MAPSI — cours 6 : Chaîne de Markov

Nicolas Thome
nicolas.thome@isir.upmc.fr

LIP6 / ISIR – Sorbonne Université, France

Introduction

- Les problèmes traités jusqu'ici :

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}, \text{ et parfois : } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

- Chaque individu $\mathbf{x} = [x_1, x_2, \cdots, x_d]$ est un vecteur i.i.d.
- **Les séquences** ne rentrent pas dans ce cadre

Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

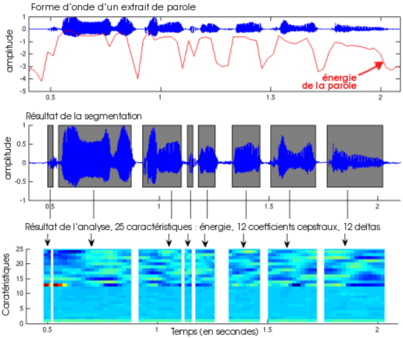
- Reconnaissance de chaine de caractères

Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

- Reconnaissance de paroles

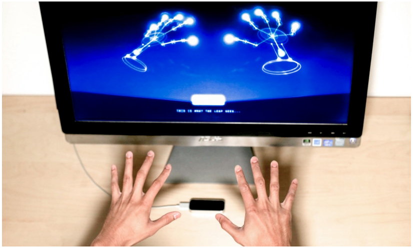


Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

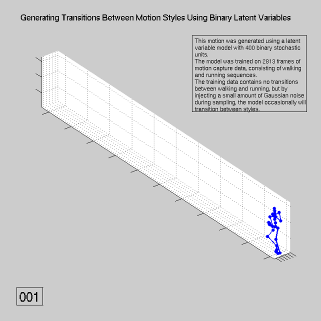
- Reconnaissance de mouvements



Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

- Génération de mouvements

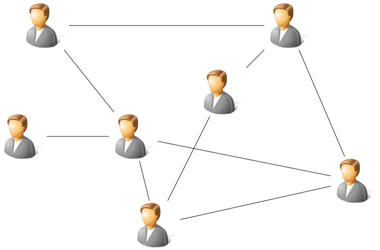


Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

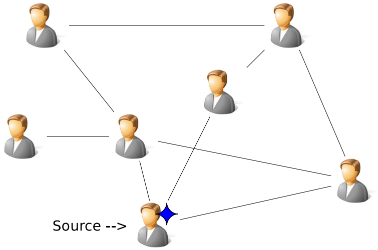
- Diffusion dans les graphes



Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

- Diffusion dans les graphes

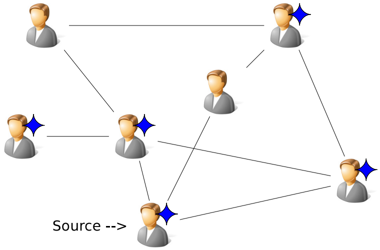


Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion

- Diffusion dans les graphes

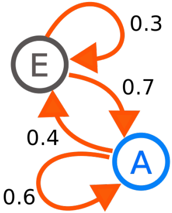


Problème

- Méthodes standards de classification/clustering ⇒ transition difficile vers les **données de taille variable**
- ⇒ modèles génératifs de séquences (de taille variable)

Approche vectorielle : $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}$

Approche séquentielle :



Modèles génératifs

Nicolas Thome
nicolas.thome@isir.upmc.fr

LIP6 / ISIR – Sorbonne Université, France

Rappel sur les modèles génératifs

1. Choix d'une modélisation des données : $p(\mathbf{x}|\theta)$
2. Apprentissage = trouver θ
3. Application possible : décision bayesienne

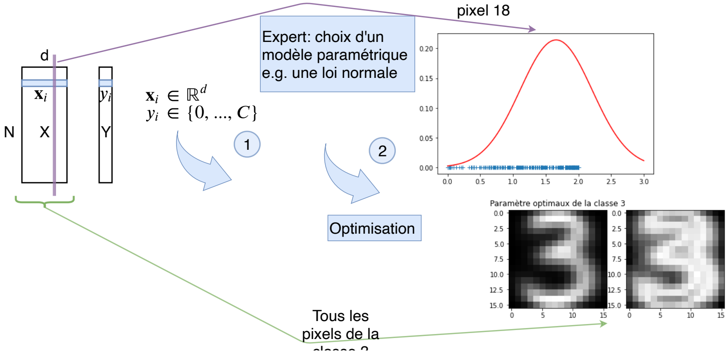
$$r(\mathbf{x}) = \arg\max_k p(\theta_k|\mathbf{x}) = \frac{p(\mathbf{x}|\theta_k)p(\theta_k)}{p(\mathbf{x})}$$

4. Application bis : génération de $\tilde{\mathbf{x}} \sim \mathcal{D}(\theta_k)$

Apprentissage d'un modèle génératif $\Leftrightarrow$ Estimation de densité

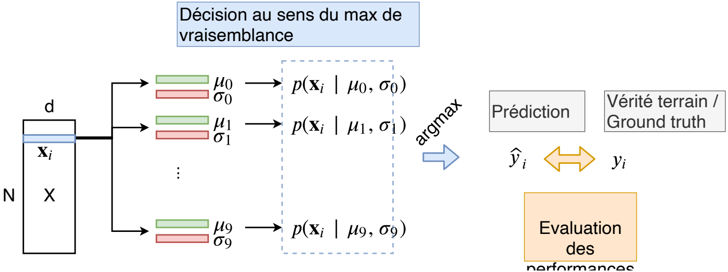
- Estimer θ_k = estimer une densité de probabilité d'une classe
- Hypothèse (forte) : les θ_k sont supposés indépendants
- Techniques d'estimation des θ_k

1. Schéma général du maximum de vraisemblance :



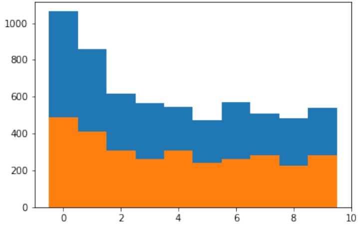
2. Inférence au sens du max de vraisemblance
3. Prise en compte des a priori
4. Génération

1. Schéma général du maximum de vraisemblance :
2. Inférence au sens du max de vraisemblance



3. Prise en compte des a priori
4. Génération

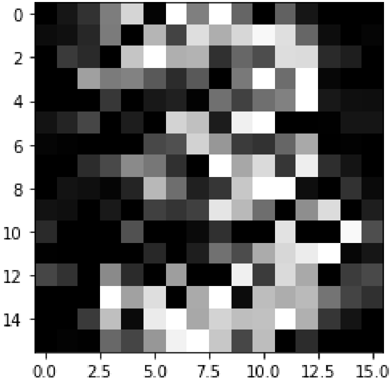
1. Schéma général du maximum de vraisemblance :
2. Inférence au sens du max de vraisemblance
3. Prise en compte des a priori



4. Génération

Retour sur USPS

1. Schéma général du maximum de vraisemblance :
2. Inférence au sens du max de vraisemblance
3. Prise en compte des a priori
4. Génération



Maximum de vraisemblance (données vec.)

- $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ exemples supposés générés par $p(\mathbf{x}|\theta)$
- Adéquation entre les données et le modèle
  - Notion de vraisemblance des observations
  - Hyp : les données sont indépendantes

$$\mathcal{L}(D, \theta) = p(D|\theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)$$

- Optimisation :

$$\theta^\star = \arg\max_\theta(\mathcal{L}(D, \theta)) = \arg\max_\theta(\log \mathcal{L}(D, \theta))$$

- Résolution :
  - Analytique : $\dfrac{\partial \mathcal{L}(D, \theta)}{\partial \theta} = 0$
  - Approchée : EM, gradient...

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \ldots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

Traitement des séquences

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \dots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance : $\log \mathcal{L} = \sum_i \log\left(p(\mathbf{x}_i|\lambda)\right)$ [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

Traitement des séquences

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \ldots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance : $\log \mathcal{L} = \sum_i \log\left(p(\mathbf{x}_i|\lambda)\right)$ [classique]
    - Apprentissage : $\lambda^\star = \text{Argmax}_\lambda \log \mathcal{L}$ [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

Traitement des séquences

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \dots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance : $\log \mathcal{L} = \sum_i \log\left(p(\mathbf{x}_i|\lambda)\right)$ [classique]
    - Apprentissage : $\lambda^\star = \text{Argmax}_\lambda \log \mathcal{L}$ [classique]
    - Inférence : $c^\star = \text{Argmax}_c\, p(\mathbf{x}_i|\lambda_c^\star)$ [classique]
  - Catégorisation (non-supervisée)

  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

Traitement des séquences

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \dots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance : $\log \mathcal{L} = \sum_i \log(p(\mathbf{x}_i|\lambda))$ [classique]
    - Apprentissage : $\lambda^\star = \text{Argmax}_\lambda \log \mathcal{L}$ [classique]
    - Inférence : $c^\star = \text{Argmax}_c\, p(\mathbf{x}_i|\lambda_c^\star)$ [classique]
  - Catégorisation (non-supervisée)
    - ... idem + bouclage EM [classique]
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

Traitement des séquences

- Modèle du jour = modèle de Markov
- Entrée = $\mathbf{x} = \{x_1, \dots, x_T\}$
  - Séquence d'observations, taille variable
  - Dépendance entre x_t et x_{t+1}
  - Hypothèse i.i.d. entre les $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle $\lambda = \{\Pi, A\}$ [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance : $\log \mathcal{L} = \sum_i \log (p(\mathbf{x}_i|\lambda))$ [classique]
    - Apprentissage : $\lambda^\star = \text{Argmax}_\lambda \log \mathcal{L}$ [classique]
    - Inférence : $c^\star = \text{Argmax}_c\, p(\mathbf{x}_i|\lambda_c^\star)$ [classique]
  - Catégorisation (non-supervisée)
    - ... idem + bouclage EM [classique]
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...
    - ... prochaine séance, modèle plus compliqué

- Outil pour faire de la prévision dans des espaces discrets
- Chaîne de Markov d'ordre k
  - Séquence de variables aléatoires $\mathbf{x} = (x_1, \ldots, x_T)$
    qui prend ses valeurs dans un ensemble fini d'états $Q = (q_1, ..., q_N)$
    et qui vérifie les propriétés dites de Markov :
    - **Chaîne de Markov d'ordre** k (horizon) :

$$p(x_{t+1} = q_j | x_1, \ldots, x_t) = p(x_{t+1} = q_j | x_{t-k+1}, \ldots, x_t)$$

  - Dans la suite : chaînes d'ordre 1.
- Exemple : météo sur un an (soleil, nuage, pluie)
  - $Q = [So, Nu, Pl]$
  - $\mathbf{x} = [x_1 = Nu, x_2 = So, \ldots, x_{365} = Pl]$

Soit un exemple très simple :

- $Q = [So, Nu, Pl]$
- $\mathbf{x} = [x_1 = Nu, x_2 = So, \ldots, x_{365} = Pl]$
- Combien de paramètres pour une chaine d'ordre 1 ?

Ordre des CM et complexité

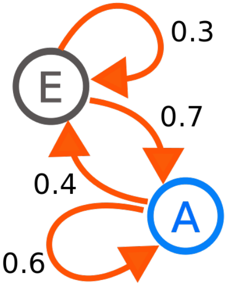
Soit un exemple très simple :

- $Q = [So, Nu, Pl]$
- $\mathbf{x} = [x_1 = Nu, x_2 = So, \ldots, x_{365} = Pl]$
- Combien de paramètres pour une chaine d'ordre 1 ?

  9 (+3)
- Combien de paramètres pour une chaine d'ordre 2 ?

