CHAPITRE 7 : CHAÎNES DE MARKOV ET MODÈLES DE MARKOV CACHÉS

Marc Sigelle

Motivation et Plan

Les chaînes de Markov forment depuis plusieurs dizaines d'années un sujet de choix aussi bien au niveau des investigations mathématiques [**?**] que des applications à base de chaînes de Markov cachées en Reconnaissance des Formes [**?**] : Traitement de la Parole, Traitement de l'Écriture, modèles de langages etc.. Au niveau théorique on considère souvent des modèles à temps discret ou continu (dans ce cas on parle de processus stochastiques de Markov) avec des espaces d'états finis ou infinis aussi bien discrets que continus. Nous nous limiterons ici au cas du **temps discret** et à un espace d'états **fini.**
Nous avons vu au Chapitre 2 que la reconnaissance consiste à comparer un échantillon de test (par exemple en effectuant une programmation dynamique) :

- avec un ensemble de référence
- par une "distance" appropriée.

Ce processus devient coûteux lorsqu'il y a beaucoup de références, par exemple dans les grandes bases de données actuelles. De plus il s'agit d'estimer ou de modéliser les distances élémentaires, ce qui n'est pas trivial. On trouve donc ici l'intérêt d'une l'approche stochastique, où :

- un "modèle " remplace l'ensemble de références.
- les probabilités sont calculées par apprentissage (ce qui remplace les distances).

Le plan de ce chapitre sera le suivant :

1) introduction : processus stochastique(s)
2) chaînes de Markov - exemples
3) chaînes de Markov cachées (HMMs) - exemples
4) apprentissage avec les HMMs - exemples

1 Modèle stochastique

Un processus aléatoire est lié à une variable d'état pouvant changer de valeur au hasard aux instants $t = 1, 2 \dots T$. On définit alors :

- la variable aléatoire (v.a.) associée : $X(t) =$ état observé au temps t, notée aussi Q_t .
- avec les notations, qui sont équivalentes : $X(t) = i$ ou $Q_t = q_t$.
- l'évolution du système est donc décrite par une suite de transitions depuis l'état initial q_1.
- $\Rightarrow$ problème : connaître les chaînes de transition $q_1 \to q_2 \to \dots q_t \quad \forall t \leq T$.

On peut pour cela calculer la loi d'évolution du système *ie.* la probabilité jointe d'une séquence d'états en utilisant la formule générale et fondamentale [1] : $P(A,\ B) = P(B\ /\ A)\ P(A)$

$$P(q_1,\ q_2,\ \dots q_T) = P(q_T\ /\ q_1 \dots q_{T-1})\ \underbrace{P(q_1 \dots q_{T-1})}$$

$$\text{''} \qquad\qquad P(q_{T-1}\ /\ q_1 \dots q_{T-2})\ P(q_1 \dots q_{T-2})$$

$$= \quad P(q_1)\ P(q_2\ /\ q_1)\ P(q_3\ /\ q_1\ q_2) \dots P(q_T\ /\ q_1 \dots q_{T-1}) \qquad (1)$$

En résumé pour calculer cette loi jointe $P(q_1,\ q_2,\ \dots q_T)$ il faut donc connaître :

- la probabilité initiale $P(q_1)$.
- les probabilités conditionnelles $P(q_t\ /\ q_1 \dots q_{t-1}) \quad \forall t = 2 \dots T$.

2 Chaîne de Markov à temps discret et espace d'états fini

On considère un espace d'états $\{1, 2 \dots M\}$ ensemble fini.

- la propriété de Markov d'ordre k dit une dépendance limitée aux k instants précédents :

$$P(q_t\ /\ q_1 \dots q_{t-1}) = P(q_t\ /\ \underline{q_{t-k} \dots q_{t-1}})$$

- en général l'ordre d'une chaîne de Markov est 1 ou 2. Dans le cas usuel $k = 1$ on a :

$$P(q_t\ /\ q_1 \dots q_{t-1}) = P(q_t\ /\ q_{t-1}) \quad \forall t$$

$$\Rightarrow P(q_1,\ q_2,\ \dots q_T) \quad = \quad P(q_1)\ P(q_2\ /\ q_1)\ P(q_3\ /\ q_2) \dots P(q_T\ /\ q_{T-1}) \quad \forall T \qquad (2)$$

[1] nous omettrons souvent dans la suite la variable aléatoire dans la probabilité que celle-ci ait une valeur donnée, par exemple : $P(Q_t = q_t) \to P(q_t)$.

