \}$

$Pagos = \{ \underbrace{+1}_{\text{gane dinero}}, \underbrace{-1}_{\text{perdí dinero}} \}$ discreto

$= \{ \text{cantidad ganada} \}$ continuo

II. Utilidad esperada

El retorno **retorno total restante** o **utilidad total restante** es una función

$$U_t = U_t(R_{t+1}, R_{t+2} \dots)$$

que mide la utilidad que recibiremos de los pagos del tiempo t hacia el futuro.

No hay manera única definir la función de utilidad. Vamos a ver la forma más común dependiendo del horizonte de la tarea

a. Tareas periódicas

- Algunas tarea de aprendizaje son de forma natural **tareas periódicas.** Estas son las tareas que **tienen un fin** y vuelven a empezar. Cada repetición de la tarea es un **episodio.**
- Para estas tareas hay un tiempo máximo T (puede depender del proceso mismo)
- En estos, casos una elección razonable de utilidad es simplemente la suma de los pagos

$$U_t = R_{t+1} + \cdots + R_T = \sum_{i=1}^{T-t} R_{t+i}$$

Ejemplos de tareas periódicas

- Juegos de mesa (ganar/perder)
- Jugar golf
- Escapar de un laberinto

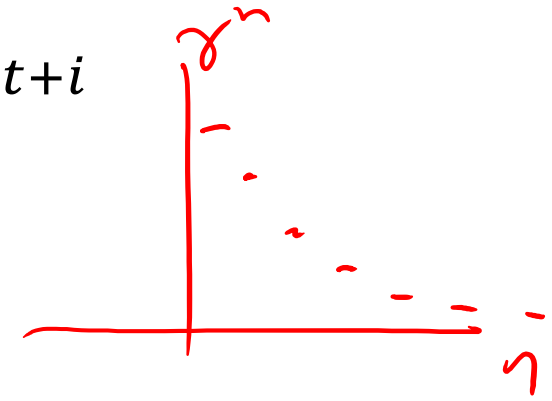
b. Tareas continuas

- Son aquellas que **no tienen un fin definido**
- Si $T = \infty$ no podemos solo sumar los retornos pues todo sería infinito
- Para poder definir una función de utilidad se usan pesos/tasas de descuento
- La solución se inspira en las ***tasas de interés*** y el *valor del dinero*

$$U_t = \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{i=1}^{\infty} \gamma^i R_{t+i}$$

con $0 < \gamma < 1$

Utilidad en valor presente



- Siempre se cumple $U_t < \infty$:

$$U_t = \sum_{i=1}^{\infty} \gamma^i R_{t+i} \leq \max_i \{R_{t+i}\} \sum_{i=1}^{\infty} \gamma^i$$

$$= \max_i \{R_{t+i}\} \frac{\gamma}{1-\gamma} < \infty$$

Ejemplos de tareas continuas

- Trading USD
- k-máquinas de dinero
- Aprender a manejar (¿?)
- ???

- No tenemos obligación de usar necesariamente estas funciones de utilidad. Podemos usar funciones de utilidad que incluyan, por ejemplo, aversión al riesgo **(Risk-sensitive Reinforcement Learning)**
- ¿Cómo es la utilidad del dinero?

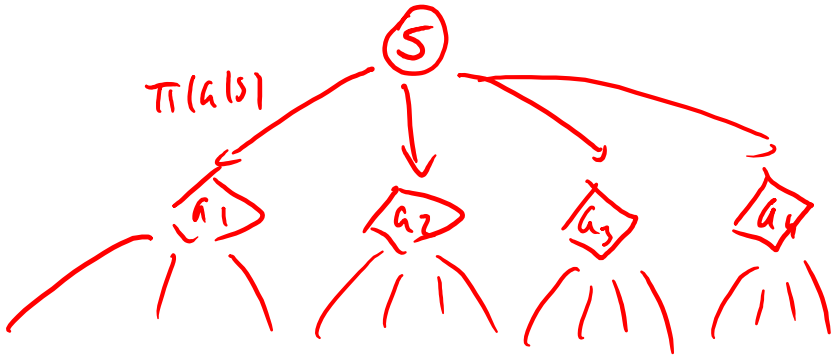
III. Políticas

- ¿Cuál es el **objetivo** del aprendizaje por reforzamiento?
- Encontrar una **política** que maximice la utilidad del agente
- Una política dicta cómo acciones el agente según el estado actual
- Es un mapeo $\pi_t(a|s) \coloneqq \mathbb{P}(A_t = a|S_t = s)$
- Usualmente π_t no depende de t y solo escribimos π

