1 Définitions des chaînes de Markov

1.1 Processus stochastique

Soit un ensemble T, engendré par l'indice t, et une variable aléatoire s_t indicée par t, on appelle processus stochastique l'ensemble : $(s_t, t \in T)$. En général, T est dénombrable et représente le temps, le processus est alors dit *discret*.

1.2 Suite stochastique

S = $\{u_1, u_2 \ldots\}$. Entre les instants t et $t+1$, le système passe aléatoirement de l'état u à l'état v.

$\Omega = \{(u, v)\}$.

$s_t = u$ ¡==¿ le système est dans l'état u au temps t.

$(s_1, s_2, \ldots s_t, \ldots)$ est une suite stochastique à ensemble discret d'états.

1.3 Chaîne de Markov

Une chaîne de Markov est une suite stochastique à ensemble discret d'états telle que :

- Suite du premier ordre

$$\begin{aligned} Prob(s_t = v/s_{t-1} = u, s_{t-2} = \ldots) \\ = Prob(s_t = v/s_{t-1} = u) \end{aligned} \tag{1}$$

- Suite du deuxième ordre

$$\begin{aligned} Prob(s_t = w/s_{t-1} = v, s_{t-2} = u, s_{t-3} = ...) \\ = Prob(s_t = w/s_{t-1} = v, s_{t-2} = u) \end{aligned} \tag{2}$$

Une chaîne est dite homogène si les probabilités de transition ne dépendent pas du temps. On peut alors poser : $a_{uv} = Prob(s_t = v/s_{t-1} = u)$ et $a_{uvw} = Prob(s_t = w/s_{t-1} = v, s_{t-2} = u)$.

Si l'ensemble des états $\mathbf{S}$ est fini et possède N états alors on dit que la chaîne est finie.

- Chaînes du premier ordre

$$\sum_{k=1}^{N} a_{jk} = 1 \quad , \; 1 \leq j \leq N$$

- Chaînes du second ordre

$$\sum_{k=1}^{N} a_{ijk} = 1 \quad avec \quad (i, j) \in \Omega$$

2 Problème de l'escadrille

Une escadrille de 4 avions effectue des missions au dessus du territoire ennemi. Elle y subit des pertes. Soit p la probabilité qu'un avion disparaisse. Elle ne part en misssion qu'avec 3 ou 4 avions sinon elle est mise au repos pendant que l'état major lui fournit d'autres avions. Chaque soir, s'il y a 0, 1 ou 2 avions valides, elle reçoit dans la nuit un avion en remplacement.

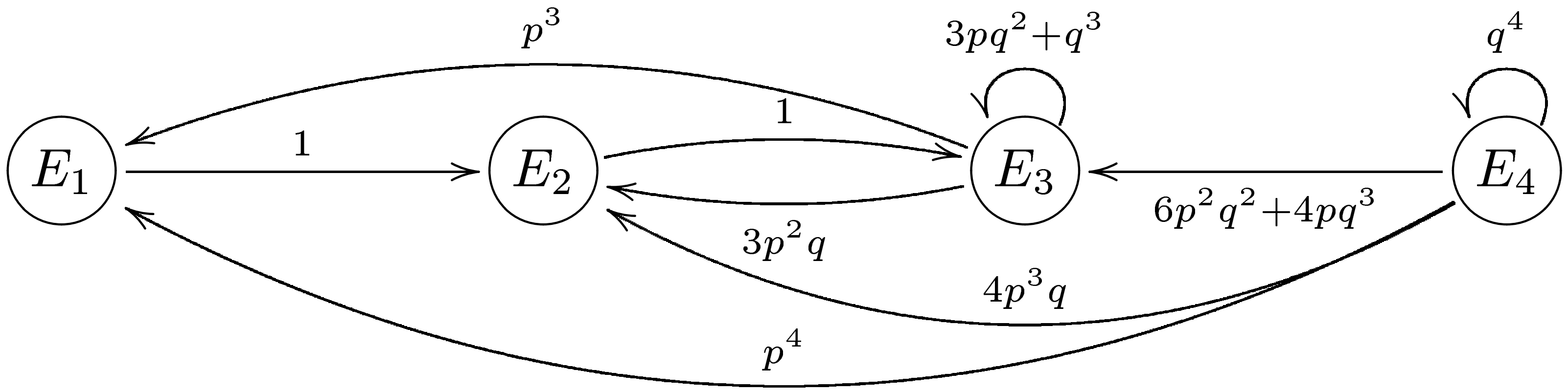


Figure 1: Graphe des états de l'escadrille

2.1 Notations et définitions

- s_t la suite de Markov finie d'ordre 1 ou 2, homogène, appelée *suite cachée*. Ce processus stochastique est entièrement défini par la matrice des probabilités de transition entre les états u et v : (a_{uv}) HMM1 ; (a_{uvw}) HMM2
- y_t le processus stochastique visible. y_t prend ses valeurs dans R^d . La densité de chaque y_t est définie par la donnée des densités $b_i()$ représentant chacun des états s_i du processus de Markov caché. On modélise $b_i()$ par une somme pondérée de M lois normales $\mathcal{N}(\mu_{im}, \Sigma_{im})$ de moyenne μ_{im} et de matrice de covariances Σ_{im}.

