Chapitre 4 : Processus de Décision de Markov (MDP)

Mohamed Anis BEN LASMAR

A.U.2024-2025

Introduction aux MDPs

Définition : Un MDP formalise un environnement pour l'apprentissage par renforcement où l'environnement est totalement observable.

- L'état actuel contient toute l'information nécessaire.
- Exemple: Tous les problèmes de RL peuvent être modélisés par le MDP.

Propriété de Markov

Définition : Un état S_t est Markov si et seulement si :

$$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, ..., S_t)$$

Conséquences :

- L'historique peut être oublié après la connaissance de l'état actuel.
- Les transitions sont définies par une matrice de probabilités.

Exemple : Student Markov Chain

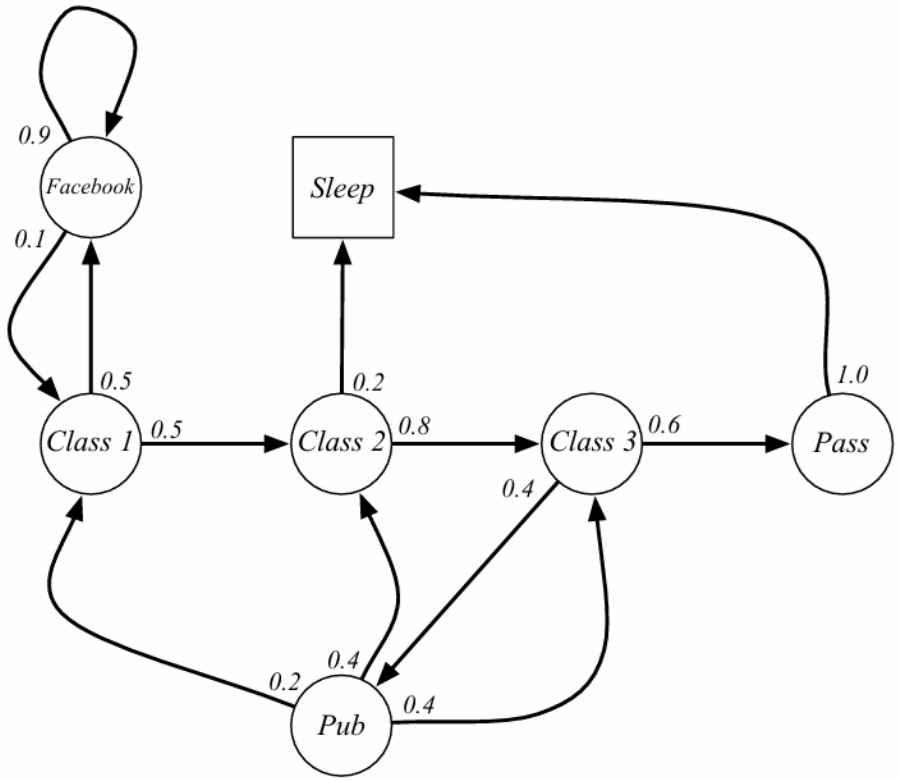
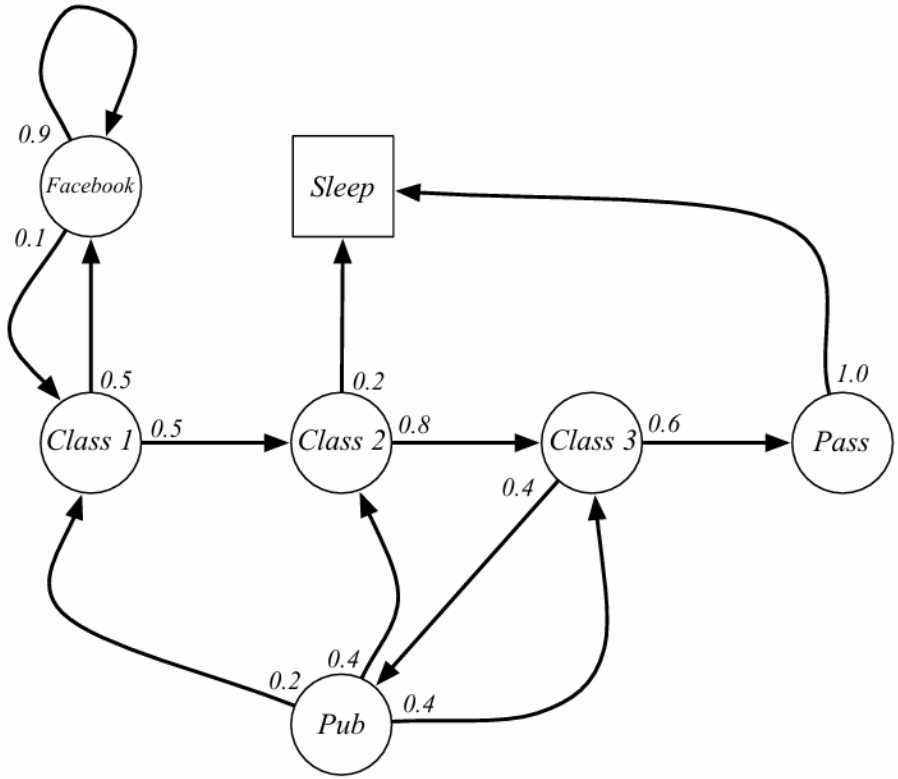


