Навчання з підкріпленням

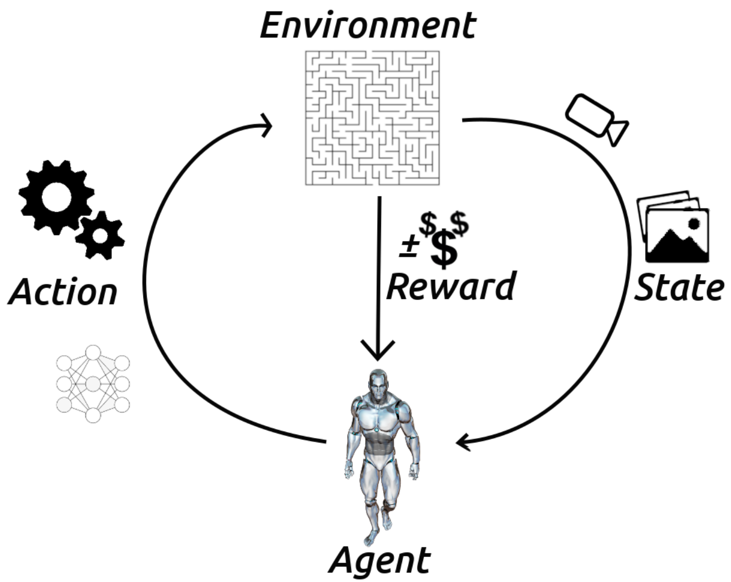
Лекція 2: Марковські процеси прийняття рішень

Кочура Юрій Петрович
iuriy.kochura@gmail.com
@y_kochura

Сьогодні

- Марківські процеси
- Марківські процеси винагороди
- Марківські процеси прийняття рішень (МППР)

Цикл взаємодії



Мета — оптимізувати загальну винагороду, отриману агентом при взаємодії з навколишнім середовищем.

Вступ до МППР

- Марківські процеси прийняття рішень формально описують середовище для навчання з підкріпленням
- Там, де середовище є повністю оглядовим
- Поточний стан агента повністю характеризує процес
- Майже всі задачі RL можна формалізувати як МППР
    - Оптимальне управління насамперед стосується безперервних МППР
    - Задачі в частково оглядовому середовищі можуть бути зведені до МППР

Властивість Маркова

> Майбутнє процесу не залежить від минулого, а залежить лише від поточного стану

Стан S_t є Марківським тоді і тільки тоді

$$\boxed{\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \cdots, S_t]}$$

- Це означає, що поточний стан агента містить все, що нам потрібно знати з його історії
- Як тільки стан стане відомим, історію можна буде відкинути
- Тобто, стан — це достатня статистика для майбутнього

Властивість Маркова

Щоб перевірити своє розуміння властивості Маркова, розглянемо декілька задач управління або задач прийняття рішень і подивимось, які з них мають властивість Маркова:

- Водіння автомобіля
- Рішення інвестувати в акції чи ні
- Вибір лікування пацієнта
- Діагностика хвороби пацієнта
- Передбачити, яка команда виграє у футбольному матчі
- Пошук найкоротшого маршруту (найкоротшого) до певного пункту призначення
- Наведення прицілу гармати на постріл у далеку мішень

Матриця зміни стану **(state transition matrix)**

Ймовірність переходу між Марківськими станами $s \to s'$, визначається так:

$$\boxed{\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]}$$

Матриця зміни стану $\mathcal{P}$ визначає ймовірності переходу між усіма станами s у всі можливі стани s':

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix},$$

де кожен рядок матриці у сумі дорівнює 1.

Марківський процес

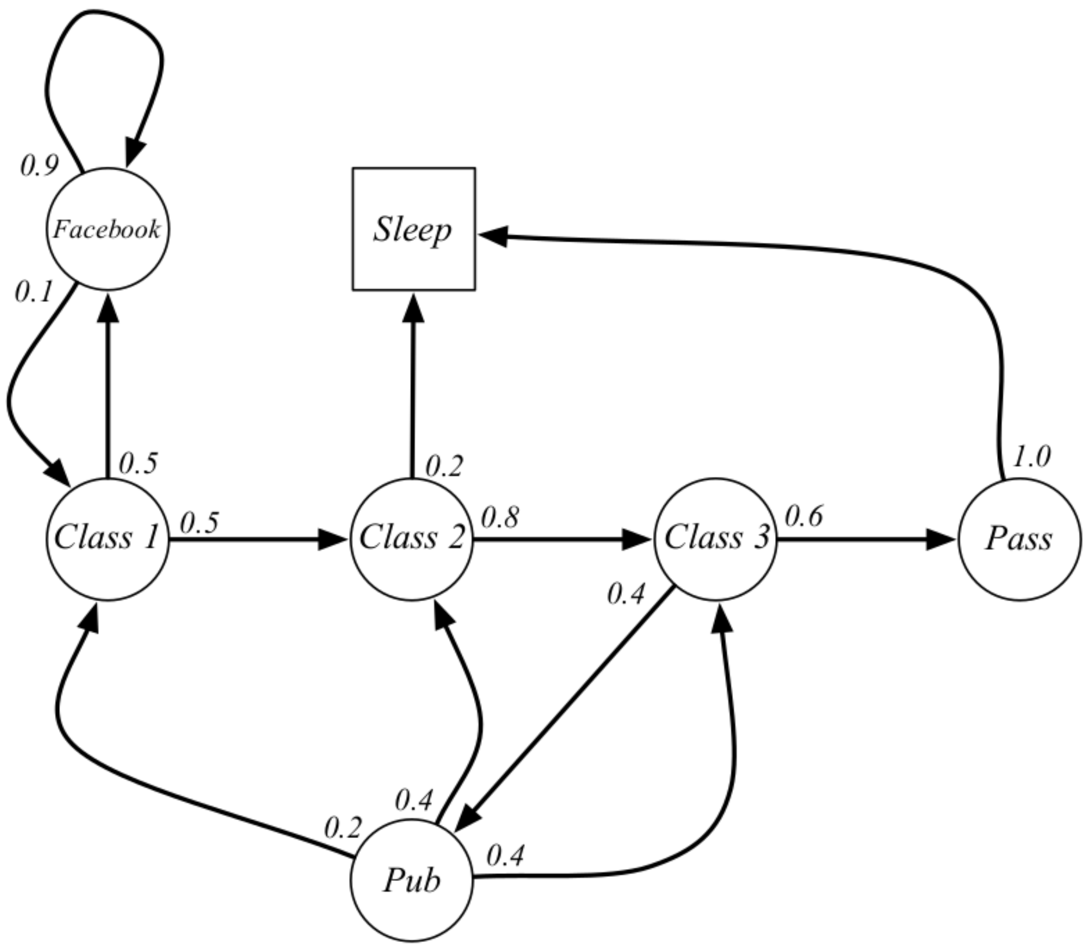
Марківський процес — це випадковий процес у якого відсутня пам'ять, тобто послідовність випадкових станів $S_1, S_2, \cdots$, які володіють властивістю Маркова.

