

Graphe

Sylvain Meignier

Fonctionnement

- 2 CM, 1 TD, 2 TP
- Evaluation
 - Un QCM en TP ou TD
 - 1h examen



Chapitre 1

Chaine de Markov



Doudou le Hamster (Wikipedia)

■ Doudou ne connaît que trois endroits :

- les copeaux où il dort
- la mangeoire où il mange
- la roue où il fait de l'exercice



Activité de Doudou

- On observe les activités de Doudou et toutes les minutes on note s'il est dans les **copeaux**, dans la **mangeoire** et dans la **roue**
 - On a les événements $\Omega = \{C, M, R\}$
 - On discrétise le temps
 - On note N le nombre de pas (d'instants)
 - On choisit une minute comme intervalle de temps
 - On estime les probabilités de passer d'un lieu à un autre
 - Calcul de fréquences d'apparition d'un événement en connaissant l'activités précédentes
 - $P(C|C), P(M|C), P(R|C), P(C|M), \dots$



Probabilités

C = copeaux, M = mangeoire, R = roue

■ On observe les événements suivants :

- $M \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow M,$
- $M \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C,$
- $C \rightarrow R, C \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C,$
- $C \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C, C \rightarrow C,$
- $C \rightarrow C, C \rightarrow C, M \rightarrow R, M \rightarrow R, M \rightarrow R,$
- $M \rightarrow C, M \rightarrow C, M \rightarrow C, M \rightarrow C, M \rightarrow C,$
- $R \rightarrow C, R \rightarrow C, R \rightarrow C, R \rightarrow C, R \rightarrow C,$
- $R \rightarrow C, R \rightarrow C, R \rightarrow C, R \rightarrow R, R \rightarrow R$



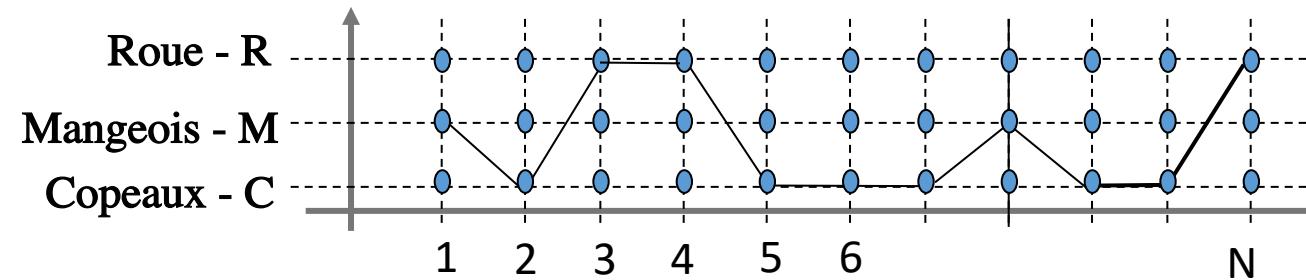
Probabilités

- Quand il dort, il a 9 chances sur 10 de ne pas se réveiller la minute suivante.
- Quand il se réveille, il y a 1 chance sur 2 qu'il aille manger et 1 chance sur 2 qu'il parte faire de l'exercice.
- Le repas ne dure qu'une minute, après il fait autre chose.
- Après avoir mangé, il y a 3 chances sur 10 qu'il parte courir dans sa roue, mais surtout 7 chances sur 10 qu'il retourne dormir.
- Courir est fatigant pour Doudou ; il y a 8 chances sur 10 qu'il retourne dormir au bout d'une minute. Sinon il continue en oubliant qu'il est déjà un peu fatigué.



Activité de Doudou

- On observe les activités de Doudou et toutes les minutes on note s'il est dans les **copeaux**, dans la **mangeoire** et dans la **roue**
 - On discrétise : note N le nombre de pas (d'instants)
 - On estime les probabilités de passer d'un lieu à un autre
 - Représentation sous forme d'un treillis (*lattice* en anglais)
 - Graphe orienté : on peut aller du pas i à i+1



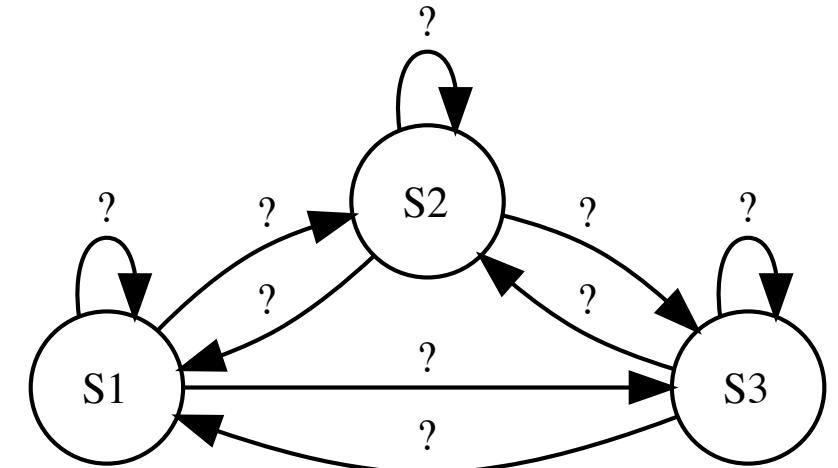
Graphe

- Faire un graphe orienté où
 - Les nœuds représentent les lieux, on notera
 - S_1 = « les copeaux »
 - S_2 = « la mangeoire »
 - S_3 = « la roue »
 - On note K le nombre de nœuds
 - Les arcs représentent le passage d'un lieu à un autre
 - Un arc est créé si la probabilité de passage d'un lieu à l'autre est non nulle



Probabilités

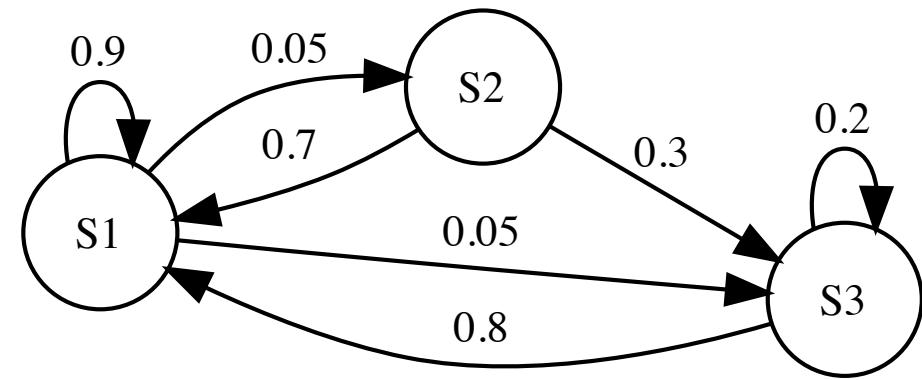
- Quand il dort, il a 9 chances sur 10 de ne pas se réveiller la minute suivante.
- Quand il se réveille, il y a 1 chance sur 2 qu'il aille manger et 1 chance sur 2 qu'il fasse de l'exercice.
- Le repas ne dure qu'une minute, après il fait autre chose.
- Après avoir mangé, il y a 3 chances sur 10 qu'il parte courir dans sa roue, mais surtout 7 chances sur 10 qu'il retourne dormir.
- Courir est fatigant pour Doudou ; il y a 8 chances sur 10 qu'il retourne dormir au bout d'une minute. Sinon il continue en oubliant qu'il est déjà un peu fatigué.
- *Dessiner le graphe*



S_1 = « les copeaux »
 S_2 = « la mangeoire »
 S_3 = « la roue »

Probabilités

- Quand il dort, il a 9 chances sur 10 de ne pas se réveiller la minute suivante.
- Quand il se réveille, il y a 1 chance sur 2 qu'il aille manger et 1 chance sur 2 qu'il fasse de l'exercice.
- Le repas ne dure qu'une minute, après il fait autre chose.
- Après avoir mangé, il y a 3 chances sur 10 qu'il parte courir dans sa roue, mais surtout 7 chances sur 10 qu'il retourne dormir.
- Courir est fatigant pour Doudou ; il y a 8 chances sur 10 qu'il retourne dormir au bout d'une minute. Sinon il continue en oubliant qu'il est déjà un peu fatigué.



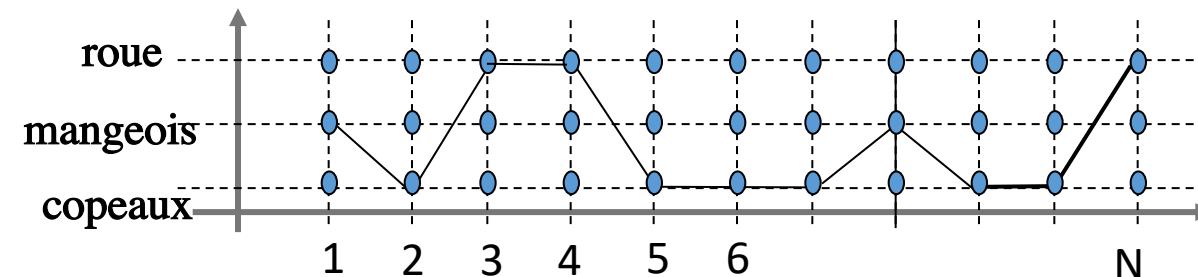
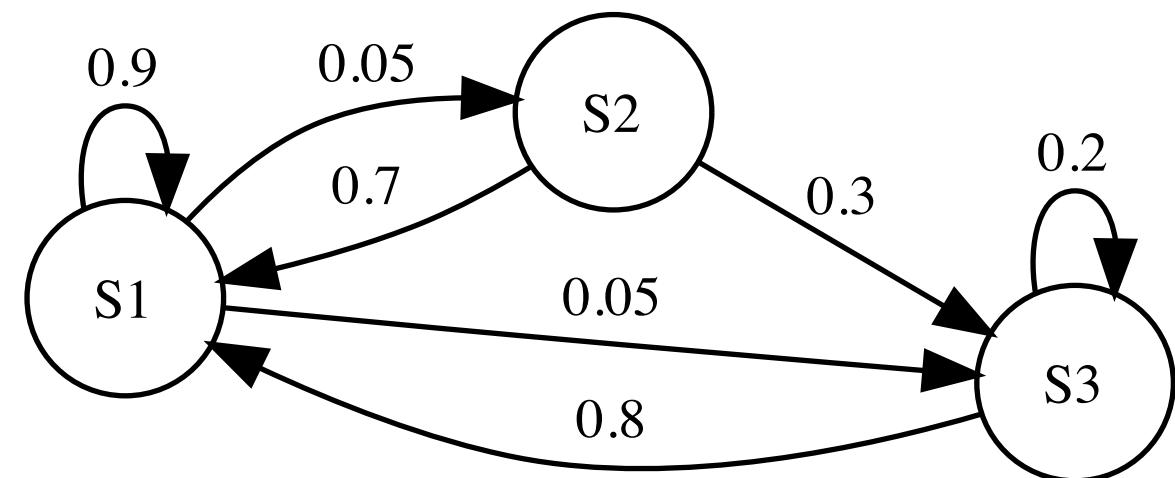
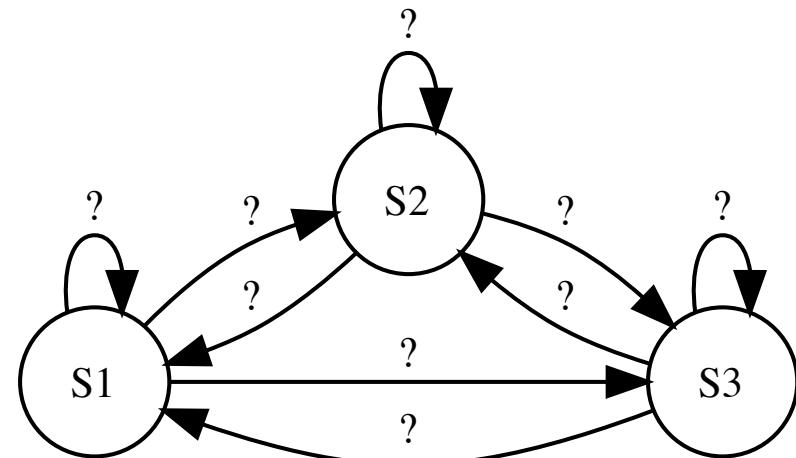
S_1 = « les copeaux »

S_2 = « la mangeoire »

S_3 = « la roue »

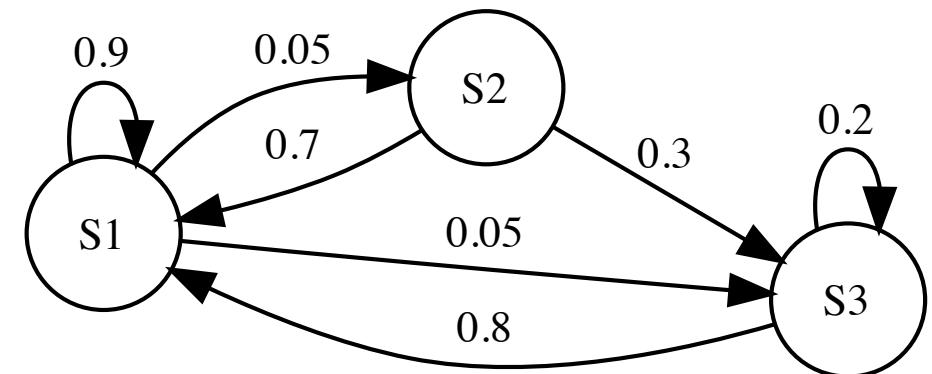
Graphe orienté probabilisé

- Graphe orienté avec des probabilités sur les arcs



Matrice de transition

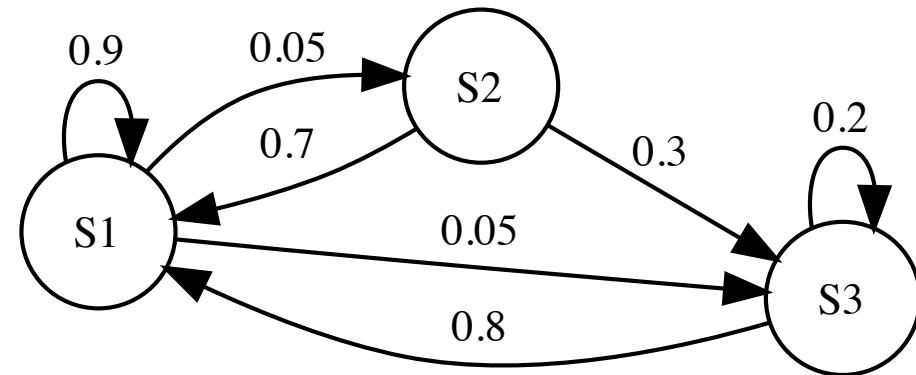
- Un graphe peut être stocké dans une matrice carrée A de dimension K identique au nombre de nœuds
- $A = (a_{i,j})$ avec i l'indice la ligne et j l'indice de colonne
- $a_{i,j}$ est la probabilité d'aller du nœud i au noeud j
 - i = ligne
 - j = colonne
- *Remplir la matrice de transition*



Matrice de transition

- $A = (a_{i,j})$ avec i l'indice la ligne et j l'indice de colonne
- $a_{i,j}$ est la probabilité d'aller du nœud i au nœud j

S_1 = « les copeaux »
 S_2 = « la mangeoire »
 S_3 = « la roue »



$$\begin{pmatrix}
 & S1 & S2 & S3 \\
 S1 & 0.9 & 0.05 & 0.05 \\
 S2 & 0.7 & 0 & 0.3 \\
 S3 & 0.8 & 0 & 0.2
 \end{pmatrix}$$

Probabilité de transition

- L'ensemble des transition = la matrice A = $(a_{i,j})$
- Propriétés
 - La somme des probabilités sortantes d'un état est égale à 1

$$\forall i, \sum_j a_{i,j} = 1$$

$$\begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.7 & 0 & 0.3 \\ 0.8 & 0 & 0.2 \end{pmatrix}$$

- Attention : ce n'est pas vrai pour les colonnes
 - $0.9+0.05+0.05 = 1$
 - $0.9+0.7+0.8 \neq 1$
- C'est une matrice à coefficients positifs ou nuls



Chaine de Markov

- On vient de construire une **chaine de Markov** (ou processus de Markov)
 - Inventeur : Andreï Markov, 1856 – 1922, Russe, Université de Saint Petersbourg. Publication en 1906.
- Un processus de Markov à temps discret est une séquence X_0, X_1, X_2, \dots de variables aléatoires à valeurs dans l'espace des états noté E.
 - Un nœud du graphe = un état
 - Un arc = une transition
 - Dans nos études E est fini, il est de cardinal K



Probabilité initiale

- Notre graphe n'a pas : d'état d'entrée, d'état de sortie
- On considère qu'il existe une transition pour passer de l'état de départ à tous les états de la chaîne
 - Un vecteur de probabilité doit être fourni

$$\pi = [\pi_1, \pi_2, \dots, \pi_K]$$
$$\sum_{i=1}^K \pi_i = 1$$

- Ou à défaut, on considère que les probabilités sont équiprobables

$$\pi = \left(\frac{1}{K} \quad \frac{1}{K} \quad \dots \quad \frac{1}{K} \right)$$



Chaine de Markov

- Une chaine de Markov est définie par 2 paramètres
 - sa matrice de transition A
 - son vecteur de probabilité initial π
- Chaine de Markov = un modèle
 - = une représentation imparfaite du monde réel
 - = simplification et approximation du réel
 - C'est un processus qui décrit vos données
- Un modèle sert :
 - à prendre des décisions



Propriété de Markov faible

- La loi de X_{n+1} ne dépend de l'histoire X_0, X_1, \dots, X_n du système que de l'état de X_n
 - = pour prédire le futur X_{n+1} , je n'ai besoin que du présent X_n
 - = le résultat d'une épreuve ne dépend que du résultat de l'épreuve précédente.

