



Навчання з підкріпленням

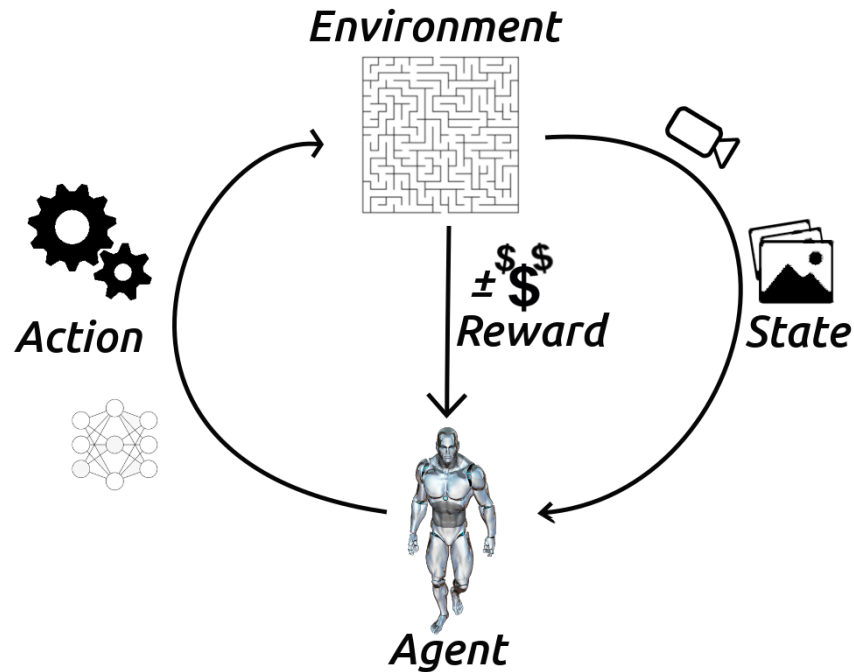
Лекція 2: Марковські процеси прийняття рішень

Кочура Юрій Петрович
iuriy.kochura@gmail.com
[@y_kochura](#)

Сьогодні

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

Цикл взаємодії



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

Вступ до МППР

- Марковські 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 Марковським тоді і тільки тоді

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

- 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 володіють властивістю 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 Марковськими станами $s \rightarrow s'$, визначається так:

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

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.

Марковський процес

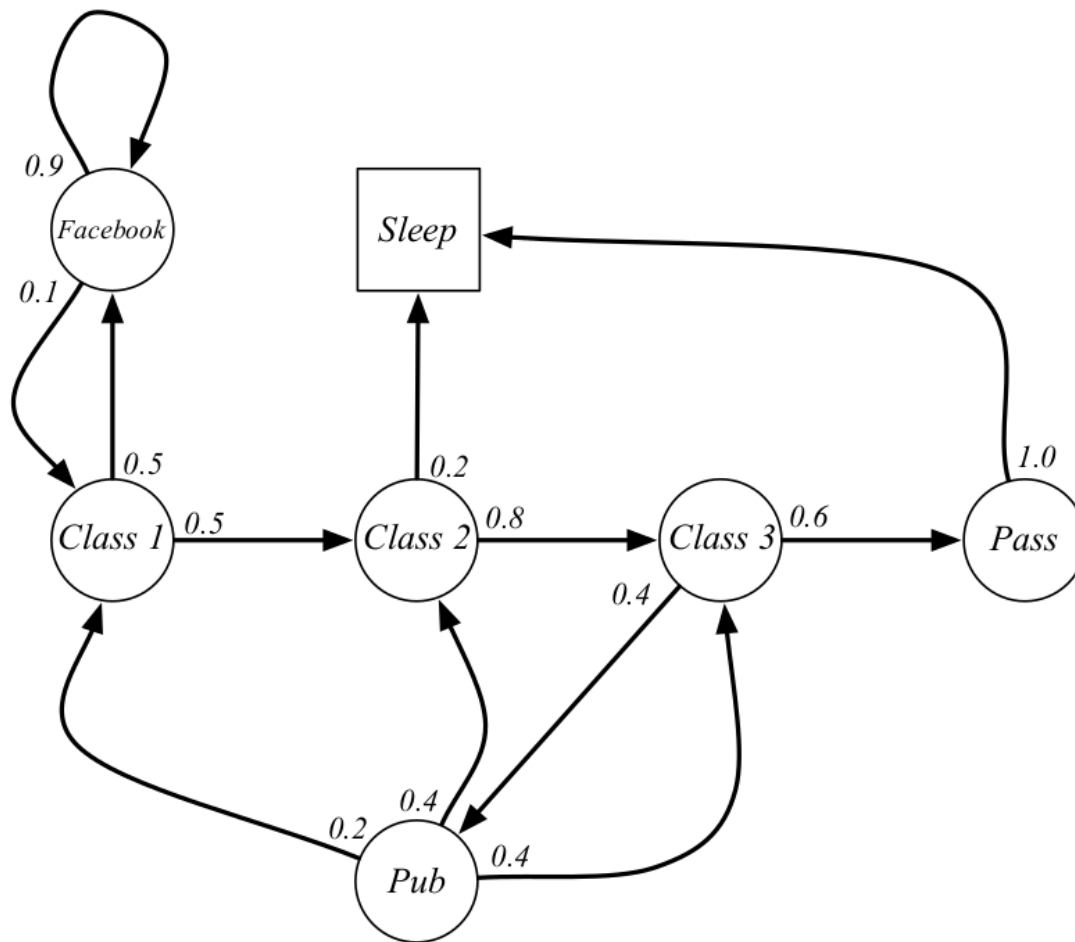
Марковський процес — це випадковий процес у якого відсутня пам'ять, тобто послідовність випадкових станів S_1, S_2, \dots , які володіють властивістю Маркова.

