Chaînes de Markov, Processus de décision markoviens et Apprentissage par renforcement

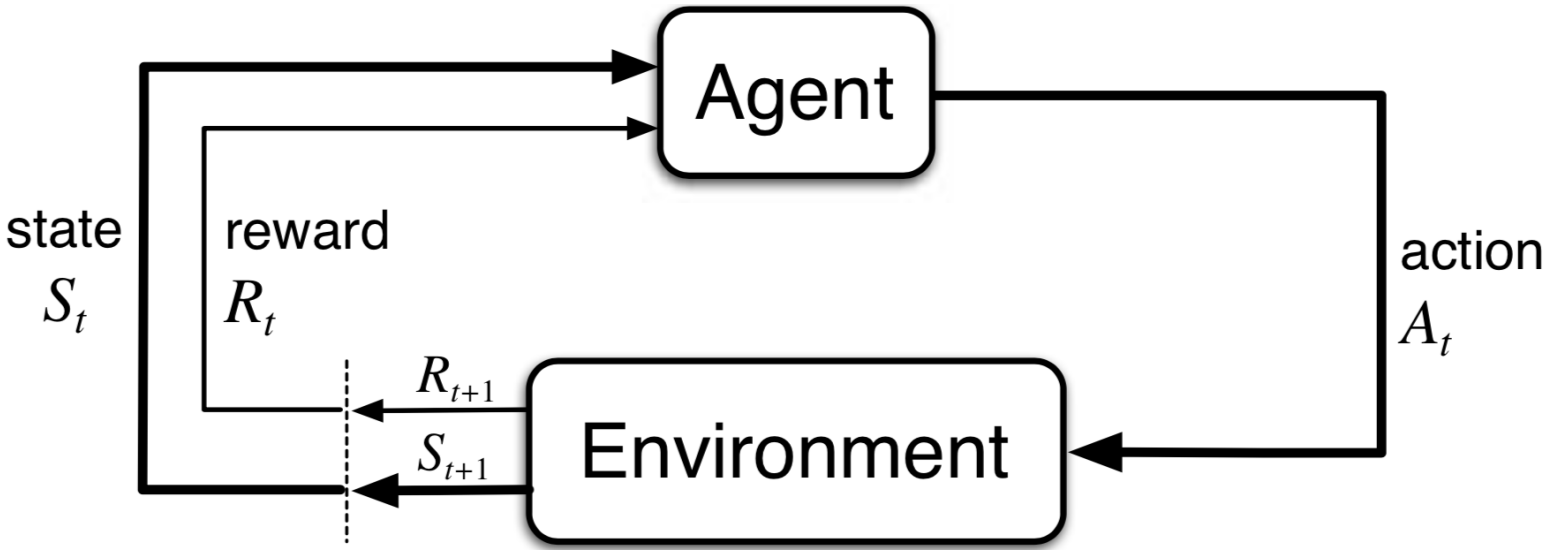
Jean-Philippe Gayon

ISIMA, filières F3 et F4 (2ème et 3ème année)

23 octobre 2024

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Apprentissage par renforcement [1]



- Etat $\rightarrow$ Action $\rightarrow$ Récompense $\rightarrow$ Etat $\rightarrow$ Action $\rightarrow$ Récompense $\rightarrow$ $\ldots$

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2 \cdots$$

- Récompenses et probabilités de transitions inconnues
- Objectif : maximiser les récompenses

1. Sutton et Barto (2018). Reinforcement Learning

Exemples

- Gym : API pour l'apprentissage par renforcement
  - `https://gymnasium.farama.org/`
- Cart-pole
  - `https://towardsdatascience.com/reinforcement-learning-concept-on-cart-pole-with-dqn-79910`
- Jeux Atari
- Faire marcher un robot humanoïde
- Frozen lake

Programme

- Chaînes de Markov : $S_0, S_1, S_2, \cdots$
- Chaînes de Markov avec récompenses (évaluer une politique)

$$S_0, R_0, S_1, R_1, S_2, R_2 \cdots$$

- Processus de décision markovien (trouver la politique optimale)

$$S_0, A_0, R_0, S_1, A_1, R_1, \cdots$$

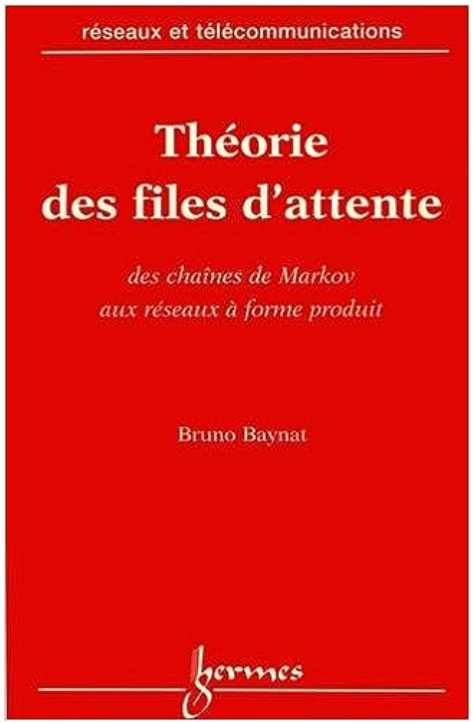
- Apprentissage par renforcement (environnement incertain)
  - Récompenses inconnues
  - Distributions inconnues

Pour aller plus loin sur l'apprentissage par renforcement

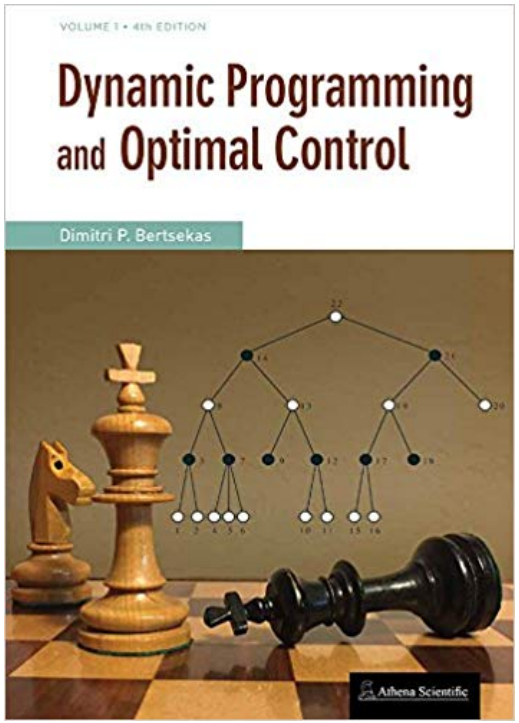
- Présentations et vidéos par David Silver
  `https://www.davidsilver.uk/teaching/`

Livres

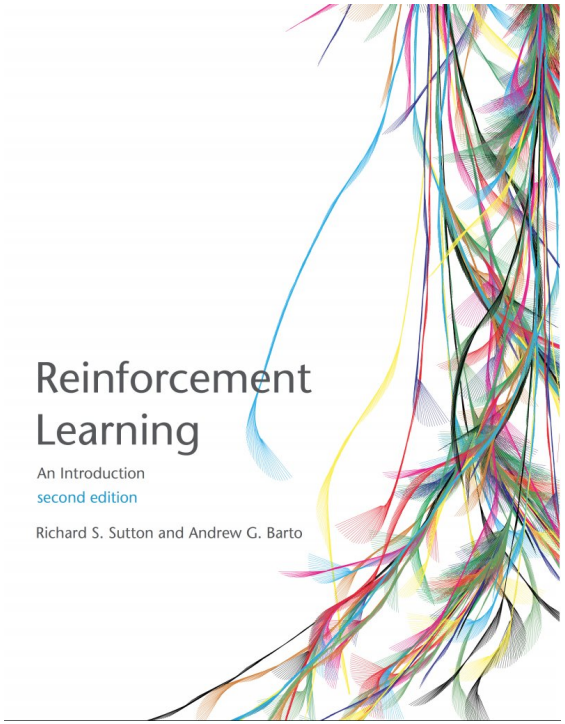
Baynat



Bertsekas



Sutton and Barto



- Extraits à lire au fur et à mesure du cours

Historique : Andreï Andreïevitch Markov (1822-1856)

- Fondateur de la théorie des processus stochastiques
- Chaînes de Markov : phénomènes aléatoires à mémoire courte
- 1913 : Analyse d'un texte de 20 000 lettres d'Alexandre Pouchkine
  - Observation : L'apparition des lettres dépend des précédentes
  - Modèle de Markov d'ordre 0 : chaque lettre apparaît indépendemment des autres
  - Modèle de Markov d'ordre 1 : la n-ième lettre ne dépend que de la $n-1$-ème lettre
  - Modèle de Markov d'ordre k : la n-ième lettre ne dépend que des k lettres précédentes $(n-1, n-2, \cdots, n-k)$
  - On peut faire la même chose avec des mots => générateur parodique
- Générateur de texte markovien : Modèle simple d'IA générative !

Processus aléatoire

- Processus aléatoire : suite de variables aléatoires $X_0, X_1, X_2, \cdots$
- De très nombreux phénomènes sont décrits par des suites de variables aléatoires inter-dépendantes
  - Cours d'une action
  - Évolution d'un stock
  - Nombre d'étudiants en retard
  - Évolution du nombre de requêtes serveur en attente d'être traitées
  - ...
- Remarque : aléatoire = stochastique

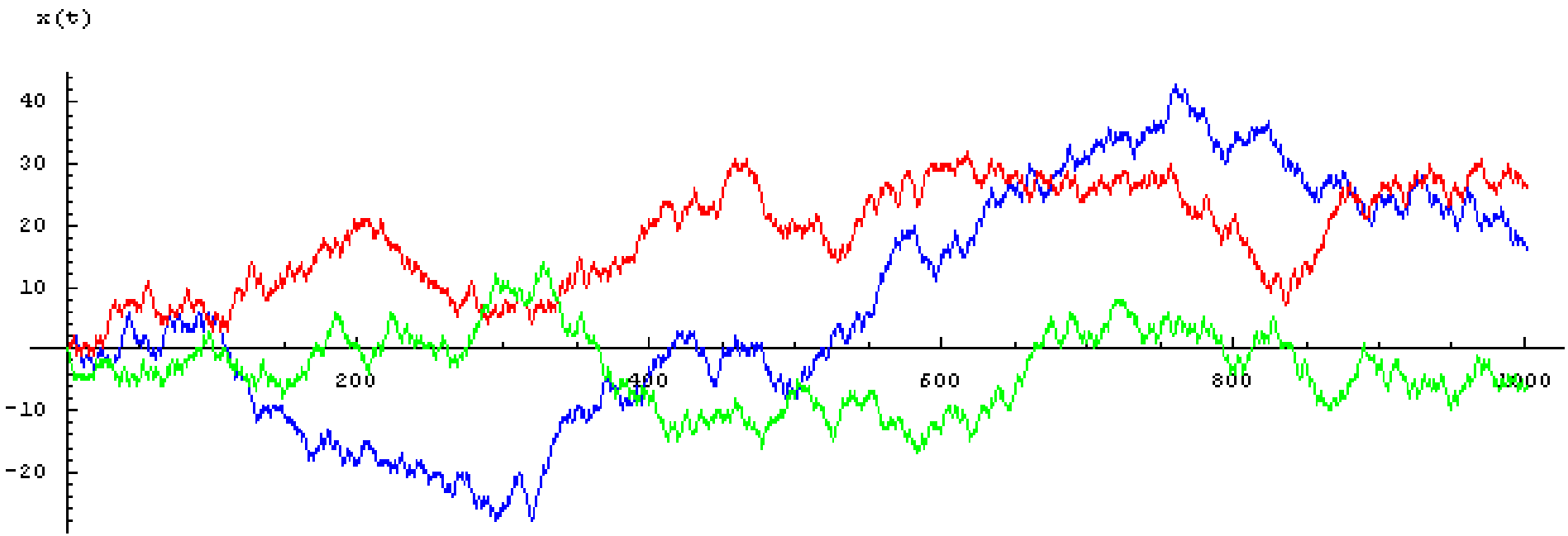
Evolution du CAC 40 sur un an [2]

- X_t : valeur de l'action le jour t
- Espace des temps discret : $t = 1, 2, \cdots, 365$ (en jours)
- Espace d'état discret : $X_t \in \{0, 0.01, 0.02, \cdots\}$ (en euros)

2. Boursorama, 10 sept. 2019

Marche aléatoire 1D [3]

- Un pas en avant (+1) avec probabilité 1/2, un pas en arrière (-1) avec probabilité 1/2
- X_t : position à la t-ième étape
- 3 exemples de trajectoires aléatoires en partant de 0



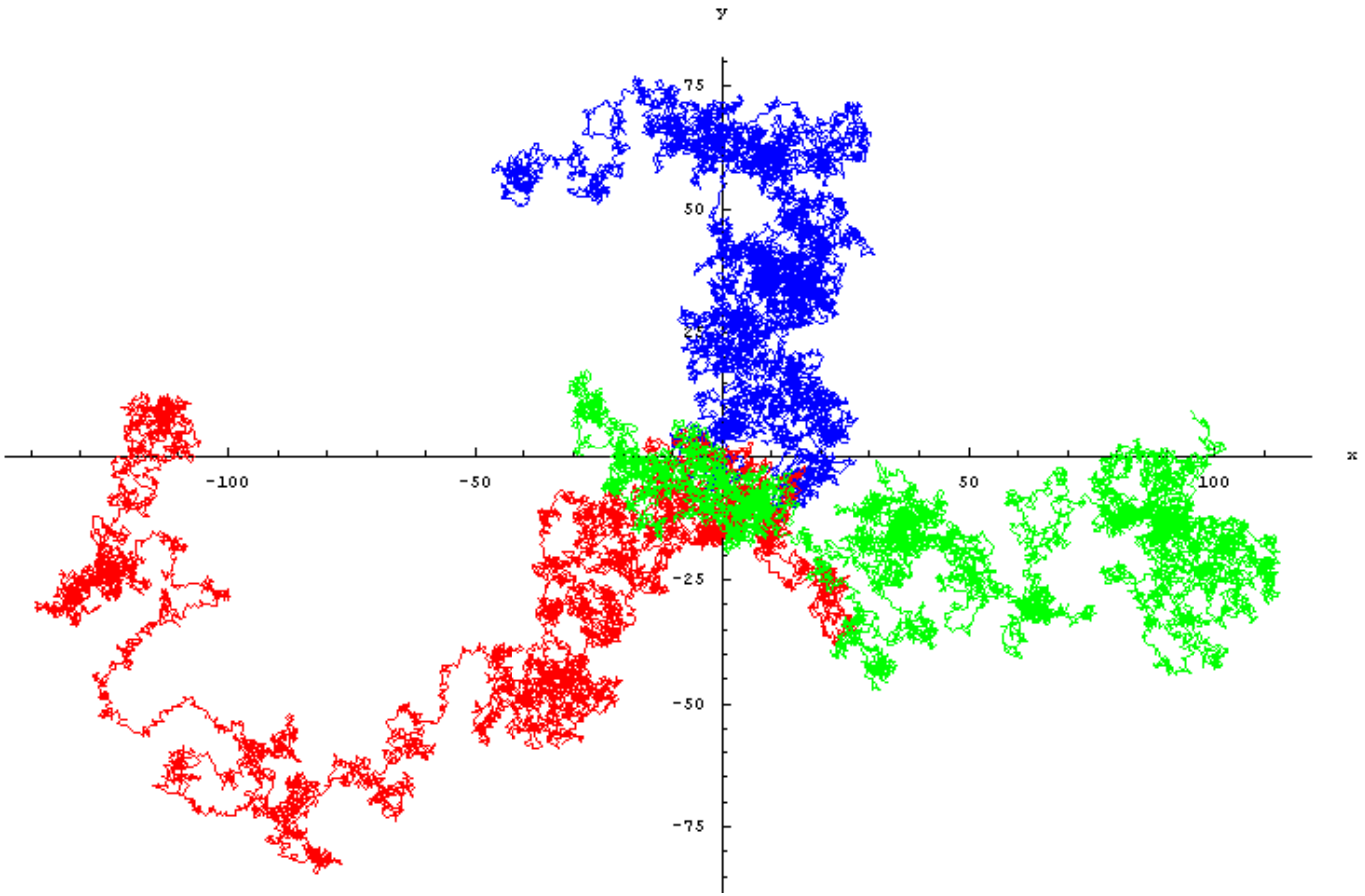
3. Wikipedia, 10 sept. 2019

Quelques questions que nous chercherons à étudier

- Distribution après n étapes : $P(X_t = k) = ?$
- Probabilité de retour à l'origine ?
  - Proba $= 1$
  - On dit que l'état initial est récurrent
- Temps moyen avant de revenir à l'origine ?
  - Infini
  - On dit que l'état initial est récurrent nul

Marche aléatoire 2D [4]

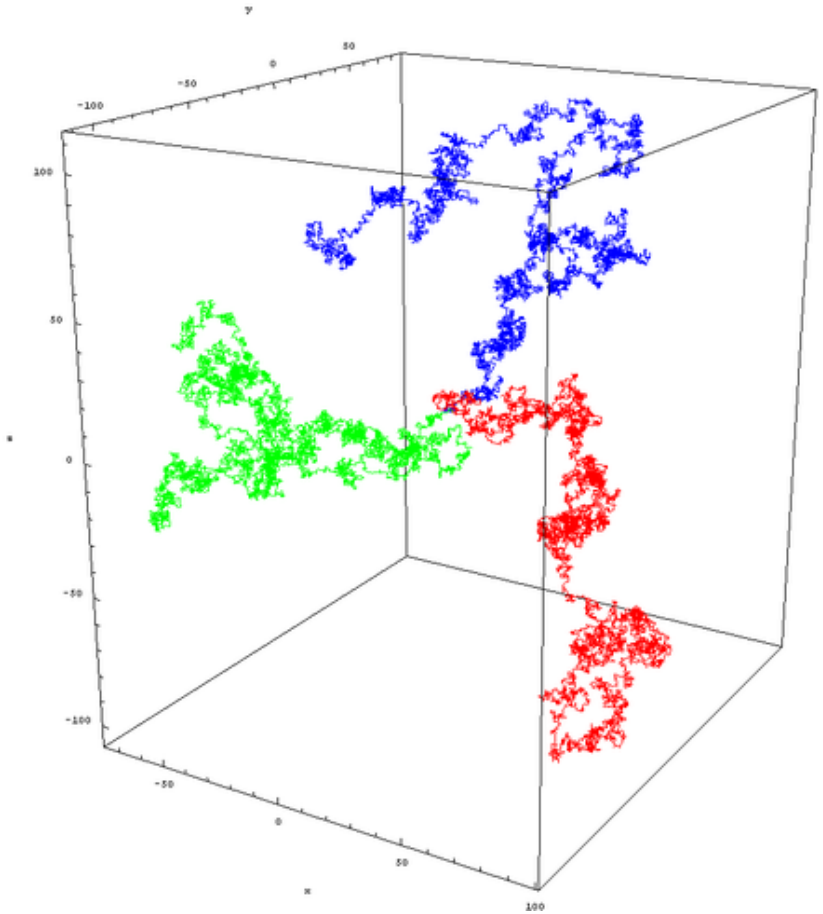
- Avant, arrière, droite, gauche avec probabilité 1/4
- $V_t = (X_t, Y_t)$: position à la t-ième étape



4. Wikipedia, 10 sept. 2019

Marche aléatoire 3D [5]

- avant, arrière, droite, gauche, haut, bas avec probabilité 1/6
- $V_t = (X_t, Y_t, Z_t)$: position à la t-ième étape
- Probabilité de revenir à l'origine < 1 (marche transitoire)



5. Wikipedia, 10 sept. 2019

Processus aléatoire (= Processus stochastique)

- Processus stochastique à temps discret : suite de variables aléatoires $\{X_n, n = 0, 1, 2, \cdots\}$
- Espace d'état E : $X_n \in E$
- Espace d'état fini (ou dénombrable)

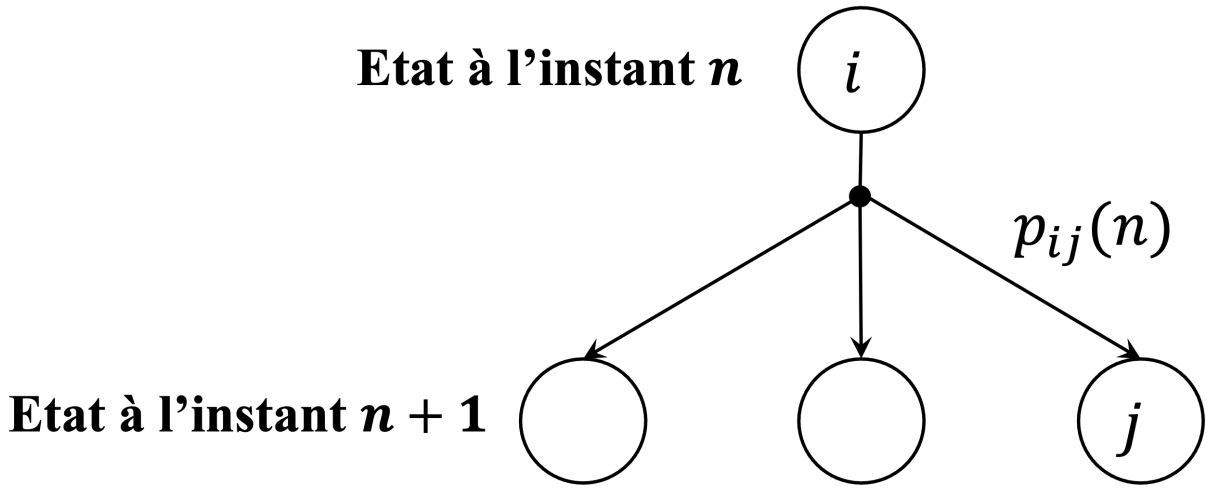
Chaîne de Markov à Temps Discret (CMTD)

- CMTD = Processus stochastique à temps discret sans mémoire

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

- L'état suivant ne dépend que de l'état courant
- Exemple : Marche aléatoire
- $p_{ij}(n) = P(X_{n+1} = j | X_n = i) =$ probabilité de transition de l'état i à l'état j à l'instant n
- CMTD homogène : $p_{ij}(n) = p_{ij}$

Dynamique

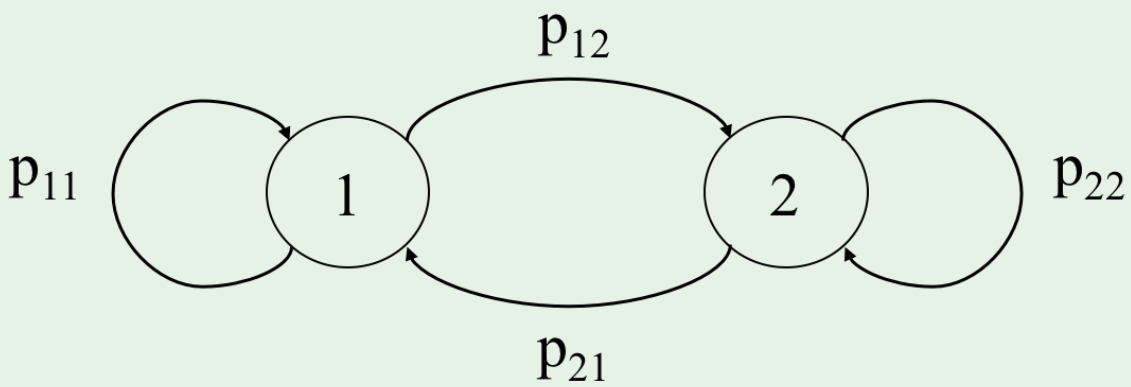


- Homogène si $p_{ij}(n) = p_{ij}$

Graphe d'une CMTD homogène

- Graphe orienté
  - État = sommet
  - Arc orienté de i à j si $p_{ij} > 0$
  - Pondération p_{ij} sur l'arc orienté (i, j)

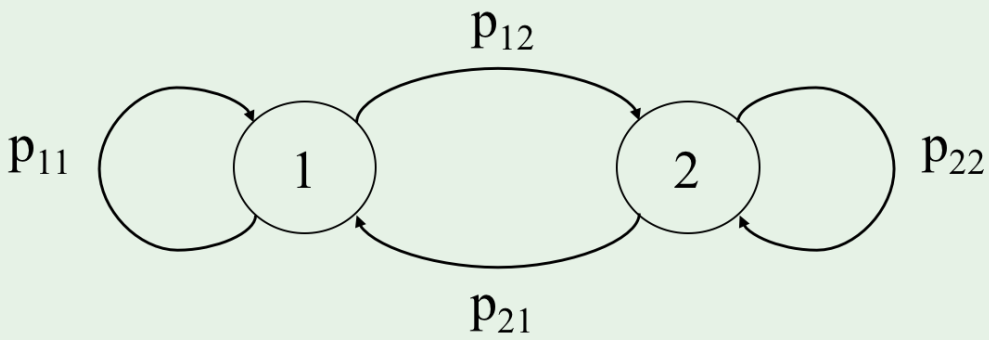
Exemple : CMTD à 2 états



Matrice de transition pour une CMTD fini et homogène

- Matrice de transition : $P = (p_{ij})$

Exemple : CMTD à 2 états



$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

- Une matrice de transition est une matrice stochastique : la somme des éléments de chaque ligne vaut 1

$$\sum_j p_{ij} = 1, \forall i$$

Propriétés d'une matrice stochastique

- Le produit de 2 matrices stochastiques est stochastique
- Corollaire : si P stochastique alors P^n stochastique.
- 1 est valeur propre
- Le rayon spectral vaut 1 : les valeurs propres sont inférieures ou égales à 1 en module

Pile ou face jusqu'à la faillite

Principe du jeu

- Les joueurs A et B possèdent respectivement α et β euros
- A chaque tour, ils misent un euro et jouent à pile au face
- Le jeu s'arrête quand l'un des deux est ruiné
- Durée du jeu = nombre de tours joués

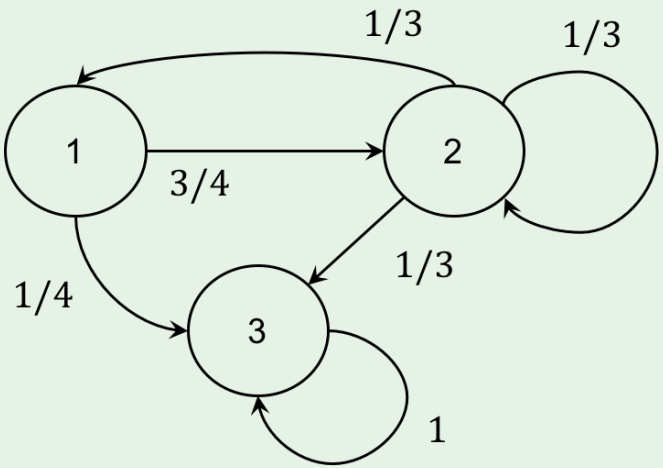
Questions

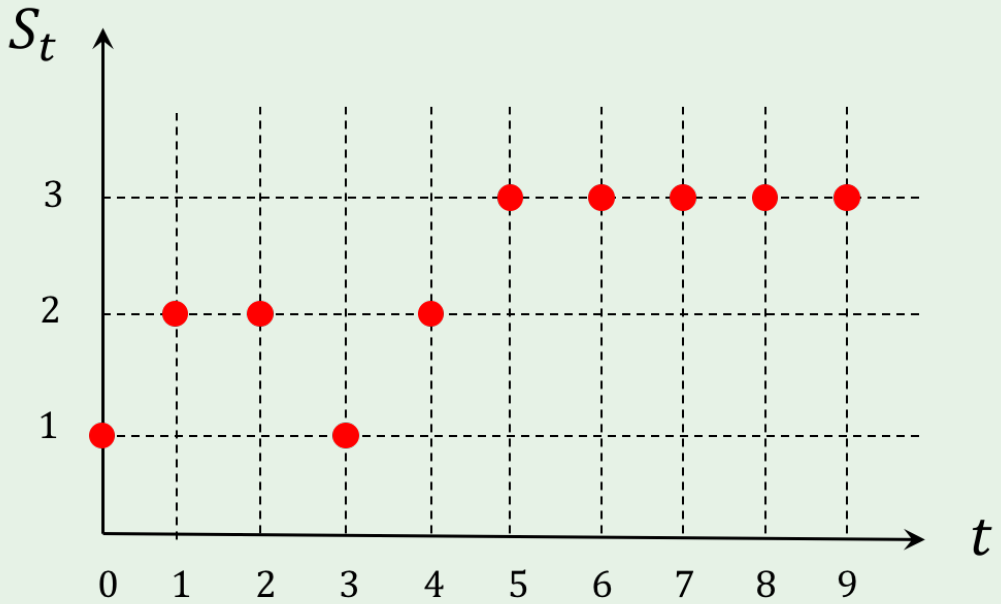
- Quelle est la probabilité pour chacun des joueurs d'être ruiné ?
- Quelle est la durée moyenne du jeu ?

Épisodes

- Épisode : $X_0, X_1, \cdots, X_T$
- T : instant final (potentiellement une variable aléatoire)

Example

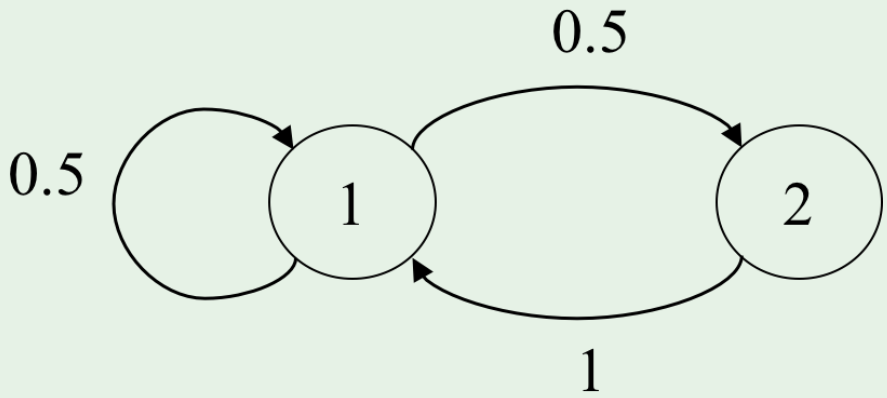




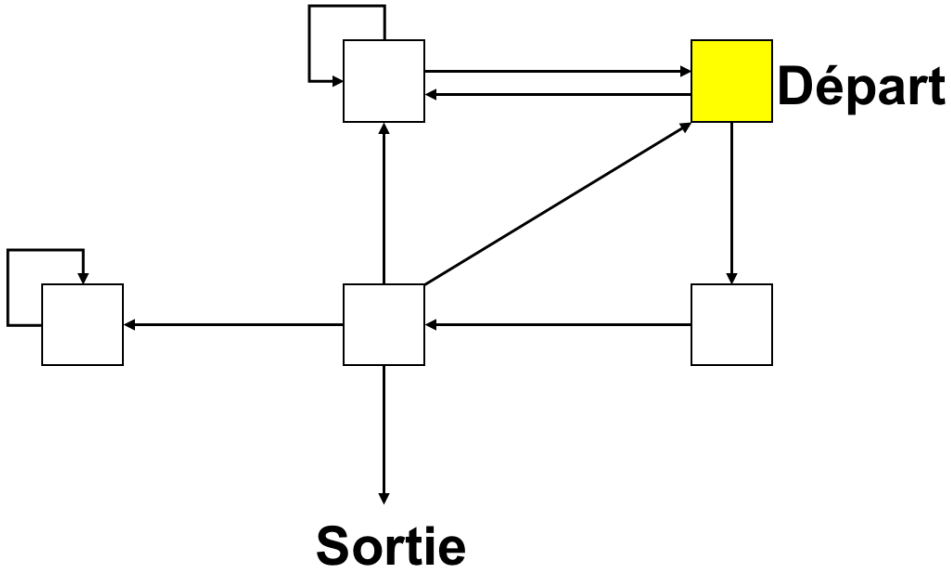
- L'épisode se termine quand on atteint l'état absorbant 3.
- Sur cette réalisation, $T = 5$.