$$b_i(y_t) = \sum_{m=1}^{M} c_{im}\mathcal{N}(y_t; \mu_{im}, \Sigma_{im}), \quad \text{avec} \quad \sum_{m=1}^{M} c_{im} = 1$$

Un modèle de Markov caché du deuxième ordre est défini par la donnée de :

- $\mathbf{S} = \{u_1, u_2, \ldots, u_N\}$, un ensemble fini de N états ;

- $\mathbf{A} = (a_{ijk})$ la matrice donnant les probabilités de transition entre états ;
- π le vecteur de probabilités initiales de la suite stochastique cachée ;
- $b_i(.)$ les lois des densités, souvent appelées abusivement *pdf*, associées aux états s_i du processus stochastique caché.

L'ensemble de ces paramètres $\theta = \{S, A, \pi, b_i(.)\}$ constitue le modèle de Markov caché θ du deuxième ordre.

2.2 Probabilité d'un alignement temporel

Un alignement temporel entre une suite d'états et une suite de trames est une correspondance (état, trame). Sa probabilité est définie par :

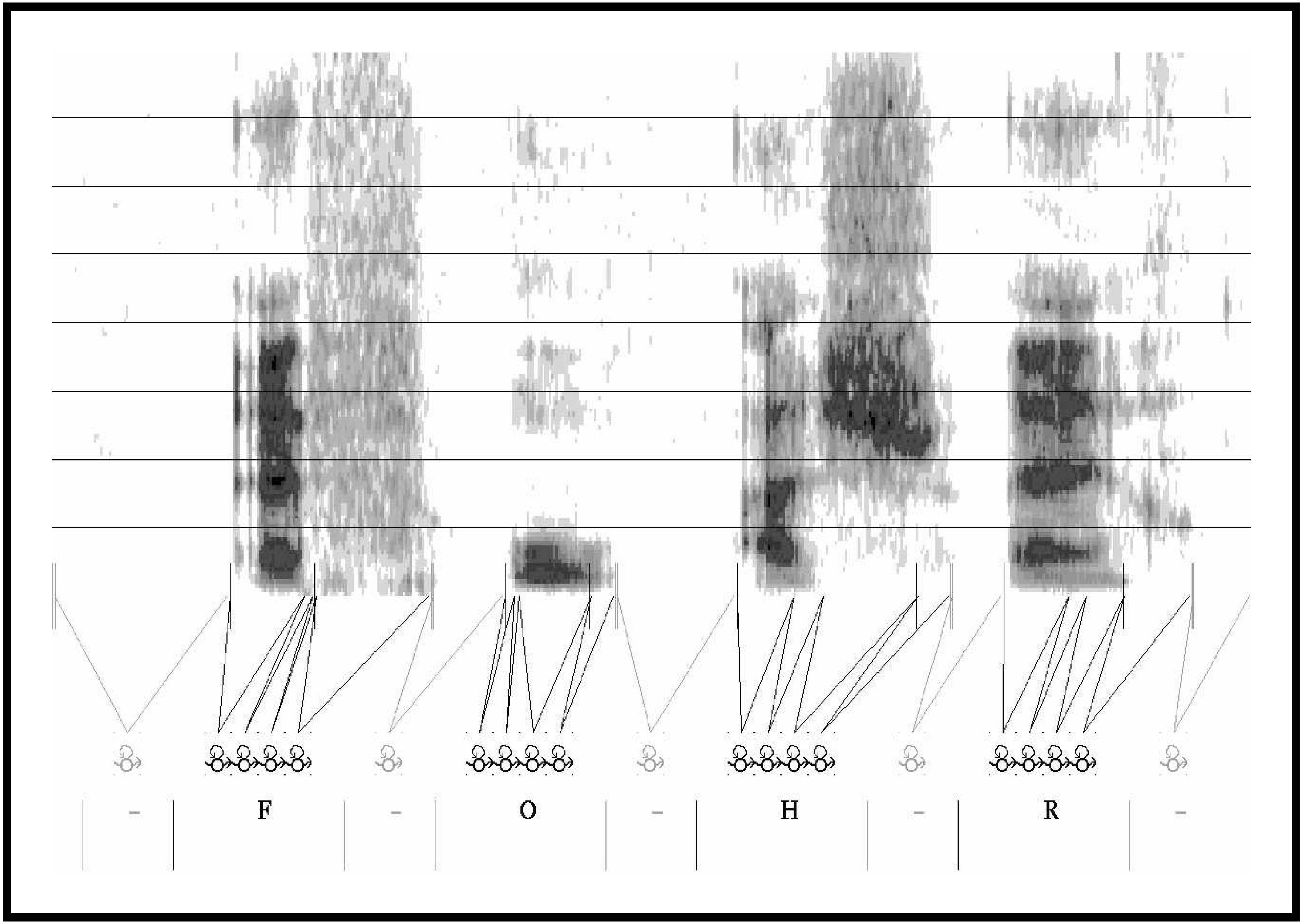
$$Prob(s, y) = \prod_t a_{s_{t-2}s_{t-1}s_t} b_{s_t}(y_t) \tag{3}$$

