Proceso de Decisión de Markov

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Septiembre de 2024

Contenido

Introducción

- Modela el ambiente en el que tiene lugar el apredizaje por refuerzo.
- Es lo suficientemente general para capturar la idea de que el ambiente en el que se toma la decisión es dinámico y cambia con las decisiones.
- Por ejemplo, en el caso no asociativo la acción óptima se elige a partir de una estimación $q(a)$ (i.e., bandidos multibrazo), mientras que en el caso asociativo es a partir de $q(a, s)$ donde s describe el estado (i.e., ambiente) en el momento de la decisión.

Introducción

- Estudiaremos la construcción paso a paso de un proceso de decision de Markov:
  1. Proceso Markoviano.
  2. Proceso Markoviano de Recompensas.
  3. Proceso de Decisión de Markov.
- Para esto utilizaremos algunos ejemplos tomados del curso de RL de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Proceso Markoviano (básico)

- Estados $S = \{Facebook, Class1, etc\}$.
- Función de transición $P : S \to \Delta(S)$

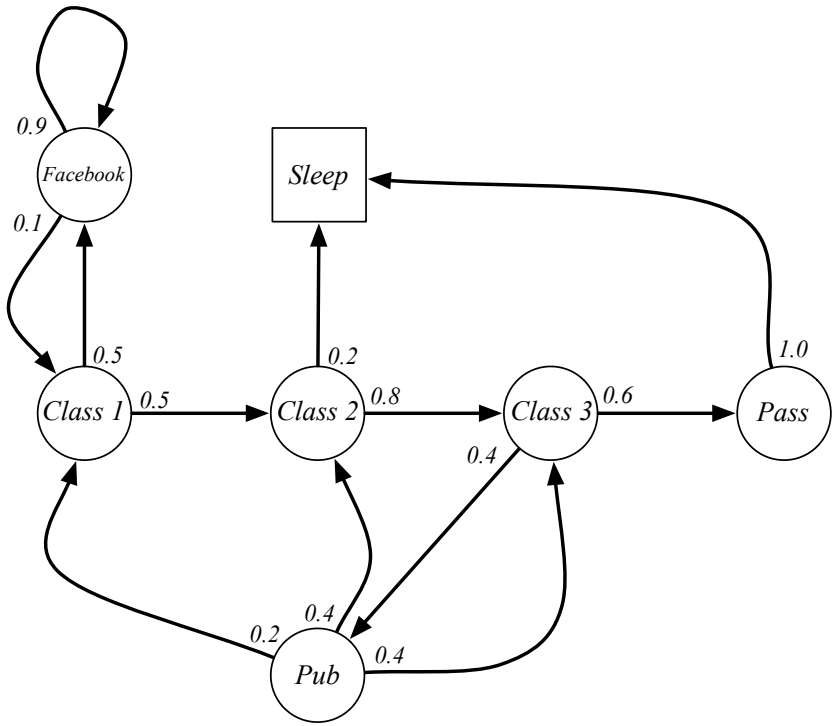
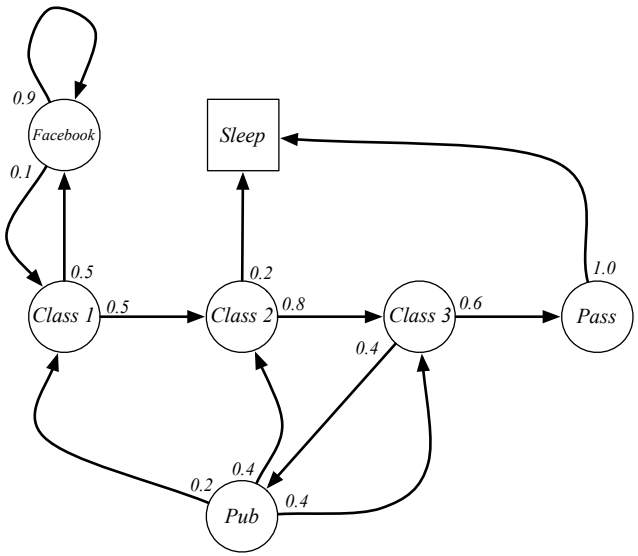


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Proceso Markoviano (básico)