Épisode sans fin

Exemple



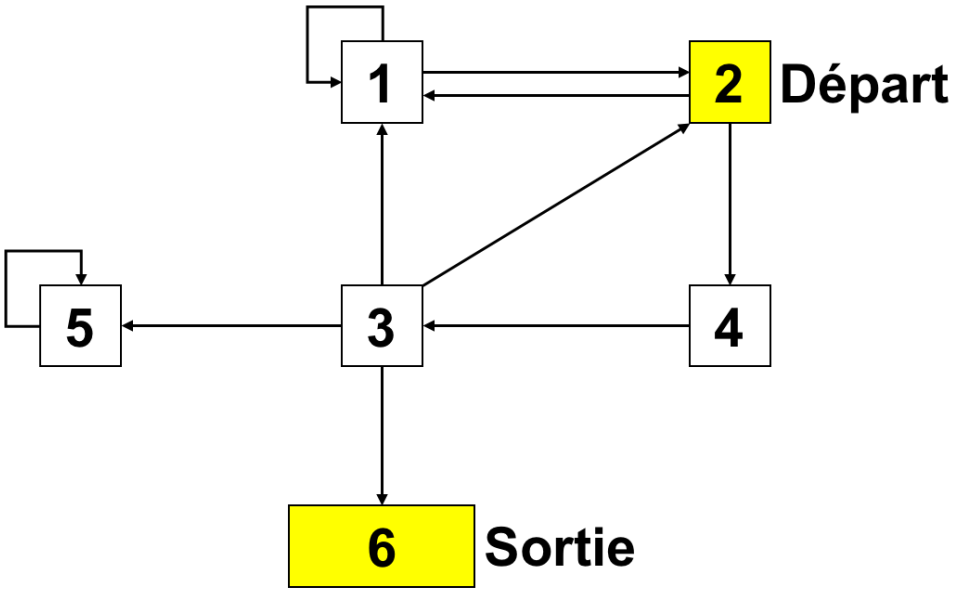
Exemple : Souris dans un labyrinthe

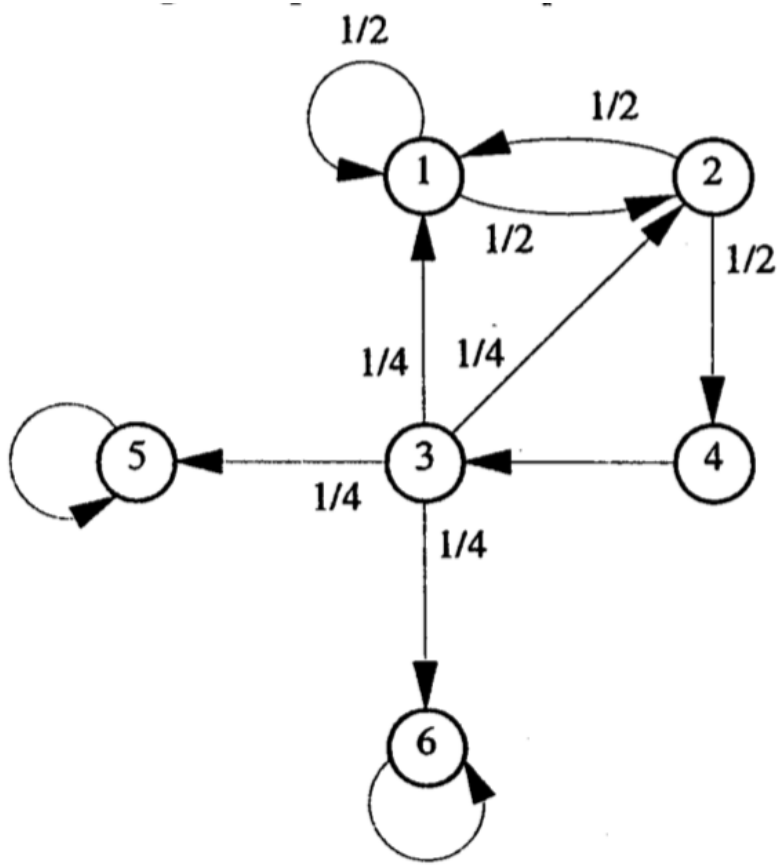


- La souris est amnésique et choisit un des couloirs de façon équiprobable
- On s'intéresse à la succession des pièces visitées (et pas au temps passé dans chaque pièce)
- Modéliser ce système par une CMTD, donner son graphe et sa matrice de transition

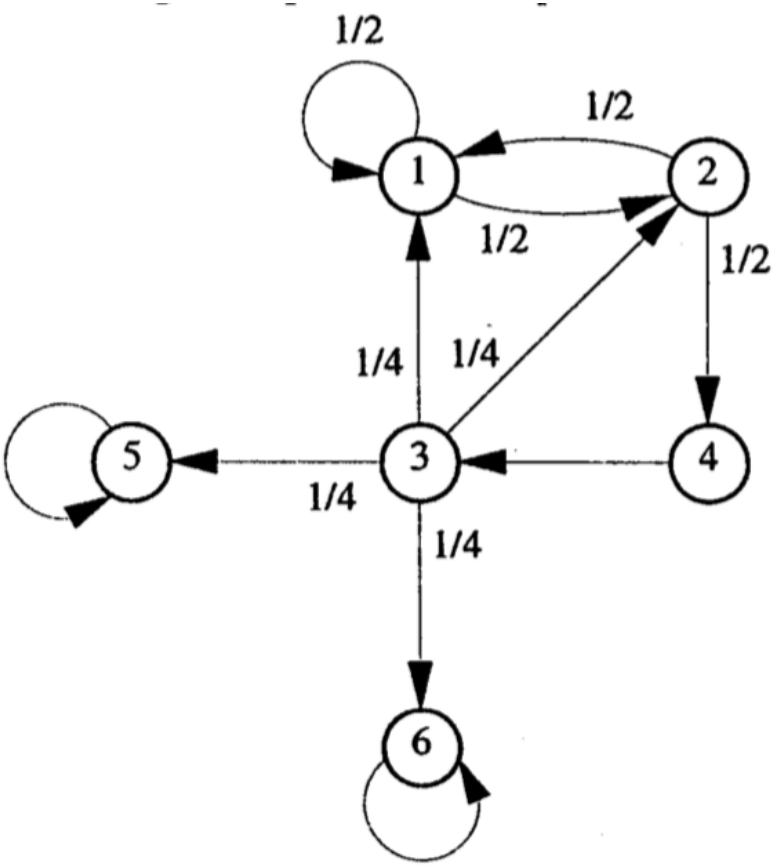
Souris dans un labyrinthe : CMTD et graphe

- X_n : position après le n-ième couloir emprunté
- $X_0 = 2$



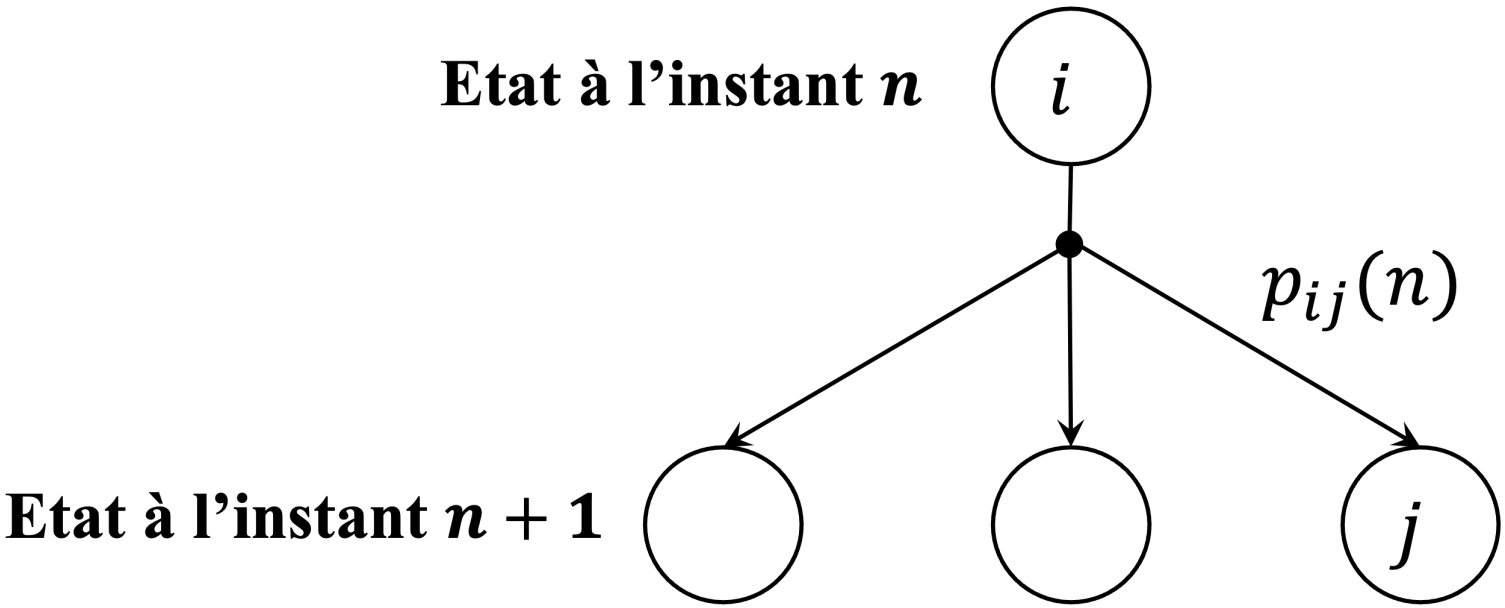


Souris dans un labyrhinte : matrice de transition



$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Régime transitoire



- $\pi_j(n) = P(X_n = j)$: probabilité d'être dans l'état j à l'instant n

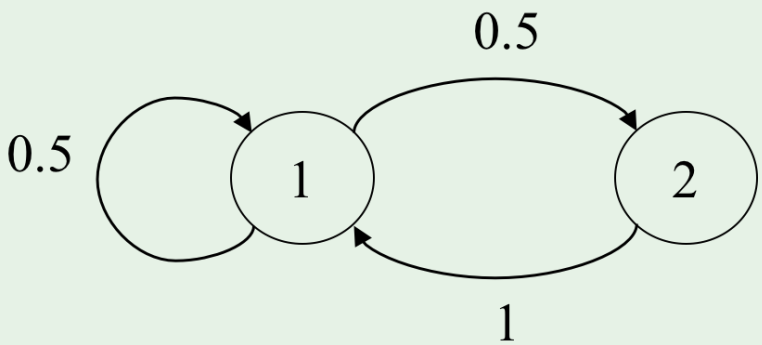
$$\pi_j(n+1) = \sum_{i \in E} p_{ij} \pi_i(n)$$

Régime transitoire (forme matricielle)

- $\pi(n) = (\pi_1(n), \pi_2(n), \cdots, \pi_{|E|}(n))$: distribution à l'instant n

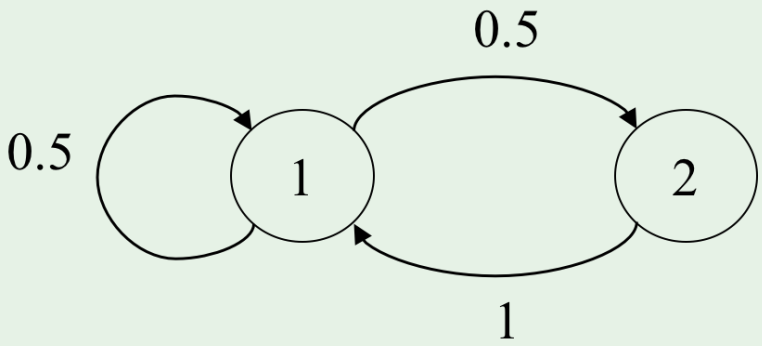
$$\pi(n+1) = \pi(n)P$$
$$\pi(n) = \pi(0)P^n$$

Exemple : CMTD à 2 états



- Avec $\pi(0) = (1, 0)$, calculer $\pi(1), \pi(2), \pi(3)$.

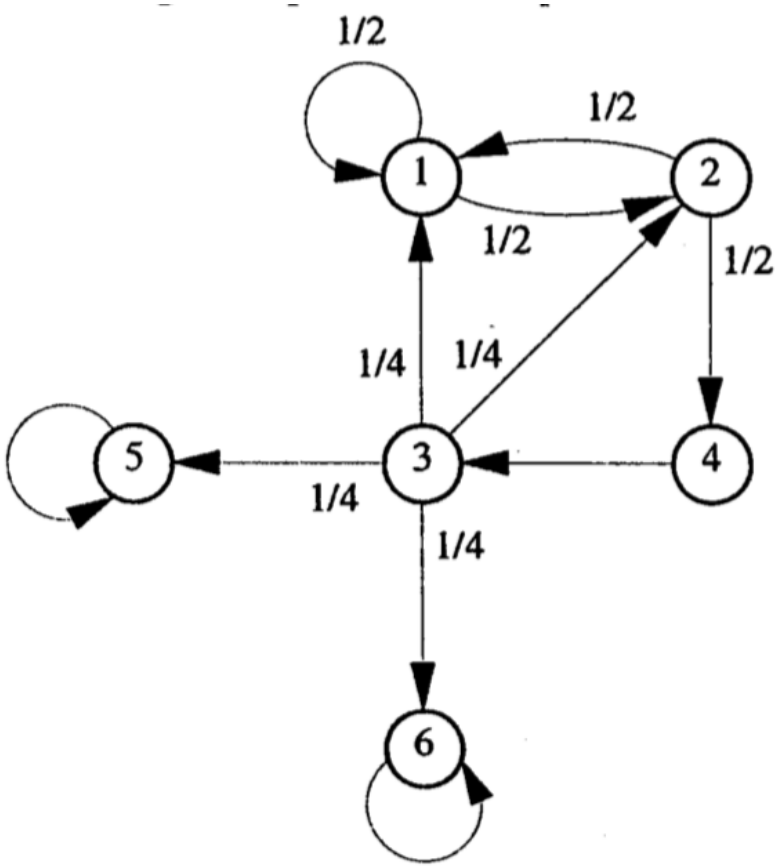
Exemple : CMTD à 2 états



$$P = \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}$$

$$\pi(0) = (1, 0)$$
$$\pi(1) = \pi(0)P = (1/2, 1/2)$$
$$\pi(2) = \pi(1)P = (3/4, 1/4)$$
$$\pi(3) = \pi(2)P = (5/8, 3/8),$$
$$\pi(10) = (0.666015625, 0.333984375)$$
$$\pi(\infty) = (2/3, 1/3)$$

Souris dans un labyrinthe : régime transitoire



- Supposons que $X_0 = 2$, calculer $\pi(0), \pi(1), \pi(2)$, $\pi(3)$ et $\lim_{n\to\infty} \pi(n)$.

Souris dans un labyrinthe : régime transitoire (réponse)

$$\pi(0) = (0, 1, 0, 0, 0, 0)$$

$$\pi(1) = \left(\frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0\right)$$

$$\pi(2) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0, 0, 0\right)$$

$$\pi(3) = \left(\frac{3}{8}, \frac{1}{4}, 0, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$$

$$\pi(\infty) = (0, 0, 0, 0, 1/2, 1/2)$$

Probabilité d'aller de i à j en n étapes

- $p_{ij}^{(n)}$: Probabilité, commençant en i, d'être en j après n transitions

$$p_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$$

$$P_{ij}^{(n)} = P^n$$

- Equations de Chapman-Kolmogorov

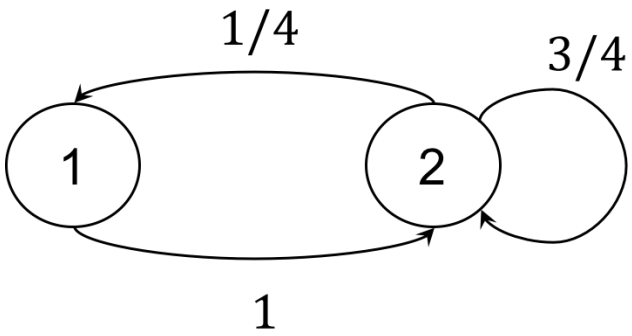
$$p_{ij}^{(n+m)} = \sum_{k \in \mathcal{S}} p_{ik}^{(n)} p_{kj}^{(m)}$$

$$P^{n+m} = P^n P^m$$

Travail à la maison

- Lire chapitre 3.1, pages 55 à 62
- Exercice 1 (téléphone arabe)
- Exercice 4 (matrices stochastiques)

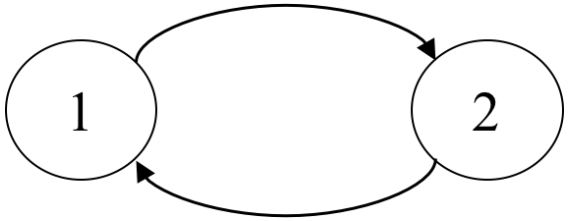
Convergence vers une distribution limite : un exemple



$$\begin{cases} \pi_1(n+1) = \frac{1}{4}\pi_2(n) \\ \pi_2(n+1) = \pi_n(1) + \frac{3}{4}\pi_2(n) \end{cases}$$

n	$\pi_1(n)$	$\pi_2(n)$
0	1	0
1	0	1
2	0.25	0.75
3	0.1875	0.8125
4	0.203125	0.796875
...	...	...
13	0.2	0.8
14	0.2	0.8

Exemple de non convergence



$$\begin{cases} \pi_1(n+1) = \pi_2(n) \\ \pi_2(n+1) = \pi_1(n) \end{cases}$$

n	$\pi_1(n)$	$\pi_2(n)$
0	1	0
1	0	1
2	1	0
3	0	1
4	1	0
...	...	...

- En revanche, convergence au sens de Césaro

$$\underbrace{\frac{1}{n}\sum_{k=1}^{n} \pi_1(k)}_{\text{Proportion de temps passé en 1 au cours des } n \text{ premières étapes}} \underset{n \to +\infty}{\longrightarrow} 0.5$$

Si convergence, alors ...

- Supposons que $\pi(n) = (\pi_1(n), \pi_2(n), \cdots)$ converge vers $\pi = (\pi_1, \pi_2, \cdots)$
- Alors

$$\begin{aligned} \pi(n+1) &= \pi(n)P \\ &\Downarrow n \to \infty \\ \pi &= \pi P \end{aligned}$$

- Questions
  - Quels sont les critères de convergence ?
  - La distribution limite dépend-elle de l'état initial ?

Distribution stationnaire

- "Stationnaire" signifie "qui ne dépend pas du temps"

$$\pi(n) = \pi$$

$$\pi(n+1) = \pi(n)P \Rightarrow \pi = \pi P$$

- Une distribution de probabilité π est dite stationnaire si et seulement si

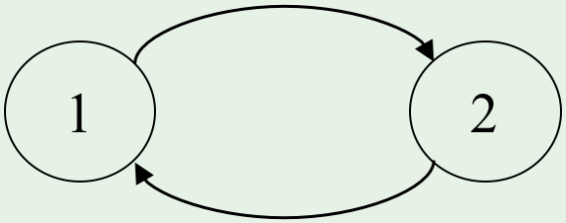
$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases}$$

Lien entre distribution stationnaire et distribution limite

Propriété

La distribution limite est la distribution stationnaire.

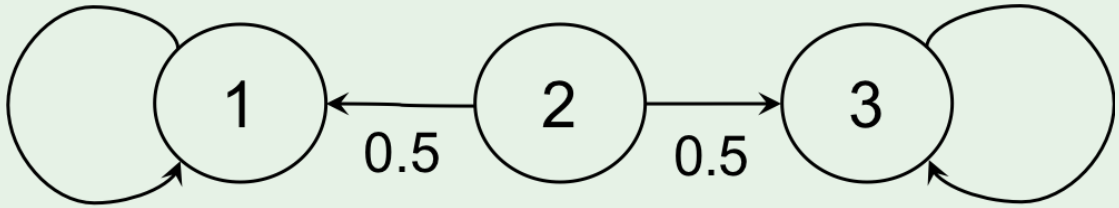
La réciproque est fausse



$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases} \Leftrightarrow \pi_1 = \pi_2 = 0.5$$

La distribution stationnaire est-elle toujours unique ?

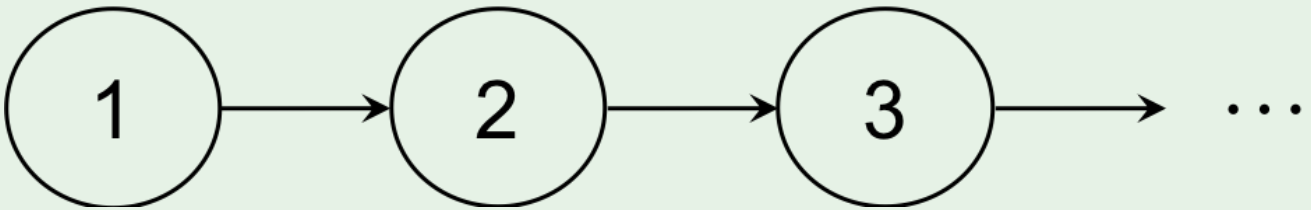
La réponse est non



$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_2 = 0 \\ \pi_1 + \pi_3 = 1 \end{cases}$$

Existe-il toujours une distribution stationnaire ?

La réponse est non

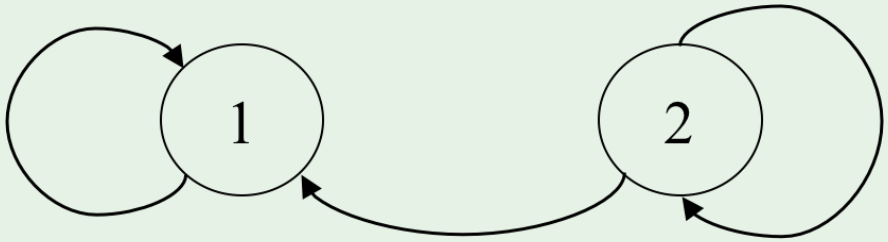


$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_i = 0, \forall i \\ \sum_i \pi_i = 1 \end{cases}$$

Classification des états

- État absorbant : $p_{jj} = 1$

Exemple

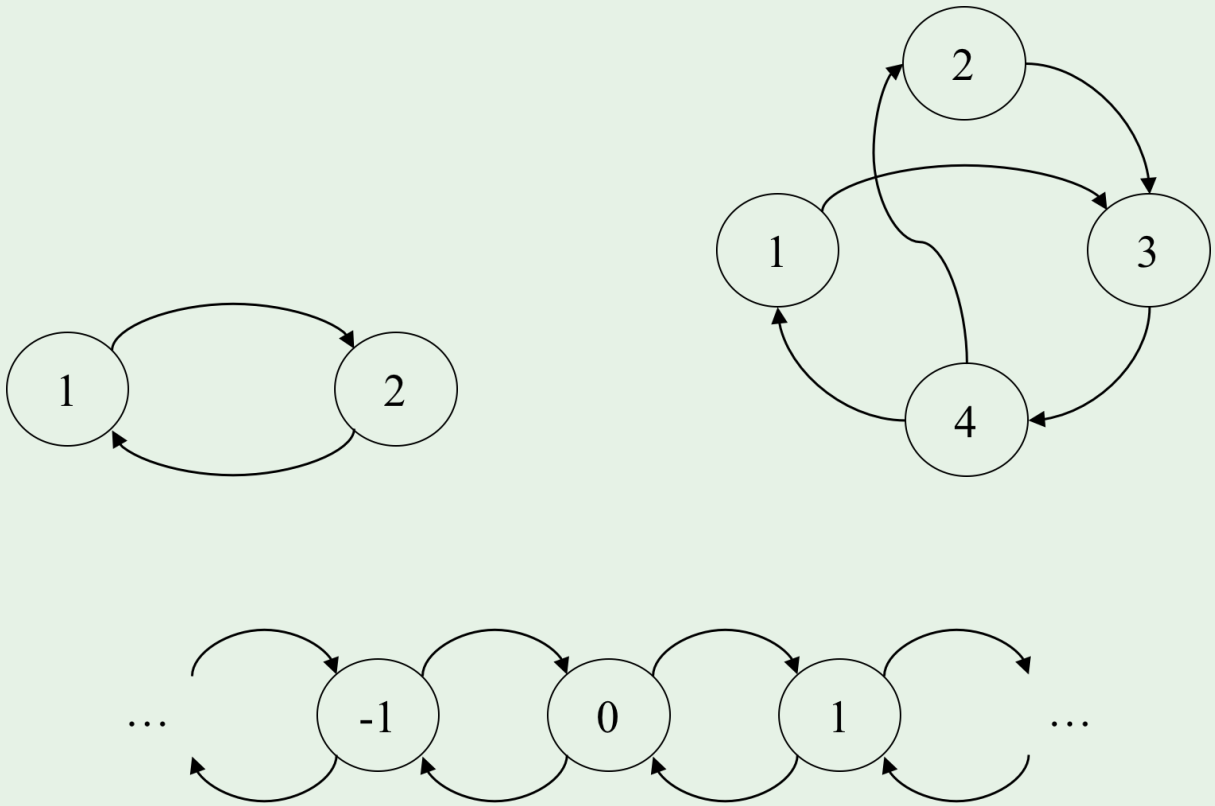


- État périodique ou apériodique
- État récurrent ou transitoire

État périodique versus état apériodique

- État périodique j : il existe $k > 1$ tel qu'on ne puisse retourner en j qu'après un nombre d'étapes multiple de k
- État apériodique : un état n'étant pas périodique

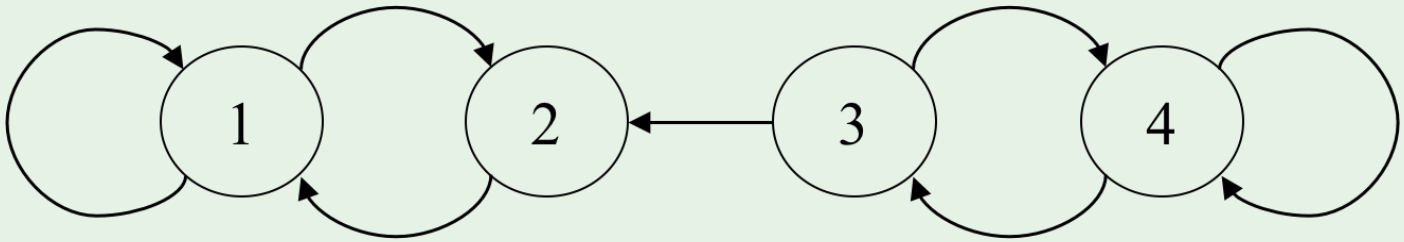
Exemple : CMTD périodiques



État récurrent versus état transitoire

- $f_{jj}^{(n)}$: probabilité que le premier retour en j, en partant de j, ait lieu en n étapes
- $f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}$: probabilité, en partant de j, de revenir un jour en j
- Etat récurrent : $f_{jj} = 1$
- Etat transitoire : $f_{jj} < 1$

Exemple : CMTD fini



- 1 et 2 sont récurrents
- 3 et 4 sont transitoires

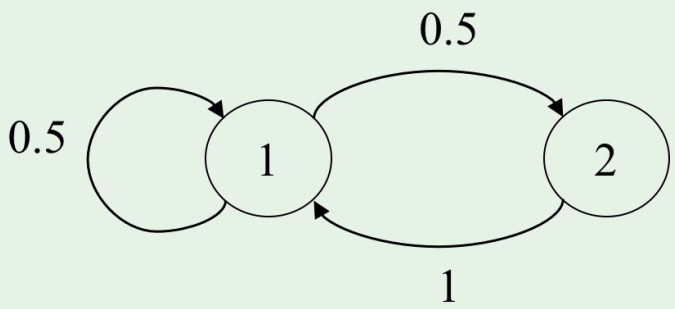
Récurrent nul versus récurrent non nul

- M_j : temps moyen entre 2 visites de j
- État récurrent non nul : $M_j < +\infty$
- État récurrent nul : $M_j = +\infty$

Proposition

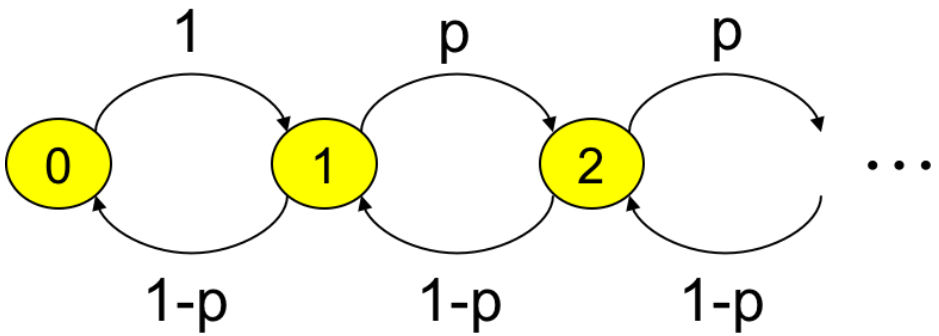
Soit E un espace d'état fini. Alors tout état récurrent est récurrent non nul.

Déterminer f_{22} et M_2



$$f_{22} = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = 1, \quad M_2 = \sum_{i=2}^{\infty} \left(\frac{1}{2}\right)^{i-1} = 3$$

Un exemple de CMTD infini



- $p > 1/2$: tous les états sont transitoires
- $p = 1/2$: tous les états sont récurrents nuls
- $0 < p < 1/2$: tous les états sont récurrents non nuls

CMTD irréductible versus irréductible

- Une CMTD est irréductible si son graphe est fortement connexe

$$\forall i, j \in E, \exists m > 1, p_{ij}^{(m)} > 0$$

- CMTD réductible = CMTD qui n'est pas irréductible

Nature des états d'une CMTD irréductible

Théorème (Admis)

Tous les états d'une CMTD irréductible sont de même nature :

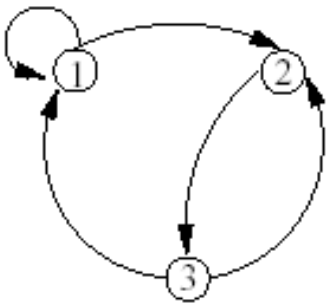
- *Tous transitoires, ou tous récurrents non nuls, ou tous récurrents nuls*
- *Tous périodiques ou tous apériodiques*

Corollaire (Nombre d'états fini)

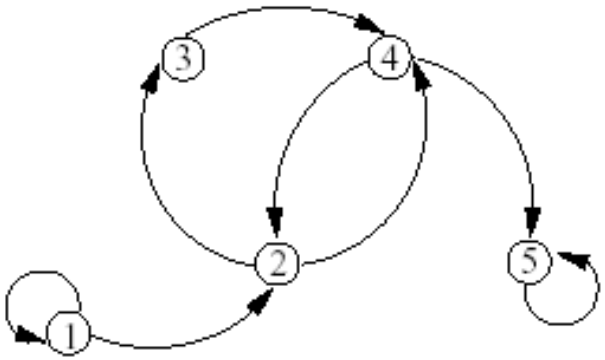
Tous les états d'une CMTD irréductible finie sont récurrents non nuls.

Exercice 5 : classification des états

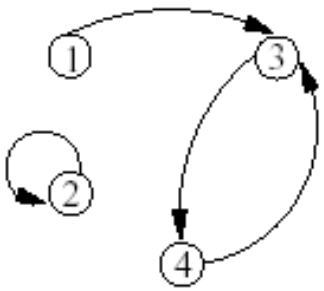
a)



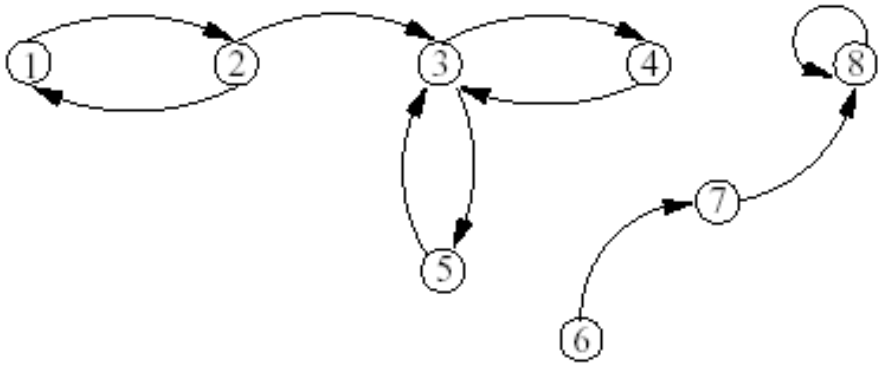
b)



c)



d)



Travail à la maison

- Lire chapitre 3.1, pages 62 à 68
- Exercice 5 (classification des états)
- Exercice 3 (marche aléatoire symétrique)

Distribution limite pour une CMTD irréductible et apériodique

Théorème (admis)

Pour une CMTD irréductible et apériodique, la distribution limite $\pi = (\pi_1, \pi_2, \cdots)$ *existe et est indépendante de la distribution initiale* $\pi(0)$ *:*

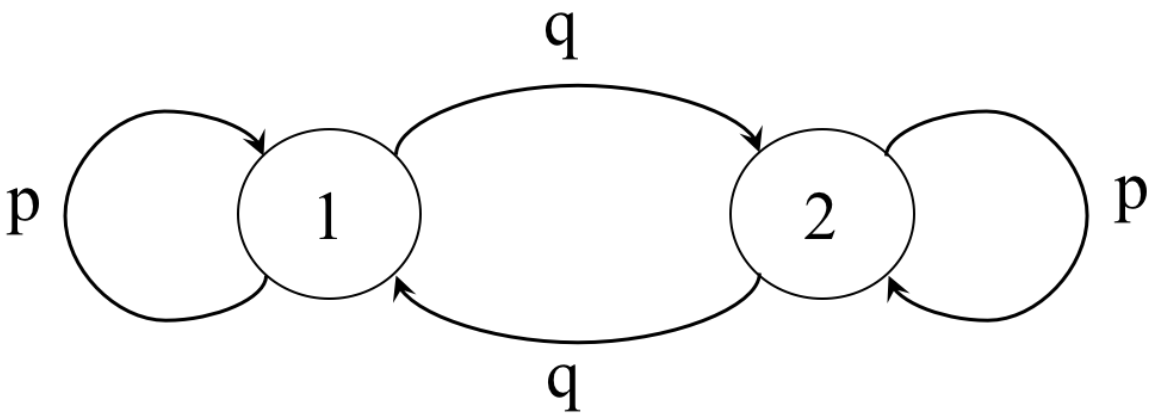
- *Si les états sont tous transitoires ou tous récurrents nuls, alors*

$$\forall j \in E, \pi_j = 0$$

- *Si les états sont tous récurrents non nuls, alors les* π_j *satisfont*

$$(S) : \begin{cases} \pi_j = \sum_{i \in E} \pi_i p_{ij}, \forall j \in E \\ \sum_{j \in E} \pi_j = 1 \end{cases}$$

Exemple : Téléphone arabe



- La CMTD est irréductible et apériodique
- Tous les états sont récurrents non nuls
- La distribution limite existe et satisfait

$$\begin{cases} \pi_1 = p\pi_1 + q\pi_2 \\ \pi_2 = p\pi_2 + q\pi_1 \\ \pi_1 + \pi_2 = 1 \end{cases} \Rightarrow \pi_1 = \pi_2 = 0.5$$

Interprétation de la distribution limite

Théorème (Césaro)

Soit (u_n) une suite de nombres réels convergeant vers L, alors la moyenne de Césaro

$$c_n = \frac{1}{n}\sum_{k=1}^{n} u_k$$

converge également vers L.

La réciproque est fausse

- $u_n = (-1)^n$ ne converge pas
- Mais c_n converge vers 0

Interprétation de la distribution limite

- Supposons que $\pi_i(n) \underset{n \to +\infty}{\longrightarrow} \pi_i$
- Alors, d'après le théorème de Césaro

$$\underbrace{\frac{1}{n} \sum_{k=1}^{n} \pi_i(k)}_{\text{Proportion de temps passé en i au cours des } n \text{ premières étapes}} \underset{n \to +\infty}{\longrightarrow} \pi_i$$

- Interprétation de π_i : proportion de temps passé en i sur un horizon infini

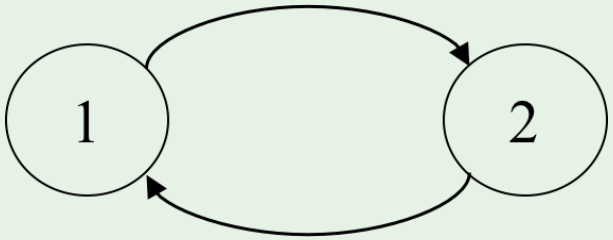
CMTD irréductible et périodique (finie)

- $\pi(n)$ n'a pas de limite

$$(S): \begin{cases} \pi_j = \sum_{i \in E} \pi_i p_{ij}, \forall j \in E \\ \sum_{j \in E} \pi_j = 1 \end{cases}$$

- (S) admet une solution unique où π_j représente la proportion de temps passé dans l'état j sur un horizon infini

Exemple



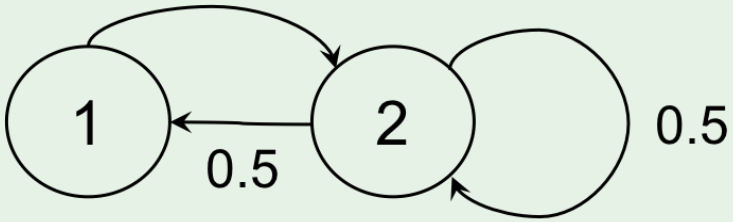
$$\begin{cases} \pi_1 = \pi_2 \\ \pi_1 + \pi_2 = 1 \end{cases} \Longrightarrow \pi_1 = \pi_2 = 0.5$$

Temps moyen entre deux visites d'un état

- M_j : temps moyen entre deux visites d'un état j

$$M_j = \frac{1}{\pi_j}$$

Exemple



Calculer M_1 et M_2 par deux méthodes (avec ou sans le résultat ci-dessus).