$a_{s_{t-2}s_{t-1}s_t}$ représente la probabilité de la contribution de la chaîne de Markov entre les instants t-2 et t (cas d'une chaîne d'ordre 2) qui définit la segmentation.

$b_{s_t}(y_t)$ il représente la vraisemblance de la trame y_t conditionnée à la densité de l'état s_t.

Dans le cas des modèles de Markov gauche-droite décrits dans la figure 8, un alignement entre états et trames définit une segmentation.

Un grand intérêt des modèles de Markov cachés est qu'ils permettent lors de la reconnaissance une recherche de tous les alignements afin de retrouver la segmentation possédant la probabilité maximale. L'algorithme utilisé est *l'algorithme de Viterbi* qui est un algorithme de programmation dynamique.



–
F
–
O
–
H
–
R
–

2.3 The Viterbi algorithm

The Viterbi algorithm plays an important role in stochastic modeling. Its purpose is to extract the best states sequence that produces the observed sequence. It is an iterative algorithm, linear in t and polynomial in N, depending on the order of the model. Its first description was given by D. Forney [**?**].

Consider a 3 state HMM (cf. Fig 2) in which the *pdf* are discrete. On state $i,\ i = 1, 3$, we can observe 2 symbols: a and b.
The matrices that define the HMM_1 λ are:

$$A = \begin{pmatrix} 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \\ 0 & 0 & 1 \end{pmatrix}$$

that defines the transition probabilities between the 3 states, and

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

in which a row represents a *pdf*. The initial state probabilities at time 0 are given by the vector:

$$\pi = \begin{pmatrix} 0.6 \\ 0.4 \\ 0 \end{pmatrix}$$

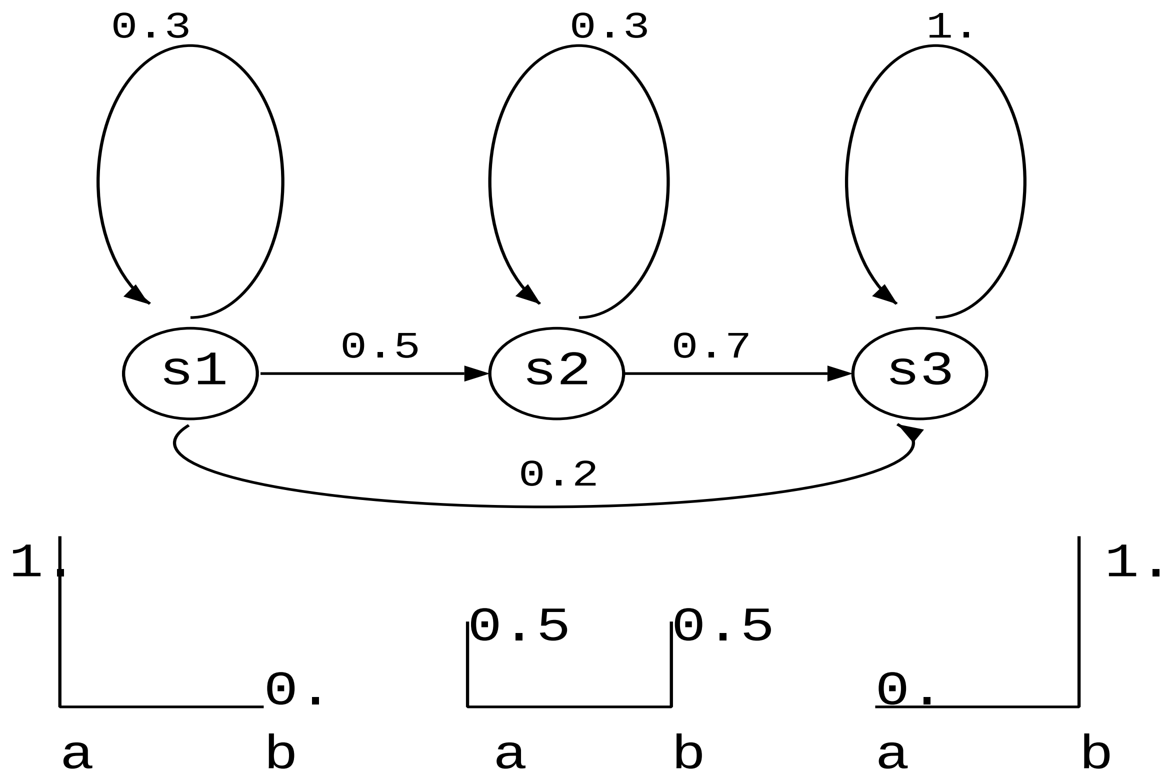


Figure 2: A simple first order HMM

Given a sequence of T observations $y = (y_1, y_2, \ldots, y_T)$, the algorithm extracts the best path

$$S = (s_1, s_2, \ldots, s_T)$$

by computing the best partial path from $t = 1$ up to time t and ending in state v for each state v. We define $\delta_t(v)$ as the probability of this best partial path.

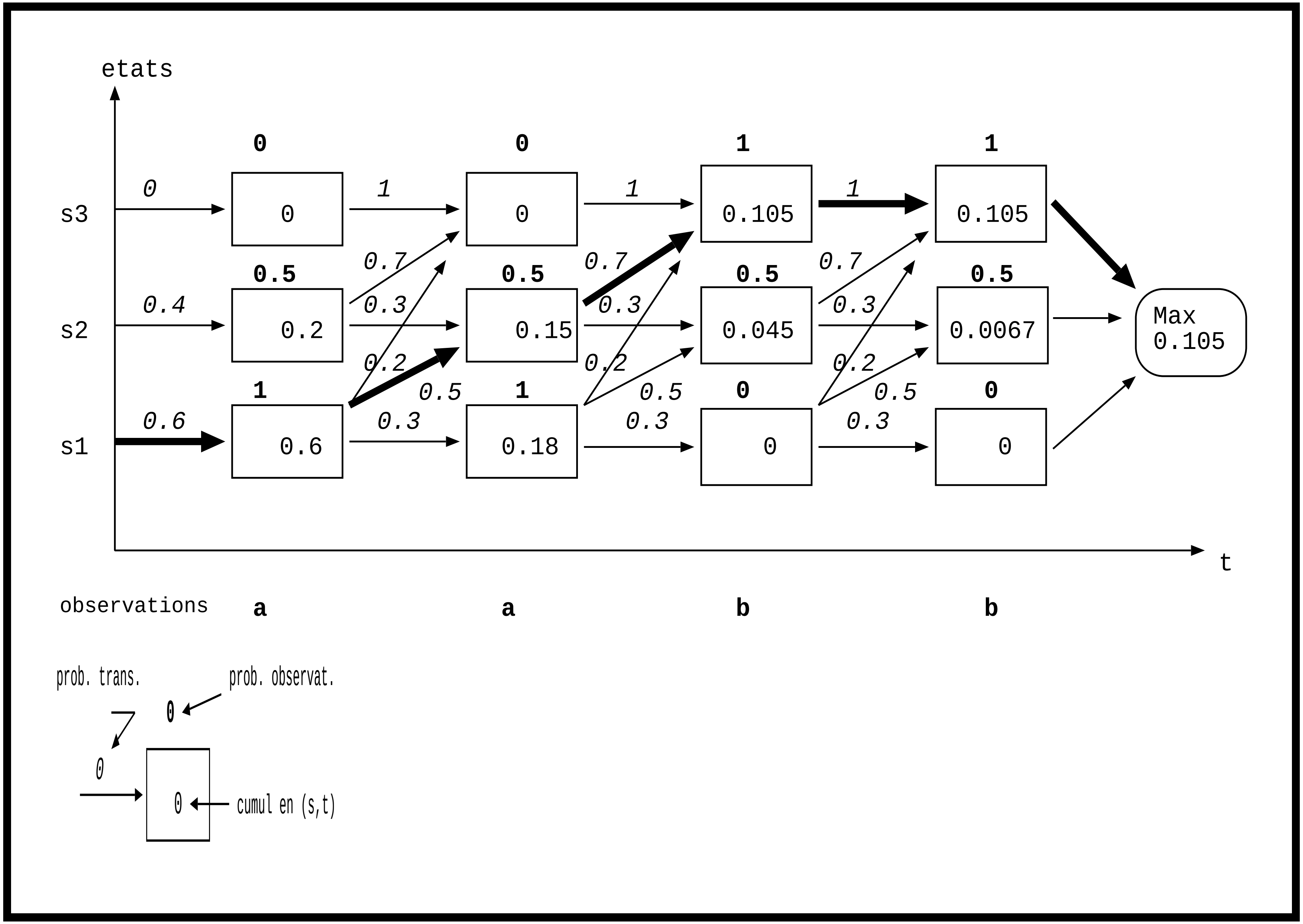
$$\delta_t(v) = \max_{s_t}[P(s_1, ..., s_{t-2}, s_{t-1}, s_t = v, y_1, ..., y_t|\lambda)], \quad \begin{matrix} 1 & \leq v & \leq N \\ 1 & \leq t & \leq T \end{matrix}$$