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

■ Exemple

$$P(X_{n+1} = "Copeaux" | X_0 = "roue", X_1 = "mangeoire", \dots, X_n = "roue") =$$

$$P(X_{n+1} = "Copeaux" | X_n = "roue")$$



Propriété de Markov faible

- Chaînes de Markov **homogènes** : le mécanisme de transition ne change pas au cours du temps

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i), \forall n \geq 0, \forall (i, j) \in E^2$$

- Exemple

$$\begin{aligned} P(X_{n+1} = "Copeaux" | X_n = "roue") &= \\ P(X_0 = "Copeaux" | \dots, X_1 = "roue"), \forall n & \end{aligned}$$



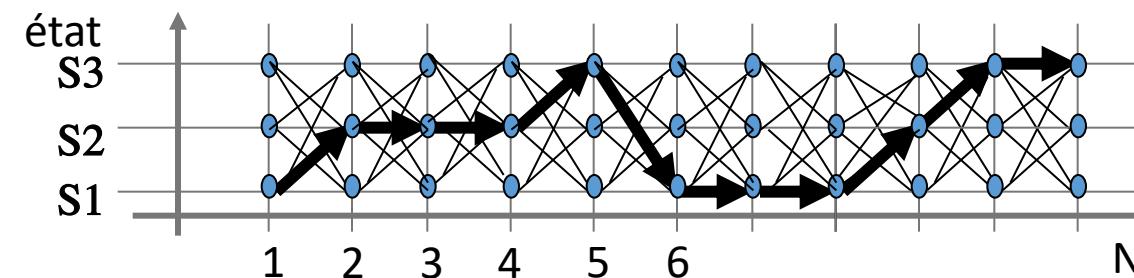
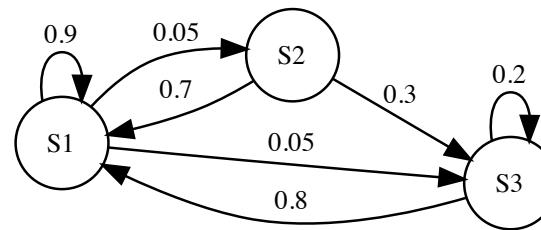
Probabilité d'un chemin

■ Observation

- Une suite de réalisations est appelée observation, noté O

■ Probabilité d'un chemin

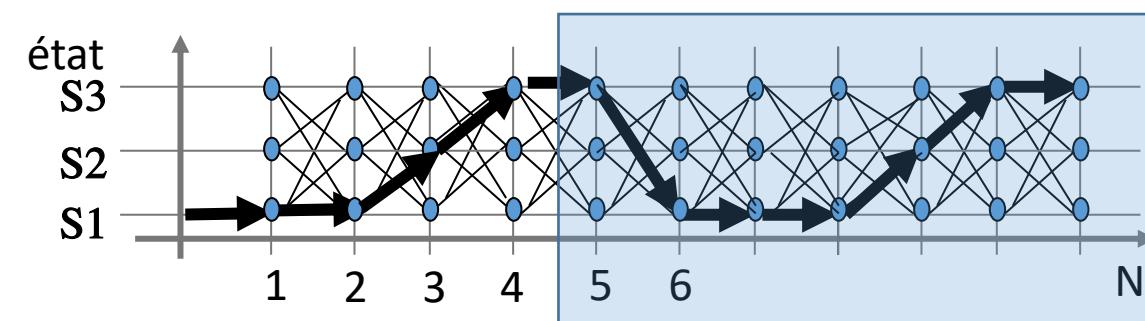
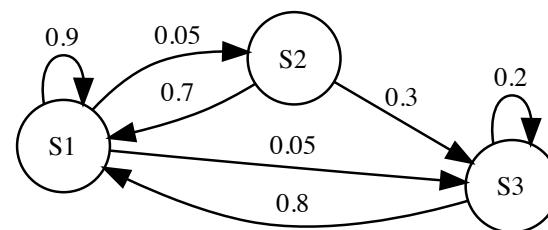
- La probabilité de réaliser un parcours fixé est le produit des probabilités des transitions situées sur le parcours



Probabilité d'un chemin

- Probabilité d'un chemin : Exemple

$$\begin{aligned}
 t &= 1, 2, 3, 4 \\
 O &= S1, S1, S2, S3 \\
 P(O|\pi, A) &= \pi_1 \times a_{1,1} \times a_{1,2} \times a_{2,3} \\
 \log(P(O|\pi, A)) &= \log(\pi_1) + \log(a_{1,1}) + \log(a_{1,2}) + \log(a_{2,3})
 \end{aligned}$$



Passer du pas n à n+1

■ Rappel :

- $P(X_{n+1} = j \mid X_n = i)$ est la probabilité de transition de l'état i à l'étape n vers l'état j à l'étape $n+1$
 - Souvent noté $a_{i,j}$
 - L'ensemble des transitions = la matrice $A = (a_{i,j})$

■ La probabilité $P(X_n = j)$

$$P(X_n = j) = \sum_{i=1}^K P(X_n = j \mid X_{n-1} = i) P(X_{n-1} = i)$$



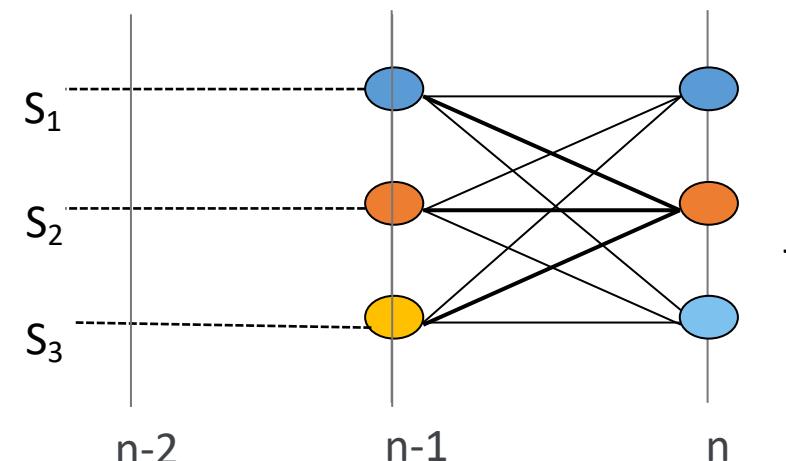
Passer du pas n à n+1

- La probabilité $P(X_n = j)$

$$P(X_n = j) = \sum_{i=1}^K P(X_n = j | X_{n-1} = i) P(X_{n-1} = i)$$

$$P(X_n = j) = \sum_{i=1}^K a_{j,i} P(X_{n-1} = i)$$

$$P(X_n = j) = a_{j,1} P(X_{n-1} = 1) + a_{j,2} P(X_{n-1} = 2) + \cdots + a_{j,k} P(X_{n-1} = k)$$



Passer du pas n à n+1

- Notation matricielle

- Soit le vecteur ligne de dimension k contenant la loi de la variable X_n , notons

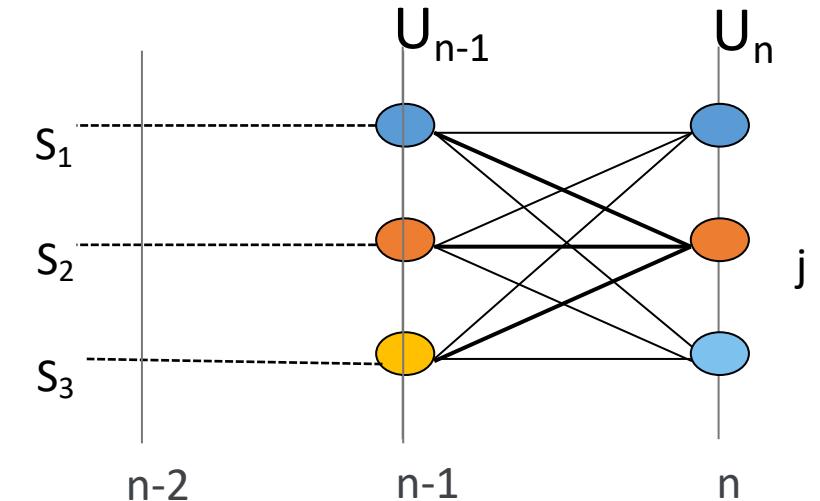
$$U_n = (P(X_n = 1), P(X_n = 2), \dots, P(X_n = k))$$

- On a :

$$U_n = U_{n-1} \times A$$

- Avec

$$U_0 = \pi$$



Passer du pas n à n+1

- On a : $U_n = U_{n-1} \times A$ avec $U_0 = \pi$

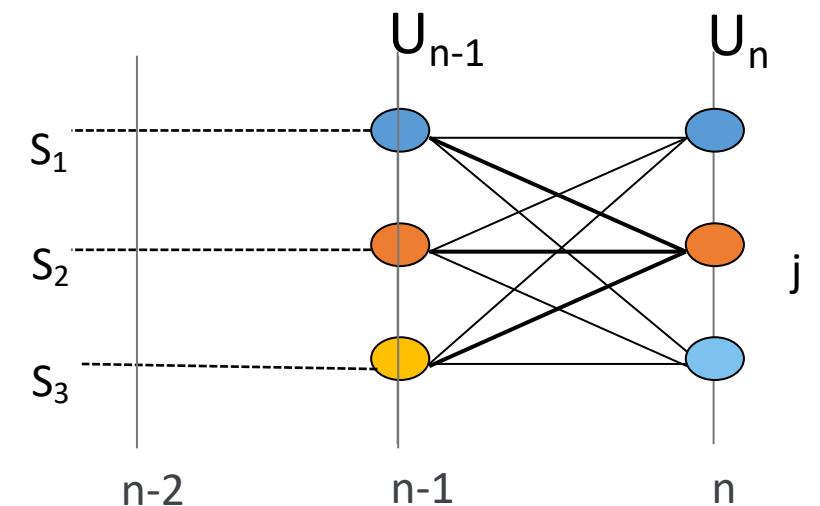
$$U_1 = U_0 \times A$$

$$U_2 = U_1 \times A$$

...

$$U_{n-1} = U_{n-2} \times A$$

$$U_n = U_{n-1} \times A$$



Passer du pas n à n+1

- On a : $U_n = U_{n-1} \times A$ avec $U_0 = \pi$

$$U_1 = U_0 \times A$$

$$U_2 = U_1 \times A$$

...

$$U_{n-1} = U_{n-2} \times A$$

$$U_n = U_{n-1} \times A$$



$$U_n = U_0 \times A^n$$

$$U_n = \pi \times A^n$$

- On a aussi

$$U_{n+m} = U_n \times A^m$$

Prédire les probabilités

- Prenons l'hypothèse que Doudou dort lors de la première minute de l'étude. $X_0 = \pi = [1, 0, 0]$.
 - Prédire les probabilités au bout d'une minute ?

$$P(X_1|A, \pi) = \pi \times A$$

- de deux minutes ?

$$P(X_2|A, \pi) = \pi \times A \times A$$

- à l'infini ?
- Cf notebook

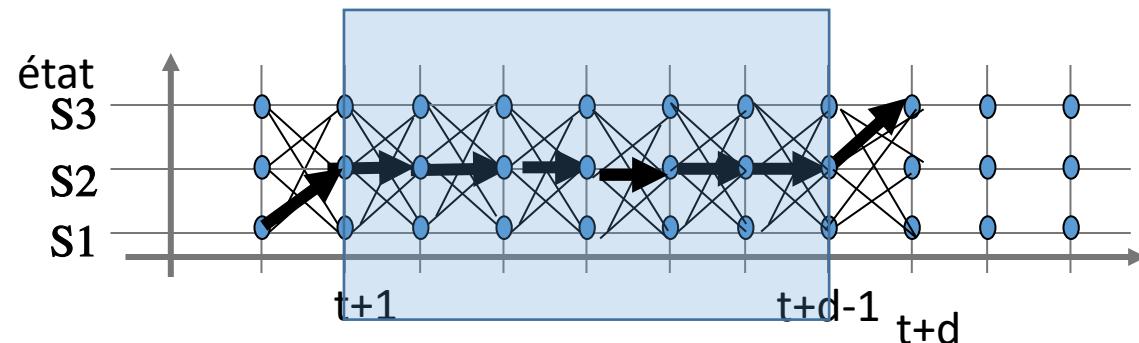


Probabilité de rester dans un état

- On suppose que nous sommes dans l'état S_3 « copeaux », quelle est la probabilité de rester d minutes ?

$$\begin{array}{ccccccccc} T & = & t & t+1, & t+2, & \dots, & t+d & t+d+1 \\ O & = & S_k \neq S_i & S_i, & S_i, & S_i, & S_i & S_j \neq S_i \end{array}$$

$$P(O|A, \pi, T = t \dots t+d) = (a_{i,i})^{d-1} \times (1 - a_{i,i})$$

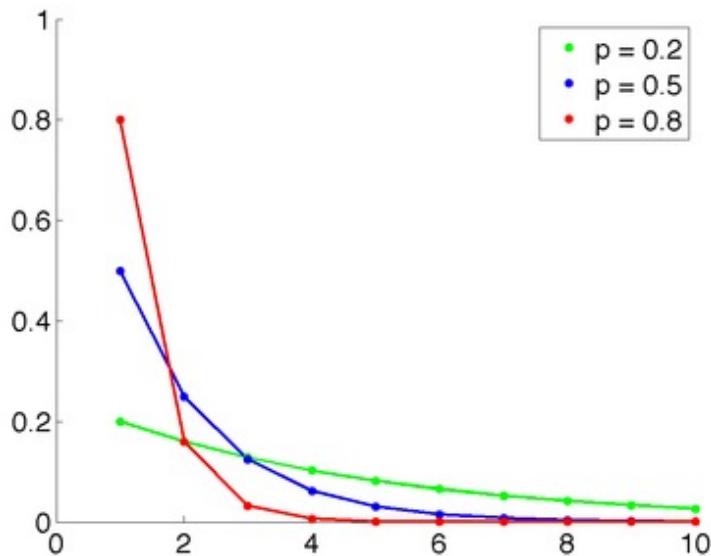


Probabilité de rester dans un état

- Durée dans un état

$$P(O|A, \pi, T = t \dots t + d) = (a_{i,i})^{d-1} \times (1 - a_{i,i})$$

- C'est une loi géométrique de paramètre $a_{i,i} = p$



Bibliographie

- Modèle de Markov :
 - <https://fr.wikipedia.org>
 - Lawrence Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, IEEE, vol 77, n° 2, feb 1989
 - **Lire l'introduction, section A et B**
- Doudou le hamster
 - <https://fr.wikipedia.org> chaine de markov
- Loi géométrique : cf <https://fr.wikipedia.org>



Chapitre 2

Chaîne de Markov

cachée



Modèle de Markov caché

- En anglais : Hidden Markov Model (HMM)
- HMM
 - un modèle statistique dans lequel le système modélisé est un **processus markovien** (chaine de Markov) de paramètres inconnus
 - sont largement utilisés en reconnaissance de formes, en intelligence artificielle ou en TALN



Bob et Alice

- Bob et Alice se téléphont tous les jours
- Bob a 3 activités : se promener, faire les courses ou faire du ménage
- Alice ne peut pas voir le temps qu'il fait chez Bob
- A partir de l'activité de Bob, Alice cherche à déterminer s'il pleut ou s'il fait beau



Bob et Alice

- Bob et Alice se téléphone tous les jours
- Bob a 3 activités : se promener (Walk), faire les courses (Shop) ou faire du ménage (Clean)
 - → Observations d'Alice à prendre dans {W, S, C}
 - Remarque : Observations discrètes
 - Alice a cette information
- A partir de l'activité de Bob, Alice cherche à déterminer s'il pleut ou s'il fait beau
 - → états = {r, s}



Bob et Alice : modèle (wikipedia)

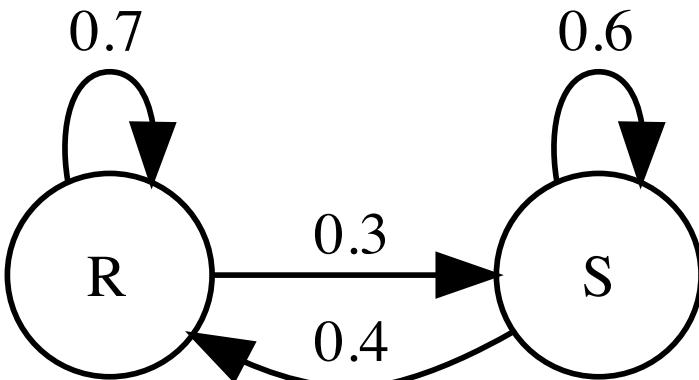
- états= ('Rainy', 'Sunny')
- Observations= ('walk', 'shop', 'clean')
- = {'Rainy': 0.6, 'Sunny': 0.4}
- Transition= {'Rainy': {'Rainy': 0.7, 'Sunny': 0.3}, 'Sunny': {'Rainy': 0.4, 'Sunny': 0.6}}
- Emission= {'Rainy': {'walk': 0.1, 'shop': 0.4, 'clean': 0.5}, 'Sunny': {'walk': 0.6, 'shop': 0.3, 'clean': 0.1}}



Probabilité du chemin

- A partir
 - des observations $O = \{W, W, S, C\}$
 - De la séquence d'états $Q = \{S, S, R, R\}$
- Quelle est la probabilité du chemin ?

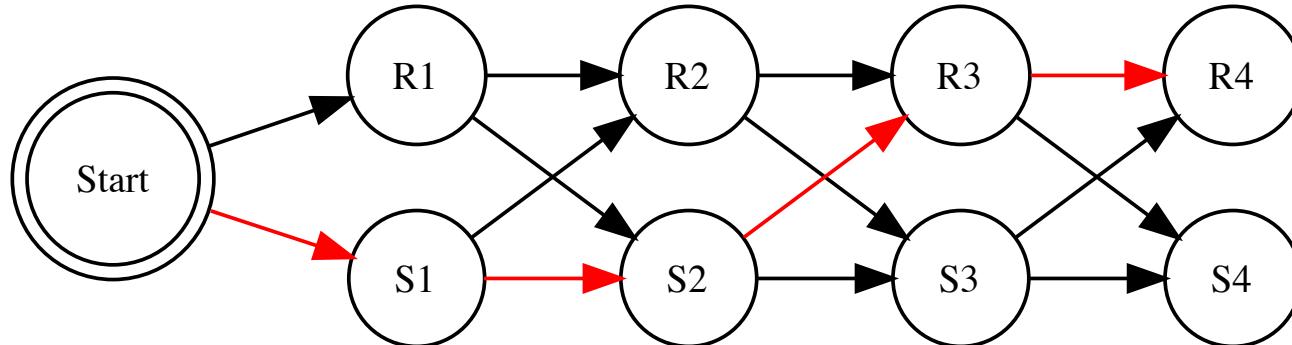
$$B_R = \begin{bmatrix} 0.1 \\ 0.4 \\ 0.5 \end{bmatrix} \begin{array}{l} \text{Walk} \\ \text{Shop} \\ \text{Clean} \end{array}$$



$$B_S = \begin{bmatrix} 0.6 \\ 0.3 \\ 0.1 \end{bmatrix} \begin{array}{l} \text{Walk} \\ \text{Shop} \\ \text{Clean} \end{array}$$

Probabilité du chemin

- A partir
 - des observations $O = \{W, W, S, C\}$
 - De la séquence d'états $Q = \{S, S, R, R\}$
- Quelle est la probabilité du chemin ?