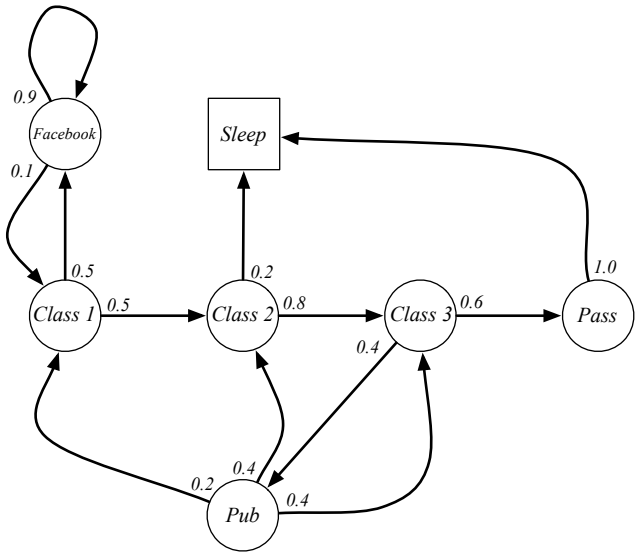
Sample episodes for Student Markov Chain starting from $S_1 = \text{C1}$

$$S_1, S_2, ..., S_T$$

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Proceso Markoviano (básico)



$$\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array} \begin{array}{c} \begin{array}{ccccccc} C1 & C2 & C3 & Pass & Pub & FB & Sleep \end{array} \\ \begin{bmatrix} & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{bmatrix} \end{array}$$

Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Proceso Markoviano de Recompensas

- Enriquecemos el modelo anterior suponiendo que hay una función de recompensa $R : S \to \mathbb{R}$.
- Con esta función definimos la función de retorno (i.e., objetivo):

$$G_t = R_{t+1} + \lambda R_{t+2} + ... + \lambda^{T-t-1} R_T \tag{1}$$

  En esta ecuación R_{t+1} queda determinado por S_t aunque sea una recompnesa inmediata en t. **La lógica de esta notación es que es necesario primero observar S_t antes de conocer R_{t+1}.**
- En algunos textos se usa R_t en vez de R_{t+1}.
- Por último, esta función juega un papel fundamental en la teoría. Función Valor de estado:
  $v(s) = E[R_{t+1} + \lambda R_{t+2} + ... + \lambda^{T-t-1} R_T \mid S_t = s]$

Proceso Markoviano de Recompensas

- La figura muestra la recompensa en cada estado (e.g., $R(Facebook) = -1$) y en rojo la funcion valor del estado (e.g., $v(C2) = 1{,}5$, véase la próxima diapositiva)

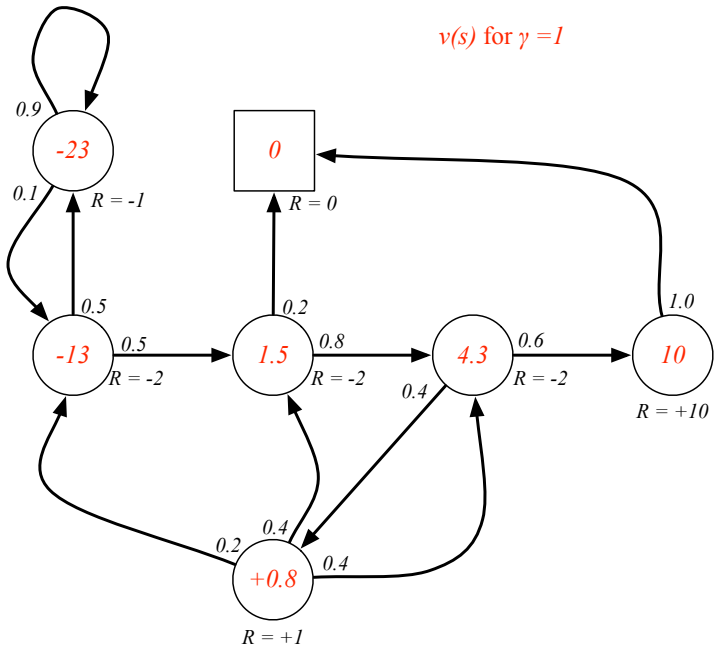


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf). $v(s)$ se estima simulando episodios desde s.

Proceso Markoviano de Recompensas

- Para calcular la función valor de cada estado se simulan varios episodios y se calcula el retorno en cada episodio. La función valor en el estado es el promedio del retorno de los episodios.
- Por ejemplo la siguiente figura ilustra el procesos para estimar $v(C1)$ con $\gamma = \frac{1}{2}$ en el ejemplo anterior.

Sample returns for Student MRP:
Starting from $S_1 = \text{C1}$ with $\gamma = \frac{1}{2}$

$$G_1 = R_2 + \gamma R_3 + ... + \gamma^{T-2} R_T$$

C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	=	-2.25
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	=	-3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	=	-3.41
C1 FB FB C1 C2 C3 Pub C1 ...	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	=	-3.20
FB FB FB C1 C2 C3 Pub C2 Sleep			

Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf). $v(s)$ se estima simulando episodios desde s.