Résolution de (S) pour une CMTD à m états (facultatif)

- Soit $1 = (1, \cdots, 1)$

$$(S) \begin{cases} \pi = \pi P \\ \pi 1^T = 1 \end{cases}$$

- A : matrice $m \times m$ avec toutes les entrées à 1 ($a_{ij} = 1$)

$$\pi A = 1$$

- Il suit avec I la matrice identité $m \times m$

$$I\pi + \pi A = \pi P + 1$$

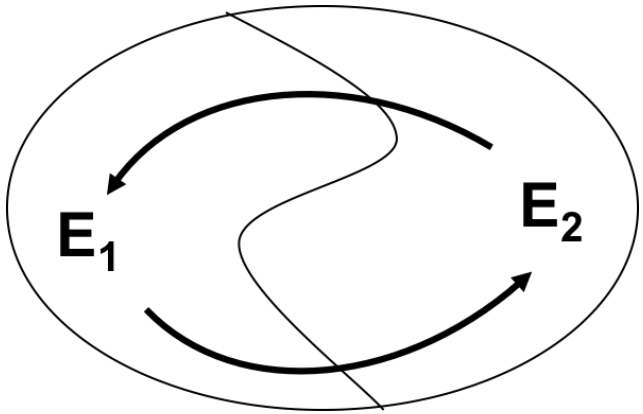
- Propriété : $I - P + A$ est inversible pour une CMTD apériodique et irréductible

$$\pi = 1(I - P + A)^{-1}$$

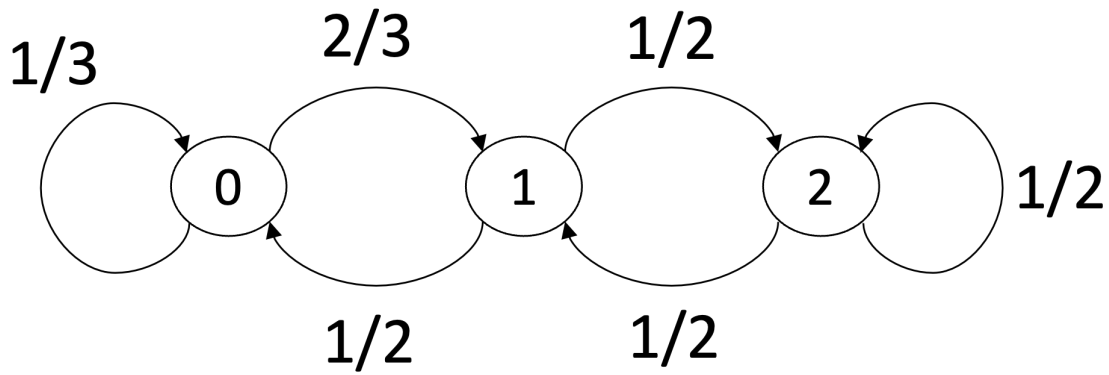
Conservation des flux

- $\pi_i p_{ij}$: Nombre de transitions de i vers j par unité de temps
- Partition des états : $E = E_1 \cup E_2$

$$\overbrace{\sum_{i \in E_1} \sum_{j \in E_2} \pi_i p_{ij}}^{\text{Flux de } E_1 \text{ vers } E_2} = \overbrace{\sum_{i \in E_2} \sum_{j \in E_1} \pi_i p_{ij}}^{\text{Flux de } E_2 \text{ vers } E_1}$$



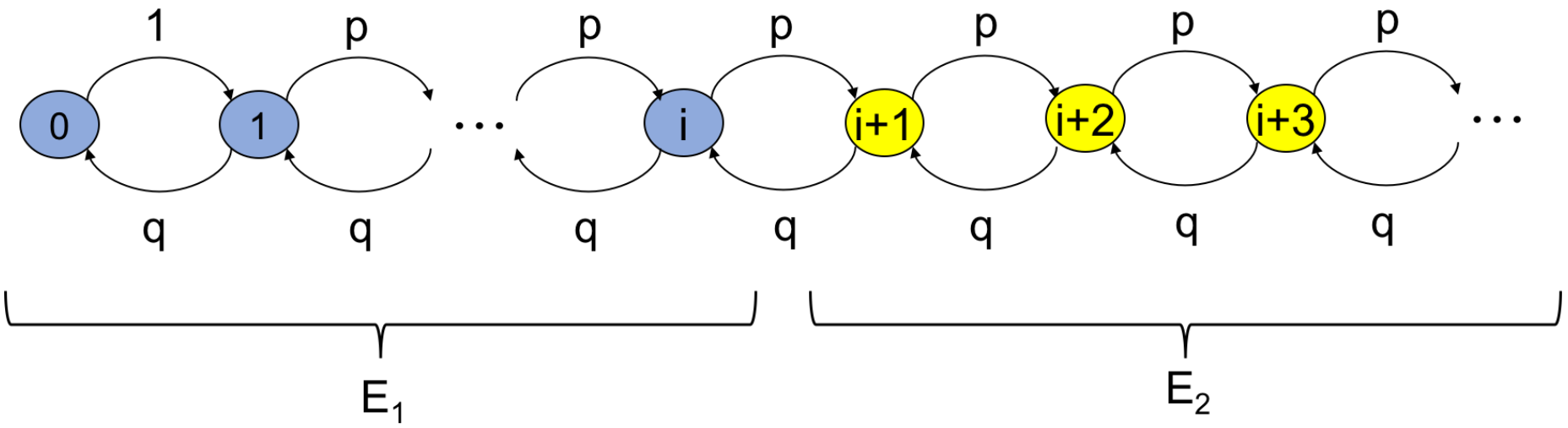
Exemple 1



- Flux de 0 vers 1 = Flux de 1 vers 0
- Flux de 1 vers 2 = Flux de 2 vers 1

$$\begin{cases} \frac{2}{3}\pi_0 = \frac{1}{2}\pi_1 \\ \frac{1}{2}\pi_1 = \frac{1}{2}\pi_2 \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{cases} \Rightarrow \begin{cases} \pi_1 = \frac{4}{3}\pi_0 \\ \pi_2 = \pi_1 = \frac{4}{3}\pi_0 \\ \pi_0(1 + \frac{4}{3} + \frac{4}{3}) = 1 \end{cases} \Rightarrow \begin{cases} \pi_0 = \frac{3}{11} \\ \pi_1 = \frac{4}{11} \\ \pi_2 = \frac{4}{11} \end{cases}$$

Exemple 2



- Flux de E_1 vers E_2 = Flux de E_2 vers E_1 $\Rightarrow$ $p\pi_i = q\pi_{i+1}$
- Soit $\rho = \frac{p}{q}$. Alors $\pi_i = \rho^i \pi_0$.
- Si $\rho < 1$, alors $\sum_{i=0}^{\infty} \pi_i = 1 \Rightarrow \pi_0 = 1 - \rho$

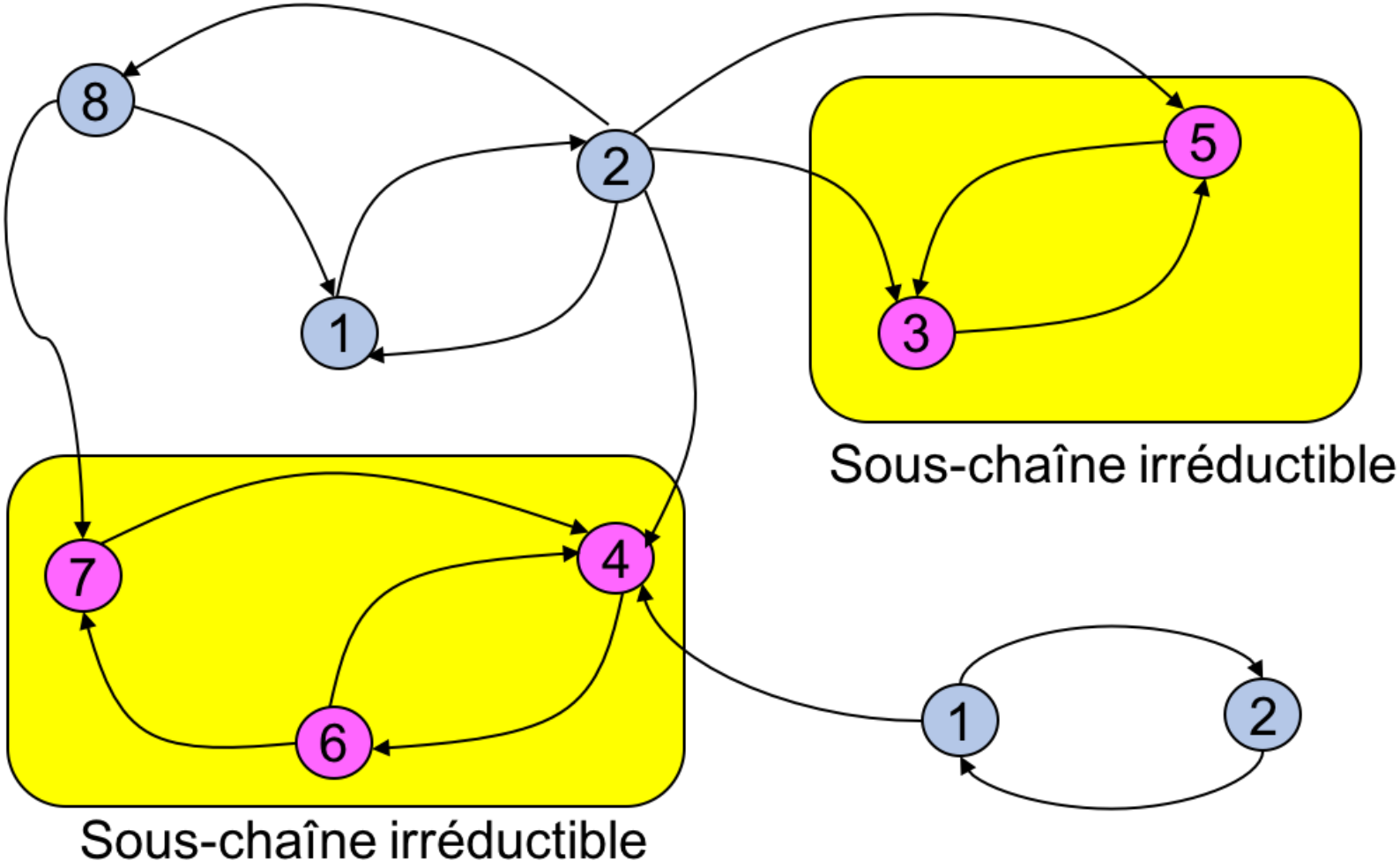
$$\pi_i = (1 - \rho)\rho^i$$

Travail à la maison

- Finir lecture du chapitre 3.1 (pages 55 à 73)
- Exercices 6 et 7
- TP sur les marches aléatoires (évaluation lors de l'examen terminal)

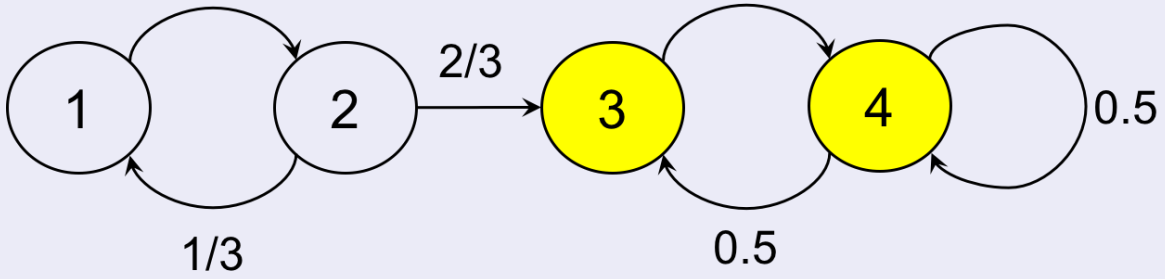
Partition d'une CMTD

- Une CMTD peut être partitionnée en sous-chaînes irréductibles, plus un ensemble d'états transitoires
  - E_T : Ensemble d'états transitoires
  - E_i : Ensemble d'états de la i-ème sous-chaîne irréductible
- Sous-chaîne irréductible = Sous-chaîne absorbante



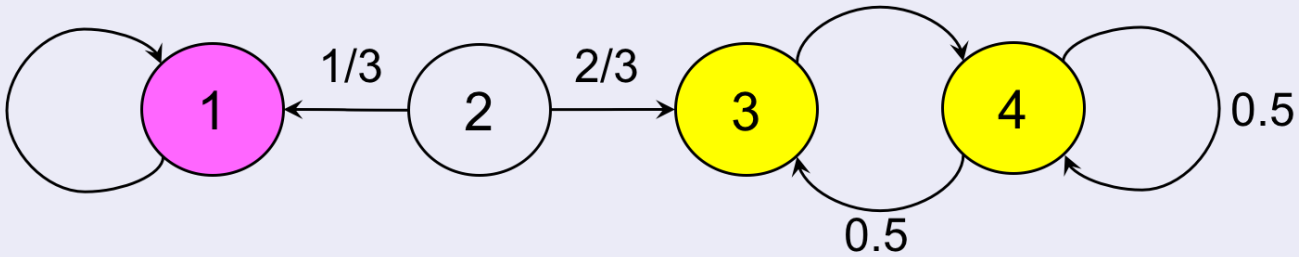
- Peut-on encore appliquer le théorème valable pour une chaîne irréductible ?
- Probabilité d'absorption par une sous-chaîne ?
- Temps moyen avant absorption par une sous-chaîne ?

Uni-chaîne : Une seule sous-chaîne irréductible



- Les résultats pour une CMTD irréductible s'appliquent encore.

Multi-chaîne : Plusieurs sous-chaîne irréductibles



- La distribution limite dépend de l'état initial
- La solution de (S) n'est pas unique

Forme canonique de la matrice de transition (facultatif)

- P_T : transitions entre les états transitoires E_T
- P_i : transitions entre les états récurrents E_i de la sous-chaîne irreducible i
- R_i : transitions des états transitoires E_T vers les états E_i de la sous-chaîne irréductible i

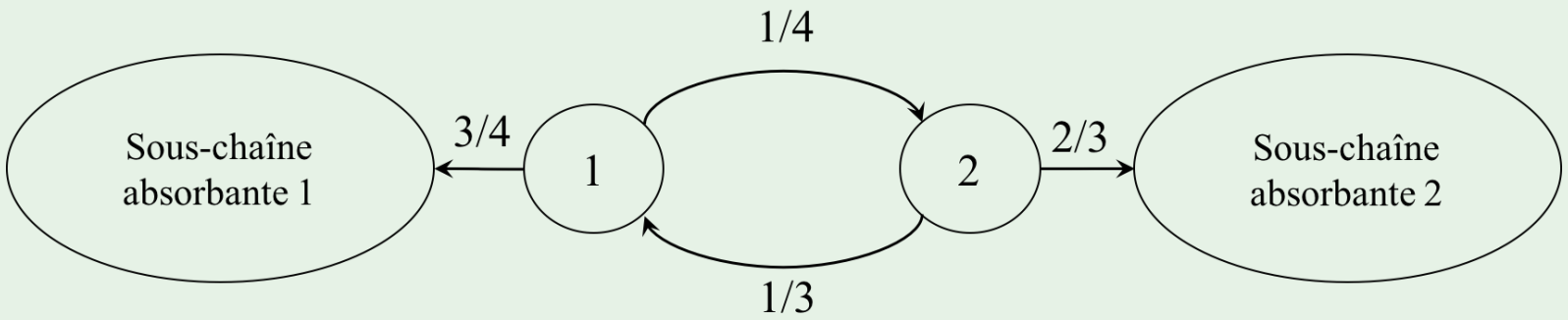
$$P = \begin{pmatrix} P_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & P_k & 0 \\ R_1 & \dots & R_k & P_T \end{pmatrix}$$

Temps moyen avant absorption par une sous-chaîne

- e_i : Temps moyen avant absorption par une sous-chaîne irréductible, en partant de l'état transitoire i

$$e_i = 1 + \sum_{j \in E_T} p_{ij} e_j$$

Exemple



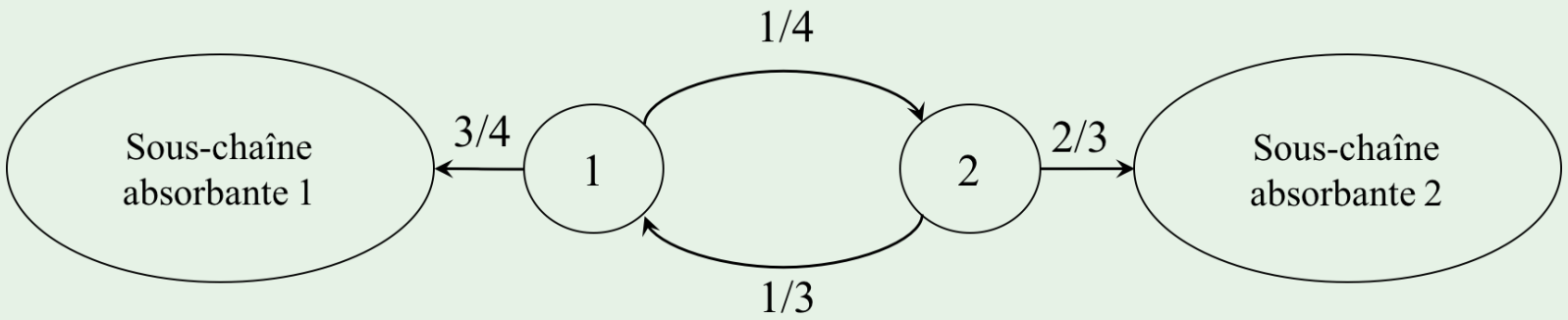
$$\begin{cases} e_1 = 1 + 1/4 e_2 \\ e_2 = 1 + 1/3 e_1 \end{cases} \Rightarrow \begin{cases} e_1 = 15/11 \\ e_2 = 16/11 \end{cases}$$

Probabilité d'absorption par une sous-chaîne

- a_{ij} : probabilité, en partant de l'état transitoire i, d'être absorbé par la sous-chaîne j

$$a_{ij} = \sum_{k \in E_j} p_{ik} + \sum_{k \in E_T} p_{ik} a_{kj}$$

Exemple



$$\begin{cases} a_{11} = 3/4 + 1/4 a_{21} \\ a_{21} = 1/3 a_{11} \end{cases} \Rightarrow \begin{cases} a_{12} = 9/11 \\ a_{21} = 3/11 \end{cases}$$

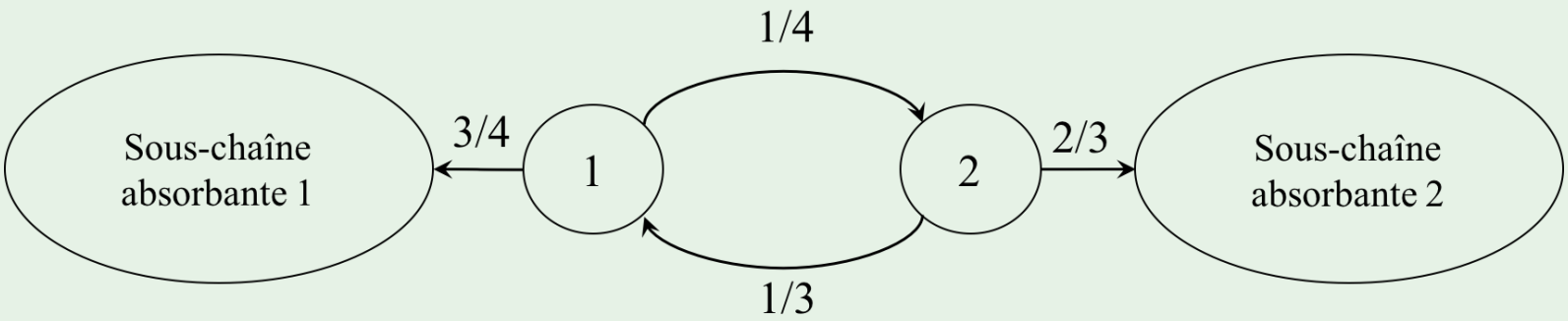
Temps moyen passé dans un état transitoire

- s_{ij} : Temps moyen passé dans j, en partant de $i \in \mathcal{S}_T$

$$s_{ij} = \delta_{ij} + \sum_{k \in \mathcal{S}_T} p_{ik} s_{kj}$$

avec $\delta_{ij} = 1$ si $i = j$ and 0 autrement

Exemple



$$\begin{cases} s_{11} = 1 + 1/4 s_{21} \\ s_{21} = 1/3 s_{11} \end{cases} \Rightarrow \begin{cases} s_{11} = 12/11 \\ s_{21} = 4/11 \end{cases}$$

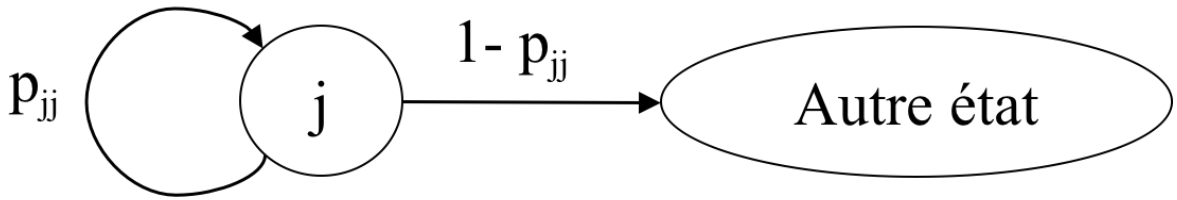
- Vérifier que $e_1 = s_{11} + s_{12}$ et $e_2 = s_{21} + s_{22}$

Exercice 8 : roulette

Roulette française versus roulette américaine

Temps de séjour dans un état (voir exercice 2)

- T_j : nombres d'étapes passées dans l'état j avant d'en sortir
- Supposons que $p_{jj} < 1$ (autrement $T_j = +\infty$)



- T_j = nombre d'expérience de Bernoulli pour obtenir un succès, avec proba de succès $(1 - p_{jj})$
- T_j distribué suivant une loi géométrique de paramètre $(1 - p_{jj})$

$$P(T_j = k) = (1 - p_{jj})p_{jj}^{k-1} \quad \text{pour } k = 1, 2, \cdots$$

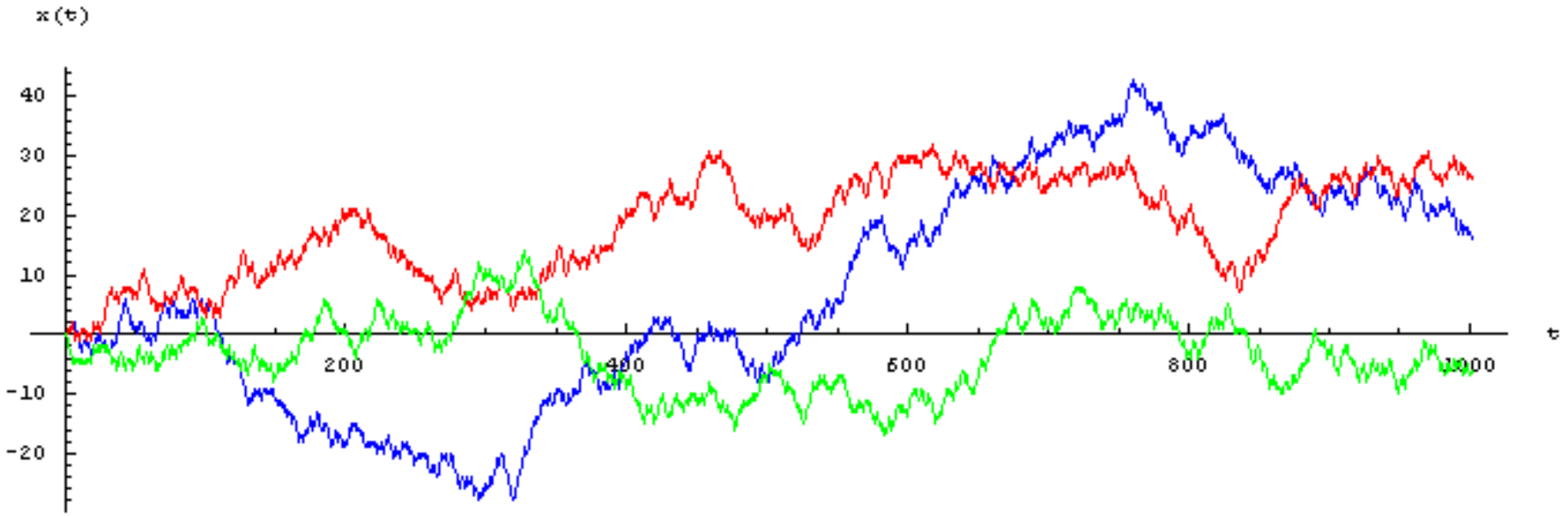
Propriété sans mémoire (ou propriété de Markov)

- Une variable aléatoire T est dite sans mémoire si

$$\forall t, x, \quad P(T > t + x | T > t) = P(T > x)$$

- La loi géométrique est sans mémoire (voir exercice 10)
- Le temps passé dans un état d'une CTMD suit donc une loi sans mémoire

Chaîne de Markov ergodique



- Hypothèse ergodique : les performances stationnaires sont égales à la performance moyenne de n'importe quelle trajectoire $(X_0, X_1, X_2, \cdots)$

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \underset{N\to+\infty}{\longrightarrow} \sum_{i\in E} f(i)\pi_i$$

- Propriété : Une CMTD irréductible (ou uni-chaîne) est ergodique
- Implication : on peut simuler le comportement d'un système ergodique avec une seule trajectoire

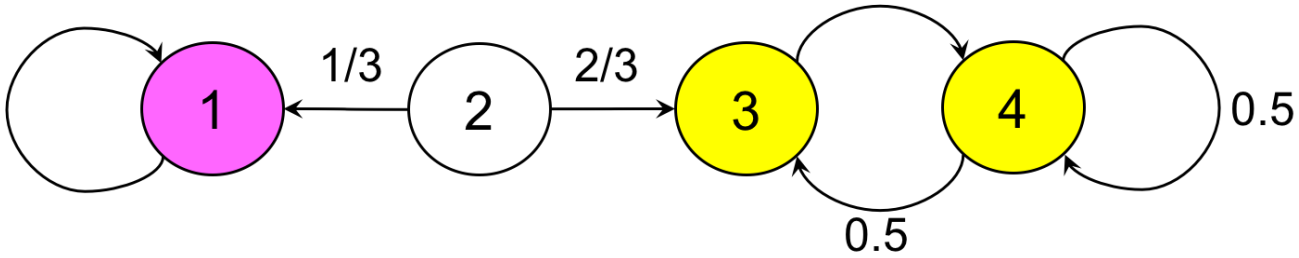
Pourcentage de temps passé dans un état

- Prenons $f(x) = 1$ si $x = j$, 0 sinon
- Alors

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N} f(X_n = j)}_{\text{Proportion de temps passé en } j \text{ sur la trajectoire } X_1, X_2, \cdots, X_N} \underset{N\to\infty}{\longrightarrow} \pi_j$$

- Conclusion : pour une CMTD ergodique, le % de temps passé dans l'état j pour une trajectoire vaut π_j, quelle que soit la trajectoire

Un exemple de chaîne non ergodique



- Si l'on part de l'état 2, il faut plusieurs simulations (en repartant de l'état 2) pour étudier les performances (par exemple estimer la probabilité d'absorption par 1)

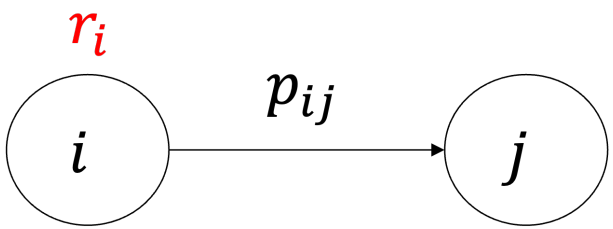
Travail à la maison

- Programme de l'évaluation bonus
  - Exercices 1 à 8
  - Exercice "Pile ou face jusqu'à la ruine" (cas particulier de l'exercice 8)
  - Chapitre 3.1
- Savoir (entre autres) :
  - Donner le graphe et la matrice de transition
  - Classer les états
  - Calculer la distribution en régime transitoire $\pi(n)$
  - Calculer la distribution limite π
  - Interpréter la distribution limite / stationnaire
  - Calculer le temps moyen avant absorption par une sous-chaîne irréductible
  - Calculer la probabilité d'absorption par une sous-chaîne irréductible

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Chaîne de Markov avec récompenses

- $\mathcal{S}$: espace d'états
- p_{ij} : probabilité d'aller d'un état i à un état j
- r_i : récompense (en espérance) à chaque visite de l'état i
- γ : Taux d'actualisation



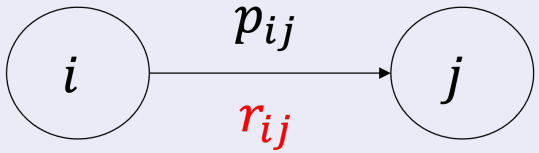
Exemple

La récompense obtenue à chaque visite de l'état i est une variable aléatoire d'espérance μ_i et d'écart-type σ_i.

$$r_i = \mu_i$$

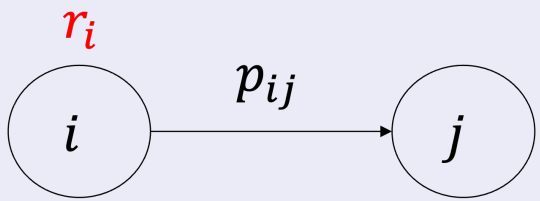
Récompenses sur les arcs / sommets

Récompenses sur les arcs



- r_{ij} : récompense si une transition de i à j se produit

Récompenses sur les sommets



- r_i : récompense à chaque visite de l'état i

- On peut passer des r_{ij} aux r_i sans modifier la récompense totale

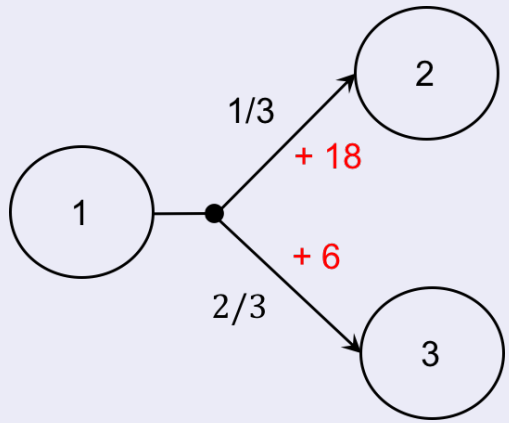
$$r_i = \sum_{j \in \mathcal{S}} p_{ij} r_{ij}$$

- Dans la suite, nous écrirons les équations avec les récompenses sur les sommets

Représentation graphique

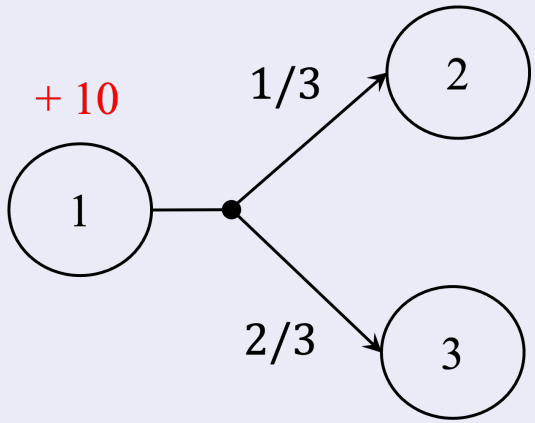
- Récompenses et coûts en rouge :
  - + R pour une Récompense
  - - C pour un Coût

Récompenses sur les arcs



$\Rightarrow$

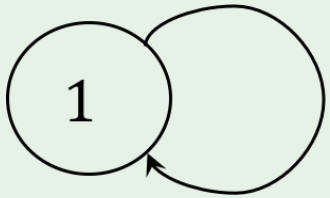
Récompenses sur les sommets



Différentes fonctions objectif (horizon infini)

- Récompense totale $= \mathbb{E}(R_0 + R_1 + R_2 + R_3 + \cdots)$
- $\gamma \in]0, 1[$: taux d'actualisation
- Récompense totale actualisée $= \mathbb{E}(R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots)$
- Gain moyen $= \lim_{T \to \infty} \mathbb{E}\left(\frac{R_0 + R_1 + \cdots + R_{T-1}}{T}\right)$

Exemple



$$\text{Récompense totale} = 5 + 5 + 5 + \cdots = +\infty$$

$$\text{Récompense totale actualisé} = 5 + 5\gamma + 5\gamma^2 + 5\gamma^3 + \cdots = \frac{5}{1-\gamma}$$

$$\text{Gain moyen (par periode)} = 5$$

Fonctions de valeur

Récompense totale

$$v(i) = \mathbb{E}\left[\sum_{t=0}^{\infty} R_t \,|\, S_0 = i\right]$$

Récompense totale actualisée ($\gamma < 1$)

$$v(i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,|\, S_0 = i\right]$$

Gain moyen

$$g(i) = \lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=0}^{T} R_t \,|\, S_0 = i\right]$$

Gain à partir de t

Definition

Le gain G_t est la récompense total actualisée à partir de t.

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$$

$$v(i) = \mathbb{E}\left[G_t \middle| S_t = i\right]$$

Récompense totale - Équations de Bellman

Définition

Un état terminal s est un état absorbant ($p_{ss} = 1$) avec récompense nulle ($r_s = 0$).

Hypothèse

Il existe un état terminal s atteint avec probabilité 1.

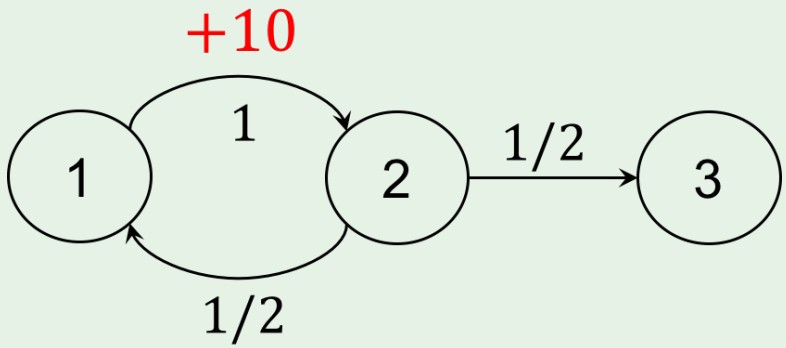
- La fonction de valeur est solution des équations de Bellman

$$v(s) = 0$$
$$v(i) = r_i + \sum_{j \in \mathcal{S}} p_{ij} v(j) \quad \text{pour } i \neq s$$

- Version avec les r_{ij}

$$v(i) = \sum_{j \in \mathcal{S}} p_{ij}(r_{ij} + v(j))$$

Exemple 1

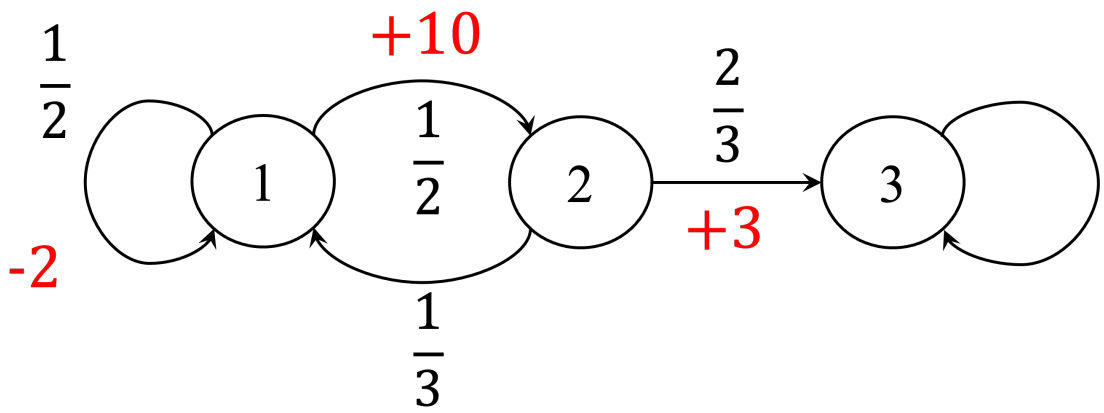


$$\begin{cases} v(1) = 10 + v(2) \\ v(2) = \frac{1}{2}v(1) + \frac{1}{2}v(3) \\ v(3) = 0 \end{cases} \Rightarrow \begin{cases} v(1) = 20 \\ v(2) = 10 \\ v(3) = 0 \end{cases}$$

Exemple 2 : Roulette (exercice 8)

Écrire un système linéaire permettant d'obtenir $v(i)$ en partant d'un état i.

Travail pour la prochaine fois



- Obtenir v par deux méthodes :
  - En résolvant un système linéaire
  - En implémentant l'algorithme d'itération sur la valeur
- Tester votre algorithme avec différents vecteurs initiaux v_0 et différents critères d'arrêt ϵ
- Devrait donner $v = \begin{bmatrix} 15 & 7 & 0 \end{bmatrix}^\top$
- Roulette (exercice 8) : résoudre les équations de Bellman de manière exacte et par itération sur la valeur. Vérifier que la récompense obtenue est cohérent avec les résultats de l'exercice 8.

Récompense actualisé

- Soit $r_{max} = \max_{i \in \mathcal{S}} |r_i|$
- La fonction de valeur est alors bornée

$$|v(i)| \leq \sum_{t=0}^{\infty} \gamma^t r_{max} = \frac{r_{max}}{1-\gamma}$$

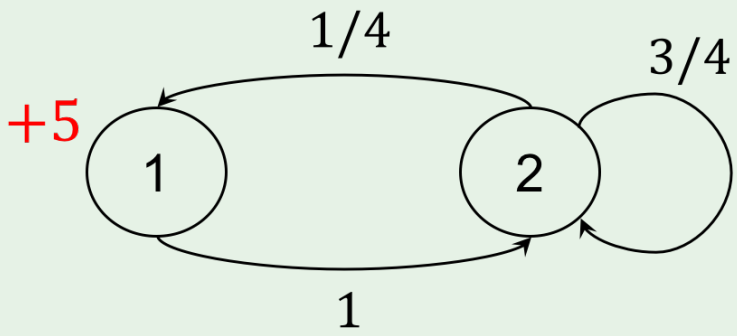
- La fonction de valeur est alors solution des équations de Bellman suivantes

$$\forall i \in \mathcal{S}, \quad v(i) = r_i + \gamma \sum_{j \in \mathcal{S}} p_{ij} v(j)$$

Pourquoi actualiser ?