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

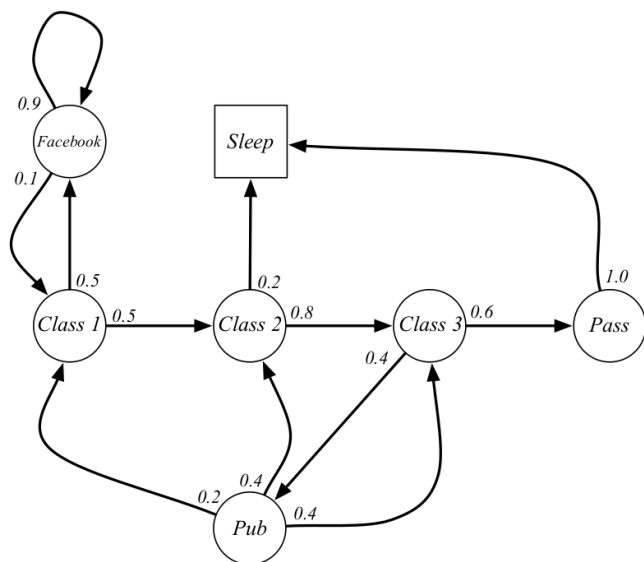
- \mathcal{S} — скінченна множина станів
- \mathcal{P} — матриця зміни стану: $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$

Приклад

Студентський ланцюг Маркова



Студентський ланцюг Маркова

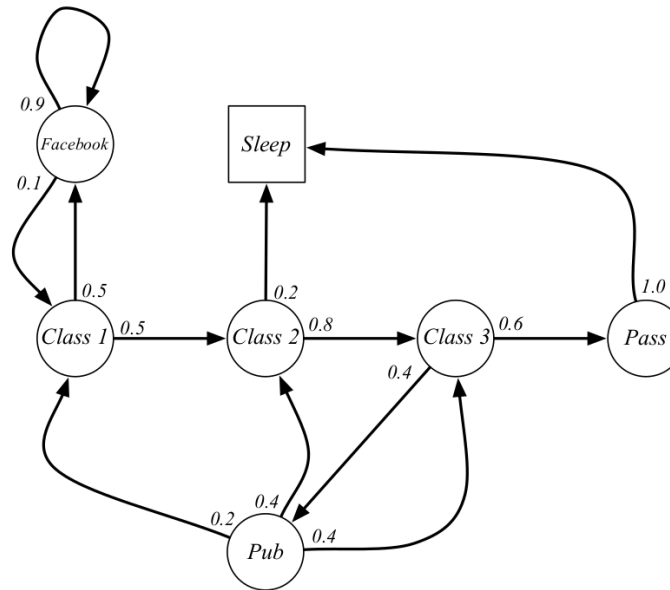


Початковий епізод починається з $S_1 = C_1$

$$S_1, S_2, \dots, S_T$$

- 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



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

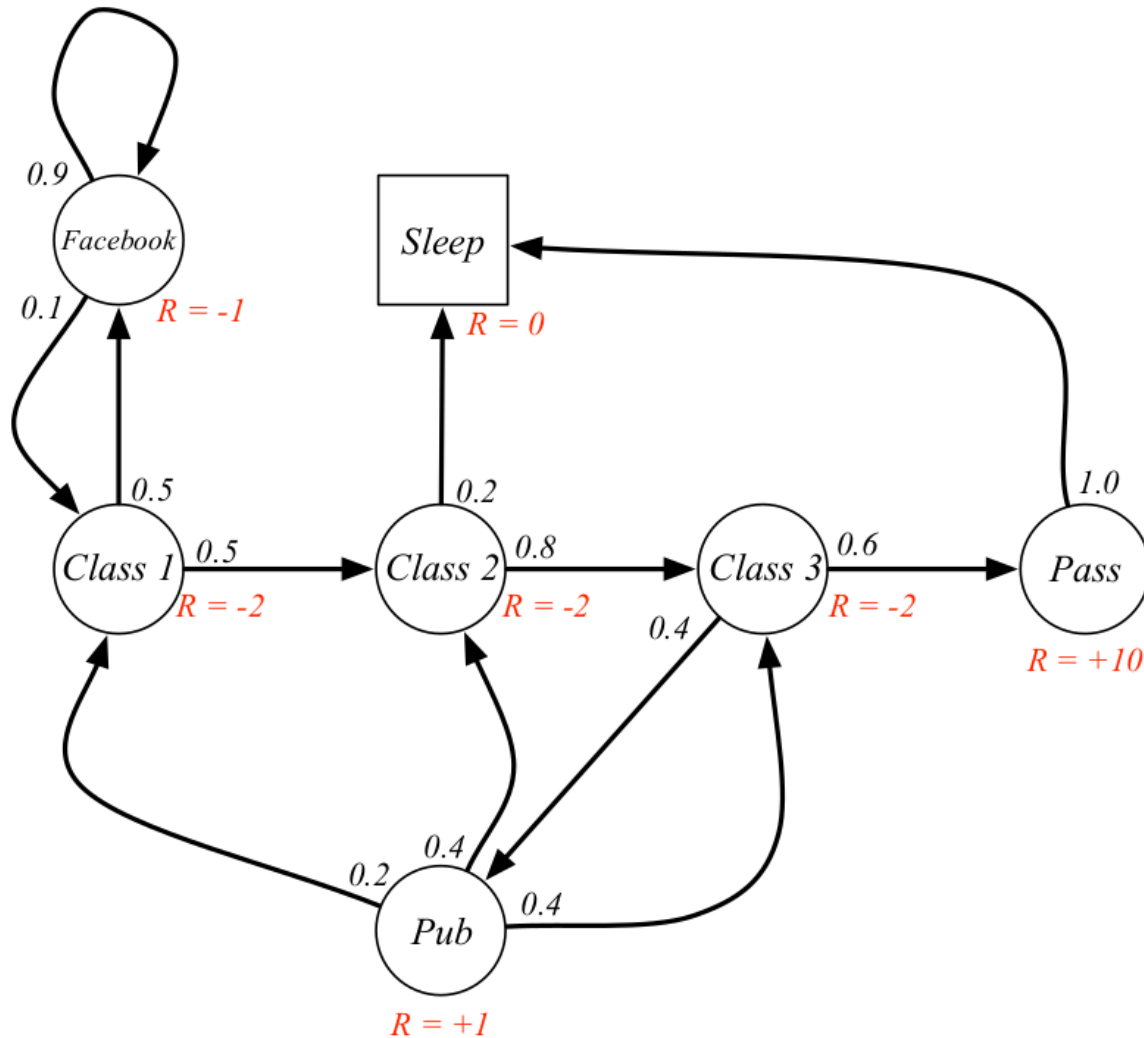
Марковські процеси винагороди

Марковський процес винагороди — ланцюг Маркова з винагородою.

Марковський 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 знецінювання, $\gamma \in [0, 1]$

Приклад: МПВ



Загальна винагорода

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

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

- Коефіцієнт знецінювання $\gamma \in [0, 1]$ визначає 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, миттєва винагорода може бути більш цікавою, ніж відтермінована
- Поведінка тварин/людини демонструє перевагу миттєвій винагороді
- Іноді можна використовувати марковський процес винагороди без знецінювання(тобто $\gamma = 1$), наприклад якщо всі послідовності закінчуються.