2.1 chaîne de Markov (d'ordre 1) stationnaire

Le système est décrit par des probabilités de transition $i \to j$ qui ne dépendent pas du temps :

$$P(Q_t = j \ / \ Q_{t-1} = i) = P(Q_{t+k} = j \ / \ Q_{t+k-1} = i) = a_{ij}$$

On définit alors :

$A = [\ a_{ij}\ ]$	matrice des probabilités de transition	$M \times M$
$\Pi = [\ \pi_i\ ]$	probabilités initiales (d'avoir une valeur d'état donnée)	M
$\pi_i = P(Q_1 = i)$		

Les formules suivantes ont bien sur lieu (normalisation) :

$$\forall i \in \{1 \dots M\} \quad 0 \leq \pi_i \leq 1 \quad \text{et} \ \sum_{i=1}^{M} \pi_i = 1$$

$$\forall i,\ j \in \{1 \dots M\} \quad 0 \leq a_{ij} \leq 1 \quad \text{et} \ \sum_{j=1}^{M} a_{ij} = 1$$

2.2 cas d'une station météo, étude de l'évolution du temps [?]

3 états : 1 = pluie, 2 = nuages , 3 = soleil . Le modèle est représenté par les matrices A et π :

$$A = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix} \quad \text{on observe 3 = soleil à } t = 1 = \text{lundi}$$

$\to$ quelle est la probabilité P que le temps au cours du reste de la semaine soit

soleil	soleil	soleil	pluie	pluie	soleil	nuages	?
3	3	3	1	1	3	2	

$$P(Q_1 = 3,\ Q_2 = 3,\ Q_3 = 3,\ Q_4 = 1,\ Q_5 = 1,\ Q_6 = 3,\ Q_7 = 2 \ / \ \text{modèle})$$

$$= \Pi_3 \ P(Q_2 = 3 \ / \ Q_1 = 3) \ P(Q_3 = 3 \ / \ Q_2 = 3) \ P(Q_4 = 1 \ / \ Q_3 = 3) \ \times$$

$$P(Q_5 = 1 \ / \ Q_4 = 1) \ P(Q_6 = 3 \ / \ Q_5 = 1) \ P(Q_7 = 2 \ / \ Q_6 = 3)$$

$$= 1 \ (a_{33})^2 \ a_{31} \ a_{11} \ a_{13} \ a_{32} = 7.68.10^{-4}$$

2.3 "durée de vie d'un état" [?]

Sachant que le système est dans l'état $i \to$ quelle est la probabilité qu'il y reste la durée d ?

$$P_i(d) = P(Q_1 = i,\ Q_2 = i \ \dots Q_{d-1} = i,\ Q_d = i,\ Q_{d+1} \neq i)$$

$$= P(Q_1 = i) \ P(Q_2 = i \ / \ Q_1 = i) \dots P(Q_d = i \ / \ Q_{d-1} = i) \ P(Q_{d+1} \neq i \ / \ Q_d = i)$$

$$= (a_{ii})^{d-1} \ (1 - a_{ii})$$

On obtient là un modèle exponentiel caractéristique. La durée moyenne de vie de l'état i est :

$$\mathbb{E}[D] = \sum_{d=1}^{+\infty} d \ P_i(d) = \sum_{d=1}^{+\infty} d \ (a_{ii})^{d-1} \ (1 - a_{ii}) = \frac{1}{1 - a_{ii}}$$

3 HMM hidden Markov models - chaînes de Markov cachées

Il s'agit de deux processus stochastiques dont l'un est caché *ie.* :

- une suite de v.a. cachée = suite d'**états** $q_1,\ q_2 \dots q_t, \dots q_T$
- l'autre, observable = suite de **symboles** (observations) produits en chacun de ces états $o_1,\ o_2 \dots o_t, \dots o_T$

La différence fondamentale avec les modèles classiques est que l'on on n'observe pas directement les états mais seulement les symboles produits par ces états. De plus on "code" l'évolution temporelle dans la séquence d'états (cachés), dont l'espace des valeurs possibles est en général de cardinal beaucoup plus faible que celui des observations.