Марківський процес (або ланцюг Маркова) — це кортеж $\langle \mathcal{S}, \mathcal{P} \rangle$:

- $\mathcal{S}$ — скінченна множина станів
- $\mathcal{P}$ — матриця зміни стану: $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$

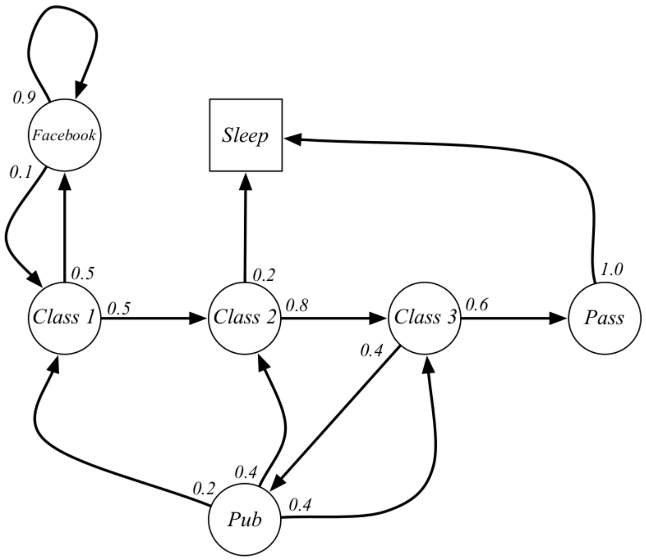
Приклад

Студентський ланцюг Маркова



—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Студентський ланцюг Маркова

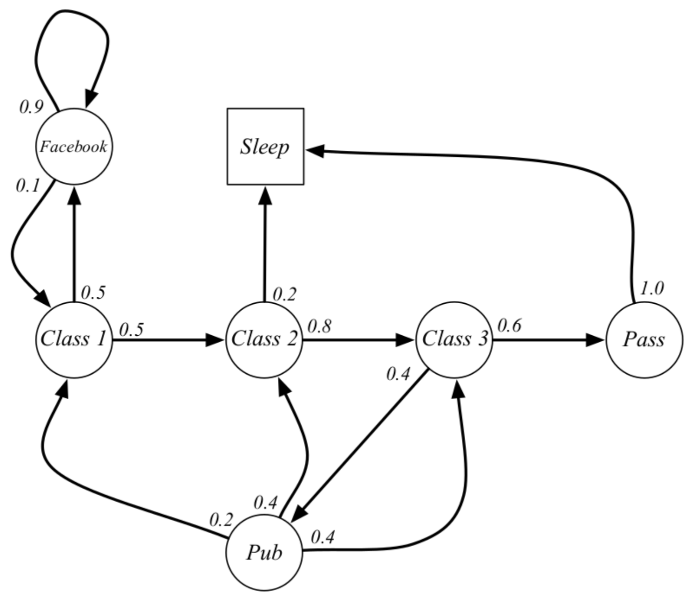


Початковий епізод починається з $S_1 = C_1$

$$S_1, S_2, \cdots, S_T$$

- 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

Студентський ланцюг Маркова: матриця зміни стану



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

—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

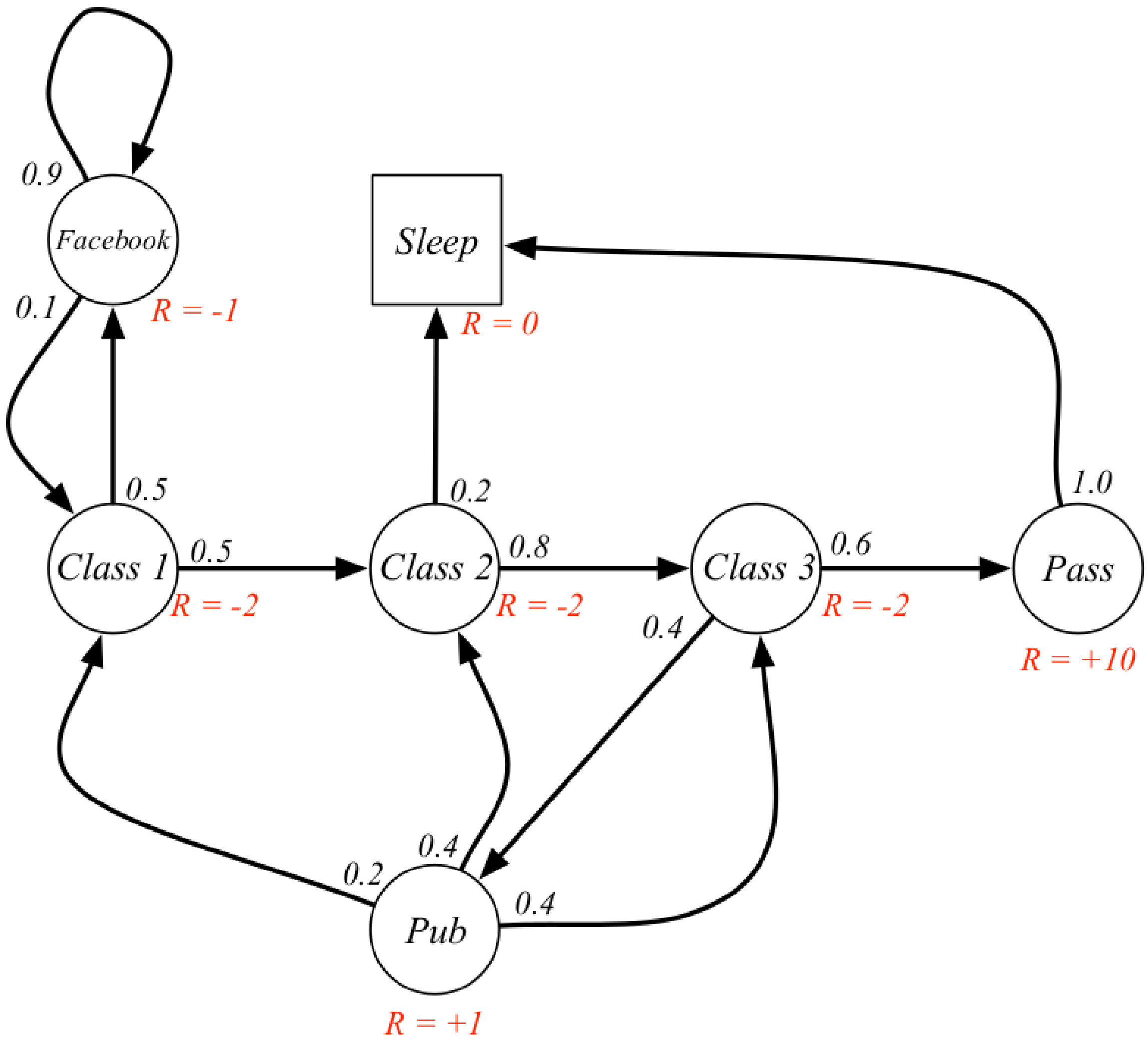
Марківські процеси винагороди

Марківський процес винагороди — ланцюг Маркова з винагородою.