A recursive computation is given by the following equation:

$$\delta_t(v) = \max_{1 \leq u \leq N}[\delta_{t-1}(u) \cdot a_{uv}] \cdot b_v(y_t), \quad \begin{matrix} 1 & \leq v & \leq N \\ 1 & \leq t & \leq T \end{matrix} \tag{4}$$

The initial condition is given by:

$$\delta_0(i) \quad = \quad \Pi_i, \qquad 1 \leq i \leq N$$


etats
s3
s2
s1
0
0.5
1
0
0.2
0.6
0.15
0.18
0.105
0.045
0.0067
Max
0.105
t
observations
a
a
b
b
prob. trans.
prob. observat.
cumul en (s,t)

Second-order HMM (HMM2) Unlike the first-order Markov chain, where the stochastic process is specified by a 2-dimensional matrix of *a priori* transition probabilities a_{ij} between states s_i and s_j, the second-order Markov chain is specified by a 3 dimensional matrix a_{ijk}.

$$\begin{aligned} a_{ijk} \stackrel{\text{def}}{=} P(q_t = s_k | q_{t-1} = s_j, q_{t-2} = s_i, q_{t-3} = ...) = \\ P(q_t = s_k | q_{t-1} = s_j, q_{t-2} = s_i) \end{aligned} \tag{5}$$

with the constraints:

$$\sum_{k=1}^{N} a_{ijk} = 1 \quad \text{with } 1 \leq i \leq N \ , \ 1 \leq j \leq N$$

where N is the number of states in the model and q_t is the actual state at time t.

The probability of the state sequence $Q \stackrel{\text{def}}{=} q_1, q_2, ..., q_T$ is defined as:

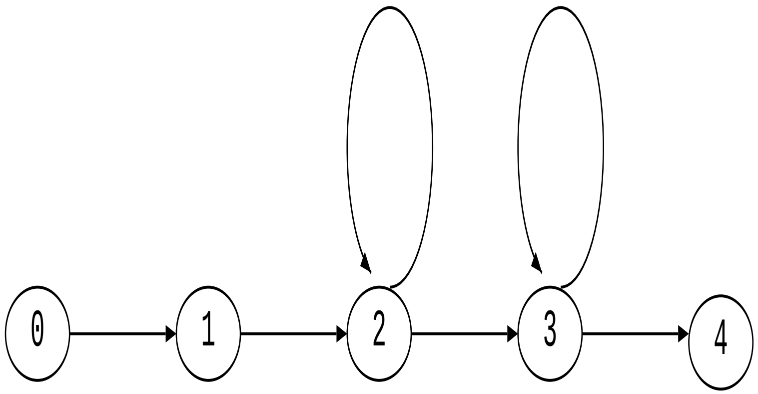
$$P(Q) = \Pi_{q_1} a_{q_1 q_2} \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} \tag{6}$$

where Π_i is the probability of state s_i at time t = 1 and a_{ij} is the probability of the transition $s_i \to s_j$ at time t = 2.

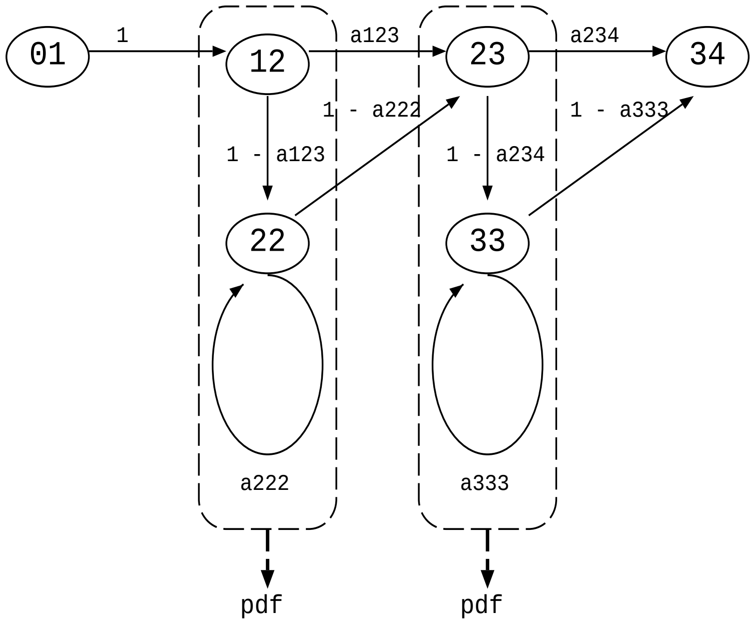
Given a sequence of observed vectors $O \stackrel{\text{def}}{=} O_1, O_2, ..., O_T$, the joint state-output probability $P(Q, O|\lambda)$, is defined as:

$$P(Q, O|\lambda) = \Pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} b_{q_t}(O_t) \tag{7}$$

Each second-order Markov model has an equivalent first-order model on the 2-fold product space **S x S**, but going back to first-order increases the number of states dramatically. For instance, figure 3(b) shows the equivalent HMM_1 associated with the HMM_2 depicted in figure 3(a). In this model the states in the same column share the same *pdf*.