3.1 caractérisation d'un HMM (λ) : deux hypothèses fondamentales

- Indépendance conditionnelle des observations :

$$P(o \ / \ q, \lambda) = P(o_1 \dots o_T \ / \ q_1 \dots q_T, \lambda) = \prod_{t=1}^{T} \underbrace{P(o_t \ / \ q_t, \lambda)}_{b_j(o_t)} \qquad \swarrow (q_t = j)$$

- La loi a priori sur la séquence d'états q est une chaîne de Markov stationnaire d'ordre 1 :

$$P(q \ / \ \lambda) = \underbrace{P(q_1 \ / \ \lambda)}_{\pi_i} \prod_{t=1}^{T-1} \underbrace{P(q_{t+1} \ / \ q_t)}_{a_{ij}} \qquad \swarrow (q_t = i \ , \ q_{t+1} = j)$$

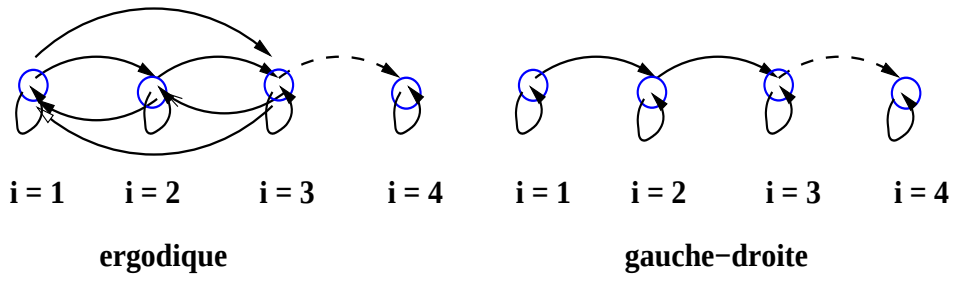
3.2 définition et notations pour un HMM à temps discret

un modèle HMM $\lambda = (A, B, \pi)$ est alors défini par :

S	=	$\{1 \dots M\}$ ensemble des états du modèle M états	
V	=	$\{V_1 \dots V_N\}$ ensemble des N symboles observables dans chaque état	$(N >> M)$
A	=	$[\ a_{ij}\ ]$ matrice des probas de transition (stationnaire)	$M \times M$
Π	=	$[\ \pi_i\ ]$ probas initiales	M
B	=	$[\ b_j(o_t)\ ]$ matrice des probas des symboles émis dans chaque état j	
avec		$b_j(o_t) = P(O_t = o_t \ / \ q_t = j)$ (stationnaire)	$M \times N$

3.3 caractéristiques des modèles HMM

La topologie désigne le graphe des transitions possibles entre états :



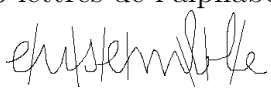
ergodique **gauche−droite**

$a_{ij} > 0 \quad \forall i, j \in \{1..M\}$ $a_{ij} > 0$ si $j = i$ ou $j = i + 1$

3.4 1er exemple : reconnaissance de l'écriture cursive (minuscules)

Les hypothèses sont les suivantes :

- il y a $M = 26$ états cachés correspondant aux 26 lettres de l'alphabet.