Марківський процес винагороди — це кортеж $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$:

- $\mathcal{S}$ — скінченна множина станів
- $\mathcal{P}$ — матриця зміни стану: $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$
- $\mathcal{R}$ — функція винагороди: $\mathcal{R}_s = \mathbb{E}[R_{t+1} | S_t = s]$
- γ — коефіцієнт зменшення (знецінювання), $\gamma \in [0, 1]$

Приклад: МПВ



—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Загальна винагорода

Загальна винагорода — сумарна винагорода отримана агентом з моменту часу t з урахування знецінювання:

$$\boxed{G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}}$$

- Коефіцієнт знецінювання $\gamma \in [0, 1]$ показує на цінність майбутніх винагород
- Значення винагороди R, отримане після $k + 1$ кроків: $\gamma^k R$
- Чим менший коефіцієнт знецінювання, тим менше агент замислюється над вигодою від майбутніх своїх дій.

Яка роль знецінювання?

- Дозволяє уникнути нескінченної загальної винагороди в циклічних марківських процесах
- Невизначеність щодо майбутнього може бути представлена не повністю
- Якщо винагорода є фінансовою, негайні винагороди можуть бути більш цікавими, ніж відстрочені винагороди
- Поведінка тварин/людини демонструє перевагу миттєвій винагороді
- Іноді можна використовувати марківський процес винагороди без знецінювання(тобто $\gamma = 1$), наприклад якщо всі послідовності закінчуються.

—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Функція цінності

Функція цінності $v(s)$ показує довгострокову цінність перебування агента у стані s

Функція цінності $v(s)$ марківського процесу винагороди — середнє значення загальної винагороди починаючи від стану s

$$\boxed{\begin{aligned} v(s) &= \mathbb{E}\left[G_t \mid S_t = s\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s\right] \end{aligned}}$$

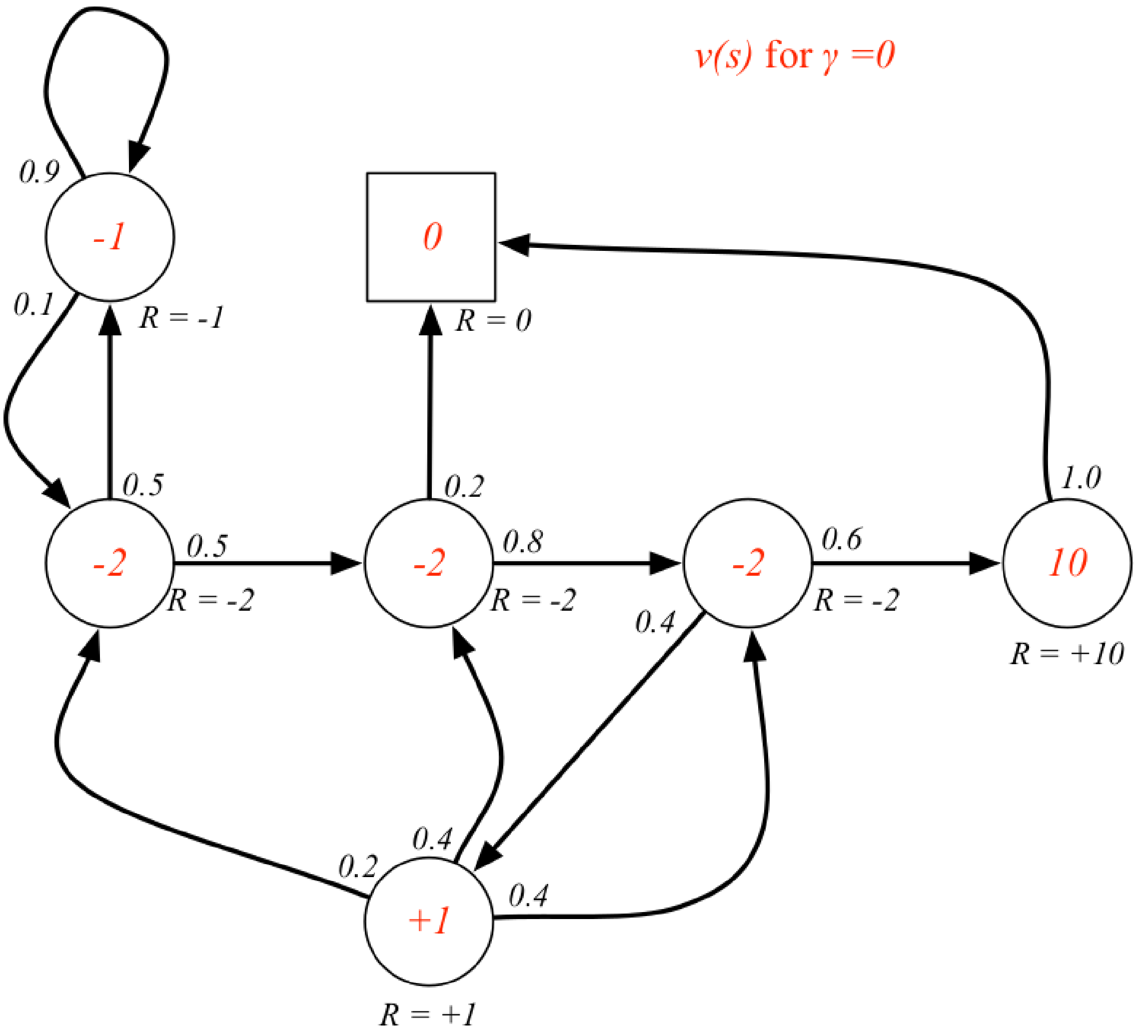
Приклад: МПВ загальна винагорода

Приклади загальної винагороди для раніше розглянутого прикладу.
Покачок з $S_1 = C_1$ з $\gamma = \frac{1}{2}$