(a) original second-order mode



(b) first-order equivalent model

Figure 3: Decreasing the order of a HMM_2.

Note that HMM_2 converge naturally to Ferguson-like models.

The transition probabilities determine the characteristic of the state duration model. In a HMM_1, whose topology is depicted in the figure 3(a), the probability $d_j(l)$ that the stochastic chain loops l times in the state j follows a geometric law of parameter a_{jj}.

$$d_j(l) = a_{jj}^{l-1} \times (1 - a_{jj})$$

In the model depicted in figure 3(b), in which the successive states are denoted by $i = j - 1,\ j,\ k = j + 1$, the duration in state j may be defined as :

$$\begin{aligned} d_j(0) &= 0 \\ d_j(1) &= a_{ijk}, \quad i \neq j \neq k \\ d_j(n) &= (1 - a_{ijk}) \cdot a_{jjj}^{n-2} \cdot (1 - a_{jjj}), \quad n \geq 2 \end{aligned} \tag{8}$$

The state duration in a HMM_2 is governed by two parameters, *i.e*, the probability of entering a state only once, and the probability of visiting a state at least twice, with the latter modeled as a geometric decay.

On cherche à maximiser la vraisemblance du modèle sur le corpus d'observations : P(observations / modele) que nous noterons $f(y/\theta) = P(Y = y/\theta)$.

$$f(y/\theta) = \sum_{s \in S^n} g(y, s/\theta)$$

avec la probabilité d'un alignement défini comme suit

$$g(y, s/\theta) = \prod_t a_{s_t s_{t+1}} b_{s_{t+1}}(y_{t+1}).$$

La log vraisemblance est définie par

$$L(y/\theta) = \log f(y/\theta).$$

L'algorithme EM construit une suite $\theta_0, \theta_1, \ldots \theta_m \ldots$ telle que $L(\theta_0/y) \leq L(\theta_1/y) \ldots \leq L(\theta_m/y)$ qui converge vers un point stationnaire. Sous certaines conditions de régularité (1983), on obtient un maximum local ou global θ_{MLE}.

$$\Pi(S, Y/\theta = \frac{g(Y, S/\theta)}{f(Y/\theta)} \tag{9}$$

$$Q(\theta/\theta') = E(\log g(Y, S/\theta)/y, \theta')$$

$Q(\theta/\theta')$ représente l'espérance de $\log g(Y, S/\theta)$ à l'aide de la répartition calculée avec y, θ'.

$$H(\theta/\theta') = \sum_{s \in S^n} \log \Pi(S/Y, \theta) \Pi(S/Y, \theta')$$

Comme $\sum_{s \in S^n} \Pi(S/Y, \theta') = 1$, on obtient

$$L(Y/\theta) = \sum_{s \in S^n} \log f(Y/\theta) \Pi(S/Y, \theta') = E(\log f(Y/\theta)/y, \theta')$$

En appliquant la définition de Π donnée dans l'équation **??** on obtient

$$E(\log f(Y/\theta)/y, \theta') = E(\log \frac{g(Y, S/\theta)}{\Pi(s/Y, \theta)}/y, \theta')$$

et donc

$$L(Y/\theta) = Q(\theta/\theta') - H(\theta/\theta').$$

Parmi toutes les valeurs de θ, choississons celle qui donne le maximum de la fonction $Q(\theta/\theta')$; on a donc $Q(\theta/\theta') \geq Q(\theta'/\theta')$. Il nous reste à appliquer l'inégalité de Jansen qui stipule

$$H(\theta/\theta) \geq H(\theta/\theta').$$

pour voir que ce choix de θ donne

$$L(y/\theta) \geq L(y/\theta').$$

Cette propriété de monotonie permet de démontrer la convergence de l'algorithme.

étape E : comme espérance ; on calcule $Q(\theta/\theta_m)$

étape M : comme maximisation ; on cherche la valeur de θ qui maximise $Q(\theta/\theta_m)$.