Proceso Markoviano de Recompensas

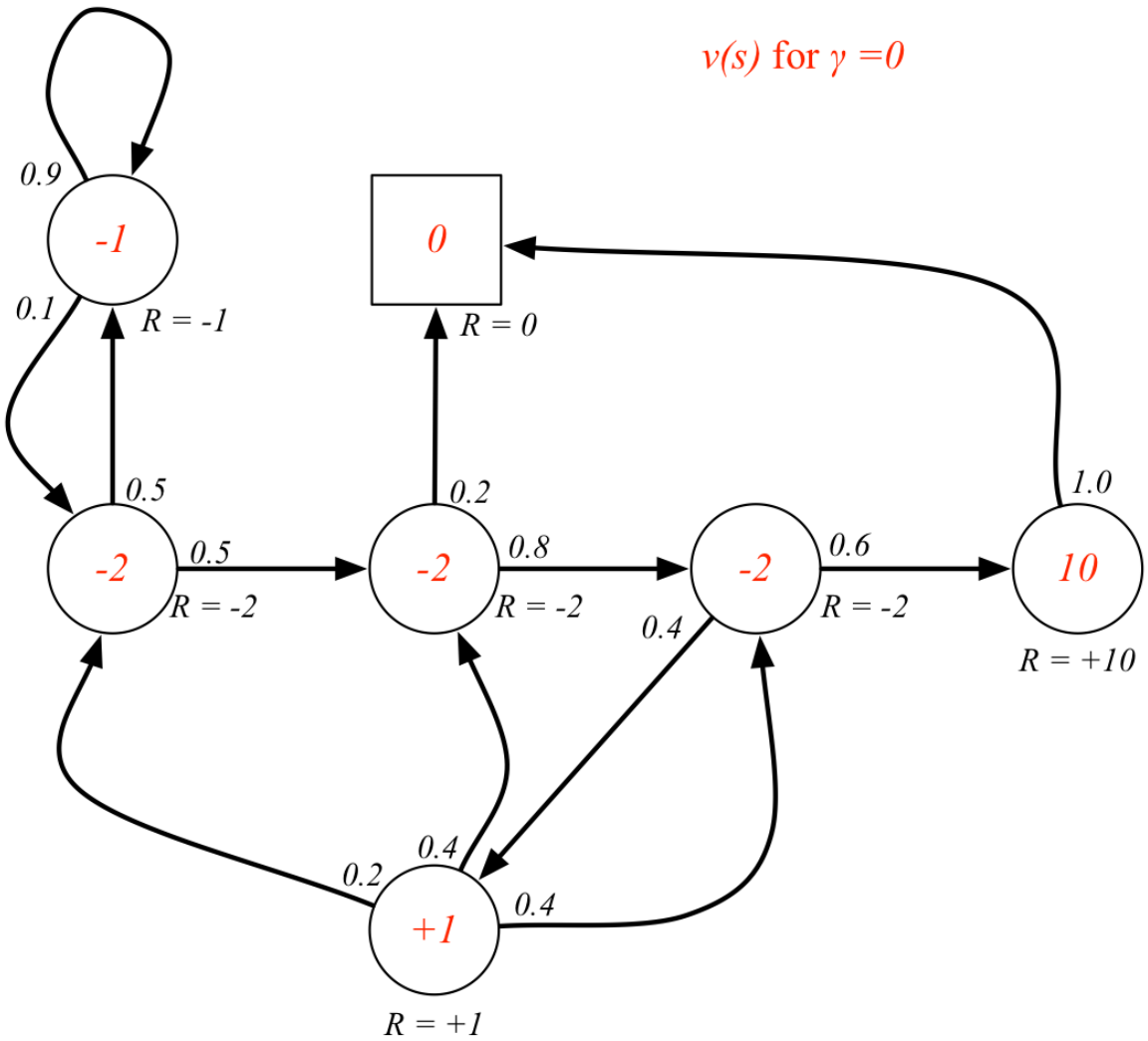


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf). $v(s)$ se estima simulando episodios desde s.