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

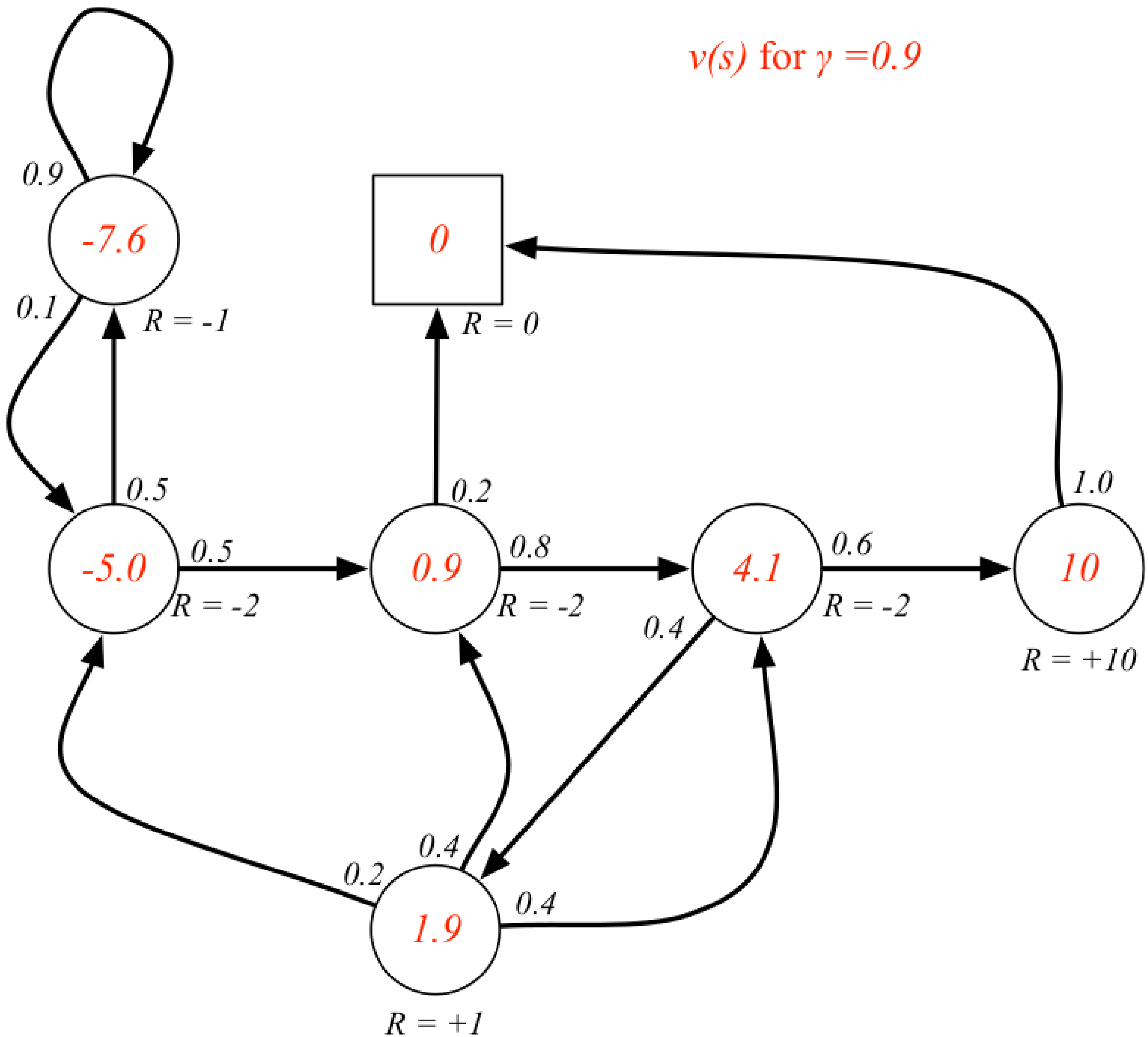
C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	$=$	-2.25
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	$=$	-3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8}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

Приклад: Функція цінності



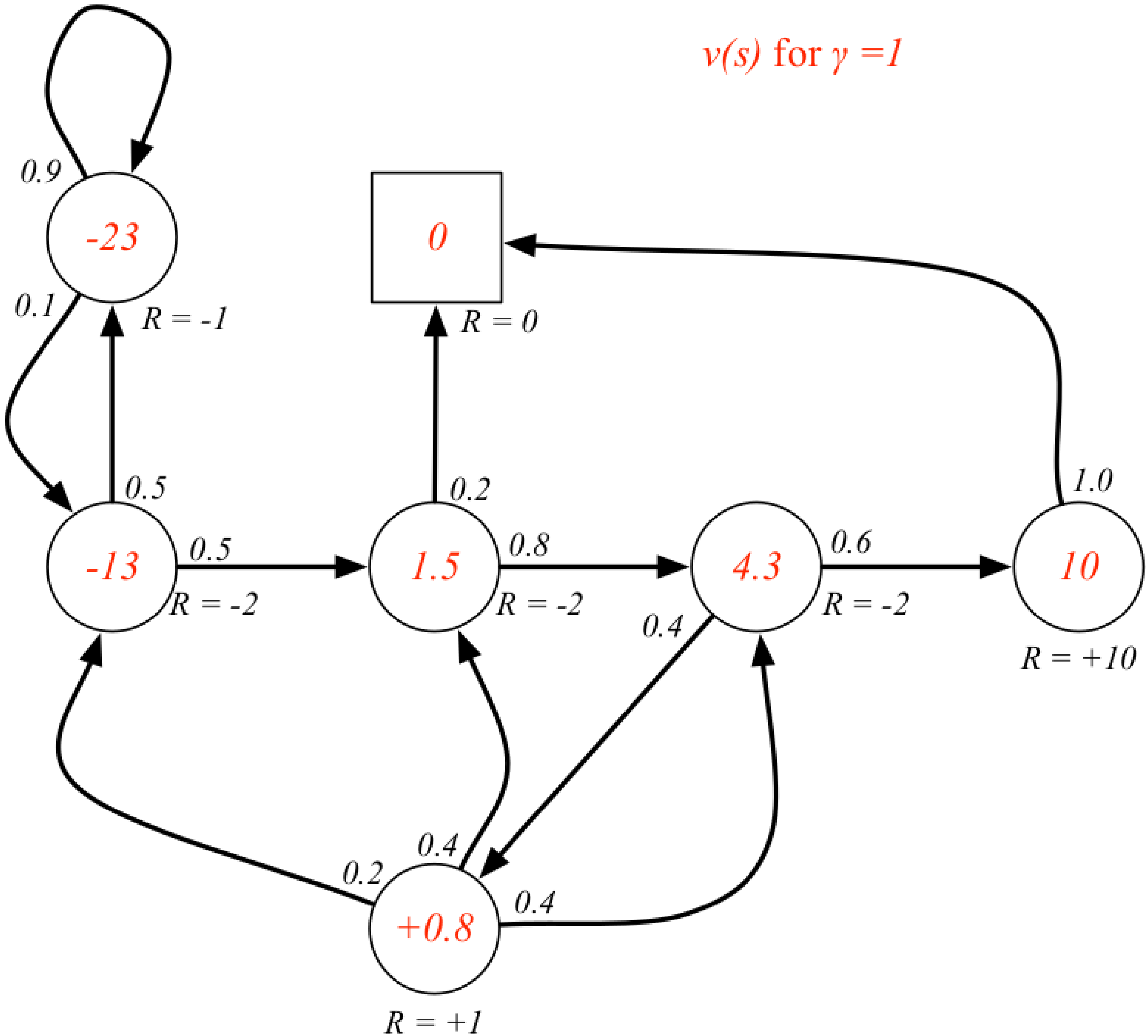
—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Приклад: Функція цінності



—

Приклад: Функція цінності



—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

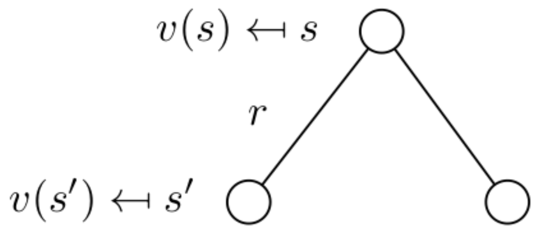
Рівняння Беллмана для МПВ

$$\boxed{\begin{aligned} v(s) &= \mathbb{E}\left[G_t \mid S_t = s\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \cdots) \mid S_t = s\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s\right] \end{aligned}}$$