2.4 Application aux HMM d'ordre 1

On a

$$g(y,s) = \prod_t a_{s_t s_{t+1}} b_{s_{t+1}} y(t+1)$$

On appelle $1_{\{s_t=u, s_{t+1}=v\}}$ la fonction indicatrice qui vaut 1 si et seulement $s_t = u$ et $s_{t+1} = v$.

$$g(y,s) = \prod_{(u,v)\in S^2} a_{uv}^{\sum_t 1_{\{s_t=u,s_{t+1}=v\}}} \prod_t b_{s_{t+1}} y(t+1)$$

$$
\begin{aligned}
Q(\theta/\theta_m) &= E(\log g(Y, S/\theta)/y, \theta_m) \\
&= \sum_{(u,v)\in S^2} E(\sum_t 1_{\{s_t=u, s_{t+1}=v\}}/y, \theta_m) \log a_{uv} + \cdots \qquad (10)
\end{aligned}
$$

L'espérance conditionnelle relativement à θ_m et y de la fonction indicatrice $1_{\{s_t=u, s_{t+1}=v\}}$ vaut

$$E(1_{\{s_t=u, s_{t+1}=v\}}/y, \theta_m) = Prob(s_t = u, s_{t+1} = v/y, \theta_m)$$

Ce qui donne

$$Q(\theta/\theta_m) = \sum_{(u,v)\in S^2} \sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m) \log a_{uv} + \cdots$$

On maximise $Q(\theta/\theta_m)$ en maximisant chacun des termes. Voyons comment maximiser le premier terme sous la contrainte

$$\sum_{v\in S} a_{uv} = 1, \quad \forall u \in S$$

On introduit la quantité (nulle) $\lambda(1 - \sum_v a_{uv})$ appelée, multiplicateur de Lagrange.

$$\begin{aligned} Q(\theta/\theta_m) = &\sum_{u \in S} \sum_{v \in S} \sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m) \times \log a_{uv} \\ &+ \sum_{u \in S} \lambda(1 - \sum_v a_{uv}) + \cdots \end{aligned} \tag{11}$$

Dérivons partiellement suivant a_{uv}, il vient

$$\frac{\partial Q(\theta/\theta_m)}{\partial a_{uv}} = \frac{\sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m)}{a_{uv}} - \lambda = 0$$

On tire la valeur de a_{uv} en fonction de λ et on somme sur tous les v, sachant que l'on a $\sum_v a_{uv} = 1$ on obtient

$$\lambda = \frac{1}{\sum_{v \in S} \sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m)}$$

d'où la valeur de a_{uv} qui maximise $Q(\theta/\theta_m)$

$$a_{uv} = \frac{\sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m)}{\sum_v \sum_t Prob(s_t = u, s_{t+1} = v/y, \theta_m)}$$

On maximise de la même façon les autres termes de $Q(\theta/\theta_m)$.

2.5 Calcul des comptes dans un HMM1

$$\alpha(t,u) = Prob(s_t = u, y1, \ldots y_t)$$

On a la relation de récurrence

$$\alpha(t,v) = \sum_u \alpha(t-1,u) a_{uv} b_v(y_{t+1}) \tag{12}$$

On peut calculer $Prob(y_1, \ldots y_T)$ à l'aide de $\alpha(T,v)$.

$$Prob(y_1, \ldots y_T) = \sum_{v \in S} \alpha(T,v) \tag{13}$$

De même, on définit la fonction $\beta(t,u)$ par

$$\beta(t,u) = Prob(t_{t+1}, \ldots y_T / s_t = u)$$

qui se calcule par récurrence

$$\beta(t,u) = \sum_{v \in S} a_{uv} b_v(y_{t+1} \beta(t+1, v) \tag{14}$$

Par des applications répétées de la formule de Bayes [**?**], on montre que

$$\begin{aligned} Prob(s_t = u, s_{t+1} = v/y) &= \frac{Prob(s_t = u, s_{t+1} = v, y)}{Prob(y_1, \ldots y_T)} \\ &= \frac{\alpha(t,u) a_{uv} b_v(y_t) \beta(t+1, v)}{\sum_{v \in S} \alpha(T, v)} \end{aligned} \tag{15}$$

2.6 Complexité et implémentation

N_{inp} le facteur de branchement d'entrée moyen.

réseau formé de N états

Le calcul de α et β demande $N \times N_{inp} \times T$ opérations.

Calcul de α

à partir du couple v on recherche l'ensemble des transitions (u, v) telle que a_{uv} soit non nul.

$$v \rightarrow (u_0, v, a_{u_0 v}), (u_1, v, a_{u_1 v}), \cdots$$

le calcul de $Prob(s_t = u, s_{t+1} = v/y)$ utilise les mêmes accès.

Calcul de β

$$u \to (u, v_0, a_{uv_0}), (u, v_1, a_{uv_1}), \cdots$$

Cette table possède N entrées et chaque entrée pointe sur une liste de N_{inp} éléments.