Proceso Markoviano de Recompensas

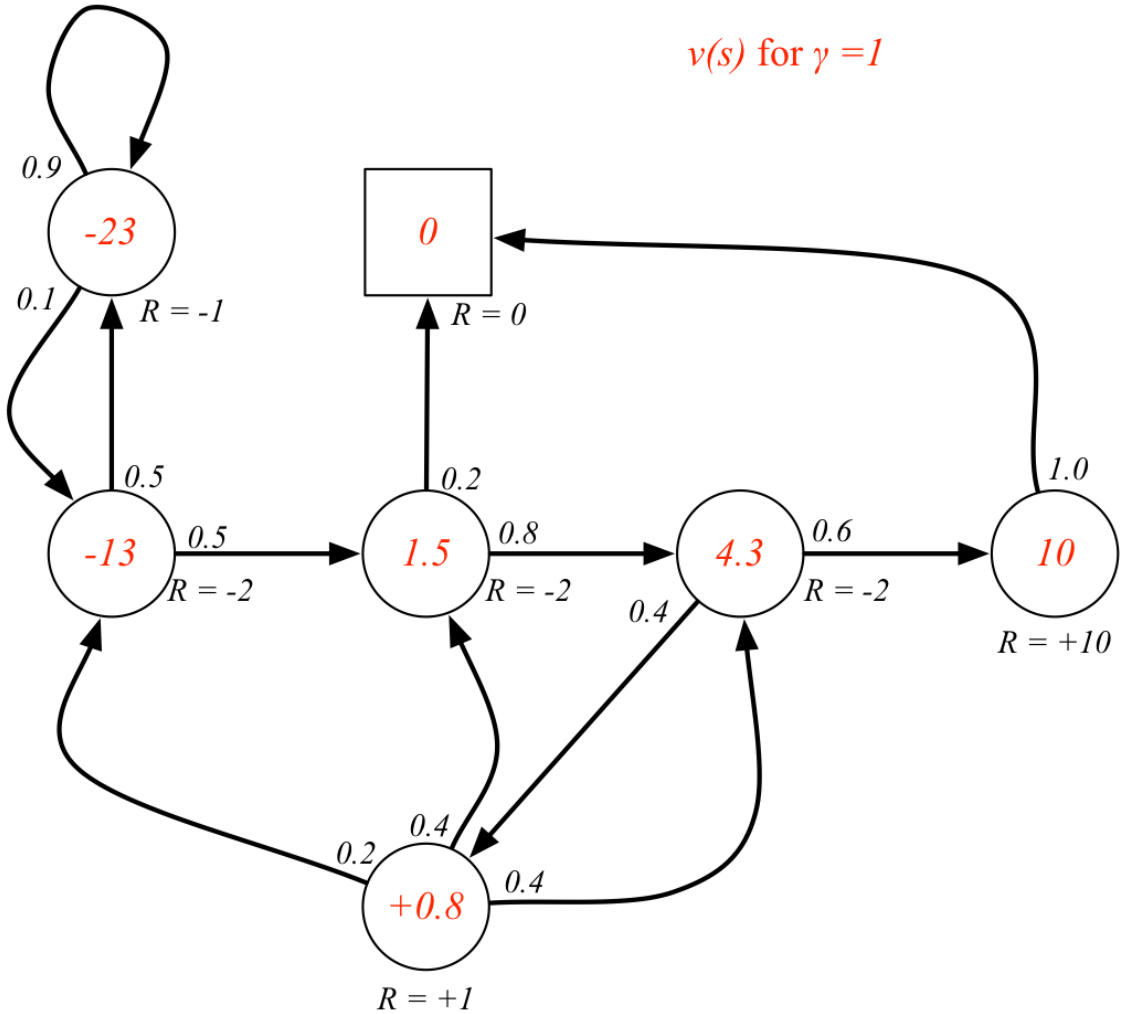


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf). $v(s)$ se estima simulando episodios desde s.

Contenido

MDP

- Los elementos principales del modelo de decisión markoviano son:
  1. El modelo del ambiente.
  2. La función de recompensa.
  3. La función valor (i.e., recompensa de varios periodos).
  4. La función de política (i.e., la decisión en cada periodo).

MDP

- Un MDP es una generalización de un proceso markoviano.
- Un **Proceso de Markov** es un proceso estocástico S_t en un conjunto de estados S tal que:

$$P[S_{t+1} = s' \mid S_t = s, S_{t-1}, ..., S_0] = P[S_{t+1} = s' \mid S_t = s] \tag{2}$$

- Denotamos esta probabilida de transición por: $P : S \to \Delta(S)$ donde $\Delta(S)$ es el conjunto de distribuciones de probabilidad sobre S.
- En un **Proceso de Decisión de Markov** (MDP), las probabilidades de transición entre los estados dependen de las acciones de un agente.
- La probabilidad de transición en este caso es de la forma: $P : S \times A \to \Delta(S)$ donde A es el conjunto de acciones del agente.

MDP

- Un MDP es (S, A, P, R, γ) donde:
  1. S es un conjunto de estados.
  2. A es conjunto de acciones.
  3. $P : S \times A \to \Delta(S)$ representa las probabilidades de transición.
  4. $R : S \times A \to \mathbb{R}$ es una función de recompensa (instantánea).
  5. $\gamma \in [0, 1]$ es un factor de descuento intertemporal de recompensas futuras.