- Si les récompenses sont financières, les récompenses immédiates ont plus de valeur
- Les êtres vivants préfèrent les récompenses immédiates
- Évite de cycler
- Lorsqu'il n'y a pas d'état terminal

Exemple avec $\gamma = 0.8$



$$\begin{cases} v(1) = 5 + \gamma \times v(2) \\ v(2) = \gamma \times \left[\frac{1}{4} v(1) + \frac{3}{4} v(2) \right] \end{cases} \Rightarrow \begin{cases} v(1) = \frac{25}{3} \\ v(2) = \frac{25}{6} \end{cases}$$

Version matricielle des équations de Bellman

- Notons $1, \cdots, n$ les états non terminaux
- $P = (p_{ij})$: matrice de transition excluant les états terminaux
- Cette matrice n'est pas stochastique

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} + \gamma \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

$$v = r + \gamma P v$$

$$v = (I - \gamma P)^{-1} r$$

Algorithme d'itération sur la valeur

- Distance entre deux fonctions de valeur u and w avec la norme ∞

$$\|u - w\| = \max_{i \in \mathcal{S}} |u(i) - w(i)|$$

Algorithme d'itération sur la valeur

- v_0 arbitraire ($v_0 \in \mathbb{R}^n$)
- $v_{k+1} = r + \gamma P v_k$
- Critère d'arrêt : $\|v_{k+1} - v_k\|_\infty < \epsilon$

- Si $\gamma < 1$ ou si il existe un état terminal, alors la suite (v_k) converge vers l'unique solution de

$$v = r + \gamma P v$$

- A quelle distance est-on de v pour un ϵ donné ? $\|v_k - v\|_\infty| \leq ?$

Preuve de la convergence si $\gamma < 1$

- Opérateur $L(v) = r + \gamma P v$
- Les équations de Bellman s'écrivent

$$v = L(v)$$

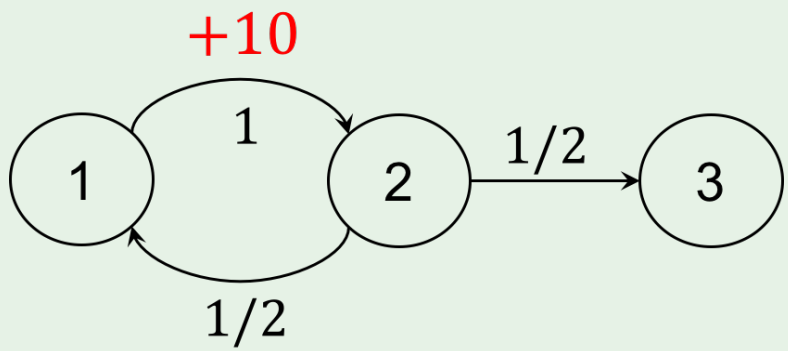
- L'opérator L est γ-contractant

$$\|L(u) - L(w)\| \leq \gamma \|u - w\|$$

- $L^k(u)$: applique k fois l'opérateur L à u
- Par le théorème du point fixe de Banach, la suite (v_k) converge vers l'unique solution de $v = L(v)$ avec une vitesse de convergence linéaire γ

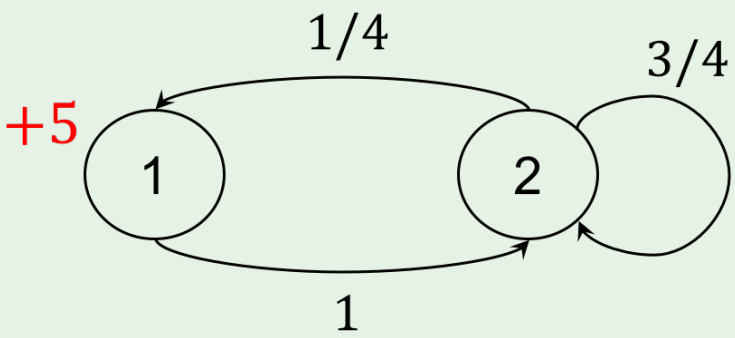
$$\|v_k - v\| = \|L^k(v_0) - L^k(v)\| \leq \gamma^k \|v_0 - v\|$$

Exemple with $\gamma = 1$ and $v_0 = 0$



k	$v_k(1)$	$v_k(2)$	$v_k(3)$
0	0	0	0
1	10	0	0
2	10	5	0
3	15	5	0
4	15	7.5	0
...	...	...	...
10	19.375	9.688	0
...	...	...	...
50	20.000	10.000	0

Exemple with $\gamma = 0.8$ and $v_0 = 0$



k	$v_k(1)$	$v_k(2)$
0	0	0
1	5	0
2	5	1
3	5.8	1.6
4	6.28	2.12
...	...	...
10	7.796	3.630
...	...	...
50	8.333	4.167
...	...	...
100	8.333	4.167

Itération sur la valeur "In-place"

- L'itération sur la valeur Synchrone stocke deux copies de la fonction de valeur

$$\forall i \in \mathcal{S}, \qquad v_{new}(i) \leftarrow r_i + \sum_{j \in \mathcal{S}} p_{ij} v_{old}(j)$$

- La version In-place stocke une seule copie

$$\forall i \in \mathcal{S}, \qquad v(i) \leftarrow r_i + \sum_{j \in \mathcal{S}} p_{ij} v(j)$$

- La version in-place est plus efficace car tient compte de la dernière mise à jour
- Son efficacité dépend de l'ordre de parcours des états

Lien entre récompense non actualisée (avec état terminal) et récompense actualisée

- Récompense totale actualisée avec probabilité s de transition (p_{ij})

$$v(i) = r_i + \gamma \sum_{j \in \mathcal{S}} p_{ij} v(j)$$

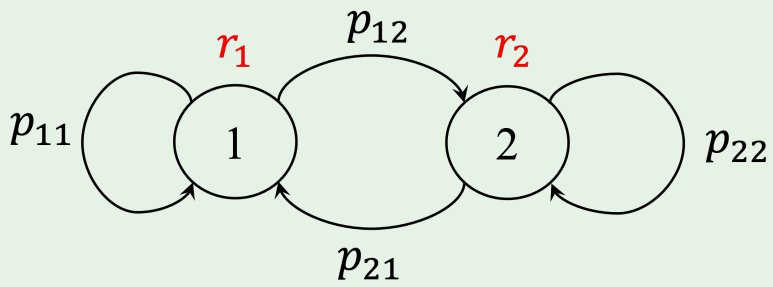
- Récompense totale avec espace d'état $\mathcal{S} \cup t$ avec t l'état terminal et les probabilités de transition suivantes $p'_{ij} = \gamma p_{ij}, \forall i, j \in \mathcal{S}$ and $p'_{it} = 1 - \gamma, \forall i \in \mathcal{S}$

$$\begin{aligned} v(i) &= r_i + \sum_{j \in \mathcal{S} \cup t} p'_{ij} v(j) \\ &= r_i + \gamma \sum_{j \in \mathcal{S}} p_{ij} v(j) \end{aligned}$$

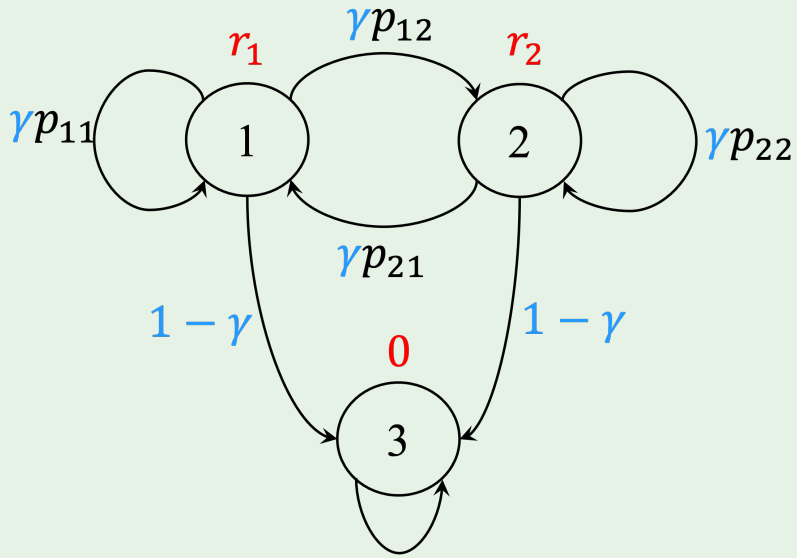
- Les équations étant les mêmes, la fonction de valeur est identique
- Récompense totale actualisé = cas particulier de la récompense totale avec état terminal

Exemple

Récompense totale actualisée avec $\gamma \in]0, 1[$



Récompense totale (non actualisée) équivalente, avec un état terminal



Gain moyen

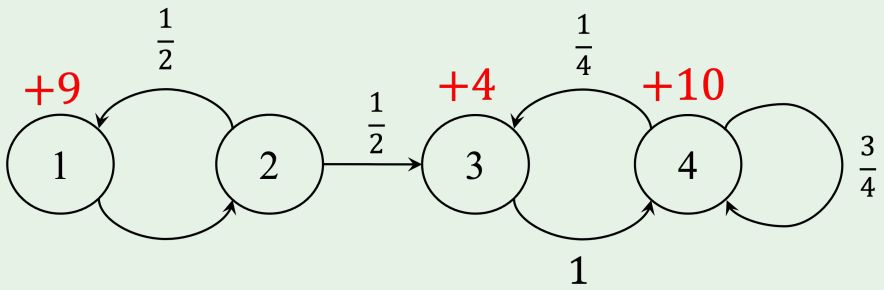
- g : gain moyen

$$g = \lim_{T \to \infty} E\left[\frac{1}{T}\sum_{t=0}^{T} R_t\right]$$

- Le gain moyen est indépendant de l'état initial si la chaîne de Markov est irréductible (ou uni-chaîne)

$$g = \sum_{i \in \mathcal{S}} \pi_i r_i$$

Exemple



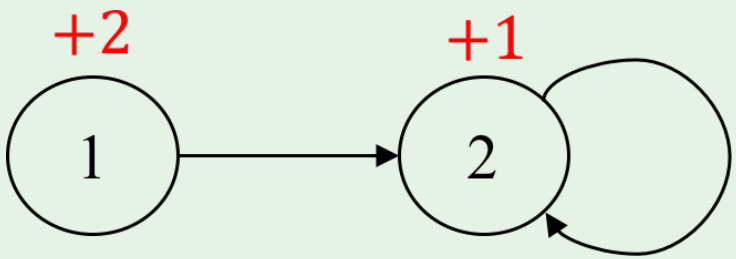
$$\begin{aligned} g &= \pi_1 r_1 + \pi_2 r_2 \\ &= 0.2 \times 5 + 0.8 \times 10 \\ &= 9 \end{aligned}$$

Lien entre gain moyen et récompense actualisée

- Hypothèse : chaîne de Markov uni-chain

$$\forall i, \quad g = \lim_{\gamma \to 1^-} (1-\gamma) v(i)$$

Exemple



$$\begin{aligned} g &= 1 \\ v(2) &= 1 + \gamma + \gamma^2 + \cdots = \frac{1}{1-\gamma} \\ v(1) &= 2 + \gamma v(2) = 2 + \frac{\gamma}{1-\gamma} \end{aligned}$$

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Récompenses et distributions inconnues

- Probabilités p_{ij} inconnues
- Récompenses r_{ij} inconnues
- Épisode $= S_0, R_0, S_1, R_1, \cdots, S_T, R_T$
- On peut simuler un grand nombre d'épisodes

Loi des grands nombres

- X : variable aléatoire d'espérance μ et d'écart-tupe σ (ex : la récompense obtenue lors d'un épisode)
- $X_1, X_2, \cdots, X_n$: variables aléatoires i.i.d. de mêmes distribution que X
- Moyenne empirique

$$\mu_n = \frac{X_1 + \cdots + X_n}{n}$$

- Elle converge vers $\mu = E(X)$

$$\mu_n \underset{n \to \infty}{\longrightarrow} \mu$$

Moyenne incrémentale et taux d'apprentissage

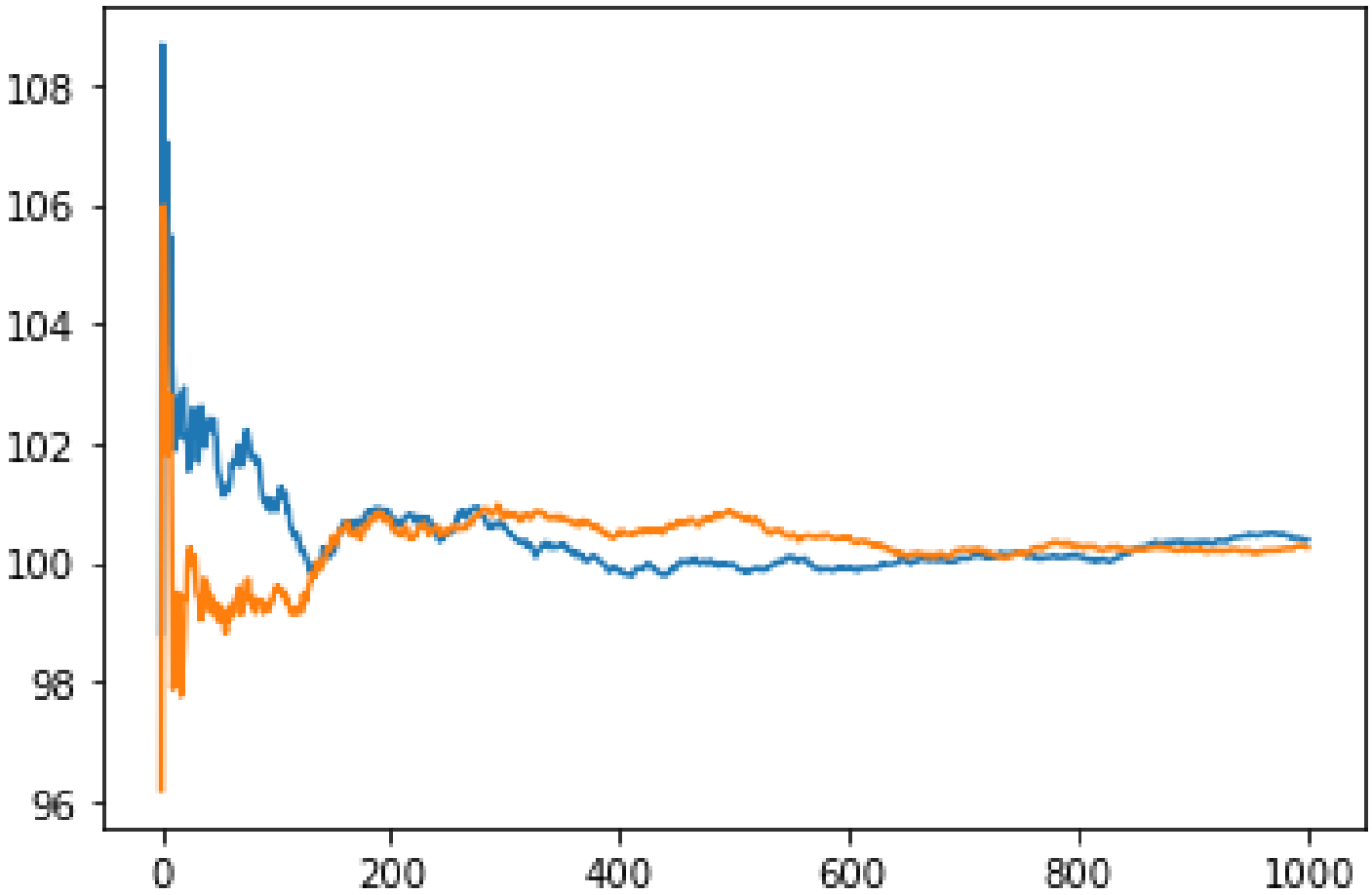
- μ_k : moyenne de $(X_1, \cdots, X_k)$
- $\mu_0 = 0$

$$\underbrace{\mu_k}_{\text{Nouvelle estimation}} = \left(1 - \frac{1}{k}\right) \cdot \underbrace{\mu_{k-1}}_{\text{Ancienne estimation}} + \underbrace{\frac{1}{k}}_{\text{Taux apprentissage}} \cdot \underbrace{X_k}_{\text{Nouvelle valeur}}$$

$$\mu_k = \mu_{k-1} + \frac{1}{k}(X_k - \mu_{k-1})$$

Exemple

- X est distribuée suivant une loi normale ($\mu = 100$ and $\sigma = 10$)
- On trace μ_k pour $k = 1, \cdots, 1000$ pour deux épisodes



Autres taux d'apprentissage

$$\underbrace{\mu_k}_{\text{Nouvelle estimation}} = (1-\alpha_k)\cdot \underbrace{\mu_{k-1}}_{\text{Ancienne estimation}} + \underbrace{\alpha_k}_{\text{Taux d'apprentissage}} \cdot \underbrace{X_k}_{\text{Ancienne valeur}}$$

$$= \mu_{k-1} + \alpha_k(X_k - \mu_{k-1})$$

- Si $\alpha_k = \frac{1}{k}$, on obtient la moyenne empirique
- Condition de convergence vers μ (Robbins-Monro) :
  - $\sum_k \alpha_k = +\infty$
  - $\sum_k \alpha_k^2 < \infty$
- If $\alpha_k = \alpha$
  - Ne converge pas vers μ
  - Préférable pour un environnement non stationnaire

Travail à la maison

- X distribuée suivant une loi normale ($\mu = 100$ et $\sigma = 20$)
- Tracer μ_k pour $k = 1$ à $k = n$
- Tester plusieurs taux d'apprentissage
  - $\alpha_k = \frac{1}{k}$
  - $\alpha_k = \alpha$ (faire varier α dans $[0, 1]$)
  - $\alpha_k = \frac{1}{k^a}$ (faire varier a)
- Faire varier n (e.g. $n = 10^2, 10^3, 10^4, \cdots$)
- Faire varier μ_0 (e.g. $\mu_0 = 0, 50, 100, 150$)
- Commenter

Simulation de Monte-Carlo

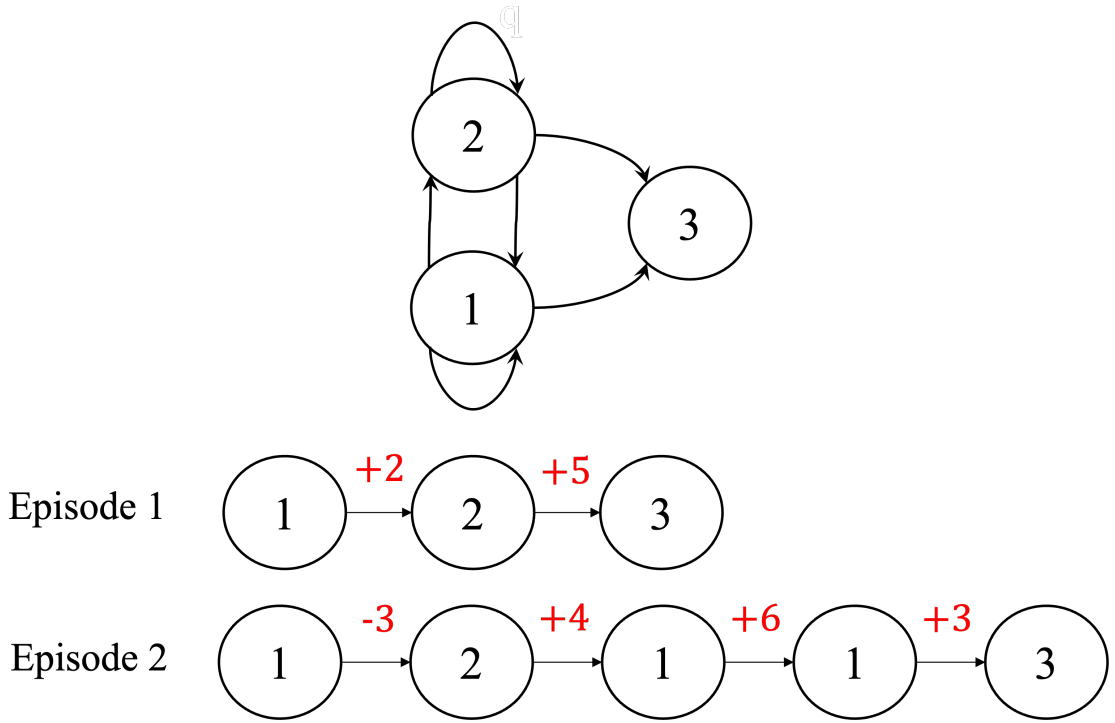
- Hypothèse : Tous les épisodes terminent
- Épisode : $S_0, R_0, S_1, R_1, \cdots$
- Récompense à partir de t :

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$$

$$v(s) = E(G_t | S_t = s)$$

- $V(s)$: estimation de $v(s)$

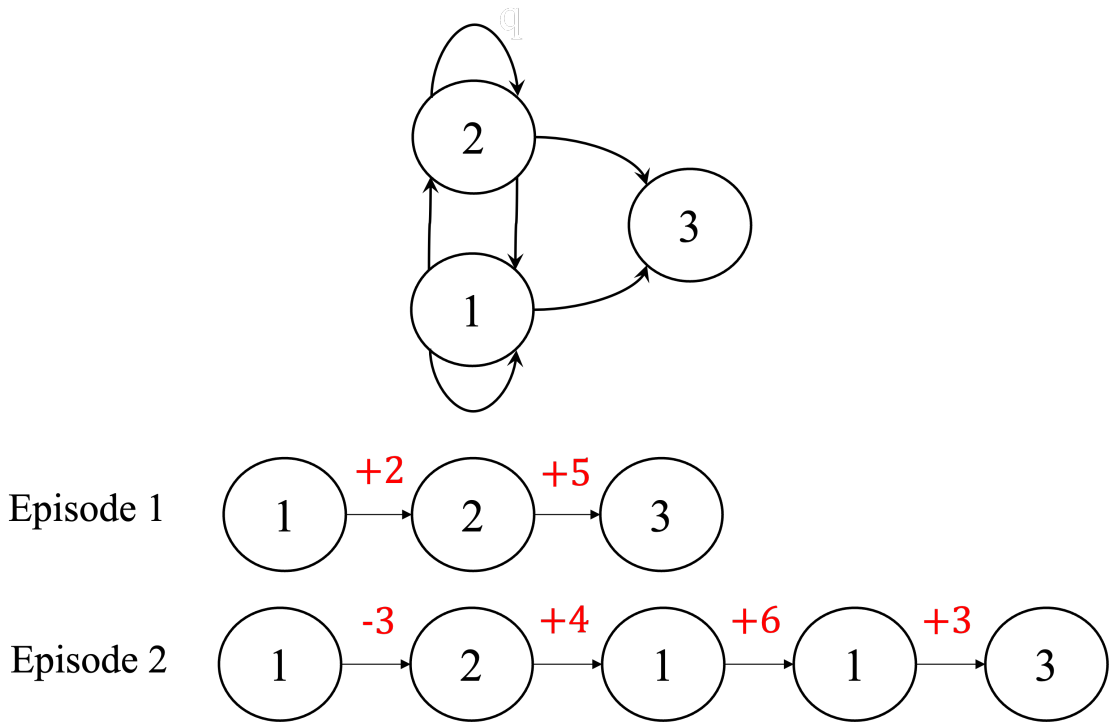
Monte-Carlo First-Visit



$$V(1) = \frac{1}{2}\left[(2+5\gamma) + (-3+4\gamma+6\gamma^2+3\gamma^3)\right]$$

$$V(2) = \frac{1}{2}\left[5 + (4+6\gamma+3\gamma^2)\right]$$

Monte-Carlo Every Visit



$$V(1) = \frac{1}{4}\left[(2+5\gamma) + (-3+4\gamma+6\gamma^2+3\gamma^3) + (6+3\gamma) + 3\right]$$

Monte-Carlo Every-Visit : algorithme

- $V(s)$: estimation de $v(s)$
- $Gains(s)$: liste vide

Boucle (pour chaque épisode)

- Générer un épisode $S_0, R_0, S_1, R_1, \cdots, S_T$
- $G \leftarrow 0$
- Pour $t = T-1, T-2, \cdots, 0$
  1. $G \leftarrow G + \gamma R_t$
  2. Ajouter G à $Gains(S_t)$
  3. $V(S_t) \leftarrow$ moyenne$(Gains(S_t))$

- V converge vers v (loi des grands nombres)

Monte-Carlo Every-Visit : version incrémentale

- $V(s)$: estimation de $v(s)$, initialisée à 0 par exemple
- $Gains(s)$: liste vide
- $N(s) \leftarrow 0$: compteur du nombre de visites

Boucle (pour chaque épisode)

- Générer un épisode $S_0, R_0, S_1, R_1, \cdots, S_T$
- $G \leftarrow 0$
- Pour $t = T-1, T-2, \cdots, 0$
  1. $G \leftarrow G + \gamma R_t$
  2. $N(S_t) \leftarrow N(S_t) + 1$
  3. $V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}(G - V(S_t))$

- Pour des problèmes non stationnaires, il est utile d'oublier les anciens épisodes

$$V(S_t) \leftarrow V(S_t) + \alpha(G - V(S_t))$$

Temporal-Difference learning : TD(0)

- $V(s)$: estimation de $v(s)$
- Initialiser par exemple avec $V(s) = 0, \forall s$ (doit être égal à 0 pour un état terminal)

Boucle (pour chaque épisode)

- Intialiser l'état S
- Tant que S n'est pas terminal
  1. Observer R et S' (récompense et état suivant)
  2. Mettre à jour V

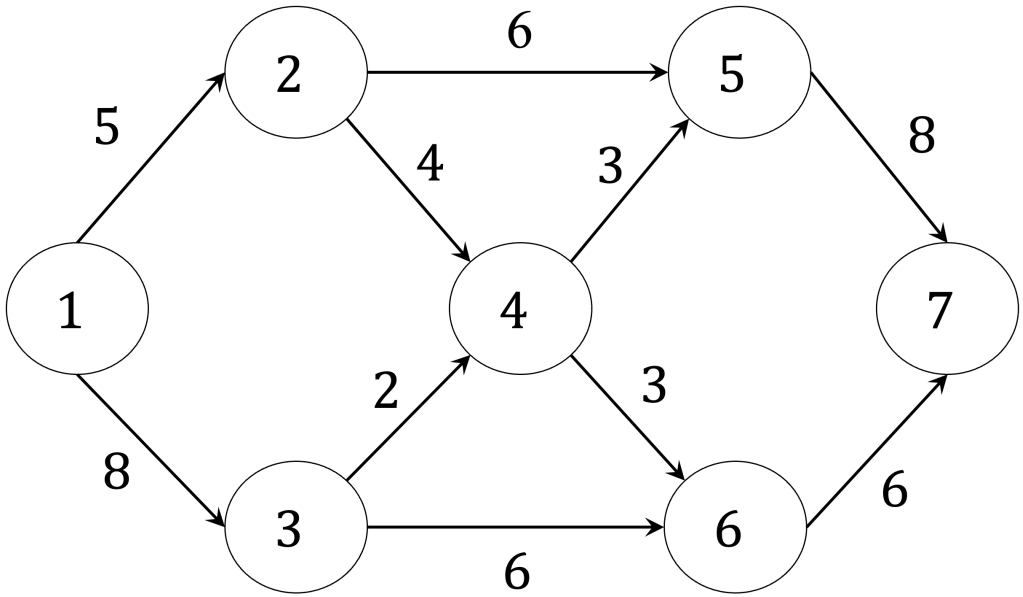
$$\underbrace{V(S)}_{\text{Nouvelle estimation}} \leftarrow (1-\alpha) \underbrace{V(S)}_{\text{Ancienne estimation}} + \underbrace{\alpha}_{\substack{\text{Taux} \\ \text{d'apprentissage}}} [R + \gamma \underbrace{V(S')}_{\substack{\text{Estimation} \\ \text{du gain futur}}}]$$

  3. $S \leftarrow S'$

- V converge vers v sous les conditions de Robbins-Monro

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Plus court chemin déterministe

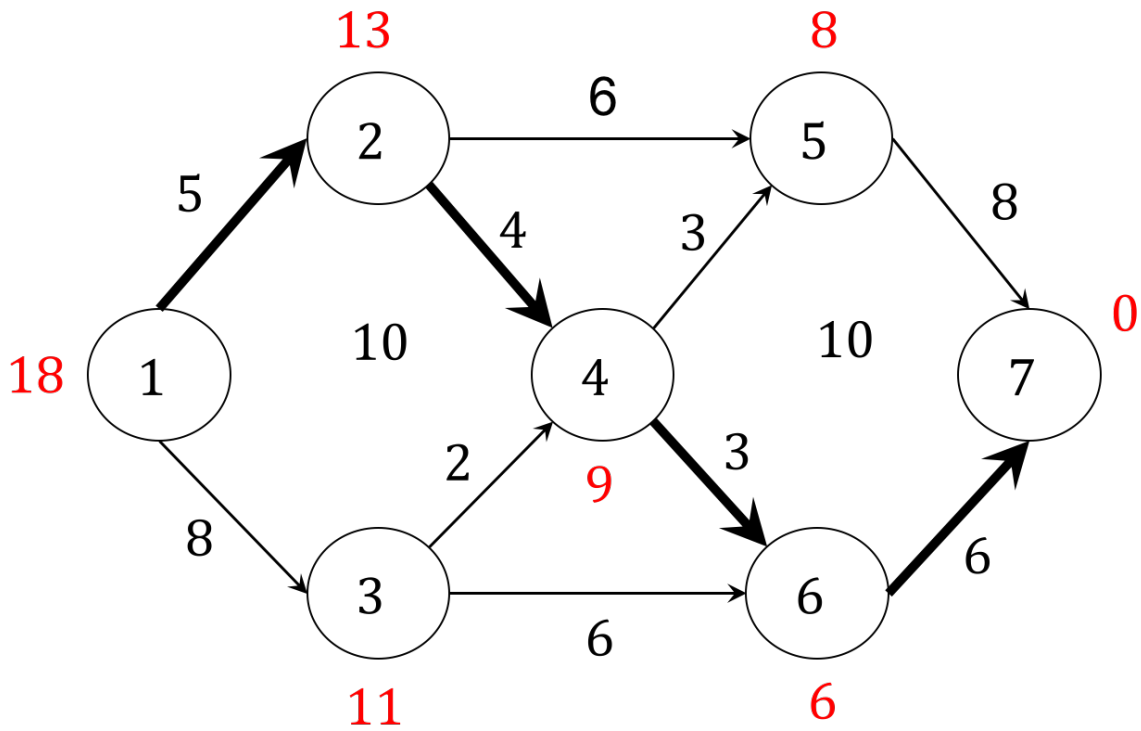


- d_{ij} : distance du sommet i au sommet j ($d_{12} = 5$)
- s : source ($s = 1$)
- t : destination ($t = 7$)
- Objectif : trouver le plus court chemin de s to t

Fonction de valeur

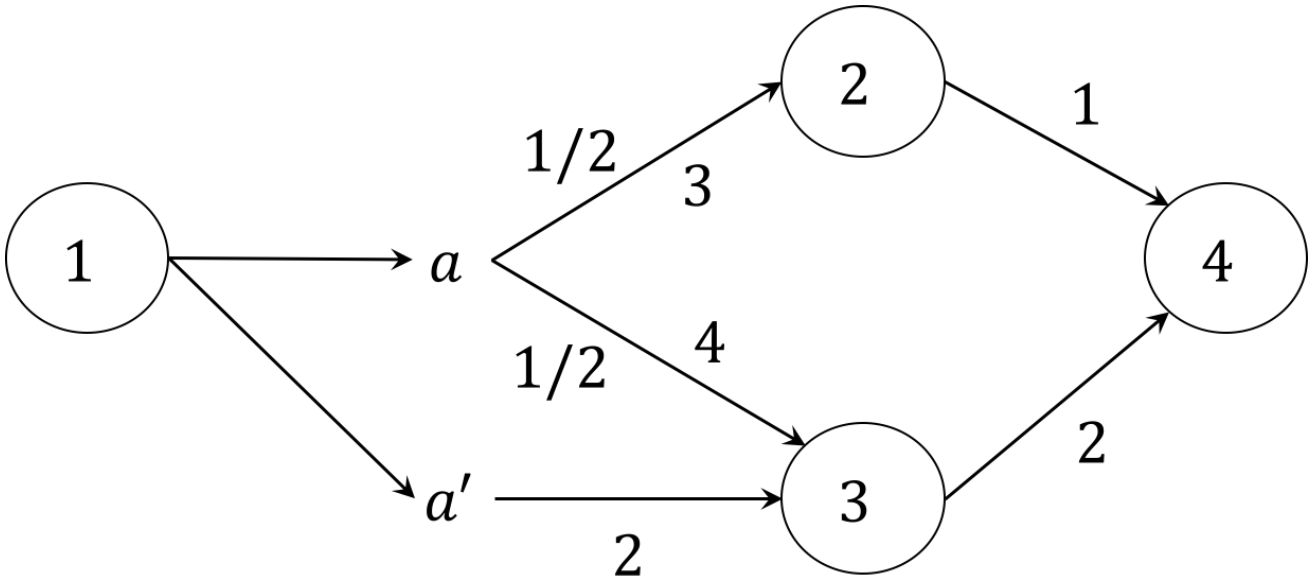
- $v^*(i)$: distance minimale entre le sommet i et la destination
- $Succ(i)$: ensemble des successeurs de i ($Suc(1) = \{2, 3\}$)
- Équations d'optimalité

$$v^*(i) = \min_{j \in Suc(i)} d_{ij} + v^*(j)$$



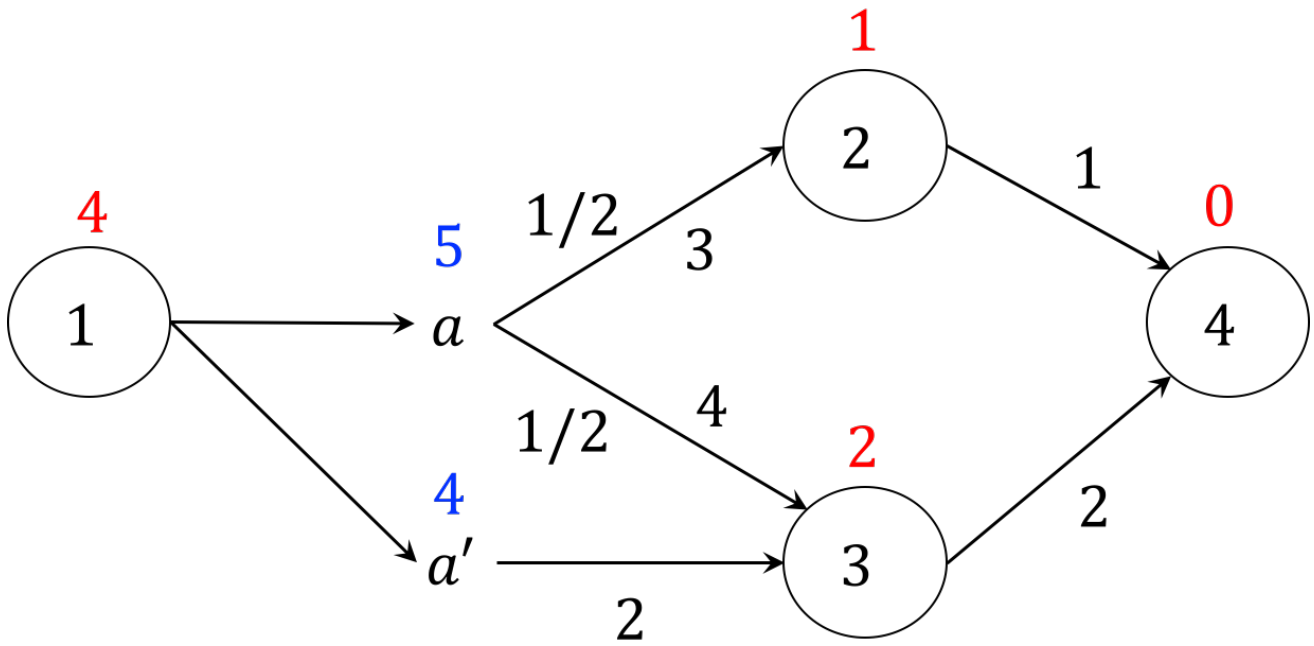
Plus court chemin stochastique

- Trouver la distance minimale (en espérance) de la source s à la t



Fonctions de valeur

- $v^*(i)$: distance minimale (en espérance) de i à la destination finale
- $q^*(i,a)$: distance minimale (en espérance) de (i,a) à la destination finale



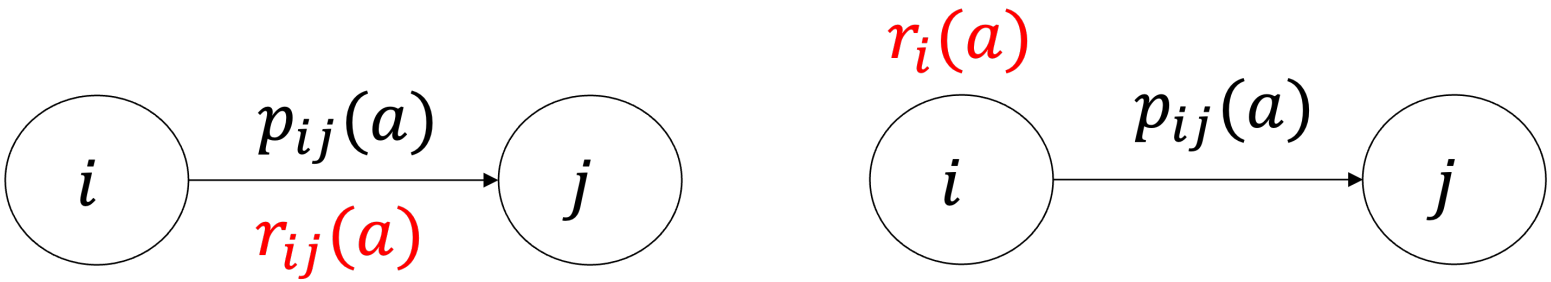
- $q^*(1,a) = \frac{1}{2}[3 + v^*(2)] + \frac{1}{2}[4 + v^*(3)] = 5$
- $q^*(1,a') = 2 + v^*(3) = 4$
- $v^*(1) = \min[q^*(1,a), q^*(1,a')] = 4$

Processus de décision markovien (=MDP)

- 5 ingrédients
  - $\mathcal{S}$: Espace d'états (que nous supposerons fini)
  - $\mathcal{A}$: Espace d'actions (que nous supposerons fini)
  - $p_{ij}(a)$: Probabilité d'aller de l'état i vers l'état j si l'action a est choisie
  - $r_i(a)$: Récompense si l'action a est choisie dans l'état i
  - Objectif : Déterminer la politique qui maximise l'espérance du gain total (taux d'actualisation γ)

- Une politique spécifie l'action à prendre à chaque instant (quel que soit l'état, l'instant, l'historique des états et des actions, ...)