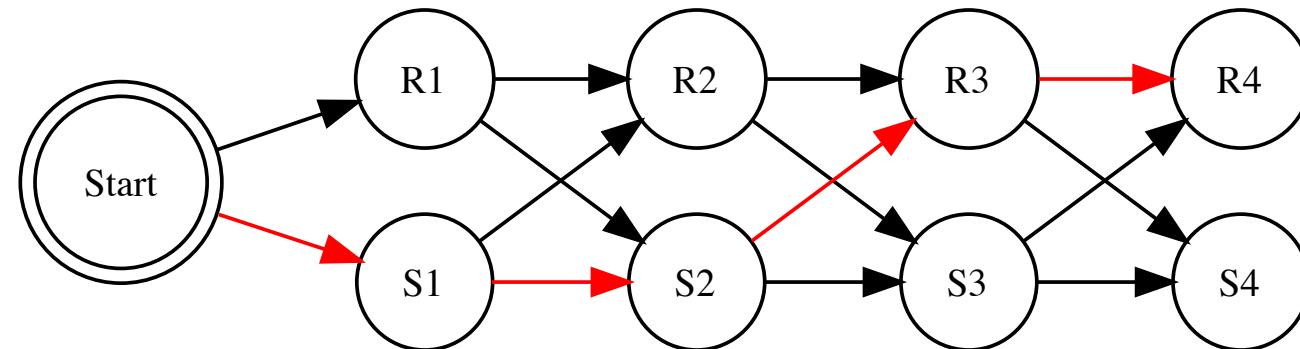
Probabilité du chemin

- Observation : $O = \{W, W, S, C\}$
- Séquence d'états : $Q = \{S, S, R, R\}$
- La liste des états : E
- Transitions : A
- Probabilités d'émissions : B
- Probabilités initiales : π

$$A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} \begin{matrix} R \\ S \end{matrix}$$

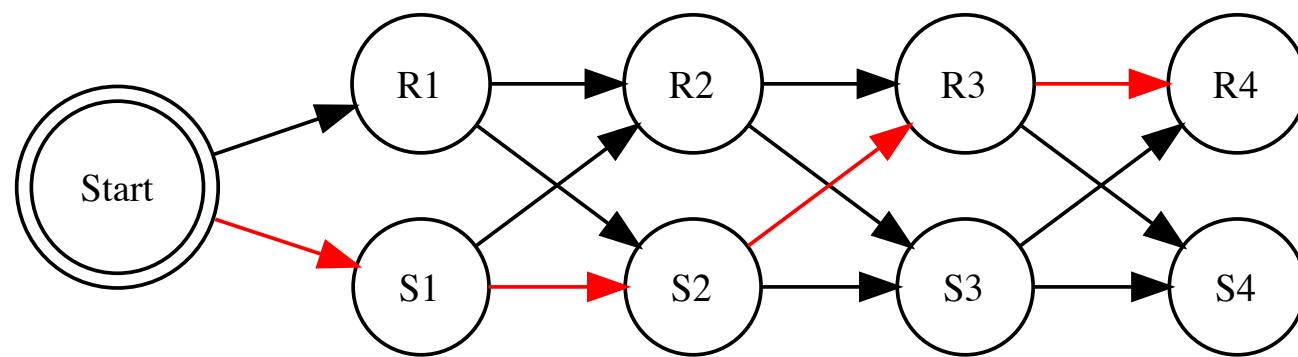
$$B = \begin{pmatrix} W & S & C \\ 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix} \begin{matrix} R \\ S \end{matrix}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix} \begin{matrix} R \\ S \end{matrix}$$



Probabilité du chemin

- Observation : $O = \{W, W, S, C\}$
- Séquence d'états : $Q = \{S, S, R, R\}$



$$1 = R \quad A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}^R_S$$

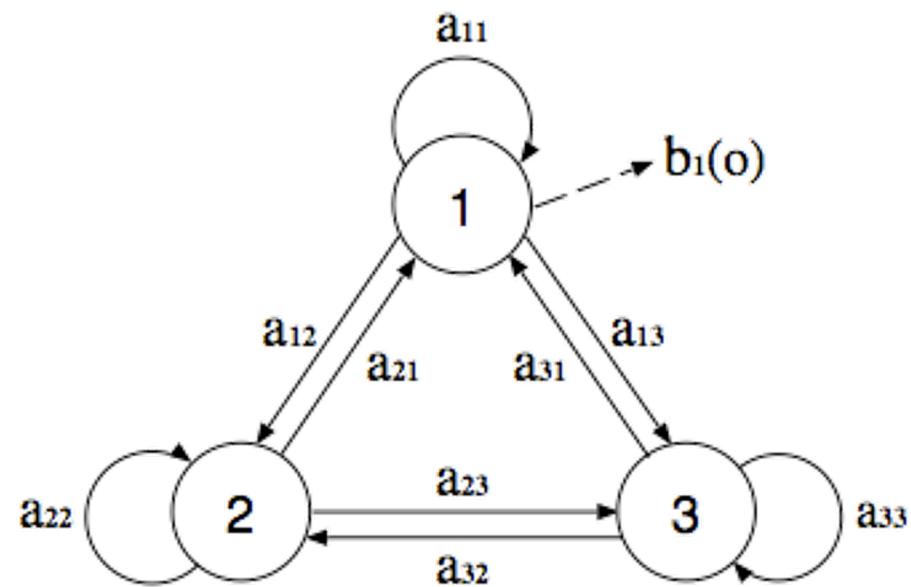
$$1 = R \quad B = \begin{pmatrix} W & S & C \\ 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}^R_S$$

$$\begin{aligned}
 P(O|Q, A, B, \pi) &= \pi_2 \times b_2(W) \times a_{2,2} \times b_2(W) \times a_{2,1} \times b_1(S) \times a_{1,1} \times b_1(C) \\
 &= 0.46 \times 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.4 \times 0.7 \times 0.5
 \end{aligned}$$

Modèle de Markov caché

- Définition : un graphe définit par
 - un ensemble d'états et une matrice de transitions A
 - les probabilités d'émissions B
 - Les probabilités initiales



Probabilité de transition

- L'ensemble des transitions = la matrice A = $(a_{i,j})$
- Propriétés
 - La somme des probabilités sortantes d'un état est égale à 1

$$\forall i, \sum_j a_{i,j} = 1$$

- C'est une matrice à coefficients positifs ou nuls



Probabilité initiale

- Notre graphe n'a pas : d'état d'entrée, d'état de sortie
- On considère qu'il existe une transition pour passer de l'état de départ à tous les états de la chaîne
 - Un vecteur de probabilité doit être fourni

$$\pi = [\pi_1, \pi_2, \dots, \pi_K]$$
$$\sum_{i=1}^K \pi_i = 1$$

- Ou à défaut, on considère que les probabilités sont équiprobables

$$\pi = \left(\frac{1}{K} \quad \frac{1}{K} \quad \dots \quad \frac{1}{K} \right)$$



Propriétés

Rappel

- La loi de X_{n+1} ne dépend de l'histoire X_0, X_1, \dots, X_n du système que de l'état de X_n
 - = pour prédire le futur X_{n+1} , je n'ai besoin que du présent X_n
 - = le résultat d'une épreuve ne dépend que du résultat de l'épreuve précédente.

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

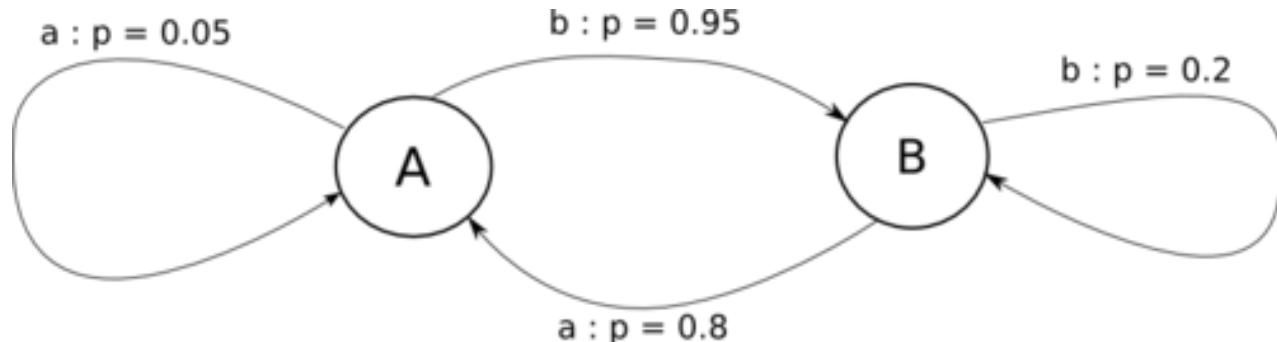
- Chaînes de Markov **homogènes** : le mécanisme de transition ne change pas au cours du temps

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i), \forall n \geq 0, \forall (i, j) \in E^2$$



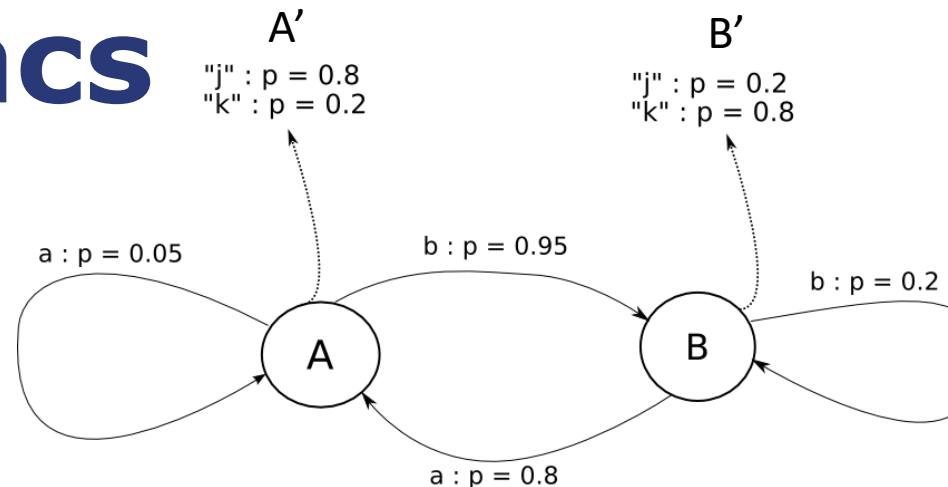
Exemple : jeux des sacs

- J'ai deux sacs noté A et B.
 - Dans le sac A, il y a 1 jeton a et 19 jeton b
 - Dans le sac B, il y a 4 jeton a et 1 jeton b
- Si je tire un jeton a, je replace le jeton et tire dans le sac A (idem pour b)
- Après n tirage : a b a b a b a a b a...



Exemple : jeux des sacs

- On ajoute 2 sacs A' et B'
 - A' contient quatre jetons j et un jeton k
 - B' contient un jeton j et quatre jetons k
- Le jeu
 - tirer un jeton dans A' , garder sa valeur, remettre le jeton
 - Tirer un jeton dans A pour connaître le groupe de sacs prochain
 - Recommencer autant de fois que le joueur le souhaite
- On génère deux séquences :
 - La sortie, connue, le résultat du jeu = une liste de k et j
 - La séquence des transitions, inconnue



Usage de HMM

■ 3 usage typiques

- 1/ Connaissant l'HMM, calculer la probabilité d'une séquence d'observation particulière
 - Algorithmes Forward et backward
- 2/ Connaissant l'HMM, trouver la séquence la plus probable d'état (caché, A & B) pour d'une séquence d'observation donnée
 - Algorithme de Viterbi
- 3/ Étant donné des séquences d'observations, calculer les probabilités de l'HMM
 - Algorithme de Baum-Welch (apprentissage du HMM, cas particulier d'EM)



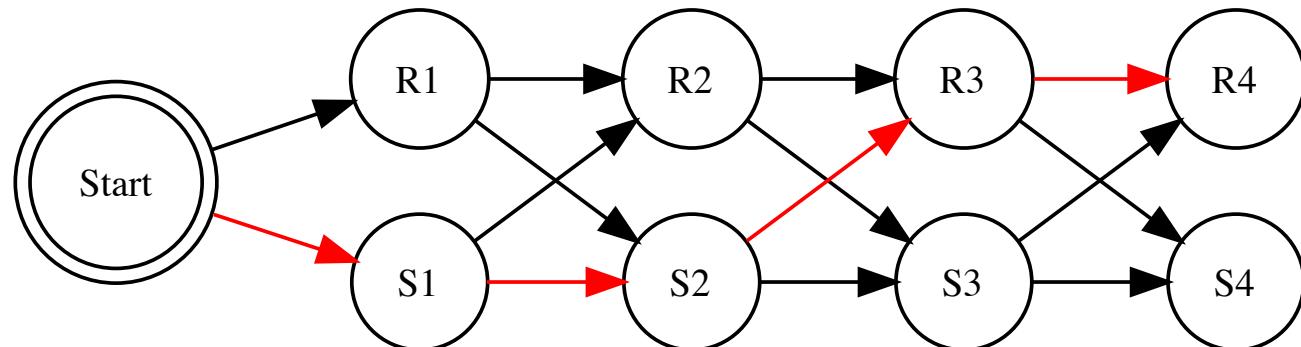
Rappel : probabilité du chemin

- Observation : $O = \{W, W, S, C\}$
- Séquence d'états : $Q = \{S, S, R, R\}$
- Transitions : A
- Probabilités d'émissions : B
- Probabilités initiales : π
- On note le modèle : $\lambda = (\pi, A, B)$

$$A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}_{S \times S}$$

$$B = \begin{pmatrix} W & S & C \\ 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}_{R \times S}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}_{S \times R}$$



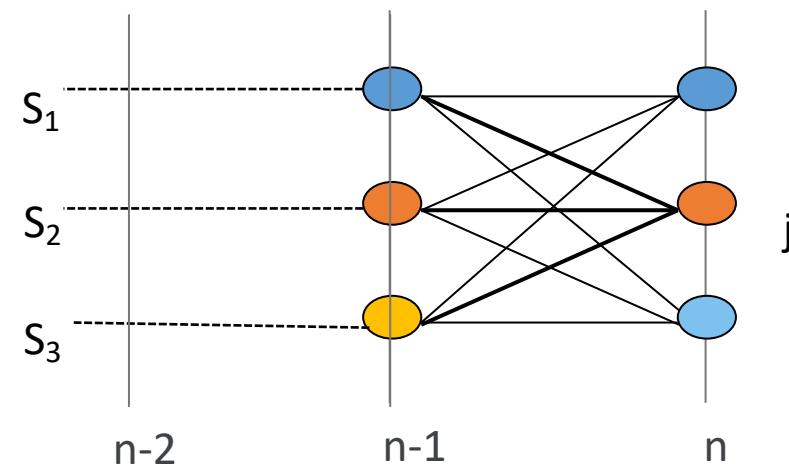
Rappel : Passer du pas n à n+1

- Pour une chaîne de Markov, la probabilité $P(X_n = j)$

$$P(X_n = j) = \sum_{i=1}^K P(X_n = j | X_{n-1} = i) P(X_{n-1} = i)$$

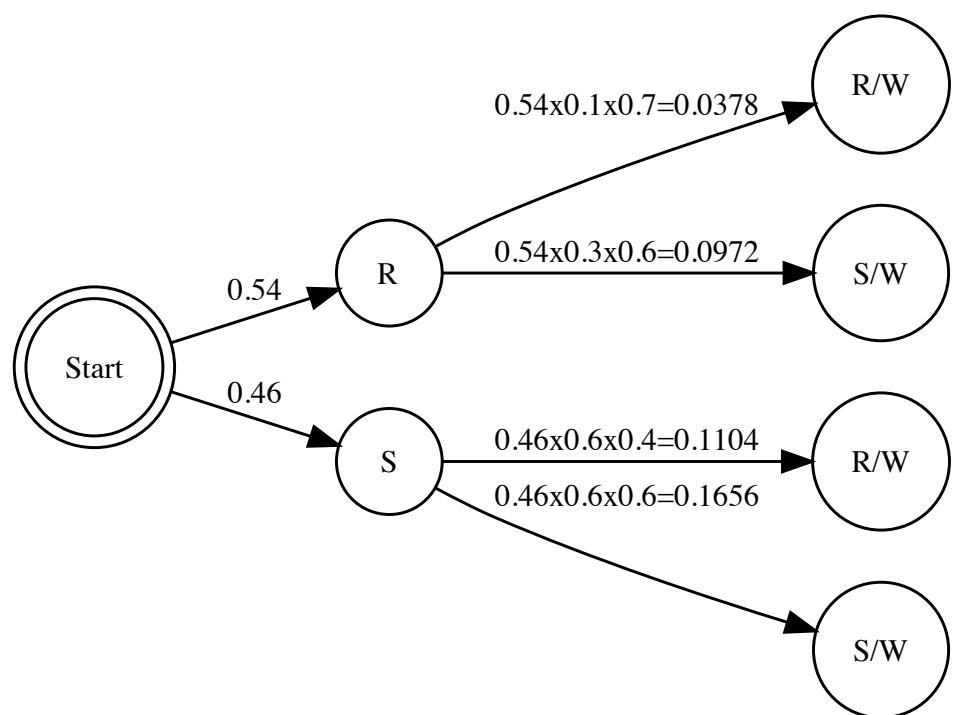
$$P(X_n = j) = \sum_{i=1}^K a_{j,i} P(X_{n-1} = i)$$

$$P(X_n = j) = a_{j,1} P(X_{n-1} = 1) + a_{j,2} P(X_{n-1} = 2) + \cdots + a_{j,k} P(X_{n-1} = k)$$