Функція цінності

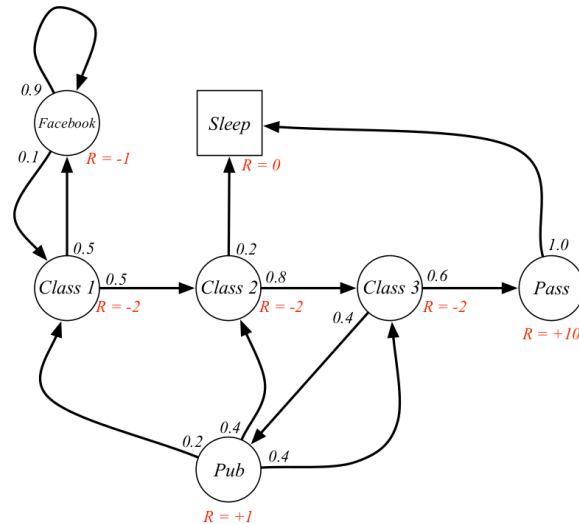
Функція цінності $v(s)$ показує довгострокову цінність перебування агента у стані s

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

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

Приклад: МПВ загальна винагорода

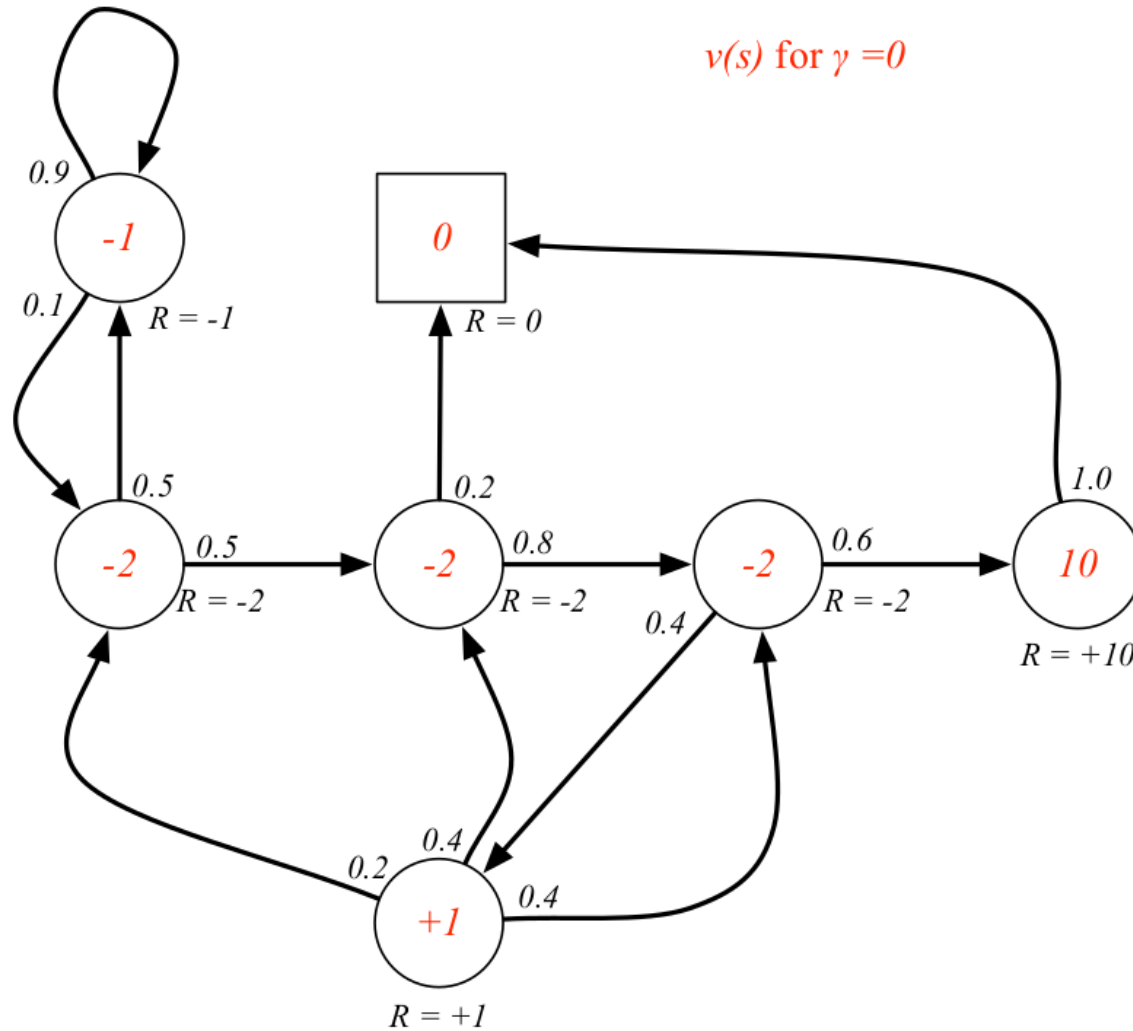
Покачок з $S_1 = C_1$ з $\gamma = \frac{1}{2}$