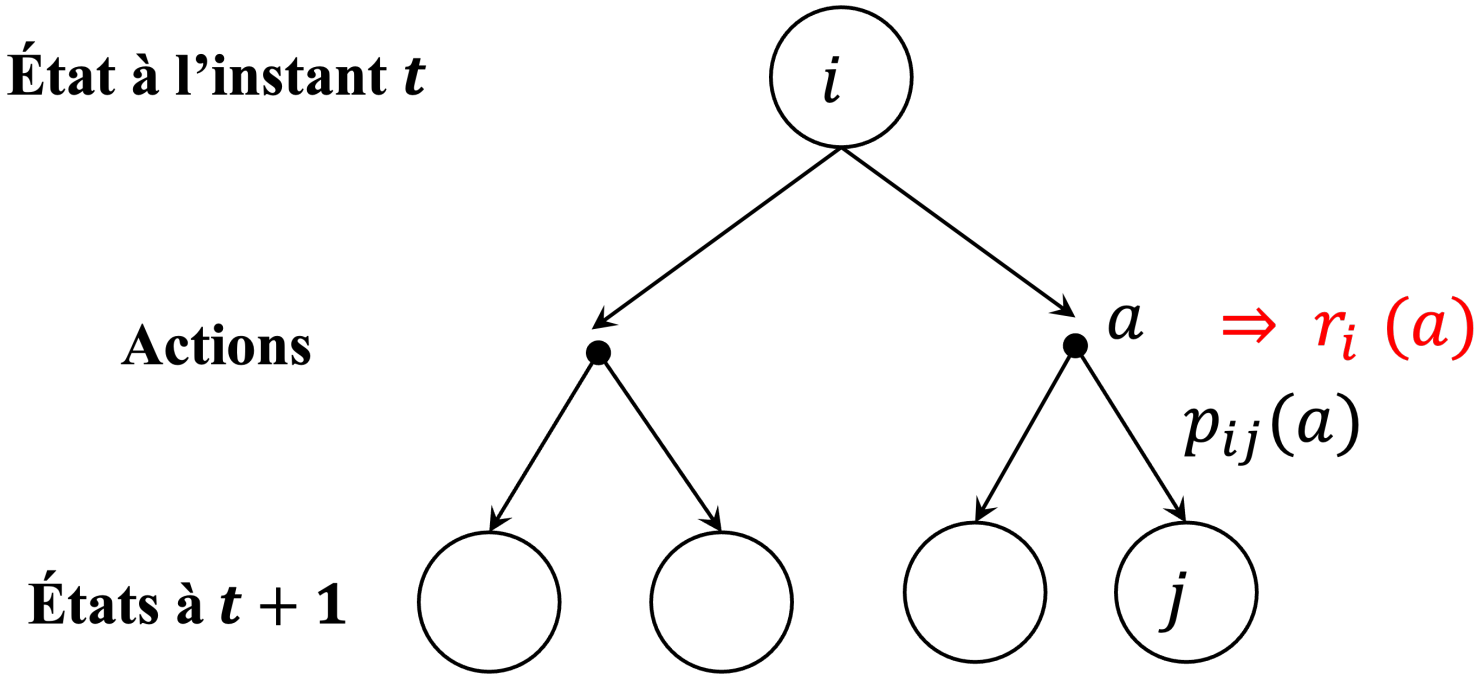
Récompenses sur les arcs ou les sommets



- On peut passer des récompenses sur les arcs à des récompenses sur les sommets sans changer le gain total

$$r_i(a) = \sum_{j \in \mathcal{S}} p_{ij}(a) r_{ij}(a)$$

Dynamique



Politiques de décision

- On peut montrer qu'il existe une politique optimale avec les propriétés suivantes :
  - markovienne (la décision ne dépend que de l'état courant)
  - déterministe (non aléatoire)
  - stationnaire (quel que soit l'instant t, la même décision est prise dans l'état i)
- Dans la suite, nous considérerons des politiques avec ces trois propriétés
- On notera π une politique
- $a(i)$: action prise dans l'état i
- Une telle politique peut être décrite simplement

État i	1	2	3	$\cdots$
Action $a(i)$	Up	Down	Left	$\cdots$

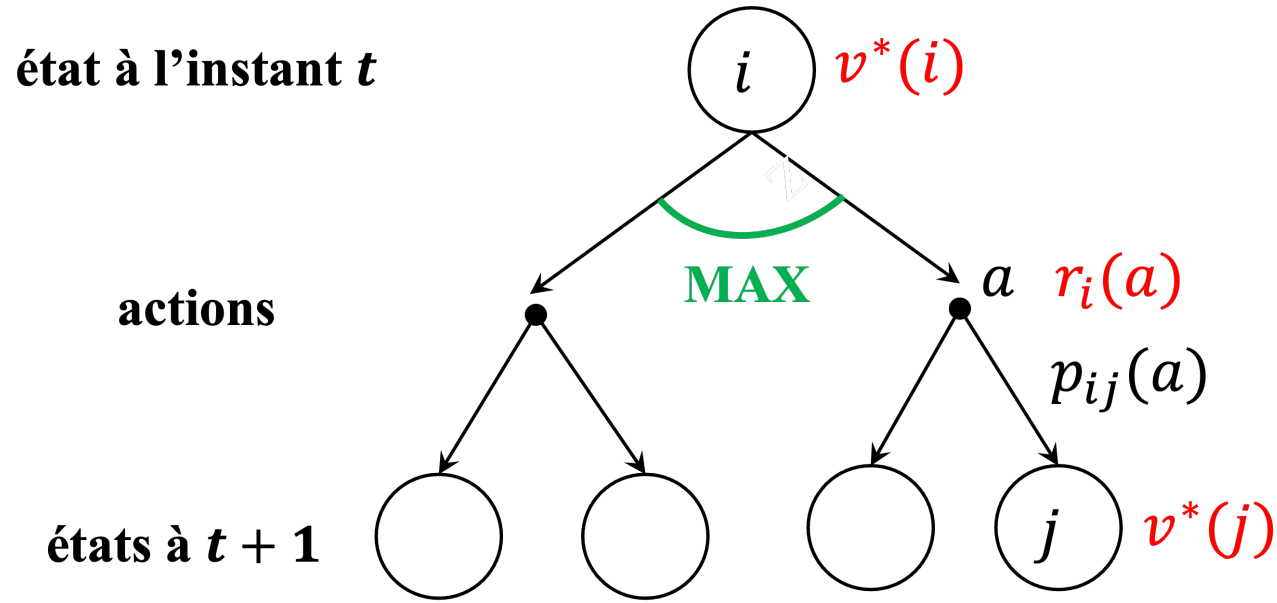
- $|\mathcal{A}|^{|\mathcal{S}|}$ politiques possibles

Fonction de valeur optimale

- $v^*(i)$: Récompense optimale (en espérance) en partant de l'état i
- $q^*(i, a)$: Récompense optimale (en espérance) en prenant l'action a dans l'état l'état i et en appliquant ensuite la politique optimale

$$v^*(i) = \max_{a \in \mathcal{A}} q^*(i, a)$$

Équations d'optimalité pour v^*

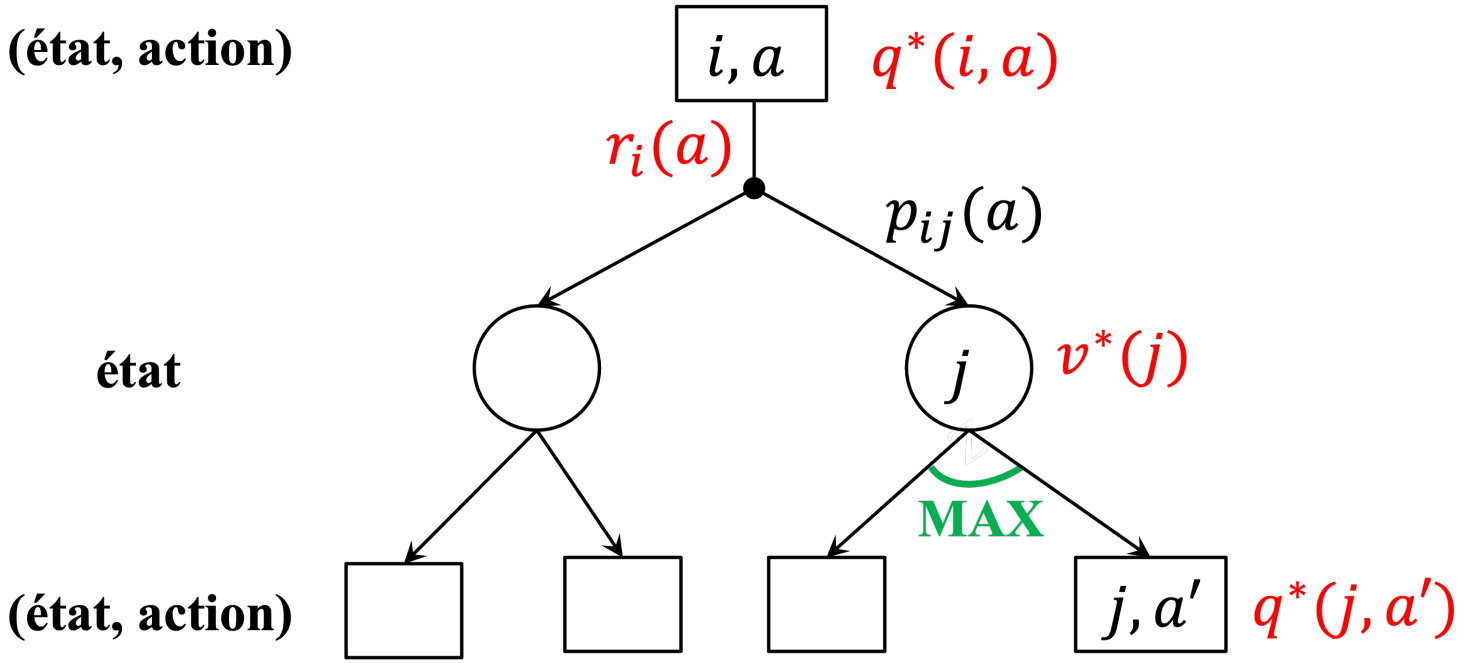


$$q^*(i,a) = \sum_j p_{ij}(a)[r_{ij}(a) + \gamma v^*(j)]$$

$$v^*(i) = \max_a q^*(i,a)$$

$$\boxed{v^*(i) = \max_a r_i(a) + \gamma \sum_j p_{ij}(a) v^*(j)}$$

Équations d'optimalité pour q^*

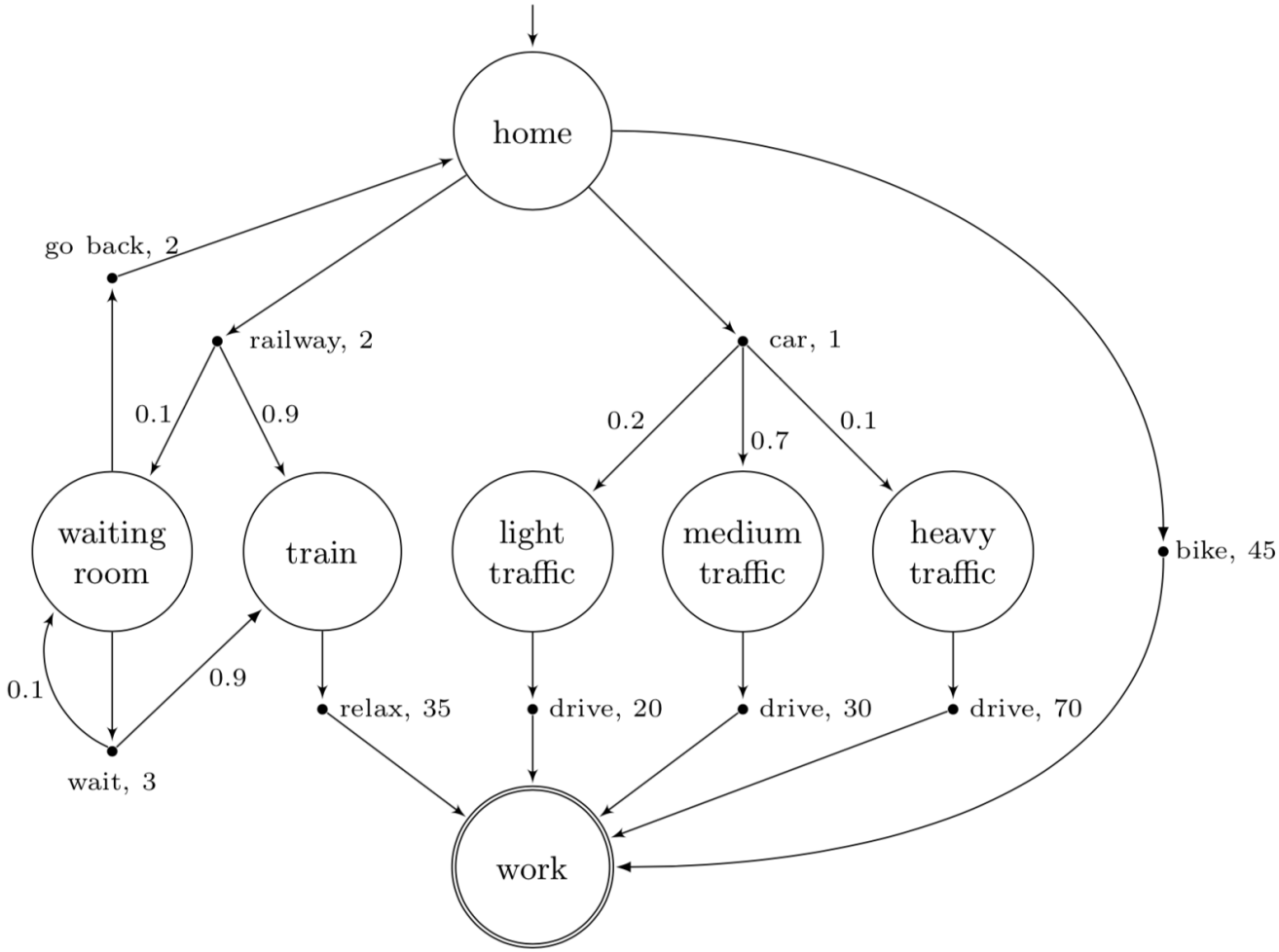


$$q^*(i, a) = r_i(a) + \gamma \sum_j p_{ij}(a) v^*(j)$$

$$v^*(j) = \max_{a'} q^*(j, a')$$

$$\boxed{q^*(i, a) = r_i(a) + \gamma \sum_j p_{ij}(a) \max_{a'} q^*(j, a')}$$

Comment aller à son travail ?



- Objectif : minimiser le temps (en espérance) pour aller à son travail
- Écrire les équations d'optimalité

Algorithme d'itération sur la valeur

- Les équations de Bellman sont des équations au point fixe

$$v^*(i) = \max_a r_i(a) + \gamma \sum_j p_{ij}(a) v^*(j)$$

- Soit la suite de fonctions de valeur v_k définie par

$$v_{k+1}(i) = \max_a r_i(a) + \gamma \sum_j p_{ij}(a) v_k(j)$$

- Si v_k converge, alors elle converge vers v^*

Algorithme d'itération sur la valeur

- Soit l'opérateur T tel que

$$v^*(i) = \underbrace{\max_a r_i(a) + \gamma \sum_j p_{ij}(a) v^*(j)}_{Tv^*(i)}$$

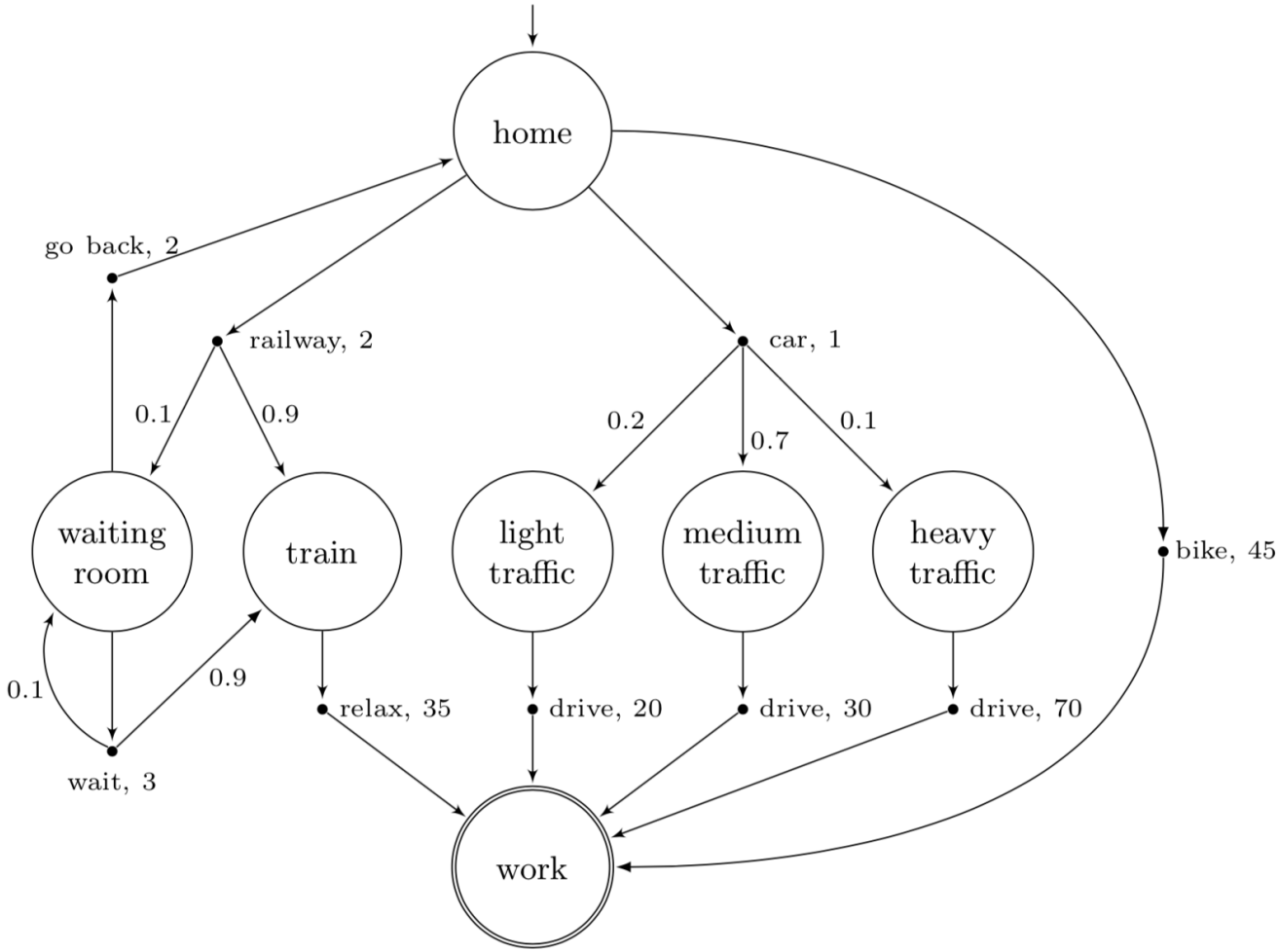
- Point fixe : $v^* = Tv^*$
- On peut montrer que l'opérateur T est γ-contractant

Algorithme d'itération sur la valeur

- v_0 arbitraire
- $v_{k+1} = Tv_k$
- Critère d'arrêt : $\|v_{k+1} - v_k\|_\infty < \epsilon$

- La suite v_k converge vers v^*
  - si $\gamma < 1$
  - si $\gamma = 1$ et qu'il existe un état terminal atteint avec probabilité 1 quelle que soit la politique

Comment aller à son travail ?



- Écrire deux itérations de l'algorithme d'itération sur la valeur, en prenant $v_0 = 0$

Comment obtenir la politique optimale à la fin de l'algorithme ?

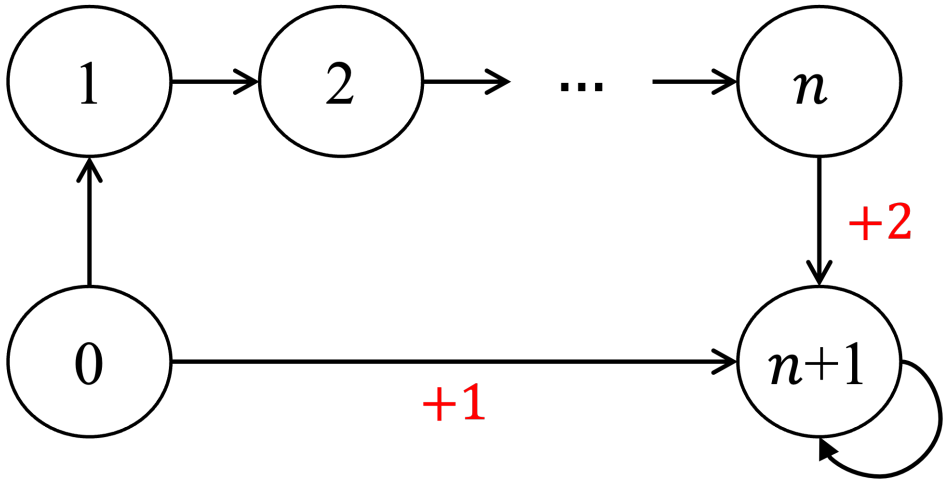
- A la fin de l'algorithme, nous avons v^* (ou du moins une bonne approximation)
- L'action optimale $a^*(i)$ dans l'état i est alors

$$a^*(i) = \arg\max_a q^*(i, a)$$

où

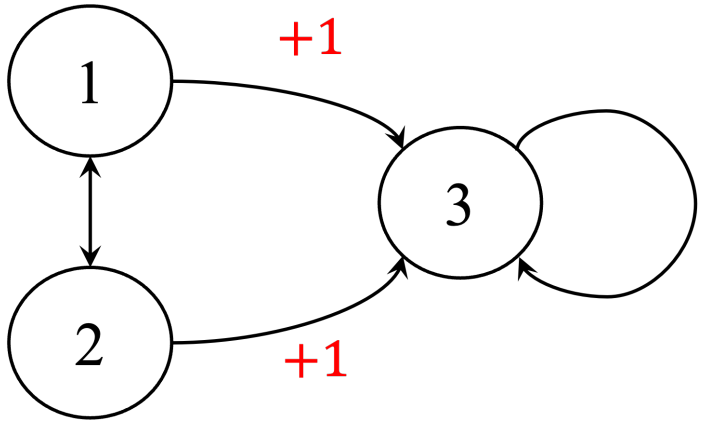
$$q^*(i, a) = r_i(a) + \gamma \sum_{j \in \mathcal{S}} p_{ij}(a) v^*(j)$$

La politique optimale peut dépendre du taux d'actualisation



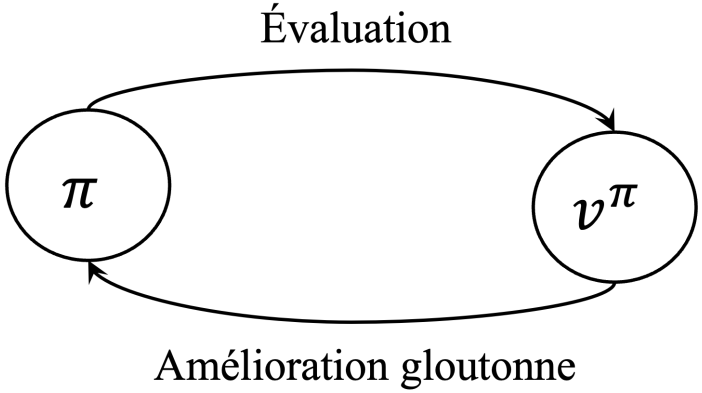
- Quel est le chemin de 0 à $(n+1)$ ayant le plus grand gain ?
- Le chemin du bas procure un gain de 1
- Le chemin du haut procure un gain de $2\gamma^n$
- Si $\gamma = 1$, il vaut mieux toujours passer par en haut
- If $\gamma < 1$, il peut être plus intéressant de passer par en bas (par exemple si $\gamma = 0.9$ et $n = 10$)

Le taux d'actualisation permet d'éviter des cycles



- De manière évidente, nous avons $v^*(1) = v^*(2) = 1$ et $v^*(0) = 0$
- Dans l'état 1,
  - si $\gamma = 1$, il est équivalent d'aller en 2 ou en 3 depuis l'état 1
  - si $\gamma < 1$, il vaut mieux toujours aller dans l'état 3

Algorithme d'itération sur la politique



Algorithme d'itération sur la politique

- Choisir une politique π arbitrairement et prendre $\pi' = \emptyset$
- Tant que π' différent de π
  1. Evaluation de la politique : déterminer v^π
  2. Amélioration de la politique : obtenir une politique π' meilleure que π

$$\forall i, v^{\pi'}(i) \geq v^\pi(i)$$

Algorithme d'itération sur la politique : Evaluation de la politique

- π : politique à l'itération k de l'algorithme
- $a(i)$: action prise dans l'état i
- Déterminer v^π en résolvant les équations de Bellman

$$v^\pi(i) = r_i(a(i)) + \gamma \sum_j p_{ij}(a(i)) v^\pi(j)$$

Itération sur la politique : Amélioration de la politique

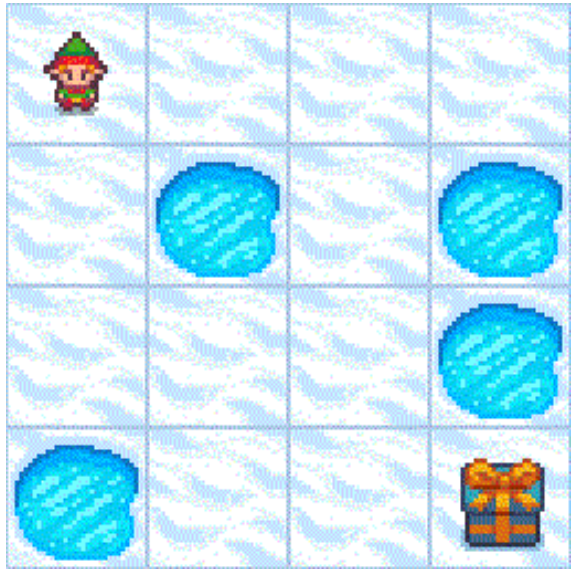
- π : politique à l'itération k de l'algorithme
- π' : politique à l'itération k de l'algorithme
- Dans l'état i, la politique π' sélectionne l'action a qui maximise

$$r_i(a) + \gamma \sum_j p_{ij}(a) v^{\pi}(j) \tag{1}$$

- On peut montrer que la politique obtenue est de récompense supérieure

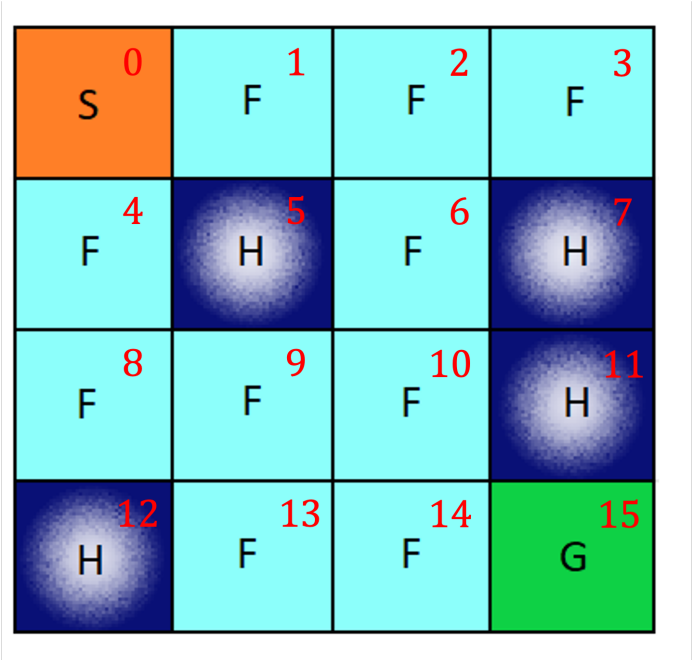
$$v^{\pi'} \geq v^{\pi}$$

Frozen lake (lac gelé)



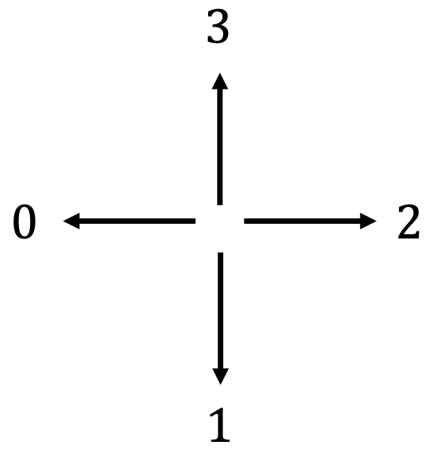
- Objectif : atteindre le cadeau (récompense de +1 si on l'atteint)

Frozen lake - États



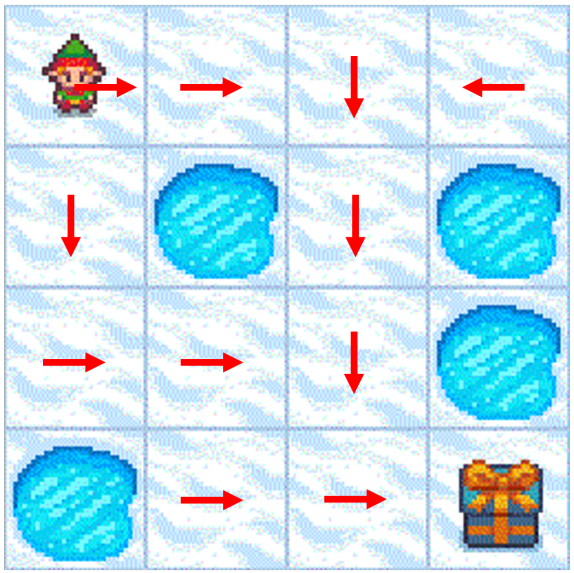
- S = Start, F = Frozen, H = Hole, G = Goal
- États : $\mathcal{S} = \{0, 1, 2, \cdots, 15\}$

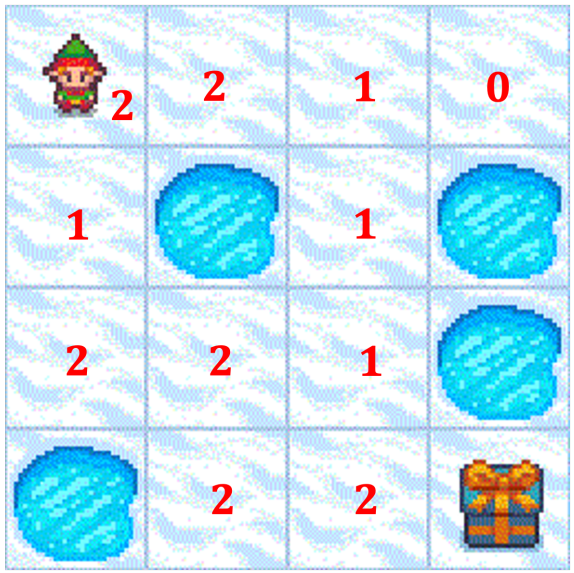
Version déterministe - Actions



- Actions : $\mathcal{A} = \{0, 1, 2, 3\}$ (left, down, right, up)
- Dans un trou H, n'importe quelle action vous laisse dans H (idem pour G)
- Aux bords, un mouvement vers l'extérieur vous laisse dans le même état

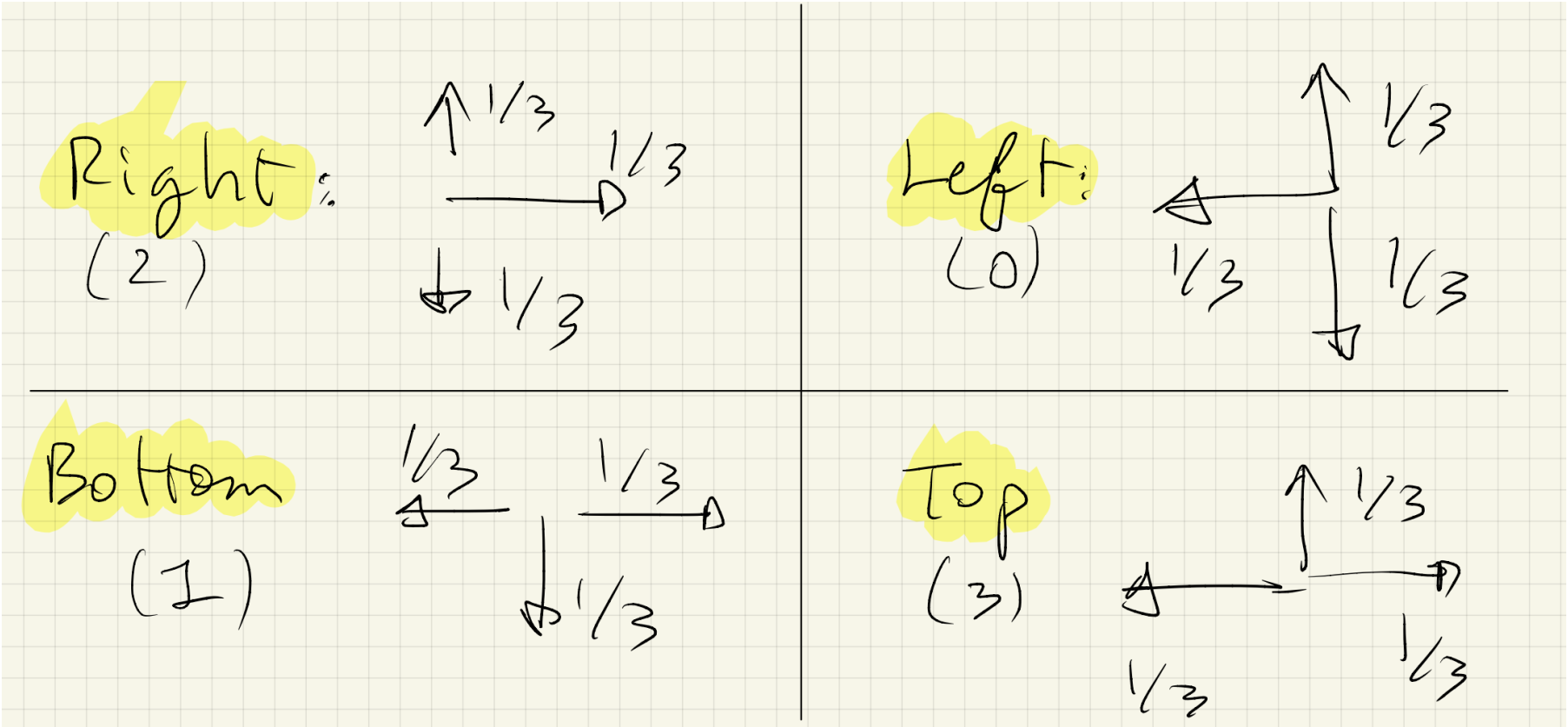
Version déterministe - Une politique optimale





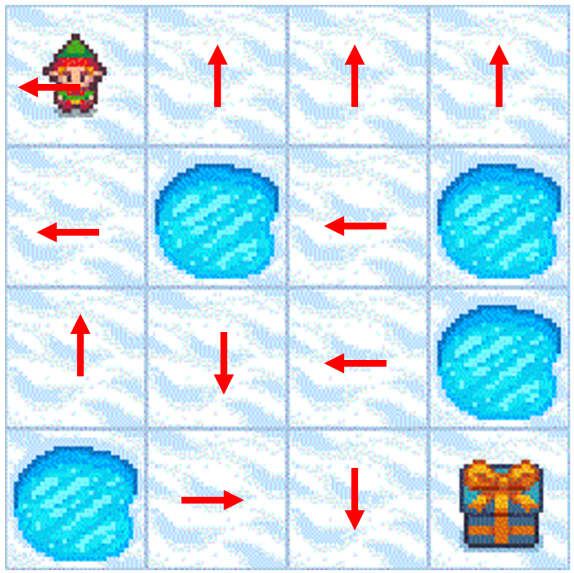
Etat i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Action $a(i)$	2	2	1	0	1	-	1	-	2	2	1	-	-	2	2	-