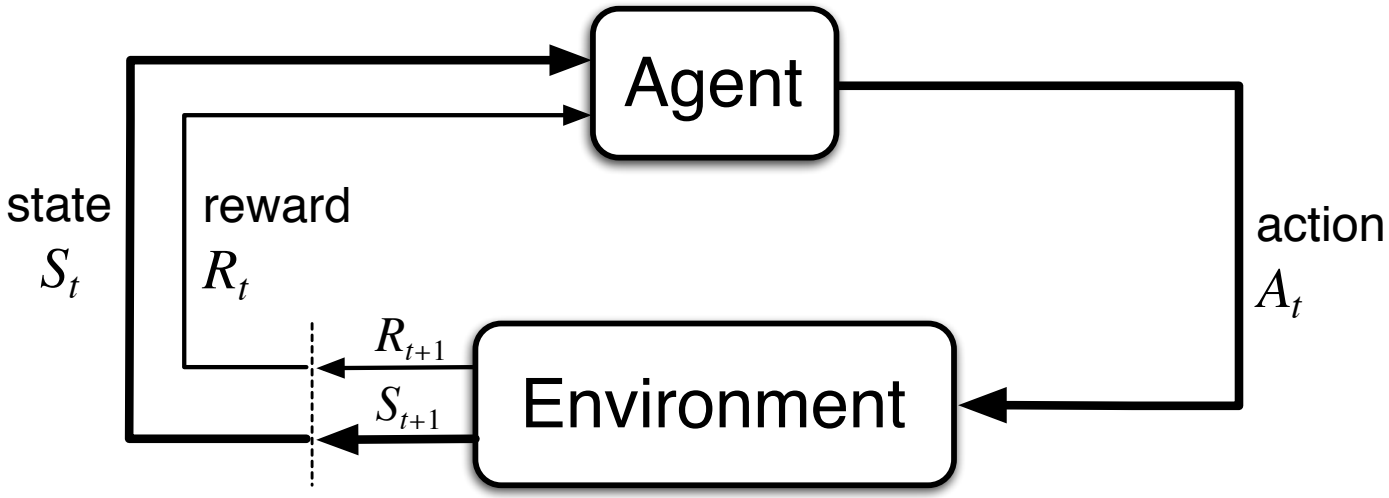


Figure 3.1: The agent–environment interaction in a Markov decision process.

- Implícito en este dibujo es la forma como se actuliza la información. (S_t, A_t) determinan la recompensa R y estado S en $t+1$, R_{t+1} y S_{t+1}.

$$(S_t, A_t) \rightarrow (S_{t+1}, R_{t+1}) \tag{3}$$

Proceso de Decisión Markoviano (básico)

- Función de transición **determinística** dados (s, a) excepto en el estado PUB.

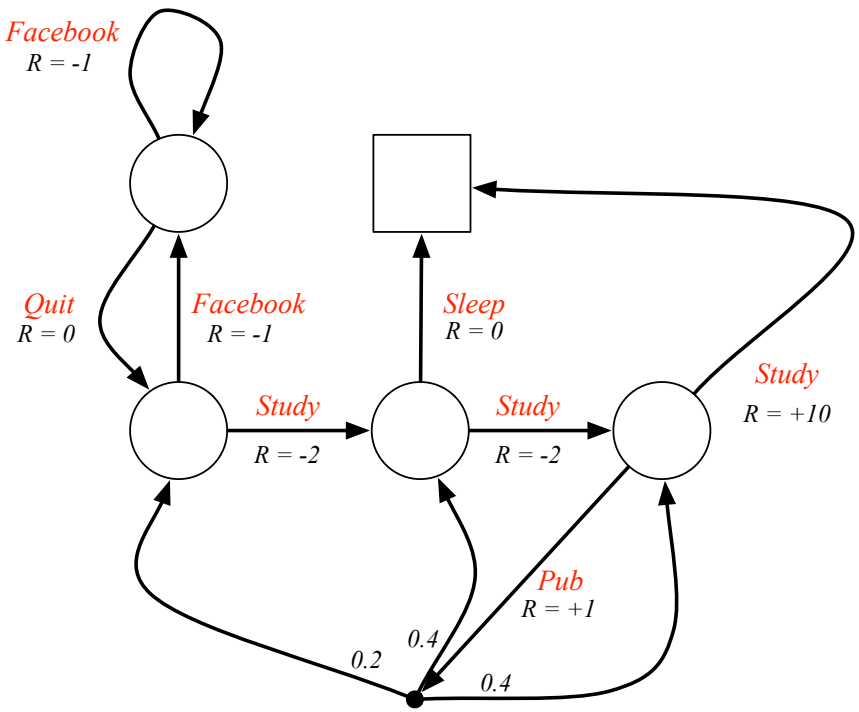


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

- La dinámica completa de un MDP se puede describir con la distribucion:

$$p(s', r' \mid s, a) = P(s_{t+1} = s', R_{t+1} = r' \mid S_t = s, A_t = a) \quad (4)$$

o frecuentemente la escribimos como:

$$p(s', r \mid s, a) = P(s_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \quad (5)$$