  Un mot est supposé déja segmenté en lettres : 

- il y a N "formes" observables, après quantification des observations. Ainsi la forme observée : '***e***' peut provenir de la lettre 'e' ou 'l' : il y a donc des confusions possibles. C'est le rôle du modèle de Markov sous-jacent de lever la confusion par la connaissance des probabilités de transitions permises entre lettres consécutives ou à l'aide d'un dictionnaire. Ainsi ***ensemble*** aura une probabilité faible d'être reconnu comme : 'enslble' .

- les formes dépendent du style et de l'algorithme de reconnaissance.

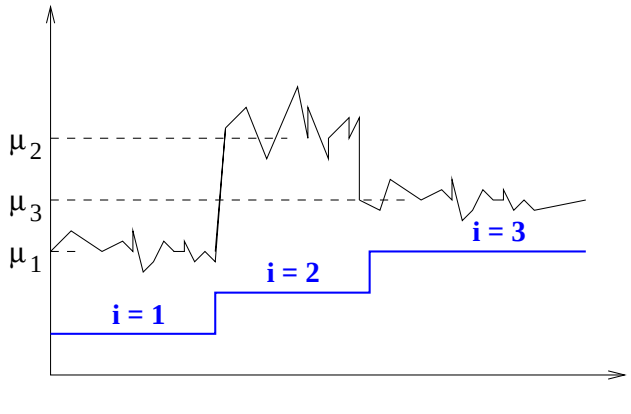
$\Rightarrow$ c'est donc un HMM avec :

- $A = \{a_{ij}\}$ probabilités de transition entre lettres : modèle de langage.
- $\Pi = \{\pi_i\}$ probabilités initiales des lettres.
- $B = \{b_j(k)\}$ probabilités des observations dans chaque état :
  cela dépend du système de reconnaissance OCR $\rightarrow$ apprentissage par comptage.

$\Rightarrow$ **but** : trouver la séquence d'états "optimale" (mot optimal).

3.5 2ème exemple : reconnaissance de la parole

Les observations sont supposées au départ continues et scalaires, pouvant être décrites en première approximation par des modèles gaussiens :



$$P(o_t \ / \ q_t = i, \ \lambda) = \frac{1}{\sqrt{2\pi}\sigma_i} \ \exp - \frac{(o_t - \mu_i)^2}{2{\sigma_i}^2}$$

En fait les descripteurs extraits à partir des observations (depuis des bancs de filtre par exemple) sont souvent vectoriels de dimension d et suivent des lois gaussiennes multi-variées :

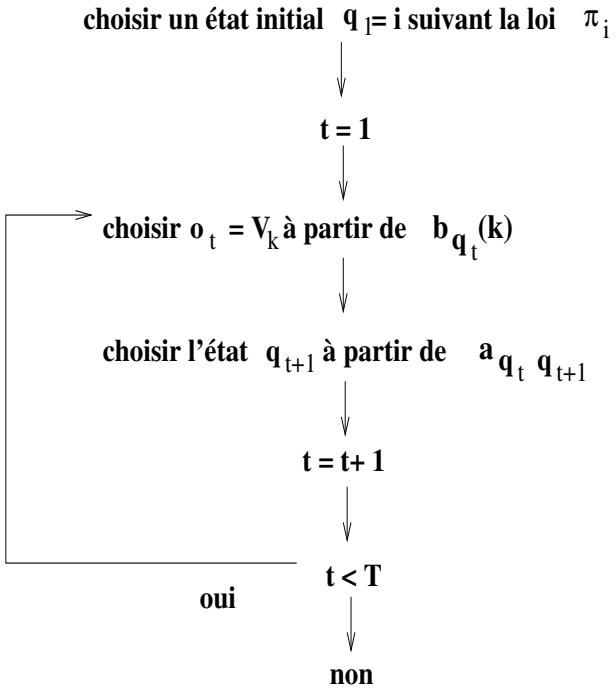
$$P(\bar{o}_t \ / \ q_t = i, \ \lambda) = \frac{1}{(2\pi)^{d/2}} \ \frac{1}{\sqrt{|\mathbf{\Sigma_i}|}} \ \exp - \frac{1}{2} \ (\bar{o}_t - \bar{\mu}_i)^t \ {\mathbf{\Sigma_i}}^{-1} \ (\bar{o}_t - \bar{\mu}_i)$$