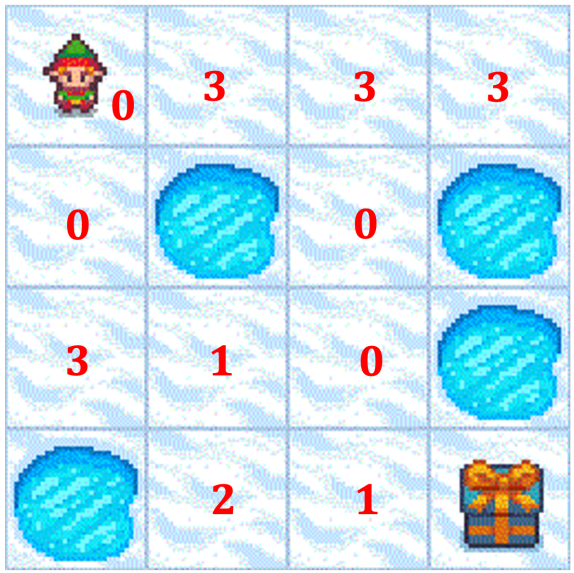
Version stochastique - Probabilités de transition



- Aux bords, un mouvement vers l'extérieur vous laisse dans le même état

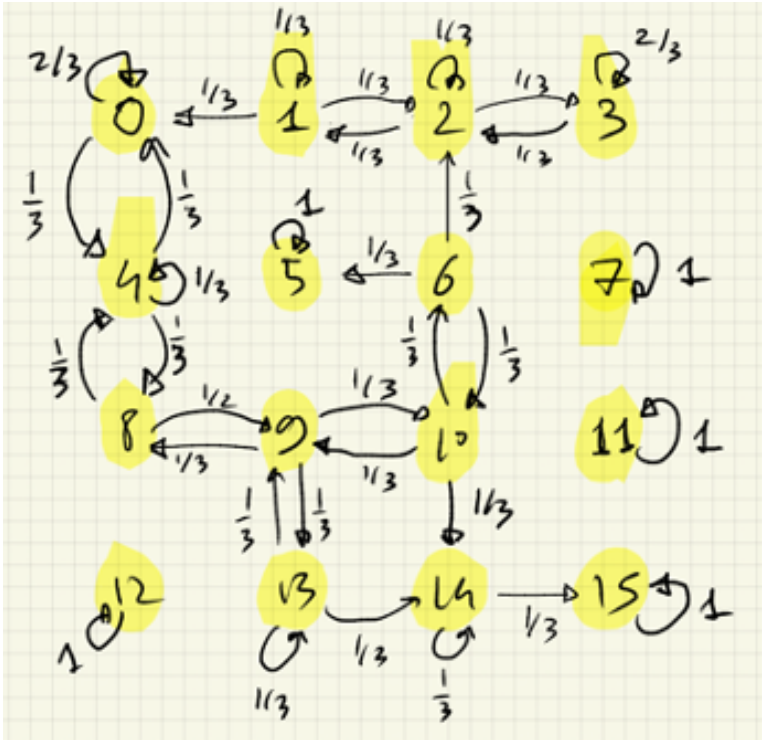
Version stochastique - Une politique optimale





État i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Action $a(i)$	0	3	3	3	0	-	0	-	3	1	0	-	-	2	1	-

Chaîne de Markov si on applique la politique optimale



- On peut montrer que la probabilité d'absorption par le sommet 15 est de $\frac{14}{17}$

Frozen lake - Notations et politiques considérées

- P = Pourcentage de fois où l'on atteint le sommet G
- N = nombre d'étapes nécessaires pour atteint le but G, quand il est atteint
- Politique aléatoire : politique qui choisit l'action $a \in \{0, 1, 2, 3\}$ avec probabilité 1/4
- Politique optimale : politique qui maximise le gain total (en espérance), c'est-à-dire qui maximise la probabilité d'atteindre le cadeau

Travail à la maison concernant Frozen lake

- On considère la version déterministe dans un premier temps
  - Écrire les équations d'optimalité avec un taux d'actualisation γ
  - Implémenter un algorithme d'itération sur la valeur qui retourne la fonction de valeur optimale ainsi que la politique optimale
  - Tester votre algorithme pour $\gamma = 1$ et $\gamma = 0.99$
  - Utiliser une simulation de Monte-Carlo avec 10 000 episodes (ou plus) pour estimer P et N
    - ⋆ Avec la politique optimale
    - ⋆ Avec la politique aléatoire
  - Commenter
- Mêmes questions pour la version stochastique

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Horizon fini

- Horizon de temps $\{0, \cdots, T\}$
- R_t : récompense à la période t
- $v^*(i)$: Récompense optimale en partant de l'état i

$$v^*(i) = \mathbb{E}[\, R_0 + \gamma R_1 + \cdots + \gamma^{T-1} R_{T-1} \,|\, S_0 = i \,]$$

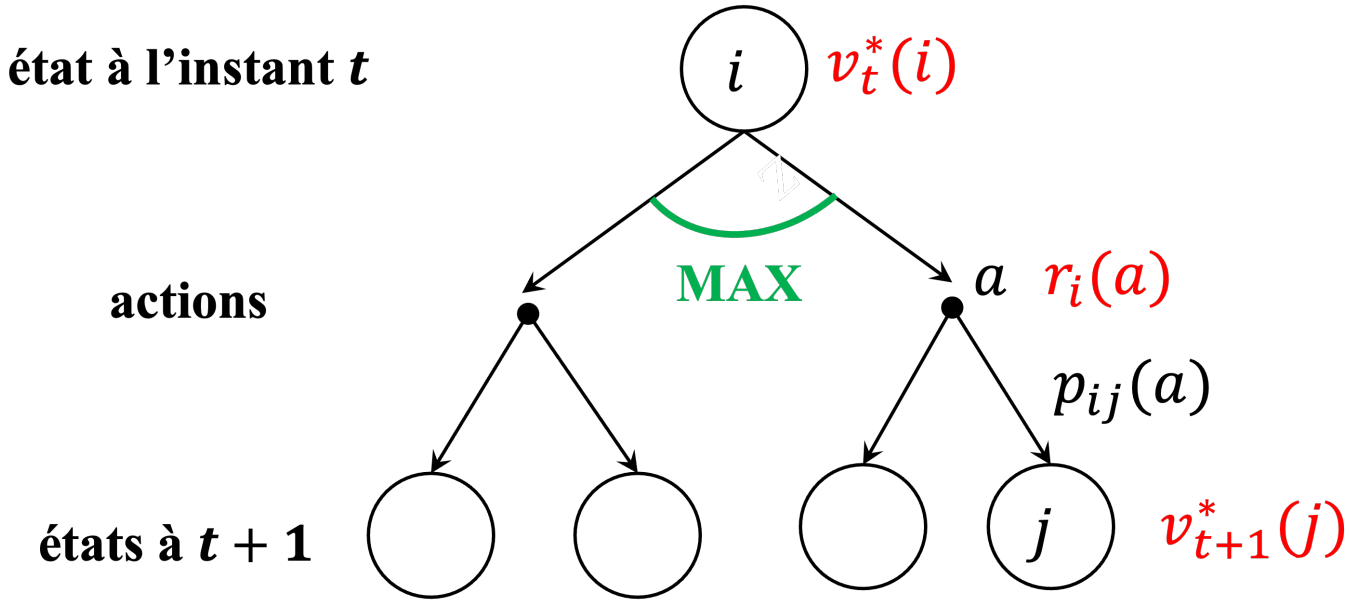
- Sous-problème de t à T
- $v_t^*(i)$: Récompense optimale de t à T en partant de l'état i à l'instant t

$$v_t^*(i) = \mathbb{E}[\, R_t + \gamma R_{t+1} + \cdots + \gamma^{T-1-t} R_{T-1} \,|\, S_t = i \,]$$

Politiques de décision

- On peut montrer qu'il existe une politique optimale avec les propriétés suivantes :
  - markovienne (la décision ne dépend que de l'état courant)
  - déterministe (non aléatoire)
- Dans la suite, nous considérerons des politiques avec ces deux propriétés
- On notera π une politique
- $a(i,t)$: action prise dans l'état i à l'instant t
- $|\mathcal{A}|^{T|\mathcal{S}|}$ politiques possibles

Relation entre v_t^* et v_{t+1}^*



$$v_t^*(i) = \max_a r_i(a) + \gamma \sum_{j \in \mathcal{S}} p_{ij} v_{t+1}^*(j)$$

$$= \max_a q_t^*(i, a)$$

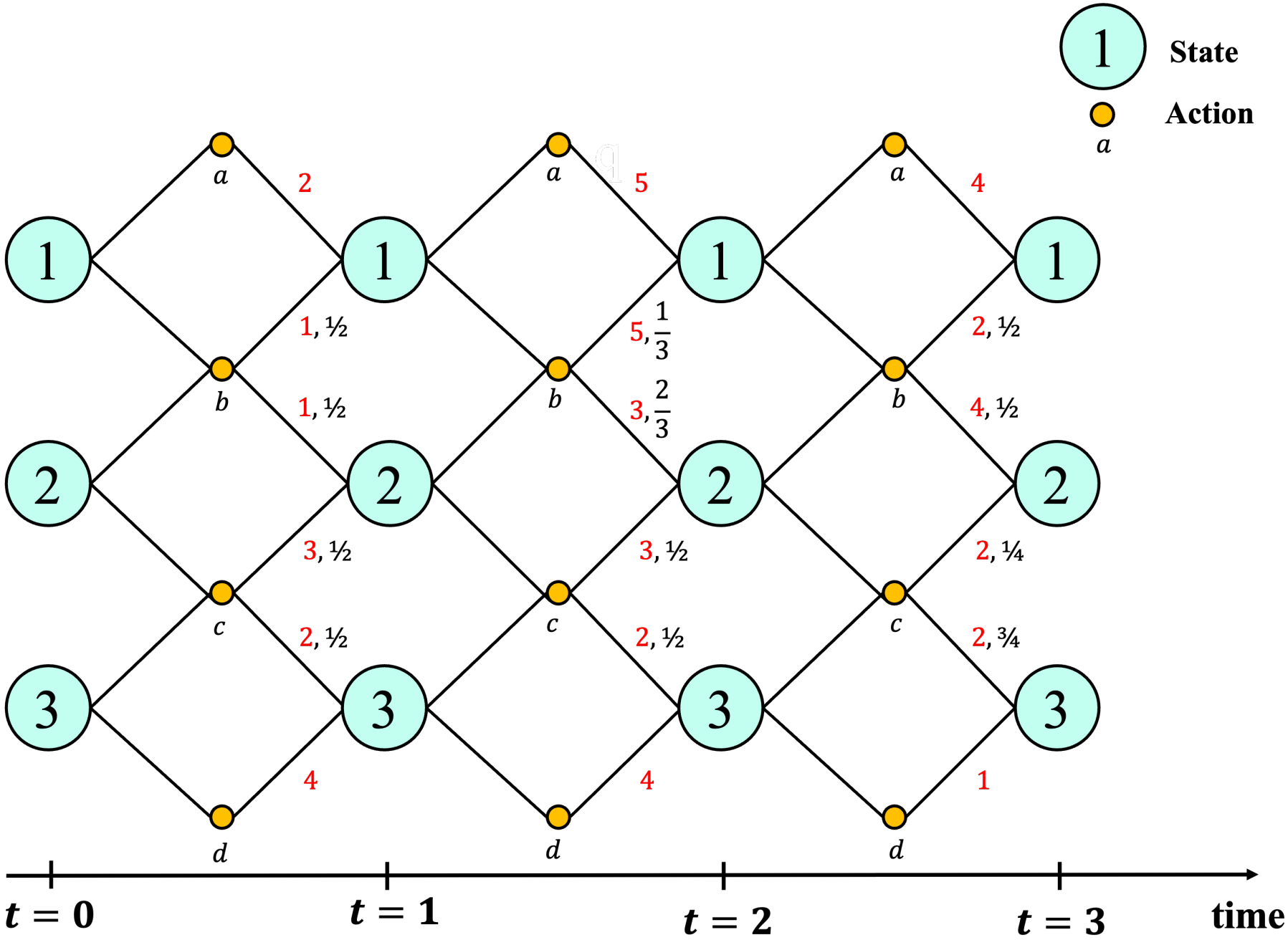
Programme dynamique en $O(|A| \times |\mathcal{S}|^2 \times T)$

Programme dynamique

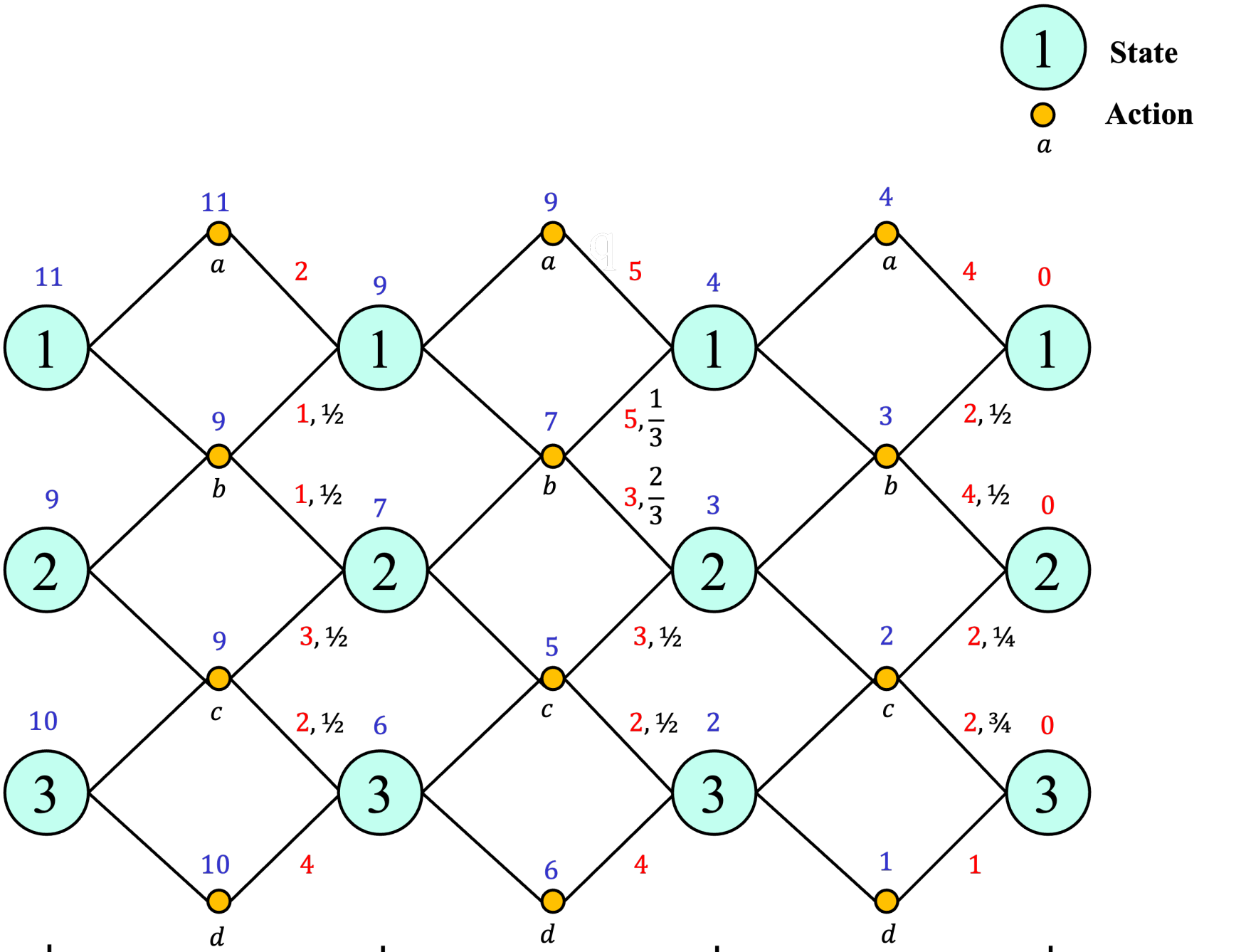
- Initialisation : $v_T^*(i) = 0$ pour $i \in \mathcal{S}$
- Pour $t = T-1, \cdots, 0$
  - Pour $i \in S$

$$v_t^*(i) = \max_a r_i(a) + \gamma \sum_{j \in \mathcal{S}} p_{ij}(a) v_{t+1}^*(j)$$

- $v* = v_0^*$


1
State
Action
a
a
a
a
2
5
4
1
1
1
1
1, ½
5, 1/3
2, ½
b
b
b
1, ½
3, 2/3
4, ½
2
2
2
2
3, ½
3, ½
2, ¼
c
c
c
2, ½
2, ½
2, ¾
3
3
3
3
4
4
1
d
d
d
t = 0
t = 1
t = 2
t = 3
time

- Calculer $v_t^*(i)$ pour chaque état i et chaque instant t
- Calcul $q_t^*(i, a)$ pour chaque paire d'état-action (i, a) et chaque instant t
- Donner une politique optimale


1
State
Action
a
11
9
4
11
a
2
9
a
5
4
a
4
0
1
1
1
1
9
1, ½
7
5, 1/3
3
2, ½
b
b
b
1, ½
3, 2/3
4, ½
9
7
3
0
2
2
2
2
9
3, ½
5
3, ½
2
2, ¼
c
c
c
2, ½
2, ½
2, ¾
10
6
2
0
3
3
3
3
10
4
6
4
1
1
d
d
d

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Problème du bandit manchot

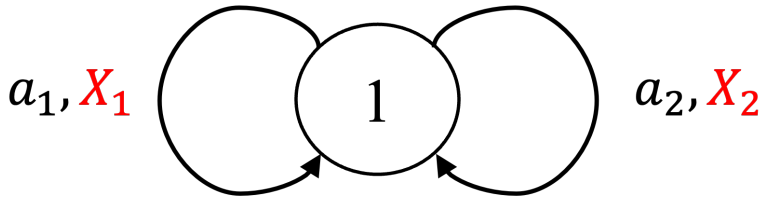
- N machines à sous
- La machine i offre une récompense X_i de loi inconnue
- $\mu_i = E(X_i)$: récompense moyenne de la machine i
- Objectif : Maximiser le gain total (en espérance) après T essais

Exemple

`https://cse442-17f.github.io/LinUCB/`

Problème du bandit manchot = MDP à 1 état

- Exemple avec deux machines



- Objectif : maximiser l'espérance du gain total (ou moyen) sur T périodes

2 algorithmes naifs

RANDOM

- Choisir une machine au hasard

GREEDY (= glouton)

- $\overline{x}_i$: Récompense moyenne obtenue par la machine i jusqu'à présent
- Jouer la machine de plus grand $\overline{x}_i$

Exprimer le gain moyen des 3 politiques suivantes

- RANDOM
- GREEDY après une phase d'initialisation avec un essai par machine
- Politique optimale

Politique optimale et regret

Politique optimale

- Sélectionner la machine ayant la plus grande espérance de récompense

$$\mu^* = \max_i \mu_i$$

- Espérance du gain : $T\mu^*$

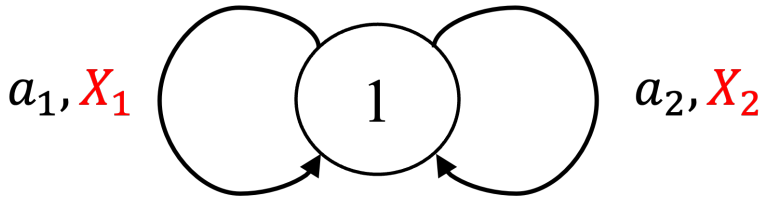
Politique d'apprentissage π

- R_t^π : récompense obtenue à l'étape t par la politique π
- Regret

$$\text{Regret}_T^\pi = \underbrace{T\mu^*}_{\text{récompense optimale}} - \underbrace{E\left(\sum_{t=1}^{T} R_t^\pi\right)}_{\text{récompense de } \pi}$$

Exemple avec deux machines

- Récompenses : $X_1 \sim \mathcal{N}(\mu_1, 1)$ et $X_2 \sim \mathcal{N}(\mu_2, 1)$
- MDP avec un état et 2 actions



- Après 100 étapes, π a choisi 75 fois l'action 1 et 25 fois l'action 2

$$t_1 = 75$$
$$t_2 = 25$$

- Les récompenses moyennes empiriques sont

$$\overline{x}_1 = 10.1$$
$$\overline{x}_2 = 9.2$$

- Quelle machine choisir à l'étape suivante ?

Exploration versus exploitation

- Exploration : Essayer d'obtenir plus d'information sur l'environnement
- Exploitation : Exploiter l'information connue pour maximiser sa récompense
- Il est important à la fois d'explorer et d'exploiter

Exemple du bandit manchot à 2 machines

- Exploitation : choisir la machine 1 qui est le meilleur choix avec l'information dont je dispose
- Exploration : choisir la machine 2

2 algorithmes fondateurs

ϵ-greedy

- Explorer avec probabilité ϵ
- Exploiter avec probabilité $(1 - \epsilon)$

UCB : Upper Confidence Bound

- Optimisme face à l'incertitude
- Intervalles de confiance sur les μ_i : $\mu_i \in [\mu_i^{min}, \mu_i^{max}]$
- Choisir la machine i de plus grand μ_i^{max}

- Les idées de ces algorithmes sont applicables à des MDP à plusieurs états

Algorithme ϵ-greedy

- $\overline{x}_i$: Récompense moyenne obtenue par la machine i jusqu'à présent

Algorithme ϵ-greedy

- Avec probabilité ϵ, jouer une machine au hasard (exploration)
- Avec probabilité $1-\epsilon$, jouer la machine de plus grand $\overline{x}_i$ (exploitation)

- Le regret est linéaire

$$\text{Regret}_T^{\epsilon} = O(T)$$

- Si $\epsilon_t = \frac{a}{t}$, le regret est logarithmique

$$\text{Regret}_T^{\epsilon_t} = O(\log T)$$

Algorithme UCB [6]

- $\overline{x}_i$: Récompense moyenne obtenue par la machine i jusqu'à présent
- t_i : nombre de fois où la machine i à été jouée
- T : nombre total d'étapes $(= \sum_i t_i)$
- c : paramètre contrôlant le degré d'exploration

Algo UCB

- Initialisation : Jouer une fois chaque machine
- Jouer la machine i qui maximise $\overline{x}_i + c\sqrt{\frac{\ln T}{t_i}}$

- Le regret est logarithmique si les récompenses sont bornées
- L'intervalle de confiance vient de l'inégalité de Hoeffding

6. Auer, Cesa-Bianchi and Fischer (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47

Inégalité de Hoeffding

- $X_1, \cdots, X_n$: v.a. indépendantes telles que

$$a_k \leq X_k \leq b_k$$

- $S_n = X_1 + \cdots + X_n$
- Alors, pour tout $\alpha > 0$

$$P\left(S_n \geq E[S_n] + \alpha\right) \leq \exp\left(-\frac{2\,\alpha^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Idée pour obtenir la borne supérieure de UCB

- Soit $X_1, \cdots, X_t$ i.i.d. avec $a_k = 0, b_k = 1$, $E(X_k) = \mu$
- En posant $u = \alpha/t$, il vient

$$P\left(\overline{X}_t \geq \mu + u\right) \leq \exp\left(-2ut^2\right)$$

- Solving $P\left(\overline{X}_t \geq \mu + u\right) = p$ gives

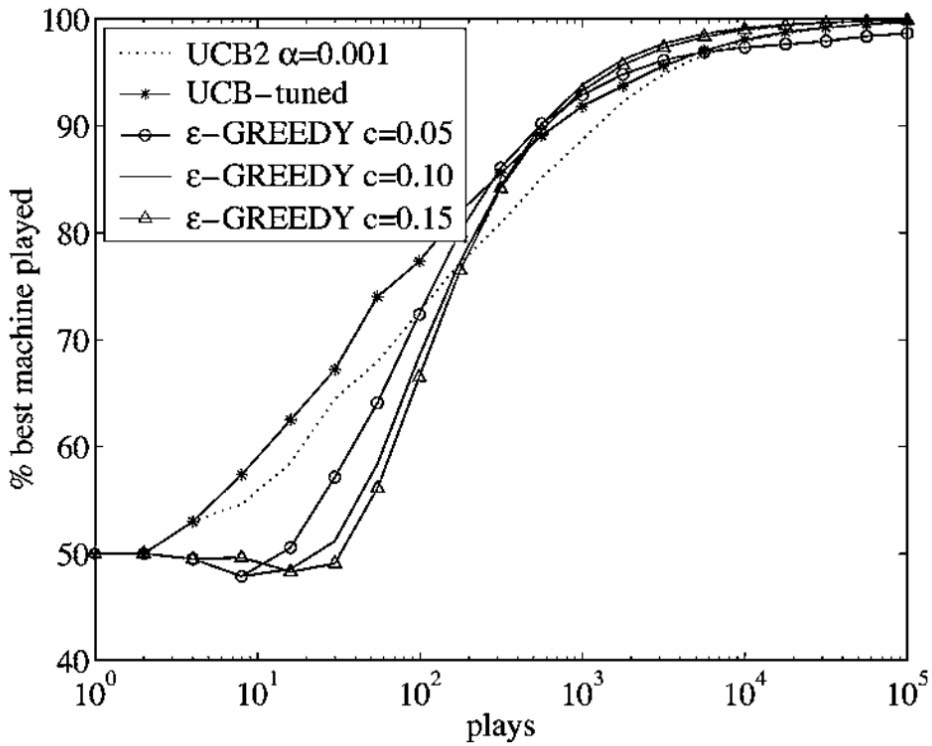
$$u = \sqrt{\frac{-\ln p}{2t}}$$

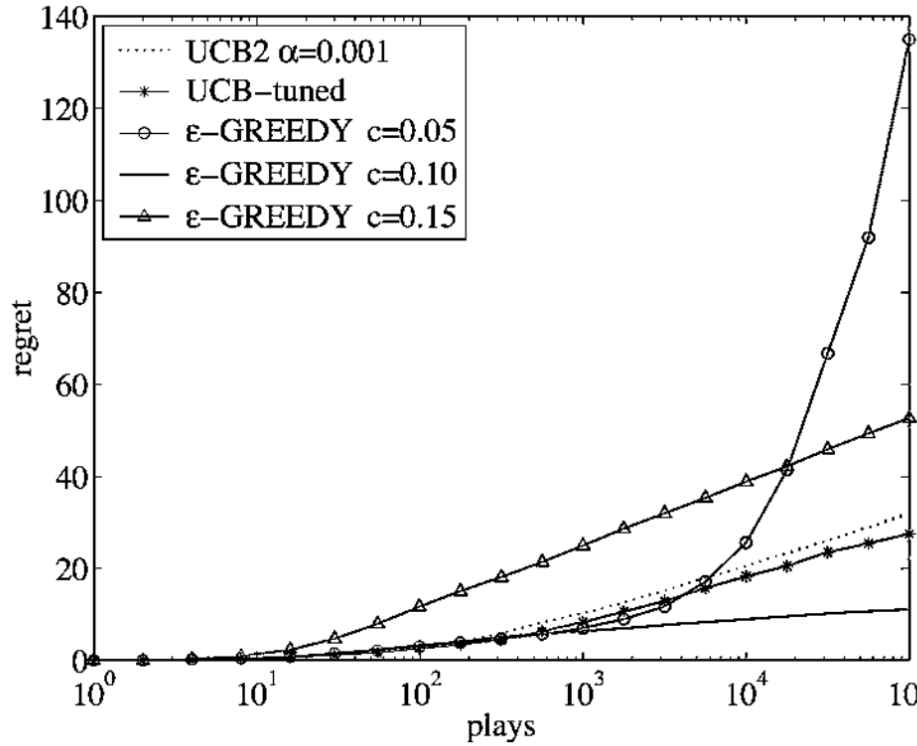
- Diminuer p avec le nombre total de parties T, e.g. $p = \frac{1}{T^4}$

$$u = \sqrt{\frac{2 \ln T}{t}}$$

Exemple de résultats numériques [7]

- 2 machines
- Récompenses distribuées suivant des lois de Bernoulli (paramètres 0.9 et 0.8)

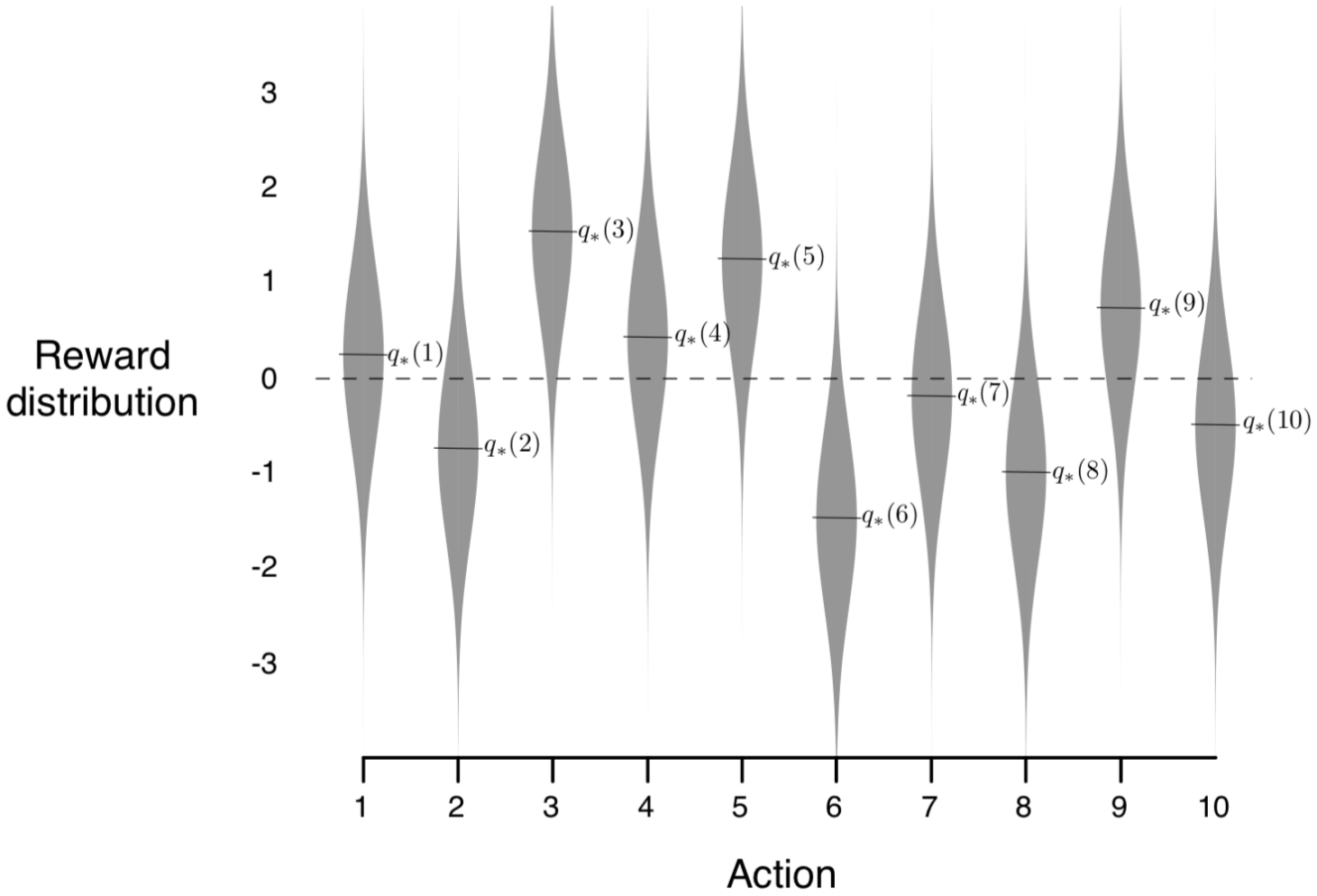




7. Auer, Cesa-Bianchi and Fischer (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47

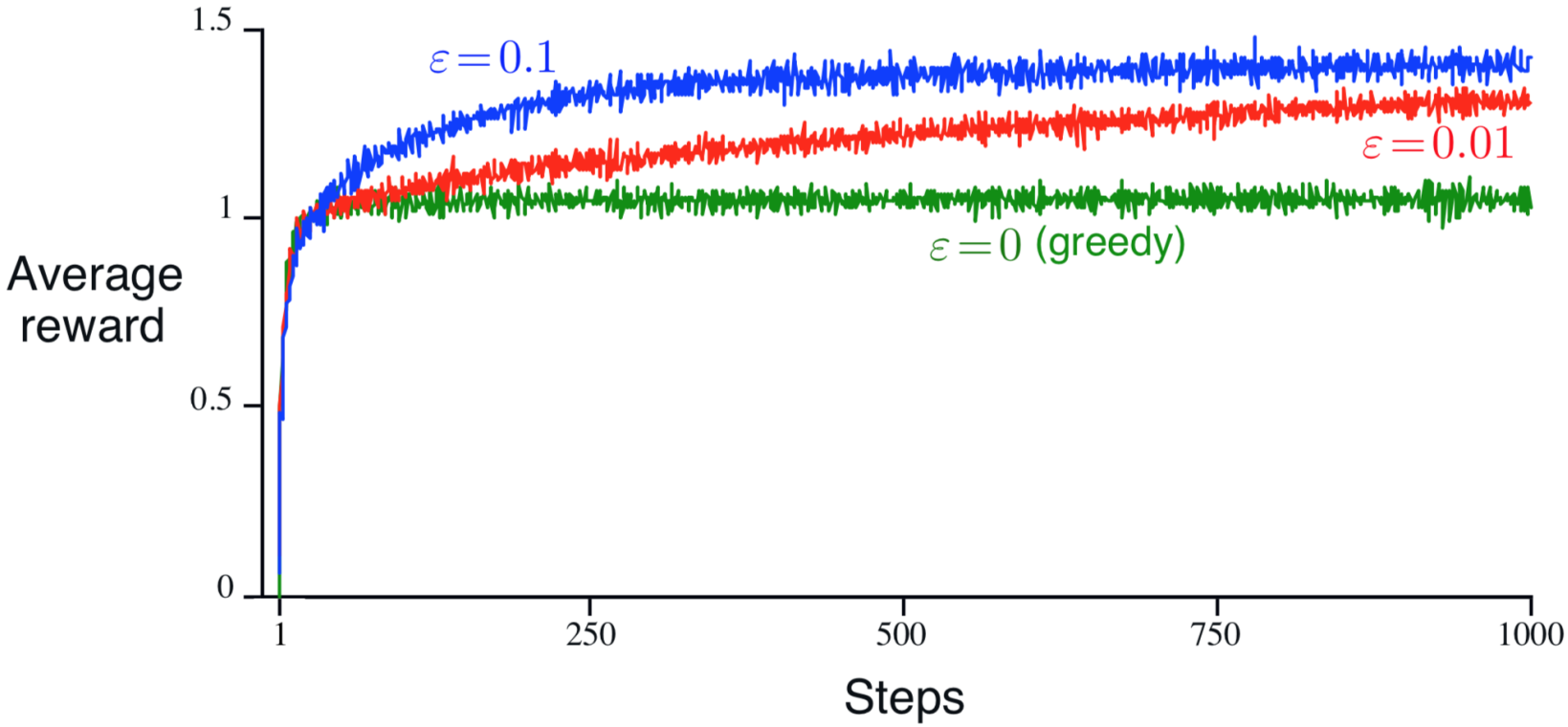
Jeu d'instances de Sutton et Barto (2018)

- 10 machines à sous
- Récompense du bras i : $X_i \sim \mathcal{N}(\mu_i, 1)$ où $\mu_i \sim \mathcal{N}(0, 1)$
- 2000 instances générées. Un exemple ci-dessous.



