IV - Hidden Markov Model

Romain HÉRAULT (P. LERAY)

INSA Rouen

Automne 2015

Markov Model / Hidden Markov Model

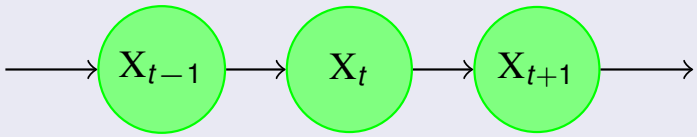
Principe

- L'état **x** du système se trouve dans un espace discret ;
- Le processus est un processus markovien du premier ordre :

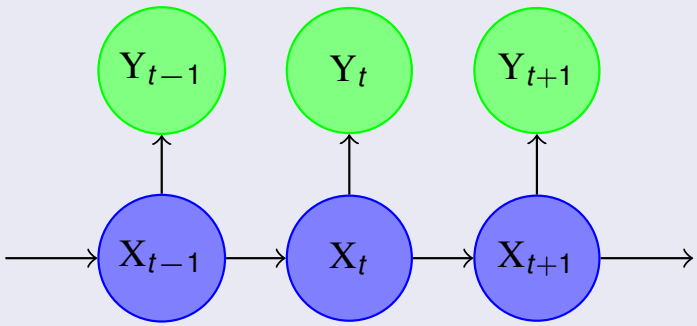
$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{x}_{t-3}, \ldots) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$$

Classification

- **Chaîne de Markov, Modèle de Markov**, l'état est directement observé ;



- **Modèle de Markov Caché**, c'est une variable **y** dépendante de l'état courant qui est observée ; **y** peut être dans un espace discret ou continu.

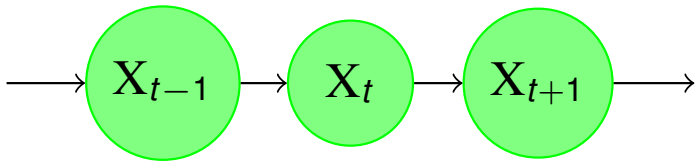


Section 1

Chaînes de Markov

Les Chaînes de Markov

(Modèle de Markov, *Markov Model*, MM)



Notations

- L'état du système suit une variable aléatoire X .
- Les supports de cette variable est un ensemble de n valeurs $\mathcal{X} = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(i)}, \ldots, \mathbf{x}_{(n)}\}$.
- Le tirage à l'instant k de cette variable est noté $\mathbf{x}_k$, il appartient à l'ensemble de supports $\mathcal{X}$.
- La probabilité que le tirage à l'instant k soit égale à la valeur numéro i est

$$P(\mathrm{X}_k = \mathbf{x}_{(i)}) \ ,$$

ou par abus de notation,

$$P(\mathbf{x}_k = i) \ .$$

- La densité de probabilité associée peut être représentée par un vecteur de dimension n,

$$p(\mathbf{x}_k) = [P(\mathrm{X}_k = \mathbf{x}_{(1)}), P(\mathrm{X}_k = \mathbf{x}_{(2)}), \ldots, P(\mathrm{X}_k = \mathbf{x}_{(n)})] \ ,$$

ou par abus de notation,

$$p(\mathbf{x}_k) = [P(\mathbf{x}_k = 1), P(\mathbf{x}_k = 2), \ldots, P(\mathbf{x}_k = n)] \ .$$

Paramètres d'une chaîne de Markov

Il existe 3 paramètres dans un modèle de chaîne de Markov.

- Le nombre d'états possibles, noté n.
- La distribution de l'état du système à l'instant 0, notée $p(\mathbf{x}_0)$ ou Π,

$$p(\mathbf{x}_0) = \Pi = [P(\mathbf{x}_0 = 1), P(\mathbf{x}_0 = 2), \ldots, P(\mathbf{x}_0 = n)] \ .$$

- La matrice de transition d'états notée $\mathbf{A}$, de dimension $n \times n$:

$$\mathbf{A} = \begin{vmatrix} P(\mathbf{x}_k = 1|\mathbf{x}_{k-1} = 1) & P(\mathbf{x}_k = 2|\mathbf{x}_{k-1} = 1) & \ldots & P(\mathbf{x}_k = n|\mathbf{x}_{k-1} = 1) \\ P(\mathbf{x}_k = 1|\mathbf{x}_{k-1} = 2) & P(\mathbf{x}_k = 2|\mathbf{x}_{k-1} = 2) & \ldots & P(\mathbf{x}_k = n|\mathbf{x}_{k-1} = 2) \\ \vdots & \vdots & \vdots & \vdots \\ P(\mathbf{x}_k = 1|\mathbf{x}_{k-1} = n) & P(\mathbf{x}_k = 2|\mathbf{x}_{k-1} = n) & \ldots & P(\mathbf{x}_k = n|\mathbf{x}_{k-1} = n) \end{vmatrix} \quad \forall k$$

On note $\lambda = \{n, \Pi, \mathbf{A}\}$ l'ensemble de ces paramètres.
Lors de la prédiction et du calcul de la vraisemblance on considérera ces paramètres connus. Alors par abus de notation,

$$p(\mathbf{x}_T|\mathbf{x}_{1:T-1}, \lambda) = p(\mathbf{x}_T|\mathbf{x}_{1:T-1}) \text{ et } p(\mathbf{x}_{1:T}|\lambda) = p(\mathbf{x}_{1:T})$$

Exemple de chaîne de Markov : Météo

Météo

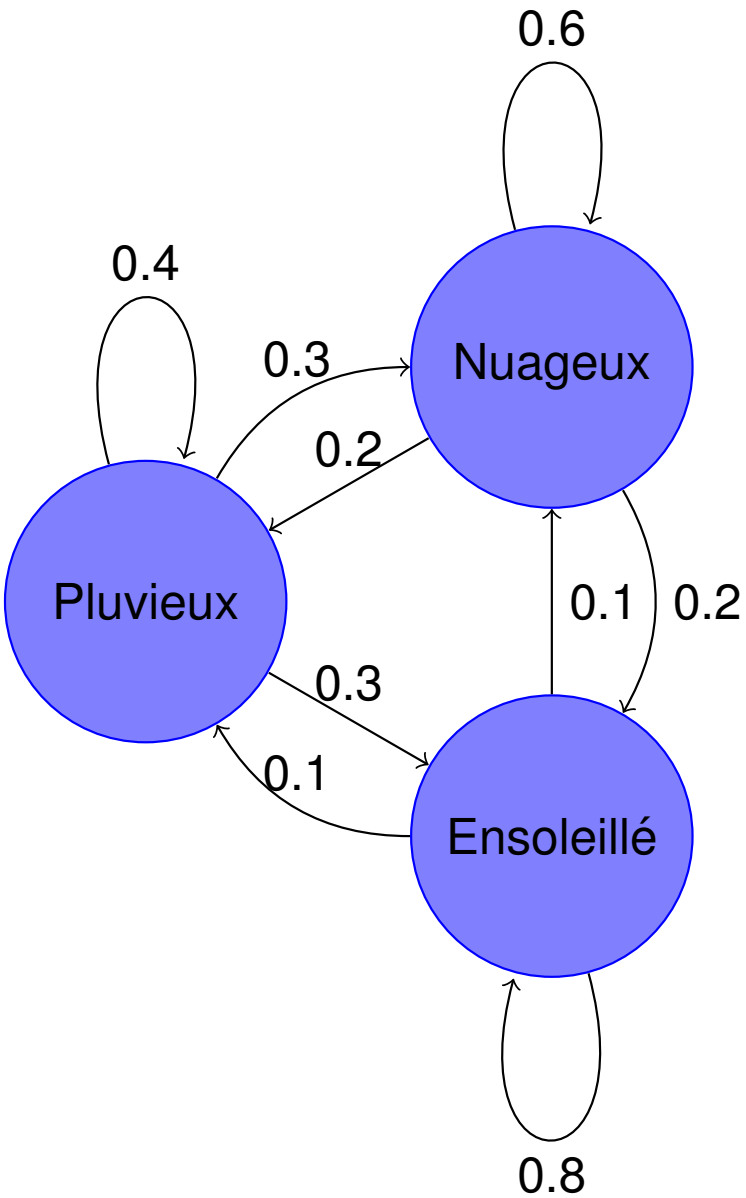
On considère 3 états possibles :

- Pluvieux
- Nuageux
- Ensoleillé

Le pas de temps correspond à un jour.

Grâce aux données historiques, on peut calculer quelle est la probabilité d'avoir un jour de la pluie et le lendemain un jour ensoleillé. On fait de même avec toutes les transitions possibles.

Exemple Météo



La matrice de transition correspond au modèle de dynamique $p(\mathbf{x}_k|\mathbf{x}_{k-1})$

$$\mathcal{X} = \{\text{Pluvieux}, \text{Nuageux}, \text{Ensoleillé}\}$$

$$\mathbf{A} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$

- Ligne : état de départ,
- Colonne : état d'arrivé.

La somme d'une ligne vaut 1 (probabilité)
Comme initialisation, on considère qu'il y a 60% de chance d'avoir de la pluie et 40% d'avoir du soleil, soit :

$$\Pi = [0.6, 0.0, 0.4]$$

Utilisation des chaînes de Markov : Prédiction

Prédiction

La distribution de l'état $\mathbf{x}_k$ connaissant l'état $\mathbf{x}_{k-1}$ est donnée par,

$$p(\mathbf{x}_k|\mathbf{x}_{k-1} = i) = \mathbf{A}_{i\bullet} \ ,$$

c'est à dire, la i-ème ligne de la matrice de transition $\mathbf{A}$.

$$\mathcal{X} = \{\text{Pluvieux}, \text{Nuageux}, \text{Ensoleillé}\}$$

$$\mathbf{A} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$

$$\Pi = [0.6, 0.0, 0.4]$$

S'il a plu, il y a

- 40% de chance qu'il pleuve le lendemain ;
- 30% qu'il fasse nuageux ;
- 30% qu'il fasse ensoleillé.

Utilisation des chaînes de Markov : Modèle génératif

Modèle génératif

On génère une séquence vraisemblable à partir de λ.

- *Initialisation :*
  On tire le premier élément $\mathbf{x}_0$ suivant la distribution Π $(= p(\mathbf{x}_0))$;
- *Itération :*
  On considère qu'on a obtenu au tour précédent $\mathbf{x}_{k-1} = \mathbf{x}_{(i)}$. On tire l'élément $\mathbf{x}_k$ suivant la distribution $p(\mathbf{x}_k | \mathbf{x}_{k-1} = i)$, ce qui correspond à la ligne $\mathbf{A}_{i\bullet}$.

Exemples non-mots

N-gramme

- Base de lexique.org
- Générateur sur lexique.org

Utilisation des chaînes de Markov : Modèle génératif

$$\mathcal{X} = \{\text{Pluvieux}, \text{Nuageux}, \text{Ensoleillé}\}$$

$$\mathbf{A} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$

$$\Pi = [0.6, 0.0, 0.4]$$

- Distribution initiale :

$$\Pi = [0.6, 0.0, 0.4]$$

On tire $\mathbf{x}_{(1)}$ pour X_0 soit *Pluvieux* à 60% de chance.

-

$$p(\mathbf{x}_1 | \mathrm{X}_0 = \mathbf{x}_{(1)}) = [0.4, 0.3, 0.3]$$

On tire $\mathbf{x}_{(3)}$ pour X_1 soit *Ensoleillé* à 30% de chance.

-

$$p(\mathbf{x}_2 | \mathrm{X}_1 = \mathbf{x}_{(3)}) = [0.1, 0.1, 0.8]$$

On tire $\mathbf{x}_{(3)}$ pour X_2 soit *Ensoleillé* à 80% de chance.

-

$$p(\mathbf{x}_3 | \mathrm{X}_2 = \mathbf{x}_{(3)}) = [0.1, 0.1, 0.8]$$

On tire $\mathbf{x}_{(2)}$ pour X_3 soit *Nuageux* à 10% de chance.

Une séquence probable peut être :

[Pluvieux, Ensoleillé, Ensoleillé, Nuageux]

Utilisation des chaînes de Markov : Vraisemblance

Formulation générale

On utilise la propriété des processus markovien :

$$
\begin{aligned}
p(\mathbf{x}_{0:T}) &= p(\mathbf{x}_T|\mathbf{x}_{0:T-1})p(\mathbf{x}_{0:T-1}) \\
p(\mathbf{x}_{0:T}) &= p(\mathbf{x}_T|\mathbf{x}_{T-1})p(\mathbf{x}_{0:T-1}) \\
p(\mathbf{x}_{0:T}) &= p(\mathbf{x}_T|\mathbf{x}_{T-1})p(\mathbf{x}_{T-1}|\mathbf{x}_{T-2})p(\mathbf{x}_{0:T-2}) \\
p(\mathbf{x}_{0:T}) &= p(\mathbf{x}_0)\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})
\end{aligned}
$$

Application aux MM

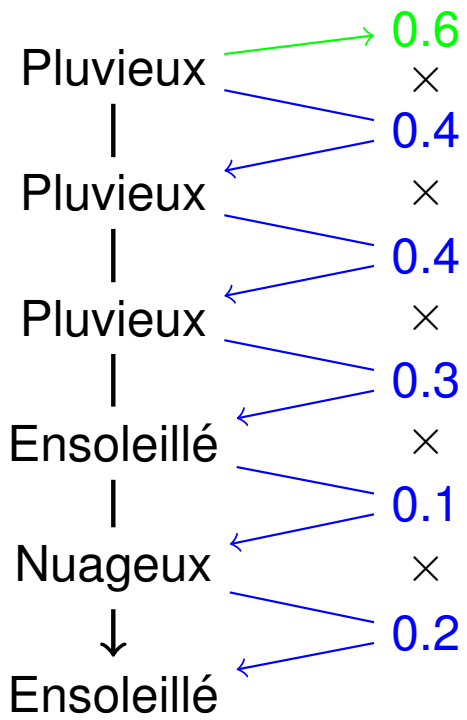
$$
\begin{aligned}
p(\mathbf{x}_{0:T}) &= \Pi_h \;|\; \mathbf{x}_0=\mathbf{x}_{(h)} \times \prod_{t=1}^{T} A_{i,j} \;|\; \mathbf{x}_{t-1}=\mathbf{x}_{(i)} \quad \text{et} \quad \mathbf{x}_t=\mathbf{x}_{(j)} \\
p(\mathbf{x}_{0:T}) &= \Pi_h \;|\; \mathbf{x}_0=h \times \prod_{t=1}^{T} A_{i,j} \;|\; \mathbf{x}_{t-1}=i \quad \text{et} \quad \mathbf{x}_t=j \\
p(\mathbf{x}_{0:T}) &= \Pi_{\mathbf{x}_0} \times \prod_{t=1}^{T} A_{\mathbf{x}_{t-1},\mathbf{x}_t}
\end{aligned}
$$

Utilisation des chaînes de Markov : Vraisemblance

$$\mathcal{X} = \{\text{Pluvieux}, \text{Nuageux}, \text{Ensoleillé}\}$$

$$\mathbf{A} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$

$$\Pi = [0.6, 0.0, 0.4]$$



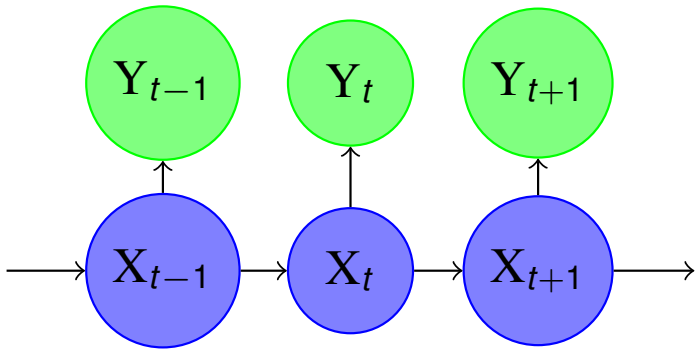
Vraisemblance : 0.000576

Section 2

Chaînes de Markov à états chachés

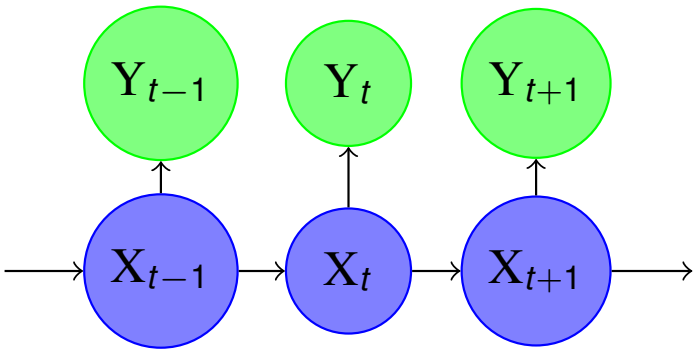
Les chaînes de Markov à états cachés

(*Hidden Markov Models*,HMM)



Extension des MM

C'est une variable Y dépendant uniquement de l'état courant qui est observée ; **y** peut être ans un espace discret ou continu.



Pour la première partie de ce cours on considérera que **y** est dans un espace **discret**.

Notations

En plus des notations indiquées pour les MM, nous avons,

- L'observation suit une variable aléatoire Y .
- Les supports de cette variable est un ensemble de m valeurs $\mathcal{Y} = \{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(i)}, \ldots, \mathbf{y}_{(m)}\}$.
- Le tirage à l'instant k de cette variable est noté $\mathbf{y}_k$, il appartient à l'ensemble de supports $\mathcal{Y}$.
- La probabilité que le tirage à l'instant k soit égale à la valeur numéro i est

$$P(\mathrm{Y}_k = \mathbf{y}_{(i)}) \ ,$$

ou par abus de notation,

$$P(\mathbf{y}_k = i) \ .$$

- La densité de probabilité associée peut être représentée par un vecteur de dimension m,

$$p(\mathbf{y}_k) = [P(\mathrm{Y}_k = \mathbf{y}_{(1)}), P(\mathrm{Y}_k = \mathbf{y}_{(2)}), \ldots, P(\mathrm{Y}_k = \mathbf{y}_{(m)})] \ ,$$

ou par abus de notation,

$$p(\mathbf{y}_k) = [P(\mathbf{y}_k = 1), P(\mathbf{y}_k = 2), \ldots, P(\mathbf{y}_k = m)] \ .$$

Paramètres d'une chaîne de Markov cachée

Il existe 5 paramètres dans un modèle de chaîne de Markov cachée.