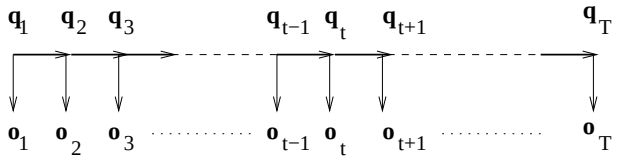
$\bar{\mu}_i$ est la moyenne de l'observation pour l'état i et $\mathbf{\Sigma_i}$ la matrice de variance-covariance associée. Cependant on affine souvent la modélisation en employant des mélanges de gaussiennes :

$$P(o_t \ / \ q_t = i, \ \lambda) = \sum_p w_p \ \mathcal{N}_p(o_t) \rightarrow \ \text{idem pour des lois multi-variées.}$$

3.6 simulation d'une HMM dont le modèle est connu

On veut synthétiser une suite d'observations suivant un modèle connu. Une séquence d'observations $o = o_1 \dots o_t \dots o_T$ est alors produite par :





4 Etapes fondamentales : apprentissage - reconnaissance

a) apprentissage

• on dispose de plusieurs modèles HMM : λ_u, en vue de leur confrontation ultérieure. Exemple : mots isolés en parole (dictionnaire), caractères isolés en écriture (A-Z,0-9).

• le nombre d'états peut dépendre du modèle, et est supposé connu par avance. Il sera toujours noté M, lorsqu'aucune confusion n'est à craindre. Exemple : chaque phonème est de durée plus ou moins longue (et possède donc plus ou moins d'états constitutifs), chaque caractère manuscrit est plus ou moins large. [2]

• on dispose pour chaque modèle d'une base d'apprentissage $\{o^{(l)}\}_{l=1..L}$, de taille L (nombre d'exemples d'apprentissage) pouvant dépendre là aussi du modèle considéré. De plus la durée de chacune des séquences peut également être variable à l'intérieur d'un même modèle.

• apprentissage = estimation des paramètres de chaque modèle λ_u : $\pi_i,\ a_{ij}\ , b_j(o_t)$

b) reconnaissance

• on dispose d'une base de test $\{o^{(r)}\}$: signaux éventuellement "dégradés".

• chaque signal de test est confronté aux divers modèles λ_u appris.

• pour un signal de test $o^{(r)}$ donné → on sélectionne le modèle le plus "vraisemblable".

• on évalue les performances de la reconnaissance globale sur l'ensemble de la base de test.

[2] L'estimation du nombre d'états (optimal) associé un modèle donné est toujours un problème ouvert.

4.1 estimation des paramètres en "données complètes"

La base de tous les raisonnements futurs est l'étude du cas (simple) où une séquence d'observation et la séquence d'états associés sont connues [3] : $(o,\ q)$. La loi **jointe** observations-états s'écrit :

$$\begin{aligned} P(o,\ q\ /\ \lambda) &= P(o\ /\ q,\ \lambda)\ P(q\ /\ \lambda) \\ &= \underbrace{P(q_1\ /\ \lambda)}_{\pi_i}\ \prod_{t=1}^{T-1}\ \underbrace{P(q_{t+1}\ /\ q_t,\ \lambda)}_{a_{ij}}\ \prod_{t=1}^{T}\ \underbrace{P(o_t\ /\ q_t,\ \lambda)}_{b_j(o_t)} \end{aligned}$$

Si on a affaire à plusieurs séquences d'observations indépendantes (*ie.* dans la phase d'apprentissage) avec leurs séquences d'états associés connues $o^{(l)}, q^{(l)}\ , l = 1..L$, alors :

$$P(o,\ q\ /\ \lambda) = \prod_{l=1}^{L} P(o^{(l)}, q^{(l)}\ /\ \lambda) = \prod_{i=1}^{M} {\pi_i}^{N_i^{(1)}}\ \ \prod_{i,j=1}^{M} {a_{ij}}^{N_{ij}}\ \ \prod_{j,k} b_j(o_t = k)^{N_{jk}} \qquad (3)$$

où :