Figure: Student Markov Chain

Exemple : Student Markov Chain



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

Processus de Récompense de Markov

Definitin

Un Processus de Recompense de Markov est un processus de Markov avec des récompenses. Il est défini par l'uplet $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$.

- $\mathcal{S}$ est l'espace d'état.
- $\mathcal{P}$ est la matrice de transition.
- $\mathcal{R}$ est la fonction de recompense $\mathcal{R}_s = E[\mathcal{R}_{t+1}|\mathcal{S}_t = s]$.
- γ est un facteur d'actualisation. $\gamma \in [0, 1]$

Processus de Récompense de Markov - Exem

Definitin

Un Processus de Recompense de Markov est un processus de Markov avec des récompenses. Il est défini par l'uplet $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$.

- $\mathcal{S}$ est l'espace d'état.
- $\mathcal{P}$ est la matrice de transition.
- $\mathcal{R}$ est la fonction de recompense $\mathcal{R}_s = E[\mathcal{R}_{t+1}|\mathcal{S}_t = s]$.
- γ est un facteur d'actualisation. $\gamma \in [0, 1]$

Le rendement

Definitin

Le rendement G_t est la recompense totale actualisée à partir du pas de temps t.

$$G_t = \mathcal{R}_{t+1} + \gamma \mathcal{R} t + 2 + \ldots = \sum_{k=0}^{\infty} \gamma^k \mathcal{R}_{t+k+1}$$

- $\gamma \in [0, 1]$ est la valeur actuelle des recompenses futures.
- La valeur des recompenses recues $\mathcal{R}$ après $k + 1$ pas est $\gamma^k \mathcal{R}$.
- Cela valorise la recompense immédiate par rapport à la recompense différée:

Pourquoi actualiser (discount) ?

La plupart des processus de récompense et de décision de Markov sont actualisés. Pourquoi ?

- Mathématiquement pratique pour actualiser les récompenses.
- Permet d'éviter des retours infinis dans les processus de Markov cycliques.
- L'incertitude sur le futur peut ne pas être pleinement représentée.
- Si la récompense est financière, les récompenses immédiates peuvent générer plus d'intérêts que celles différées.
- Le comportement humain/animal montre une préférence pour les récompenses immédiates.
- Il est parfois possible d'utiliser des processus de récompense de Markov *non actualisés* (i.e. $\gamma = 1$), par exemple si toutes les séquences se terminent.

La fonction valeur

La fonction de valeur $v(s)$ donne la valeur à long terme de l'état s.

Définition

La *fonction de valeur d'état* $v(s)$ d'un processus de récompense de Markov (MRP) est le retour espéré en partant de l'état s :