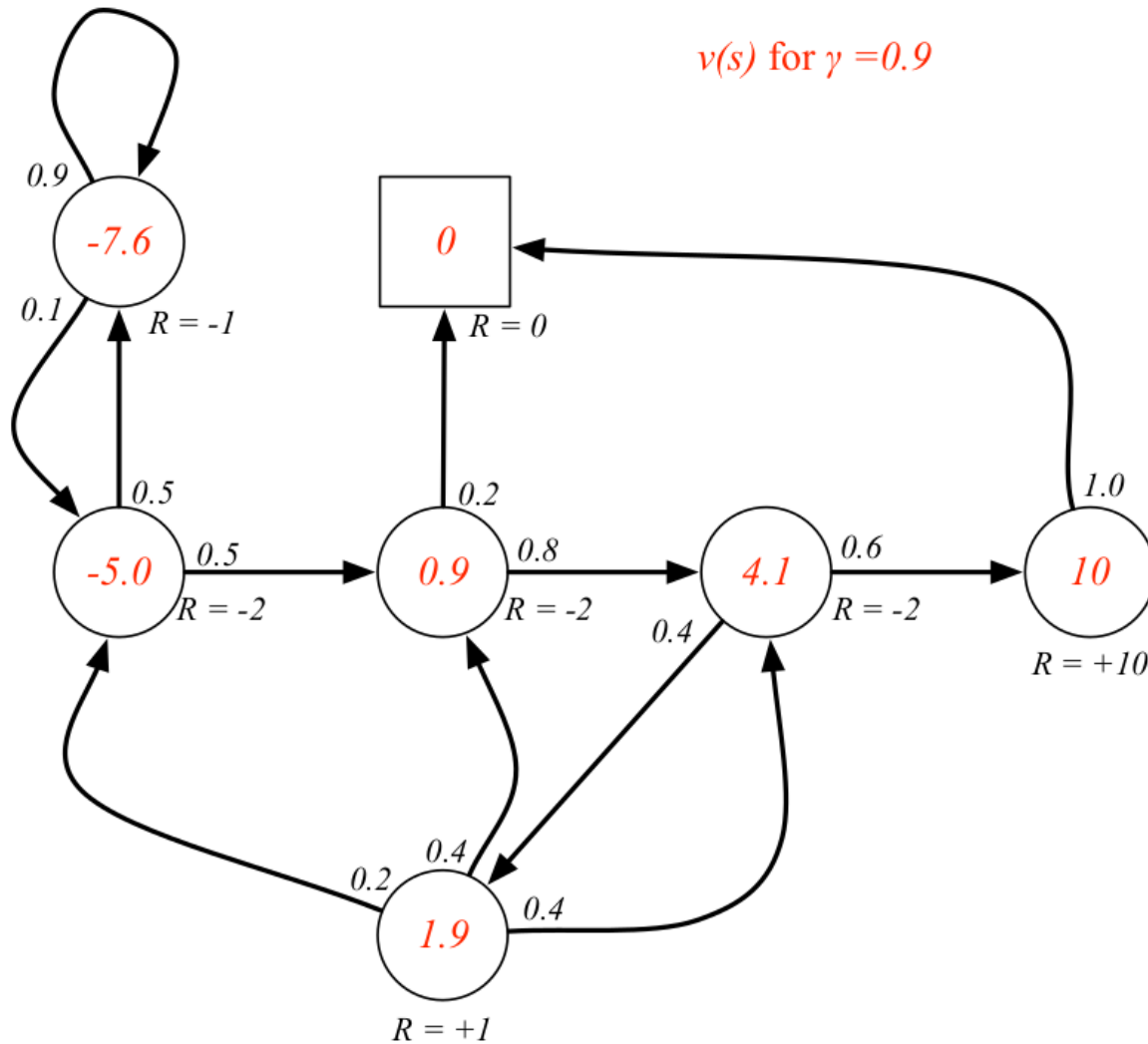
$$G_1 = R_2 + \gamma R_3 + \dots + \gamma^{T-2} R_T$$