  27 (+3)

- Une chaîne de Markov d'ordre 1 est entièrement spécifiée par la donnée :



  - d'une matrice de transition

$$\mathbf{A} = [a_{ij} = p(x_{t+1} = q_j | x_t = q_i)]$$

  - et des probabilités initiales :

$$\Pi = [\pi_i = p(x_1 = q_i)]$$

- Probabilité d'une séquence

$$p(\mathbf{x}|\lambda) = p(x_1, \ldots, x_T|\lambda)$$

Hypothèse markovienne d'ordre 1

Décomposition du calcul

$$\begin{aligned}
p(\mathbf{x}|\lambda) &= p(x_1, \ldots, x_T|\lambda) \\
&= p(x_T|x_1, \ldots, x_{T-1}, \lambda) \times p(x_1, \ldots, x_{T-1}|\lambda) \\
&= p(x_T|x_1, \ldots, x_{T-1}, \lambda) \times p(x_{T-1}|x_1, \ldots, x_{T-2}, \lambda) \ldots \\
&\quad \times p(x_1, \ldots, x_{T-2}|\lambda) \\
&= \prod_{t=2}^{T} p(x_t|x_1, \ldots, x_{t-1}, \lambda) p(x_1|\lambda)
\end{aligned}$$

Après hypothèse d'ordre 1 :

$$\begin{aligned}
p(\mathbf{x}|\lambda) &= \prod_{t=2}^{T} p(x_t|x_1, \ldots, x_{t-1}, \lambda) p(x_1|\lambda) = \prod_{t=2}^{T} p(x_t|x_{t-1}, \lambda) p(x_1|\lambda) \\
&= \pi_{x_1} \prod_{t=2}^{T} a_{x_{t-1}, x_t}
\end{aligned}$$

Représentation matricielle

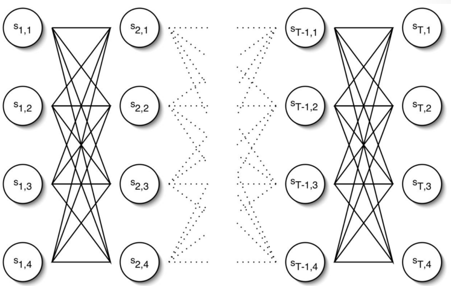
- Matrice de transition $\mathbf{A}$ de taille $N \times N$
  - Lignes $\mathbf{x}_t$, colonnes $\mathbf{x}_{t+1}$
  - Matrice stochastique, chaque ligne somme à 1 : $\sum_{j=1}^{N} a_{ij} = 1$

Représentation matricielle

- Matrice de transition $\mathbf{A}$ de taille $N \times N$
  - Lignes $\mathbf{x}_t$, colonnes $\mathbf{x}_{t+1}$
  - Matrice stochastique, chaque ligne somme à 1
- $\mathbf{p}_t := p(x_t = q_i | x_1, \ldots, x_{t-1}, \lambda), \forall i \in \{1; N\}$
  - $\mathbf{p}_t$ vecteur ligne, taille $1 \times N$
- $\mathbf{p}_{t+1} := p(x_{t+1} = q_i | x_1, \ldots, x_t, \lambda), \forall i \in \{1; N\}$
  - $\mathbf{p}_{t+1}$ vecteur ligne, taille $1 \times N$

- Matrice de transition $\mathbf{A}$ de taille $N \times N$
- $\mathbf{p}_t$ / $\mathbf{p}_{t+1}$ vecteurs de probabilté à t / $t+1$
- Treillis pour calculer le vecteur de probabilité $\mathbf{p}_{t+1}$ $\forall j \in \{1; N\}$
  - $p(x_{t+1} = q_j) = \sum_{i=1}^{N} p(x_t = q_i) p(x_{t+1} = q_j | x_t = q_i) = \sum_{i=1}^{N} p(x_t = q_i)\ a_{ij}$

$$\mathbf{p}_{t+1} = \mathbf{p}_t \times \mathbf{A}$$



Il fait soleil... Quel temps fera-t-il dans N jours ? (distribution de probabilités)

Représentation matricielle : exemple

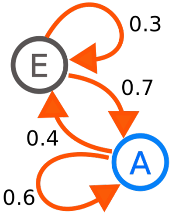
Il fait soleil... Quel temps fera-t-il dans N jours ? (distribution de probabilités)

1. Je rentre sur la ligne *soleil*... $x_0 = S$
2. A $t = 1$, $\{a_{x \cdot}\}$ me donne la distribution des probabilités des états $p(x_1 = q_i) = a_{S,i} \Rightarrow \mathbf{p}_1 = \left\{a_{S,i}\right\}_{i \in \{1;N\}}$

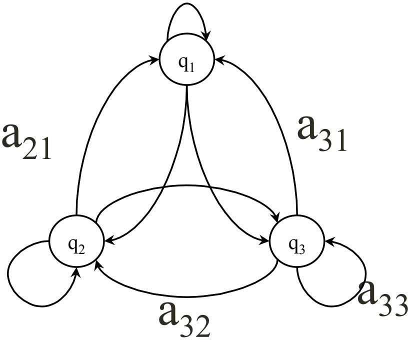
Il fait soleil... Quel temps fera-t-il dans N jours ? (distribution de probabilités)

1. Je rentre sur la ligne *soleil*... $x_0 = S$
2. A $t = 1$, $\{a_{x\cdot}\}$ me donne la distribution des probabilités des états $p(x_1 = q_i) = a_{S,i} \Rightarrow \mathbf{p}_1 = \left\{a_{S,i}\right\}_{i\in\{1;N\}}$
3. Ensuite on itère :
   - $\mathbf{p}_2 = \mathbf{p}_1 \times \mathbf{A}$
   - $\mathbf{p}_N = \mathbf{p}_1 \times \mathbf{A}^{N-1}$

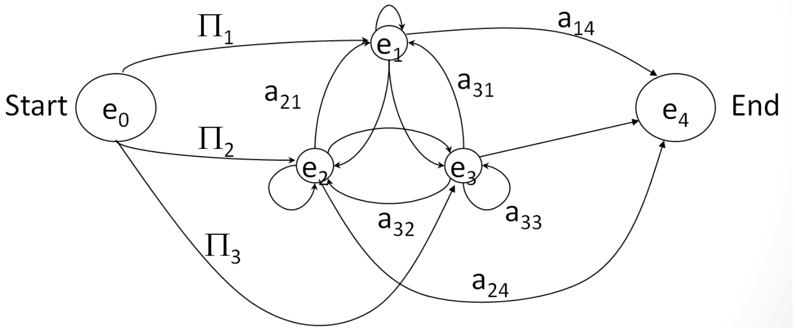
Automate basique :



Comment introduire la notion de commencement ?



Avec des noeuds identifiés de début/fin



Algorithme génaratif

Algorithm 1: Génération d'une séquence $\mathbf{x}$

Data: A, Π
Result: $\mathbf{x}$
$\mathbf{x} \leftarrow []$;
Tirer x_1 en fonction de Π;
$x_t \leftarrow x_1$, $t \leftarrow 1$;
$\mathbf{x} \leftarrow [\mathbf{x},\ x_{courant}]$;
while x_t *n'est pas l'état final* **do**
    $x_{t+1} \leftarrow$tirage selon $(A(x_t, :))$;
    $t \leftarrow t + 1$;

- Plusieurs variantes dans la clause du *while*
- Comment effectuer un tirage selon une loi de probabilité discrète ?

Outil pour le tirage aléatoire selon une loi discrète

Soit la loi :

A	1	2	3
$P(A)$	0.3	0.2	0.5

Comment effectuer un tirage selon $P(A)$?

1. Faire la somme cumulée de la loi

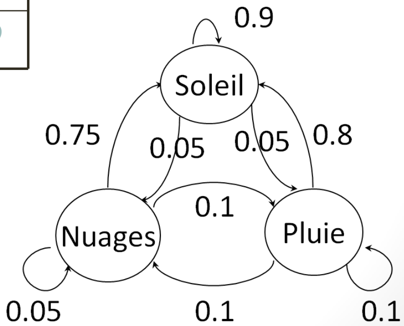
A	1	2	3
$cumsum$	0.3	0.5	1

2. Tirer un nombre x entre 0 et 1 selon la loi uniforme
3. Initialiser $vx = 1$
4. Tant que $cumsum[vx] < x$
   - $vx + +$

Génération (sur un exemple)

- q_1 = Pluie, q_2 = Nuages, q_3 = Soleil
- A =

0.1	0.1	0.8
0.1	0.15	0.75
0.05	0.05	0.9



PSSSSSSSSSSSSNSSSSSSSSSSNSSSSSSSNSSSSNSS

SSSSSSSSSSSSSSPSSSSSSSNPSSSSSPNSSSSSSNSSS

SSSSPSSSSSSSSSSPSSSSSSSSSSSSSSPSSSSSSSSS

Problèmes alternatifs ⇒ analyses du modèle

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la probabilité d'observer une séquence de soleil de longueur d (étant donné que l'on se trouve au premier jour de soleil) ?
Loi géométrique
Notons la longueur de la sous-séquence de soleil D_S,

$$P(D_S = d) = a_{ss}^{d-1}(1 - a_{ss})$$

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?
Espérance de la loi géométrique

$$E[D_S] = \sum_{d=1}^{\infty} d \; a_{ss}^{d-1}(1 - a_{ss}) = \frac{1}{1 - a_{ss}}$$

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?
Espérance de la loi géométrique

$$E[D_S] = \sum_{d=1}^{\infty} d\; a_{ss}^{d-1}(1-a_{ss}) = \frac{1}{1-a_{ss}}$$

Sketch of proof (voir TD) avec $k = d-1$ et $p = 1 - a_{ss}$:

$$\begin{aligned}
\mathrm{E}(Y) &= \textstyle\sum_{k=0}^{\infty}(1-p)^k p \cdot (k+1) \\
&= p \textstyle\sum_{k=0}^{\infty}(1-p)^k (k+1) \\
&= p \left[\frac{d}{dp} \left(-(1-p) \textstyle\sum_{k=0}^{\infty}(1-p)^k \right) \right] \\
&= -p \frac{d}{dp} \frac{1-p}{p} = \frac{1}{p}.
\end{aligned}$$

Stationnarité : CM converge-t-elle $\rightarrow$ mesure stationnaire $\boldsymbol{\mu}$?

- **Point fixe**, proba des états ne change plus au cours du temps
  - $\boldsymbol{\mu}$ = **pondération stationnaire** (inchangée après une transition)

$$\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{A}$$

  - si $\forall i,\ \mu_i \geq 0,\ \sum_i \mu_i = 1$: μ est alors une **distribution stationnaire**
  - Distribution moyenne des états
- CM : paramètres $\lambda := (\mathbf{A}, \mathbf{\Pi})$
- **Existence :** conditions pour que la CM converge (ou pas)
- **Unicité :** point fixe unique (indépendant de $\mathbf{\Pi}$)

$$\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{A}$$

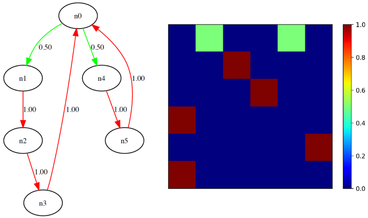
- Stationnarité : $\mu^T = \mathbf{A}^T \mu^T$
- **Condition nécessaire :** Solution donnée par le sous-espace propre (SEP) associé à la de valeur propre 1 de $\mathbf{A}^T$
  - 1 est valeur propre (VP) de $\mathbf{A}^T$ car 1 est VP de $\mathbf{A}$ ($\mathbf{A}^T$ et $\mathbf{A}$ ont même VPs). $\mathbf{A}$ est une matrice stochastique $\rightarrow$ (1, ..., 1) vecteur propre associé à VP 1.
- **Condition suffisante ?**
  1. SEP de dimension 1, et vecteur propre proba : solution unique
  2. SEP de dimension >1 : solution non unique

Quel condition sur la CM pour que 1 soit vérifiée ?

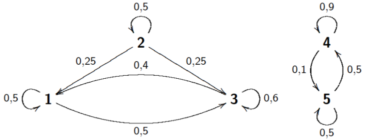
CM irreducible

- **CM irréductible :** chaque état est accessible à partir de chaque autre état.
- Graphe *G* de la CM fortement connexe, pas d'état final/absorbant
- **Chaîne réductible : plusieurs composantes** du graphe

CM irréductible



CM réductible



$$\begin{array}{c c}
 & \begin{array}{ccccc} \mathbf{1} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{2} \end{array} \\
\begin{array}{c} \mathbf{1} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \\ \mathbf{2} \end{array} & \begin{pmatrix} 0,5 & 0,5 & 0 & 0 & 0 \\ 0,4 & 0,6 & 0 & 0 & 0 \\ 0 & 0 & 0,1 & 0,9 & 0 \\ 0 & 0 & 0,5 & 0,5 & 0 \\ 0,25 & 0,25 & 0 & 0 & 0,5 \end{pmatrix}
\end{array}$$

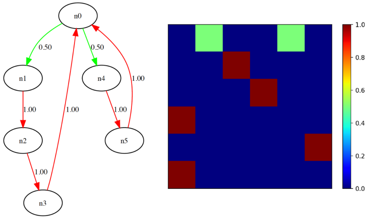
Périodicité des CM

Périodicité

- Période d'un état : PGCD des longueurs des cycles
  - Longueur cycle : temps n pour revenir dans un état, $P(X_n = i | X_0 = i) > 0$
- Période CM : PGCD de la période de tous ses états
  - Période $k > 1$: chaîne périodique
  - Période $k = 1$: chaîne apériodique
- Propriété : états de chaque composante CM ont même période

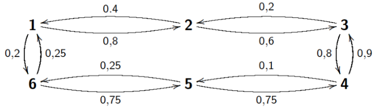
CM irréductible apériodique	CM irréductible périodique

CM irréductible apériodique



Etat 0 : PGCD(3,4)=1
Etats 1-3 : PGCD(4,7)=1
Etat 4-5 : PGCD(3,7)=1

CM irréductible périodique



$$
\begin{array}{c}
\\
\mathbf{1}\\ \mathbf{3}\\ \mathbf{5}\\ \mathbf{2}\\ \mathbf{4}\\ \mathbf{6}
\end{array}
\begin{array}{c}
\begin{array}{cccccc} \mathbf{1} & \mathbf{3} & \mathbf{5} & \mathbf{2} & \mathbf{4} & \mathbf{6} \end{array}\\
\begin{pmatrix}
0 & 0 & 0 & 0,8 & 0 & 0,2\\
0 & 0 & 0 & 0,2 & 0,8 & 0\\
0 & 0 & 0 & 0 & 0,75 & 0,25\\
0,4 & 0,6 & 0 & 0 & 0 & 0\\
0 & 0,9 & 0,1 & 0 & 0 & 0\\
0,25 & 0 & 0,75 & 0 & 0 & 0
\end{pmatrix}
\end{array}
$$

Période = 2

Ergodicité

Un processus ergodique est un processus stochastique pour lequel les statistiques peuvent être approchées par l'étude d'une seule réalisation suffisamment longue.