—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Рівняння Беллмана: усереднення

$$v(s) = \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s\right]$$

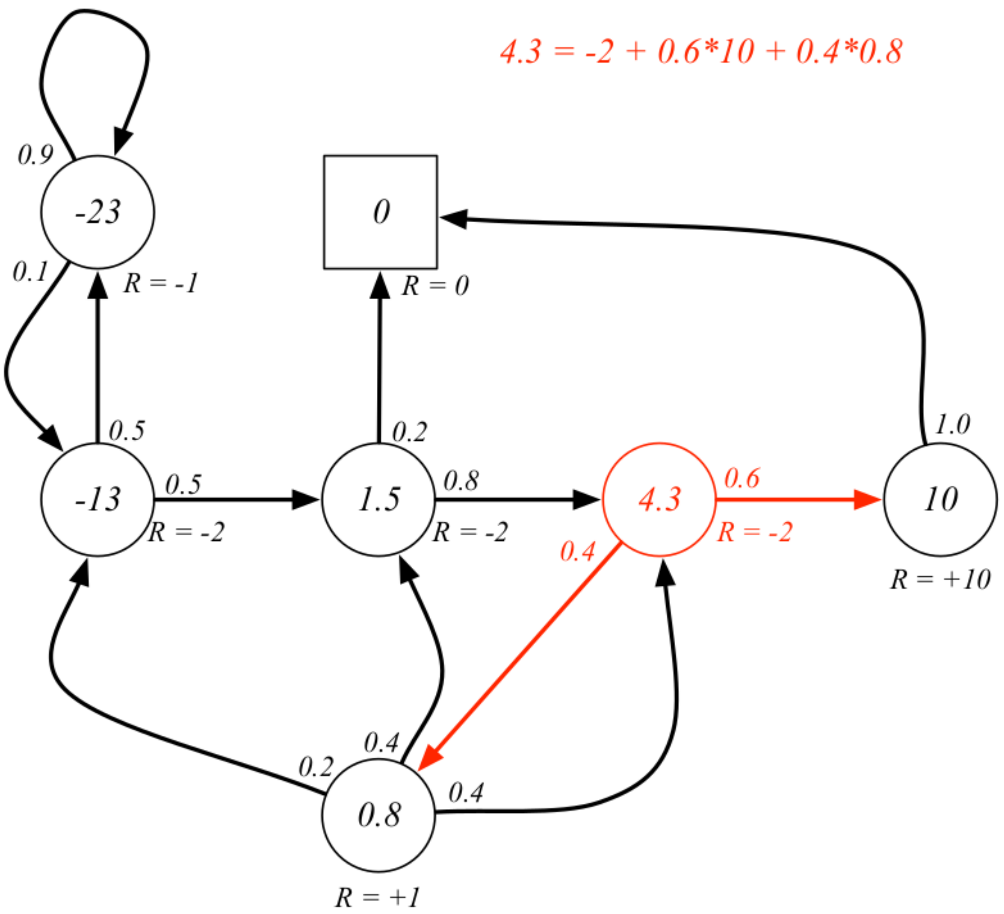


$$v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s')$$

—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Приклад усереднення рівняння Беллмана



Матрична форма рівняння Беллмана

Рівняння Беллмана можна виразити у матричній формі:

$$v = \mathcal{R} + \gamma \mathcal{P} v,$$

де v — вектор-стовпець з одним записом для кожного стану.

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

Розв'язок рівняння Беллмана

- Рівняння Беллмана є лінійним рівнянням
- Його можна розв'язати точних методів (алгебраїчним способом):

$$
\begin{aligned}
v &= \mathcal{R} + \gamma \mathcal{P} v \\
v(1 - \gamma \mathcal{P}) &= \mathcal{R} \\
v &= (1 - \gamma \mathcal{P})^{-1} \mathcal{R}
\end{aligned}
$$

- Обчислювальна складність становить $O(n^3)$ для n станів
- Алгебраїчний спосіб розв'язку можливий лише для малих МПВ ($n \sim 10^4$)
- Існує багато ітераційних методів для великих МПВ ($n \sim 10^7$)
    - Динамічне програмування
    - Оцінка Монте-Карло
    - Навчання часових різниць

Марківські процеси прийняття рішень (МППР)

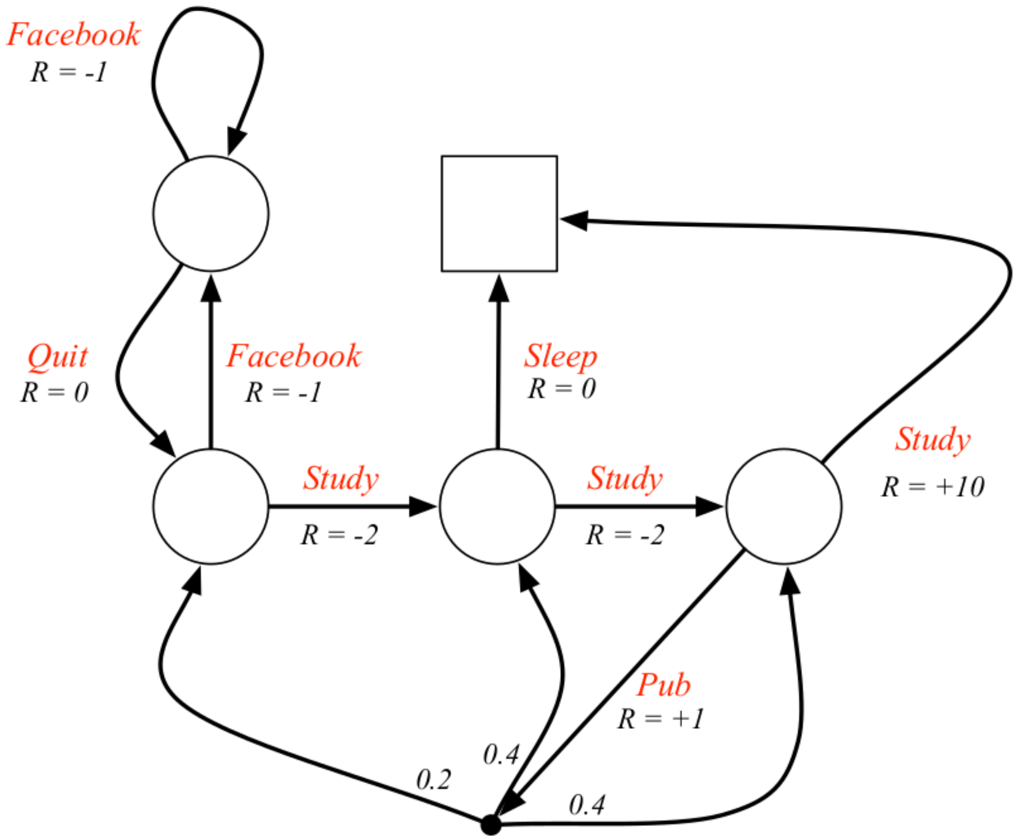
МППР

Марківський процес прийняття рішень (МППР) — марківський процес винагороди з рішеннями (прийнятими діями). Це середовище, у якому всі стани є марківськими.