$$v(s) = \mathbb{E}\left[G_t \mid S_t = s\right] \tag{1}$$

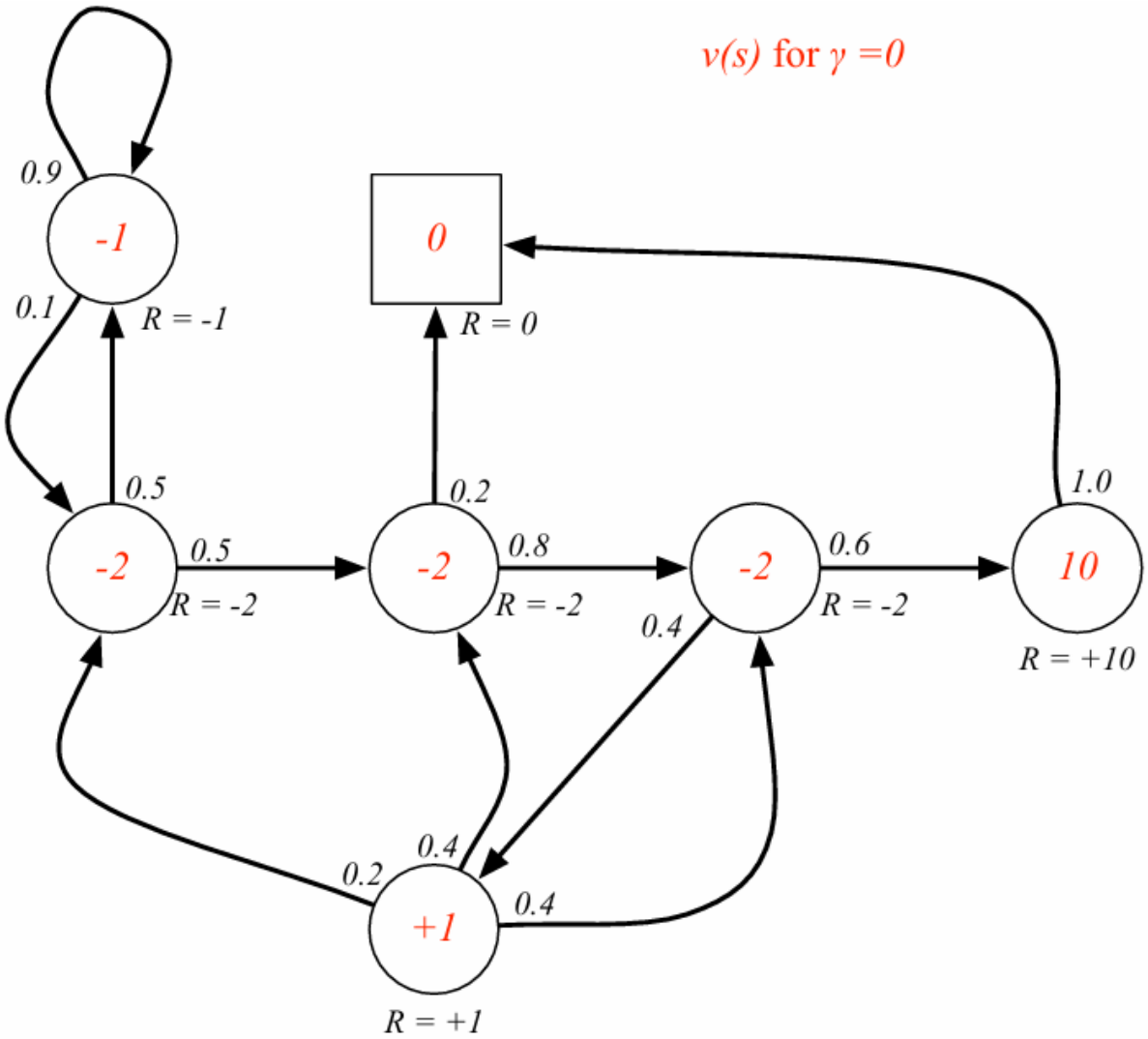
Exemple : Retours pour le MRP étudiant

Exemples de retours pour le MRP étudiant : En commençant à partir de $S_1 = C1$ avec $\gamma = \frac{1}{2}$