- La forma más común de utilizar la dinámica del MDP es:

$$p(s' \mid s, a) = \sum_r P(s_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \quad (6)$$

Contenido

Función de Política

- El **objetivo** es maximizar el valor esperado del retorno (i.e., objetivo):

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^{T-t-1} R_T$$

  donde T es la fecha de finalización de la interacción con el ambiente. Por ejemplo el momento en el que termina un juego. Cada juego o interacción completa se denomina un episodio.
- Una **función de política** es $\pi : S \to \Delta(A)$. Es una función independiente del tiempo.

Función Valor del Estado

- Podemos definir dos funciones valor: (1) Función valor del estado para la política π y (2). Función valor de la acción para la política π.
- La función valor del estado es:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

  donde se inicia en el estado s y a partir de ese momento se continua usando la función de política π.
- La notación aterior resalta que el valor esperado depende de π más alla del MDP que está en el trasfondo.
- Más precisamente, **una simulación del proceso estocástico subyacente** sería:

$$s \rightarrow a = \pi(s) \rightarrow s' \sim p(s' \mid s, a) \rightarrow \ldots \tag{7}$$

Función Valor de la Acción

- La función valor de la acción:

$$q_\pi(s, a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

donde el sistema comienza en el estado s, la primera acción es a y a partir de ese momento se sigue la política π.

Contenido

Función valor óptima

- La función valor optima se define como:

$$v_*(s) = \max_{\pi} v_\pi(s)$$

$$v_*(s) = \max_{\pi} E_\pi[R_{t+1} + \gamma R_{t+2} + ... + \gamma^{T-t-1} R_T \mid s] \quad (8)$$

Función de política óptima

- La función de política óptima π_* es el argumento que resuleve el anterior problema. Es decir:

$$\pi_*(s) = argmax_\pi v_\pi(s) \tag{9}$$

- Entonces: $v_*(s) = v_{\pi_*}(s)$.

Proceso de Decisión Markoviano (básico)

- Funcion valor del estado óptima para ejemplo anterior (gran parte del desarrollo de la teoría es aprender a calcularla de forma eficiente).

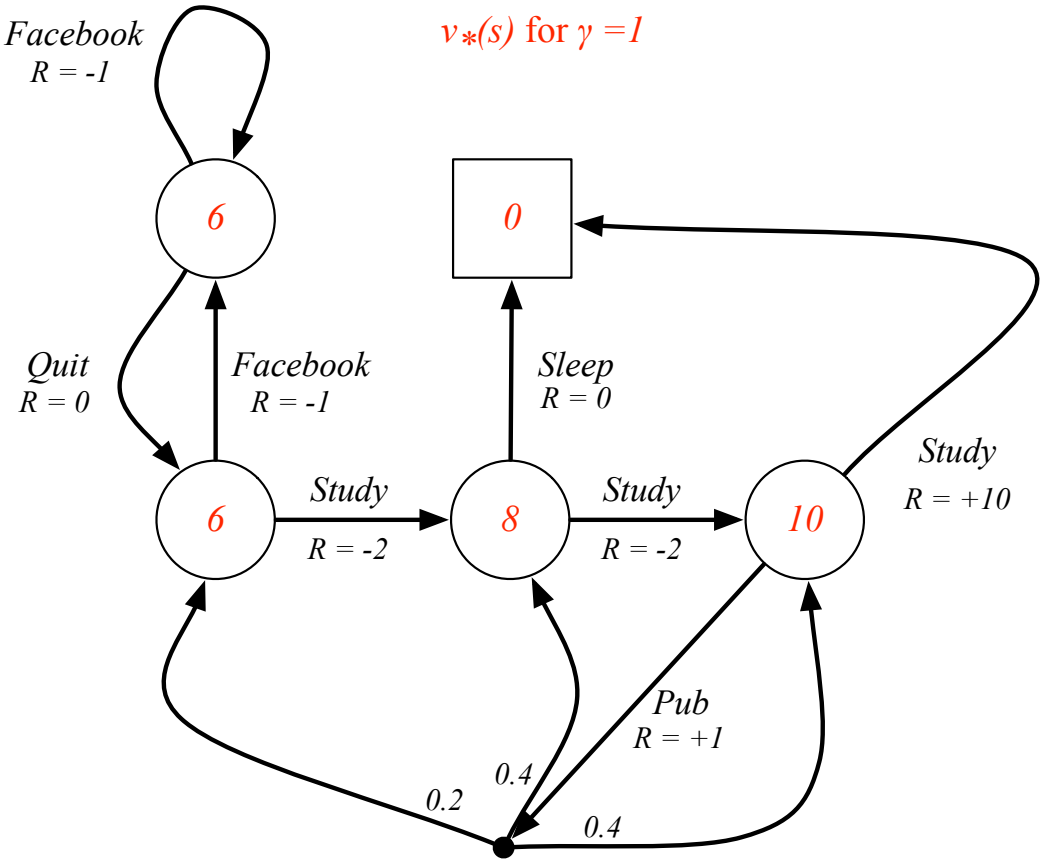


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Proceso de Decisión Markoviano (básico)

- Funcion de política óptima para ejemplo anterior.