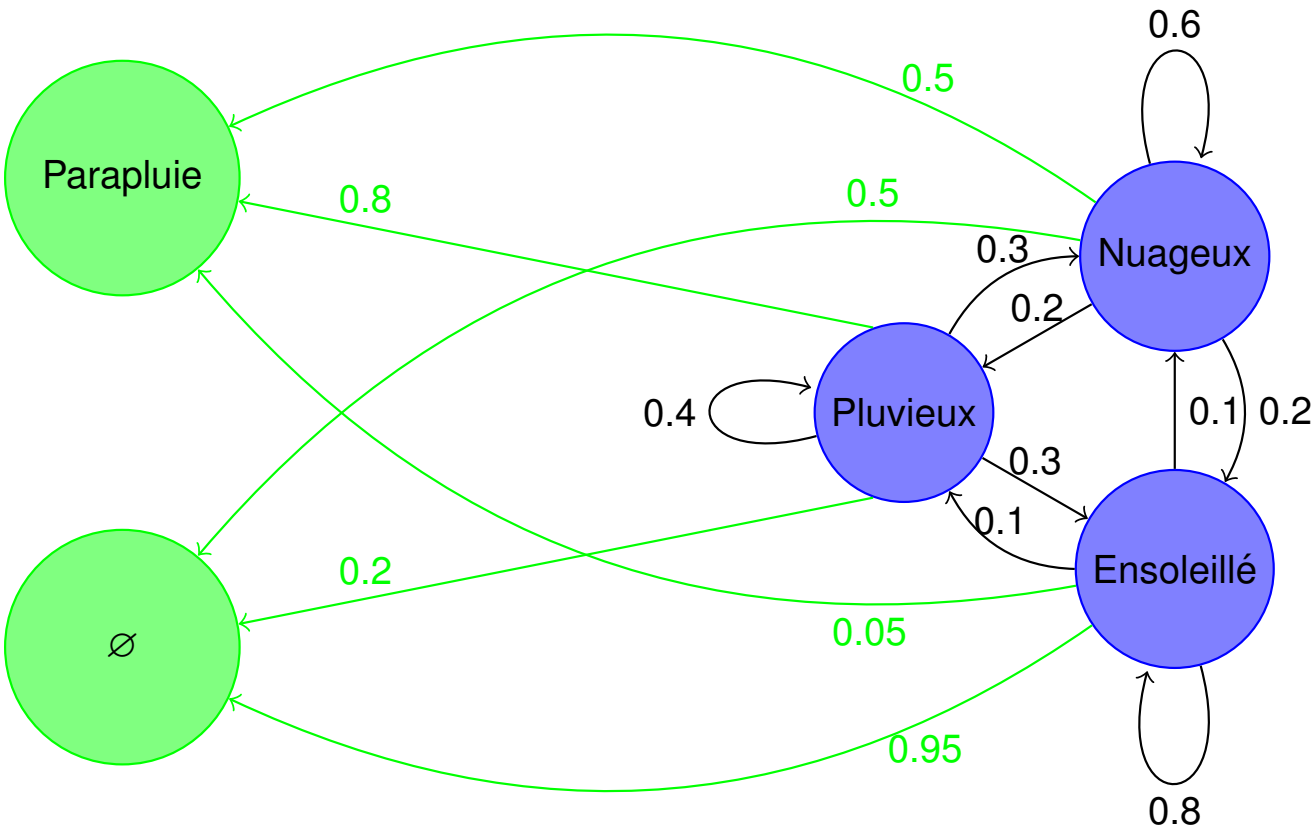
- Les 3 paramètres d'un MM,
  - Le nombre d'états possibles, noté n.
  - La distribution de l'état du système à l'instant 0, notée $p(\mathbf{x}_0)$ ou Π.
  - La matrice de transition d'états notée $\mathbf{A}$:
- Le nombre d'observations possibles, noté m.
- La matrice d'émission ou d'observations notée $\mathbf{B}$, de dimension $n \times m$:

$$\mathbf{B} = \begin{vmatrix} P(\mathbf{y}_k = 1|\mathbf{x}_k = 1) & P(\mathbf{y}_k = 2|\mathbf{x}_k = 1) & \dots & P(\mathbf{y}_k = m|\mathbf{x}_k = 1) \\ P(\mathbf{y}_k = 1|\mathbf{x}_k = 2) & P(\mathbf{y}_k = 2|\mathbf{x}_k = 2) & \dots & P(\mathbf{y}_k = m|\mathbf{x}_k = 2) \\ \vdots & \vdots & \vdots & \vdots \\ P(\mathbf{y}_k = 1|\mathbf{x}_k = n) & P(\mathbf{y}_k = 2|\mathbf{x}_k = n) & \dots & P(\mathbf{y}_k = m|\mathbf{x}_k = n) \end{vmatrix} \quad \forall k$$

On note $\lambda = \{n, \Pi, \mathbf{A}, m, \mathbf{B}\}$ l'ensemble de ces paramètres.

Exemple : météo suite

Imaginons que nous sommes enfermés dans un bâtiment et que nous ne pouvons voir directement le ciel. On peut simplement savoir si les personnes portent un parapluie ou non.



Exemple de tâche, connaissant uniquement une séquence du type [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie], il faut deviner la météo.

Calcul de la vraisemblance ou **Évaluation**

Vraisemblance : Forme générale

$$p(\mathbf{y}_{1:T}) = \int \dots \int_{\mathbf{x}_{0:T}} p(\mathbf{x}_0) \left(\prod_{k=1}^{T} p(\mathbf{x}_k|\mathbf{x}_{k-1}) \right) \left(\prod_{k=1}^{T} p(\mathbf{y}_k|\mathbf{x}_k) \right) d_{\mathbf{x}_{0:T}}$$

$$p(\mathbf{y}_{1:T}) = \int \dots \int_{\mathbf{x}_{0:T}} p(\mathbf{x}_0) \prod_{k=1}^{T} (p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{y}_k|\mathbf{x}_k)) \, d_{\mathbf{x}_{0:T}}$$

Vraisemblance : Forme HMM avec états et observations discrètes

$$p(\mathbf{y}_{1:T}) = \sum_{\mathbf{x}_0=1}^{n} \sum_{\mathbf{x}_1=1}^{n} \dots \sum_{\mathbf{x}_T=1}^{n} p(\mathbf{x}_0) \prod_{k=1}^{T} (p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{y}_k|\mathbf{x}_k))$$

$$p(\mathbf{y}_{1:T}) = \sum_{\mathbf{x}_0=1}^{n} \sum_{\mathbf{x}_1=1}^{n} \dots \sum_{\mathbf{x}_T=1}^{n} p(\mathbf{x}_0) \prod_{k=1}^{T} \mathbf{A}_{\mathbf{x}_{k-1},\mathbf{x}_k} . \mathbf{B}_{\mathbf{x}_k,\mathbf{y}_k}$$

Complexité $n^T \times T \Rightarrow$ Intraitable

Calcul de la vraisemblance

Programmation dynamique

- On note $\alpha_k(i) = p(\mathbf{y}_{1:k}, \mathbf{x}_k = i|\lambda)$ la probabilité jointe d'être dans l'état i et d'observer la séquence $\mathbf{y}_{1:k}$.
- La vraisemblance de la séquence est donnée par $p(\mathbf{y}_{1:T}|\lambda) = \sum_{i=1}^{n} p(\mathbf{y}_{1:T}, \mathbf{x}_T = i|\lambda) = \sum_{i=1}^{n} \alpha_T(i)$

Algorithme Forward

- Initialisation :

$$\alpha_0(i) = p(\mathbf{x}_0 = i) = \Pi(i) \quad \forall i \in [1..n]$$

- Itération $\forall k \in [1..T], \forall j \in [1..n]$:

$$\begin{aligned}
\alpha_k(j) &= p(\mathbf{y}_{1:k}, \mathbf{x}_k = j|\lambda) \ , \\
\alpha_k(j) &= p(\mathbf{y}_k|\mathbf{x}_k = j)p(\mathbf{y}_{1:k-1}, \mathbf{x}_k = j) \ , \\
\alpha_k(j) &= p(\mathbf{y}_k|\mathbf{x}_k = j) \sum_{i=1}^{n} p(\mathbf{x}_k = j|\mathbf{x}_{k-1} = i)p(\mathbf{y}_{1:k-1}, \mathbf{x}_{k-1} = i) \ , \\
\alpha_k(j) &= \mathbf{B}_{j,\mathbf{y}_k} \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \ .
\end{aligned}$$

- Terminaison :

$$p(\mathbf{y}_{1:T}|\lambda) = \sum_{i=1}^{n} p(\mathbf{y}_{1:T}, \mathbf{x}_T = i|\lambda) = \sum_{i=1}^{n} \alpha_T(i)$$

Complexité $n^2 \times T$

Forward représentation graphique

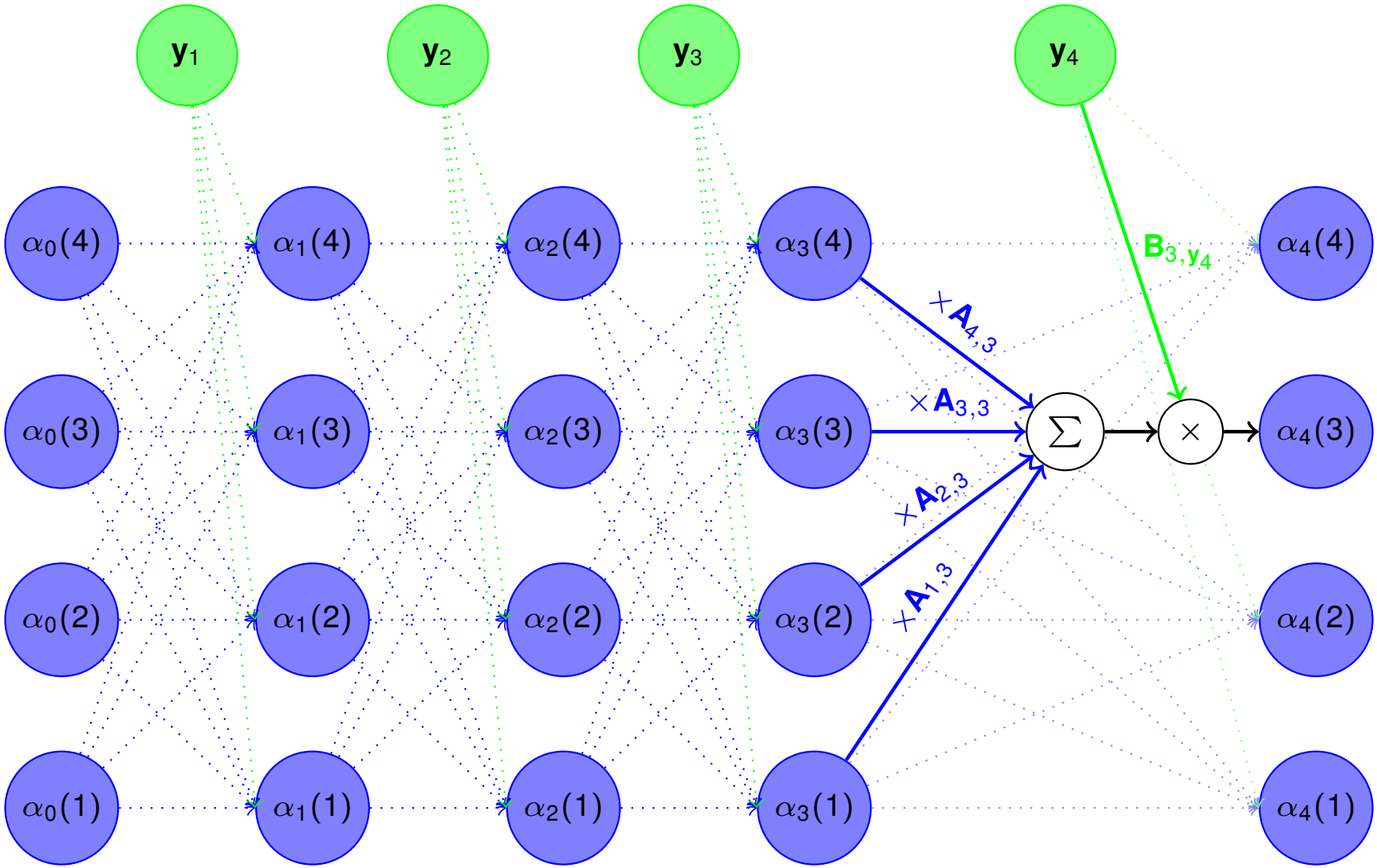


Illustration vraisemblance

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie],
Vraisemblance 0.000135

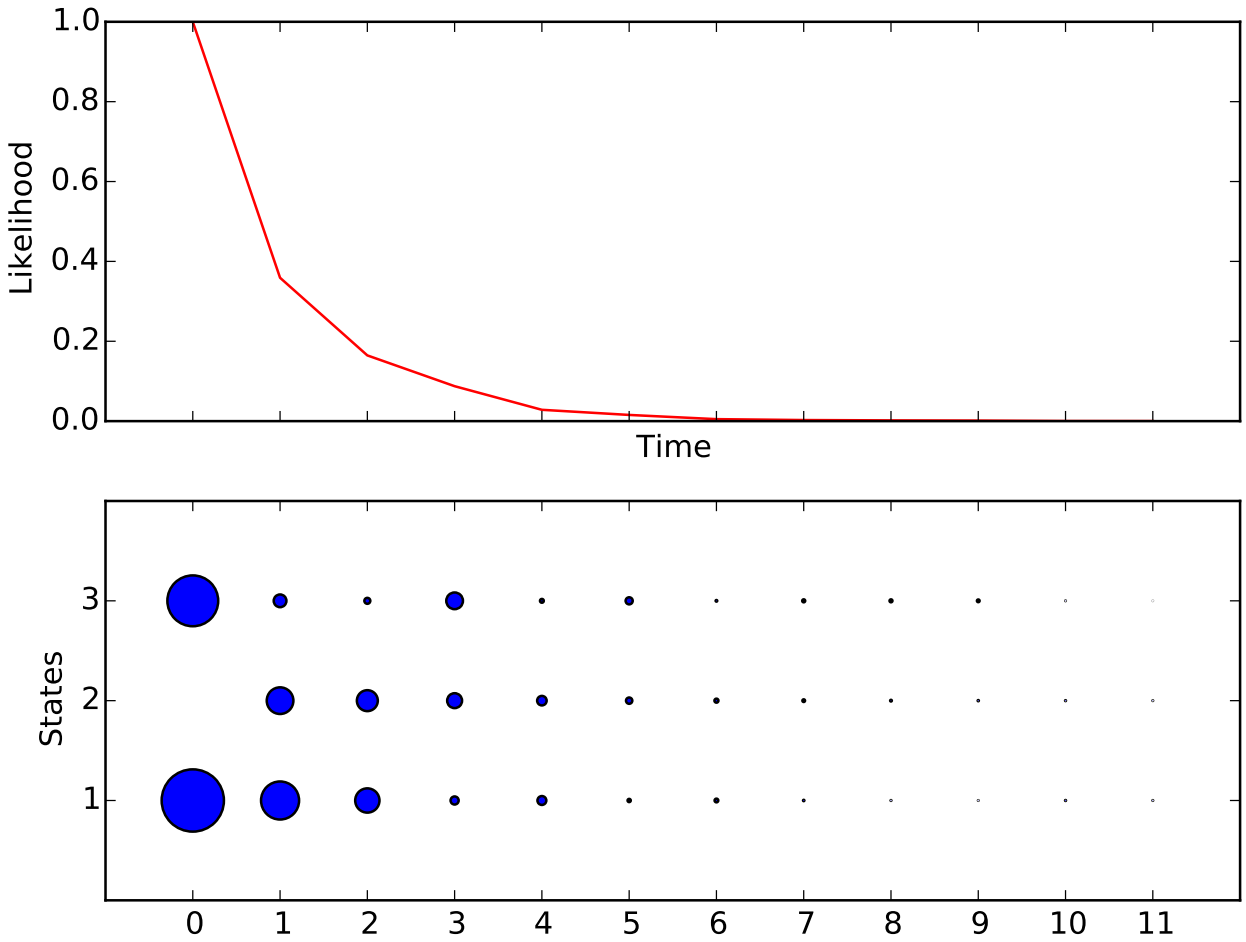
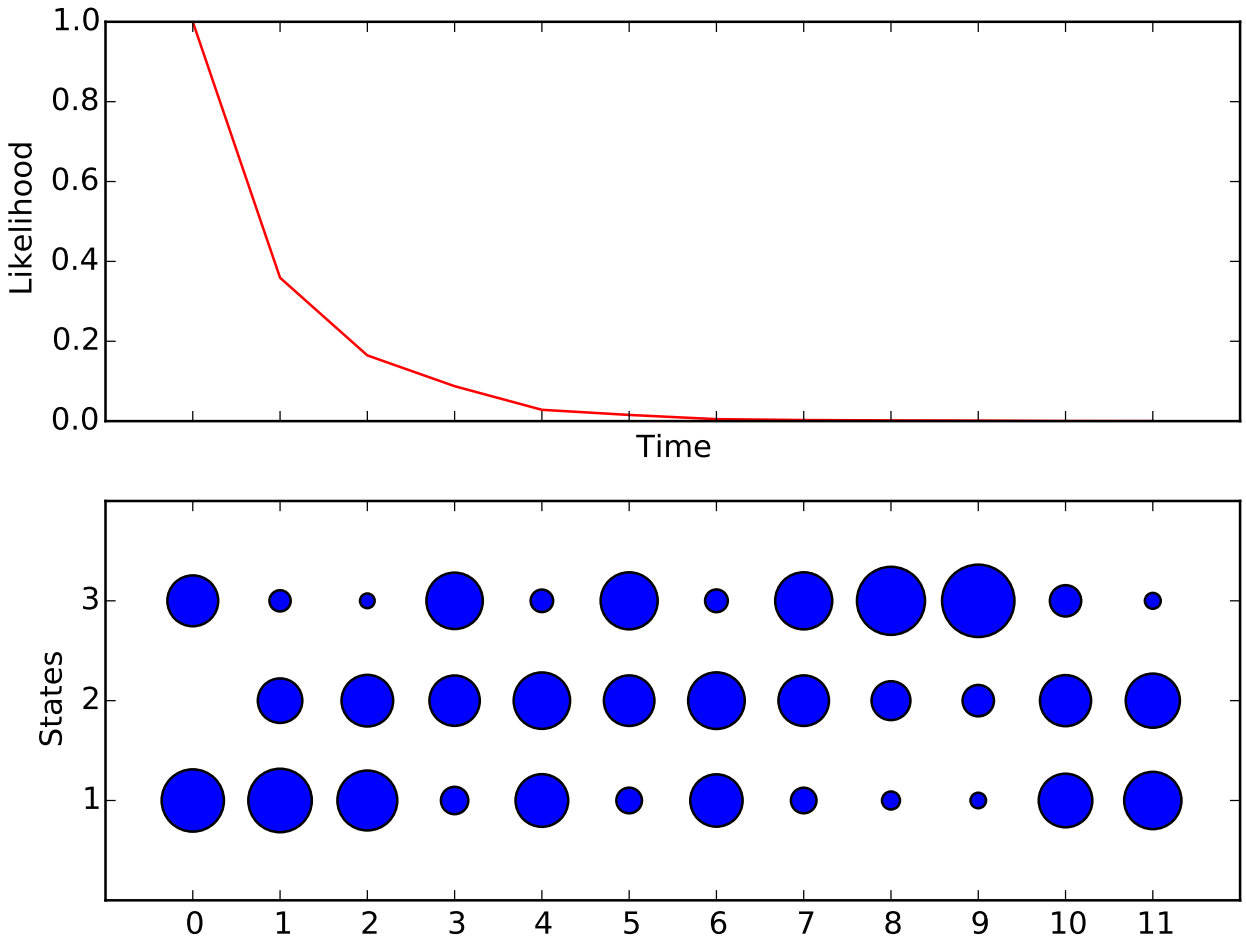


Illustration vraisemblance

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie],
Vraisemblance 0.000135



Explication ou **Décodage**

Tâche

Trouver la séquences d'états maximisant la vraisemblance.