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			

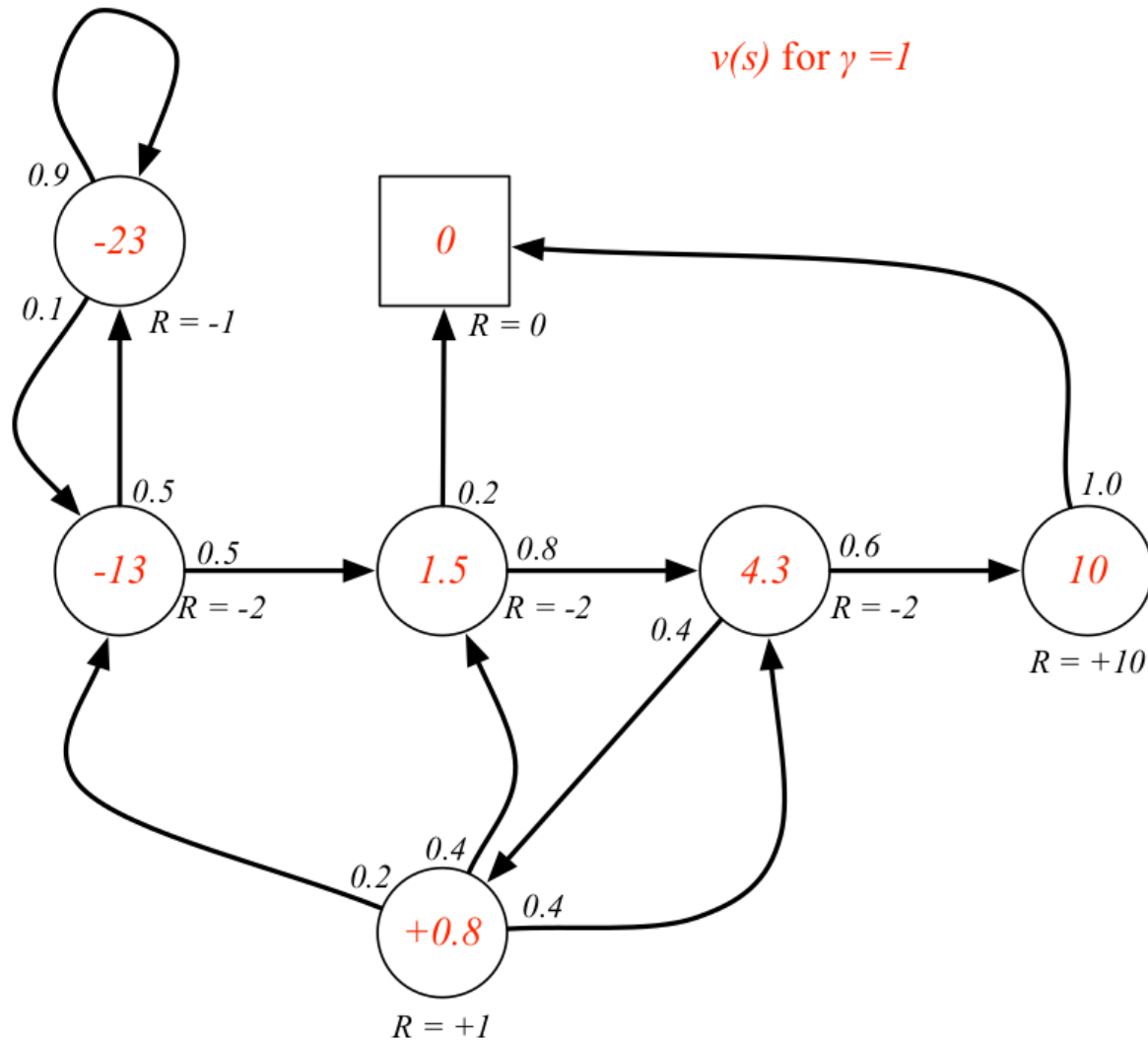
Приклад: Функція цінності



Приклад: Функція цінності



Приклад: Функція цінності

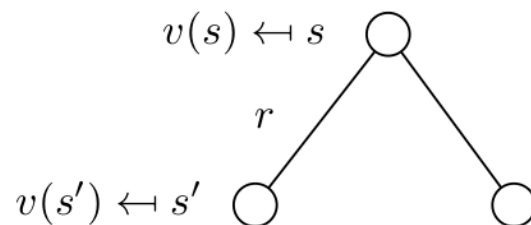


Рівняння Беллмана для МПВ

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

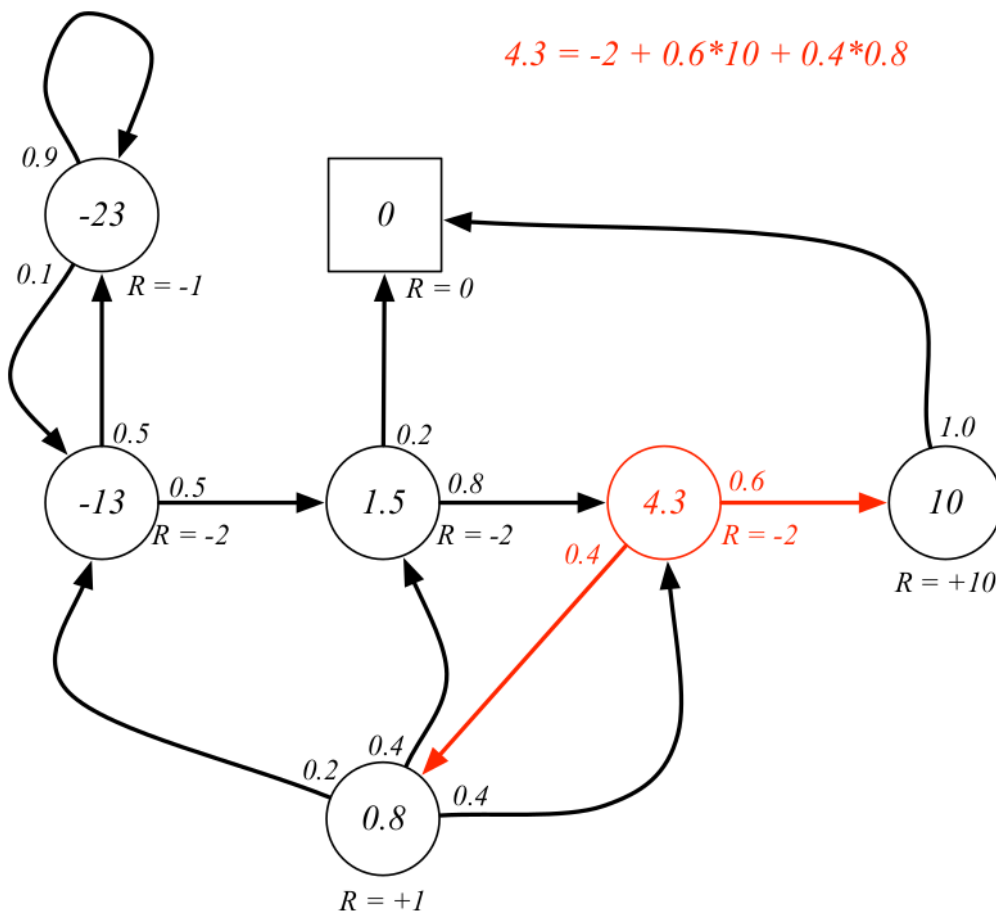
Рівняння Беллмана: усереднення

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



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

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.

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

Розв'язок рівняння Беллмана

- Рівняння Беллмана є лінійним рівнянням
- Його можна розв'язати точними методами (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

Марковські процеси прийняття рішень (МППР)

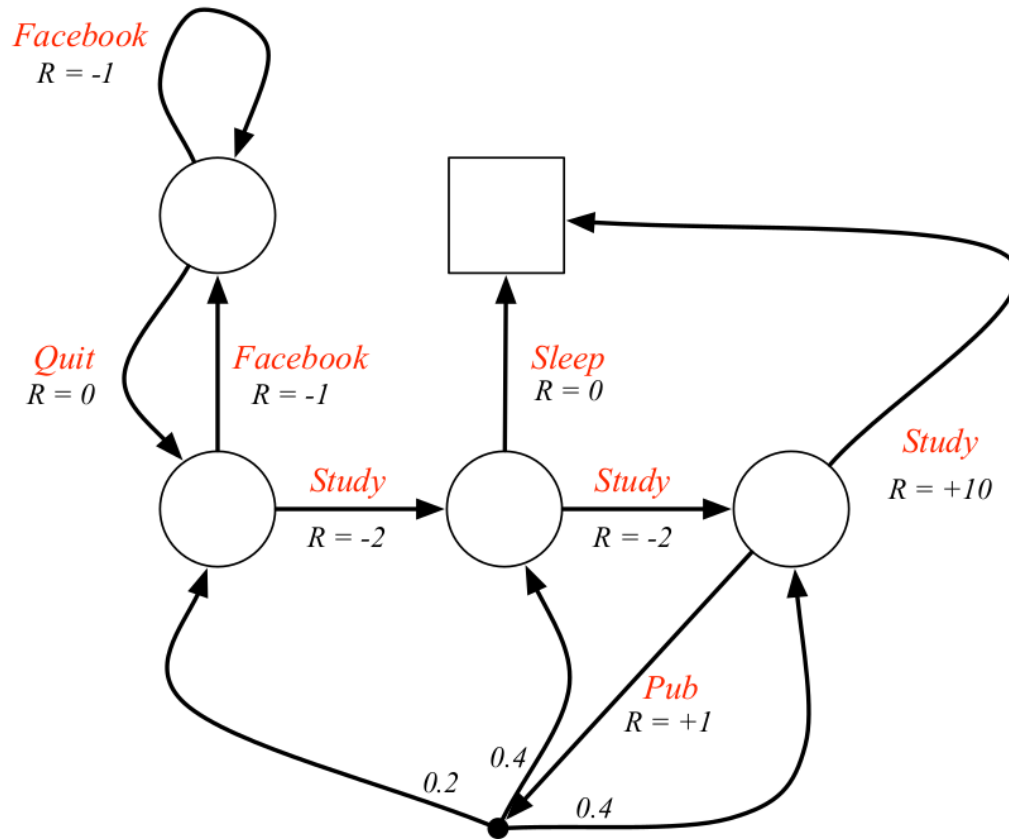
МППР

Марковський процес прийняття рішень (МППР) — марковський процес винагороди з рішеннями (прийнятими діями). Це середовище, у якому всі стани є марковськими.