- $N_i^{(1)} = \sum_{l=1}^{L} \mathbb{1}_{q_1^{(l)}=i}$ est le nombre de séquences d'apprentissage dont l'état initial est i .
- $N_{ij} = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} \mathbb{1}_{q_t^{(l)}=i,\ q_{t+1}^{(l)}=j}$ est le nombre de fois dans l'ensemble des séquences d'apprentissage où l'on rencontre le couple d'états (consécutifs) : $q_t^{(l)} = i,\ q_{t+1}^{(l)} = j$.
- $N_{jk} = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=j,\ o_t^{(l)}=k}$ est le nombre de fois dans l'ensemble des séquences d'apprentissage où l'on rencontre le couple simultané état-observation : $q_t^{(l)} = j,\ o_t^{(l)} = k$.

On peut alors estimer les paramètres de la chaîne de Markov sous-jacente au maximum de vraisemblance, *ie.* en maximisant la probabilité jointe $P(o,\ q\ /\ \lambda)$ par rapport aux paramètres du modèle sous contraintes. En effet, ces paramètres étant des probabilités, donc normalisés à 1, l'emploi d'autant de multiplicateurs de Lagrange associés $\{\nu\}$ permet d'obtenir les estimateurs suivants.

- Ainsi, pour les probabilités initiales, définissons le lagrangien suivant :

$$\mathcal{L} = \log P(o,\ q\ /\ \lambda) - \nu \left(\sum_{i=1}^{M} \pi_i - 1 \right)$$

De (**??**) nous déduisons :

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial \pi_i} - \nu \quad = \quad \frac{N_i^{(1)}}{\pi_i} - \nu = 0$$

$$\Rightarrow \frac{N_i^{(1)}}{\pi_i} = \nu = \frac{\sum_{i=1}^{M} N_i^{(1)}}{\sum_{i=1}^{M} \pi_i}\ \text{(proportions !)} = L \ \ \Rightarrow \ \ \widehat{\pi}_i = \frac{N_i^{(1)}}{L} = \frac{\sum_{l=1}^{L} \mathbb{1}_{q_1^{(l)}=i}}{L} \qquad (4)$$

[3] on peut considérer q comme une segmentation, supposée connue, du signal observé o.

On retrouve bien pour l'état initial i la probabilité empirique en tant que quotient du nombre de séquences qui commencent par cet état sur le nombre de séquences d'apprentissage total.

• De même, pour les probabilités de transitions, définissons le lagrangien suivant :

$$\mathcal{L} = \log P(o,\ q\ /\ \lambda) - \sum_{i=1}^{M} \nu_i \left(\sum_{j=1}^{M} a_{ij} - 1 \right)$$

De (**??**) nous déduisons :

$$\frac{\partial \mathcal{L}}{\partial a_{ij}} = \frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial a_{ij}} - \nu_i \quad = \quad \frac{N_{ij}}{a_{ij}} - \nu_i = 0$$

$$\Rightarrow \frac{N_{ij}}{a_{ij}} = \nu_i = \frac{\sum_{j=1}^{M} N_{ij}}{\sum_{j=1}^{M} a_{ij}} \text{ (proportions !)} = N_i^* \ \Rightarrow\ \widehat{a_{ij}} = \frac{N_{ij}}{N_i^*} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} \mathbb{1}_{q_t^{(l)}=i,\ q_{t+1}^{(l)}=j}}{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} \mathbb{1}_{q_t^{(l)}=i}} \quad (5)$$

On trouve le quotient du nombre total de couples d'instants consécutifs où l'on rencontre les états successifs i et j sur le nombre d'instants (excepté l'instant final) où l'on est sur l'état i.