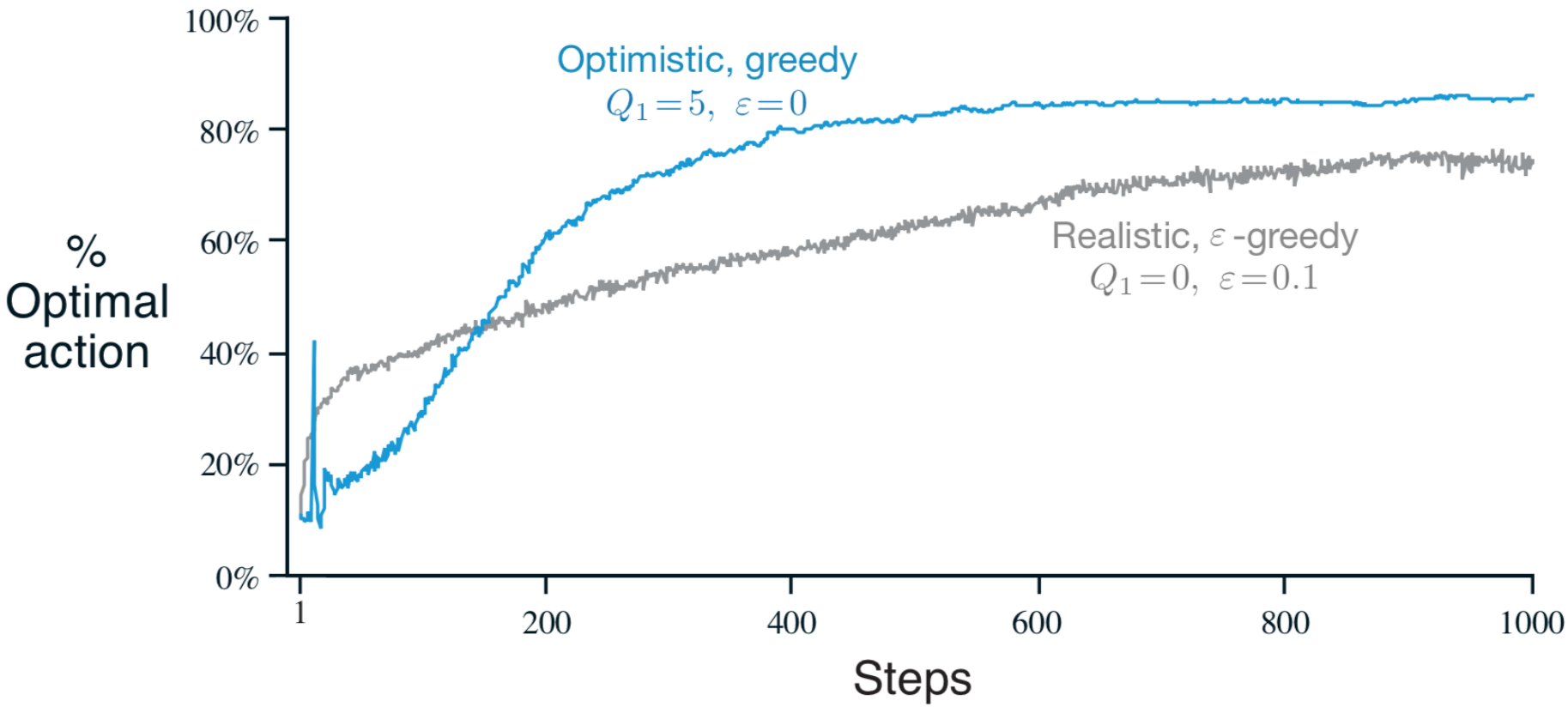
Performances de ϵ-greedy

- Initialisation : $Q(a) = 0$ pour chaque machine a

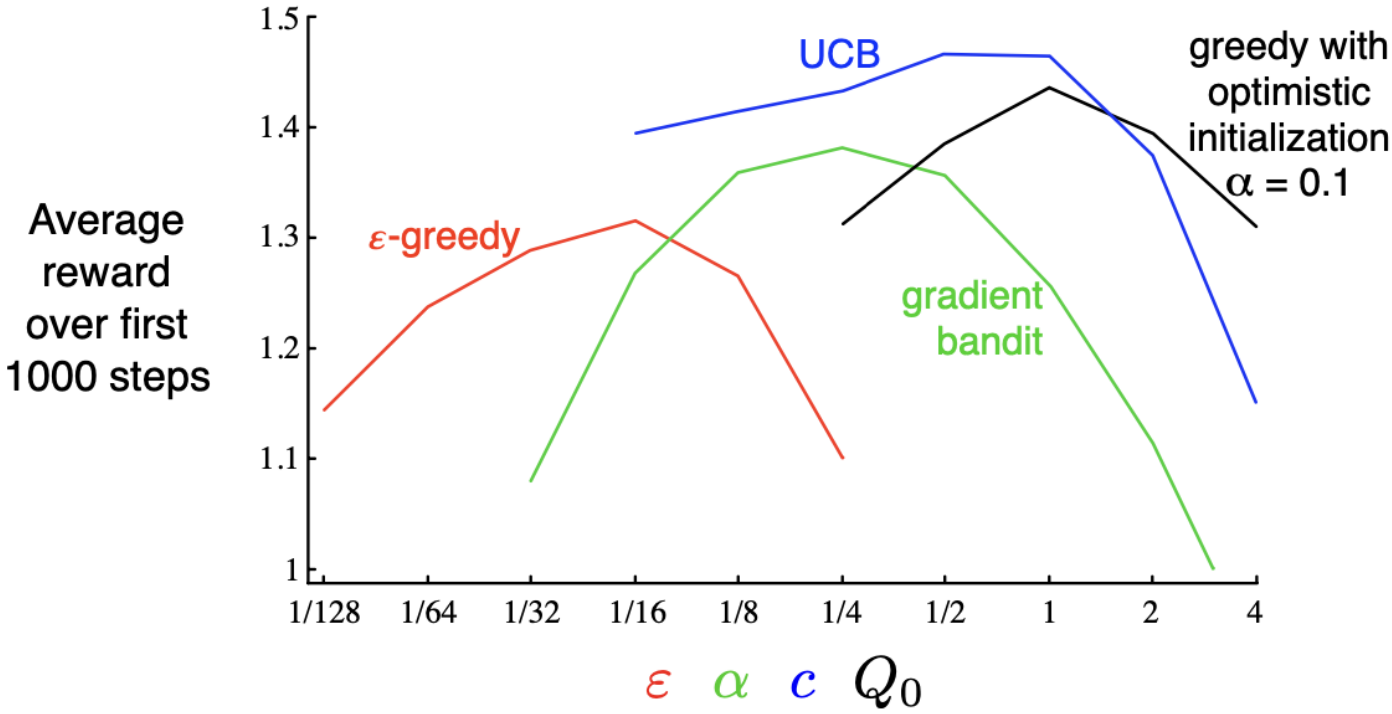


Estimation initiale optimiste

- Initialisation : $Q(a) = +5$ pour chaque machine a



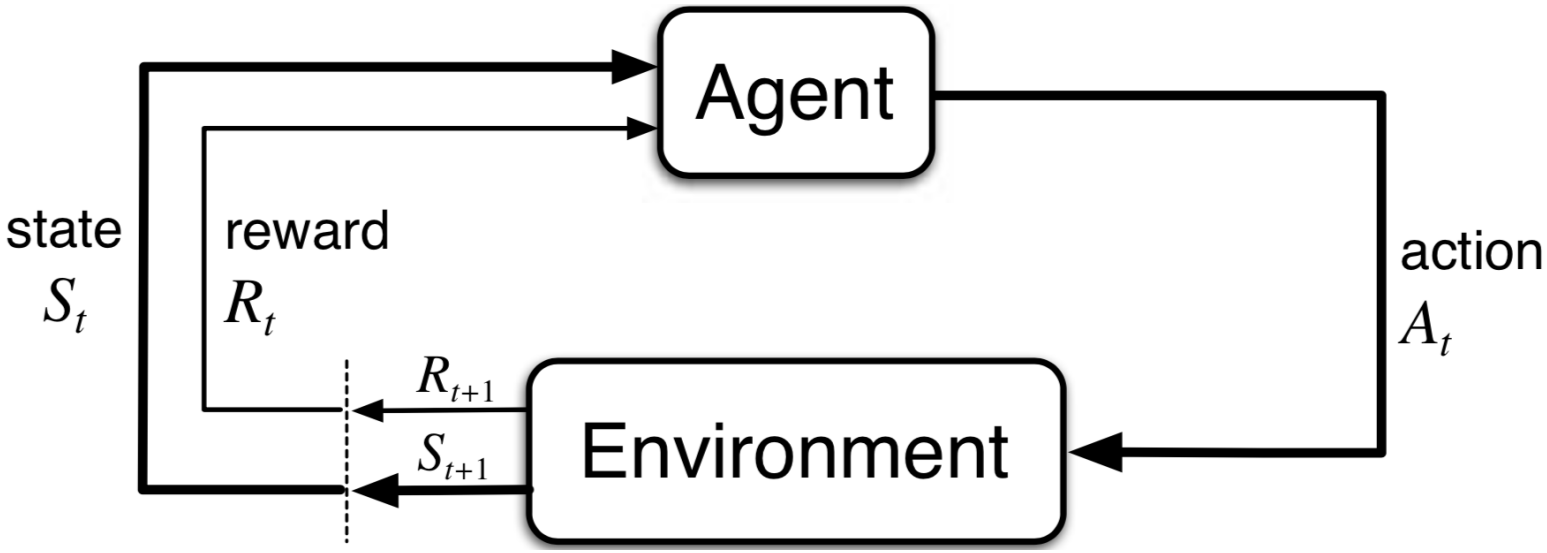
Le paramétrage est un art [Sutton et Barto, 2018]



- Chaque point représente la récompense moyenne pour un algorithme donné avec un paramétrage donné

1. Introduction
2. Chaînes de Markov à temps discret
3. Chaînes de Markov avec récompenses (évaluer une politique)
   - Distributions connues
   - Distributions inconnues
4. Processus de décision markovien (déterminer la politique optimale)
   - Horizon infini
   - Horizon fini
5. Apprentissage par renforcement (environnement inconnu)
   - Bandit manchot
   - Processus de décision markovien
6. Chaîne de Markov à temps continu

Apprentissage par renforcement



- Etat $\rightarrow$ Action $\rightarrow$ Récompense $\rightarrow$ Etat $\rightarrow$ Action $\rightarrow$ Récompense $\rightarrow$ ...
- Récompenses et probabilités de transitions inconnues
- Objectif : maximiser le gain total
- Apprentissage par renforcement = Reinforcement learning = RL

Apprentissage par renforcement dans un MDP

On considère un processus de décision markovien (MDP = Markov decision process)

- $\mathcal{S}$ et $\mathcal{A}$: Espace d'états et d'actions
- $p_{ij}(a)$: Probabilités de transition inconnues
- $r_i(a)$: Récompenses inconnues
- Objectif : Déterminer la politique qui minimise l'espérance du gain total (taux d'actualisation γ)

Le MDP est une boite noire

Modèle avec simulateur

- Choisir un état S et une action A
- Observer la récompense R et l'état suivant S'

Modèle navigant

- Initialiser l'état S
- Tant que S n'est pas terminal
  1. Choisir une action A
  2. Observer la récompense R et l'état suivant S'
  3. $S \leftarrow S'$

Regret d'une politique

- π^* : politique optimale
- π : une politique
- R_t^π : récompense obtenue en période t par la politique π
- Récompense totale (en espérance) de la politique π sur T périodes

$$v_T^\pi = E\left[\sum_{t=1}^{T} R_t^\pi\right]$$

Définition

Le regret de la politique π est la différence entre les récompenses de π^* et π.

$$Regret_T^\pi = v_T^* - v_T^\pi$$

Politique no-regret

Définition

Une politique π est dite no-regret si son regret est sous-linéaire.

$$\frac{Regret_T^\pi}{T} \underset{T \to +\infty}{\longrightarrow} 0$$

- Si la politique est no-regret, le ratio de compétitivité tend vers 1

$$\frac{v_T^\pi}{v_T^*} \underset{T \to +\infty}{\longrightarrow} 1$$

Politique ϵ-greedy

- $Q(s,a)$: estimation de $q^*(s,a)$
- Politique gloutonne

$$\text{argmax}_a Q(s,a)$$

- Politique ϵ-greedy

$$a \leftarrow \begin{cases} \text{argmax}_a Q(s,a) & \text{avec proba } 1-\epsilon \quad \text{(exploitation)} \\ \text{random action} & \text{avec proba } \epsilon \quad \text{(exploration)} \end{cases}$$

Algorithme de Q-learning

- Initialiser arbitrairement $Q(s, a)$ pour tout s, a (doit valoir 0 pour les états terminaux)

Pour chaque épisode

- Initialiser l'état S
- Tant que S n'est pas terminal
  1. Choisir une action A selon la politique ϵ-greedy
  2. Observer la récompense R et l'état suivant S'
  3. Mettre à jour Q

$$\underbrace{Q(S,A)}_{\text{Nouvelle estimation}} \leftarrow (1-\alpha)\underbrace{Q(S,A)}_{\text{Ancienne estimation}} + \underbrace{\alpha}_{\substack{\text{Taux} \\ \text{d'apprentissage}}} [R + \gamma \underbrace{\max_{a'} Q(S',a')}_{\substack{\text{Estimation} \\ \text{des récomenses futures}}}]$$

  4. $S \leftarrow S'$

Q-learning : quelques remarques

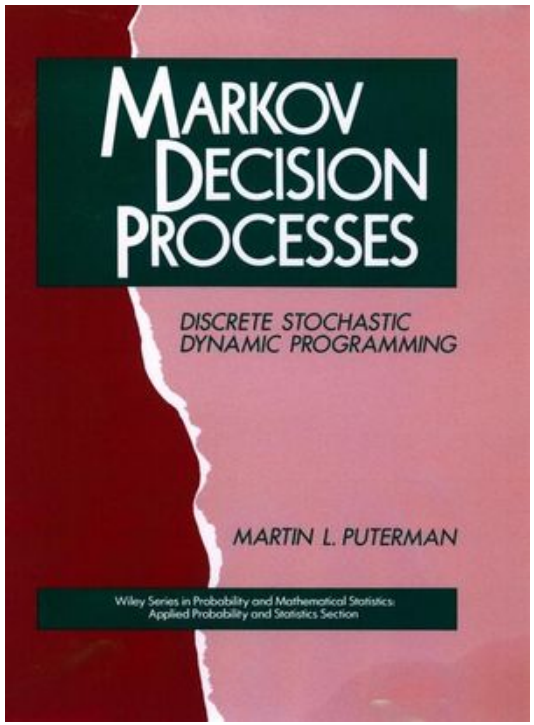
- Mélange intelligent de simulation et de programmation dynamique
- On-line : propose une heuristique à tout instant

$$\text{argmax}_a Q(s, a)$$

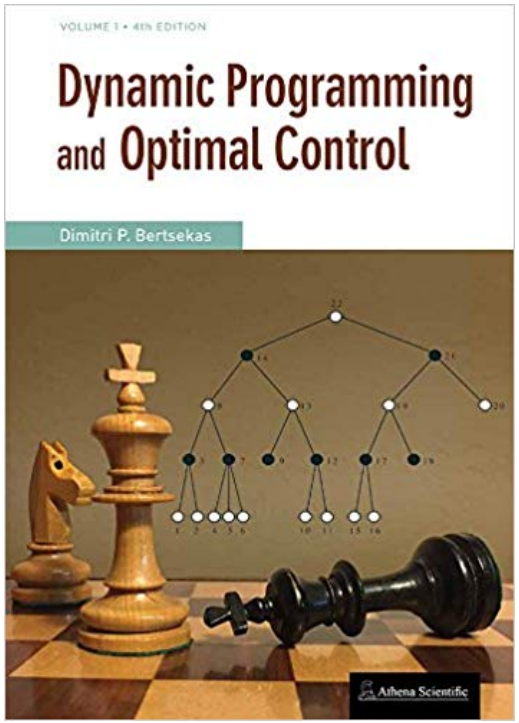
- Converge (doucement) vers la politique optimale sous certaines conditions
- Peut fonctionner en environnement stationnaire
- "Deep Q-learning" : Lorsque l'espace d'états devient grand, on doit approcher la fonction $Q(s, a)$ en utilisant par exemple des réseaux de neurones

Pour aller plus loin

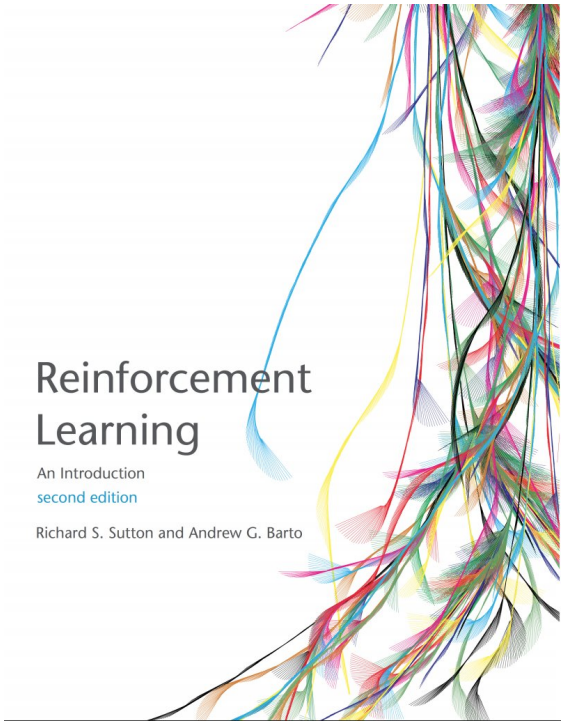
Puterman



Bertsekas



Sutton and Barto

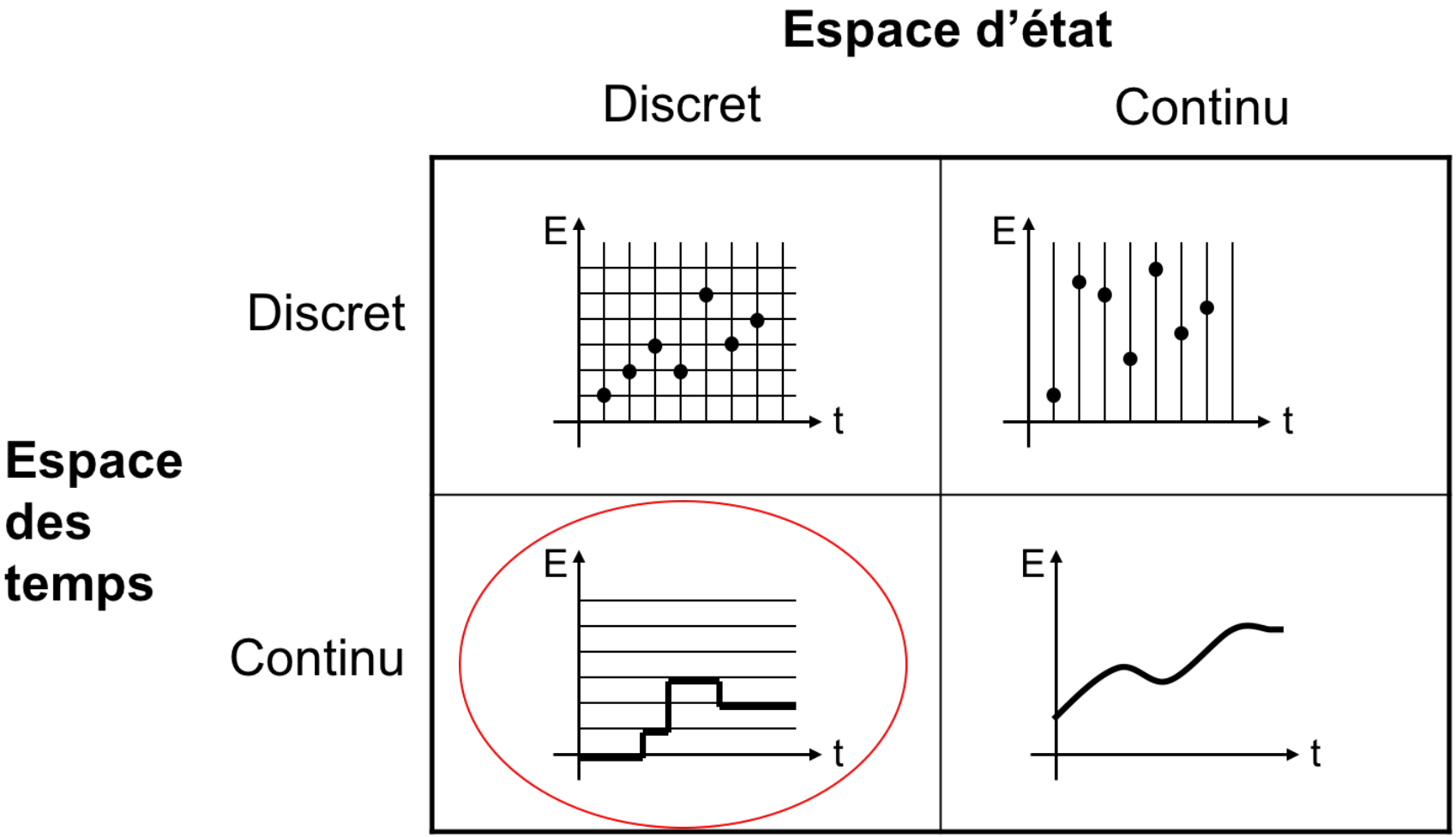


Présentations et vidéos par David Silver sur le RL :
`https://www.davidsilver.uk/teaching/`

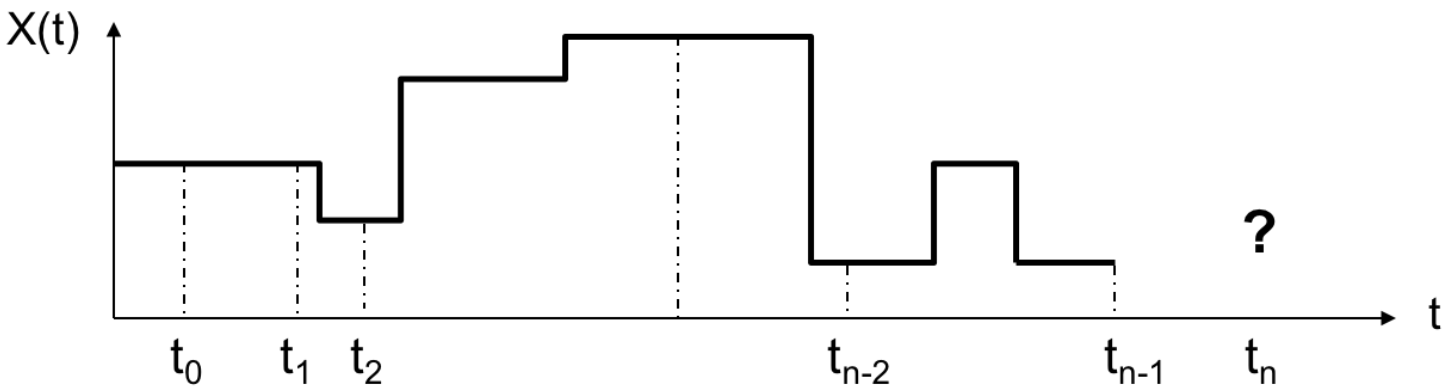
Programme de l'examen

- 1/3 sur les TP
- 1/3 tiré du polycopié d'exercices
- 1/3 surprise

Classification des processus aléatoires



Chaîne de Markov à Temps Continu (CMTC)



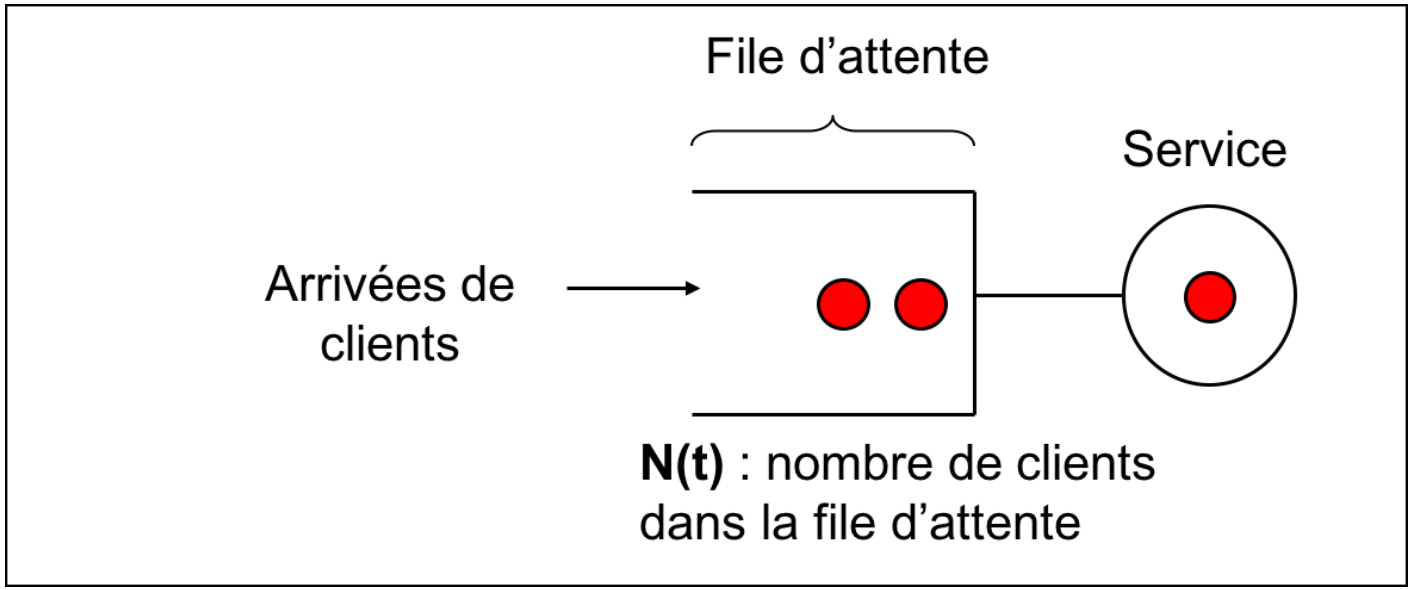
Définition

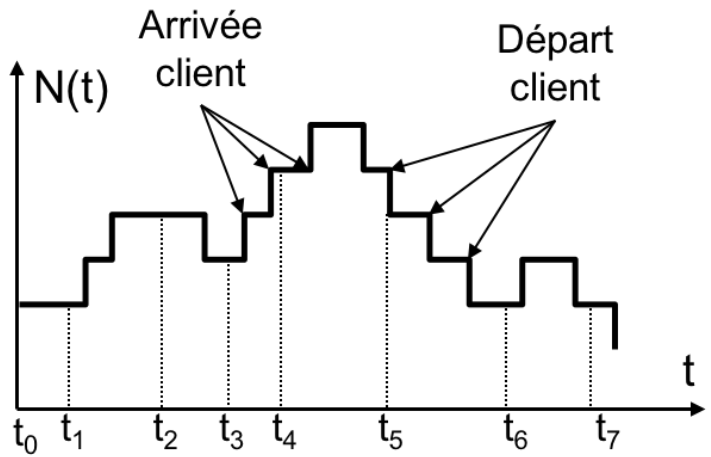
Le processus $(X(t))_{t>0}$ *est une CMTC si,* $\forall t_0 < t_1 < \cdots < t_n$,

$$P[X(t_n) = i_n \,|\, X(t_{n-1}) = i_{n-1}, X(t_{n-2}) = i_{n-2}, \cdots, X(t_0) = i_0]$$
$$= P[X(t_n) = i_n \,|\, X(t_{n-1}) = i_{n-1}]$$

- Une information détaillée du passé ne fournit pas plus d'info sur l'évolution future que la connaissance de la dernière observation

Exemple : file d'attente (supermarché, serveur, ...)





CMTC homogène

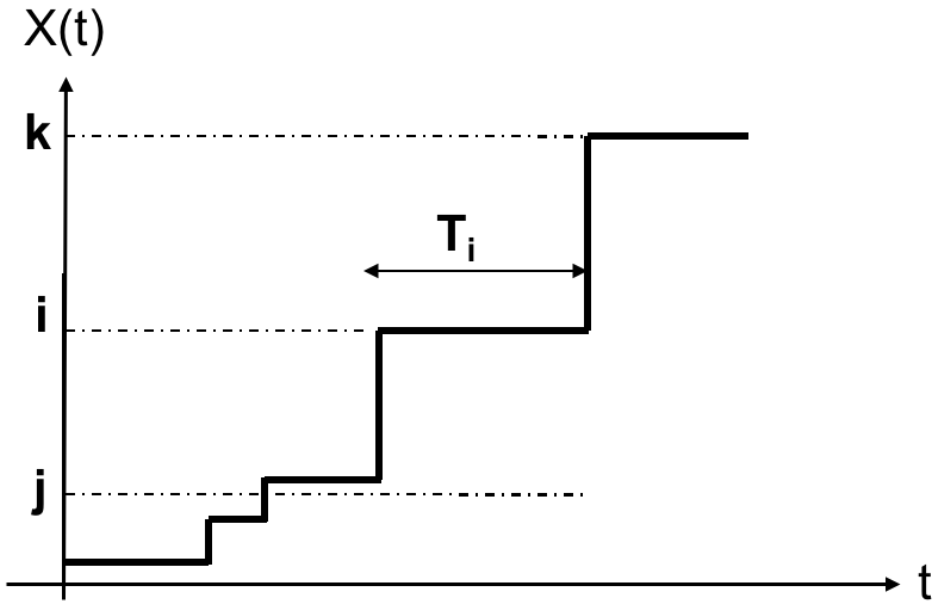
Définition

Une CMTC est dite homogène si les probabilités de transition ne dépendent que de la durée entre 2 observations

$$P[X(t+s) = j \,|\, X(s) = i] = P[X(t) = j) \,|\, X(0) = i]$$

- Par la suite, uniquement des CMTC homogènes

Temps de séjour dans un état



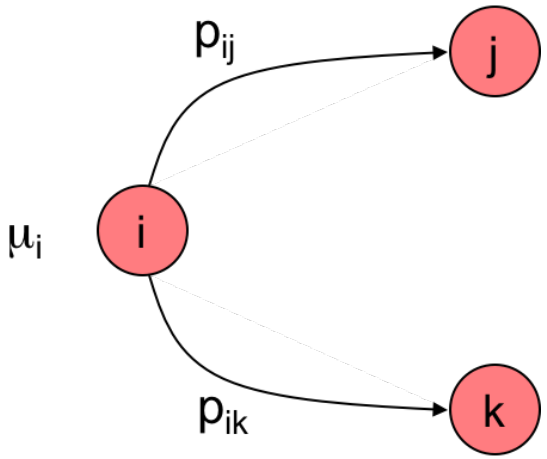
Propriété

Le temps de séjour T_i dans l'état i d'une CMTC homogène suit une loi exponentielle de taux μ_i.

$$T_i \sim \exp(\mu_i)$$

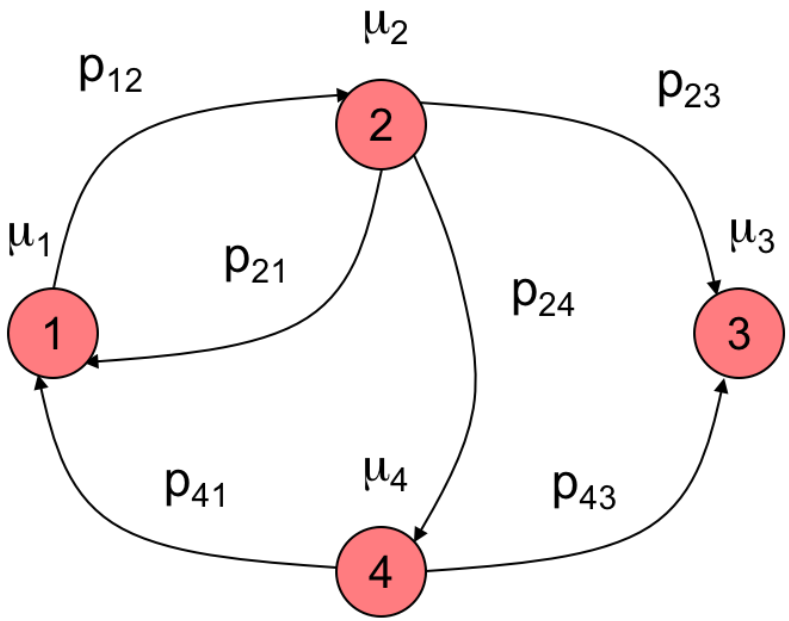
1ère caractérisation d'une CMTC homogène

- $T_i \sim \exp(\mu_i)$: temps de séjour dans l'état i
- p_{ij} : probabilité de transition de i à j, lorsqu'on quitte i



1ère caractérisation d'une CMTC homogène

- Une CMTC homogène est entièrement définie par les $\{\mu_i\}$ et les $\{p_{ij}\}$



Lois exponentielles

- Soit $T \sim \exp(\lambda)$.
- Densité

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{sinon} \end{cases}$$

- Interprétation de la densité

$$f(t)dt \simeq P(t < T \leq t + dt)$$

- Fonction de répartition

$$F(t) = P(T \leq t) = \begin{cases} 1 - e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Lois exponentielles (suite)

- Espérance, écart-type et coefficient de variation

$$E(T) = \sigma(T) = \frac{1}{\lambda}$$

$$cv(T) = \frac{\sigma(T)}{E(T)} = 1$$

- Le coefficient de variation est sans dimension et permet de comparer la variabilité de variables aléatoires dans des unités différentes

Lois exponentielles (suite)

Propriété

Soit deux v.a. indépendantes $X_1 \sim \exp(\lambda_1)$ *et* $X_2 \sim \exp(\lambda_2)$. *Alors :*

- $\min(X_1, X_2) \sim \exp(\lambda_1 + \lambda_2)$
- $P(X_1 < X_2) = \dfrac{\lambda_1}{\lambda_1 + \lambda_2}$

- Généralisation à n lois exponentielles indépendantes :

$$\min(X_1, \cdots, X_n) \sim \exp(\lambda_1 + \cdots + \lambda_n)$$

$$P(X_1 < \min(X_2, \cdots, X_n)) = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_n}$$

Taux de panne

Définition

Le taux de panne d'une v.a. de densité f et de fonction de répartition F est la fonction

$$\tau(t) = \frac{f(t)}{1 - F(t)}.$$

- Interprétation en fiabilité : soit T la durée de vie d'un appareil.

$$\tau(t)dt \simeq \text{proba qu'une panne se produise entre } t \text{ et } t + dt$$
$$\text{sachant que l'appareil fonctionne encore à } t$$

- Interprétation pour les CMTC : soit T le temps de séjour dans un état.

$$\tau(t)dt \simeq \text{proba de quitter l'état entre } t \text{ et } t + dt$$
$$\text{sachant que l'on est encore dans cet état à } t$$

Taux de panne d'une loi exponentielle

Propriété

Pour une loi exponentielle de taux λ, le taux de panne est constant :

$$\tau(t) = \lambda$$

Propriété

La loi exponentielle est sans mémoire :

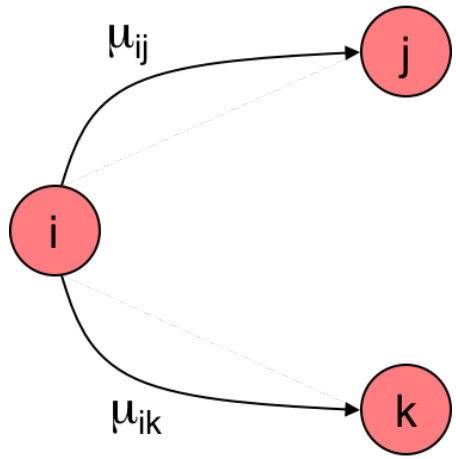
$$P(T > s + t | T > s) = P(T > t)$$

Travail à la maison

- Lire le chapitre 3.2 jusqu'à la page 75
- Exercices 9, 10, 11

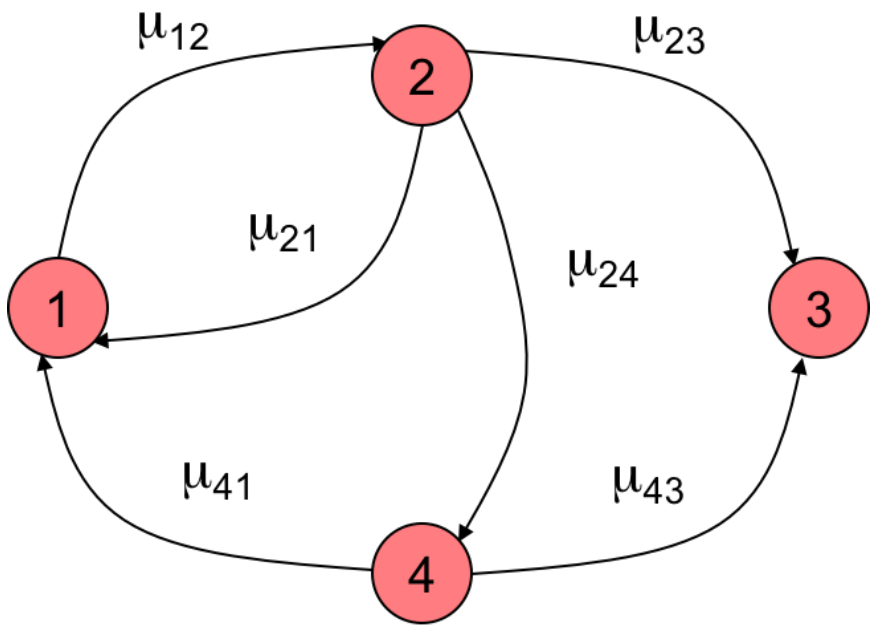
2ème caractérisation d'une CMTC homogène (la plus couramment utilisée)

- $T_{ij} \sim \exp(\mu_{ij})$: temps au bout duquel on passe de l'état i à l'état j



- Si $T_{ij} < T_{ik}$ pour tout $k \neq j$, aller dans l'état j

2ème caractérisation d'une CMTC homogène

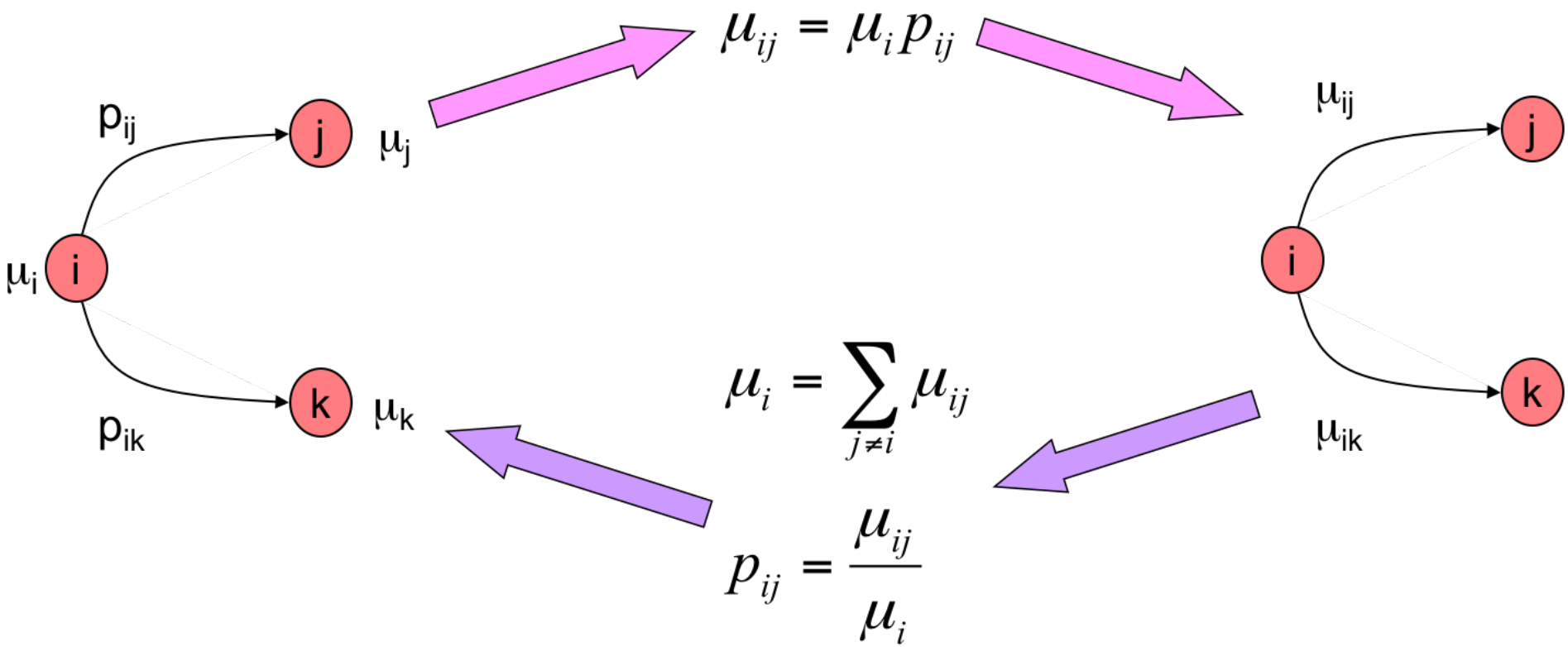


- μ_{ij} : taux de transition de i vers j

$$\mu_{ij} = \frac{1}{E(T_{ij})}$$

- Unité des taux de transition : $[s^{-1}]$

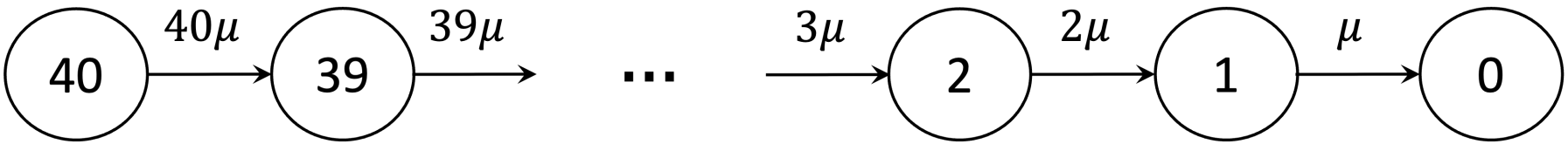
Equivalence entre les 2 caractérisations



Exemple : Nombre d'étudiants dans la salle

- Hypothèses et notations
  - $N(t)$: nombre d'étudiants dans la salle à l'instant t
  - $N(0) = 40$
  - Le i-ième étudiant de la salle sort après un temps T_i exponentiel de taux μ
  - Les dates de départ $T_1, \cdots, T_{40}$ sont supposées indépendantes les unes des autres
- Pourquoi $N(t)$ est une CMTC ?
- Donner son graphe pour la 2-ème caractérisation.