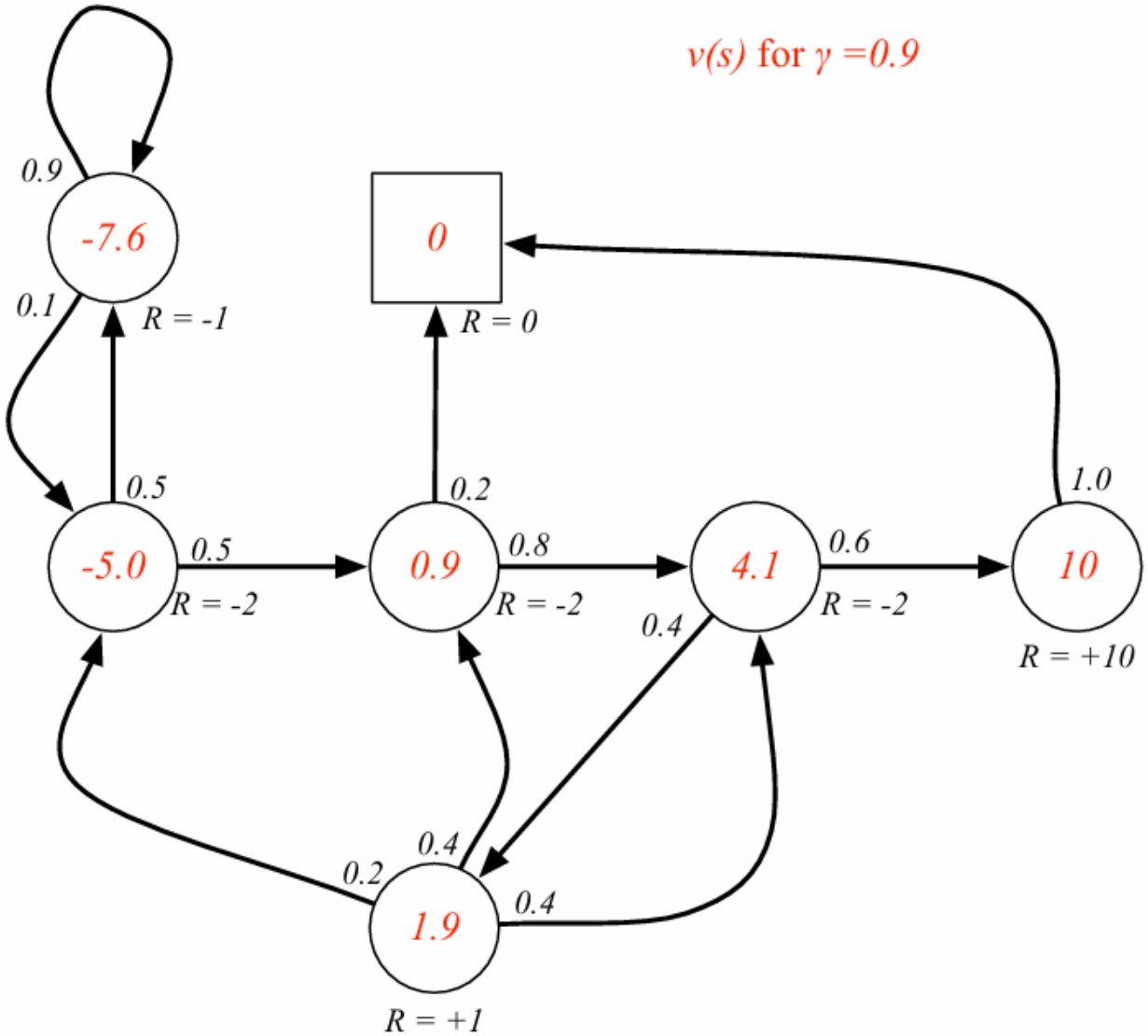
$$G_1 = R_2 + \gamma R_3 + \cdots + \gamma^{T-2} R_T \tag{2}$$