Usage HMM : principe

- Générer toutes les solutions :
 - Construire un arbre des solutions
 - parcours en profondeur



$$A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}_{\text{R}}_{\text{S}}$$

$$B = \begin{pmatrix} W & S & C \\ 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}_{\text{R}}_{\text{S}}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}_{\text{R}}_{\text{S}}$$

Usage HMM : principe

- Parcours en largeur basé sur le principe de la programmation dynamique
 - résoudre les sous-problèmes en stockant les résultats intermédiaires
- Utilise la propriété d'homogénéité

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

- Calculer les scores de l'instant t en fonction de $t-1$



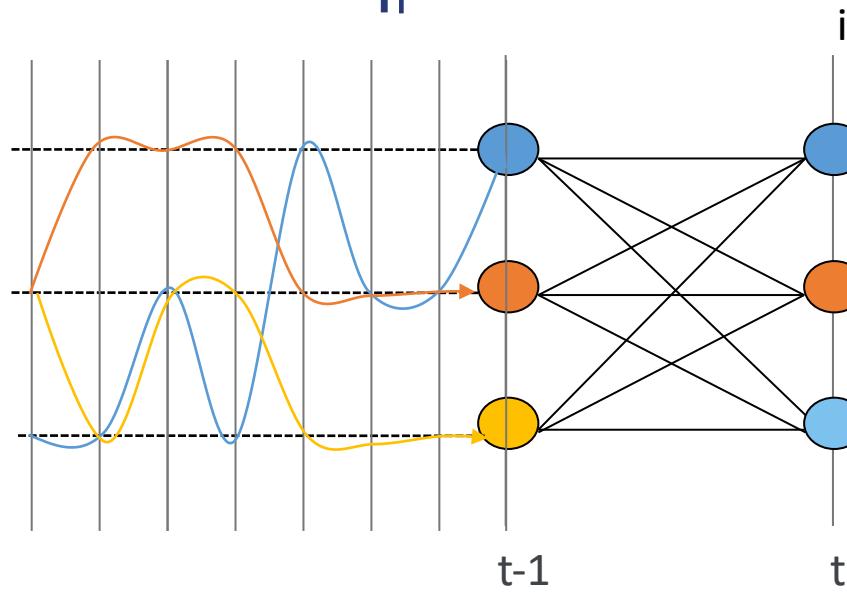
Forward Backward

- 1/ Connaissant l'HMM, calculer la probabilité d'une séquence d'observation particulière
 - Avec : $\lambda = (\pi, A, B)$ et $O = o_1, o_2, \dots, o_t$
 - Calculer $P(O|\lambda)$
 - Algorithme Forward, algorithme Backward



Forward

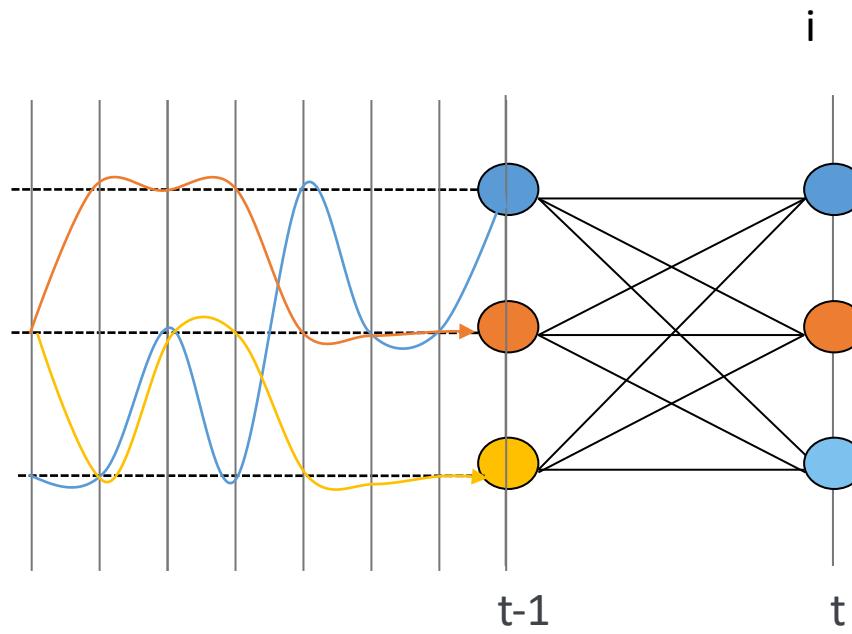
- La probabilité de la séquence d'états pour expliquer O à l'instant t dépend seulement de la probabilité de la séquence à $t-1$
- Pour chaque état q_i et à chaque instant i , sommer les probabilités arrivant en q_i



Forward

- On définit la variable suivante

$$\alpha_t(i) = P(o_1, \dots, o_t, q_t = i | \pi, A, B)$$

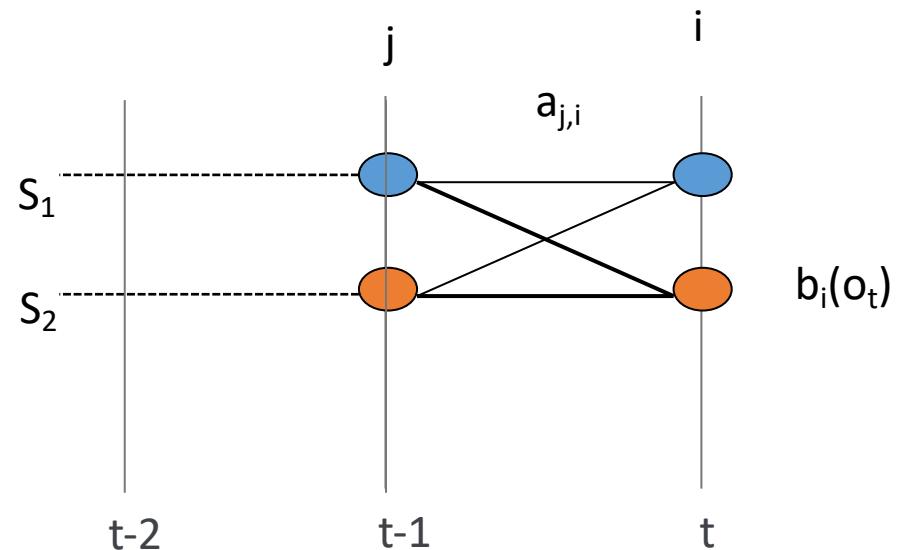


Forward

- On définit la variable suivante

$$\alpha_t(i) = P(o_1, \dots_N, o_t, q_t = i | \pi, A, B)$$

$$\alpha_t(i) = b_i(o_t) \sum_{j=1} \alpha_{t-1}(j) a_{j,i}$$



$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}_{S^R}$$

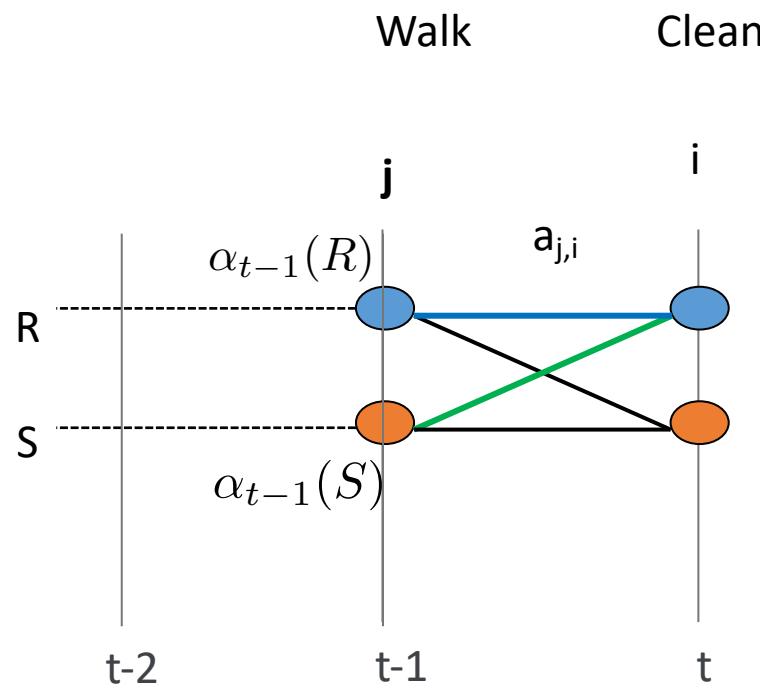
$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}_{S^R}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}_{S^R}$$

$$\alpha_t(i) = b_i(o_t) \sum_{j=1}^N \alpha_{t-1}(j) a_{j,i}$$

Observations=('walk','shop','clean')

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$



$$b_{i=R}(o_t = \text{Clean}) = 0.5$$

$$b_{i=S}(o_t = \text{Clean}) = 0.1$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

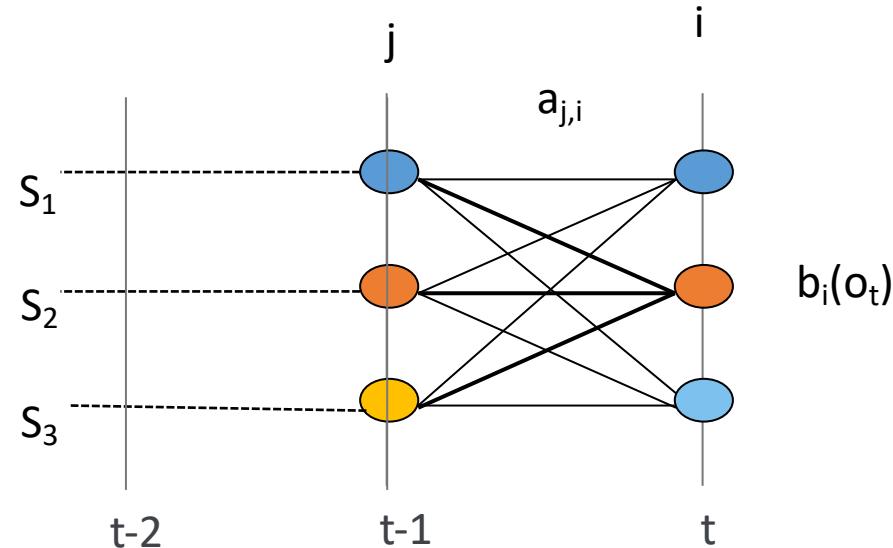
$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}$$

$$\alpha_t(R) = 0,5 \times (\alpha_{t-1}(R) \times 0,7 + \alpha_{t-1}(S) \times 0,4)$$

Forward

- On définit la variable suivante

$$\alpha_t(i) = b_i(o_t) \sum_{j=1}^N \alpha_{t-1}(j) a_{j,i} \quad P(O|\lambda) = \sum_{i=1}^N \alpha_T(i)$$

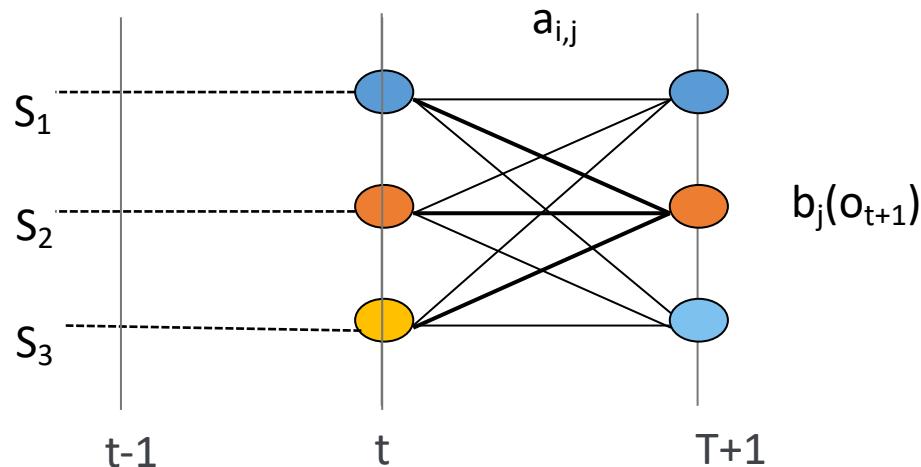


Backward

$$\alpha_t(i) = b_i(o_t) \sum_{j=1}^N \alpha_{t-1}(j) a_{j,i}$$

- Même principe mais la progression est faite de T vers 1
- Généralement le nom de la variable est $\beta_t(i)$

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_{i,j} b_j(o_{t+1})$$



Baum-Welch

- Calculer : $\alpha_t(i)$ $\beta_t(i)$
- estimer les coefficients $\gamma_t(i)$, la probabilité d'être dans l'état i à l'instant t sachant les observations et le modèle

$$\gamma_i(t) = P(X_t = i \mid Y, \theta) = \frac{P(X_t = i, Y \mid \theta)}{P(Y \mid \theta)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^N \alpha_j(t)\beta_j(t)}$$

$$\xi_{ij}(t) = P(X_t = i, X_{t+1} = j \mid Y, \theta) = \frac{P(X_t = i, X_{t+1} = j, Y \mid \theta)}{P(Y \mid \theta)} = \frac{\alpha_i(t)a_{ij}\beta_j(t+1)b_j(y_{t+1})}{\sum_{k=1}^N \sum_{w=1}^N \alpha_k(t)a_{kw}\beta_w(t+1)b_w(y_{t+1})}$$



Baum-Welch

- Estimer les paramètres du modèle
- L'exemple est donnée pour une séquence. Généralement, nous avons plusieurs séquences R

$$\pi_i^* = \gamma_i(1) \quad a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}$$

$$b_i^*(v_k) = \frac{\sum_{t=1}^T \mathbf{1}_{y_t=v_k} \gamma_i(t)}{\sum_{t=1}^T \gamma_i(t)}$$

$$\begin{aligned}\pi_i^* &= \frac{\sum_{r=1}^R \gamma_{ir}(1)}{R} \\ a_{ij}^* &= \frac{\sum_{r=1}^R \sum_{t=1}^{T-1} \xi_{ijr}(t)}{\sum_{r=1}^R \sum_{t=1}^{T-1} \gamma_{ir}(t)}, \\ b_i^*(v_k) &= \frac{\sum_{r=1}^R \sum_{t=1}^T \mathbf{1}_{y_{tr}=v_k} \gamma_{ir}(t)}{\sum_{r=1}^R \sum_{t=1}^T \gamma_{ir}(t)}\end{aligned}$$



Viterbi

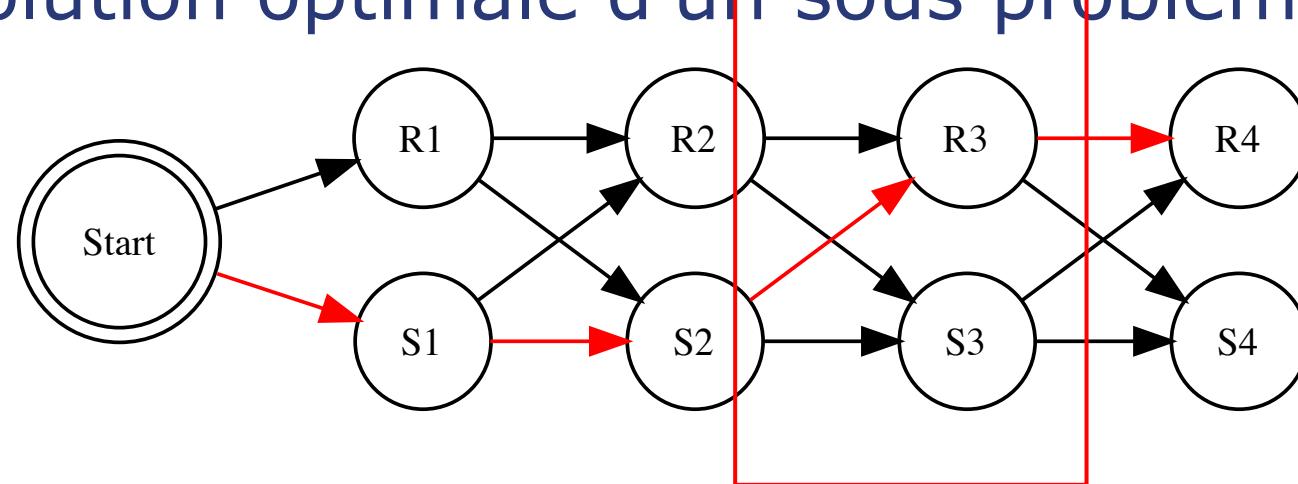
- 2/ Connaissant l'HMM, trouver la séquence la plus probable d'état (caché) pour d'une séquence d'observation donnée
 - Avec : $\lambda = (\pi, A, B)$ et $O = o_1, o_2, \dots, o_t$
 - Trouver la séquence $Q_i = \{q_1^i, q_2^i, \dots\}$ la plus probable parmi tous les séquences d'états possibles $\mathbb{Q} = \{Q_1, Q_2, \dots\}$

$$\max_{i \in \mathbb{Q}} P(Q_i | O, \lambda)$$



Viterbi

- La séquence d'états la plus probable pour expliquer O à l'instant t dépend seulement de la séquence la plus probable à $t-1$
- Pour chaque état et à chaque instant, garder l'état le plus probable
- Déduire la solution optimale d'un problème à partir d'une solution optimale d'un sous problème