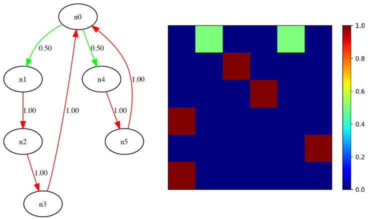
- Une chaîne de Markov est ergodique si π_n converge, indépendamment de π_0.
- π_n converge alors vers $\pi^\star$, la distribution stationnaire.
- L'espace propre associé à la VP 1 de $\mathbf{A}^T$ dimension 1

Théorème

Les chaines **irréductible** et **apériodique** sont **ergodiques**

- Dans ce cas :
  $\boldsymbol{A}^\infty = \left(\boldsymbol{\Pi_1}, \quad ...\boldsymbol{\Pi_k}..., \quad \boldsymbol{\Pi_N}\right)$, avec $\boldsymbol{\Pi_k} = (\Pi_k, ..., \Pi_k) \in \mathbb{R}^N$
- et $\forall \boldsymbol{x}$ tq $\sum_{i=1}^{N} x_i = 1$, $\boldsymbol{xA}^\infty = (\Pi_1, ..., \Pi_N)$ (convergence vers $\boldsymbol{\Pi}$)

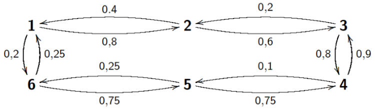
CM irréductible apériodique
$\Rightarrow$ **ergodique**



Etat 0 : PGCD(3,4)=1
Etats 1-3 : PGCD(4,7)=1
Etat 4-5 : PGCD(3,7)=1

$\Rightarrow$ **Convergence**

CM irréductible périodique
$\Rightarrow$ **non ergodique**



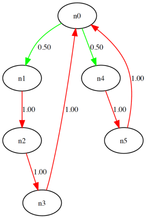
$$\begin{array}{c|c} & \begin{array}{cccccc} \mathbf{1} & \mathbf{3} & \mathbf{5} & \mathbf{2} & \mathbf{4} & \mathbf{6} \end{array} \\ \begin{array}{c} \mathbf{1} \\ \mathbf{3} \\ \mathbf{5} \\ \mathbf{2} \\ \mathbf{4} \\ \mathbf{6} \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0,8 & 0 & 0,2 \\ 0 & 0 & 0 & 0,2 & 0,8 & 0 \\ 0 & 0 & 0 & 0 & 0,75 & 0,25 \\ 0,4 & 0,6 & 0 & 0 & 0 & 0 \\ 0 & 0,9 & 0,1 & 0 & 0 & 0 \\ 0,25 & 0 & 0,75 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

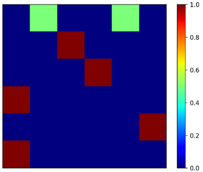
Période = 2
Sûr $\neq$ (2, 4, 6) pour n pair
Sûr $\neq$ (3, 5, 7) pour n impair

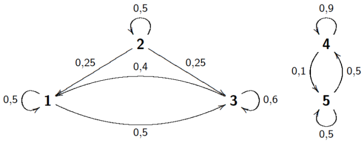
$\Rightarrow$ **Pas de convergence**

CM irréductible apériodique
$\Rightarrow$ **ergodique**

CM réductible apériodique
$\Rightarrow$ **non ergodique**







$$\begin{array}{c|c} & \begin{array}{ccccc} \mathbf{1} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{2} \end{array} \\ \begin{array}{c} \mathbf{1} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \\ \mathbf{2} \end{array} & \begin{pmatrix} 0,5 & 0,5 & 0 & 0 & 0 \\ 0,4 & 0,6 & 0 & 0 & 0 \\ 0 & 0 & 0,1 & 0,9 & 0 \\ 0 & 0 & 0,5 & 0,5 & 0 \\ 0,25 & 0,25 & 0 & 0 & 0,5 \end{pmatrix} \end{array}$$

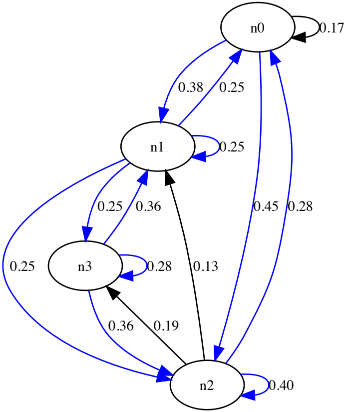
1 VP de $\mathbf{A}^T$

2 VPs de $\mathbf{A}^T$

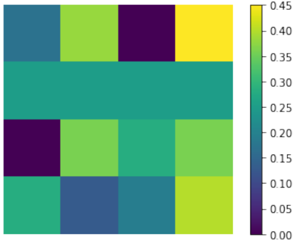
$\Rightarrow$ **Convergence indépendante de Π**

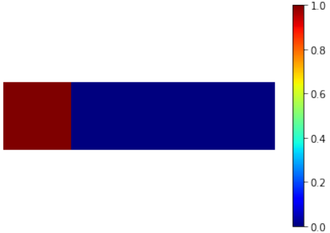
$\Rightarrow$ **Convergence dépendante de Π**

CM irréductible ? Apériodique ? Ergodique ?



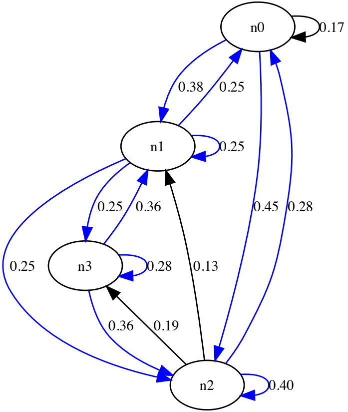
Matrice de transition :



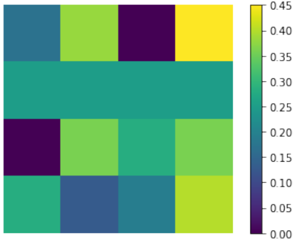


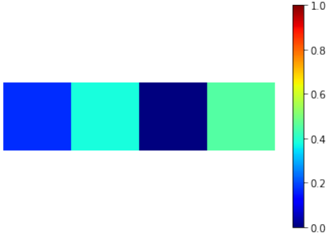
$t = 0$

CM irréductible ? Apériodique ? Ergodique ?



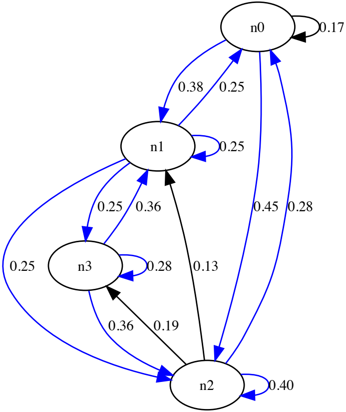
Matrice de transition :



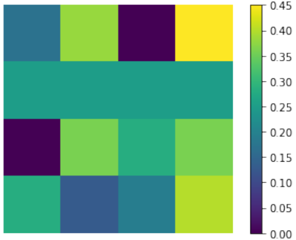


$t = 1$

CM irréductible ? Apériodique ? Ergodique ?

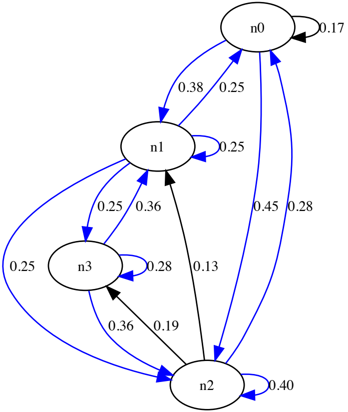


Matrice de transition :

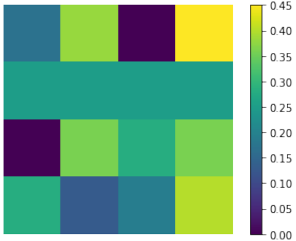


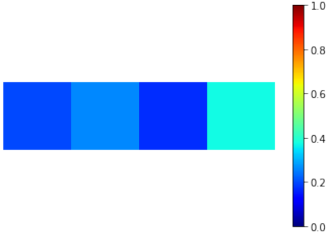
$t = 2$

CM irréductible ? Apériodique ? Ergodique ?



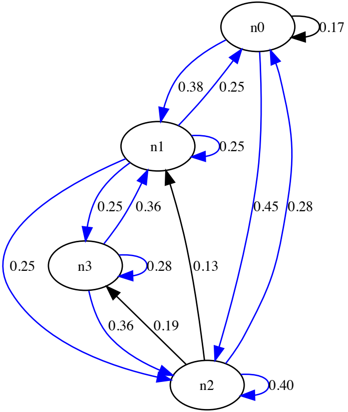
Matrice de transition :



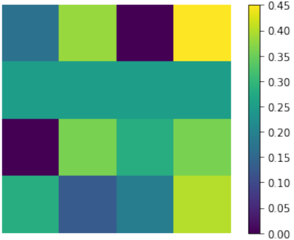


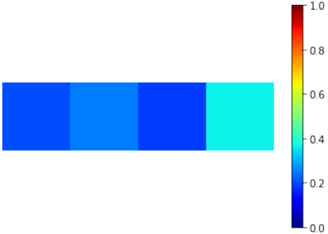
$t = 3$

CM irréductible ? Apériodique ? Ergodique ?



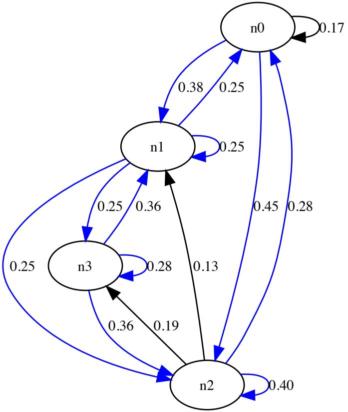
Matrice de transition :



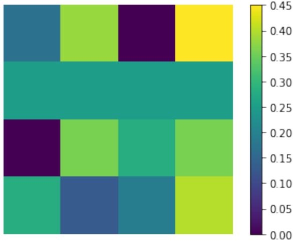


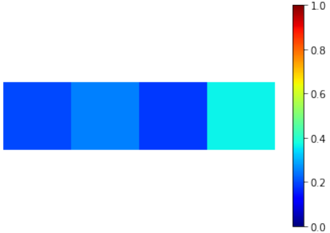
$t = 4$

CM irréductible ? Apériodique ? Ergodique ?



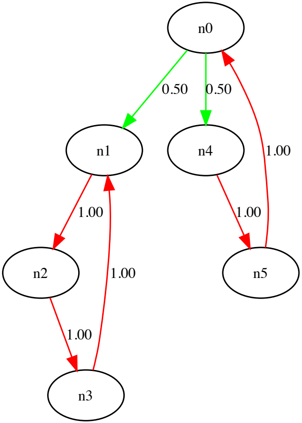
Matrice de transition :



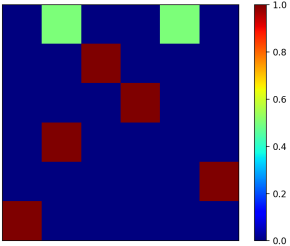


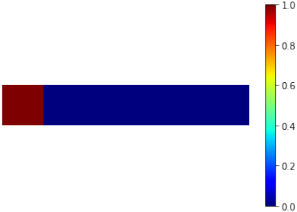
$t = 5$

CM irréductible ? Apériodique ? Ergodique ?