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

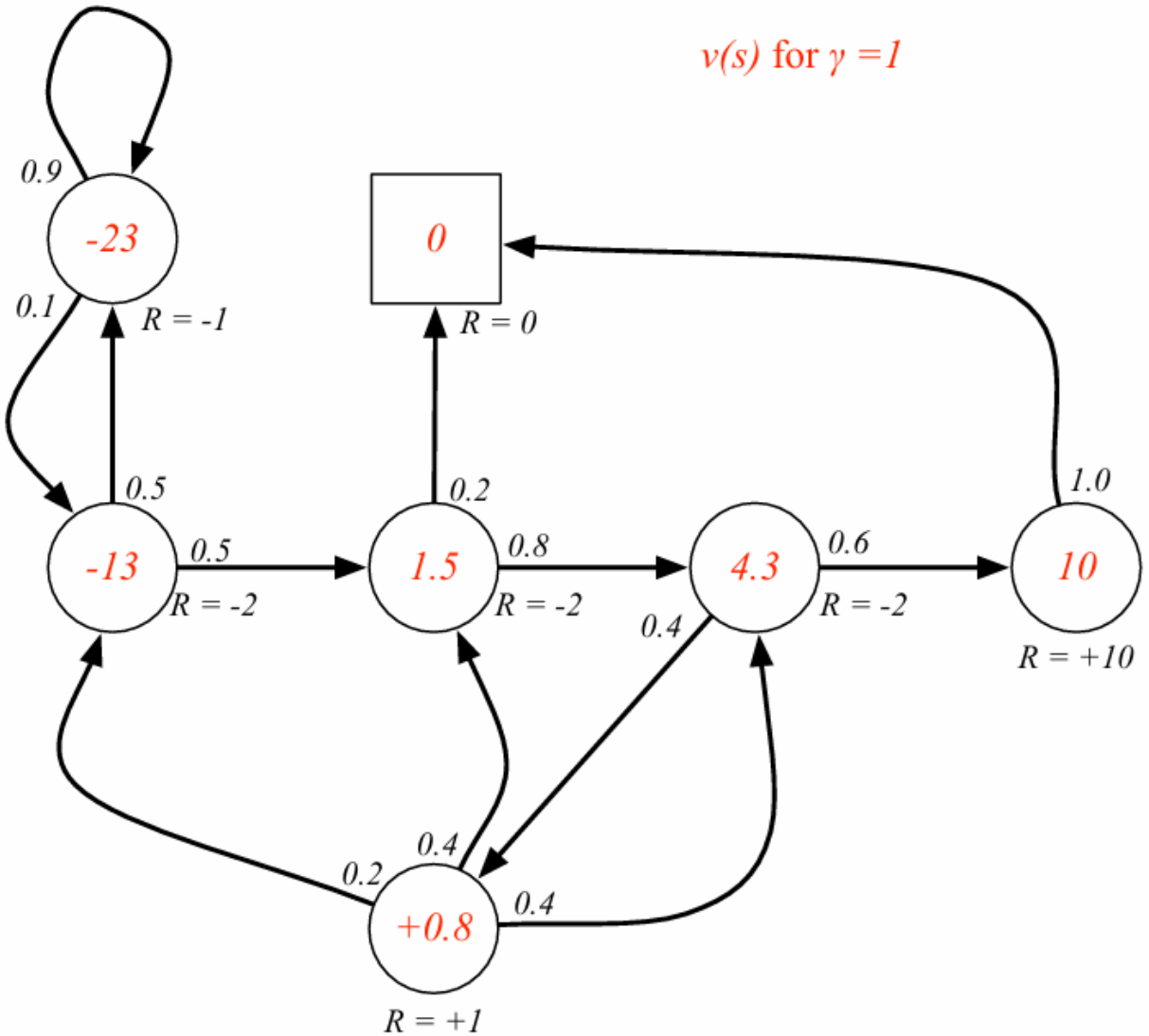
Exemple :La fonction de valeur d'état pour le MRP étudiant



Exemple :La fonction de valeur d'état pour le MRP étudiant



Exemple :La fonction de valeur d'état pour le MRP étudiant



Equation de Bellman pour les processus MRP

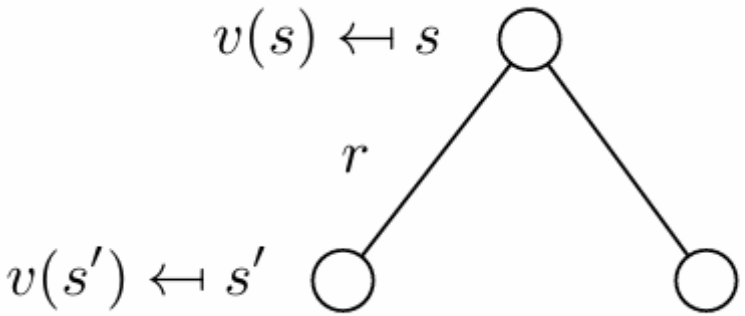
La fonction de valeur peut être décomposée en deux parties :

- Récompense immédiate R_{t+1}
- Valeur actualisée de l'état successeur $\gamma v(S_{t+1})$