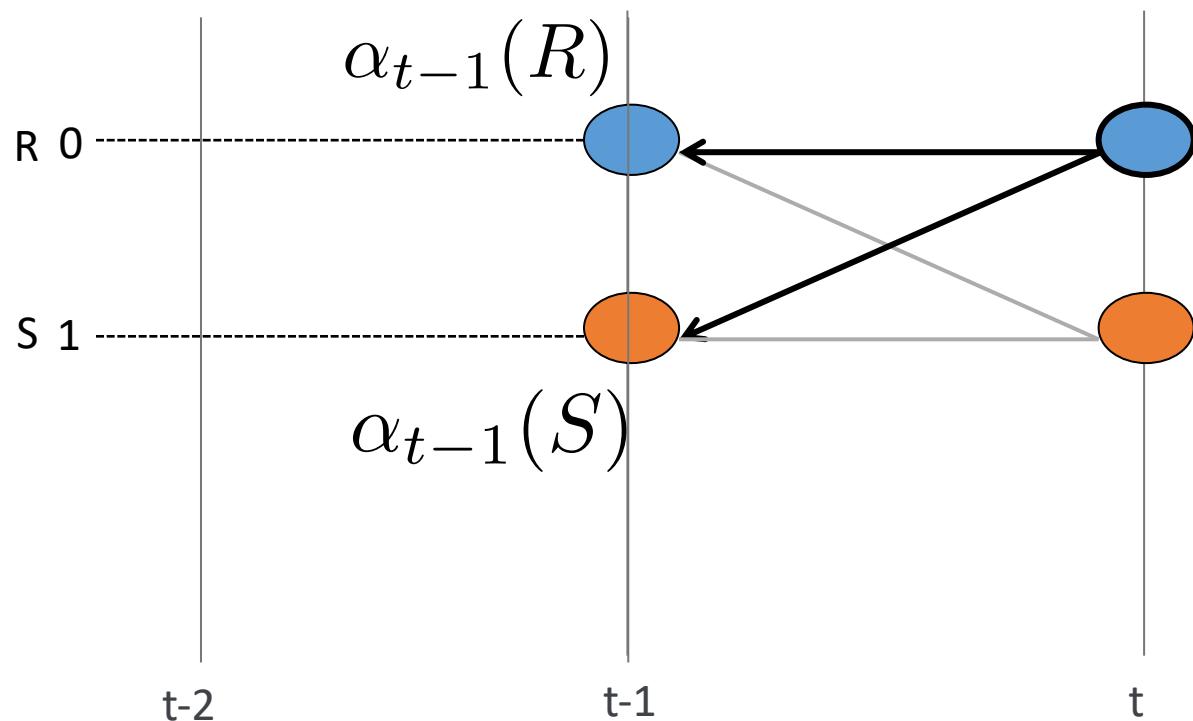


$$\alpha_t(i) = b_i(o_t) \max_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

$$\beta_t(i) = b_i(o_t) \operatorname{argmax}_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

$O_{t-1} = \text{Walk}$
j

$O_t = \text{Clean}$
i



0, 1, 2
Observations=('walk','shop','clean')

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$



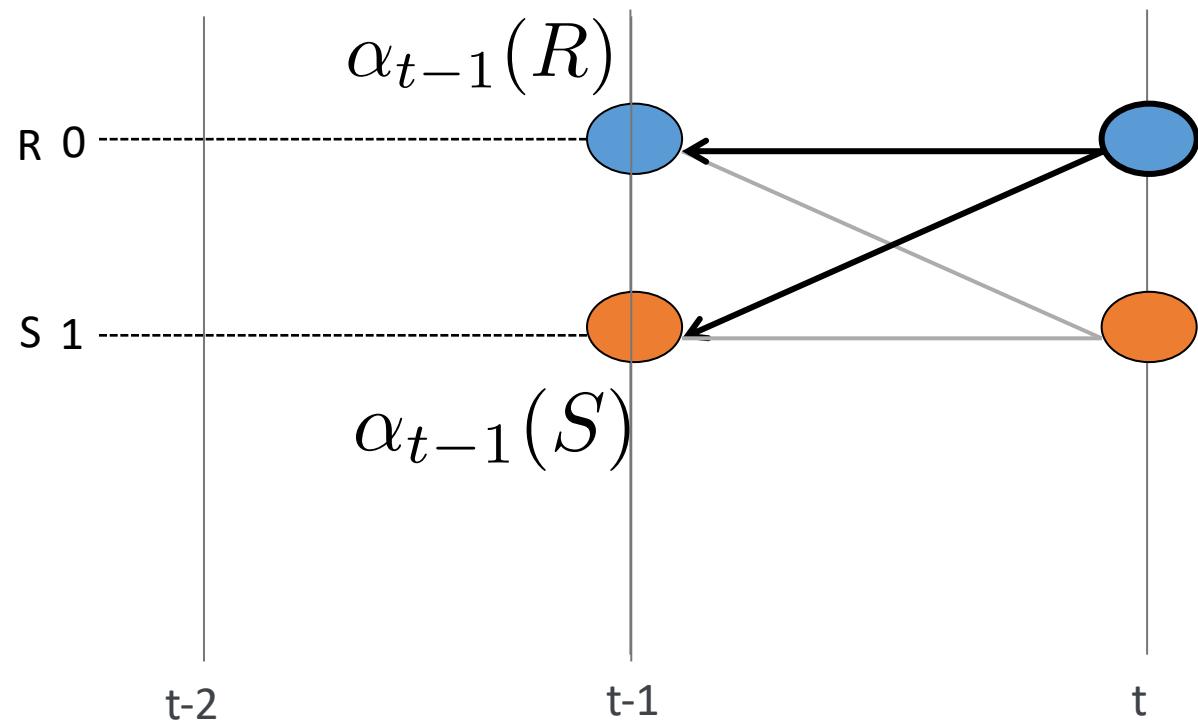
$$\alpha_t(i) = \boxed{b_i(o_t)} \max_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

0, 1, 2
Observations=('walk','shop','clean')

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

$O_{t-1} = \text{Walk}$ $O_t = \text{Clean}$
 j i



$$\alpha_t(R) = \boxed{0.5} \times \boxed{0.7} \times \alpha_{t-1}(R)$$

$$\alpha_t(R) = \boxed{0.5} \times \boxed{0.4} \times \alpha_{t-1}(S)$$



$$\alpha_t(i) = \boxed{b_i(o_t)} \max_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

0, 1, 2
Observations=('walk','shop','clean')

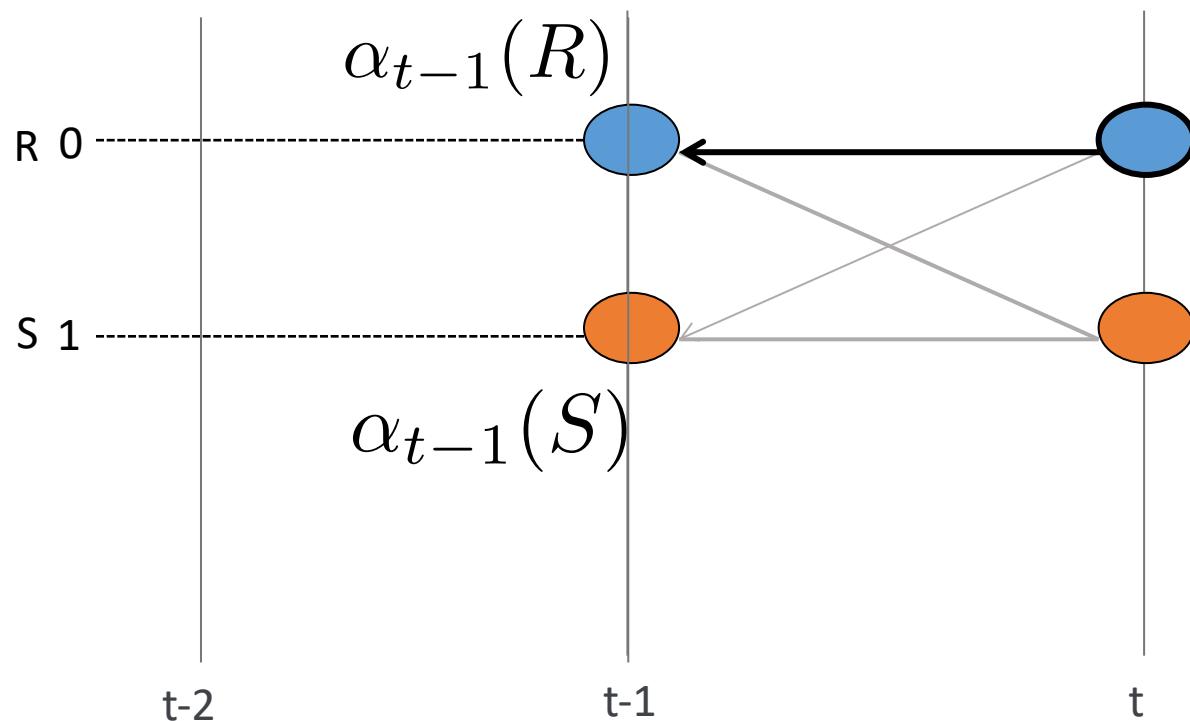
$$\beta_t(i) = b_i(o_t) \operatorname{argmax}_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

$O_{t-1} = \text{Walk}$
 j

$O_t = \text{Clean}$
 i

$$A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$



$$\alpha_t(R) = \boxed{0.5} \times \boxed{0.7} \times \alpha_{t-1}(R)$$

$$\alpha_t(R) = \boxed{0.5} \times \boxed{0.4} \times \alpha_{t-1}(S)$$

$$\beta_t(R) = R$$



$$\alpha_t(i) = \boxed{b_i(o_t)} \max_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

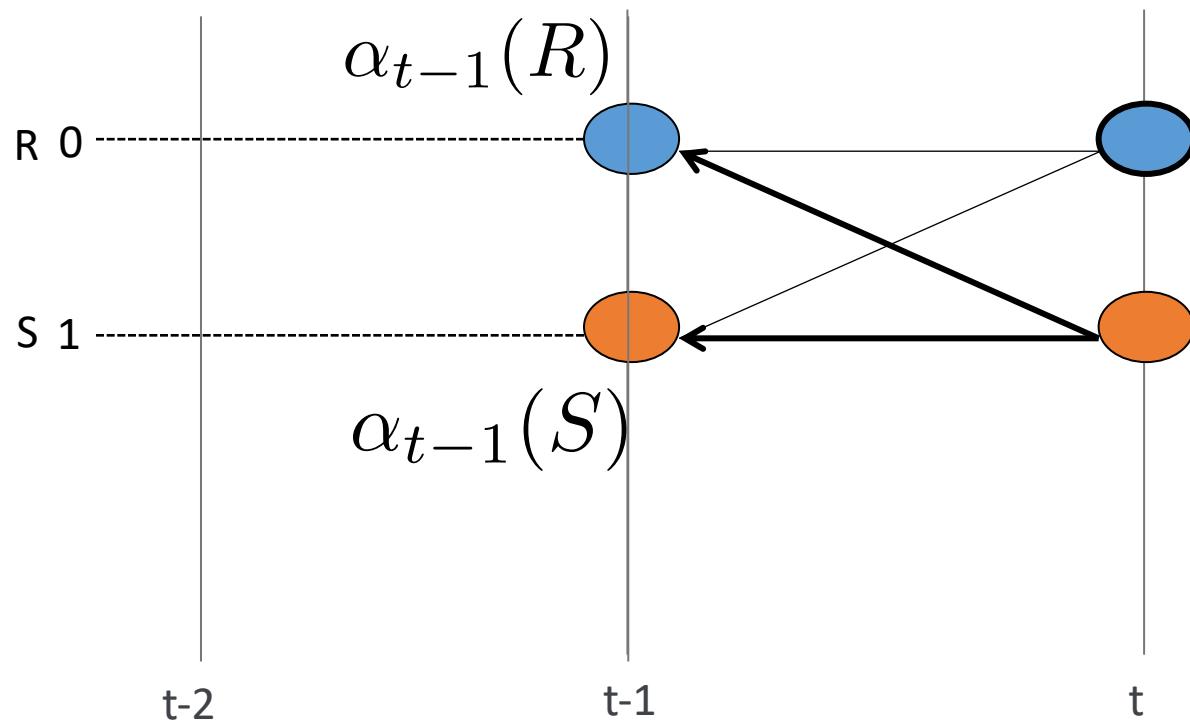
0, 1, 2
Observations=('walk','shop','clean')

$$\beta_t(i) = b_i(o_t) \operatorname{argmax}_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

$$\begin{array}{c} O_{t-1} = \text{Walk} \\ j \\ \hline O_t = \text{Clean} \\ i \end{array}$$

$$A = \begin{pmatrix} R & S \\ \boxed{0.7} & 0.3 \\ 0.4 & \boxed{0.6} \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & \boxed{0.1} \end{pmatrix}$$



$$\begin{aligned} \alpha_t(S) &= \boxed{0.1} \times \boxed{0.4} \times \alpha_{t-1}(R) \\ \alpha_t(S) &= \boxed{0.1} \times \boxed{0.3} \times \alpha_{t-1}(S) \end{aligned}$$



$$\alpha_t(i) = \boxed{b_i(o_t)} \max_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

0, 1, 2
Observations=('walk','shop','clean')

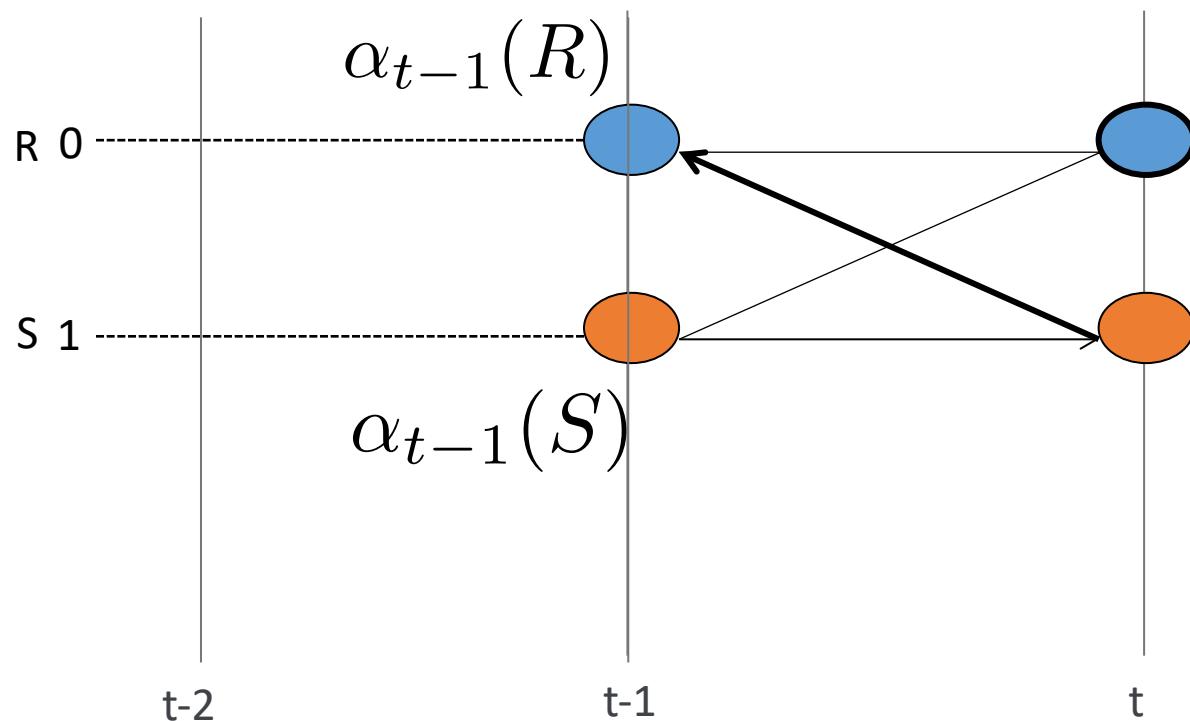
$$\beta_t(i) = b_i(o_t) \operatorname{argmax}_{j=1}^N a_{j,i} \times \alpha_{t-1}(i)$$

$O_{t-1} = \text{Walk}$
 j

$O_t = \text{Clean}$
 i

$$A = \begin{pmatrix} R & S \\ 0.7 & 0.3 \\ \boxed{0.4} & 0.6 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & \boxed{0.1} \end{pmatrix}$$



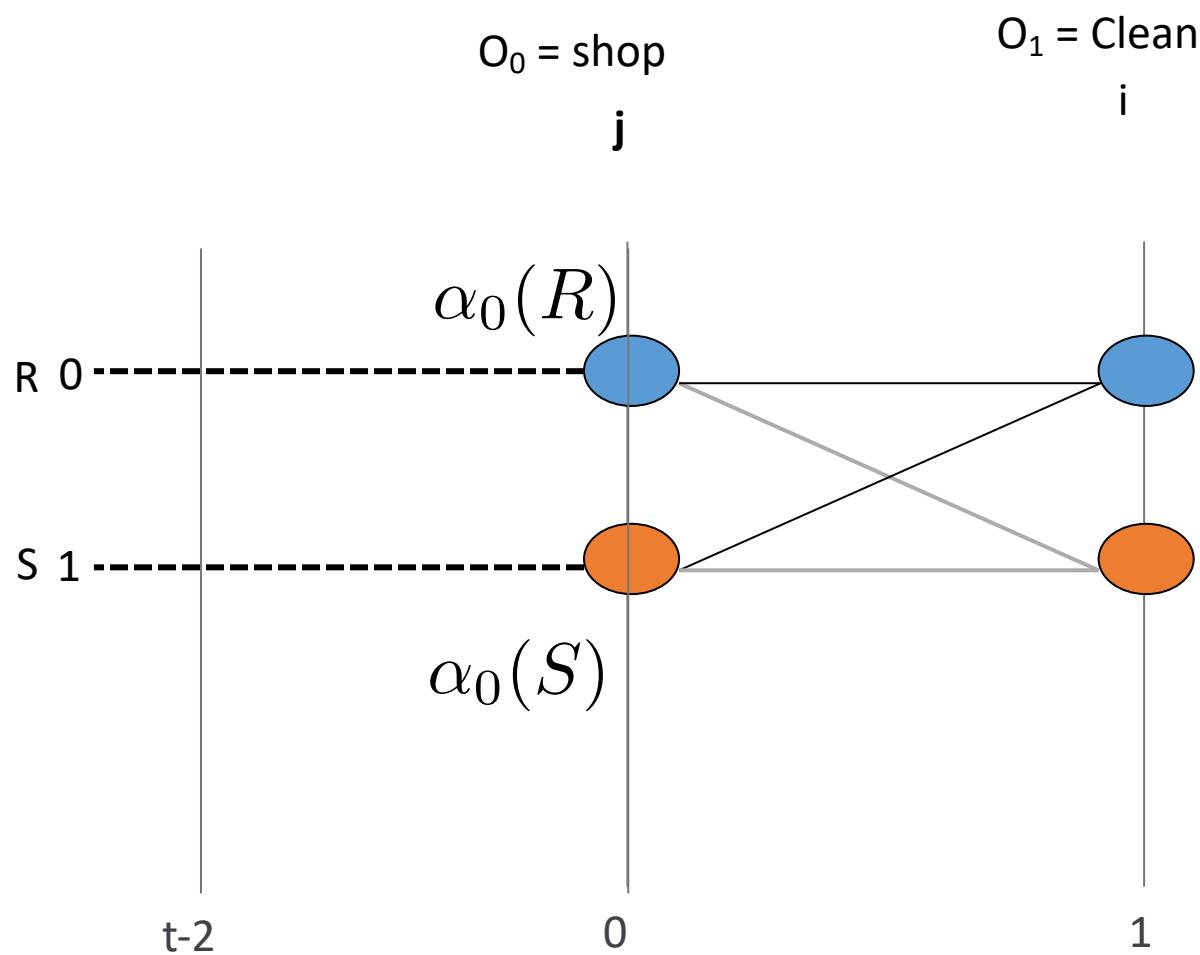
$$\alpha_t(S) = \boxed{0.1} \times \boxed{0.4} \times \alpha_{t-1}(R)$$

$$\alpha_t(S) = \boxed{0.1} \times \boxed{0.3} \times \alpha_{t-1}(S)$$

$$\beta_t(S) = R$$



$$\alpha_0(i) = \pi_i \times b_i(O_0)$$



Observations= ('walk','shop','clean')