Première approche

- On calcule $\alpha_k(i) = p(\mathbf{y}_{1:k}, \mathbf{x}_k = i|\lambda)$ et $\beta_k(i) = p(\mathbf{y}_{k+1:T}|\mathbf{x}_k = i, \lambda)$;
- On en déduit $\gamma_k(i) = p(\mathbf{x}_k = i|\lambda, \mathbf{y}_{1:T})$;
- La vraisemblance d'une séquence est alors accessible $p(\mathbf{x}_{1:T}|\lambda, \mathbf{y}_{1:T}) = \prod_{k=1}^{T} p(\mathbf{x}_k|\lambda, \mathbf{y}_{1:T})$;
- On teste toutes les séquences possibles et on garde la plus vraisemblable.

Algorithme Backward

On note $\beta_k(i) = p(\mathbf{y}_{k+1:T}|\mathbf{x}_k = i, \lambda)$ la probabilité d'observer la séquence $\mathbf{y}_{k+1:T}$ connaissant l'état i à l'instant k.

- Initialisation :

$$\beta_T(i) = 1 \quad \forall i \in [1..n]$$

- Itération $\forall k \in [T-1..0]$ par pas de -1, $\forall i \in [1..n]$:

$$\begin{aligned}
\beta_k(i) &= p(\mathbf{y}_{k+1:T}|\mathbf{x}_k = i, \lambda) \ , \\
\beta_k(i) &= \sum_{j=1}^{n} p(\mathbf{x}_{k+1} = j|\mathbf{x}_k = i) p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1} = j) p(\mathbf{y}_{k+2:T}|\mathbf{x}_{k+1} = j) \ , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} \mathbf{B}_{j,\mathbf{y}_{k+1}} \times \beta_{k+1}(j) \ .
\end{aligned}$$

Complexité $n^2 \times T$

Algorithme Forward-Backward

- On fait une passe *forward*
- On fait une passe *backward*
- Puis . . .

Résolution

$\forall k \in [1..T], \forall i \in [1..n]$

$$\begin{aligned}
\gamma_k(i) &= p(\mathbf{x}_k = i|\lambda, \mathbf{y}_{1:T}) \\
\gamma_k(i) &= \frac{p(\mathbf{x}_k = i, \mathbf{y}_{1:T}|\lambda)}{p(\mathbf{y}_{1:T}|\lambda)} \\
\gamma_k(i) &= \frac{p(\mathbf{y}_{1:k}, \mathbf{x}_k = i|\lambda)p(\mathbf{y}_{k+1:T}|\mathbf{x}_k = i, \lambda)}{p(\mathbf{y}_{1:T}|\lambda)} \\
\gamma_k(i) &= \frac{\alpha_k(i)\beta_k(i)}{\sum_j \alpha_k(j)\beta_k(j)}
\end{aligned}$$

Propriété :

$$\forall k \quad \sum_i \gamma_k(i) = 1$$

Et finalement ?

Lissage OK, on connaît $p(\mathbf{x}_k = i|\lambda, \mathbf{y}_{1:T})$.

Attention

Connaissant la séquence $\mathbf{y}_{1:T}$, l'état le plus vraisemblable à l'instant k , $\arg\max_{\mathbf{x}_k} p(\mathbf{x}_k|\mathbf{y}_{1:T})$, n'appartient pas forcément à la séquence d'états la plus vraisemblable, $\arg\max_{\mathbf{x}_{0:T}} p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$.

On doit tester toutes les séquences possibles et on garde la plus vraisemblable,

$$\begin{aligned}\arg\max_{\mathbf{x}_{0:T}} p(\mathbf{x}_{1:T}|\lambda, \mathbf{y}_{1:T}) &= \arg\max_{\mathbf{x}_{0:T}} \prod_{k=1}^{T} p(\mathbf{x}_k|\lambda, \mathbf{y}_{1:T}) \\ &= \arg\max_{\mathbf{x}_{0:T}} \prod_{k=1}^{T} \gamma_k(\mathbf{x}_k)\end{aligned}$$

Complexité

$n^T \Rightarrow$ Intraitable

Solution : Algorithme de Viterbi (1967)

On pose $\delta_k(i)$ la vraisemblance jointe de

- la séquence d'observation $\mathbf{y}_{1:k}$,
- l'état $\mathbf{x}_k$,
- et de la séquence d'états $\mathbf{x}_{0:k-1}$ la plus vraisemblable qui mène à $\mathbf{x}_k$.

Autrement dit,

$$\delta_k(i) = \max_{\mathbf{x}_{0:k-1}} p(\mathbf{y}_{1:k}, \mathbf{x}_{0:k-1}, \mathbf{x}_k = i) \ ,$$

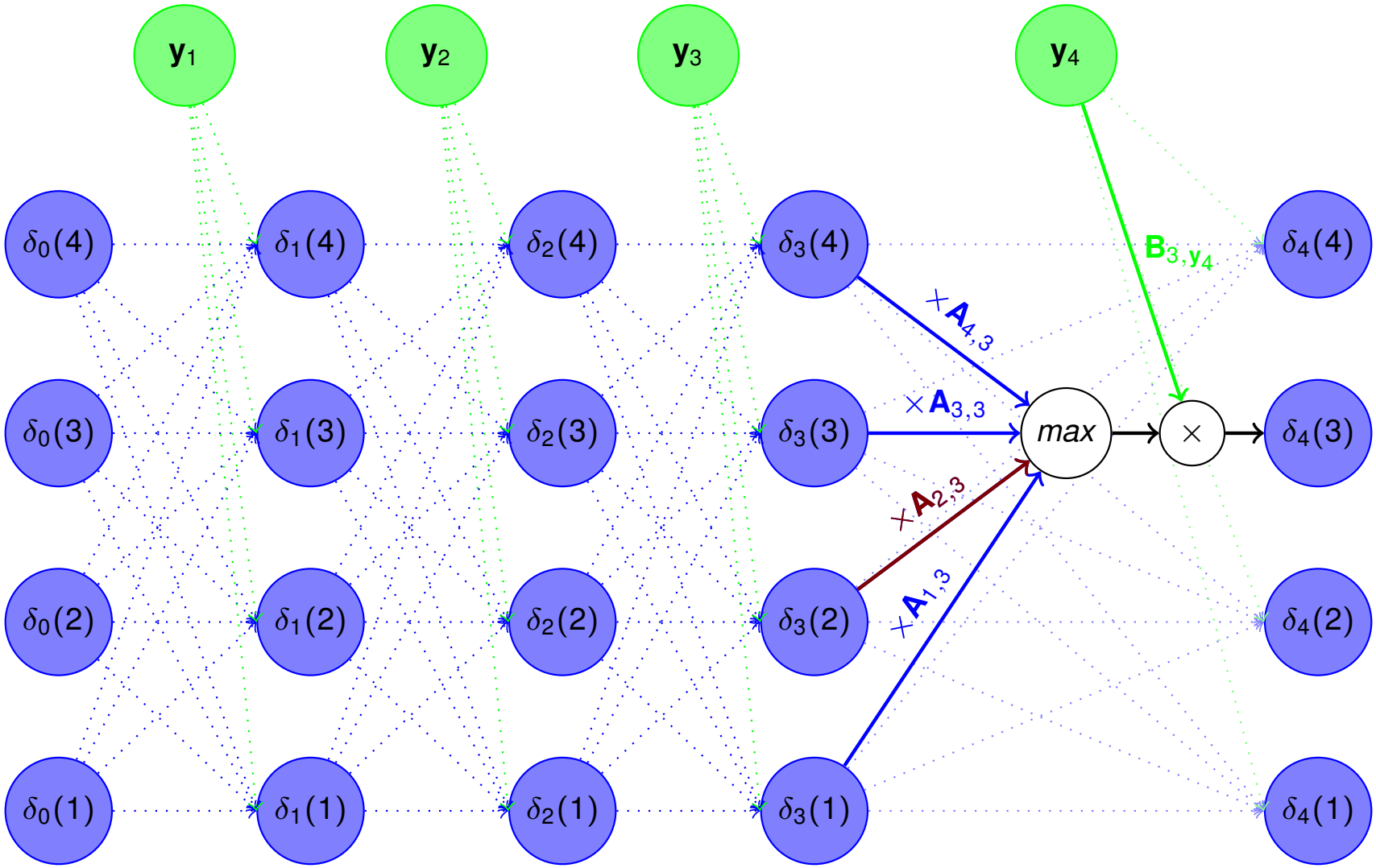
qui peut se calculer itérativement,

$$\begin{aligned} \delta_k(j) &= \left[\max_i \delta_{k-1}(i) p(\mathbf{x}_k = j | \mathbf{x}_{k-1} = i)\right] p(\mathbf{y}_k | \mathbf{x}_k = j) \\ \delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \ . \end{aligned}$$

En outre, pour chaque $\delta_k(j)$, on va garder le dernier état $\mathbf{x}_{k-1}$ de la séquence d'états $\mathbf{x}_{0:k-1}$ la plus vraisemblable qui mène à lui.

$$\psi_k(j) = \arg\max_i \delta_{k-1}(i) \mathbf{A}_{i,j} \ .$$

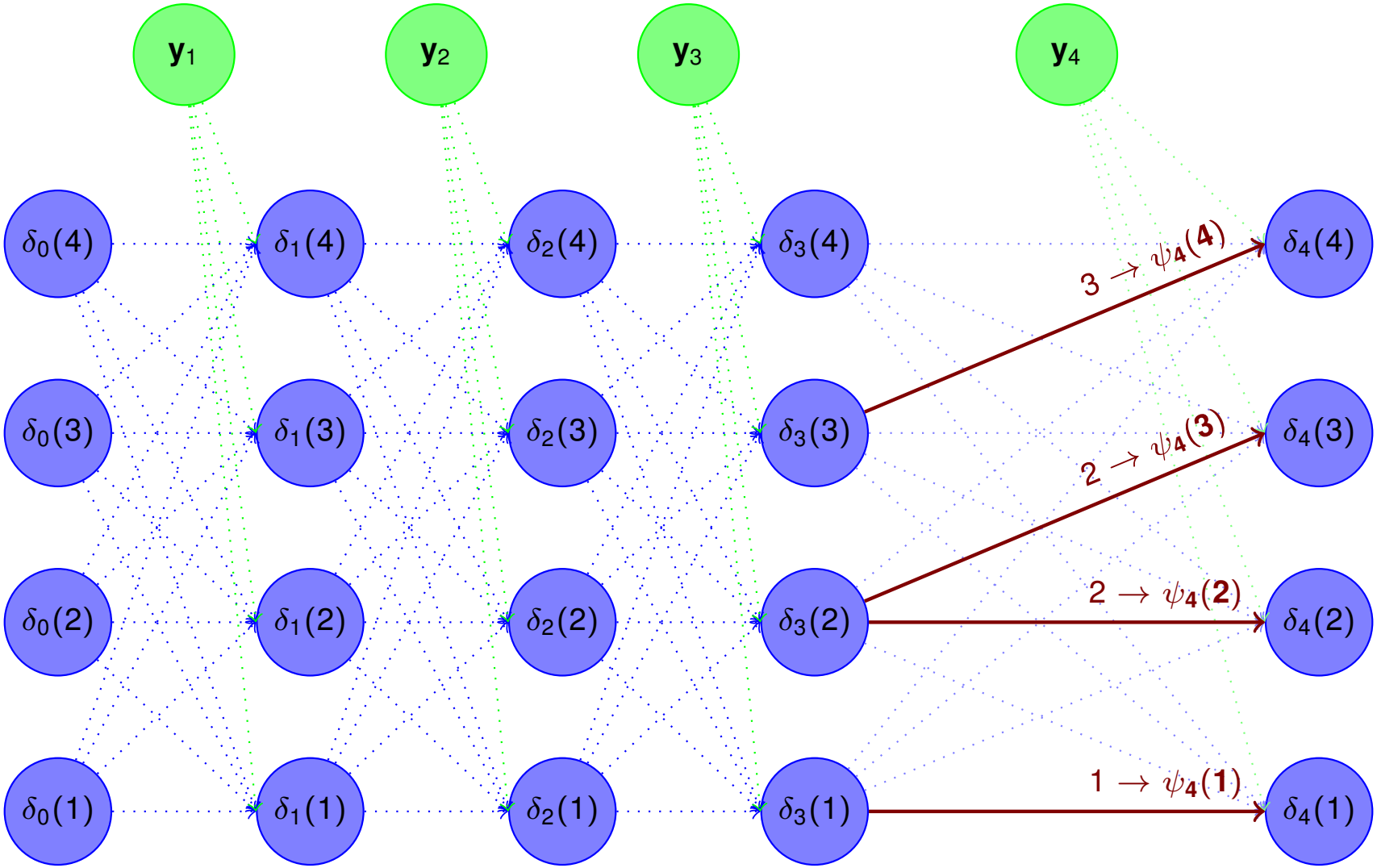
Viterbi, phase 1 :**Itération**



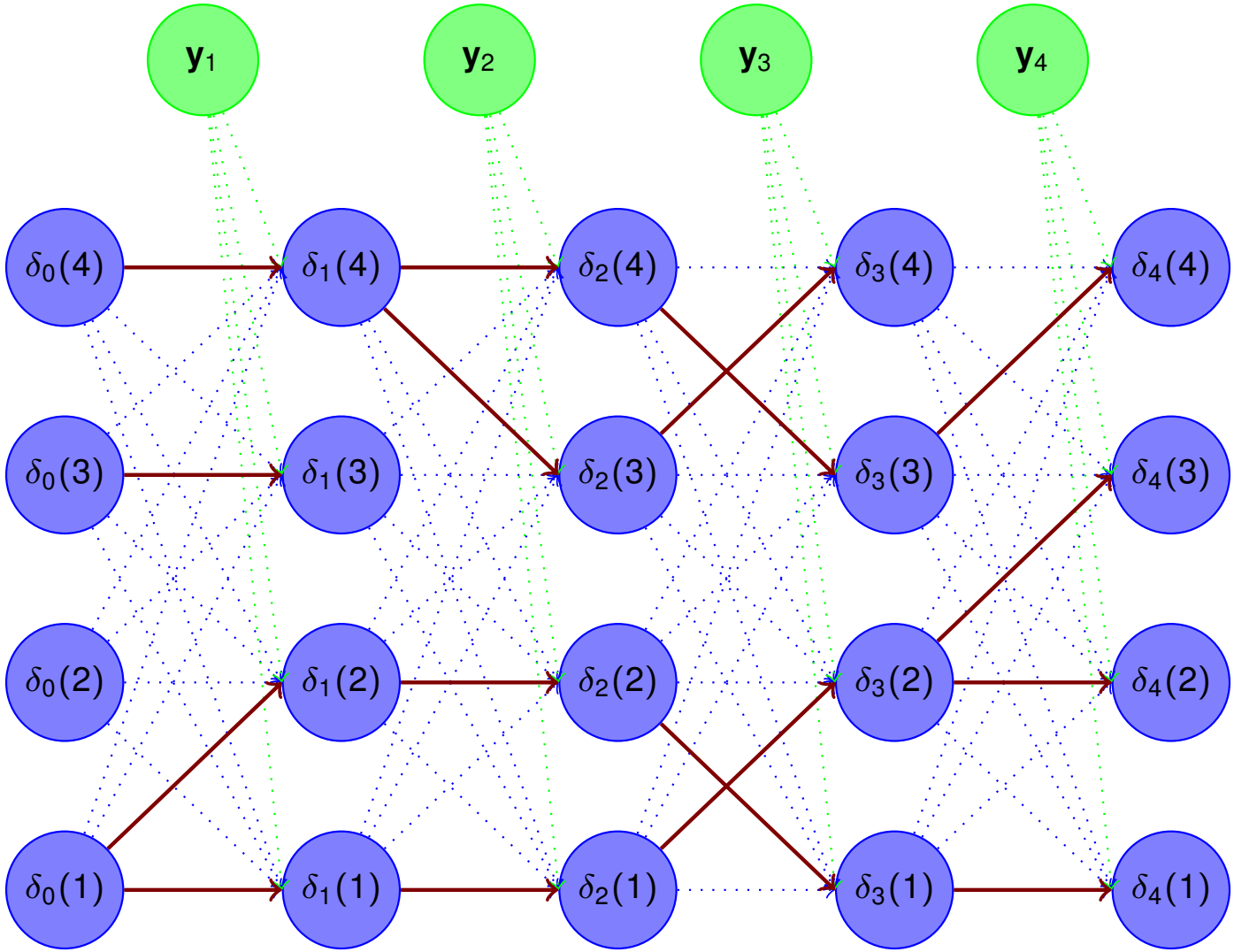
Viterbi, phase 1 : **Itération**

A l'instant k, $\psi_k(i)$ et $\psi_k(j)$ avec $i \neq j$ peuvent être identiques !