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: $\mathcal{P}_{ss'}^a = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
- \mathcal{R} — функція винагороди: $\mathcal{R}_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- γ — коефіцієнт 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):

$$A_t \sim \pi(\cdot | S_t), \forall t > 0$$

- 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, \dots — марковський процес $\langle \mathcal{S}, \mathcal{P}^\pi \rangle$
- Послідовність зі станів та винагород S_1, R_1, S_2, \dots — марківський процес винагород $\langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$

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

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

Функція цінності

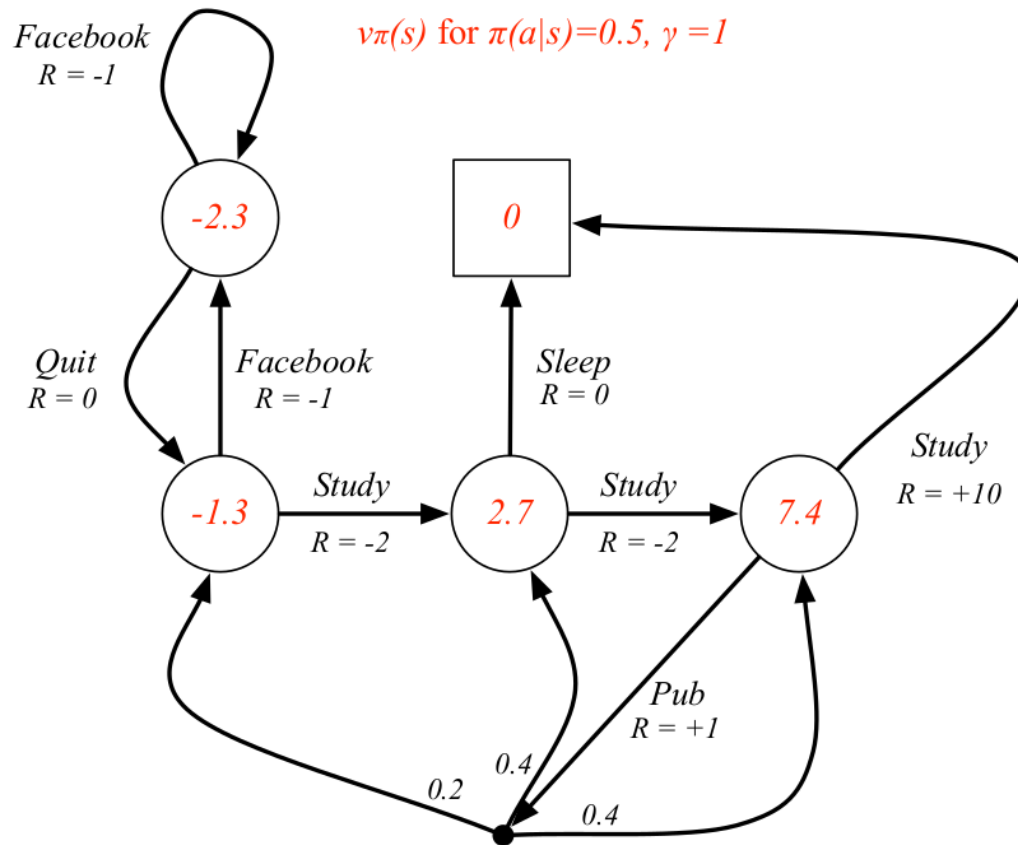
Функція цінності $v_\pi(s)$ МППР — середнє значення загальної винагороди починаючи від стану s при дотриманні заданої стратегії π

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

Q-функція:

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

Приклад функції цінності

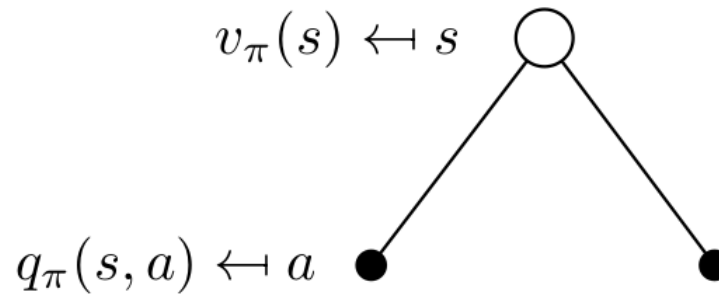


Рівняння Беллмана для МППР

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

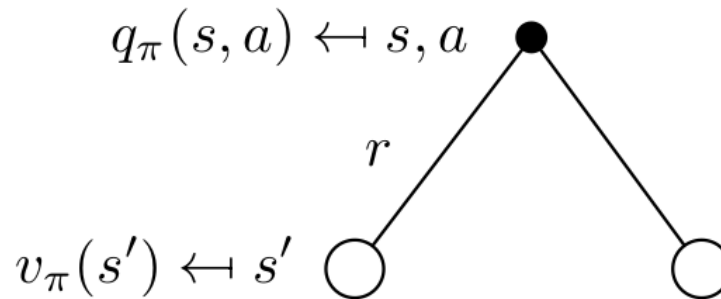
$$\begin{aligned} q_{\pi}(s, a) &= \mathbb{E} [G_t \mid S_t = s, A_t = a, \pi] = \\ &= \mathbb{E} [R_{t+1} + \gamma q_{\pi}(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a, \pi] \end{aligned}$$

Рівняння Беллмана v_π



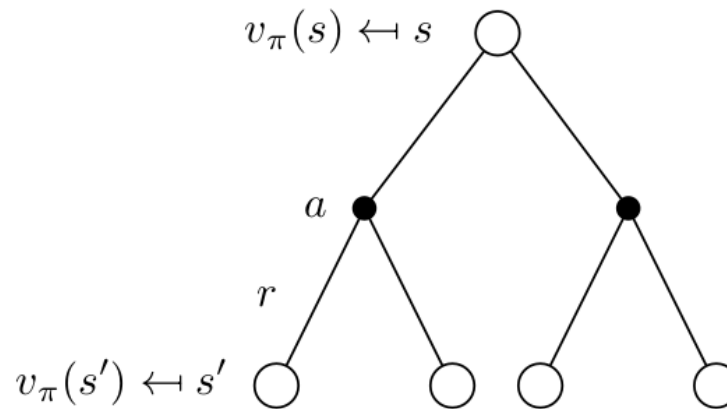
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$$