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

$$\pi = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix}$$

$$\alpha_0(R) = 0.54 \times 0.4$$

$$\alpha_0(S) = 0.46 \times 0.3$$



Chapitre 3 : étiquettagé



Problème d'étiquettes

- Nous avons en entrée une phrase
 - L'astrophysicien Stephen Hawking est mort à 76 ans
- Nous souhaitons annoter le texte en fonction de :
 - Catégorie grammaticale (POS tagging)
 - DET NC NP NP V ADJ PREP NUM NC
 - Entités nommées
 - [L'astrophysicien] [Stephen Hawking] est mort à [76 ans]
 - FCT-B FCT-I PERS-B PERS-I O O O NUM-B NUM-I
- La séquence générée à la même longueur
- Problème d'étiquettage = associer à chaque mot une étiquette



Apparté

- L'astrophysicien Stephen Hawking est mort à 76 ans
- 1^{er} version du linguiste :
 - L' → déterminant
 - astrophysicien → substantif
 - Stephen Hawking → nom propre
 - est mort → verbe au passé composé
 - à → préposition
 - 76 → déterminant numéral cardinal
 - ans → substantif



Apparté

- L'astrophysicien Stephen Hawking est mort à 76 ans
- 2^{er} version du linguiste :
 - L' → déterminant
 - astrophysicien → substantif
 - Stephen Hawking → nom propre
 - est → verbe
 - mort → adjectif attribut
 - à → préposition
 - 76 → déterminant numéral cardinal
 - ans → substantif

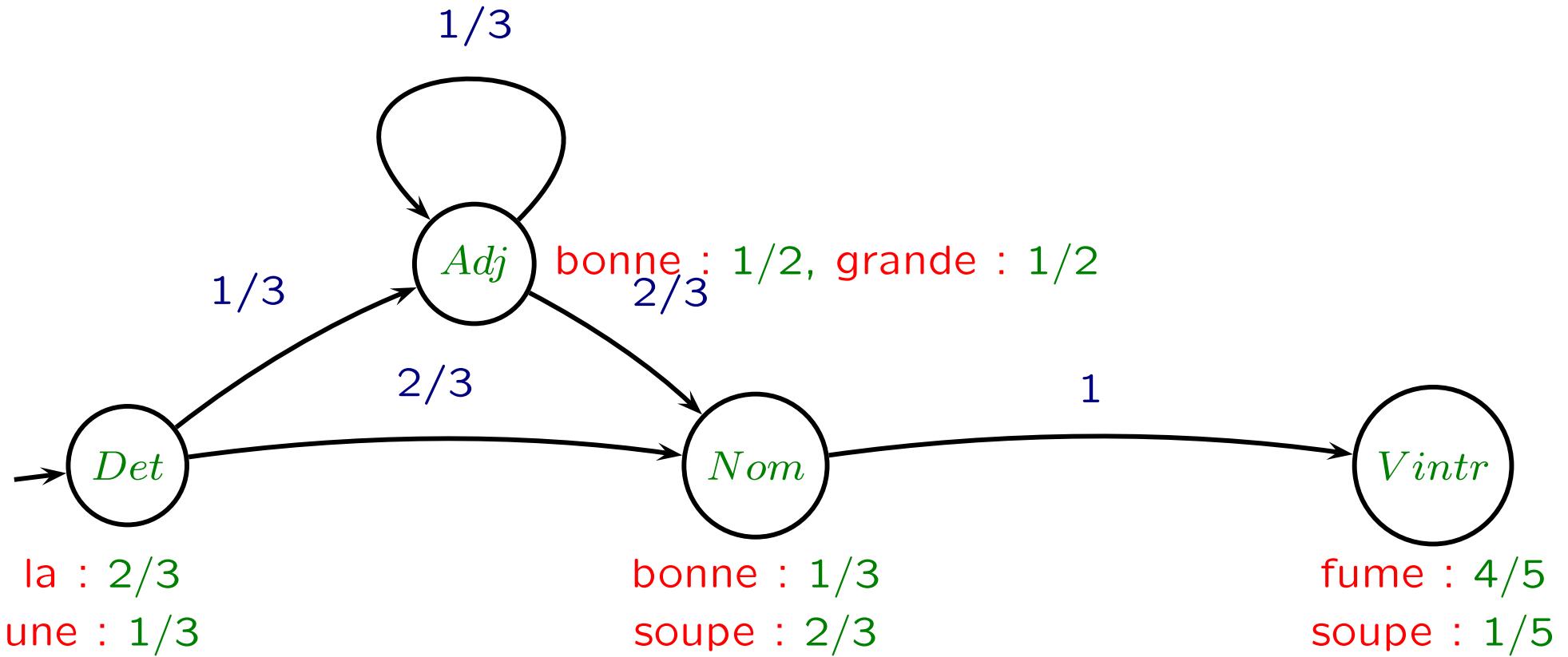


Problème d'étiquettes

- Attention pour une même tache le jeux d'étiquettes peut être différent d'un outils à l'autre !
- Problème d'ambiguité des étiquettes
- Besoin de données d'apprentissage étiquetées
- Mots
 - Tous connus
 - Mots peu fréquent
 - Un même mot peut avoir plusieurs étiquettes
 - Ex : paris

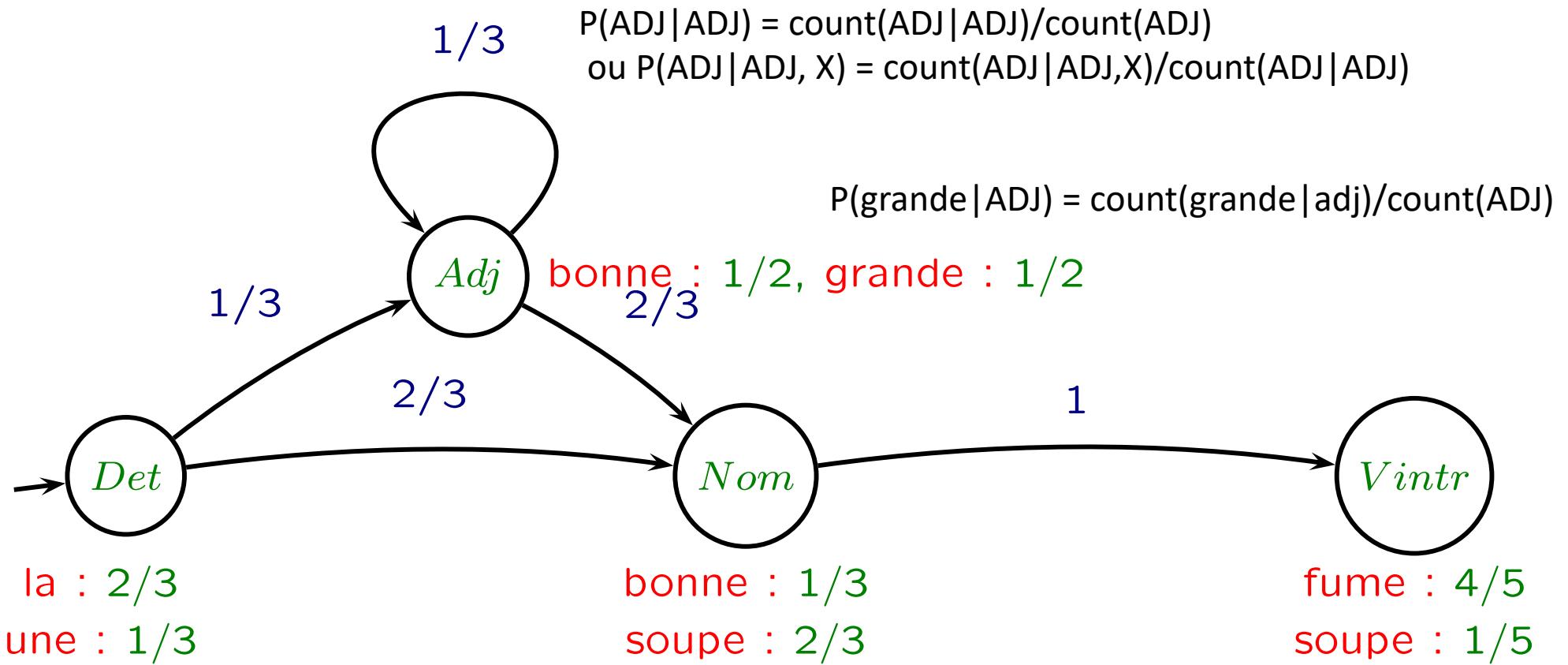


HMM et étiquettes



Source : cours Isabelle Tellier – Paris 3

HMM, étiquettes et probabilités



Source : cours Isabelle Tellier – Paris 3

Entités nommées

- = une expression linguistique référentielle
 - Souvent associée aux noms propres et au nom commun (ou locution nominale) décrivant un individu ou un objet déterminé unique.
- Objets considérés :
 - Les personnes (ou anthroponymes) entités humaines, réelles ou fictives, contemporaines ou historiques,
 - Lieux (ou toponymes) : entités localisées géographiquement,
 - Organisations (ou ergonymes) : sociétés, institutions, gouvernements, etc.
 - Les dates, quantités,



Lecture

- <http://www.cs.columbia.edu/~mcollins/hmms-spring2013.pdf>
- Youtube : [NLP - 05. Tagging Problems, and HMM](#)



Chapitre 4

Modèles gaussiens



Gaussienne à une dimension

- Distribution gaussienne spécifiée par 2 paramètres

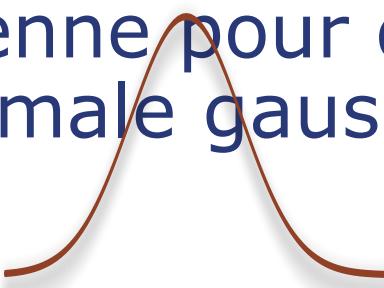
- la moyenne

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- σ : son écart type, σ^2 la variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$$

- On parlera de gaussienne pour désigner une distribution de loi normale gaussienne



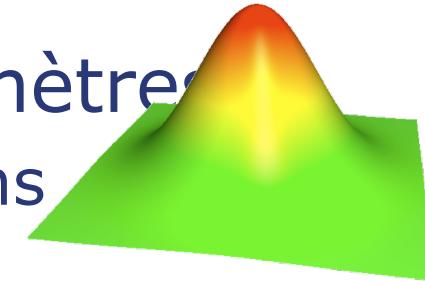
Gaussienne à d dimensions

- Distribution spécifiée par 2 paramètres
 - Moyenne, un vecteur à D dimensions

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_d \end{bmatrix}$$

avec

$$\mu_i = \frac{1}{N} \sum_{k=1}^N x_i^k$$



- Matrice de covariance

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \cdots & \cdots & \cdots & \sigma_{1,d} \\ \vdots & \ddots & & & \vdots \\ \vdots & & \sigma_{i,j} & \sigma_{i,i} & \vdots \\ \vdots & & & \ddots & \vdots \\ \sigma_{d,1} & \cdots & \cdots & \cdots & \sigma_{d,d} \end{bmatrix}$$

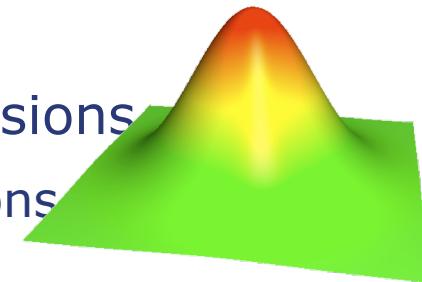
avec

$$\sigma_{i,j} = \frac{1}{N} \sum_{k=1}^N x_i^k x_j^k - \mu_i \mu_j$$



Gaussienne à N dimensions

- Cas plein : Σ une matrice à $N \times N$ dimensions
 - Exprime la corrélation entre les dimensions
 - Matrice symétrique, positive
- Cas diagonal : Σ une matrice à $N \times N$ dimensions avec des 0 sauf sur la diagonale



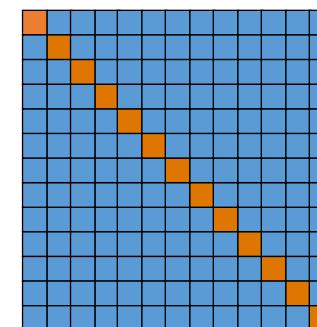
■ Exemple

- 13 attributs = dimensions 13

La moyenne :
Vecteur à 13 dimensions

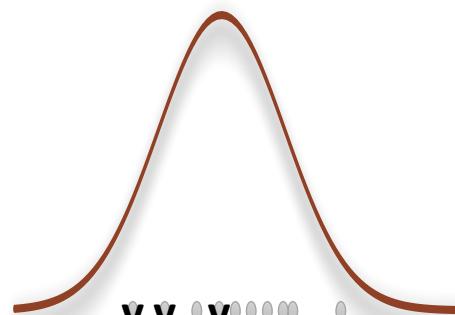
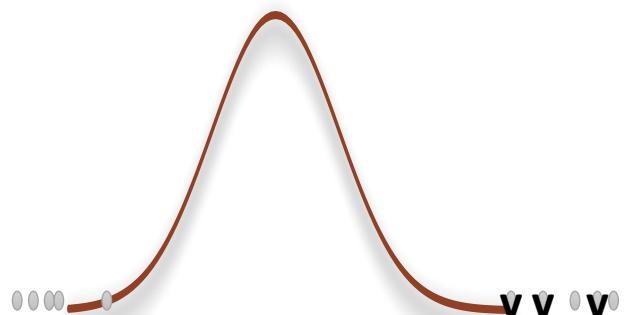


La covariance :
Une matrice à
 13×13 dimensions



Gaussienne et vraisemblance

- Un échantillon d'observations issues d'une
- distribution
 - Peu vraisemblable que la distribution soit à l'origine de l'échantillon
 - Les observations sont dans une région où la densité de probabilité est faible
 - → petite valeur
- Vraisemblable que l'échantillon soit issu de la distribution
- Les observations sont dans une région où la densité de probabilité est forte
- → grande valeur



Gaussienne et vraisemblance

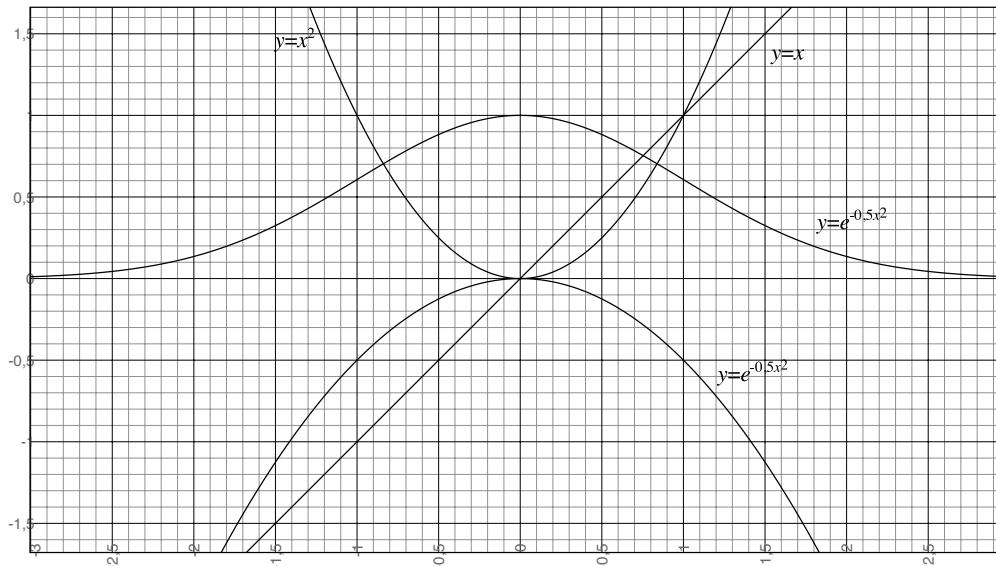
■ Pour une gaussienne

$$L(x|\mu, \Sigma) = \frac{1}{(2\Pi)^{D/2}|\Sigma|^{1/2}} e^{\frac{-1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}$$

■ Remarques

$$L(x|\mu, \Sigma) > 0$$

- Plus $L(o|\mu, \Sigma)$ est grand plus il est vraisemblable que x soit issu de la gaussienne



Mixture de gaussiennes

- Une combinaison linéaire de M gaussiennes

$$X = \{w_i, \mu_i, \Sigma_i\}, \forall i \in 1 \dots M$$

- w_i : le poids de la $i^{\text{ème}}$ composante

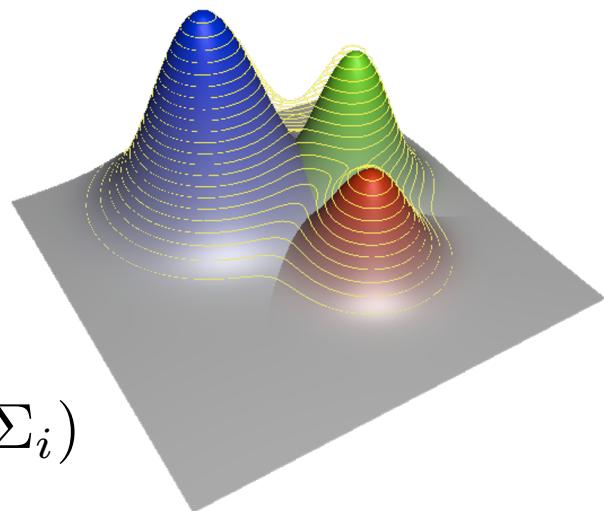
$$\sum_{i=1}^M w_i = 1$$

- La log vraisemblance

$$\log L(o|X) = \sum_{i=1}^M w_i \log L(o|\mu_i, \Sigma_i)$$

- Remarques

- La gaussienne peut être multidimensionnelle ou non
- Les matrices de covariance peuvent être diagonales ou pleines
- On parle de GMM : Gaussien Mixture Model