Viterbi, phase 1 : **Itération**



Viterbi, phase 1 : **Itération**



Viterbi, phase 2 : **Backtracking**

Le dernier état $\mathbf{x}_T^*$ de la séquence $\mathbf{x}_{0:T}^*$ la plus vraisemblable connaissant les observations $\mathbf{y}_{1:T}$ correspond au maximum de $\delta_T(j)$.

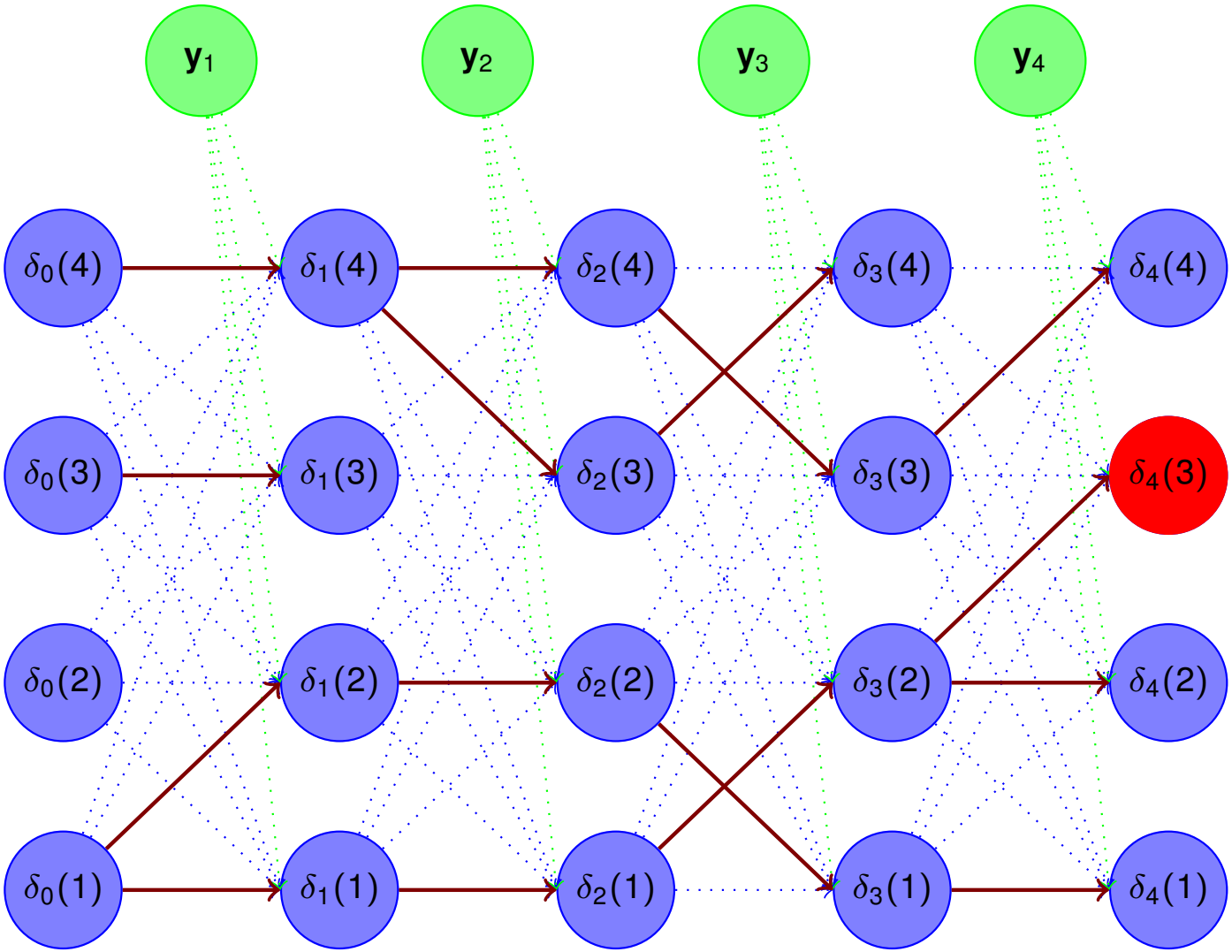
$$\mathbf{x}_T^* = \arg\max_j \delta_T(j) \ .$$

L'avant-dernier état $\mathbf{x}_{T-1}^*$ de la séquence $\mathbf{x}_{0:T}$ la plus vraisemblable est alors donné par $\psi_T(\mathbf{x}_T^*)$.

Par récursion, on a $\forall k \in [T-1..1]$ par pas de -1,

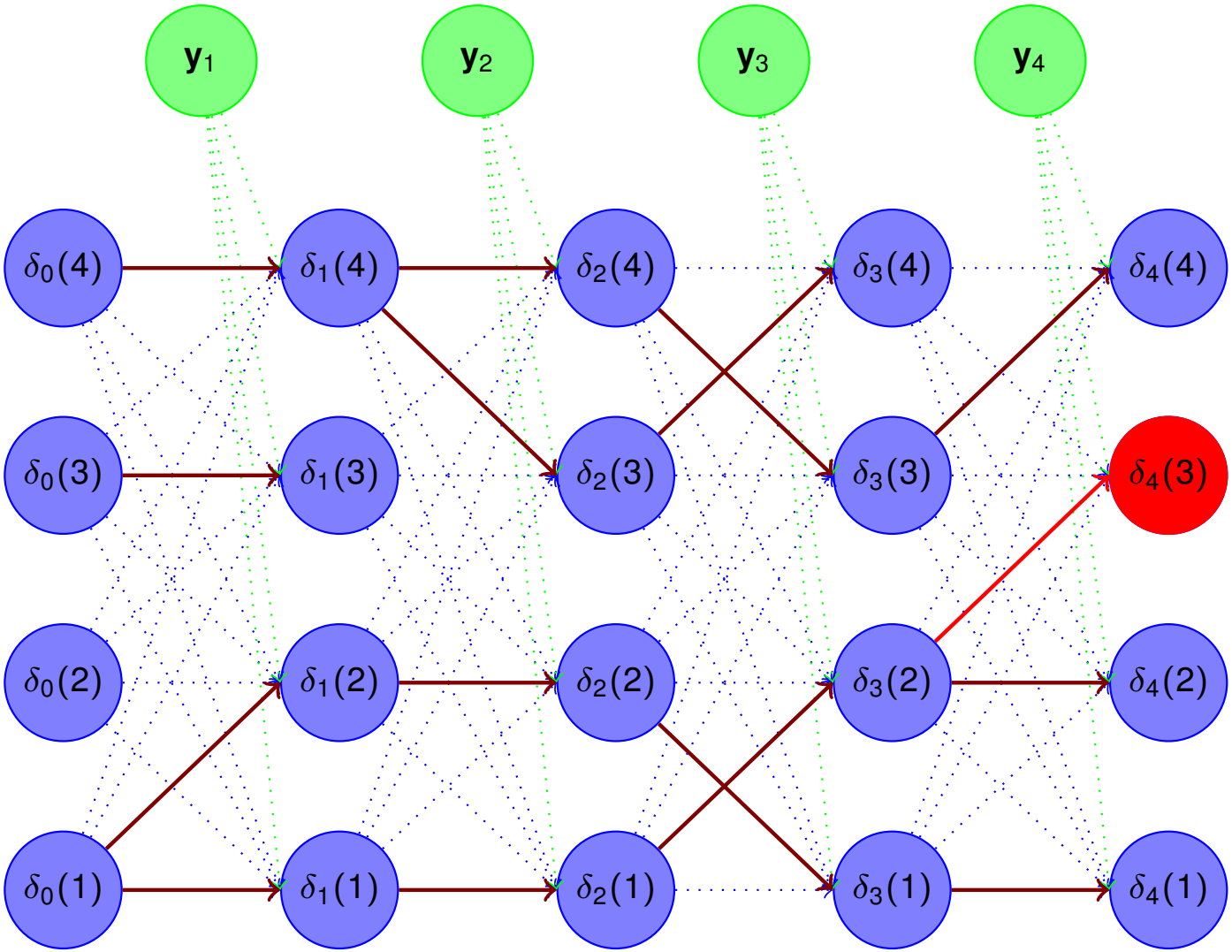
$$\mathbf{x}_k^* = \psi_{k+1}(\mathbf{x}_{k+1}^*)$$

Viterbi, phase 2 : **Backtracking**



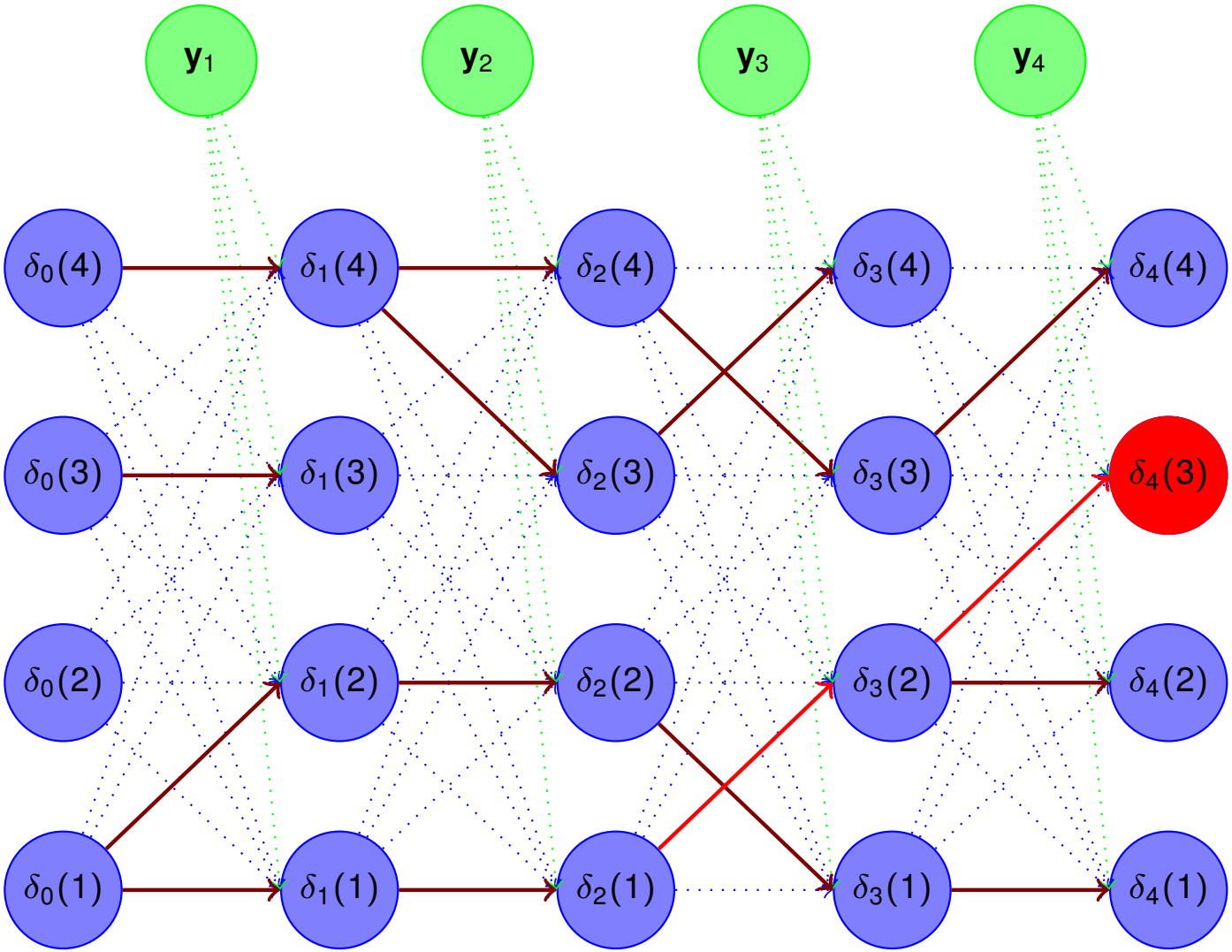
$\arg\max_j \delta_4(j) = 3$

Viterbi, phase 2 : **Backtracking**



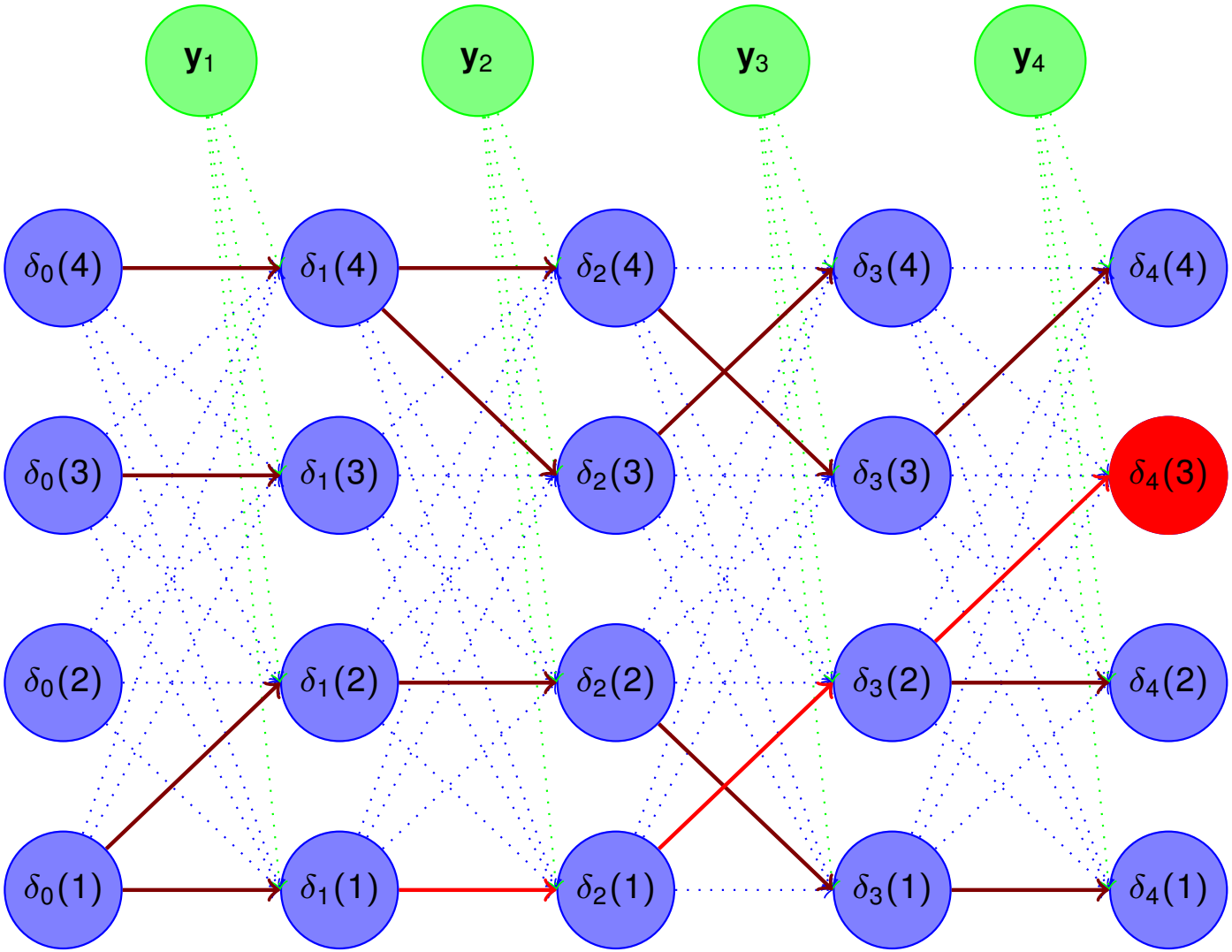
$\arg\max_j \delta_4(j) = 3$

Viterbi, phase 2 : **Backtracking**



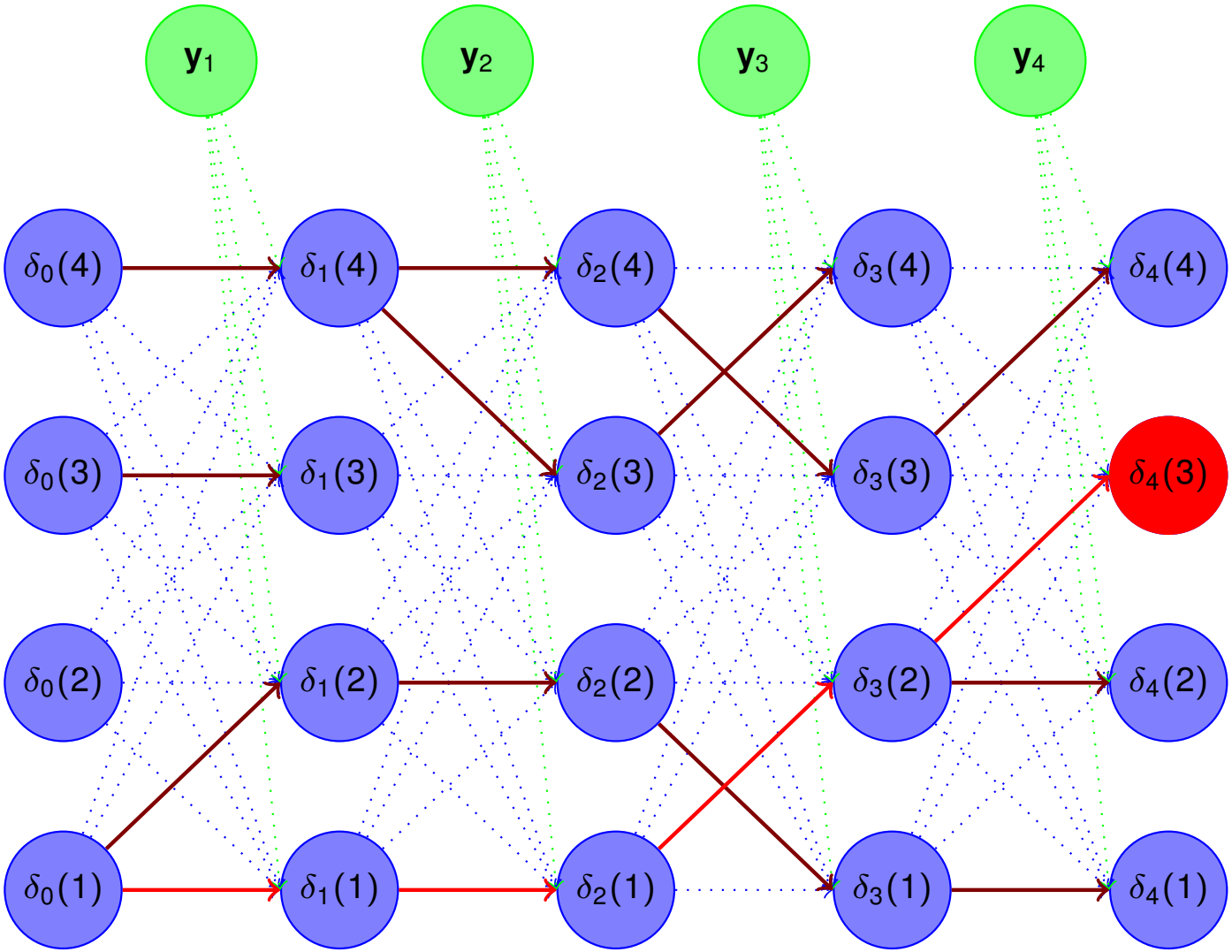
$\arg\max_j \delta_4(j) = 3$

Viterbi, phase 2 : **Backtracking**



$\arg\max_j \delta_4(j) = 3$

Viterbi, phase 2 : **Backtracking**



$\arg\max_j \delta_4(j) = 3$

Séquence la plus vraisemblable $[1, 1, 1, 2, 3]$.