Exemple : Nombre d'étudiants dans la salle

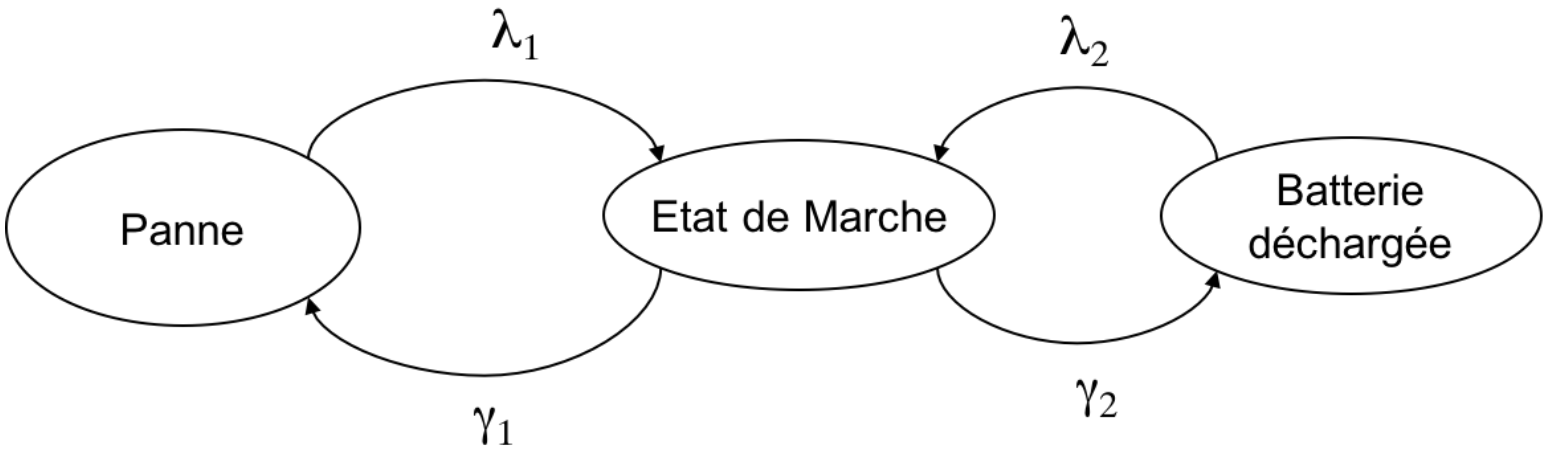


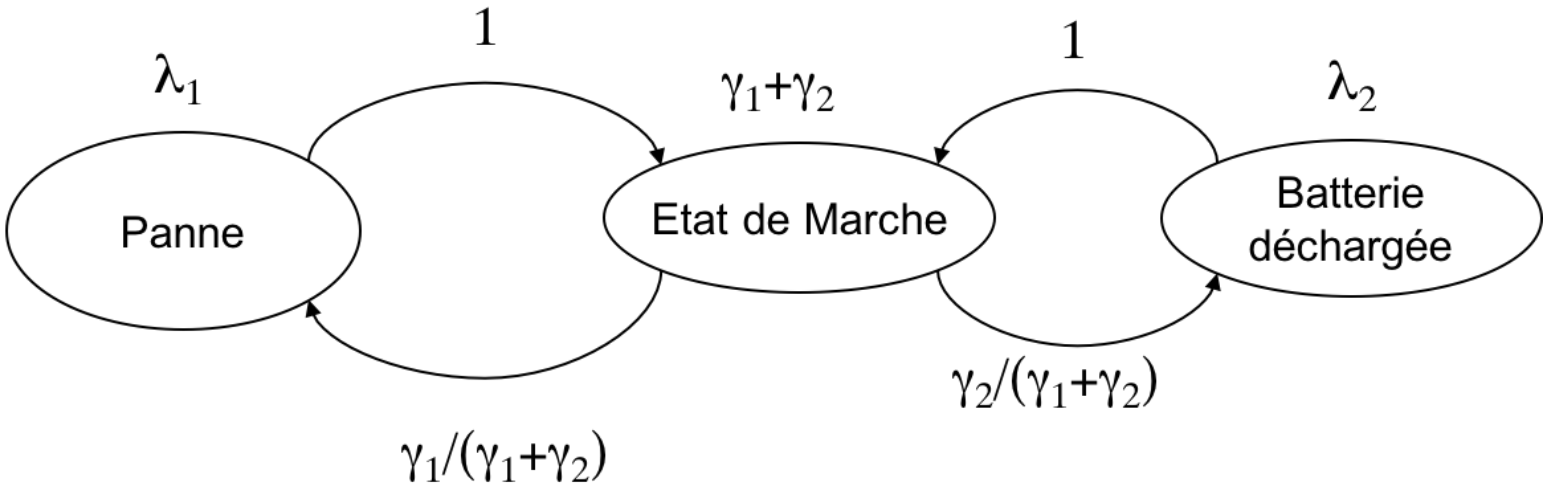
Exemple : téléphone portable

- Panne après un temps exponentiel de taux γ_1
- Réparation après un temps exponentiel de taux λ_1
- Déchargement de batterie après un temps exponentiel de taux γ_2
- Chargement après un temps exponentiel de taux λ_2

- Donner son graphe (suivant les 2 caractérisations)

Exemple : téléphone portable





Analyse en régime transitoire

- $t \in \mathbb{R}^+$
- $\pi_i(t)$: probabilité d'être dans l'état i à l'instant t

$$\pi(t) = (\pi_1(t), \pi_2(t), \cdots)$$

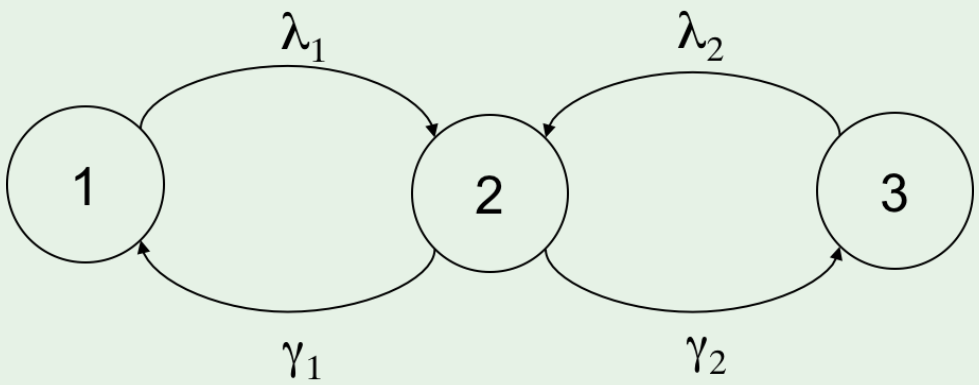
Générateur infinitésimal

- Soit une CMTC de taux de transition (μ_{ij})
- La matrice $Q = (q_{ij})$ avec

$$\begin{cases} q_{ij} & = \mu_{ij} \text{ si } i \neq j \\ q_{ii} & = -\mu_i = -\sum_{j \neq i} \mu_{ij} \end{cases}$$

est appelé le générateur infinitésimal de la CMTC

Exemple : Téléphone portable



$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 \\ \gamma_1 & -(\gamma_1 + \gamma_2) & \gamma_2 \\ 0 & \lambda_2 & -\lambda_2 \end{pmatrix}$$

Equations en régime transitoire

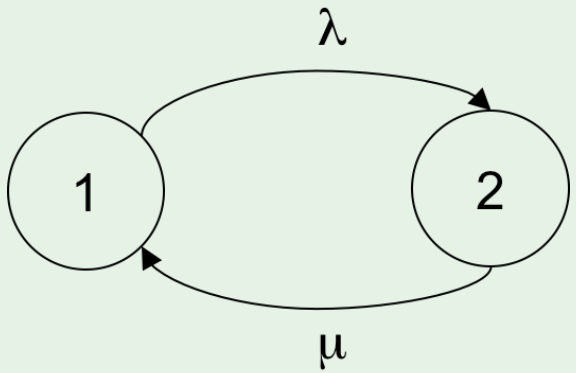
- Cf poly pour la démonstration :

$$\forall j \in E, \quad \frac{d\pi_j(t)}{dt} = \sum_i q_{ij}\pi_i(t)$$

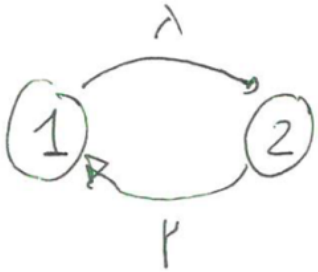
- Sous forme matricielle

$$\frac{d\pi(t)}{dt} = \pi(t)Q$$

Exemple



Exprimer $\pi_1(t)$ et $\pi_2(t)$ en fonction de λ, μ et t



$\pi_1(0) = 0$
$\pi_2(0) = 1$

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

$$\begin{cases} \pi_1'(t) = -\lambda \pi_1(t) + \mu \pi_2(t) \\ \pi_2'(t) = \lambda \pi_1(t) - \mu \pi_2(t) \\ \pi_1(t) + \pi_2(t) = 1 \end{cases}$$

$$\Rightarrow \quad \pi_1'(t) = -\lambda \pi_1(t) + \mu(1 - \pi_2(t)) \iff \pi_1'(t) + (\lambda+\mu)\pi_1(t) = \mu$$

Solution sans 2nd membre = ~~$\pi_1(t)$~~ $= A e^{-(\lambda+\mu)t}$

Sol. part. avec ——— $= \dfrac{\mu}{\lambda+\mu}$

$$\pi_1(t) = A e^{-(\lambda+\mu)t} + \frac{\mu}{\lambda+\mu}$$

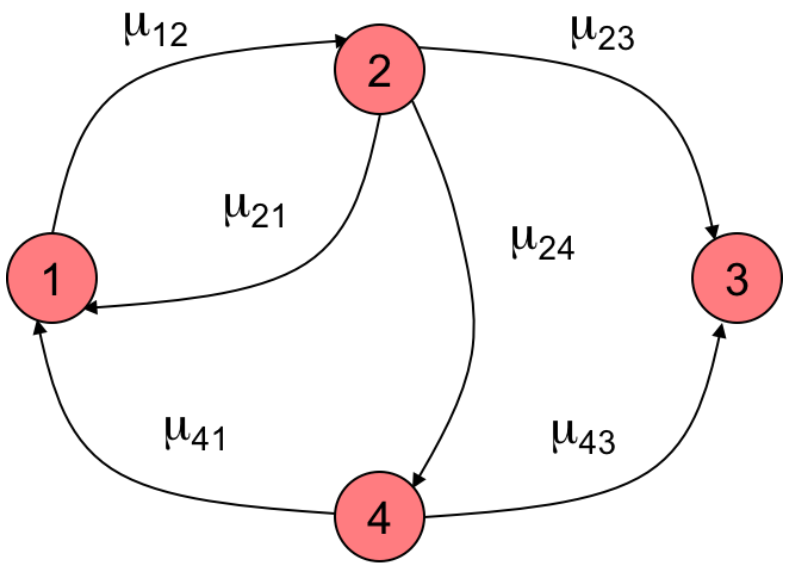
$$+ \pi_1(0) = 0$$

D'où
$$\begin{cases} \pi_1(t) = \dfrac{\mu}{\lambda+\mu}\left(1 - e^{-(\mu+\lambda)t}\right) \xrightarrow[t\to+\infty]{} \dfrac{\mu}{\lambda+\mu} \\ \pi_2(t) = \dfrac{\lambda + \mu e^{-(\mu+\lambda)t}}{\lambda+\mu} \xrightarrow[t\to+\infty]{} \dfrac{\lambda}{\lambda+\mu} \end{cases}$$

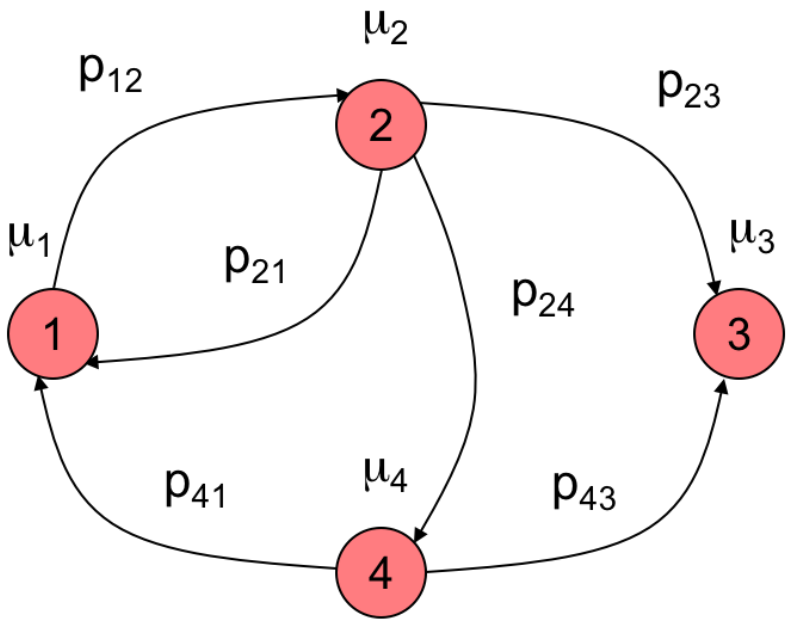
Travail à la maison

- Polycopié : lire jusqu'à la page 81
- Exercices 12 et 13

2ème caractérisation d'une CMTC (rappel)

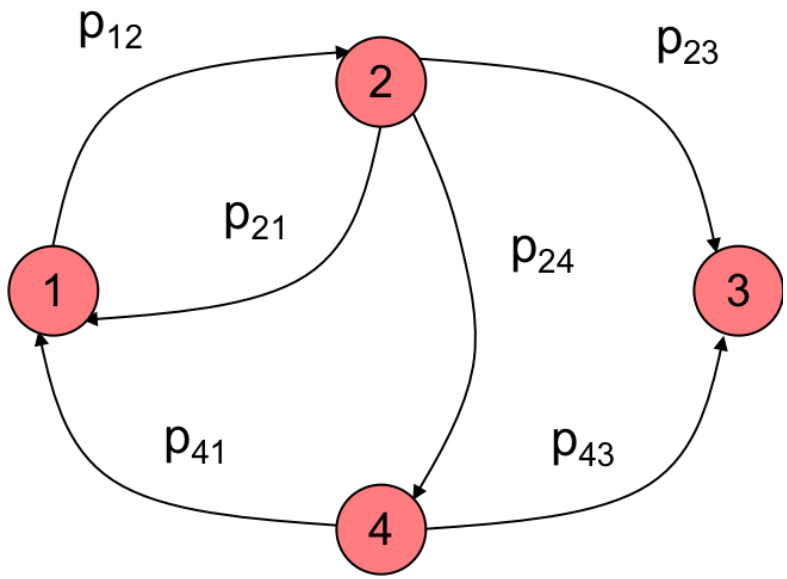


1ère caractérisation d'une CMTC (rappel)



CMTD incluse associée à une CMTC

- CMTD incluse : CMTD de matrice de transition (p_{ij})



Classification des états d'une CMTC

- La nature d'un état est défini de la même manière que pour une CMTD
  - $f_{jj} < 1$: état transitoire
  - $f_{jj} = 1$, $M_j < \infty$: état récurrent non nul
  - $f_{jj} = 1$, $M_j = \infty$: état récurrent nul
- Pas de notion de périodicité dans une CMTC
- Une CMTC est dite irréductible si sa CMTD incluse est irréductible

Proposition

- *Un état i d'une CMTC est transitoire (respectivement récurrent) ssi l'état i de la CMTD incluse est transitoire (respectivement récurrent)*
- *Les états d'une CMTC irréductible sont de même nature : transitoires / récurrents nuls / récurrents non nuls*
- *Les états d'une CMTC finie et irréductible sont récurrents non nuls*

Distribution limite

- $\pi(t) = (\pi_1(t), \pi_2(t), \cdots)$
- Equations en régime transitoire pour une CMTC finie

$$\frac{d\pi(t)}{dt} = \pi(t)Q$$

- Supposons que $\pi(t)$ converge vers π quand t tend vers l'infini, alors

$$\pi Q = 0$$

Distribution limite pour une CMTC irréductible

Théorème (admis)

Pour une CMTC irréductible , la distribution limite $\pi = (\pi_1, \pi_2, \cdots)$ *existe et est indépendante de la distribution initiale* $\pi(0)$ *:*

- *Si les états sont tous transitoires ou tous récurrents nuls, alors*

$$\forall j \in E, \pi_j = 0$$

- *Si les états sont tous récurrents non nuls, alors les* π_j *satisfont*

$$(S) : \begin{cases} \sum_{i \in E} \pi_i q_{ij} = 0, \forall j \in E \\ \sum_{j \in E} \pi_j = 1 \end{cases}$$

- Pour une CMTC finie

$$\sum_{i \in E} \pi_i q_{ij} = 0 \Leftrightarrow \pi Q = 0$$

Flux sortant d'un état = flux entrant dans cet état

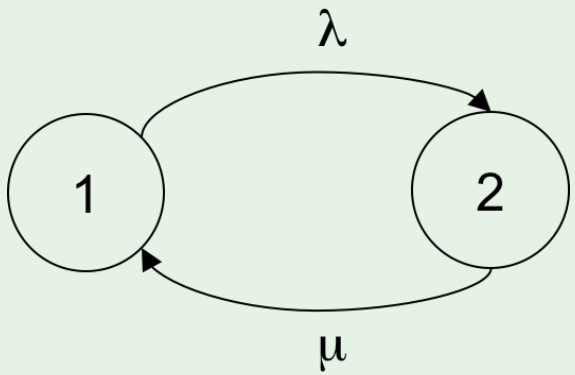
- L'équation $\sum_{i \in E} \pi_i q_{ij} = 0$ peut s'écrire aussi

$$\overbrace{\pi_i \sum_{j \neq i} \mu_{ij}}^{\text{Flux sortant de } i} = \overbrace{\sum_{j \neq i} \mu_{ji} \pi_j}^{\text{Flux entrant en } i}$$

- $\mu_{ji} \pi_j$: nombre de transitions par unité de temps de j vers i

Exemple : Marche / Panne

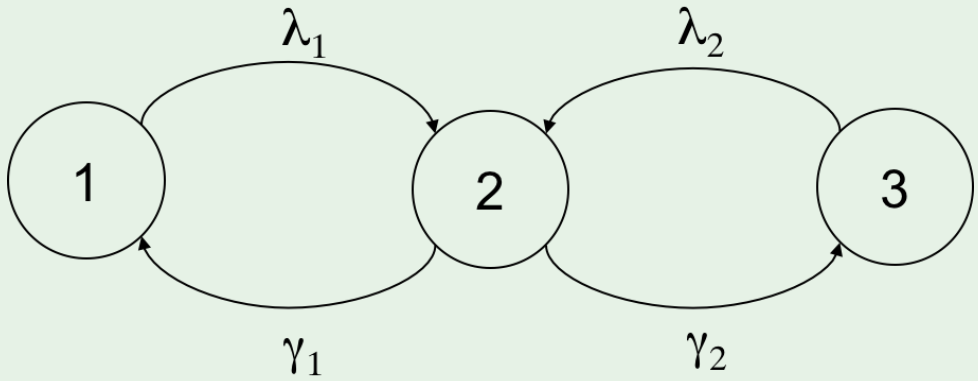
- Un appareil tombe en panne après un temps $\sim \exp(\lambda)$
- En cas de panne, l'appareil est réparé au bout d'un temps $\sim \exp(\mu)$
- Quelle est la probabilité que l'appareil fonctionne ?



$$(S) : \begin{cases} \lambda\pi_1 = \mu\pi_2 & \text{(flux sortant de 1 = flux entrant en 1)} \\ \mu\pi_2 = \lambda\pi_1 & \text{(flux sortant de 2 = flux entrant en 2)} \\ \pi_1 + \pi_2 = 1 \end{cases}$$

$$\Rightarrow \pi_1 = \frac{\mu}{\lambda + \mu}, \pi_2 = \frac{\lambda}{\lambda + \mu}$$

Exemple : Téléphone portable



Déterminer π_1, π_2 et π_3.
Il faut résoudre le système linéaire suivant :

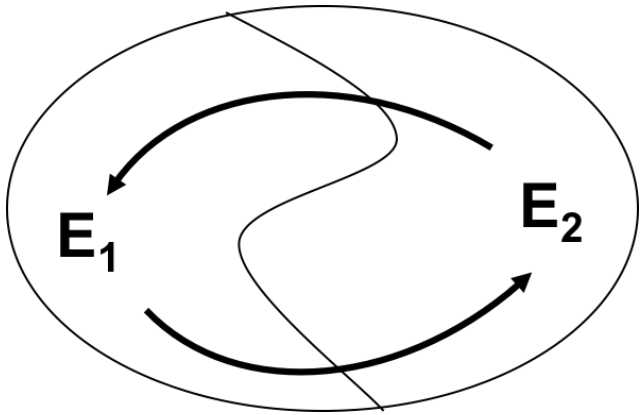
$$(S) : \begin{cases} \lambda_1\pi_1 = \gamma_1\pi_2 & \text{(flux sortant de 1 = flux entrant en 1)} \\ (\gamma_1 + \gamma_2)\pi_2 = \lambda\pi_1 + \lambda_2\pi_3 & \text{(flux sortant de 2 = flux entrant en 2)} \\ \lambda_2\pi_3 = \gamma_2\pi_2 & \text{(flux sortant de 3 = flux entrant en 3)} \\ \pi_1 + \pi_2 + \pi_3 = 1 \end{cases}$$

$$\Rightarrow \cdots \Rightarrow \pi_1 = ?, \pi_2 = ?, \pi_3 = ?$$

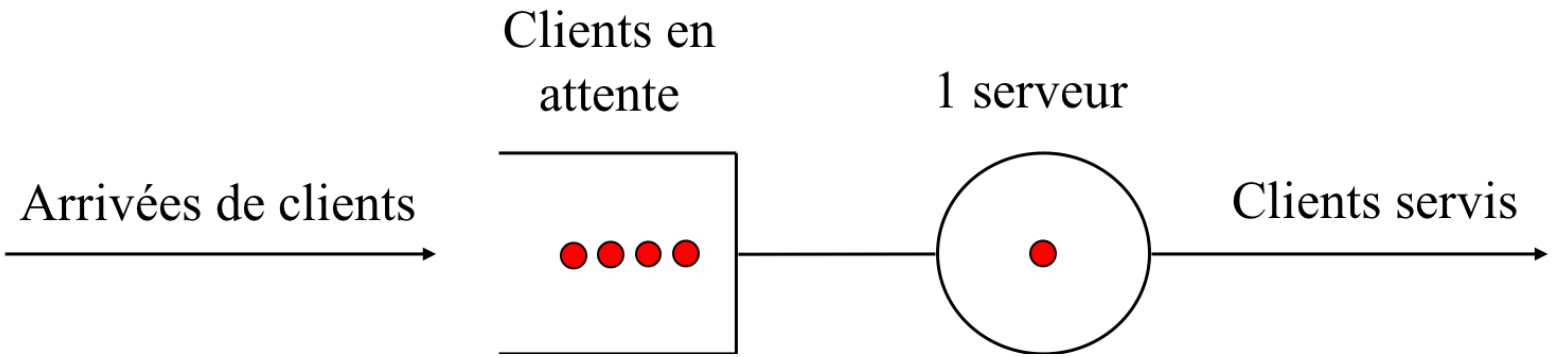
Conservation des flux

- Partition des états : $E = E_1 \cup E_2$

$$\overbrace{\sum_{i \in E_1} \sum_{j \in E_2} \pi_i \mu_{ij}}^{\text{Flux de } E_1 \text{ vers } E_2} = \overbrace{\sum_{i \in E_2} \sum_{j \in E_1} \pi_i \mu_{ij}}^{\text{Flux de } E_2 \text{ vers } E_1}$$



Exemple : file d'attente $M/M/1$



- T_s : temps pour servir un client

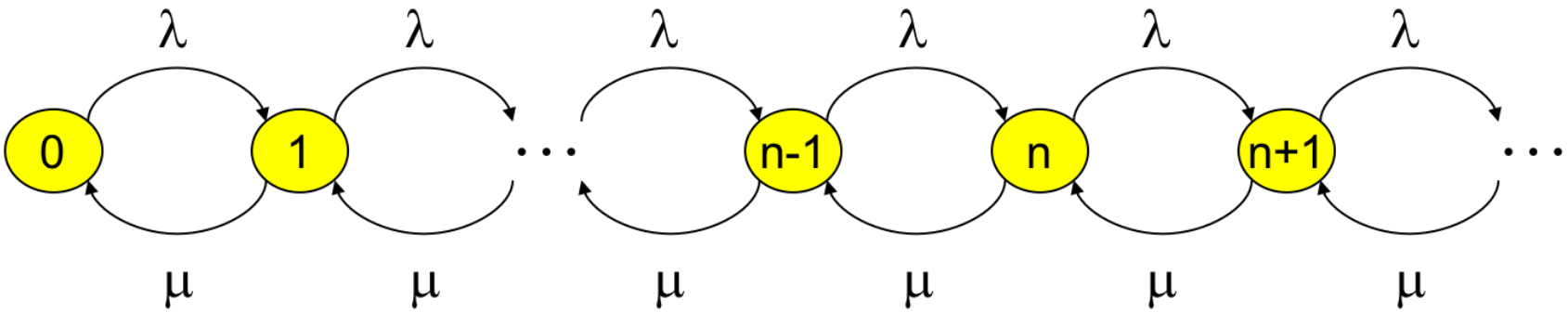
$$T_s \sim exp(\mu)$$

- T_a : temps entre 2 arrivées de clients

$$T_a \sim exp(\lambda)$$

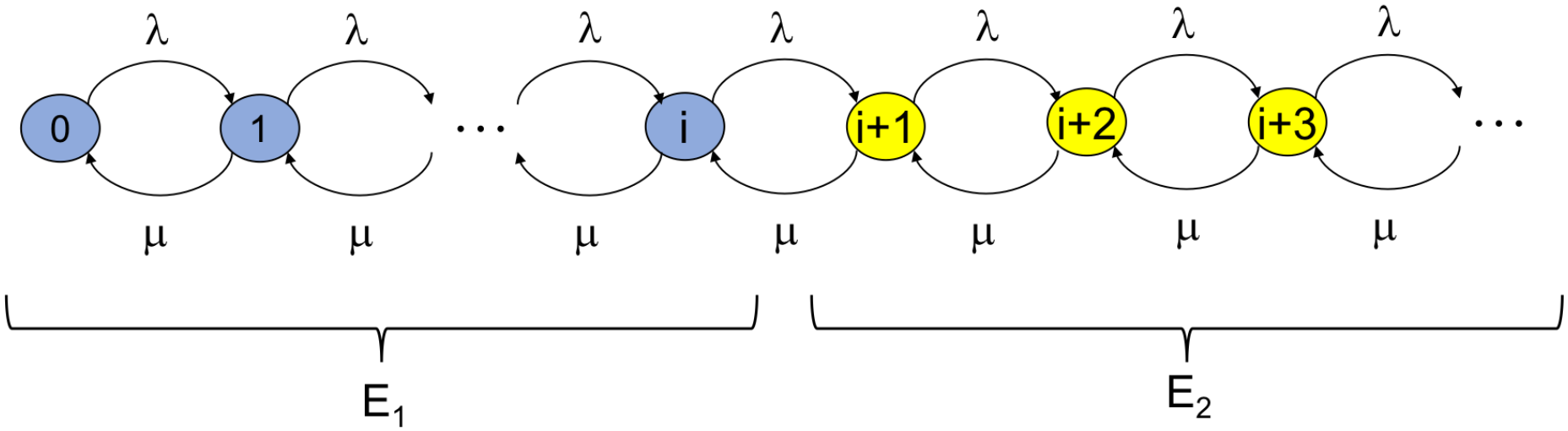
- $N(t)$: nombre de clients dans le système (en attente + en service)

$N(t)$ est une CMTC



- Cette CMTC est irréductible
  - $\mu < \lambda$: états tous transitoires et les $\pi_i = 0$
  - $\mu = \lambda$: états tous récurrents nuls et les $\pi_i = 0$
  - $\mu > \lambda$: états tous récurrents non nuls et les et π_i sont solution de (S)

Distribution limite pour $\mu > \lambda$



- Flux de E_1 vers E_2 = Flux de E_2 vers E_1 $\Rightarrow$ $\lambda \pi_i = \mu \pi_{i+1}$
- Soit $\rho = \dfrac{\lambda}{\mu}$ le taux d'utilisation de la file d'attente. Alors

$$\pi_i = \rho^i \pi_0$$

- Si $\rho < 1$, alors $\sum_{i=0}^{+\infty} \pi_i = 1$ donne $\pi_0 = 1 - \rho$ puis

$$\pi_i = (1 - \rho)\rho^i$$

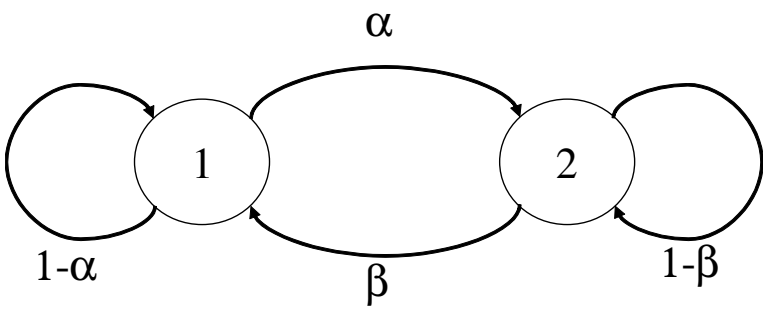
Travail à la maison

- Finir chapitre 3.2 sur les CMTC
- Exercices 14 et 15
- TP file d'attente

Récapitulatif

	CMTD	CMTC
Transitions	p_{ij}	μ_{ij}
Régime transitoire	$\pi(n+1) = \pi(n)P$	$\frac{d\pi(t)}{dt} = \pi Q$
Régime permanent	$\pi = \pi P$	$0 = \pi Q$
Flux de i vers j	$\pi_i p_{ij}$	$\pi_i \mu_{ij}$
Temps passé dans un état	géométrique	exponentiel

Convergence rate : An exemple (facultatif)



$$\pi(0) = (1, 0), P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

- Give π : $\pi_1 = \dfrac{\beta}{\alpha+\beta}$
- Give $\pi_1(n)$: $\pi_1(n) = \dfrac{\beta}{\alpha+\beta} + \dfrac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n$
- Give $|\pi_1(n) - \pi_1|$: $|\pi_1(n) - \pi_1| = \dfrac{\alpha}{\alpha+\beta}|1-\alpha-\beta|^n$
- Study the rate of convergence : converge geometrically but very slowly when $\alpha \simeq \beta \simeq 1$ (quasi-periodic) or if $\alpha \simeq \beta \simeq 0$ (quasi not connected).
- Compute eigenvalues of P : 1 and $1-\alpha-\beta$ ($\lambda_1 = 1$ and $\lambda_1 + \lambda_2 = Tr(P) = 2-\alpha-\beta$).

Upper bound on the convergence rate (facultatif)

- Eingenvalues of P : $\lambda_1 = 1 \geq |\lambda_2| \geq \cdots \geq |\lambda_{\mathcal{S}}|$
- For an aperiodic and irreducible Markov chain

$$|\lambda_2| < 1$$

  - The convergence to the stationary distribution is geometric in the second eigenvalue. There exists $C > 0$ such that

$$|\pi_j(n) - \pi_j| \leq C|\lambda_2|^n$$

- Previous result is trivial when P is diagonalisable : $P^k = AD^kA^{-1}$

Convergence rate : Birth death process (facultatif)

$p = 1 - q$, $|\mathcal{S}|$ states

$$P = \begin{pmatrix} q & p & 0 & \dots & 0 \\ q & 0 & p & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & q & 0 & p \\ 0 & \dots & 0 & q & p \end{pmatrix}$$

$$\lambda_1 = 1, \lambda_j = 2\sqrt{pq}\cos\left(\frac{(j-1)\pi}{|\mathcal{S}|}\right)$$

λ_2 is maximal for $p = q = 0.5$ and goes to 1 when $|\mathcal{S}|$ goes to infinity.
Will be an issue for queuing systems close to instability.