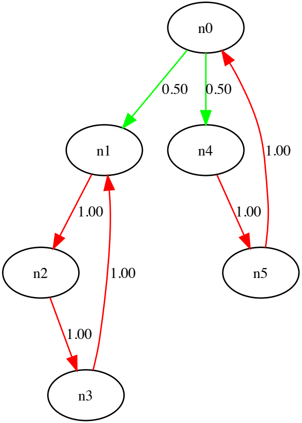
Matrice de transition :



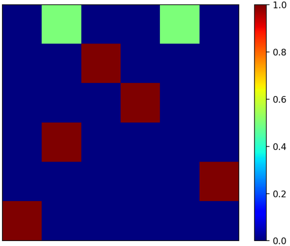


$t = 0$

CM irréductible ? Apériodique ? Ergodique ?



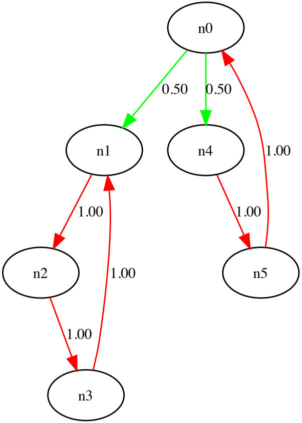
Matrice de transition :

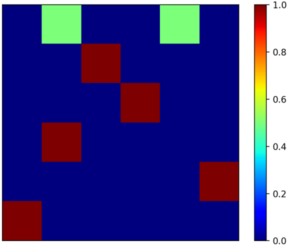


$t = 1$

CM irréductible ? Apériodique ? Ergodique ?

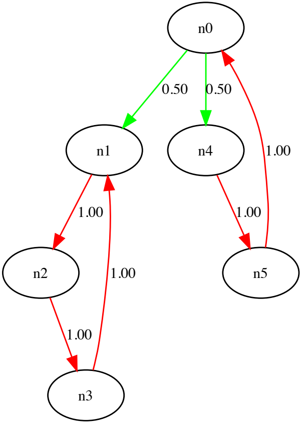
Matrice de transition :



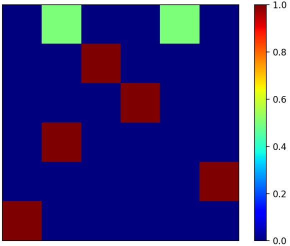


$t = 2$

CM irréductible ? Apériodique ? Ergodique ?

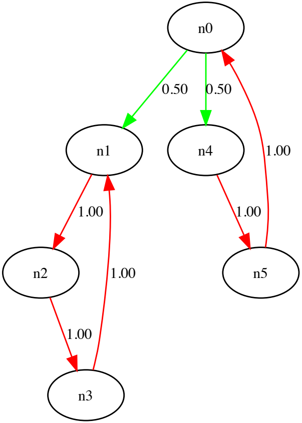


Matrice de transition :

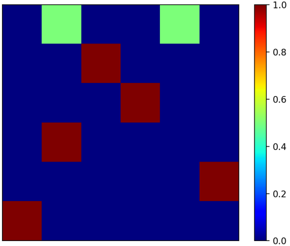


$t = 3$

CM irréductible ? Apériodique ? Ergodique ?

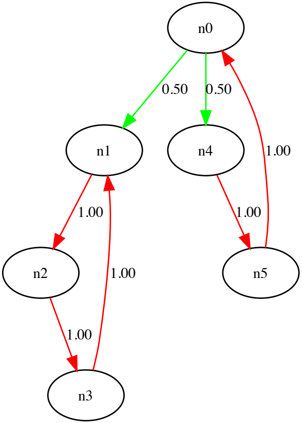


Matrice de transition :

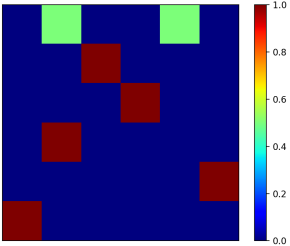


$t = 4$

CM irréductible ? Apériodique ? Ergodique ?

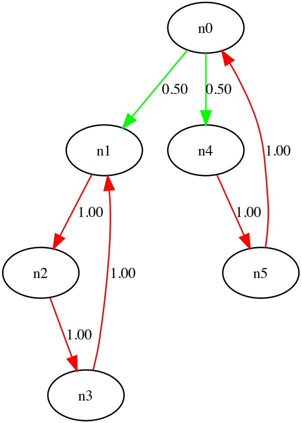


Matrice de transition :

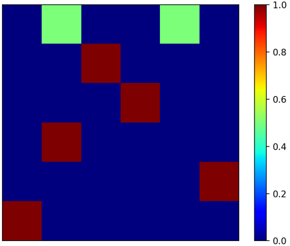


$t = 5$

CM irréductible ? Apériodique ? Ergodique ?

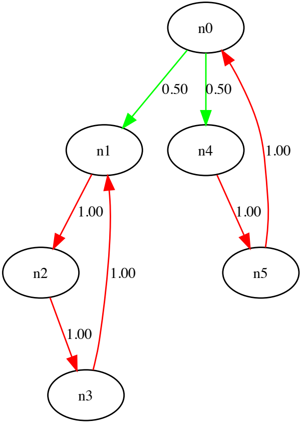


Matrice de transition :

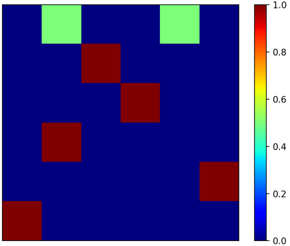


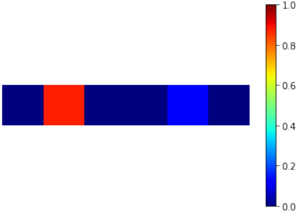
$t = 6$

CM irréductible ? Apériodique ? Ergodique ?



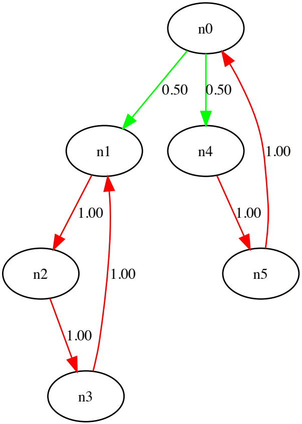
Matrice de transition :



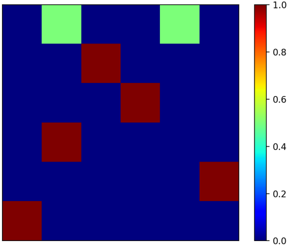


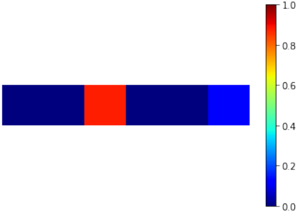
$t = 7$

CM irréductible ? Apériodique ? Ergodique ?



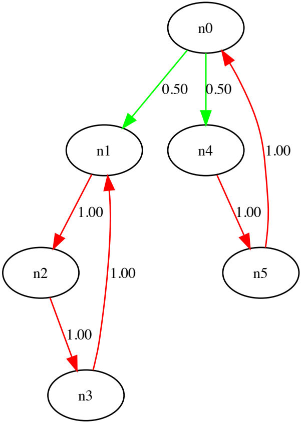
Matrice de transition :



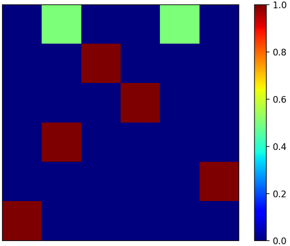


$t = 8$

CM irréductible ? Apériodique ? Ergodique ?



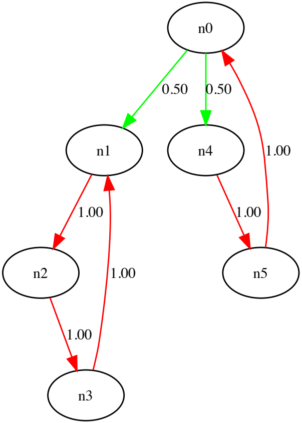
Matrice de transition :

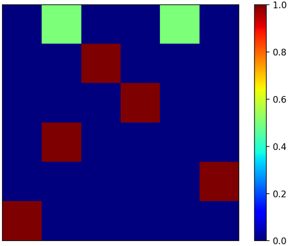


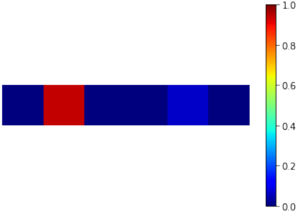
$t = 9$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :



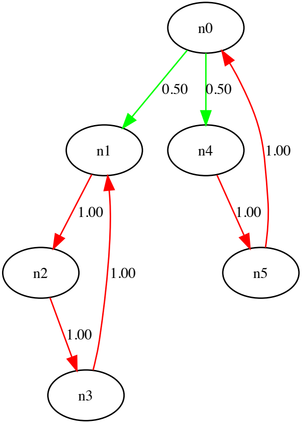


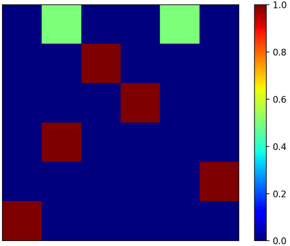


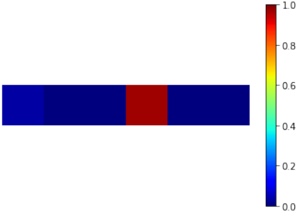
$t = 10$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

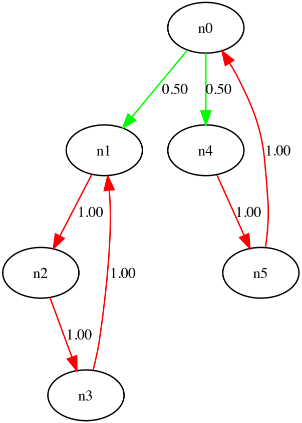




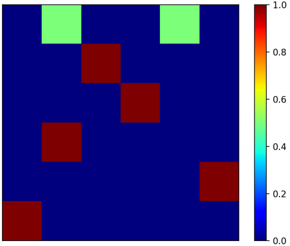


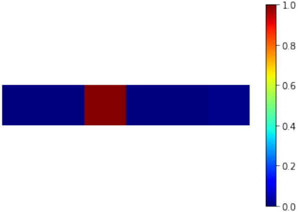
$t = 15$

CM irréductible ? Apériodique ? Ergodique ?



Matrice de transition :

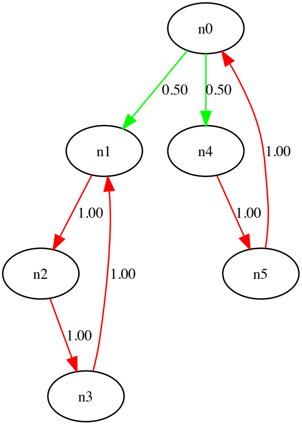




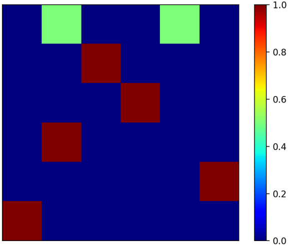
$t = 20$

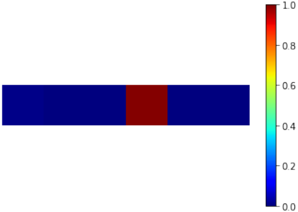
Exemples & discussion (2)

CM irréductible ? Apériodique ? Ergodique ?



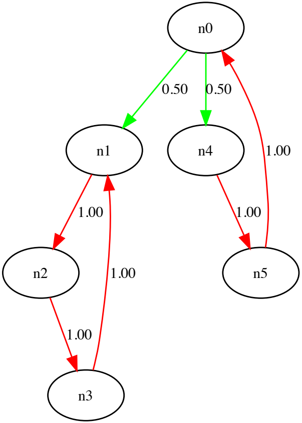
Matrice de transition :



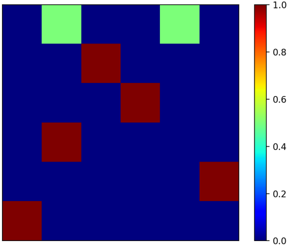


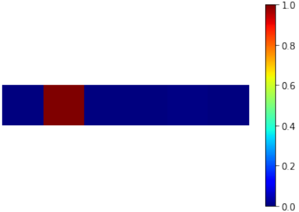
$t = 30$

CM irréductible ? Apériodique ? Ergodique ?