Viterbi, full picture

(1) Initialisation

$$\delta_0(i) \quad = \quad \Pi_i$$

(2) Itération

$\forall k \in [1..T]$ et $\forall j \in [1..n]$

$$\begin{aligned} \delta_k(j) &= \left[\max_i \delta_{k-1}(i)\mathbf{A}_{i,j}\right]\mathbf{B}_{j,\mathbf{y}_k} \ , \\ \psi_k(j) &= \arg\max_i \delta_{k-1}(i)\mathbf{A}_{i,j} \ . \end{aligned}$$

(3) Terminaison

$$\begin{aligned} p(\mathbf{x}^*_{0:T}, \mathbf{y}_{1:T}) &= \max_i \delta_T(i) \ , \\ \mathbf{x}^*_T &= \arg\max_i \delta_T(i) \ . \end{aligned}$$

où $\mathbf{x}^*_{0:T}$ est la séquence la plus vraisemblable.

(4) Backtracking

$\forall k \in [T-1..1]$ par pas de -1,

$$\mathbf{x}^*_k = \psi_{k+1}(\mathbf{x}^*_{k+1})$$

Illustration Viterbi

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie]

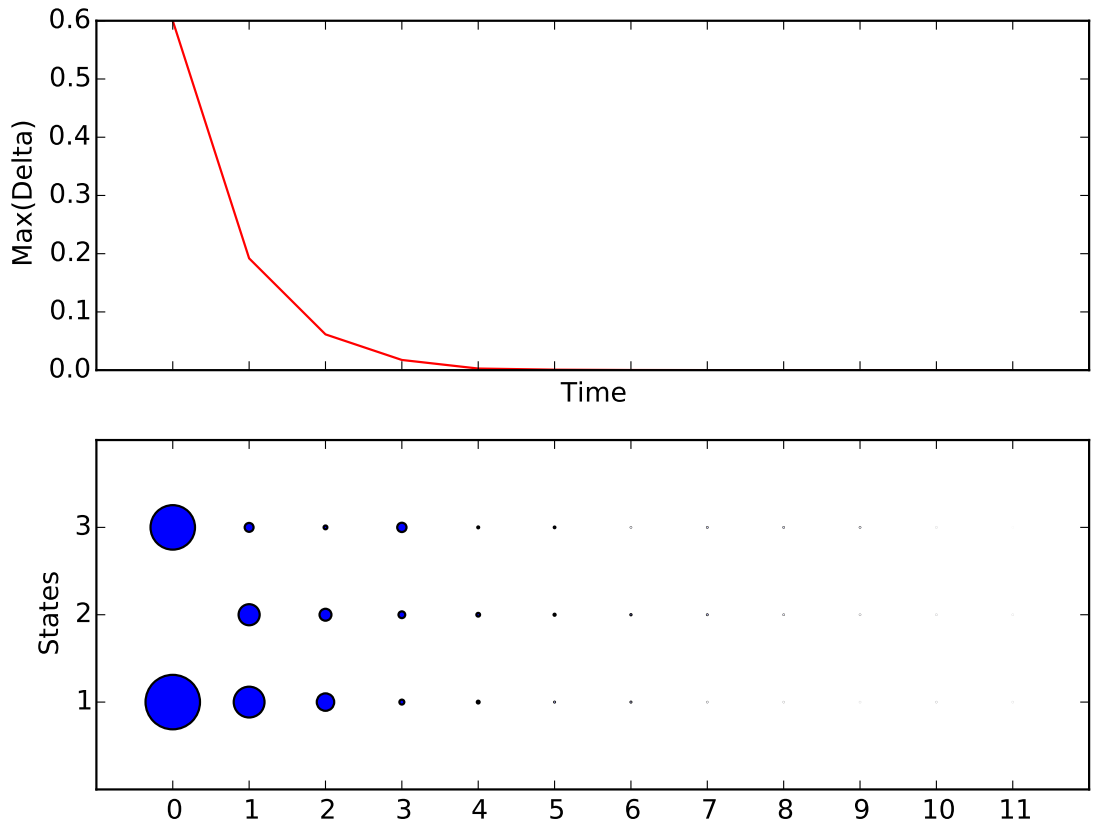


Illustration Viterbi

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie]

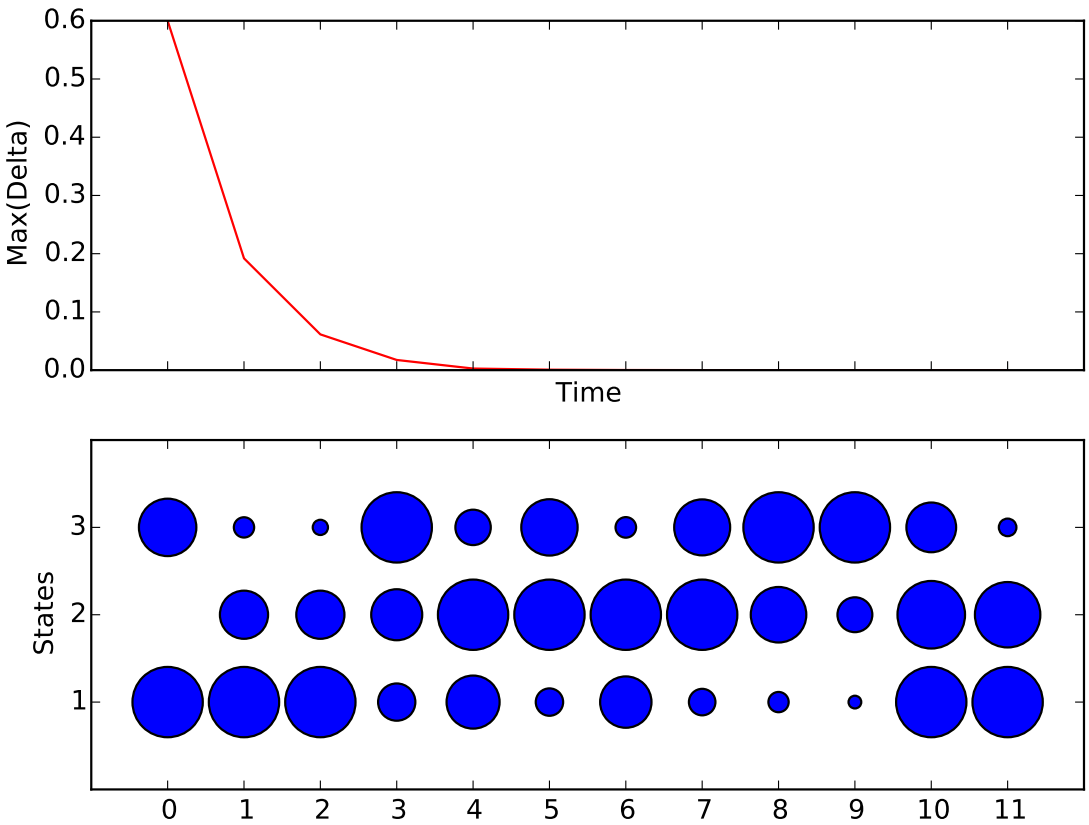


Illustration Viterbi

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie]

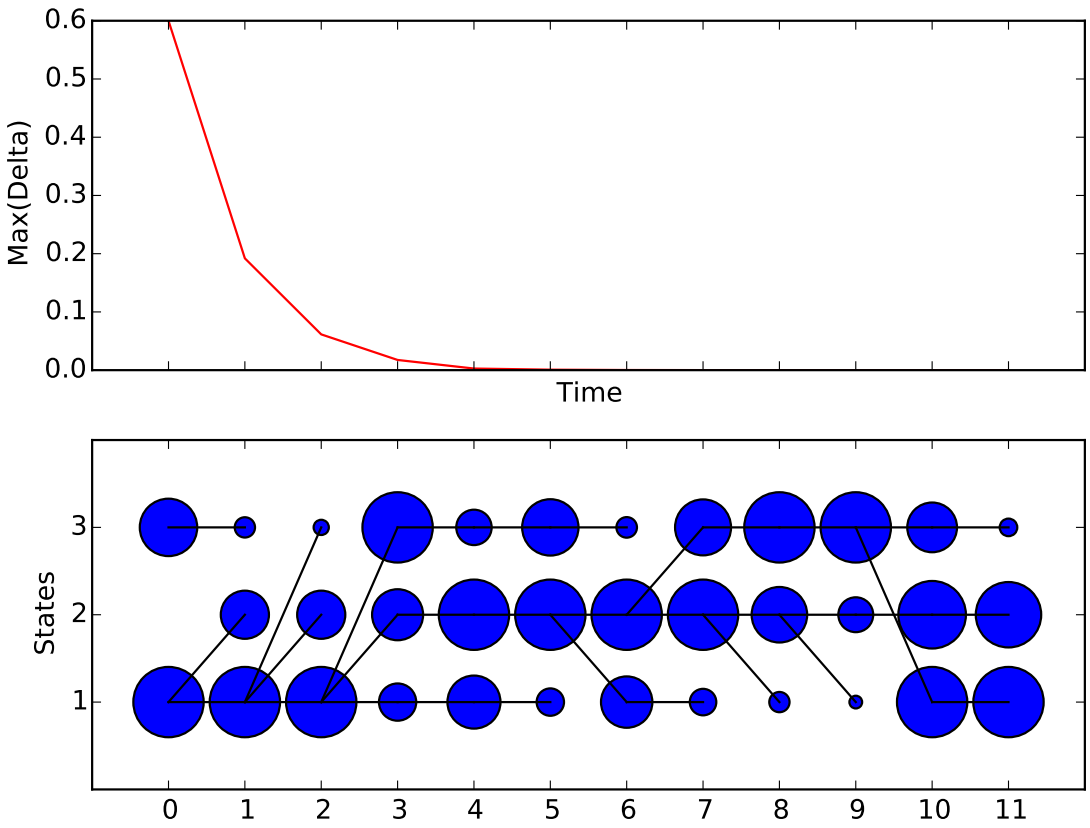
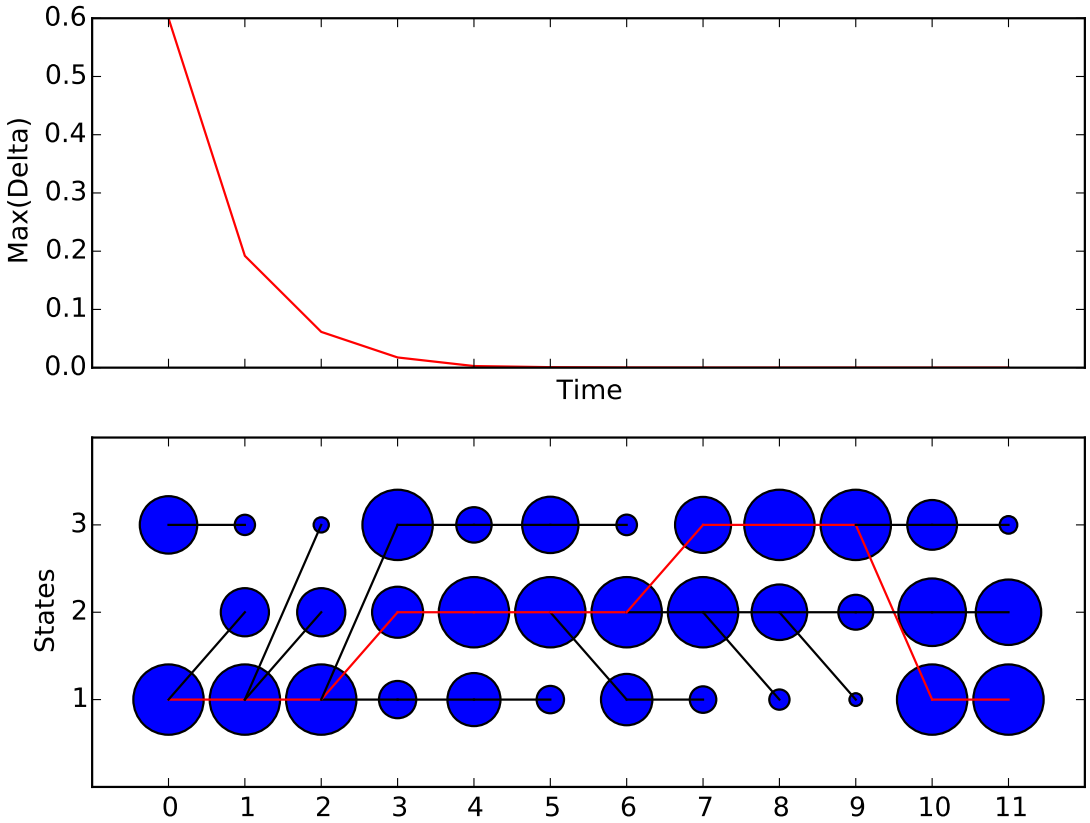


Illustration Viterbi

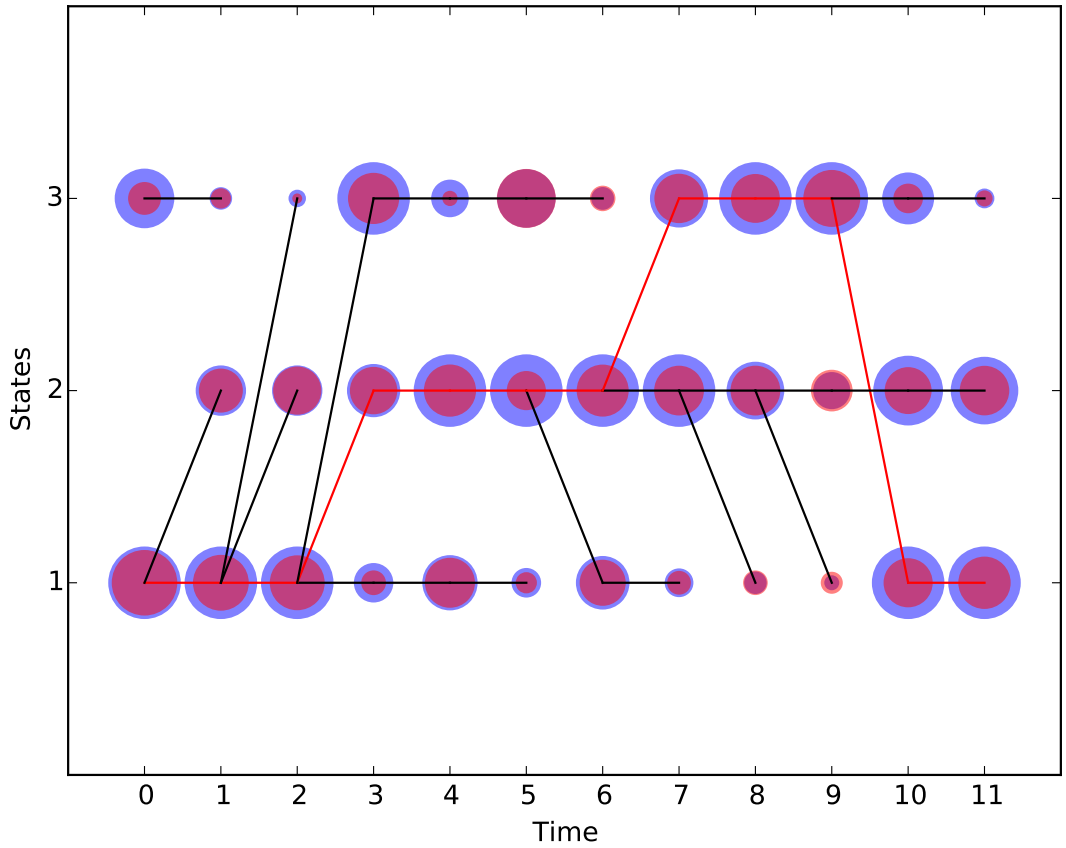
Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie]



[Pluvieux, Pluvieux, Pluvieux, Nuageux, Nuageux, Nuageux, Nuageux, Ensoleillé, Ensoleillé, Ensoleillé, Pluvieux, Pluvieux]

Illustration Viterbi

Séquence [parapluie, parapluie, ∅, parapluie, ∅, parapluie, ∅, ∅, ∅, parapluie, parapluie]



[Pluvieux, Pluvieux, Pluvieux, Nuageux, Nuageux, Nuageux, Nuageux, Ensoleillé, Ensoleillé, Ensoleillé, Pluvieux, Pluvieux]

Section 3

Apprentissage des chaînes de Markov à états cachés

Apprentissage des HMM

Comment faire lorsque l'on dispose de S séquences pour apprendre un modèle $\lambda = \Pi, A, B$ correspondant à ces séquences ?

Deux cas

- Apprentissage supervisé : pour chacune des séquences, on dispose à la fois de la séquence d'état $\mathbf{x}_{0:T}$ et de la séquence d'observation $\mathbf{y}_{1:T}$,
- Apprentissage non-supervisé : on ne dispose que des observations $\mathbf{y}_{1:T}$.

On notera pour la séquence s, les états $\mathbf{x}^{(s)}$, les observations $\mathbf{y}^{(s)}$ et la durée de la séquence $T^{(s)}$.

Apprentissage supervisé

Comment fait-on ?