IV. Valor esperado, Funciones de valor

- El **valor** de una acción/estado va a depender de la política y está relacionado con la utilidad
- Es la cantidad más importante para **evaluar** la calidad de una acción/estado

a. Valor de una acción



estados + pagos + valor a empezar

- Para poder aprender, necesitamos saber la utilidad esperada dada una acción, en función de una política
- Es costumbre usar la letra q para la función de valor y hacer explícita la dependencia de la política π_t

todos los pagos del futuro

$$q_{\pi_t}(s,a) = \mathbb{E}_{\pi_t}(U\_t|S_t = s, A_t = a)$$

- En problemas homogéneos no hay dependencia del tiempo y simplemente escribimos q_π
- El problema de aprendizaje del agente es inferir q_π para cada valor (s,a)

q en el caso homogéneo

$$G_t = R_{t+1} + \dots + R_T$$

- Caso por episodios

$$q_\pi(s,a) = \mathbb{E}_\pi(G_t \mid S_t = s, A_t = a) \overset{\text{saca sumas}}{=} \sum_{i=1}^{T} \mathbb{E}(R_{t+i} \mid S_t = s, A_t = a)$$

$$= \sum_{i=1}^{T} r_{t+i}(s,a)$$

- Caso continuo $G_t = \sum_{i=1}^{\infty} \gamma^i R_{t+i}$

$$q_\pi(s,a) = \sum_{i=1}^{\infty} \gamma^i r_{t+i}(s,a)$$

b. Valor de un estado

- El valor de los estados es una función que calcula la utilidad esperada dado un estado, en función de las posibles acciones de la política

$$v_{\pi_t}(s) = \mathbb{E}_{\pi_t}(U_t|S_t = s_t) = \sum_{a\in\mathcal{A}(s)} \mathbb{E}(U_t|S_t = s_t, A_t = a)\mathbb{P}(A_t = a|S_t = s)$$

esperanza total

$$= \sum_{a\in\mathcal{A}(s)} q_{\pi_t}(s,a)\,\pi_t(a|s)$$

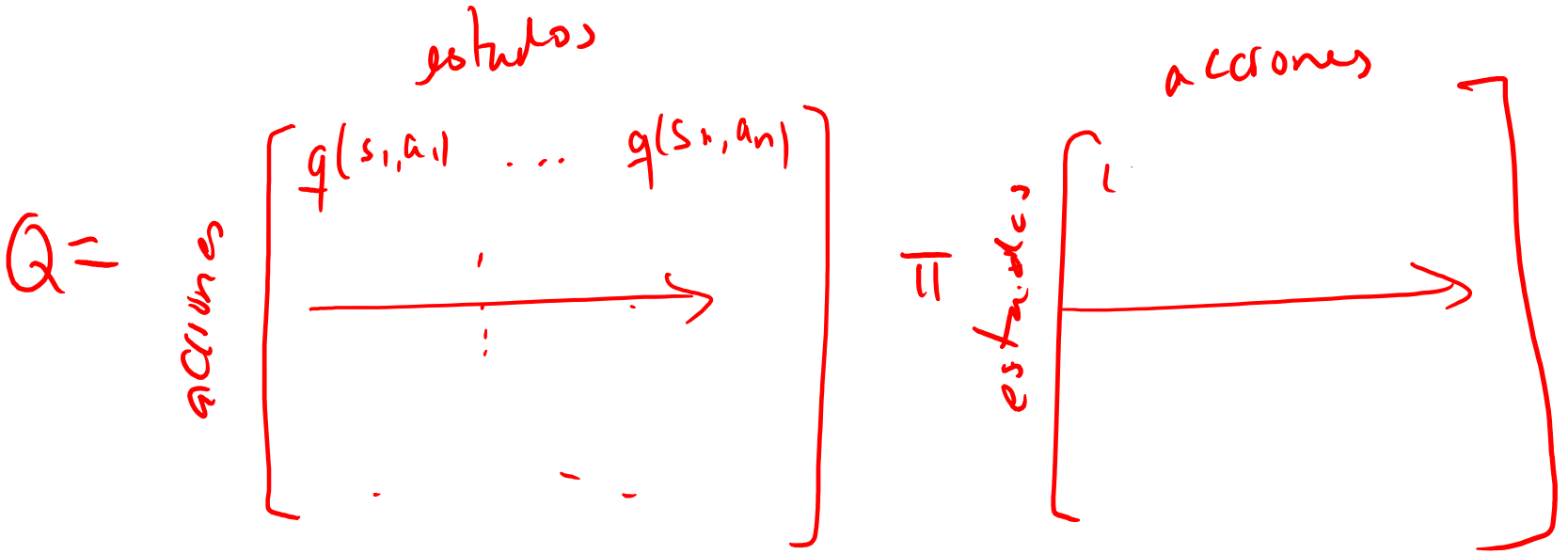
- En problemas homogéneos la función de valor no depende de t y escribe simplemente $v_\pi(s)$

En forma matricial…

$$V_\pi = Q\pi$$

$$V_\pi(s) = Q_s \cdot \pi_s$$

donde



Tarea individual 1

- Buscar un ejemplo de problema de aprendizaje por reforzamiento distinto a los que hemos dado
- Identificar los elementos de un problema de aprendizaje por reforzamiento y MDP
- Argumentar si se cumple la propiedad markoviana
- Dar uno o varios ejemplos de políticas
- Máximo dos cuartillas a computadora o a mano
- Entregar escrito/impreso la próxima clase