Matrice de transition :

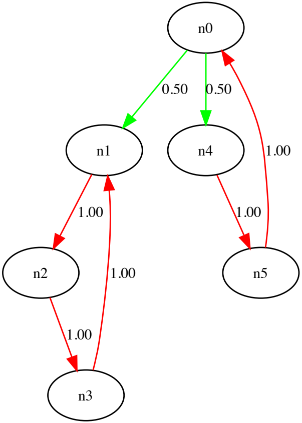


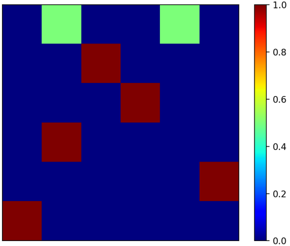


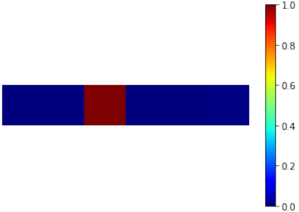
$t = 40$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :



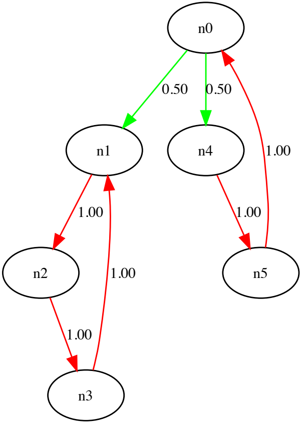


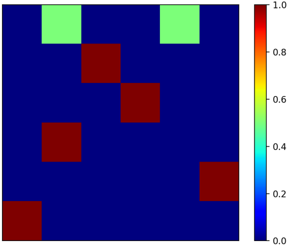


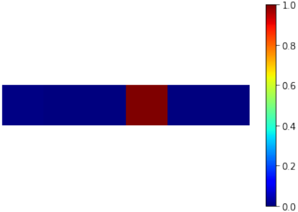
$t = 50$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :



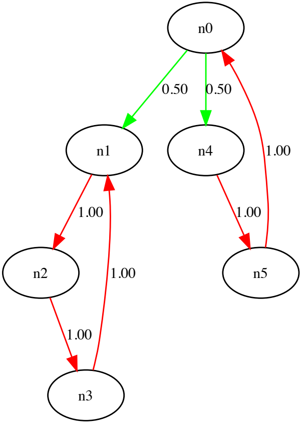


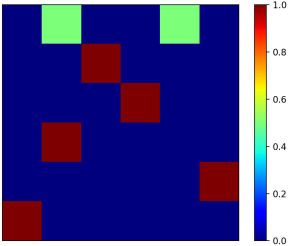


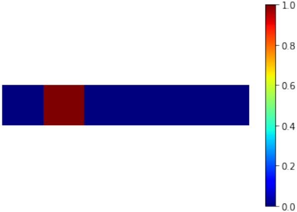
$t = 60$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

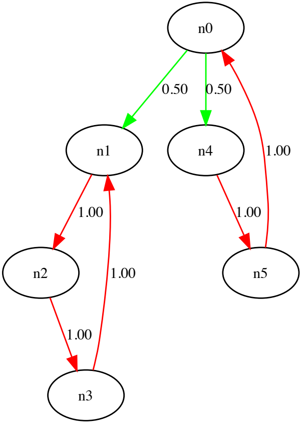




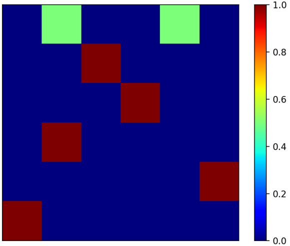


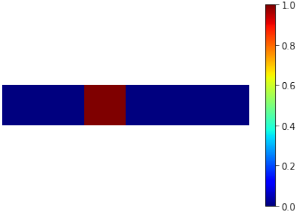
$t = 70$

CM irréductible ? Apériodique ? Ergodique ?



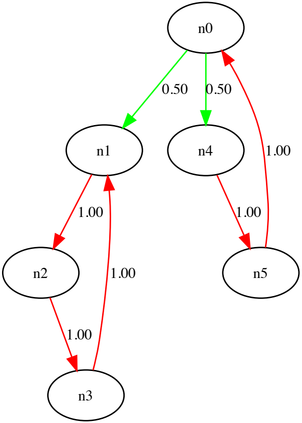
Matrice de transition :



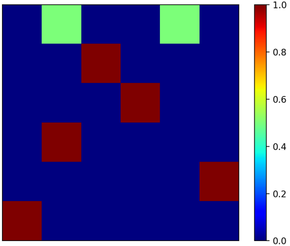


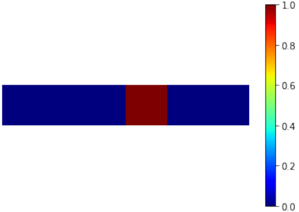
$t = 80$

CM irréductible ? Apériodique ? Ergodique ?



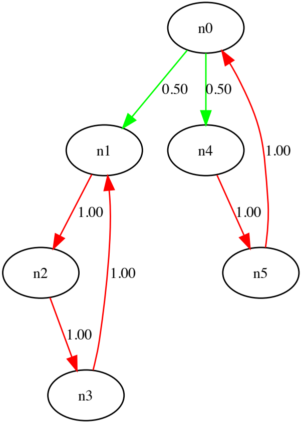
Matrice de transition :



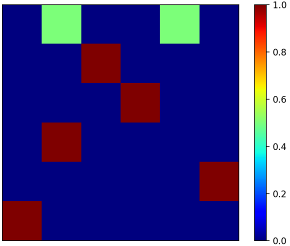


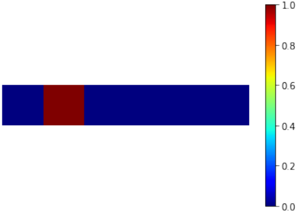
$t = 90$

CM irréductible ? Apériodique ? Ergodique ?



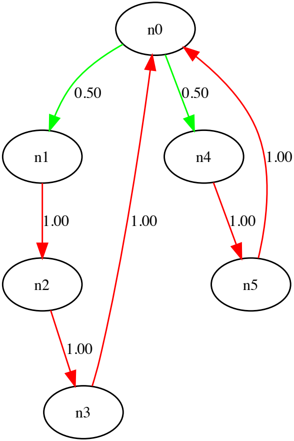
Matrice de transition :



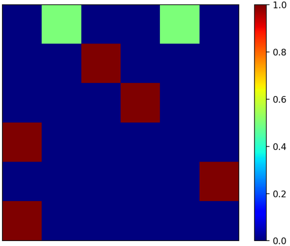


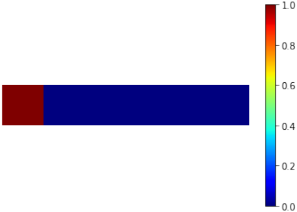
$t = 100$

CM irréductible ? Apériodique ? Ergodique ?



Matrice de transition :

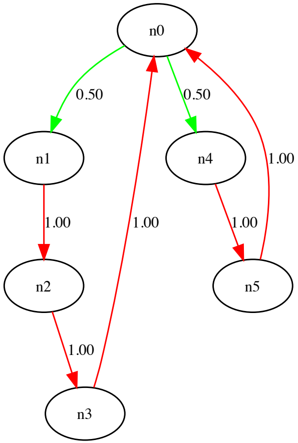


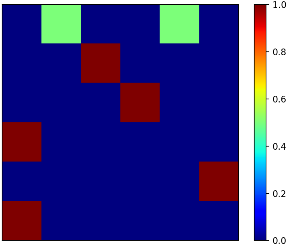


$t = 0$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

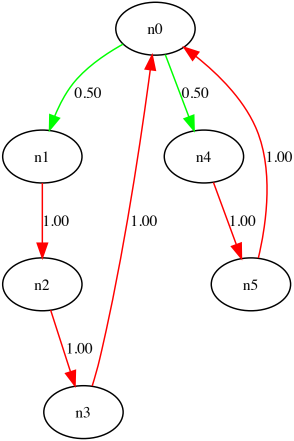


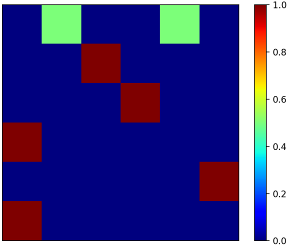


$t = 1$

CM irréductible ? Apériodique ? Ergodique ?

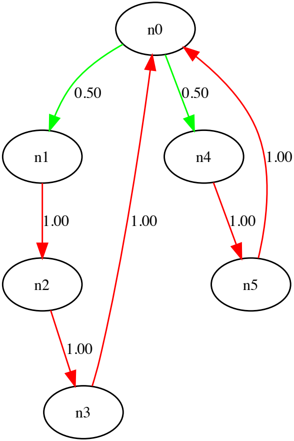
Matrice de transition :



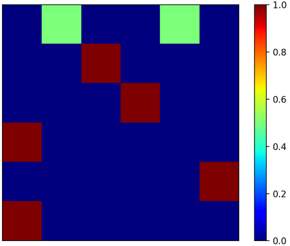


$t = 2$

CM irréductible ? Apériodique ? Ergodique ?

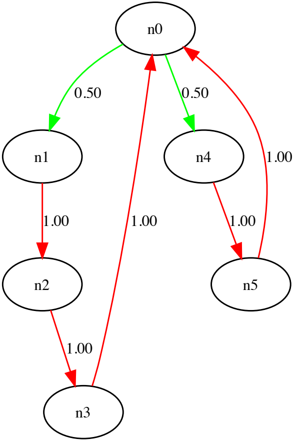


Matrice de transition :

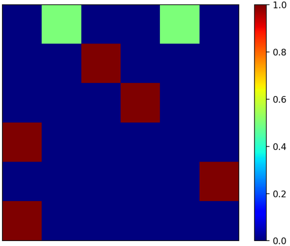


$t = 3$

CM irréductible ? Apériodique ? Ergodique ?

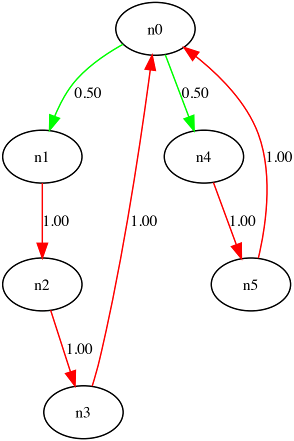


Matrice de transition :

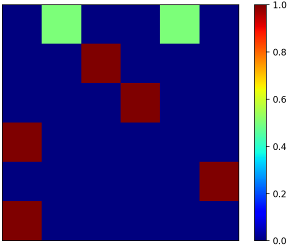


$t = 4$

CM irréductible ? Apériodique ? Ergodique ?



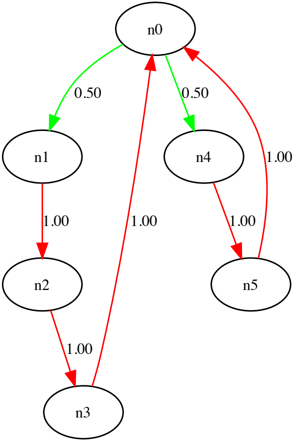
Matrice de transition :

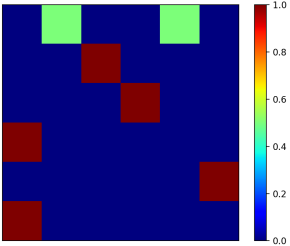


$t = 5$

CM irréductible ? Apériodique ? Ergodique ?

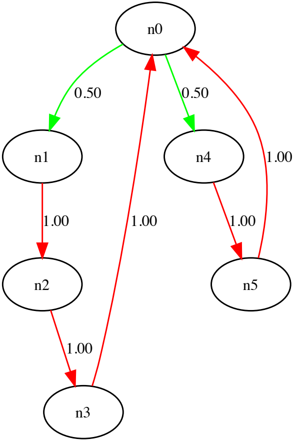
Matrice de transition :



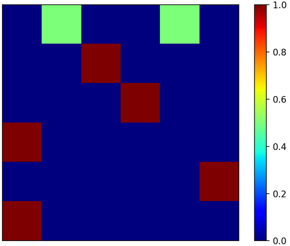


$t = 6$

CM irréductible ? Apériodique ? Ergodique ?



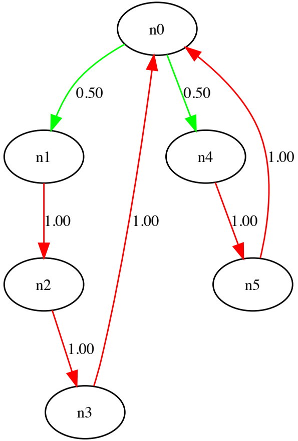
Matrice de transition :

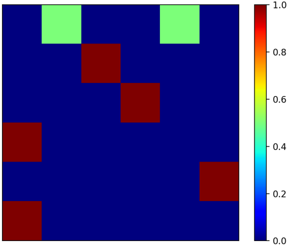


$t = 7$

CM irréductible ? Apériodique ? Ergodique ?