C'est un simple comptage !

$$\mathbf{A}_{i,j} = \frac{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}_k = i, \mathbf{x}_{k+1} = j)}{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}_k = i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_{k=1}^{T} \mathbb{I}(\mathbf{x}_k = i, \mathbf{y}_k = j)}{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}_k = i)}$$

$$\Pi_{\mathbf{x}_0} = 1 \quad \text{et} \quad \Pi_i = 0 \quad \forall i \neq \mathbf{x}_0$$

Avec une seule observation.

Apprentissage supervisé

Comment fait-on ?

C'est un simple comptage !

$$\mathbf{A}_{i,j} = \frac{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{(s)} = i, \mathbf{x}_{k+1}^{(s)} = j)}{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{(s)} = i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_s \sum_{k=1}^{T^{(s)}} \mathbb{I}(\mathbf{x}_k^{(s)} = i, \mathbf{y}_k^{(s)} = j)}{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{(s)} = i)}$$

$$\Pi_i = \frac{\sum_s \mathbb{I}(\mathbf{x}_0^{(s)} = i)}{\sum_s 1} = \frac{\sum_s \mathbb{I}(\mathbf{x}_0^{(s)} = i)}{S}$$

Avec S observations.

Apprentissage non-supervisé

Pas si simple

Il nous manque à la fois :

- λ ;
- et $\mathbf{x}^{(s)}$.

Si on connaît l'un on sait déjà estimer l'autre. Mais comment faire pour estimer les deux en même temps ?

On considérera le nombre d'états cachés n comme un hyper-paramètre à fixer avant l'apprentissage. Il pourra par exemple être choisi par validation croisée.
Les transitions d'états autorisés ou non seront aussi connus avant l'apprentissage ($\mathbf{A}$ est-elle creuse ? si oui où ?).

Rappel : Algorithme EM

Vous avez déjà vu un cas où on cherche à la fois les paramètres d'un modèle et l'étiquetage issu de ce modèle :

Le clustering par mélange de gaussienne.

Ce problème est résolu par l'algorithme *Espérance-Maximisation* (*Expectation-Maximization*).

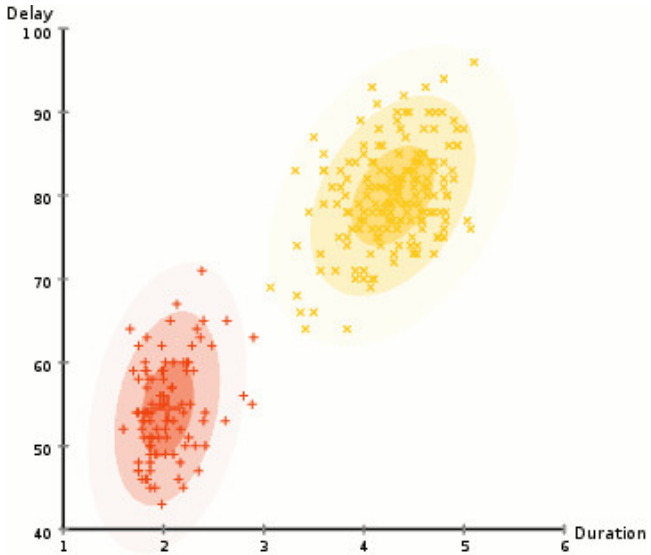


FIGURE: Old faithfull

Source Wikipedia, Auteur Chire, Licence CC BY-SA 3.0

Rappel : Algorithme EM

Le modèle est composé de G gaussiennes.

$$p(\mathbf{x}|\lambda) = \sum_{g=1}^{G} \pi_g f_{\mathcal{N}}(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

où π_g dénote la probabilité *a priori* de g, $\boldsymbol{\mu}_g$ son centre, $\boldsymbol{\Sigma}_g$ sa covariance, et $f_{\mathcal{N}}$ la d.d.p. d'une loi normale,

$$f_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_g|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_g)^\intercal \boldsymbol{\Sigma}_g^{-1} (\mathbf{y} - \boldsymbol{\mu}_g)\right) \quad .$$

Les paramètres sont $\lambda = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \quad \forall \ g \ in \ [1..G]\}$.
On a,

$$\sum_{g=1}^{G} \pi_g = 1 \quad .$$

Quel est le jeux de paramètre λ qui maximise la vraisemblance $p(\mathbf{x}|\lambda)$ des K observations $\mathbf{x}_k$?
Problème non-convexe $\rightarrow$ maximum local.

Rappel : Algorithme EM

On itère sur deux étapes jusqu'à obtention d'un plateau sur la vraisemblance.

Espérance

Les paramètres λ fixés, on calcule la vraisemblance globale et l'appartenance $o_{k,g}$ à une gaussienne de chaque observation k.

$$
\begin{aligned}
p(\mathbf{x}_k|\lambda) &= \sum_{g=1}^{G} \pi_g f_{\mathcal{N}}(\mathbf{x}_k|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \\
o_{k,g} &= \frac{\pi_g f_{\mathcal{N}}(\mathbf{x}_k|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{p(\mathbf{x}_k|\lambda)} \\
p(\mathbf{x}|\lambda) &= \prod_k p(\mathbf{x}_k|\lambda)
\end{aligned}
$$

Maximisation

A partir des appartenances, on calcule les paramètres λ qui maximise la vraisemblance.

Pour chaque gaussien g,

$$
\begin{aligned}
\pi_g &= \frac{1}{K}\sum_k o_{k,g} \ , \\
\boldsymbol{\mu}_g &= \frac{\sum_k o_{k,g}\mathbf{x}_k}{\sum_k o_{k,g}} \ , \\
\boldsymbol{\Sigma}_g &= \frac{\sum_k o_{k,g}(\mathbf{x}_k - \boldsymbol{\mu}_g)(\mathbf{x}_k - \boldsymbol{\mu}_g)^\mathsf{T}}{\sum_k o_{k,g}} \ .
\end{aligned}
$$

(Attention maximum local ! dépend de l'initialisation)

Apprentissage par EM des HMM

On reprend le même principe pour les HMM, on itère :

- A partir de $\lambda = \{\Pi, \mathbf{A}, \mathbf{B}\}$ fixé, on estime les états $\mathbf{x}$;
- A partir des estimation des états, on calcule $\lambda = \Pi, \mathbf{A}, \mathbf{B}$.

Cela converge vers un maximum local qui dépend de l'initialisation de λ.

Deux approches

Il existe deux approches quand à la prise en compte de l'estimation des états,

- soit on traite la séquence d'état le plus vraisemblable $\mathbf{x}^*_{0:T}$,
  i.e. appartenance $o \in \{0, 1\}$, $\Rightarrow$ apprentissage par Viterbi ;
- soit on s'appuie sur la distribution des états $\gamma_k(i)$,
  i.e. appartenance $o \in [0 \quad 1]$, $\Rightarrow$ apprentissage par Baum-Welch.

Apprentissage par Viterbi

Initialisation

On tire de façon aléatoire $\lambda = \{\Pi, \mathbf{A}, \mathbf{B}\}$ (en prenant compte la structure de $\mathbf{A}$).

Espérance

A λ fixé, on applique le décodage de Viterbi, on récupère la séquence d'états optimale $\mathbf{x}^*_{0:T}$.

Maximisation

$$\mathbf{A}_{i,j} = \frac{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}^*_k = i, \mathbf{x}_{k+1} = j)}{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}^*_k = i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_{k=1}^{T} \mathbb{I}(\mathbf{x}^*_k = i, \mathbf{y}_k = j)}{\sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}^*_k = i)}$$

$$\Pi_{\mathbf{x}^*_0} = 1 \quad \text{et} \quad \Pi_i = 0 \quad \forall i \neq \mathbf{x}^*_0$$

Avec une seule observation.

Apprentissage par Viterbi

Initialisation

On tire de façon aléatoire $\lambda = \{\Pi, \mathbf{A}, \mathbf{B}\}$ (en prenant compte la structure de $\mathbf{A}$).

Espérance

A λ fixé, on applique le décodage de Viterbi, on récupère les S séquences d'états optimale $\mathbf{x}_{0:T}^{*,(s)}$.

Maximisation

On met à jour λ grâce un comptage à partir des séquences $\mathbf{x}_{0:T}^{*,(s)}$.

$$\mathbf{A}_{i,j} = \frac{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{*,(s)} = i, \mathbf{x}_{k+1}^{*,(s)} = j)}{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{*,(s)} = i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_s \sum_{k=1}^{T^{(s)}} \mathbb{I}(\mathbf{x}_k^{*,(s)} = i, \mathbf{y}_k^{(s)} = j)}{\sum_s \sum_{k=0}^{T^{(s)}-1} \mathbb{I}(\mathbf{x}_k^{*,(s)} = i)}$$

$$\Pi_i = \frac{\sum_s \mathbb{I}(\mathbf{x}_0^{*,(s)} = i)}{\sum_s 1} = \frac{\sum_s \mathbb{I}(\mathbf{x}_0^{*,(s)} = i)}{S}$$

Avec S observations.

Apprentissage par Baum-Welch

Initialisation

On tire de façon aléatoire $\lambda = \{\Pi, \mathbf{A}, \mathbf{B}\}$ (en prenant compte la structure de $\mathbf{A}$).

Espérance

A λ fixé, on applique l'algorithme *forward-backward*, on récupère

- $$\begin{aligned} \gamma_k(i) &= p(\mathbf{x}_k = i | \lambda, \mathbf{y}_{1:T}) \\ \gamma_k(i) &= \frac{\alpha_k(i)\beta_k(i)}{\sum_l \alpha_k(l)\beta_k(l)} \end{aligned}$$
- $$\begin{aligned} \xi_k(i,j) &= p(\mathbf{x}_k = i, \mathbf{x}_{k+1} = j | \lambda, \mathbf{y}_{1:T}) \\ \xi_k(i,j) &= \frac{\alpha_k(i)\mathbf{A}_{i,j}\mathbf{B}_{j,\mathbf{y}_{k+1}}\beta_{k+1}(j)}{\sum_l \alpha_k(l)\beta_k(l)} \end{aligned}$$

Apprentissage par Baum-Welch

Maximisation

On met à jour λ à partir des distributions $\gamma_k(i)$ et $\xi_k(i,j)$.

$$\mathbf{A}_{i,j} = \frac{\sum_{k=0}^{T} \xi_k(i,j)}{\sum_{k=0}^{T-1} \gamma_k(i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_{k=1}^{T} \mathbb{I}(\mathbf{y}_k = j)\gamma_k(i)}{\sum_{k=1}^{T-1} \gamma_k(i)}$$

$$\Pi_i = \gamma_0(i)$$

Maximisation : S observations

On met à jour λ à partir des distributions $\gamma_k^{(s)}(i)$ et $\xi_k^{(s)}(i,j)$.

$$\mathbf{A}_{i,j} = \frac{\sum_s \sum_{k=0}^{T} \xi_k^{(s)}(i,j)}{\sum_s \sum_{k=0}^{T-1} \gamma_k^{(s)}(i)}$$

$$\mathbf{B}_{i,j} = \frac{\sum_s \sum_{k=1}^{T} \mathbb{I}(\mathbf{y}_k^{(s)} = j)\gamma_k^{(s)}(i)}{\sum_s \sum_{k=1}^{T-1} \gamma_k^{(s)}(i)}$$

$$\Pi_i = \frac{1}{S} \sum_s \gamma_0^{(s)}(i)$$

Section 4

Observations dans un espace continu

Principe des HMM à observations dans un espace continu

Les observations $\mathbf{y}$ prennent leur valeur dans un espace continu, par exemple $\mathbb{R}^m$. Ces observations sont tirées d'une distribution (d'émission) qui ne dépend uniquement de l'état $\mathbf{x}$ quelque soit l'instant k. Ainsi,

$$\mathbf{y}_k \sim p(\mathbf{y}_k|\mathbf{x}_k) \forall k \quad .$$

Utilisation

On abandonne $\mathbf{B}$! Lors des phase de forward-backward ou Viterbi pour le calcul de la vraisemblance ou du décodage, on remplace simplement $\mathbf{B}_{i,j}$ par $p(\mathbf{y}|\mathbf{x} = i)$.

Apprentissage

Si la distribution $p(\mathbf{y}_k|\mathbf{x}_k, \theta)$ dépend de paramètres θ (par exemple le centre $\boldsymbol{\mu}$ et l'écart type $\boldsymbol{\Sigma}$ pour une loi normale), on apprendra les paramètres θ en complément des paramètres de λ par les algorithmes EM précédemment évoqué.

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= \mathbf{B}_{j,\mathbf{y}_k} \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} \mathbf{B}_{j,\mathbf{y}_{k+1}} \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= p(\mathbf{y}_k | \mathbf{x}_k = j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} \mathbf{B}_{j,\mathbf{y}_{k+1}} \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} \mathbf{B}_{j,\mathbf{y}_{k+1}} \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} p(\mathbf{y}_{k+1} | \mathbf{x}_{k+1} = j) \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} f_{\mathcal{N}}(\mathbf{y}_{k+1}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] \mathbf{B}_{j,\mathbf{y}_k} \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} f_{\mathcal{N}}(\mathbf{y}_{k+1}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] p(\mathbf{y}_k | \mathbf{x}_k = j) \quad .
\end{aligned}$$

Exemple de distribution d'émission : une distribution normale (1)

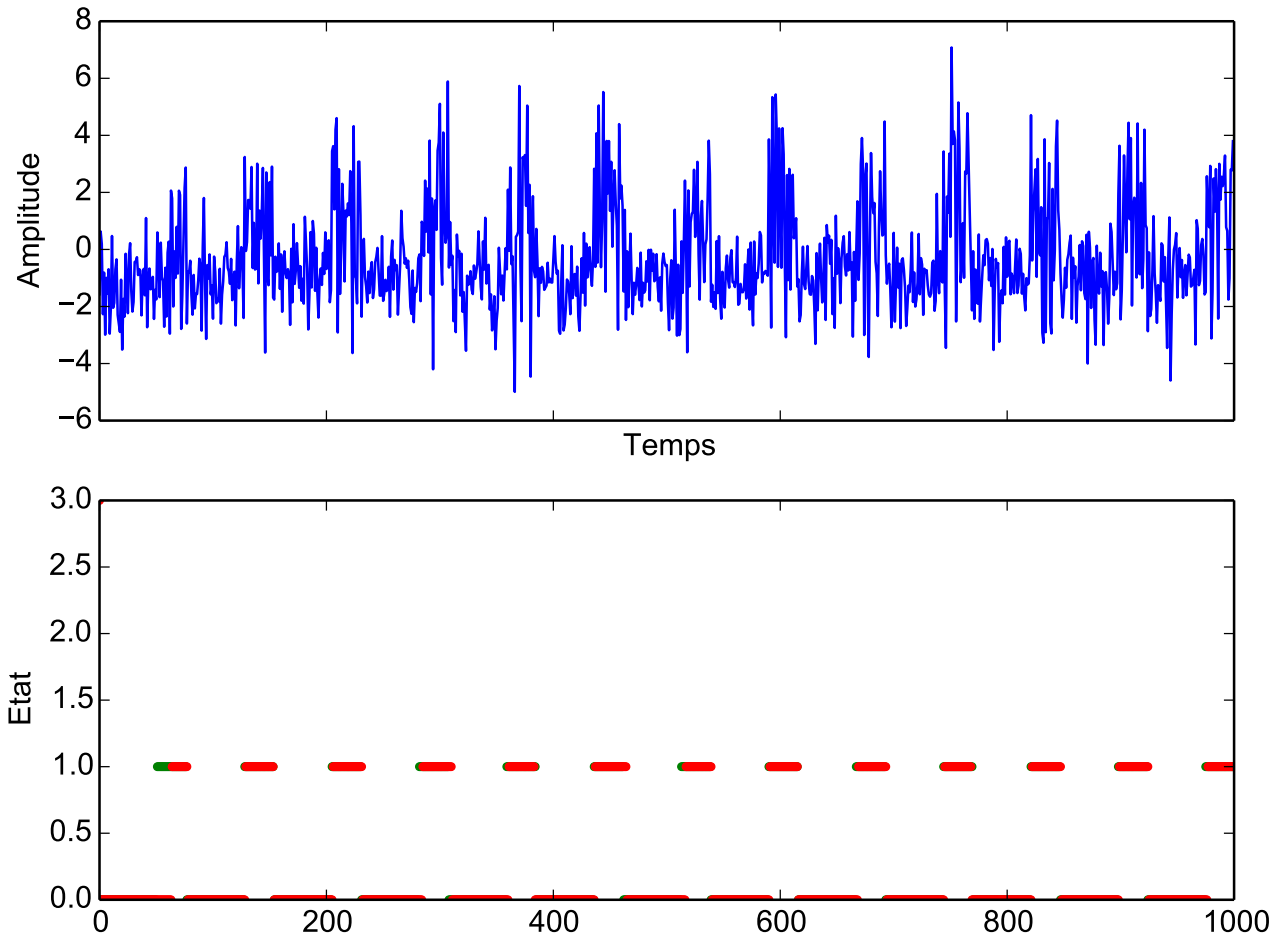
Si il y a n états possibles alors nous avons n distributions d'émissions $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ où $i \in \{1..n\}$ et $\mathbf{y} \in \mathbb{R}^m$.
La fonction de répartition de chacune de ces distributions est donnée par,

$$f_{\mathcal{N}}(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{m/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\mathbf{y} - \boldsymbol{\mu}_i)\right) \quad .$$

Utilisation

$$\begin{aligned}
\alpha_k(j) &= f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sum_{i=1}^{n} \mathbf{A}_{i,j} \times \alpha_{k-1}(i) \quad , \\
\beta_k(i) &= \sum_{j=1}^{n} \mathbf{A}_{i,j} f_{\mathcal{N}}(\mathbf{y}_{k+1}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \times \beta_{k+1}(j) \quad , \\
\delta_k(j) &= \left[\max_i \delta_{k-1}(i) \mathbf{A}_{i,j}\right] f_{\mathcal{N}}(\mathbf{y}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad .
\end{aligned}$$

Illustration



Exemple de distribution d'émission : une distribution normale (2)

Apprentissage (Baum-Welch)

Espérance

$$\gamma_k(i) = \frac{\alpha_k(i)\beta_k(i)}{\sum_l \alpha_k(l)\beta_k(l)} ,$$

$$\xi_k(i,j) = \frac{\alpha_k(i)\mathbf{A}_{i,j} f_{\mathcal{N}}(\mathbf{y}_{k+1}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\beta_{k+1}(j)}{\sum_l \alpha_k(l)\beta_k(l)} .$$