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

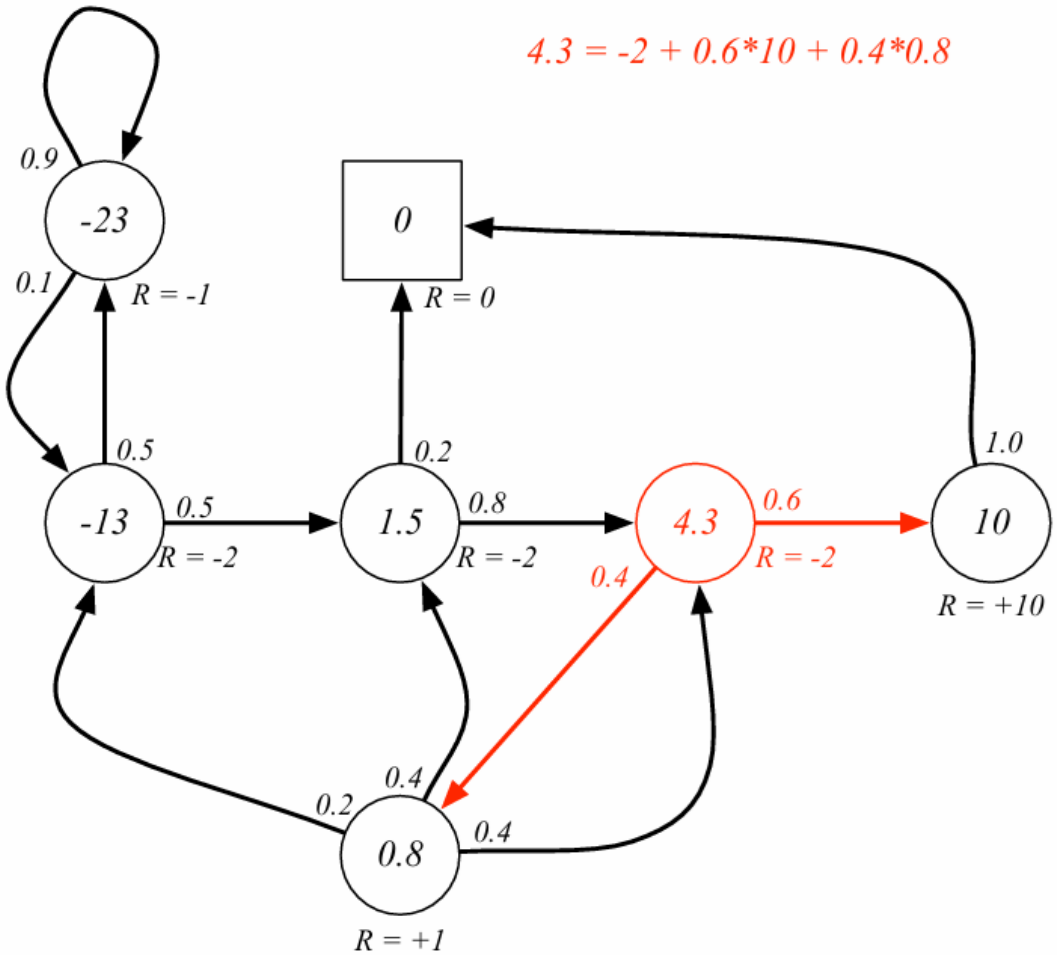
Equation de Bellman pour les processus MRP

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



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

Equation de Bellman pour les processus MRP



Equation de Bellman - Forme matricielle

L'équation de Bellman peut être exprimée de manière concise en utilisant des matrices :

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

où $\mathbf{v}$ est un vecteur colonne avec une entrée par état :

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

Résolution de l'Equation de Bellman

- L'équation de Bellman est une équation linéaire.
- Elle peut être résolue directement :

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

$$(I - \gamma \mathcal{P})\mathbf{v} = \mathcal{R}$$

$$\mathbf{v} = (I - \gamma \mathcal{P})^{-1}\mathcal{R}$$

- La complexité computationnelle est de $\mathcal{O}(n^3)$ pour n états.
- La solution directe est uniquement possible pour de petits MRP.
- Il existe plusieurs méthodes itératives pour les grands MRP, par exemple :
  - Programmation dynamique.
  - Évaluation Monte-Carlo.
  - Apprentissage par différence temporelle (Temporal-Difference Learning).

Chapitre 4 : Processus de Décision de Markov (MDP)

Mohamed Anis BEN LASMAR

A.U.2024-2025