МППР — це кортеж $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$:

- $\mathcal{S}$ — скінченна множина станів
- $\mathcal{A}$ — скінченна множина дій
- $\mathcal{P}$ — матриця зміни стану: $\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$
- $\mathcal{R}$ — функція винагороди: $\mathcal{R}_{s}^{a} = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- γ — коефіцієнт зменшення (знецінювання), $\gamma \in [0, 1]$

Приклад: МППР



Стратегія

Стратегія

$$\pi(a|s) = \mathbb{P}(A_t = a|S_t = s)$$

- Стратегія повністю визначає поведінку агента
- Стратегія у МППР залежить від поточного стану, а не від історії
- Тобто, стратегія є стаціонарною (не залежить від часу): $A_t \sim \pi(\cdot|S_t), \forall t > 0$

—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

- Для заданого МППР $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ та стратегії π
- Послідовність станів $S_1, S_2, \cdots$ &dmash; марківський процес $\langle \mathcal{S}, \mathcal{P}^\pi \rangle$
- Послідовність зі станів та винагород $S_1, R_2, S_2, \cdots$ &dmash; марківський процес винагород $\langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$

$$\mathcal{P}^\pi = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^a_{ss'}$$

$$\mathcal{R}^\pi = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^a_s$$

—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Функція цінності

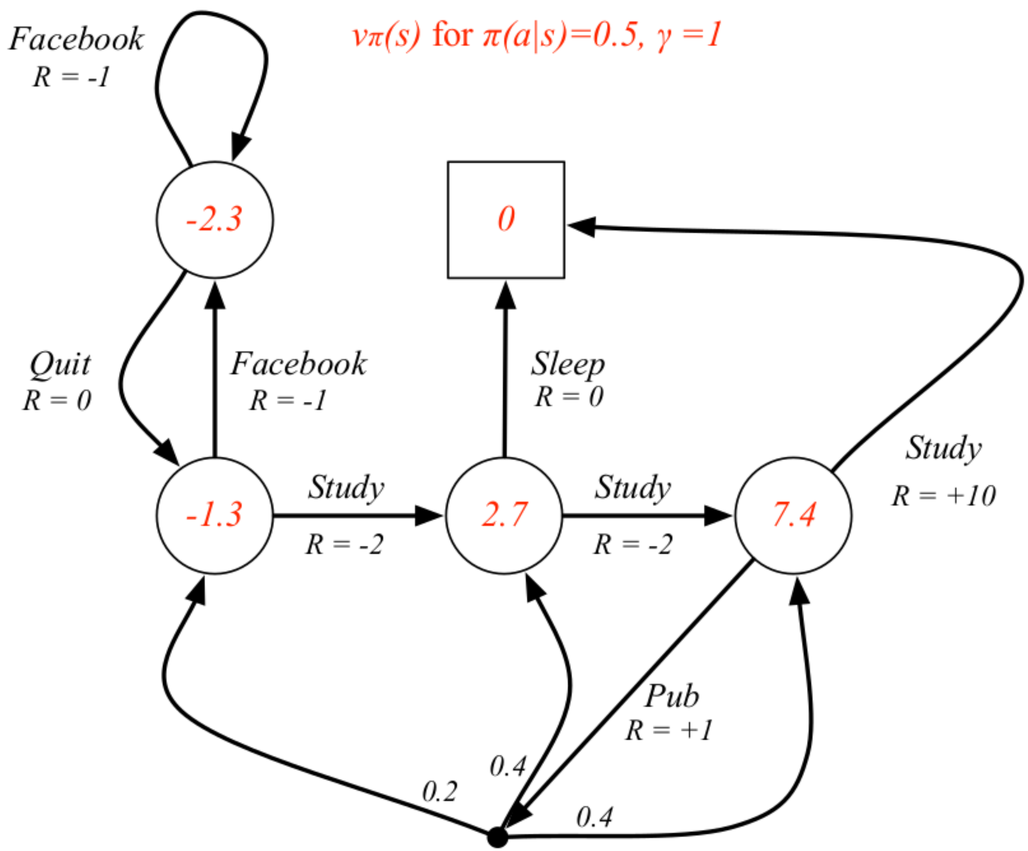
Функція цінності $v_\pi(s)$ МППР — середнє значення загальної винагороди починаючи від стану s при дотриманні заданої стратегії π

$$\boxed{\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t \mid S_t = s, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s, \pi\right] \end{aligned}}$$

Q-функція:

$$\boxed{\begin{aligned} q_\pi(s, a) &= \mathbb{E}\left[G_t \mid S_t = s, A_t = a, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s, A_t = a, \pi\right] \end{aligned}}$$

Приклад функції цінності



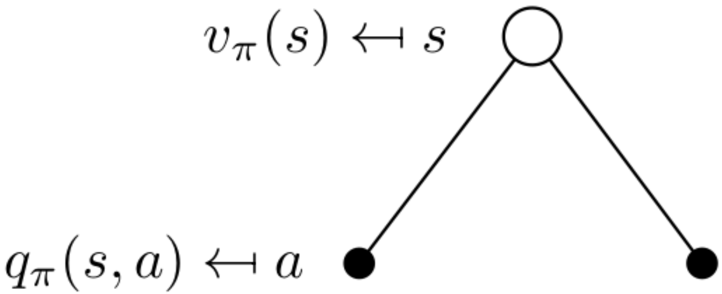
Рівняння Беллмана для МППР

$$\boxed{\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t \mid S_t = s, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) \mid S_t = s, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, \pi\right] \end{aligned}}$$

$$\boxed{\begin{aligned} q_\pi(s,a) &= \mathbb{E}\left[G_t \mid S_t = s, A_t = a, \pi\right] = \\ &= \mathbb{E}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a, \pi\right] \end{aligned}}$$

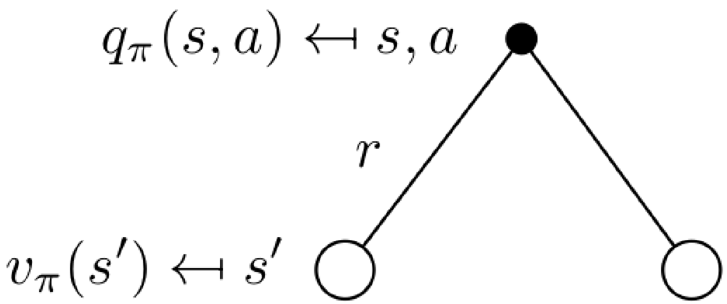
—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Рівняння Беллмана v_π



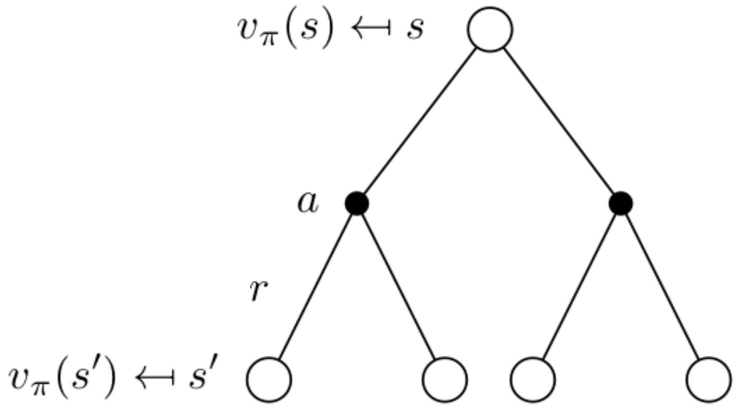
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$

Рівняння Беллмана q_π



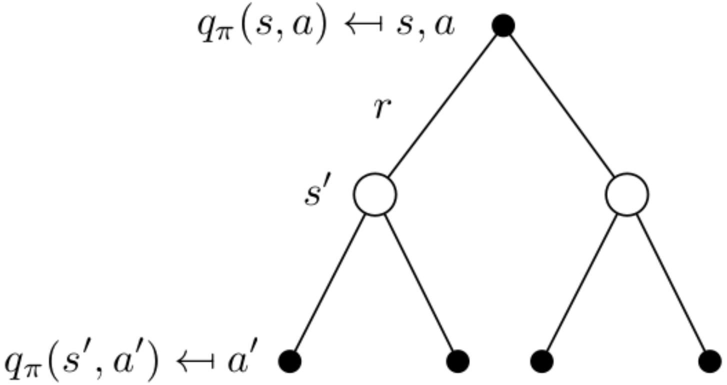
$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$