Maximisation

$$\mathbf{A}_{i,j} = \frac{\sum_{k=0}^{T} \xi_k(i,j)}{\sum_{k=0}^{T-1} \gamma_k(i)} ,$$

$$\Pi_i = \gamma_0(i) ,$$

$$\boldsymbol{\mu}_i = \frac{\sum_k \gamma_k(i)\mathbf{y}_k}{\sum_k \gamma_k(i)} ,$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_k \gamma_k(i)(\mathbf{y}_k - \boldsymbol{\mu}_i)(\mathbf{y}_k - \boldsymbol{\mu}_i)^\intercal}{\sum_k \gamma_k(i)} .$$

Exercices

1. Extension à S observations ?
2. Extension à un mélange de G gaussiennes ?

Durée de maintient dans un état avant transition

Probabilité de rester dans l'état *i* pendant *d* instants

$$p(\mathbf{x}_{k,k+d-1}) = \mathbf{i}, \mathbf{x}_{k+d} \neq i|\lambda) = p(\mathbf{x}_{k,k+d-1} = \mathbf{i})p(\mathbf{x}_{k+d} \neq i|\mathbf{x}_{k+d-1} = i\lambda)$$

$$p(\mathbf{x}_{k,k+d-1}) = \mathbf{i}, \mathbf{x}_{k+d} \neq i|\lambda) = \mathbf{A}_{i,i}^{(d-1)} * (1 - \mathbf{A}_{i,i})$$

Durée moyenne

$$\bar{d} = \sum_{d=1}^{+\infty} d * p(\mathbf{x}_{k,k+d-1})$$

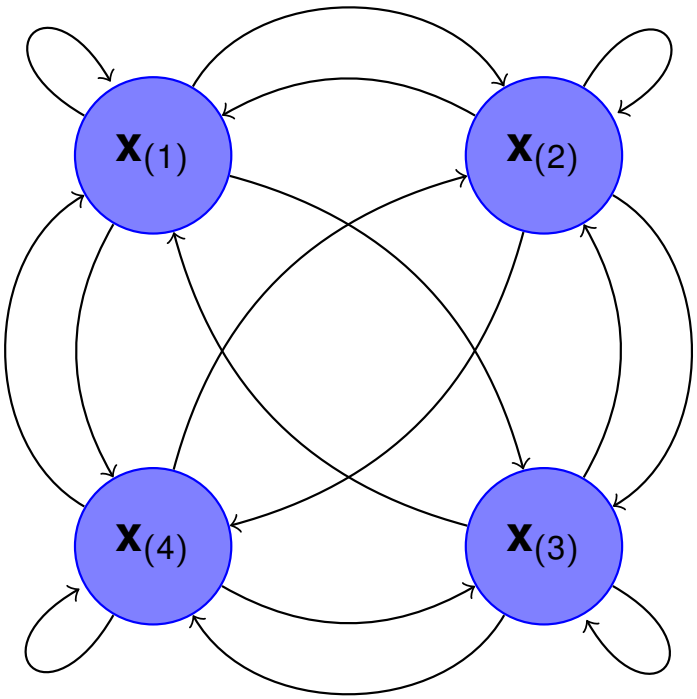
$$\bar{d} = \sum_{d=1}^{+\infty} d * \mathbf{A}_{i,i}^{(d-1)} * (1 - \mathbf{A}_{i,i})$$

$$\bar{d} = \frac{1}{1 - \mathbf{A}_{i,i}}$$

Section 5

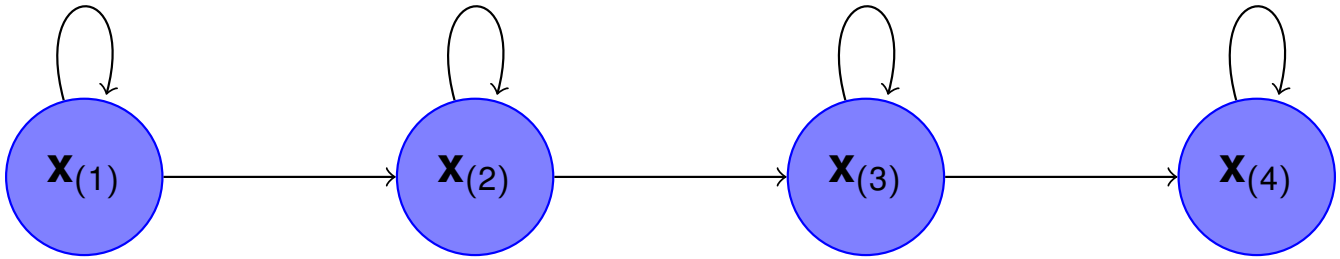
Topologie des HMM

Ergodique



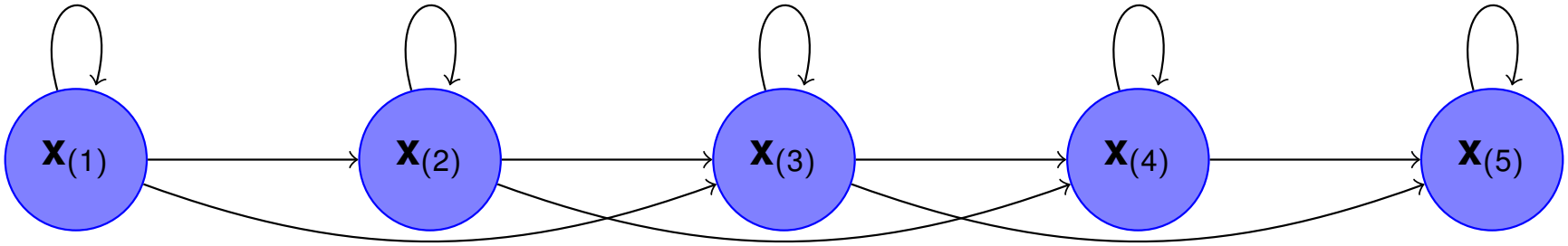
$$\mathbf{A} = \begin{vmatrix} a11 & a12 & a13 & a14 \\ a21 & a22 & a23 & a24 \\ a31 & a32 & a33 & a34 \\ a41 & a42 & a43 & a44 \end{vmatrix}$$

Linéaire



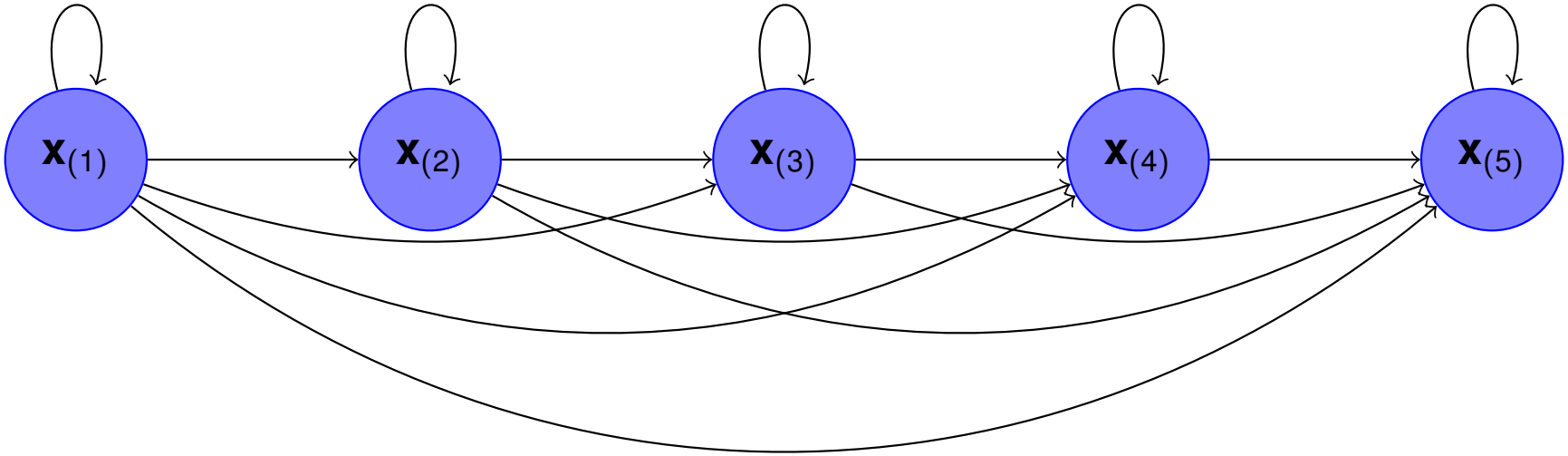
$$\mathbf{A} = \begin{vmatrix} a11 & a12 & 0 & 0 \\ 0 & a22 & a23 & 0 \\ 0 & 0 & a33 & a34 \\ 0 & 0 & 0 & a44 \end{vmatrix}$$

Bakis



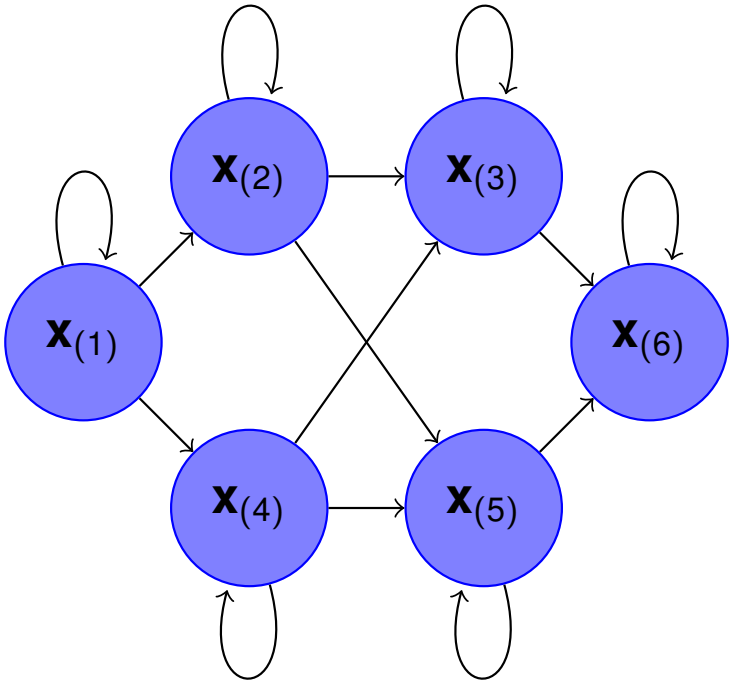
$$\mathbf{A} = \begin{vmatrix} a11 & a12 & a13 & 0 & 0 \\ 0 & a22 & a23 & a24 & 0 \\ 0 & 0 & a33 & a34 & a35 \\ 0 & 0 & 0 & a44 & a45 \\ 0 & 0 & 0 & 0 & a55 \end{vmatrix}$$

Gauche à Droite *Left to right*



$$\mathbf{A} = \begin{vmatrix} a11 & a12 & a13 & a14 & a15 \\ 0 & a22 & a23 & a24 & a25 \\ 0 & 0 & a33 & a34 & a35 \\ 0 & 0 & 0 & a44 & a45 \\ 0 & 0 & 0 & 0 & a55 \end{vmatrix}$$

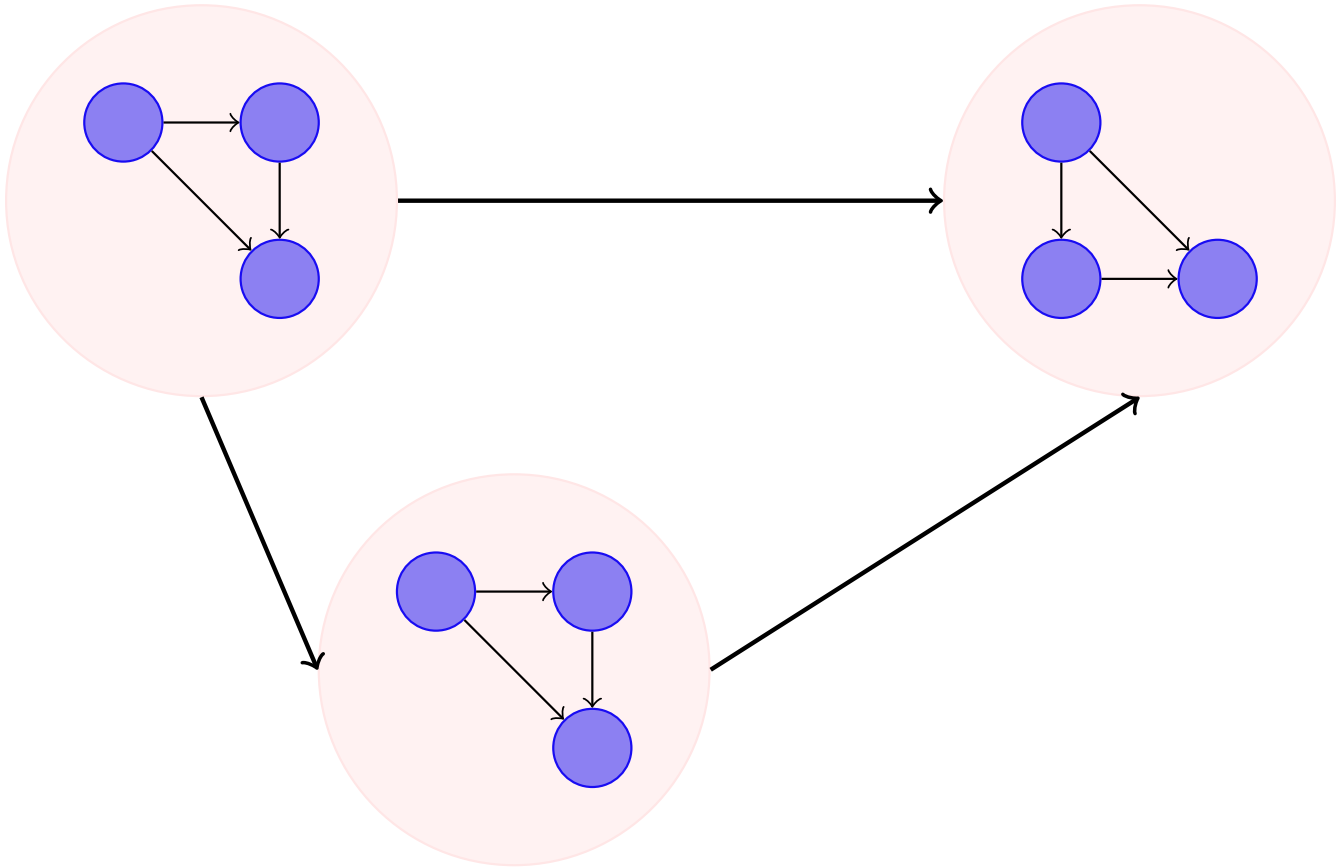
Parallèle



$$\mathbf{A} = \begin{vmatrix} a11 & a12 & 0 & a14 & 0 & 0 \\ 0 & a22 & a23 & 0 & a25 & 0 \\ 0 & 0 & a33 & 0 & 0 & a36 \\ 0 & 0 & a43 & a44 & a45 & 0 \\ 0 & 0 & 0 & 0 & a55 & a56 \\ 0 & 0 & 0 & 0 & & a66 \end{vmatrix}$$

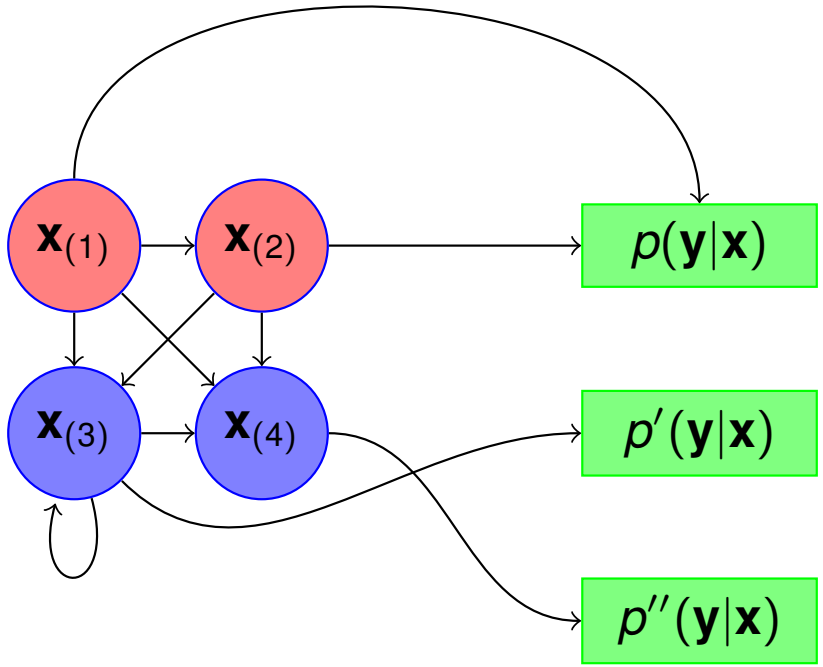
Composition des HMMs

Les HMMs peuvent être composés entre eux pour former un nouveau HMM.



Distributions d'émission liées

On peut forcer plusieurs états à partager les mêmes distributions d'émission.



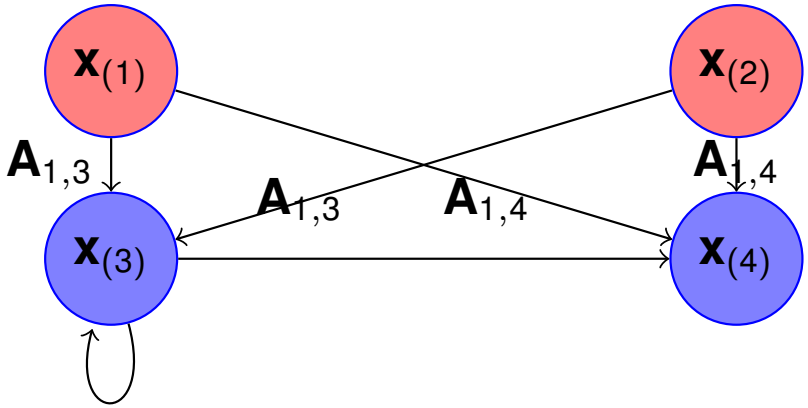
On parle de distributions d'émission liées (ou *tied*).

$$\begin{aligned} \mathbf{B}_{i,.} &= \mathbf{B}_{j,.} \\ p(\mathbf{y}|\mathbf{x}=i) &= p(\mathbf{y}|\mathbf{x}=j) \end{aligned}$$

Exemple d'utilisation : modélisation d'un même caractère dans des mots différents.

Distributions de transition liées

De la même façon, on peut forcer plusieurs états à partager les mêmes distributions de transition.