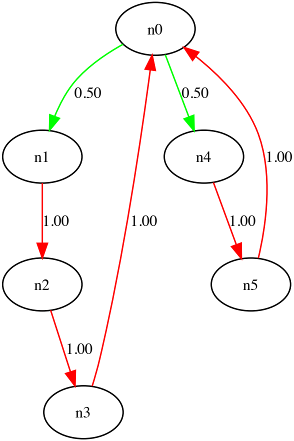
Matrice de transition :



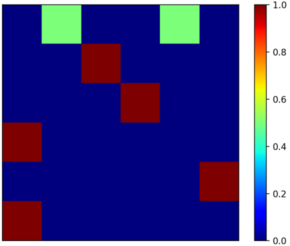


$t = 8$

CM irréductible ? Apériodique ? Ergodique ?



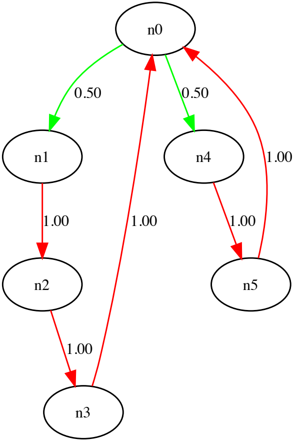
Matrice de transition :

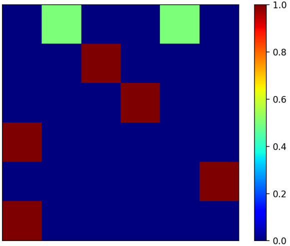


$t = 9$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

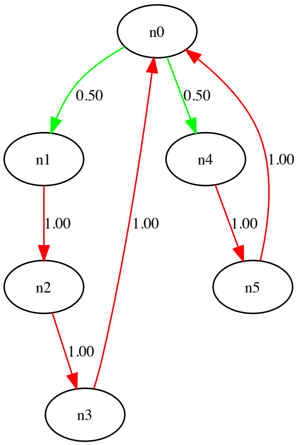


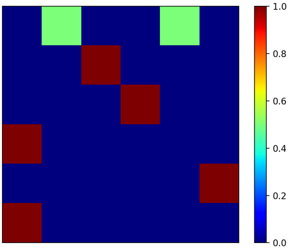


$t = 10$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

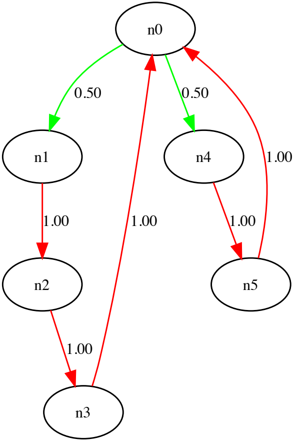


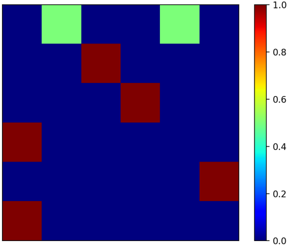


$t = 15$

CM irréductible ? Apériodique ? Ergodique ?

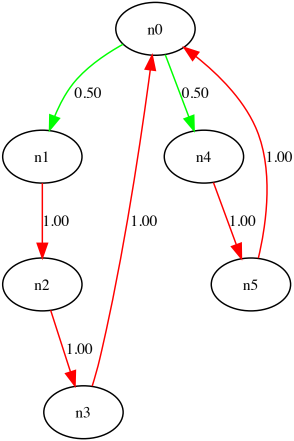
Matrice de transition :



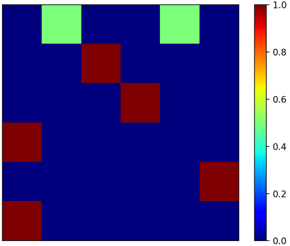


$t = 20$

CM irréductible ? Apériodique ? Ergodique ?



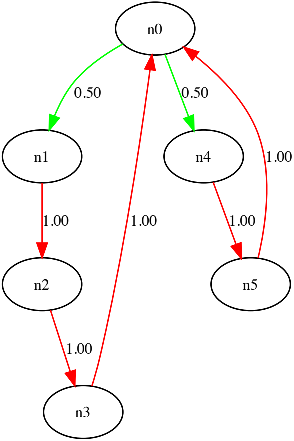
Matrice de transition :

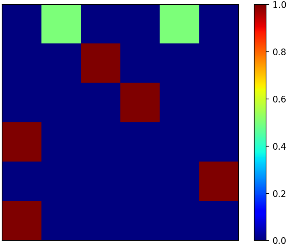


$t = 30$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

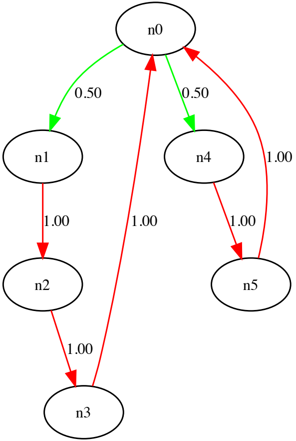


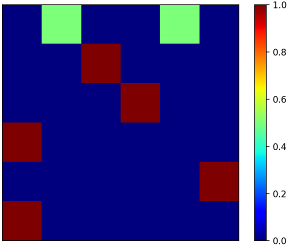


$t = 50$

CM irréductible ? Apériodique ? Ergodique ?

Matrice de transition :

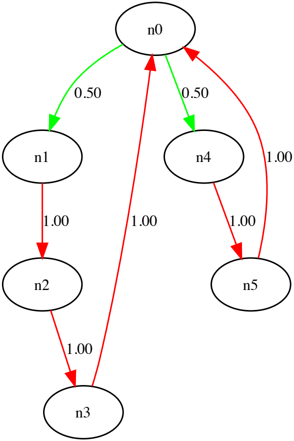


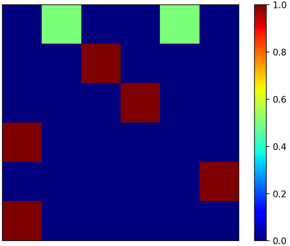


$t = 50$

CM irréductible ? Apériodique ? Ergodique ?

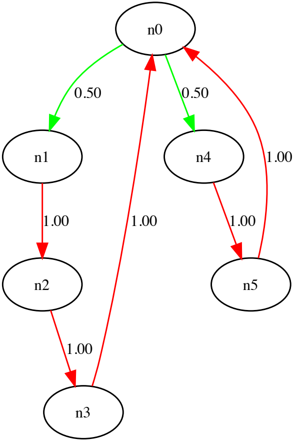
Matrice de transition :



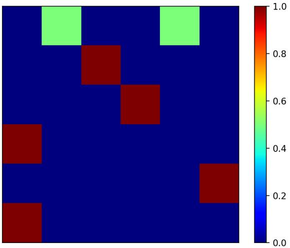


$t = 60$

CM irréductible ? Apériodique ? Ergodique ?



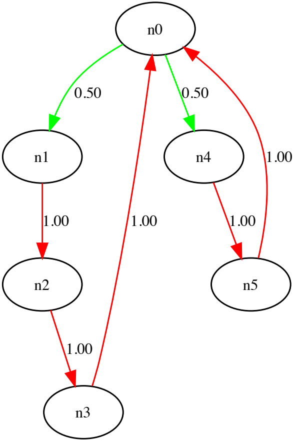
Matrice de transition :

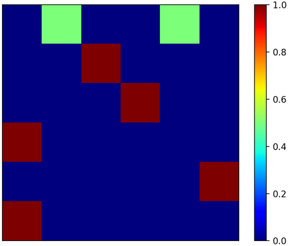


$t = 70$

CM irréductible ? Apériodique ? Ergodique ?

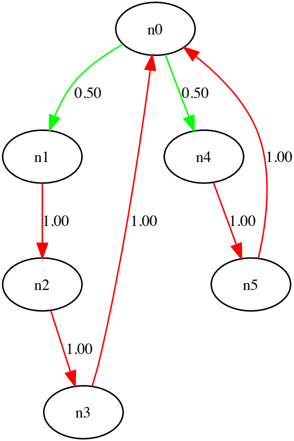
Matrice de transition :



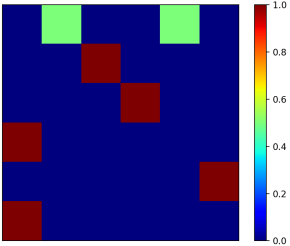


$t = 80$

CM irréductible ? Apériodique ? Ergodique ?



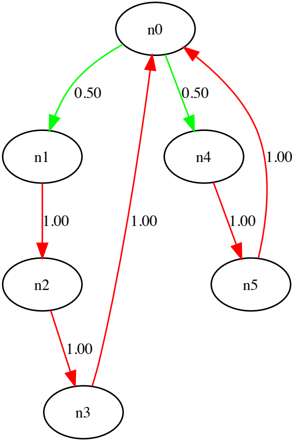
Matrice de transition :



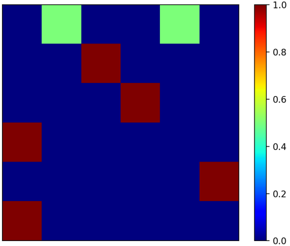
$t = 90$

Exemples & discussion (3)

CM irréductible ? Apériodique ? Ergodique ?

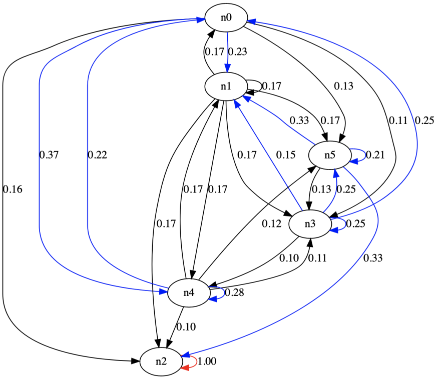


Matrice de transition :

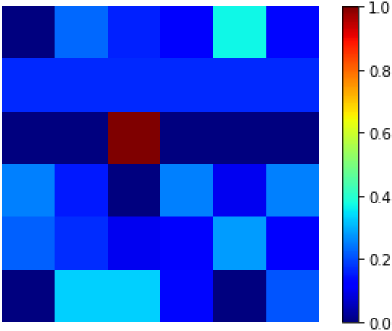


$t = 100$

CM irréductible ? Apériodique ? Ergodique ?

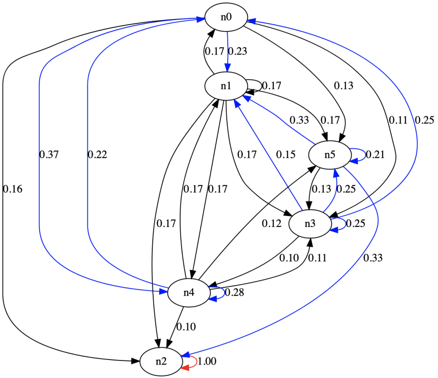


Matrice de transition :

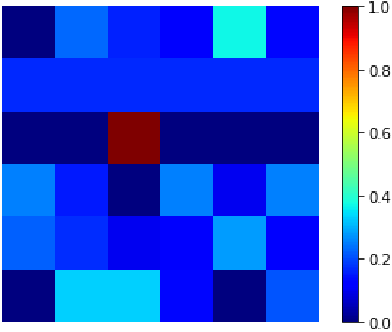


Exemples & discussion (4)

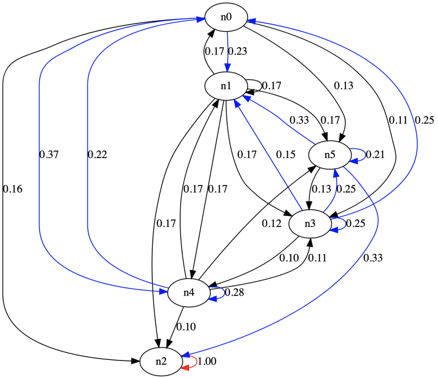
CM irréductible ? Apériodique ? Ergodique ?