Classification par modèle

- On va remplacer
 - le centre de gravité par une gaussienne de paramètres μ et Σ
 - La distance euclidienne par la vaissamblance entre la gaussienne et chaque individu



Extension K-mean

- Étapes 1 :
 - Affecter chaque individu à la classe avec la plus grande vraisemblance
- Étapes 2 :
 - Mettre à jour les gaussiennes de chaque classe
 - En fonction des paramètres affectés, calculer la moyenne et la covariance
 - Le poids sera de : nombre d'individus affectés à la classe / K
- Test d'arrêt
 - Reprendre à l'étape 1, tant que la vraisemblance entre deux itérations augmente de plus de ϵ
 - Sinon, on stop



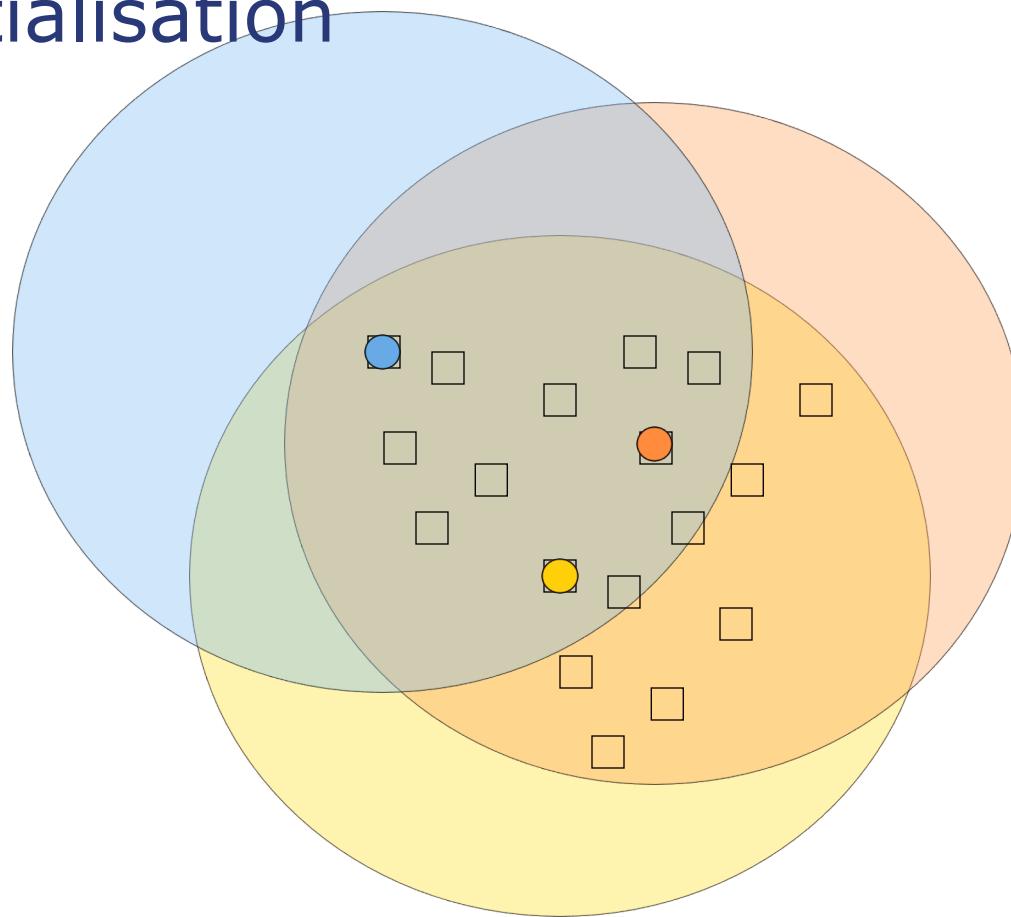
Exemple

- Initialisation

$$\mathcal{W} = 1/3$$

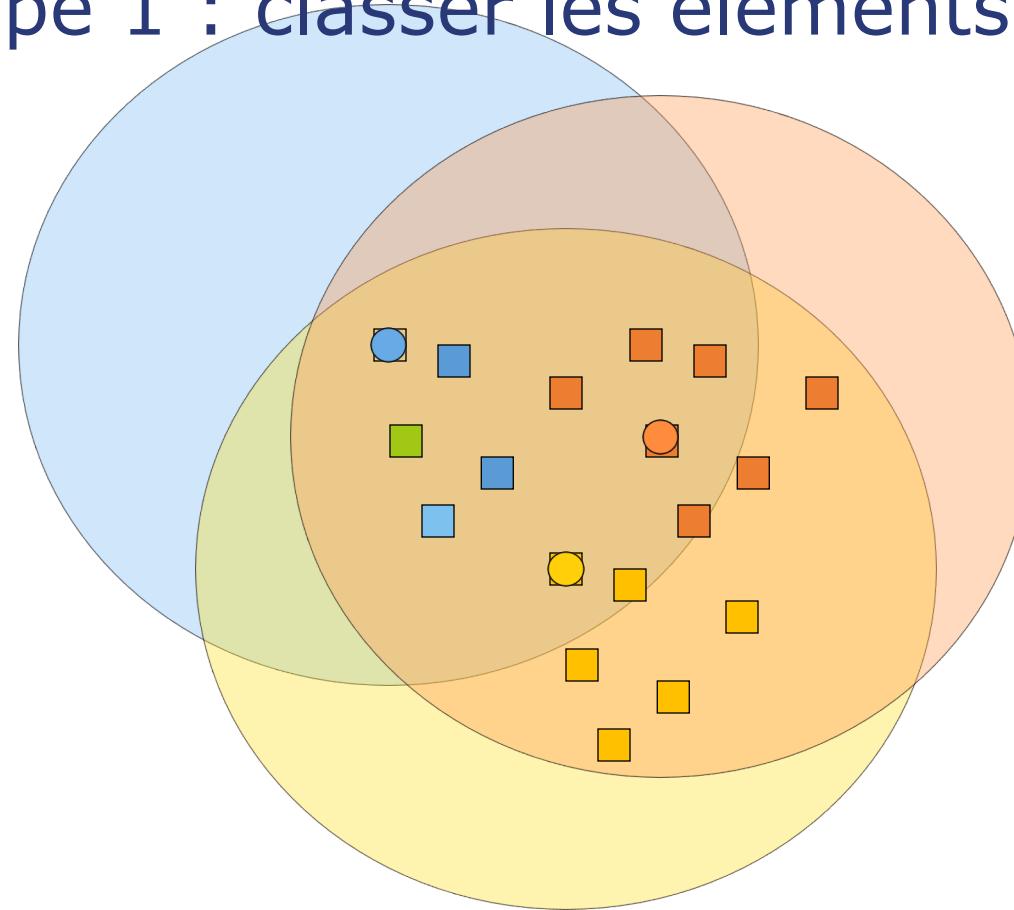
$$\mathcal{W} = 1/3$$

$$\mathcal{W} = 1/3$$



Exemple

- Etape 1 : classer les éléments



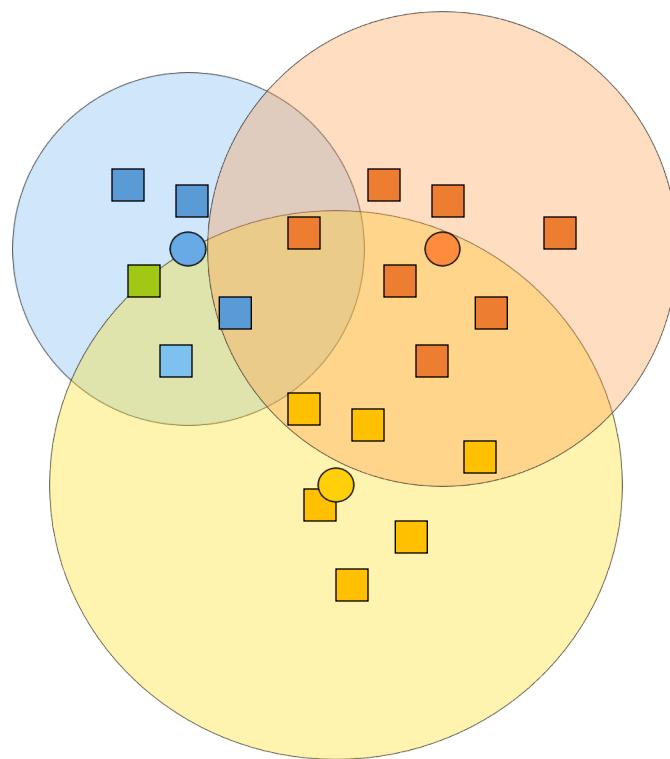
Exemple

- Etape 2 : mettre à jour les modèles

$$\mathcal{W} = 4/18$$

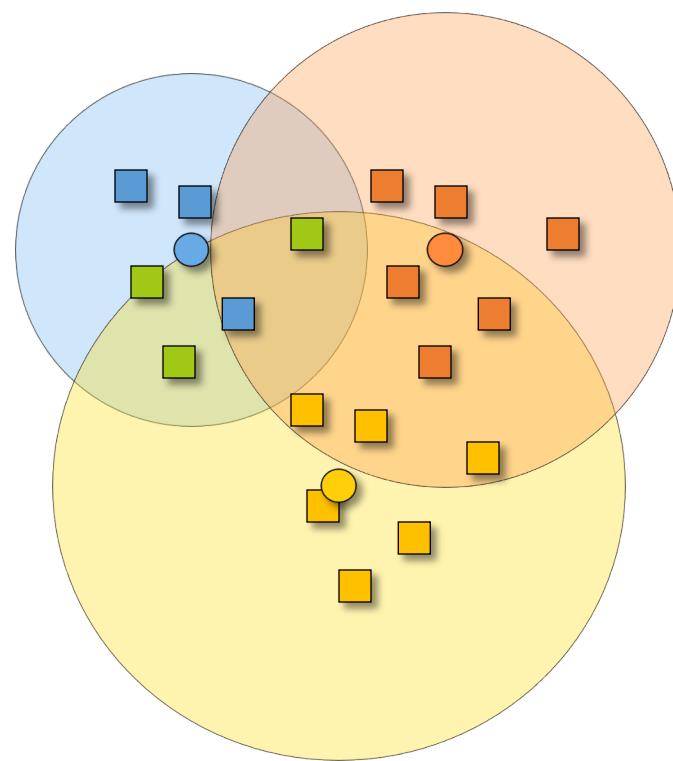
$$\mathcal{W} = 7/18$$

$$\mathcal{W} = 7/18$$



Exemple

- Etape 1



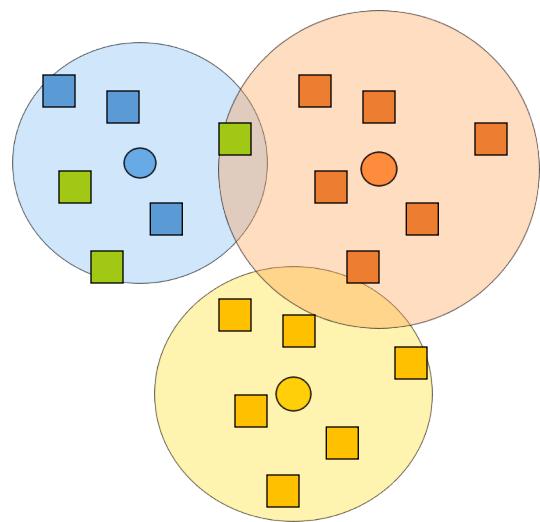
Exemple

- Etape 2

$$\mathcal{W} = 6/18$$

$$\mathcal{W} = 6/18$$

$$\mathcal{W} = 6/18$$



GMM : EM

- EM : Expectation – Maximization
- But : trouver un maximum de vraisemblance d'un GMM
 - Minimum local
- Algorithme itératif, alterne entre 2 étapes
 - E : calcul l'espérance de la vraisemblance en tenant compte des dernières variables observées
 - M : estimation du maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E



GMM : EM

- Converge vers un optimum local
- EM est un algorithme qui estime ces paramètres du GMM
- EM n'est donc pas spécifiquement un algorithme de classification, mais il peut être utilisé pour faire de la classification
- C'est une méthode de gradient : EM maximise la vraisemblance que les données résultent d'une certaine mixture



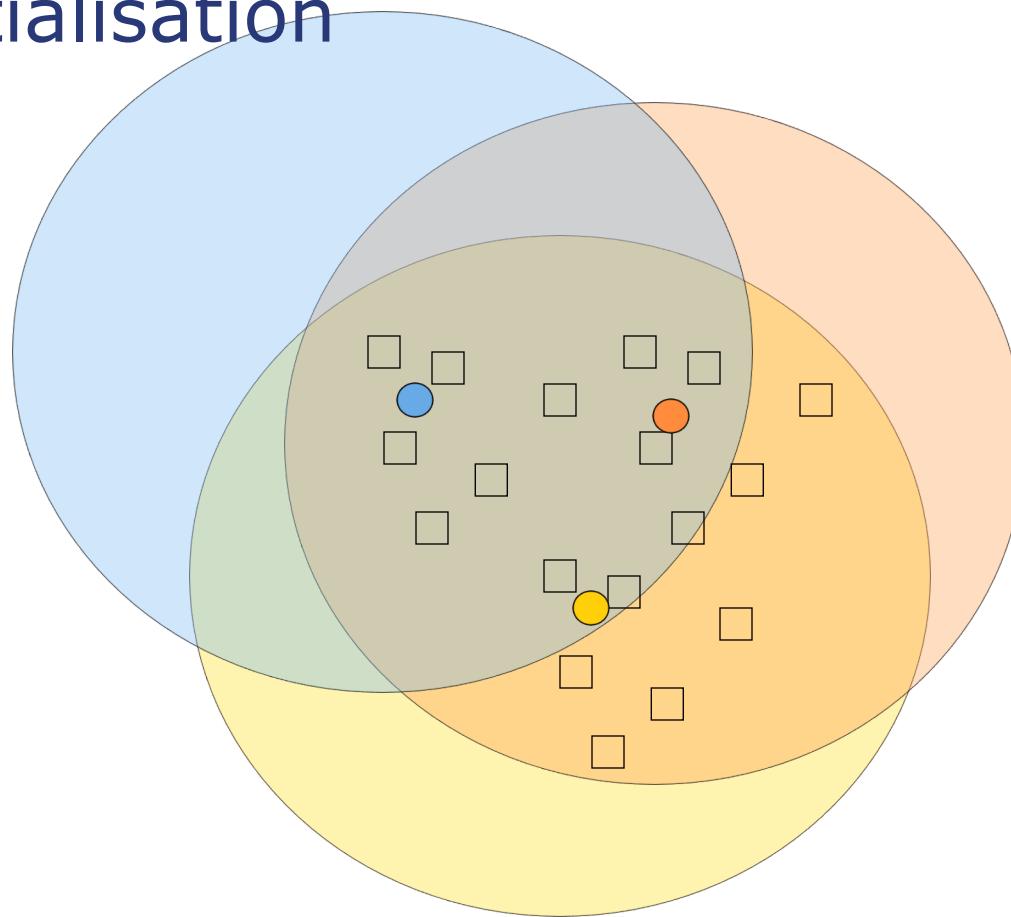
Exemple

- Initialisation

$$\mathcal{W} = 1/3$$

$$\mathcal{W} = 1/3$$

$$\mathcal{W} = 1/3$$

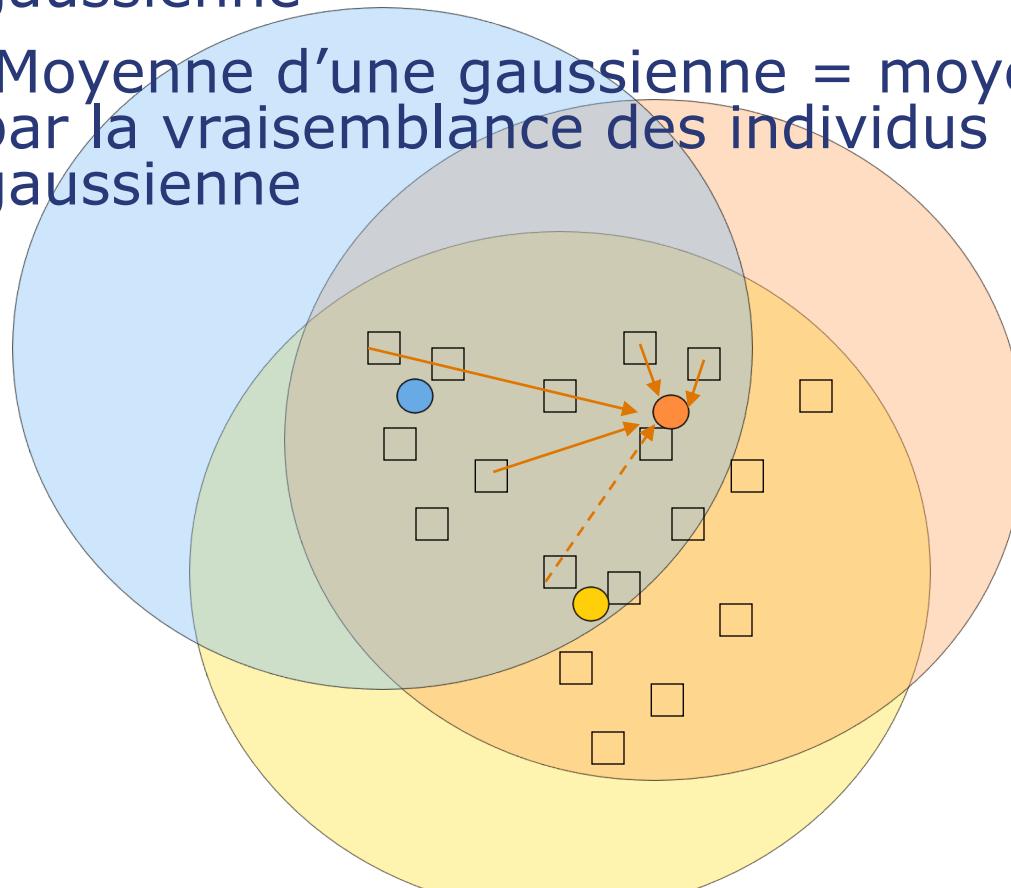


Exemple

■ Étape E

- Chaque individu contribue au calcul de chaque gaussienne
- Moyenne d'une gaussienne = moyenne pondérée par la vraisemblance des individus par rapport à la gaussienne

$$\begin{aligned} w &= 1/3 \\ w &= 1/3 \\ w &= 1/3 \end{aligned}$$



EM v1

- Soit le modèle multigaussien (GMM) :