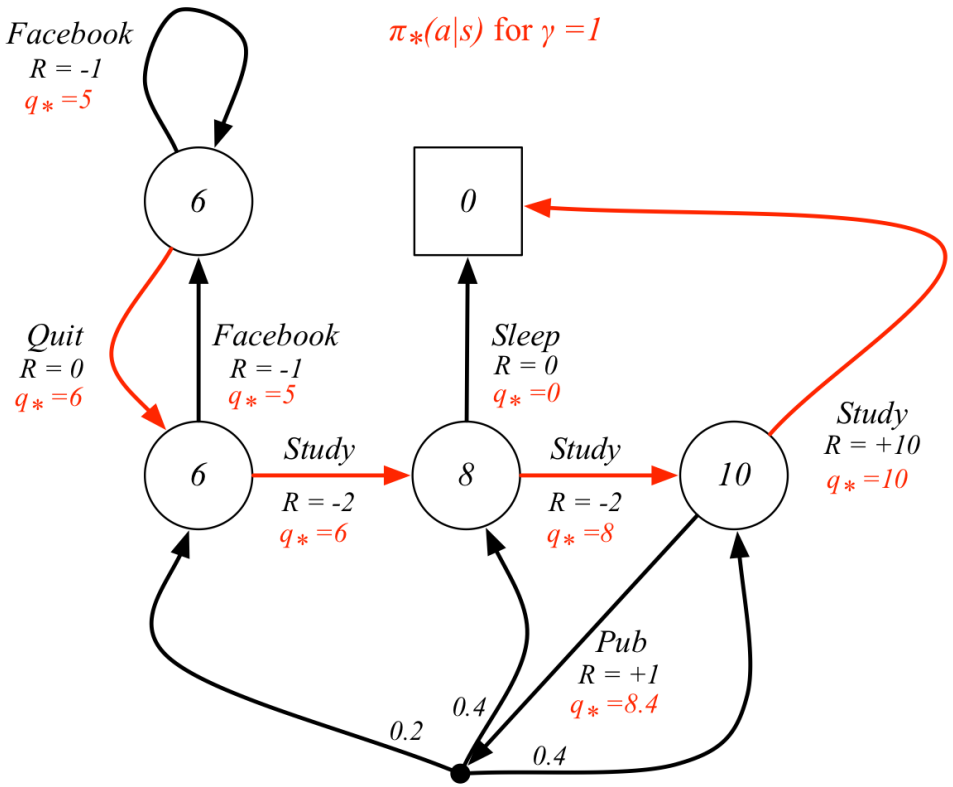


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Función valor óptima del estado y acción

- La función de valor óptima del estado y acción se define como:

$$q_*(s, a) = \max_{\pi} q_\pi(s, a)$$

Proceso de Decisión Markoviano (básico)

- Funcion valor óptima de la acción para ejemplo anterior.