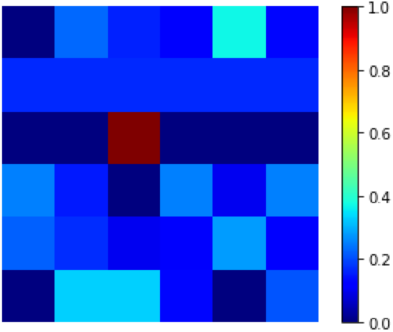
Matrice de transition :



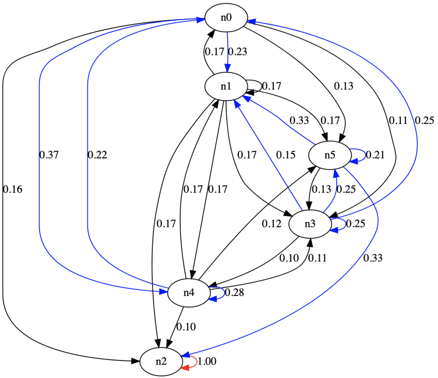
CM irréductible ? Apériodique ? Ergodique ?



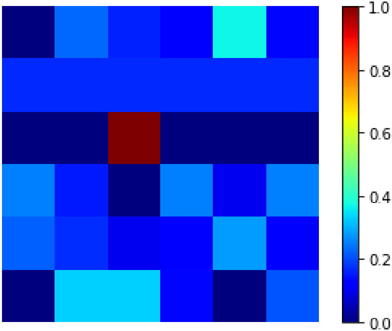
Matrice de transition :

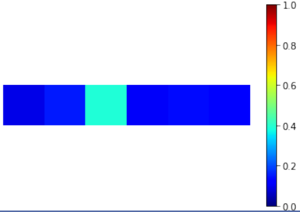


CM irréductible ? Apériodique ? Ergodique ?

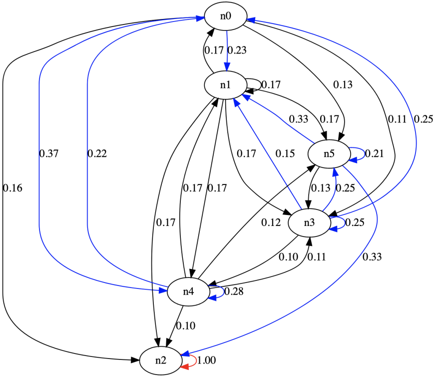


Matrice de transition :

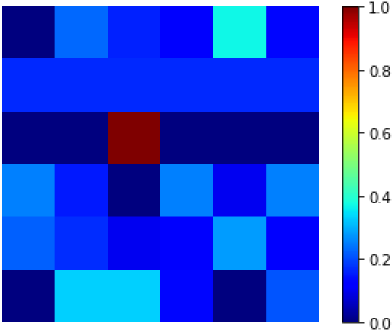


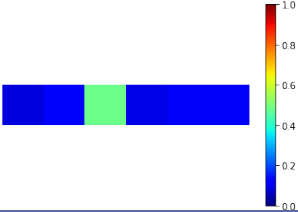


CM irréductible ? Apériodique ? Ergodique ?

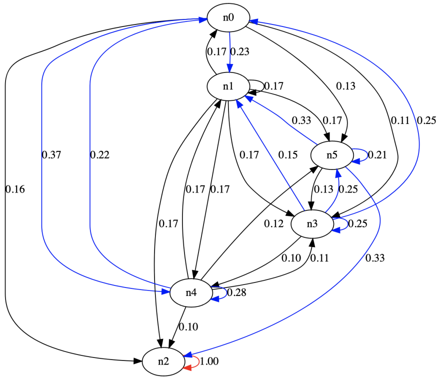


Matrice de transition :

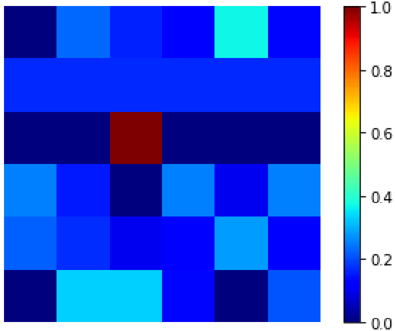


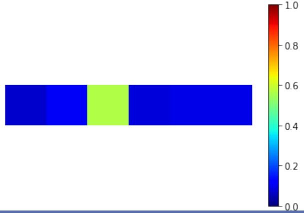


CM irréductible ? Apériodique ? Ergodique ?

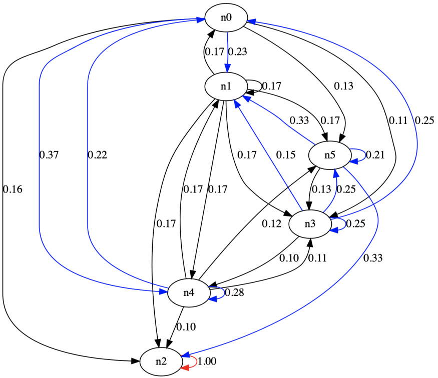


Matrice de transition :

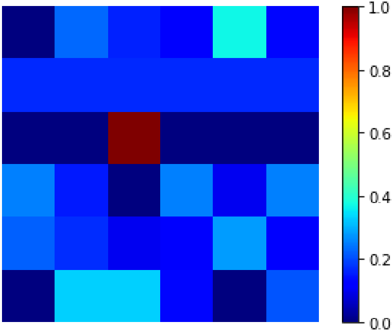


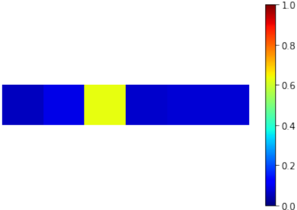


CM irréductible ? Apériodique ? Ergodique ?

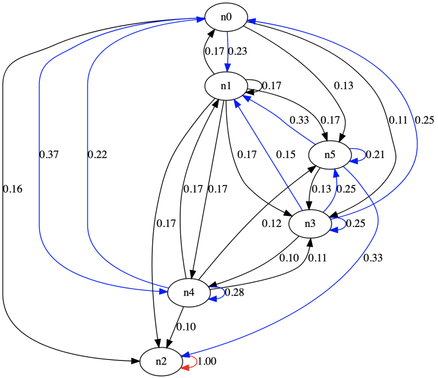


Matrice de transition :

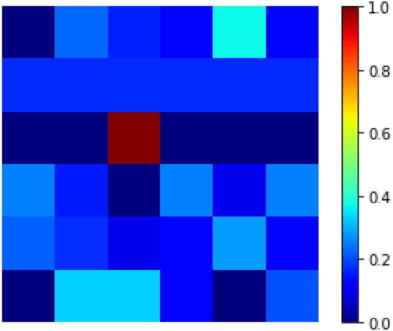


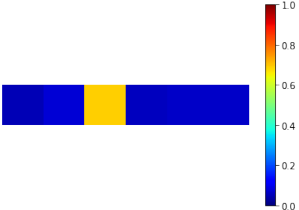


CM irréductible ? Apériodique ? Ergodique ?

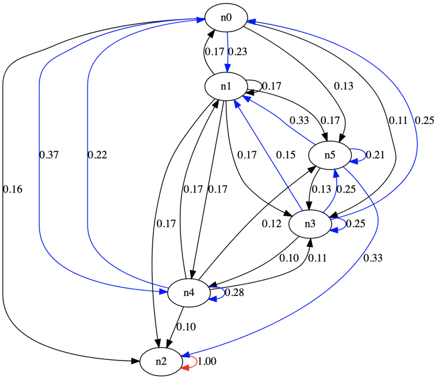


Matrice de transition :

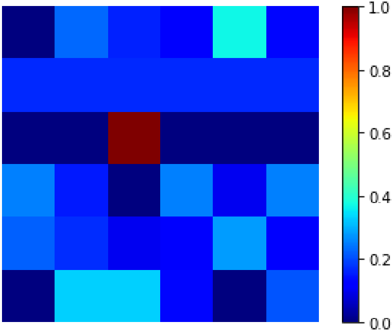


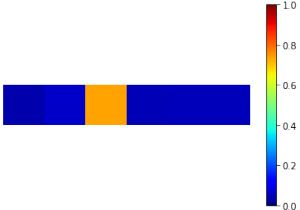


CM irréductible ? Apériodique ? Ergodique ?

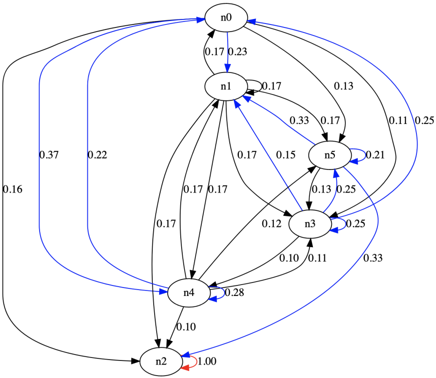


Matrice de transition :

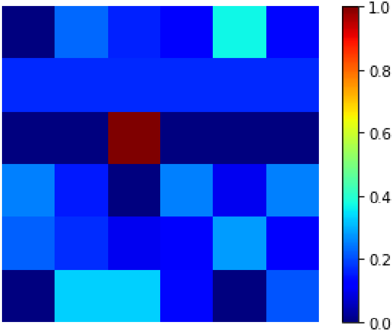


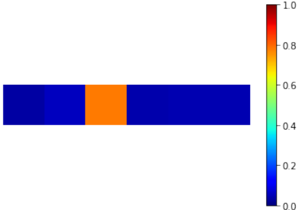


CM irréductible ? Apériodique ? Ergodique ?



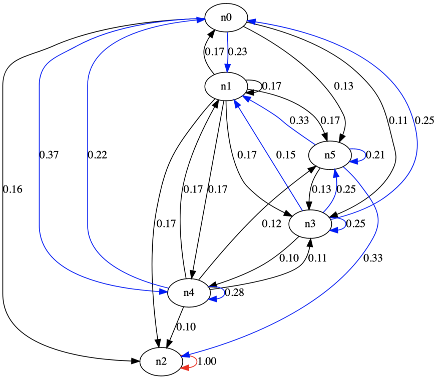
Matrice de transition :



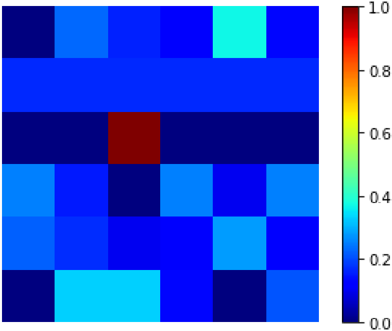


Exemples & discussion (4)

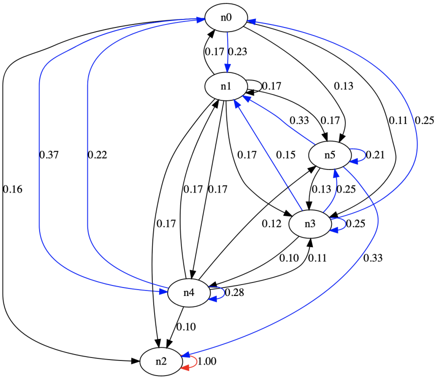
CM irréductible ? Apériodique ? Ergodique ?



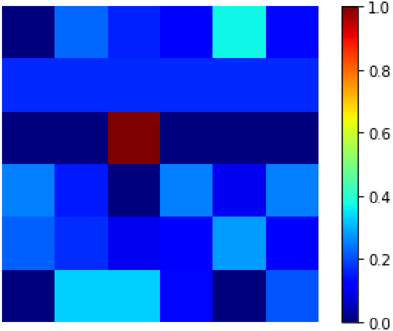
Matrice de transition :

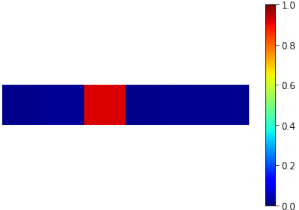


CM irréductible ? Apériodique ? Ergodique ?

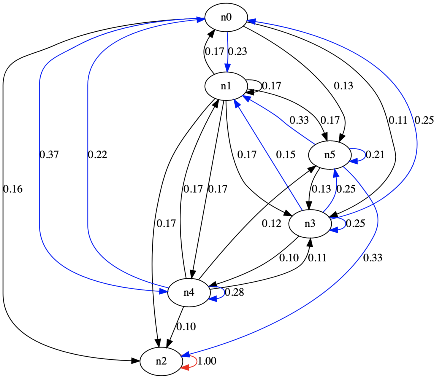


Matrice de transition :

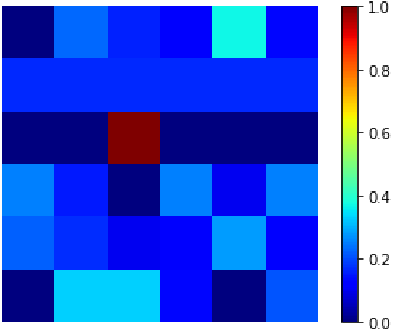


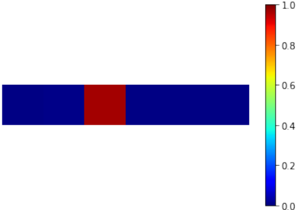


CM irréductible ? Apériodique ? Ergodique ?