Рівняння Беллмана — **2** v_π



$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s') \right)$$

Рівняння Беллмана — 2 q_π

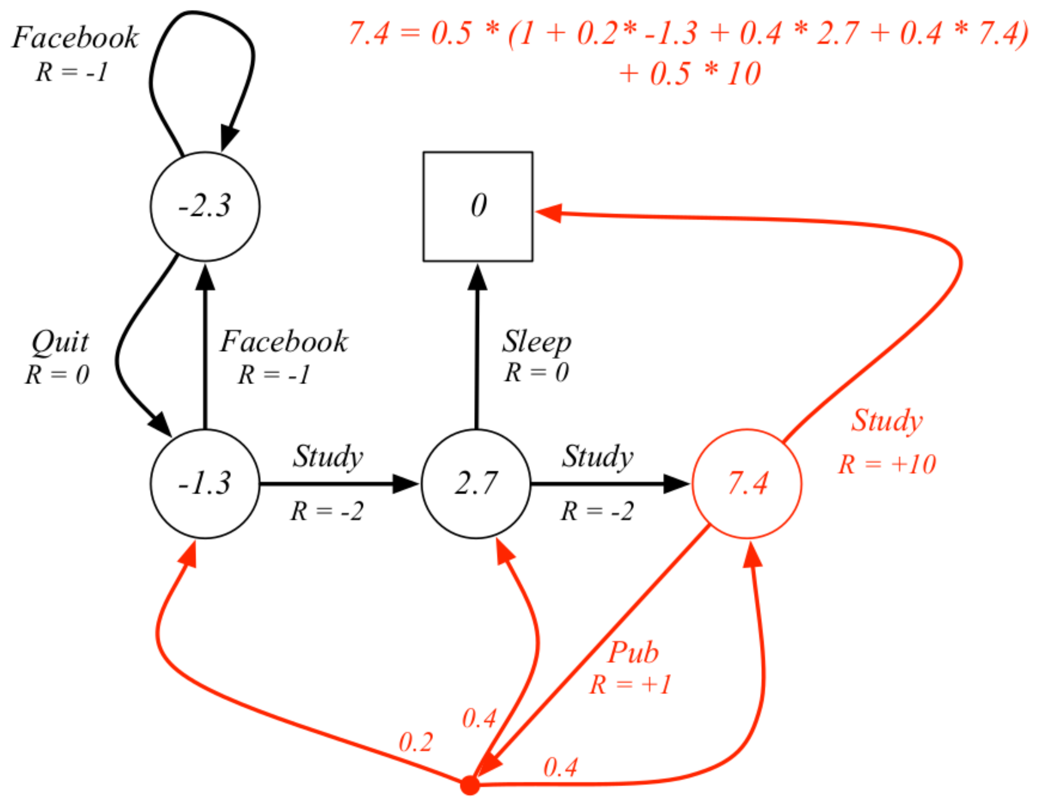


$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a')$$

—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Приклад рівняння Беллмана для МППР



—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Матрична форма рівняння Беллмана для МППР

Рівняння Беллмана можна виразити у матричній формі:

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi,$$

де v_π — вектор-стовпець з одним записом для кожного стану.

$$\begin{bmatrix} v_\pi(1) \\ \vdots \\ v_\pi(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}^\pi_1 \\ \vdots \\ \mathcal{R}^\pi_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}^\pi_{11} & \cdots & \mathcal{P}^\pi_{1n} \\ \vdots & & \\ \mathcal{P}^\pi_{n1} & \cdots & \mathcal{P}^\pi_{nn} \end{bmatrix} \begin{bmatrix} v_\pi(1) \\ \vdots \\ v_\pi(n) \end{bmatrix}$$

Точний розв'язок:

$$v_\pi = (1 - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$

—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Оптимальна функція цінності

Оптимальна функція цінності $v_*(s)$ — це максимальне значення функції серед усіх стратегій:

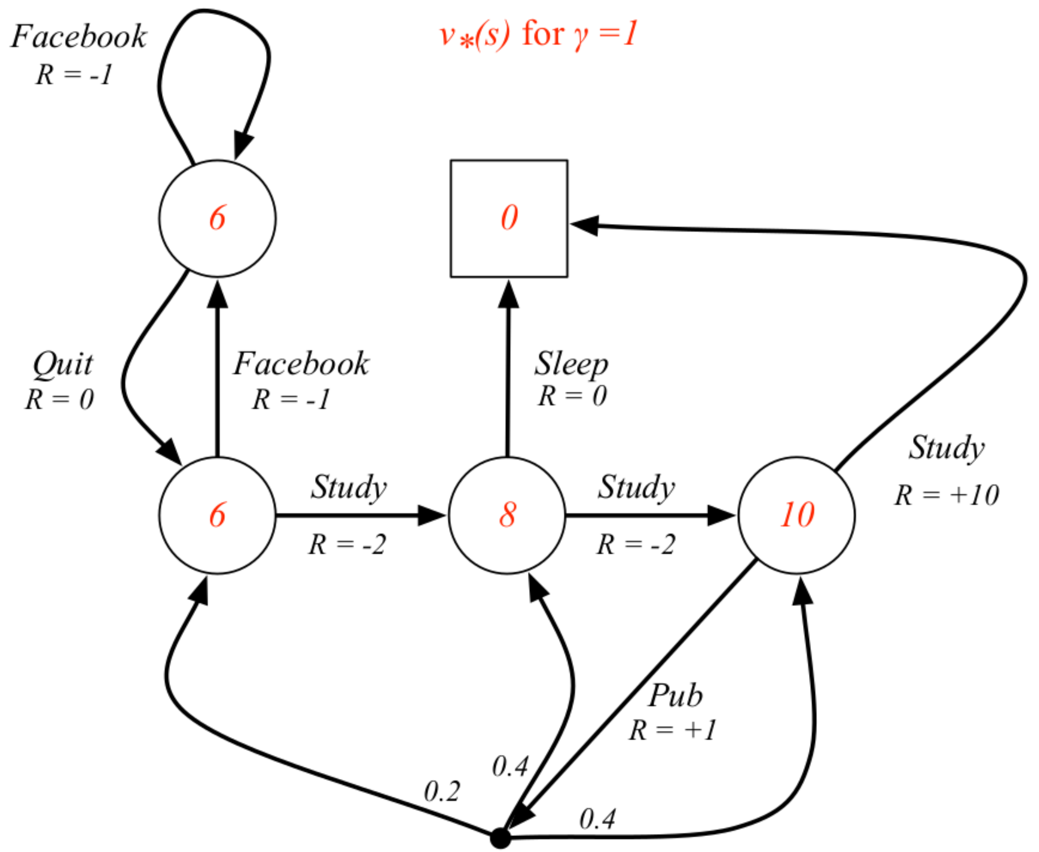
$$v_*(s) = \max_{\pi} v_\pi(s)$$

Оптимальна Q-функція $q_*(s,a)$ — це максимальне значення функції серед усіх стратегій:

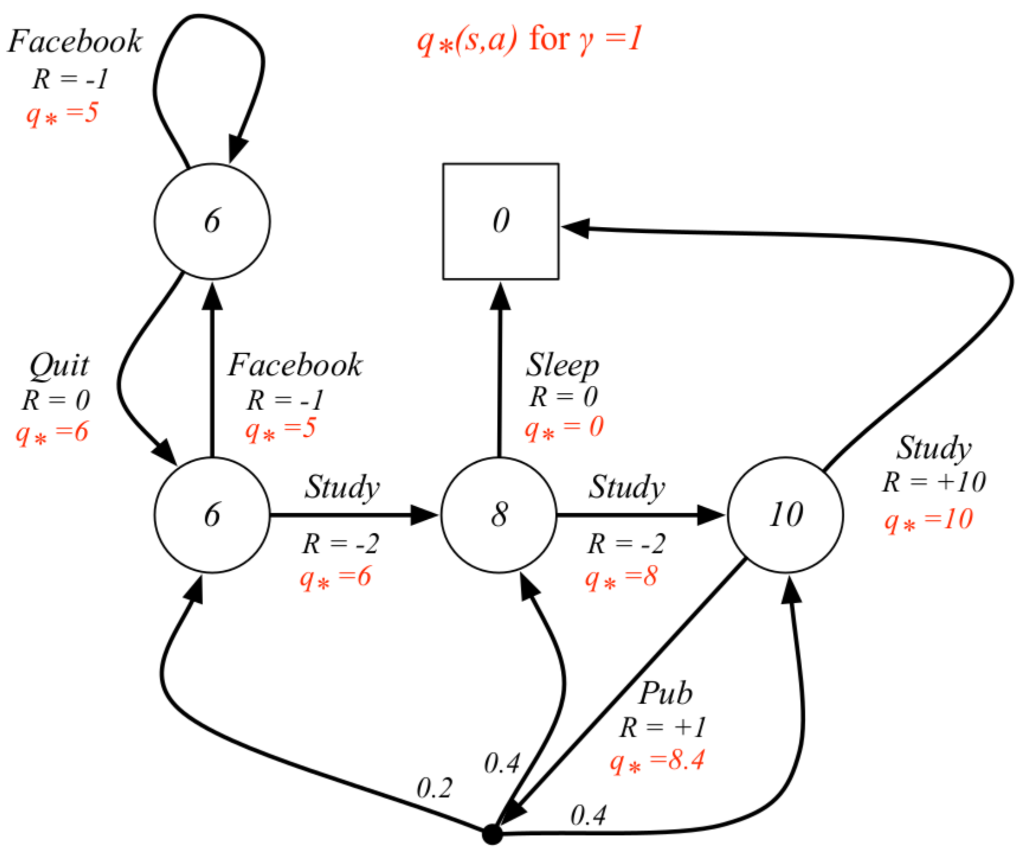
$$q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

- Оптимальна функція цінності вказує на найкращу з можливих продуктивностей у МППР.
- МППР є ''вирішиним'', коли ми знаємо оптимальне значення функції цінності.

Приклад: оптимум $v_*(s)$



Приклад: оптимум $q_*(s, a)$



Оптимальна стратегія

Упорядкування стратегій:

$\pi > \pi'$ якщо $v_\pi(s) > v'_\pi(s), \forall s$

Теорема. Для будь-якого МППР

- існує оптимальна стратегія π_*, яка краща або не гірша за інші стратегії: $\pi_* > \pi, \forall \pi$
- усі оптимальні стратегії досягають оптимальної функції цінності: $v_{\pi_*}(s) = v_*(s)$
- усі оптимальні стратегії досягають оптимального значення Q-функції: $q_{\pi_*}(s, a) = q_*(s, a)$

—
Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

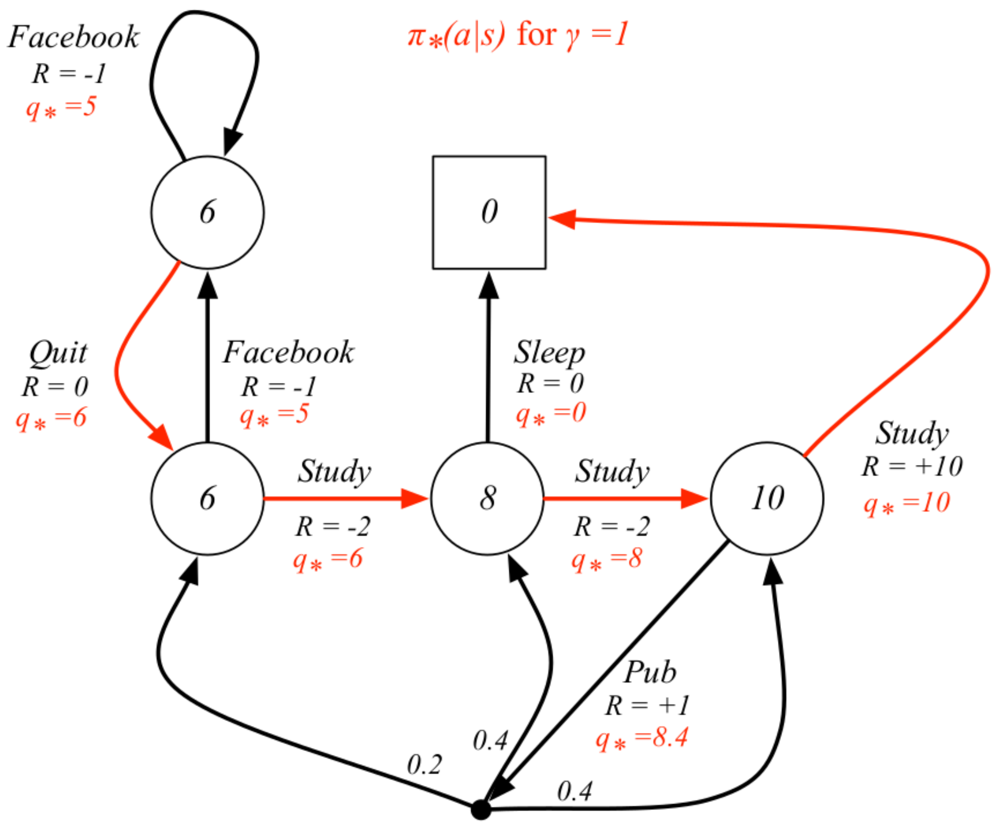
Пошук оптимальної стратегії

Оптимальна стратегія може бути знайдена, шляхом знаходження максимуму $q_*(s,a)$

$$\pi_*(a|s) = \begin{cases} 1, & \text{if}\, a = arg\max_{a\in\mathcal{A}} q_*(s,a) \\ 0, & \text{else} \end{cases}$$

- Для будь-якого МППР завжди існує детермінована оптимальна стратегія
- Якщо відомо $q_*(s,a)$, ми одразу маємо оптимальну стратегію

Приклад: оптимальна стратегія для МППР



—

Джерело слайду: Lecture 2: Markov Decision Processes [video], David Silver.

Кінець

Література

- David Silver, Lecture 2: Markov Decision Processes