Рівняння Беллмана q_π



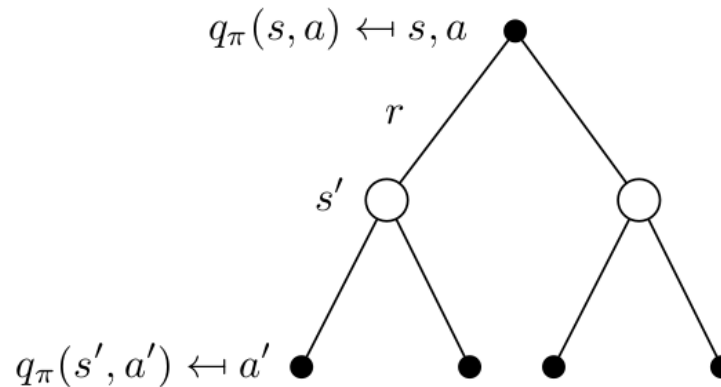
$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$

Рівняння Беллмана – $2 v_\pi$



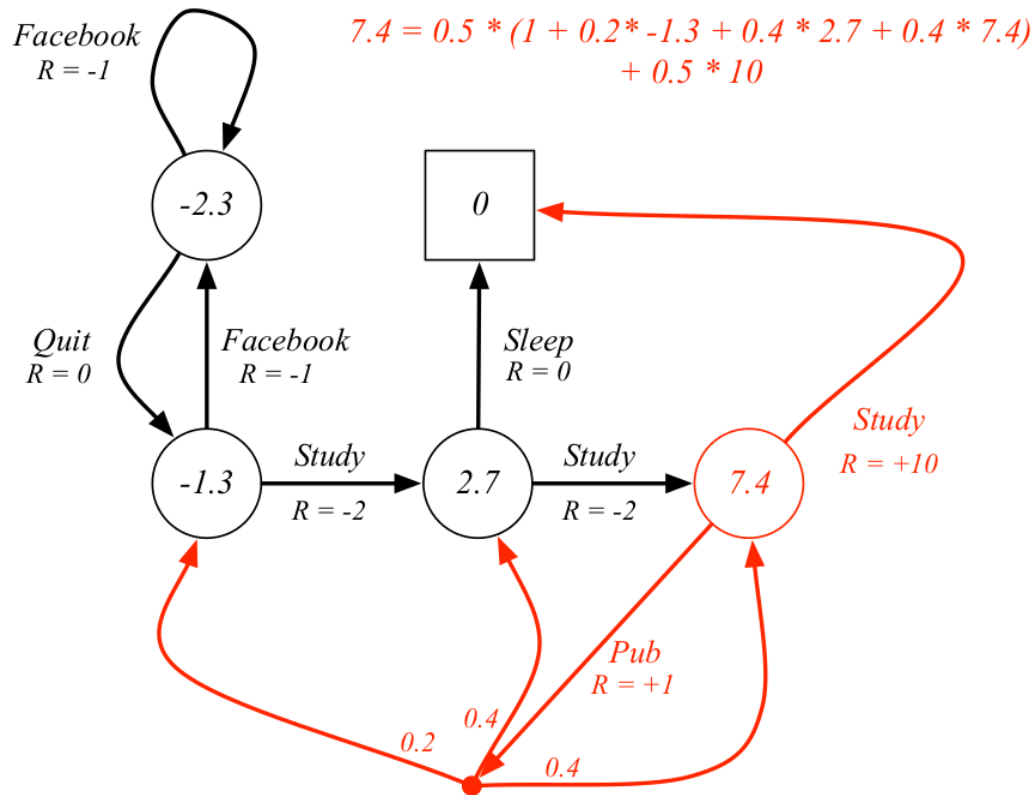
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s') \right)$$

Рівняння Беллмана – 2 q_π



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

Приклад рівняння Беллмана для МППР



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.

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

Точний розв'язок:

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

Оптимальна функція цінності

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

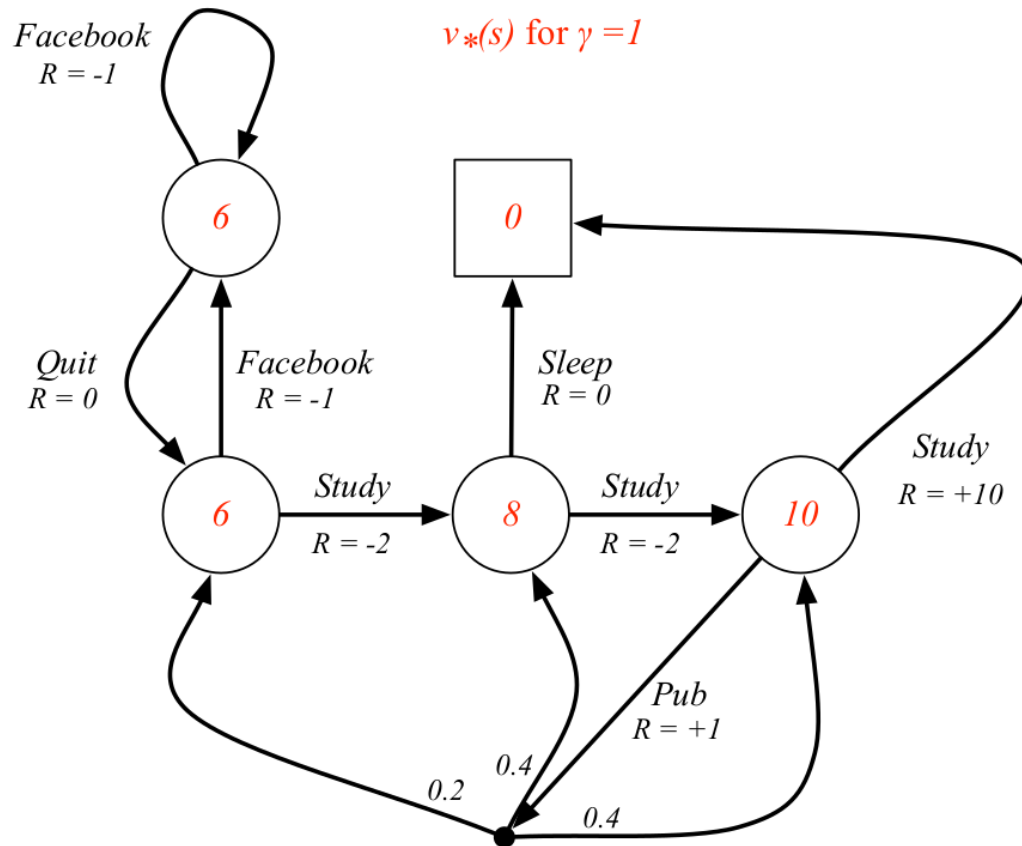
$$v_*(s) = \max_{\pi} v_{\pi}(s)$$

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

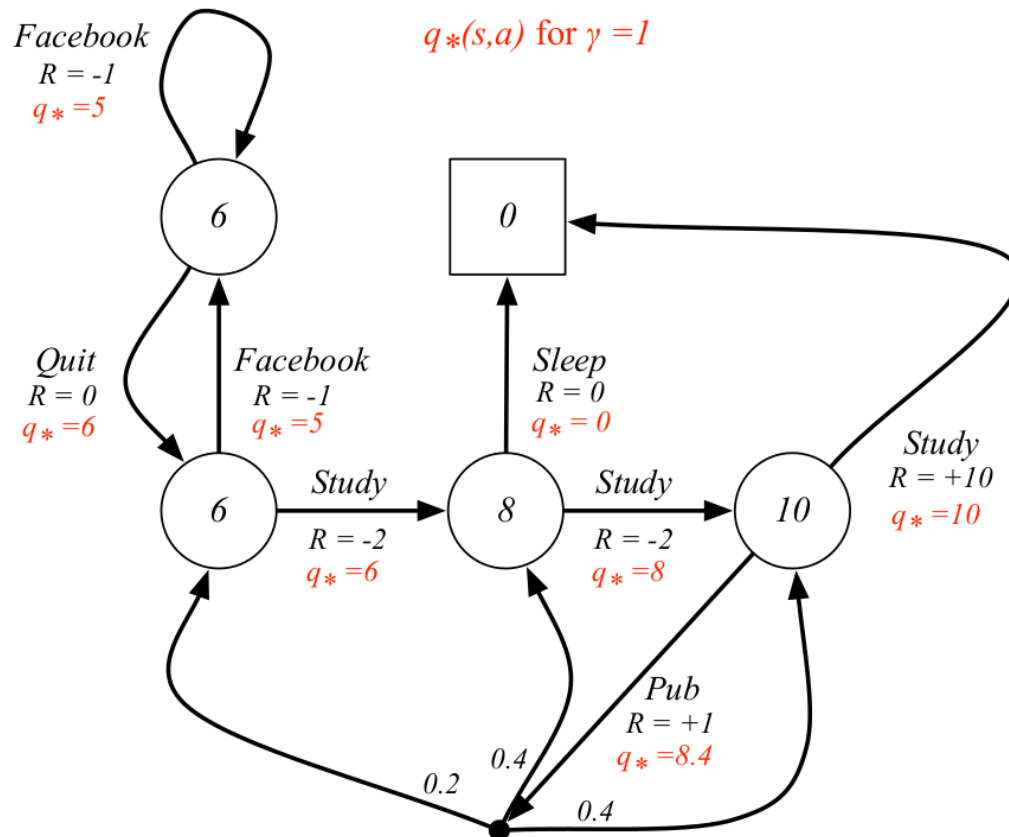
$$q_*(s, a) = \max_{\pi} q_{\pi}(s, a)$$

- Оптимальна функція цінності вказує на найкращу з можливих продуктивностей у МППР.
- МППР є "вирішеним", 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:
 $\pi_* \geq \pi, \forall \pi$
- усі оптимальні стратегії досягають оптимальної функції цінності:
 $v_{\pi_*}(s) = v_*(s)$
- усі оптимальні стратегії досягають оптимального значення Q-функції:
 $q_{\pi_*}(s, a) = q_*(s, a)$

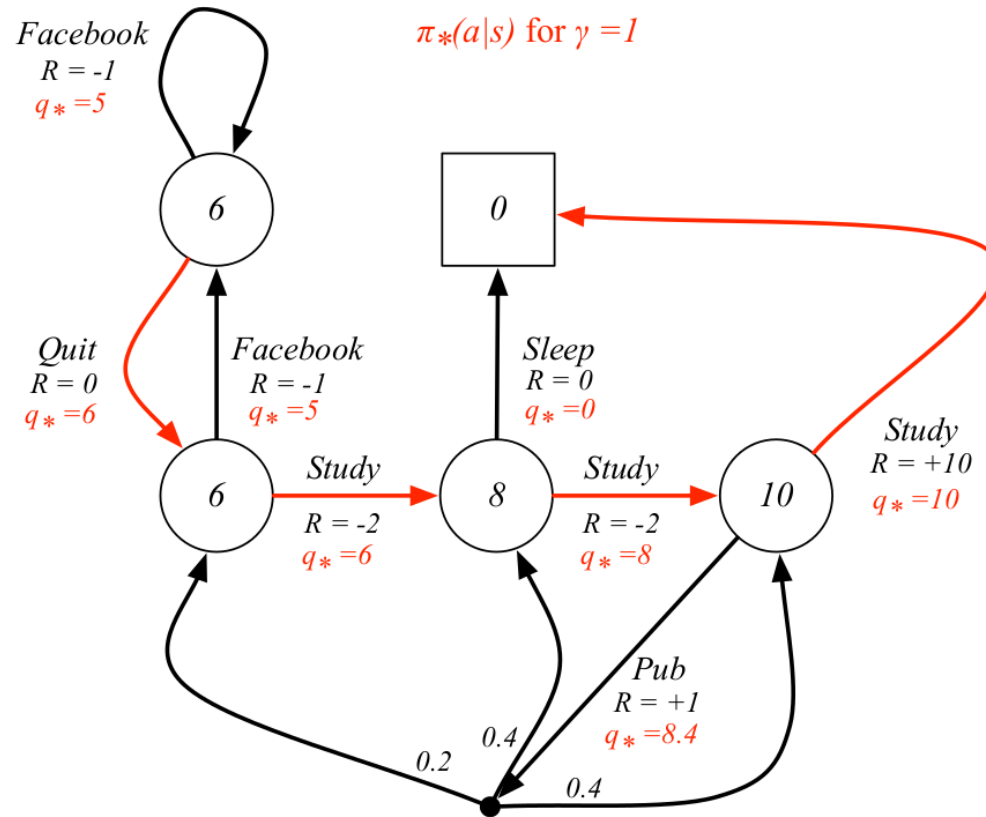
Пошук оптимальної стратегії

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



Кінець

Література

- David Silver, [Lecture 2: Markov Decision Processes](#)