On parle de distributions de transition liées (ou *tied*).

$$\mathbf{A}_{i,.} = \mathbf{A}_{j,.}$$

Attention les transitions menant aux états peuvent être différentes.

$$\mathbf{A}_{.,i} \neq \mathbf{A}_{.,j}$$

Exemple d'utilisation : modélisation d'une même liaison dans des mots différents.

Stratégies pour la reconnaissance de forme (1)

On dispose d'un ensemble d'apprentissage des exemples de la forme $(\mathbf{y}, z)$ où $\mathbf{y}$ est une séquence d'observations et z est une étiquette dans un espace discret.

Apprentissage

- On isole les séquences d'une même étiquette z ;
- On apprend un HMM sur cet ensemble de séquences, ce qui donne un modèle λ_z spécifique de l'étiquette z.

Test 1 : Par la vraisemblance

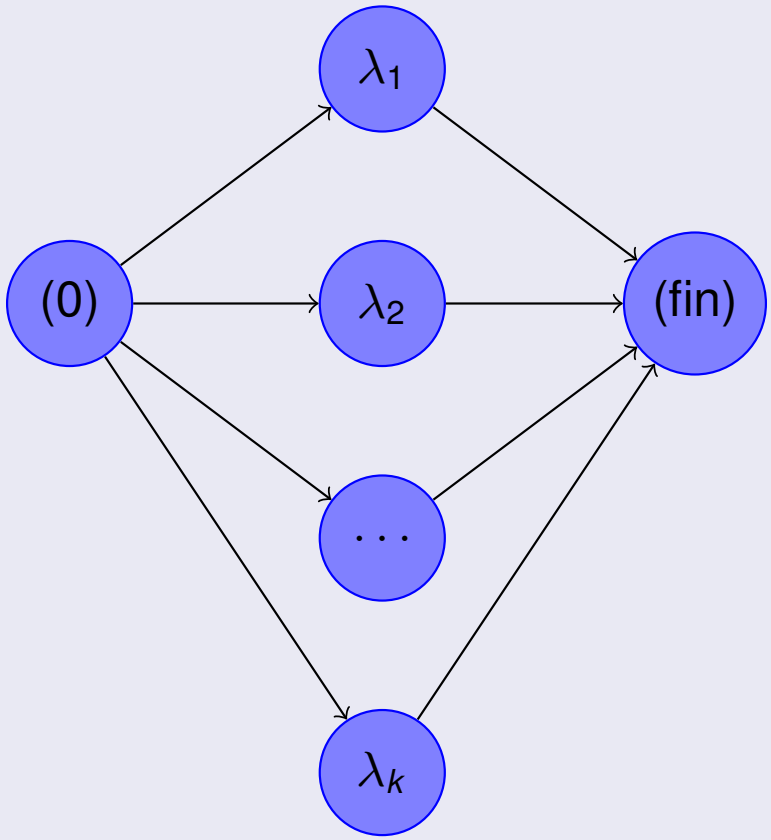
On attribue à la séquence l'étiquette du modèle qui donne la plus grande vraisemblance,

$$z_i = \arg\max_z p(\mathbf{y}_i|\lambda_z) \ .$$

Stratégies pour la reconnaissance de forme (2)

Test 2 : Par décodage de Viterbi

On met tout les HMM en // dans un grand HMM.



On applique le décodage Viterbi pour trouver le modèle le plus vraisemblable.

Section 6

Applications

Techniques spécifiques

Reconnaissance de la parole

- Fenêtre glissante
- Caractéristiques : LPC, ...

$\Rightarrow$ Segmentation ? Phonème ? Mot ? Phrase ?

Reconnaissance de l'écriture

- En-ligne ou *On-line*
  - On connaît la position d'un outil scripteur à travers le temps. On dispose de triplets (x, y, t).
  - Caractéristiques issues par exemple d'analyses morphologiques.
- Hors-ligne ou *Off-line*
  - On dispose d'une image résultat de l'action de l'écriture.
  - Fenêtre glissante
  - Caractéristiques issues de l'analyse d'une image.

$\Rightarrow$ Segmentation ? Caractère ? Mot ? Phrase ?

FIGURE: Fenêtre glissante

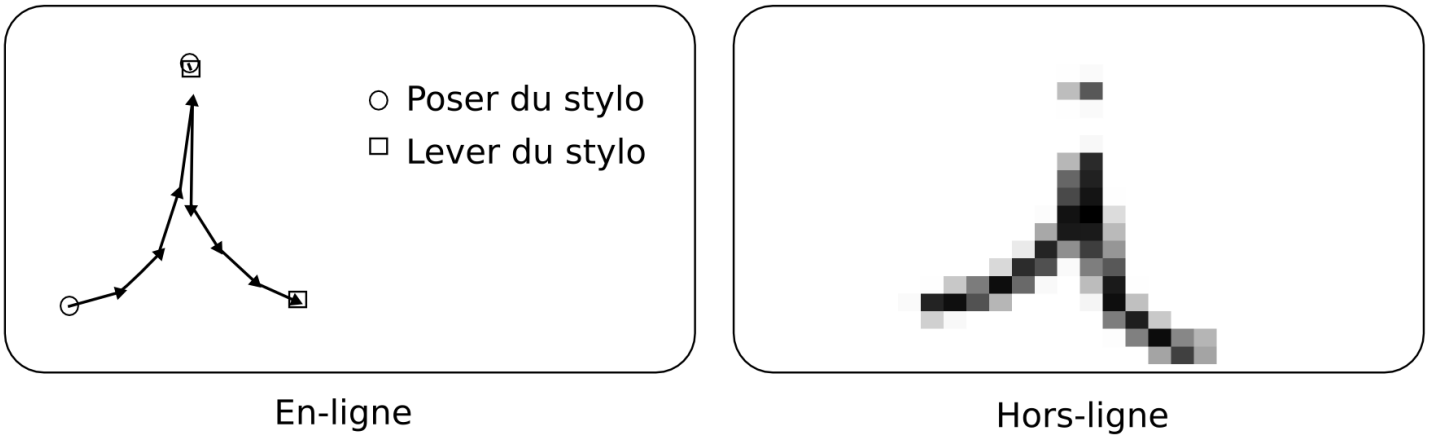


FIGURE: En-ligne Hors-ligne

Source : Wikipedia CC-BY-SA Manproc

La malédiction de la segmentation (1)

Problème

- Pour segmenter, il faut comprendre ce que l'on sépare.
- Pour reconnaître, il faut que la segmentation soit correcte.

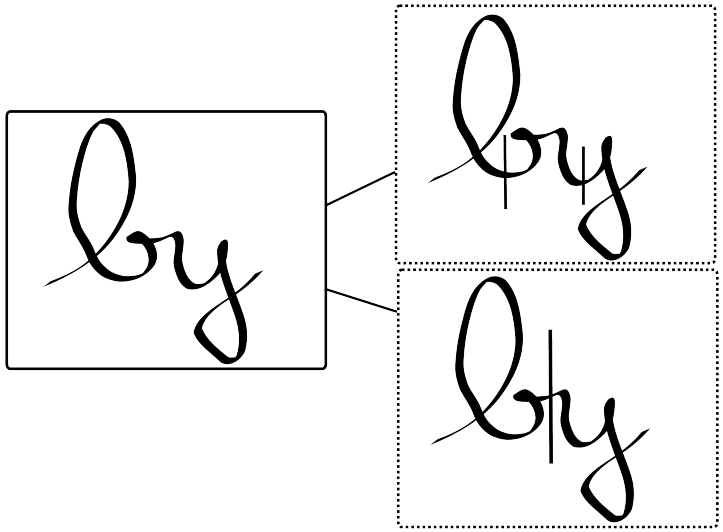


FIGURE: En-ligne Hors-ligne

Source : Wikipedia CC-BY-SA Manproc

La malédiction de la segmentation (2)

Solution

- On segmente (avec ou sans HMM) avant d'apprendre les classifieurs
- On apprend un classifieur pour un mot complet (approche globale ou *holistic*)
- On apprend la segmentation et la classification en même temps (*embeding*)

Pré-segmentation

On dispose d'un algorithme de segmentation caractère par caractère ou phonème par phonème (basé sur des HMM ou non).

Apprentissage

- On applique la segmentation en unité (caractère ou phonème)
- On apprend un modèle de séquence HMM par type d'unité (26 modèles pour les caractères)

Test

- On applique la segmentation en unité (caractère ou phonème)
- On classifie les unités à partir des modèles de séquences.

$\Rightarrow$ Quid du contexte ? Détection de mot impossible.

Approche globale *holistic*

On découpe mot par mot les séquences à apprendre ou à reconnaître.

Apprentissage

- On récupère la base d'apprentissage mot par mot
- On apprend un modèle de séquence HMM par mot (autant de modèles que de mots dans la langue !).

Test

- On segmente mot par mot la séquence (plus simple que caractères par caractères).
- On classifie le mot dans son entièreté

$\Rightarrow$ Beaucoup de modèles appris sans mutualisation de l'information.

Embedding

On reprend le principe de l'approche globale, mais on mutualise l'apprentissage là où c'est possible.

Construction

- On construit autant de modèles qu'il y a de caractères.
- On regroupe les modèles des caractères dans un grand modèle gauche-droite pour former un modèle de mot.

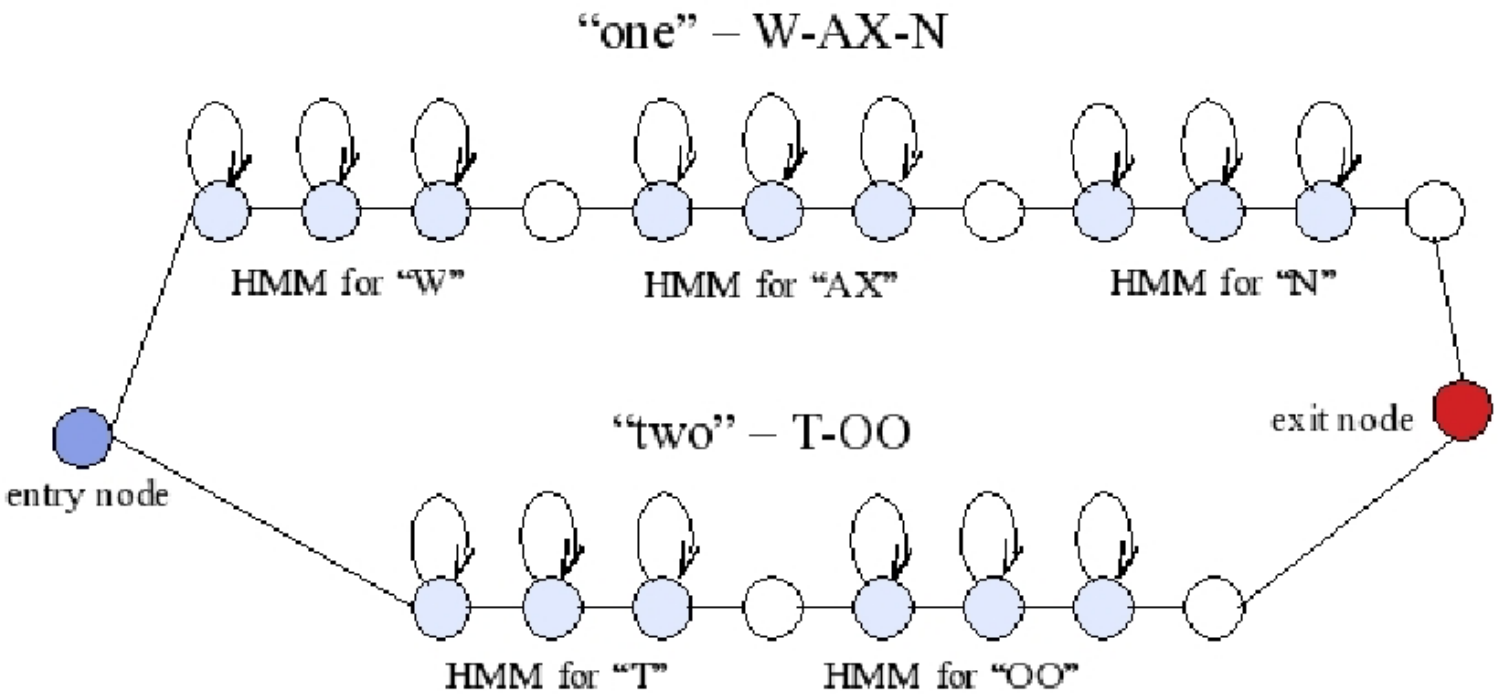
$\rightarrow$ On construit autant de *grands modèles* qu'il y a de mots.

Apprentissage

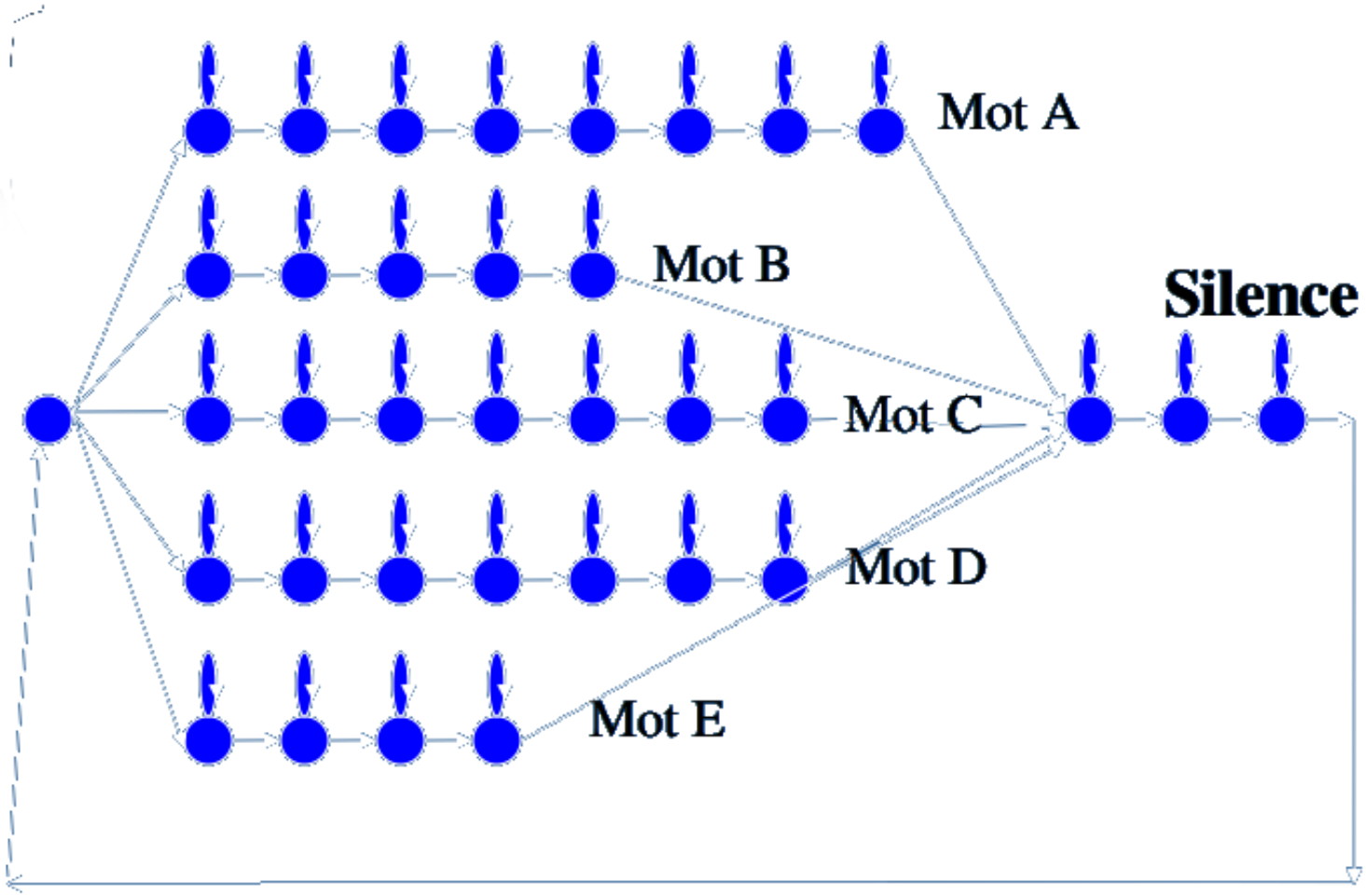
Entre tous les grands modèles,

- les distributions de transitions et d'émissions d'un même modèle de caractère sont liés et ;
- les distributions de transitions entre caractères identiques sont liés.

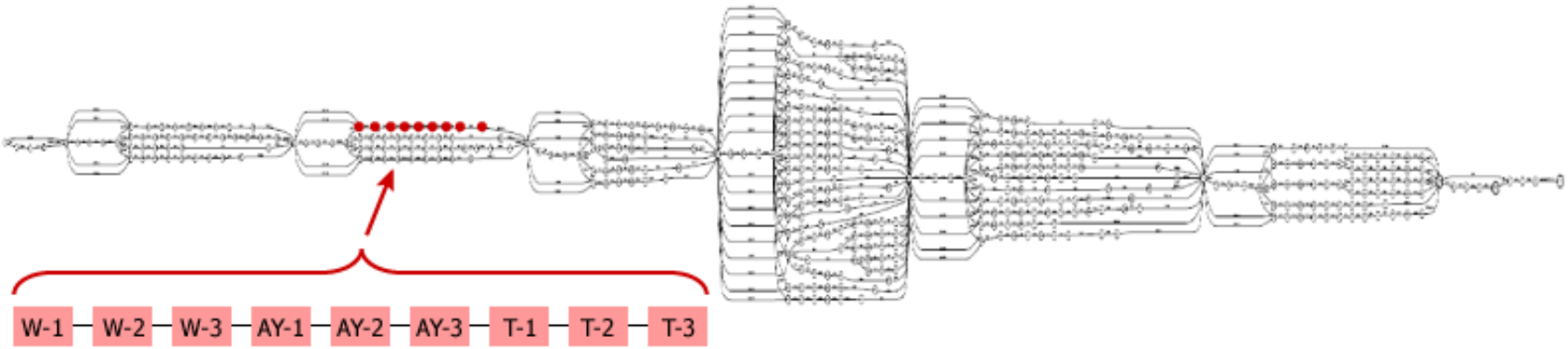
Exemple de décodage



Modèle de langue



Modèle de langue



Section 7

Au delà des HMM