- l'indice de la gaussienne : $g \in [1..G]$
 - w_g, μ_g, Σ_g sont le poids, la moyenne et la covariance de la gaussienne g
 - La somme des poids = 1
- La probabilité qu'un individu x_t soit émis par \mathcal{M} = à la somme pondérée des vraisemblances

$$p(x_t | \mathcal{M}) = \sum_{g=0}^G w_g p_g(x_t | \mu_g, \Sigma_g)$$

avec

$$p_g(x_t | \mu_g, \Sigma_g) = 2\pi^{-D/2} |\Sigma|^{-1/2} \exp\left(\frac{1}{2}(x_t - \mu_g)' \Sigma_g^{-1} (x_t - \mu_g)\right)$$



EM v1

- La probabilité d'un ensemble d'individus

$$x = \{x_1, \dots, x_t, \dots, x_T\}$$

$$l(x|\mathcal{M}) = \sum_{t=0}^T \log p(x_t|\mathcal{M})$$

- Un GMM est estimé à partir d'un ensemble d'individus en optimisant le critère du maximum de vraisemblance
- L'algorithme est itératif
- A chaque itération i , on construit un nouveau modèle $\mathcal{M}^{(i)}$ en garantissant que

$$l(x|\mathcal{M}^{(i)}) > l(x|\mathcal{M}^{(i-1)})$$



EM v1

■ Algorithme en deux étapes

- E : calcul des statistiques
 - Contribution d'une gaussienne

$$\gamma_g^{(i)}$$

$$\gamma_g^{(i)}(x_t) = \frac{w_g^{(i)} p_g(x_t | \mu_g^{(i)}, \Sigma_g^{(i)})}{\sum_{j=1}^G w_j^{(i)} p_j(x_t | \mu_j^{(i)}, \Sigma_j^{(i)})}$$

$$N_g^{(i)}(x) = \sum_{t=1}^T \gamma_g^{(i)}(x_t)$$

$$F_g^{(i)}(x) = \sum_{t=1}^T \gamma_g^{(i)}(x_t) x_t$$

$$S_g^{(i)}(x) = \sum_{t=1}^T \gamma_g^{(i)}(x_t) x_t x_t'$$



Em v1

- Etape de Maximisation : mise à jour du modèle à partir des statistiques

$$\begin{aligned} w_g^{(i+1)} &= \frac{1}{G} N_g^{(i)}(x) \\ \mu_g^{(i+1)} &= \frac{1}{N_g^{(i)}(x)} F_g^{(i)}(x) \\ \Sigma_g^{(i+1)} &= \frac{1}{N_g^{(i)}(x)} S_g^{(i)}(x) - \mu_g^{(i)} \mu_g^{(i)'} \end{aligned}$$

- Critère d'arrêt :
 - Nombre d'itérations
 - Gain en vraisemblance entre deux itérations inférieur à un seuil

$$l(x|\mathcal{M}^{(i+1)}) - l(x|\mathcal{M}^{(i)}) < \delta$$



EM v2 : Etape E

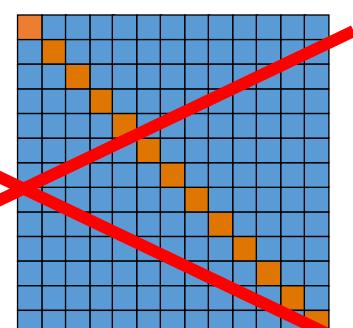
```
for (int f = 0; f < P; f++) {  
    for(int g = 0; g < K; g++) {  
        L = vraisemblance (x[f], gmm[g]);  
        sw[g] += L;  
        sgw += L;  
        for (int i = 0; i < N; i++) {  
            sm[g][i] += L * x[f][i];  
  
            sc[g][i] += L * x[f][i] * x[f][i];  
        }  
    }  
}
```

La moyenne :



Vecteur à n dimensions

La covariance :
Une matrice à
N x N dimensions



Ici la variance :
Vecteur à N dimensions



EM v2 : Étape M

```
for(int g = 0; g < K; g++) {  
    w[g] = sw[g] / sgw;  
    for (int i = 0; i < N; i++) {  
        m[g][i] = sm[g][i] / sw[g];  
  
        c[g][i] = sc[g][i] / sw[g] -  
        m[g][i]*m[g][i];  
    }  
}
```



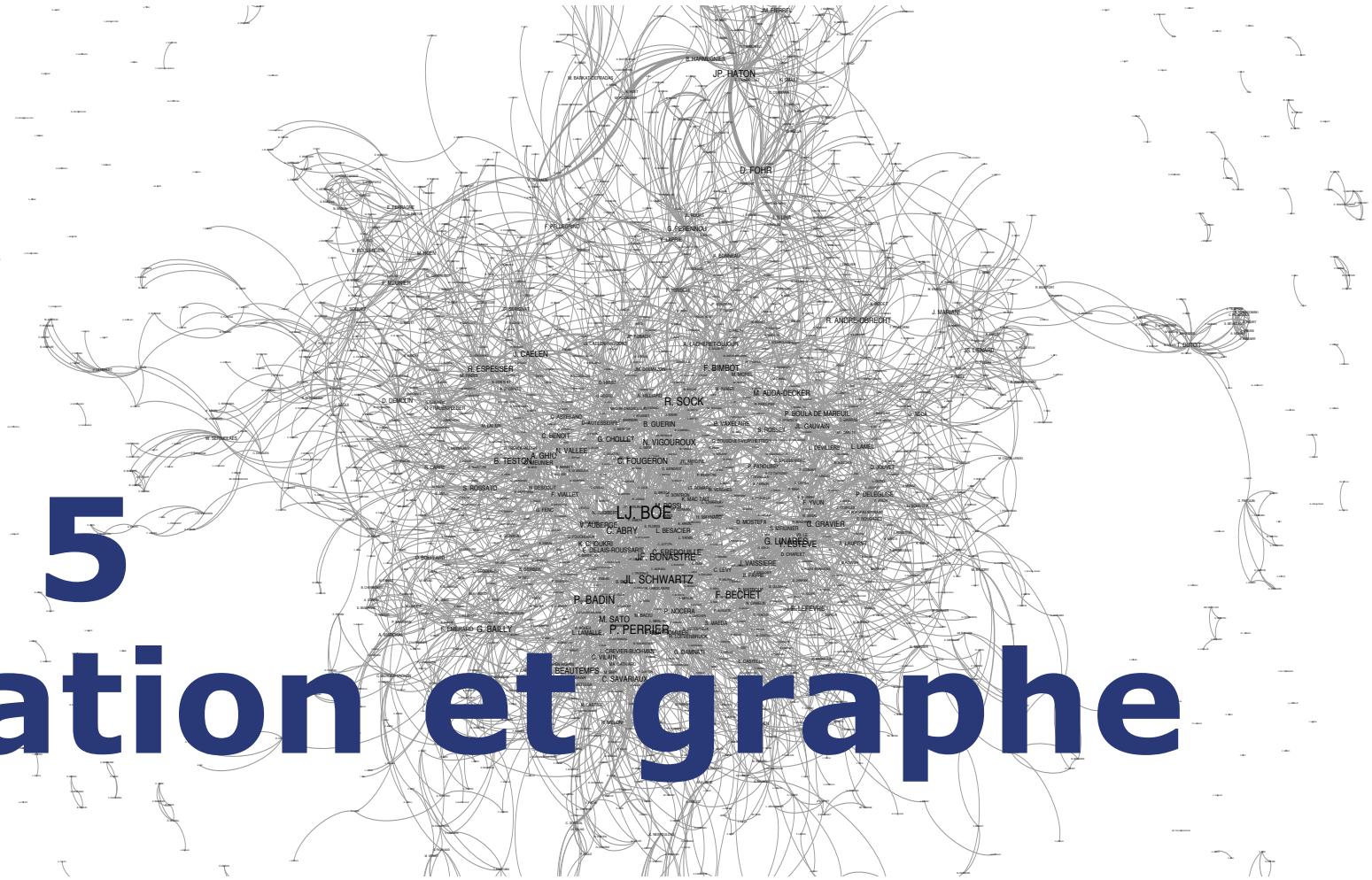
HMM et GMM

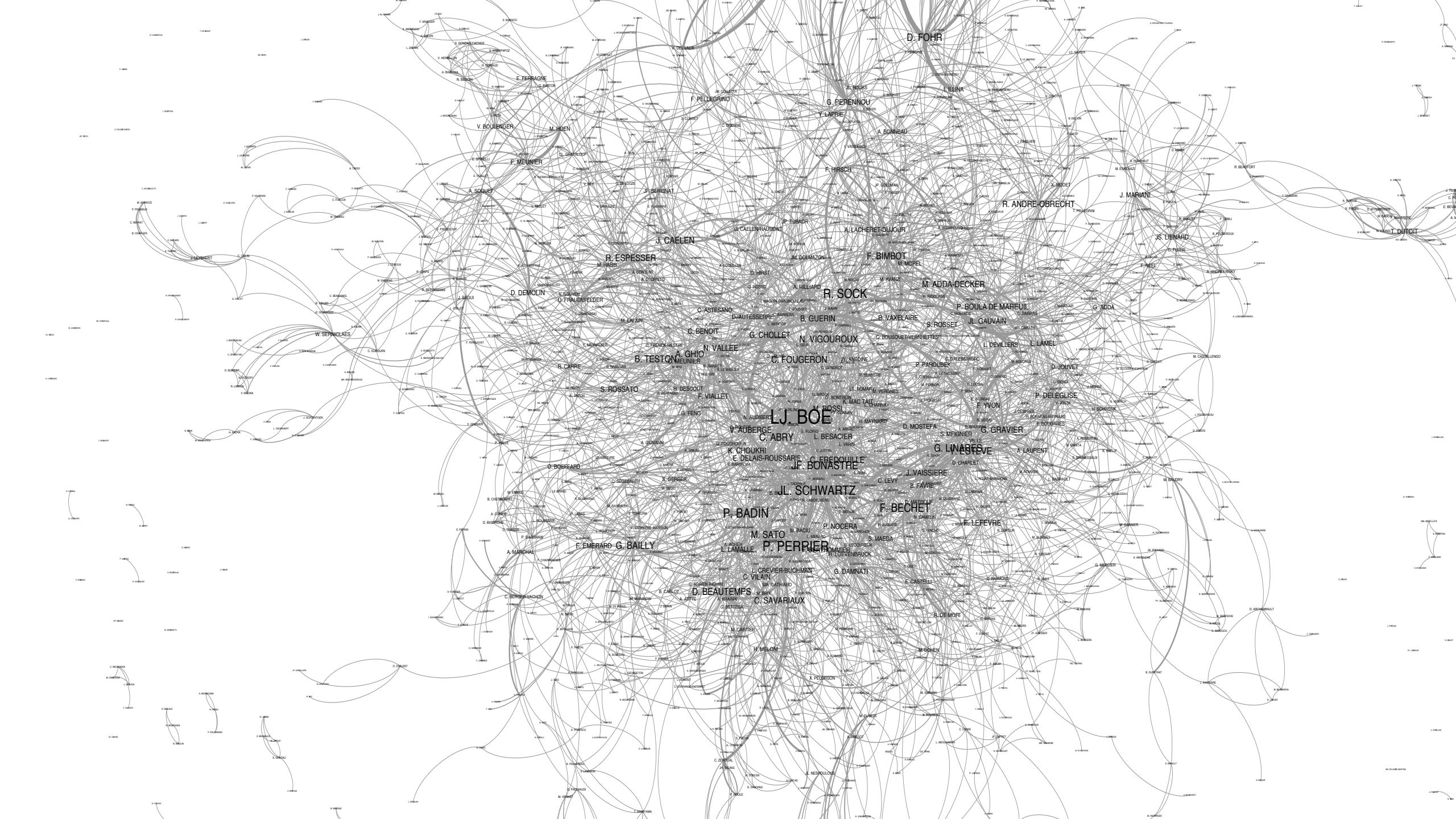
- Alice a accès à la station de météo de Bob
 - Capteur de température
 - Capteur d'humidité
 - ...
- On a des valeurs continues (un *float*)
- La distribution de probabilité B n'est plus discrète
- La distribution de probabilité peut être estimée avec un GMM
 - Attention : la vraisemblance (nombre réel) vs probabilité [0,1]



Chapitre 5

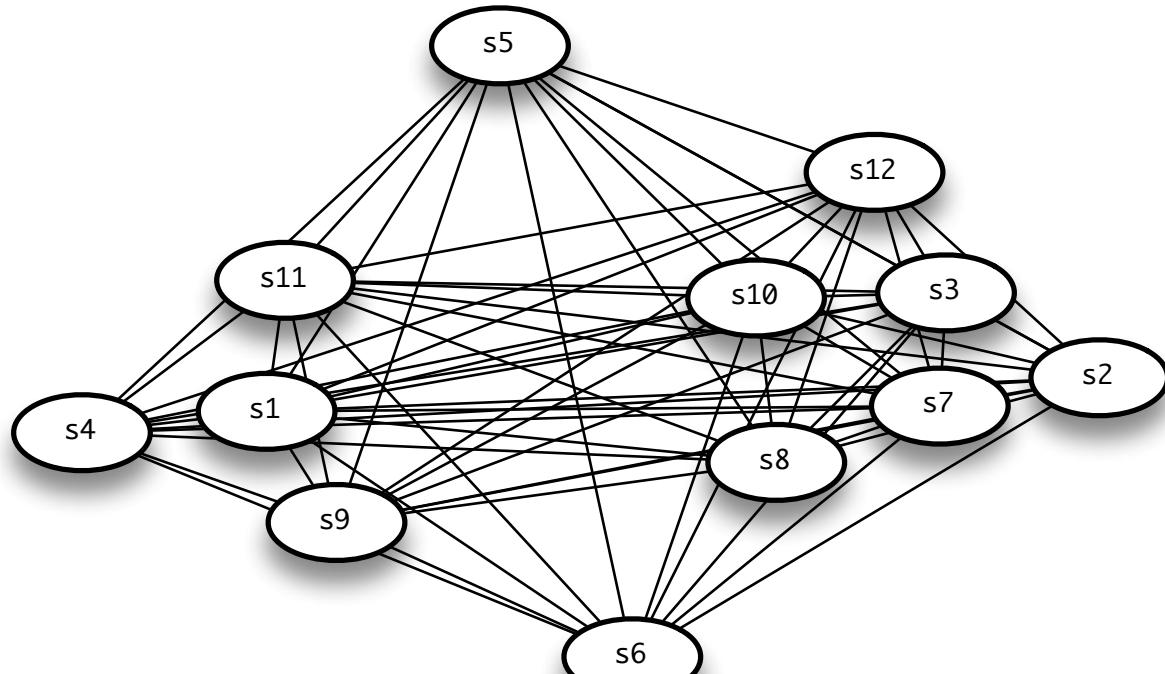
Classification et graphe





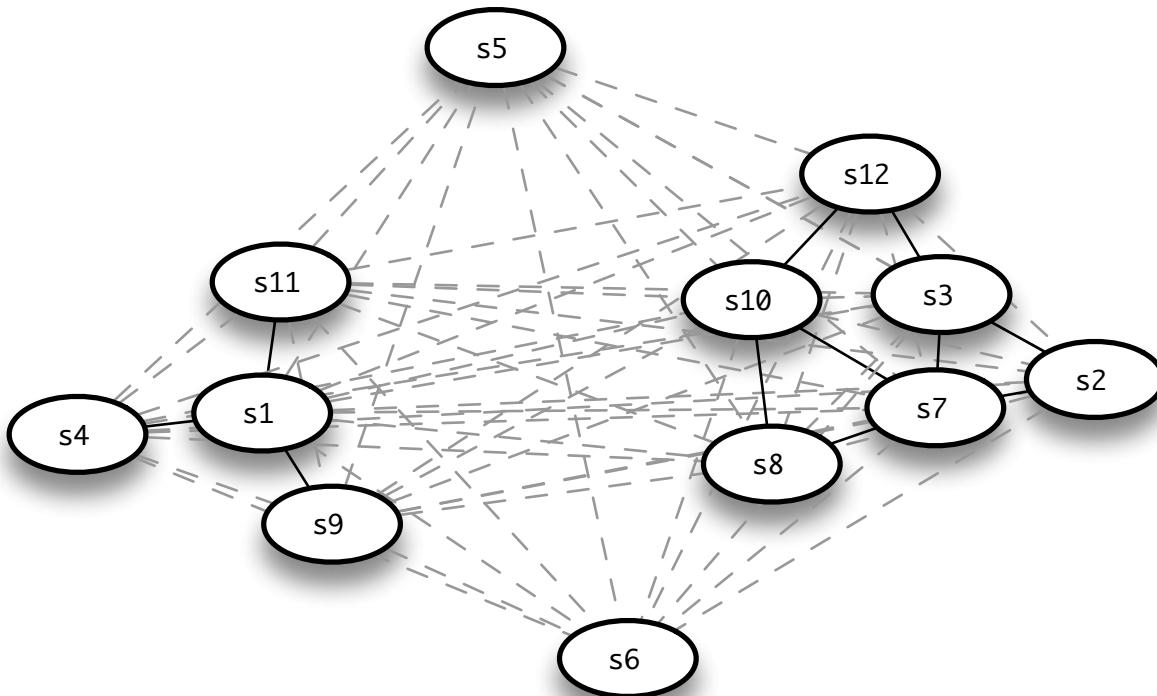
Matrice de distance et graphe

- Chaque noeud = un individu
- chaque lien = la distance entre deux individus



Graphe et seuil

- On grade les liens dont la distance est inférieure à un seuil



Graphe et composantes connexes

- On cherche les composantes connexes = sous graphe
- On cherche les graphes en étoile