Représentation des matrices α et β

La matrice α a une taille maximale de N réels $\times T$. Elle est calculée pendant la phase aller (*Forward* en anglais). Pendant la phase retour, elle ne sert qu'aux calculs des comptes représentés par $Prob(s_t = u, s_{t+1} = v/y)$.

le calcul de β au temps t se fait en utilisant les seules valeurs de β au temps $t+1$.

Représentation des réels

Débordement dans la représentation des probabilités.

Normalisation $\alpha(t,u)$. Le maximum est sauvegardé dans un tableau $scale[t]$.

$\alpha'(t,u) = \frac{\alpha(t,u)}{scale[t]}$

$\beta'(t,v) = \frac{\beta(t,v)}{scale[t]}$.

$$Prob(s_t = u, s_{t+1} = v/y) = \frac{\alpha'(t,u) a_{uv} b_v(y_t) \beta'(t+1,v)}{\sum_{v \in S} \alpha'(T,v)}$$

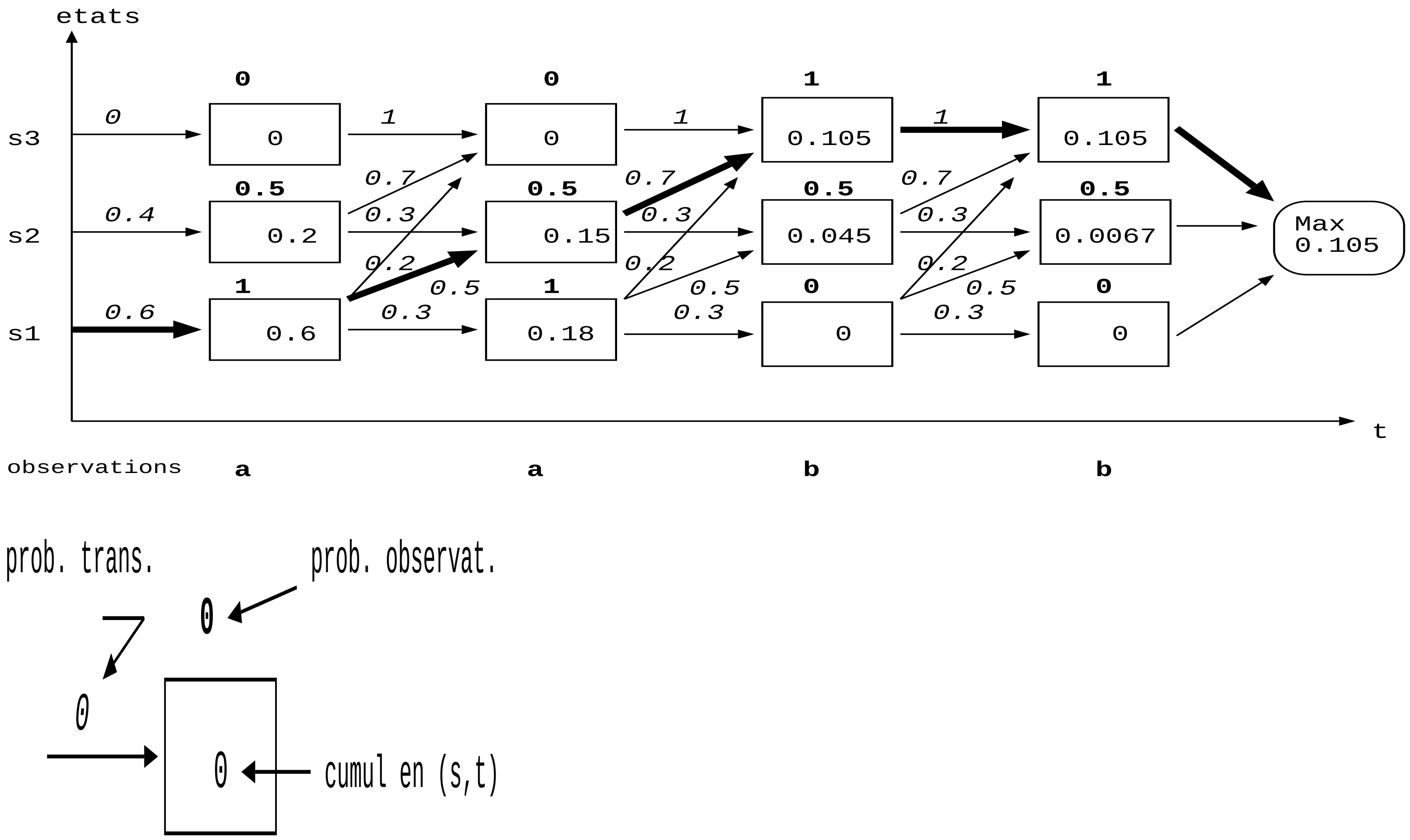


Figure 4: Calcul de δ pour la suite d'observations *aabb*