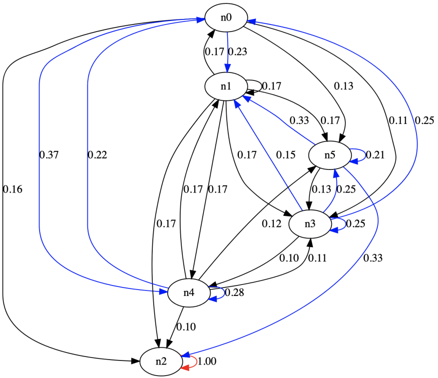


Matrice de transition :

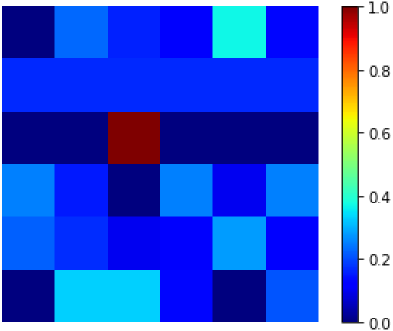


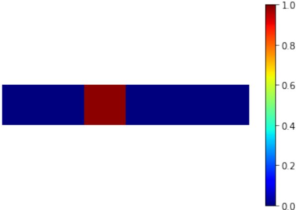


CM irréductible ? Apériodique ? Ergodique ?

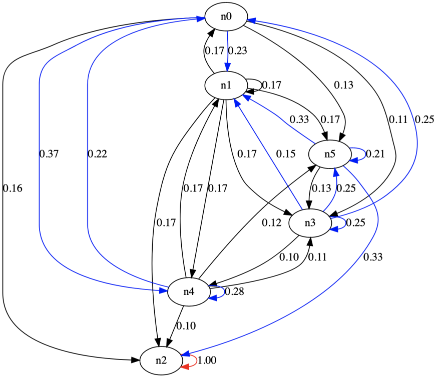


Matrice de transition :

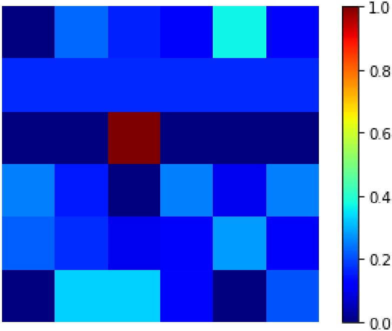


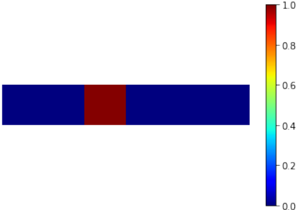


CM irréductible ? Apériodique ? Ergodique ?

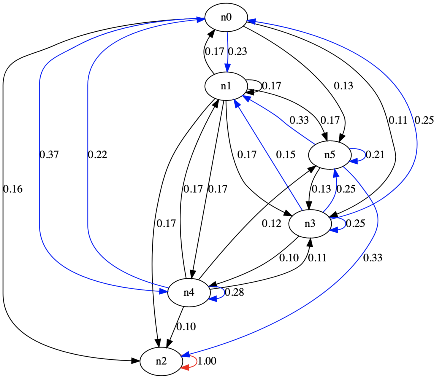


Matrice de transition :

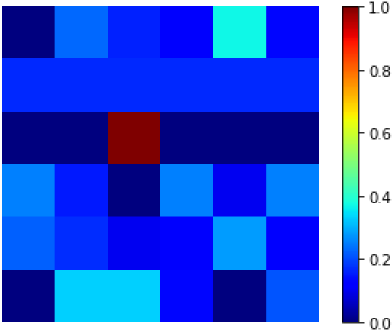


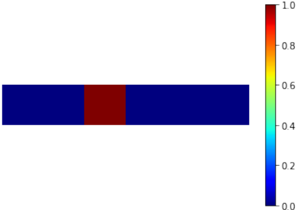


CM irréductible ? Apériodique ? Ergodique ?



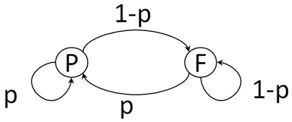
Matrice de transition :



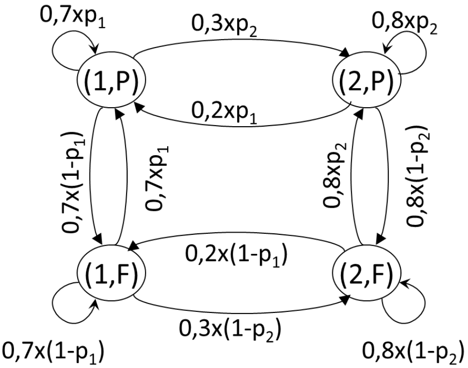


- Séquence de lancers de pièce(s)...

p : probabilité de faire *pile*



p_k : probabilité de faire *pile* avec la pièce k



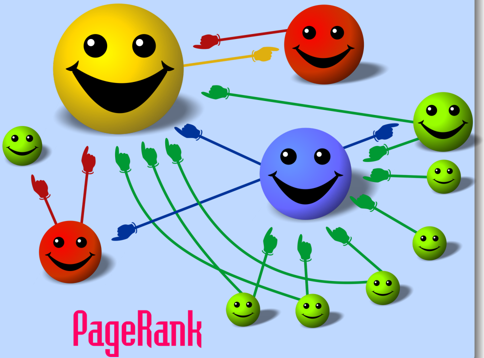
Modélisation (3)

- Modélisation de parcours utilisateur sur un site web
  - Catégorisation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

- Modélisation de parcours utilisateur sur un site web
  - Catégorisation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

Modèle

- **Etats** :
  page du site
- **Transitions** :
  hyperliens



Apprentissage des chaines de Markov

Nicolas Thome
nicolas.thome@isir.upmc.fr

LIP6 / ISIR – Sorbonne Université, France

Questions

- Etant donnée une séquence d'états, calculer sa probabilité Vu précédemment :

$$
\begin{aligned}
p(X|\lambda) &= \prod_{t=2}^{T} p(x_t|x_1,\ldots,x_{t-1},\lambda)p(x_1|\lambda) = \prod_{t=2}^{T} p(x_t|x_{t-1},\lambda)p(x_1|\lambda) \\
&= \pi_{x_1} \prod_{t=2}^{T} a_{x_{t-1},x_t}
\end{aligned}
$$

- Comment apprendre une CM à partir d'exemples ?
- Comment faire de la classification de séquences avec des CM ?

Apprentissage

- Soit une base de séquences $B = \{X^1, \dots, X^K\}$ (N états possibles)
- Critère de vraisemblance

$$\log \mathcal{L}(B, \lambda) = \log(\prod_{k=1}^{K} p(X^k|\lambda)) = \sum_{k} \log(p(X^k|\lambda))$$

- Optimisation :

$$\lambda^\star = \arg\max_{\lambda} \log \mathcal{L}(B, \lambda)$$

- Contraintes :

$$\forall i \in [1,\ N], \sum_{j=1}^{N} a_{ij} = 1$$

$$\sum_{j=1}^{N} \pi_j = 1$$

- Critère intégrant les contraintes (Lagrangien) :

$$\mathcal{C}(\lambda) = \mathcal{L}(B, \lambda) - \sum_{i=1}^{N} \nu_i \left(\sum_{j=1}^{N} a_{ij} - 1 \right) - \eta \left(\sum_{j=1}^{N} \pi_j - 1 \right)$$

- Si la dérivée par rapport au coefficient de contrainte est nulle, la contrainte est satisfaite :

$$\frac{\partial \mathcal{C}(\lambda)}{\partial \eta} = 0 \Leftrightarrow \sum_{j=1}^{N} \pi_j - 1 = 0$$

- Ecriture du Maximum de Vraisemblance :

$$L(B,\lambda) = \prod_{k=1}^{K} p(X^k|\lambda) = \prod_{k=1}^{K} \pi_{x_1^k} \prod_{t=2}^{T} a_{x_{t-1}^k, x_t^k}$$
$$L(B,\lambda) = \prod_{i=1}^{N} \pi_i^{I_i} \prod_{j=1}^{N} a_{ij}^{n_{ij}} \tag{1}$$

- n_{ij} # transitions (ij) observées, I_i # état initiaux i observés
- $\ell = log\left(L(B,\lambda)\right) = \sum_{i=1}^{N} \left[I_i \; log(pi_i) + \sum_{j=1}^{N} n_{ij} \; log(a_{ij}) \right]$
- Résolution : $\frac{\partial \ell}{\partial a_{ij}} = 0$, $\frac{\partial \ell}{\partial \pi_i} = 0$ + contraintes :

$$a_{ij} = \frac{n_{ij}}{n_{i.}} \quad \pi_j = \frac{I_j}{K}, \text{ avec : } n_{i.} = \sum_j n_{ij}$$

En pratique

- Approche par comptage
- Calcul des fréquences des évènements = solution au sens MV
- Chaque ligne de A est une distribution (sommant à 1)
- En faisant une hypothèse de stationnarité (ergodicité), il est possible d'estimer les π_j sur toute la base de données :

$$\text{classique : } \pi_j = \frac{I_j}{K} \quad \text{alternative stationnaire : } \pi_j = \frac{n.j}{\sum_{ij} n_{ij}}$$

Distance entre séquences

Nicolas Thome
nicolas.thome@isir.upmc.fr

LIP6 / ISIR – Sorbonne Université, France

Idée générale

- Similarité/distance = outil de base
  - k-plus proches voisins...
- La similarité peut concerner une partie seulement du signal
  - Reflexion sur les besoins spécifiques

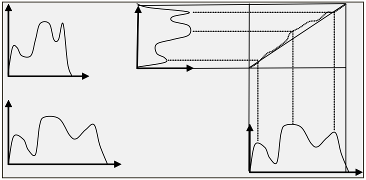
Dynamic Time Warping (DTW)

- L'analyse du signal est locale (cf. stationnarité)
- Unités de reconnaissance plus globales (phonèmes, mots, ...)
- $\Rightarrow$ Nécessité de comparer des séquences de vecteurs
- **DTW = distance entre séquences**
  - ayant des longueurs différentes
  - insensible à certaines variabilités d'élocution
  - calculable efficacement

Distance entre séquences

Idée :

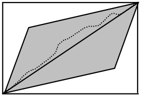
- Existence d'une distance entre séquences capable de
  - Prendre en compte les différences de rythme dans les séquences
  - Comparer des séquences de longueur différente
- Dynamic Time Warping (DTW)

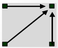


La distance entre les séquences est la somme des distances entre éléments mis en correspondance par l'alignement

Appariement avec contraintes

- ❑ Continuité locale
- ❑ Alignement quasi linéaire
- ❑ Début et fin synchrones





Distance entre séquences (2)

Notion de chemin d'alignement :

- Chemin : $c = \{(i_k, j_k)\}_{k=1,\dots,K}$ tel que :

$$\forall k,\ (i_k, j_k) = \begin{cases} (i_{k-1} - 1, j_{k-1}) & i_1 = j_1 = 1 \\ (i_{k-1} - 1, j_{k-1} - 1) & j_K = T_2 \\ (i_{k-1}, j_{k-1} - 1) & i_K = T_1 \end{cases}$$

- Distance suivant un alignement :

$$D_c(S_1, S_2) = \sum_{k=1}^{K} d_{c(k)}(S_1[i(k)], S_2[j(k)])$$

- Distance entre 2 séquences

$$D(S_1, S_2) = \min_c D_c(S_1, S_2)$$

Phase avant :

- calcul des $\forall i, j,\ d(S_1[i], S_2[j])$
- sommes cumulées

3	2	5	2	4	2	2
2	2	3	2	1	4	4
2	1	2	2	2	3	4
1	1	2	1	1	3	2
1	1	3	3	3	3	4

2	2					
1	2					

3	3	4				
2	2	4				
1	2	5				

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Phase retour :

- Chemin correspondant au cout minimum

9	7	10	8	10	9	← 11
6	5	6	6	← 7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

3	2	5	2	4	2	2
2	2	3	2	1	4	4
2	1	2	2	2	3	4
1	1	2	1	1	3	2
1	1	3	3	3	3	4

2	2					
1	2					

3	3	4				
2	2	4				
1	2	5				

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Phase avant

Phase arrière

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Alignement final et distance

x_5	3	2	5	2	4	2	2
x_4	2	2	3	2	1	4	4
x_3	2	1	2	2	2	3	4
x_2	1	1	2	1	1	3	2
x_1	1	1	3	3	3	3	4
	y_1	y_2	y_3	y_4	y_5	y_6	y_7