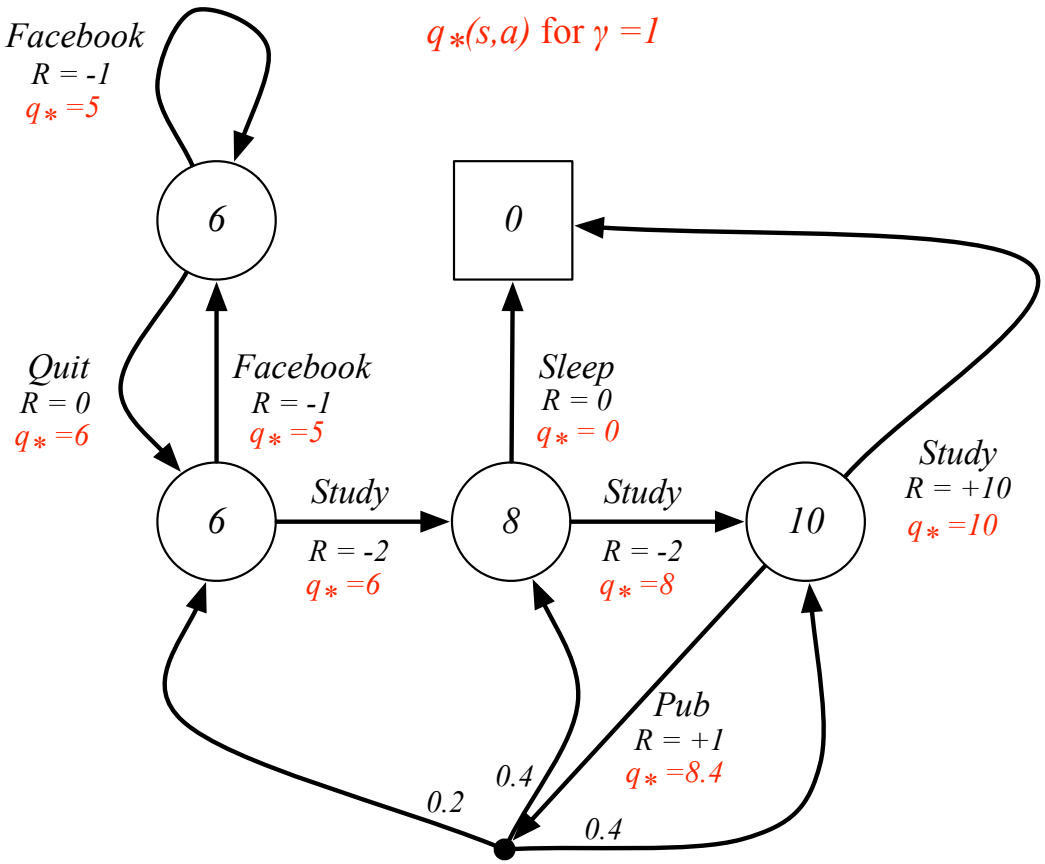


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Lo que viene...

- Gran parte de la teoría que desarrollaremos tiene como propósito enseñar a calcular $v_*(s)$, $\pi_*(s)$ y $q_*(s, a)$.
- Por el momento vamos a dar algunos ejemplos económicos relevantes de este formalismo.

Contenido

Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- El problema de crecimiento económico es:

$$\max_{(c_t)_{t=0,1,..}} \sum_{t=0}^{\infty} \beta^t \log(c_t) \tag{10}$$

$$k_{t+1} = k_t^{\alpha} - c_t, k_0 \text{ dado.} \tag{11}$$

donde $\alpha \in (0,1)$, k_t es el stock de capital de una economía y c_t es el consumo agregado.

Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- Como problema de RL:
  - Sea $S = R_+$, $A = R_+$, $R(k, c) = \log(c)$, donde k es el estado y c es la acción.
  - La dinámica es determinística: $k_{t+1} = k_t^{\alpha} - c_t$.
  - $s_t = k_{t-1}, a_t = c_{t-1}$, $T = \infty$, $\gamma = \beta$.
  - El problema de RL:

$$\max_{\pi} \log(c_t) + \gamma \log(c_{t+1}) + ... \tag{12}$$

Modelo básico de crecimiento Brock - Mirman (1972) estocástico

- El problema de crecimiento económico es:

$$\max_{(c_t)_{t=0,1,..}} E[\sum_{t=0}^{\infty} \beta^t \log(c_t)] \tag{13}$$

$$k_{t+1} = z_t k_t^{\alpha} - c_t, k_0 \text{ dado.} \tag{14}$$

donde $\alpha \in (0,1)$, k_t es el stock de capital de una economía y c_t es el consumo agregado, $\log(z_t) \sim i.i.d, N(0, \sigma^2)$ es la productividad total de los factores.

Modelo básico de crecimiento Brock - Mirman (1972) estocástico

- Como problema de RL:
  - Todo igual al ejemplo anterior excepto que la dinámica ahora es estocástica. $p(k_{t+1} \mid k_t, c_t)$ depende de la variable aleatoria z_t.

Robot que debe alcanzar estado terminal en grilla

- 15 estados.
- Las acciones son moverse en cualquiera de los cuatro direcciones cardinales. Sin embargo, si en algun caso la acción saca a el agente del tablero, entonces se queda donde estaba.
- Si la acción es determinística la transición también lo es.

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

Table 1: Small Gridworld: A simple example for MDP. In this Small Gridworld, each square with number on it is a non-terminal state. The two shaded squares are the terminal state. The goal is to make a robot move to the terminal state from any non-terminal state.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

- 15 estados, $\gamma = 1$, recompensa $= -1$ siempre hasta llegar a estado terminal.
- Acciones $= \{$N,W,E,S$\}$ si la acción saca al robot de la grilla (e.g., una función de política aleatoria) entonces no cambia de estado.

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

Table 1: Small Gridworld: A simple example for MDP. In this Small Gridworld, each squa number on it is a non-terminal state. The two shaded squares are the terminal state. The to make a robot move to the terminal state from any non-terminal state.