• En ce qui concerne les lois d'observation on pourrait obtenir des formules similaires pour un ensemble d'observations discret fini [4], que nous laissons au lecteur à titre d'exercice. Comme on l'a vu précédemment, les observations sont souvent continues et modélisées, dans le cas le plus simple, par des gaussiennes pures. La composante de la log-vraisemblance dépendant de ces paramètres peut s'écrire assez facilement à l'aide des fonctions indicatrices d'état :

$$LL = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \sum_{i=1}^{M} \mathbb{1}_{q_t^{(l)}=i} \ \left(-\frac{(o_t - \mu_i)^2}{2\sigma_i^2} - \log \sigma_i \right)$$

Il est alors assez facile de montrer en dérivant la log-vraisemblance par rapport aux divers paramètres des lois gaussiennes (moyenne et variance dans chaque état) que l'on on obtient les moyennes et variances empiriques suivantes :

$$\widehat{\mu_i} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \mathbb{1}_{q_t^{(l)}=i}}{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}} \qquad \widehat{(\sigma_i)^2} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \left(o_t^{(l)} - \widehat{\mu_i}\right)^2\ \mathbb{1}_{q_t^{(l)}=i}}{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}} \quad (6)$$

On laisse le soin au lecteur de trouver les formules équivalentes dans le cas des lois gaussiennes multi-variées.

[4] par exemple quantifié à l'aide d'un codebook.

5 Trois problèmes fondamentaux liés aux HMM [?]

Etant donnés : un modèle $\lambda = (A, B, \pi)$ (exemple : mot ou caractère isolé) et une séquence d'observations $o = o_1 \dots o_t \dots o_T$:

1) $\rightarrow$ calculer $P(o \ / \ \lambda)$.

2) $\rightarrow$ trouver une séquence d'états "optimale" a posteriori $\widehat{q} = q_1 \dots q_t \dots q_T = \arg\max_q P(q \ / \ o, \ \lambda)$

3) $\rightarrow$ étant données plusieurs séquences d'observations $o^{(l)}$ (de longueur fixe ou variable) et un modèle λ courant , réestimer les paramètres du modèle $\lambda = (A, B, \pi)$ pour maximiser (en fait augmenter) la "vraisemblance" des observations $P(\dots o^{(l)} \dots \ / \ \lambda)$.

Ce programme en trois points peut être réalisé à partir des variables backward-forward que nous allons maintenant introduire.

5.1 introduction aux variables backward-forward

On va dans ce qui suit utiliser l'aspect "**graphique**" des HMM. On rappelle ce qui connu :

- une séquence d'observations o de durée T.
- un modèle λ

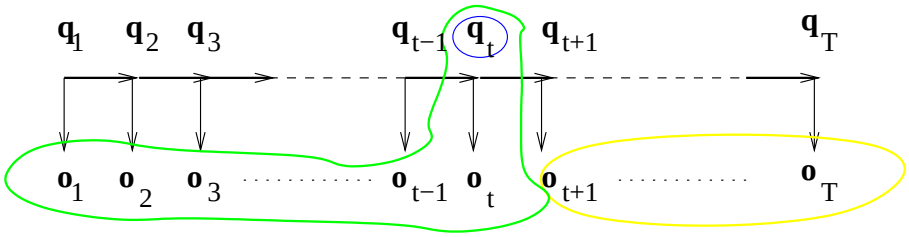


FIG. 1: aspect graphique des HMM

Essayons d'analyser l'expression suivante (Fig. **??**) :

$$P(o_{t+1} \dots o_T \ / \ o_1 \dots o_t, \ q_t)$$

Si nous pensons en termes de simulation (cf. paragraphe **??**), on voit que pour simuler la séquence $(o_{t+1} \dots o_T)$ connaissant $(o_1 \dots o_t, \ q_t)$, il suffit de connaître la valeur de q_t [5] puisque celle-ci permettra de générer les valeurs de q_{t+1}, o_{t+1} etc.. Cette remarque permet alors de définir les expressions suivantes pour chaque instant $1 \leq t \leq T$:

$$P(o_{t+1} \dots o_T \ / \ o_1 \dots o_t, \ q_t) = \boxed{P(o_{t+1} \dots o_T \ / \ \mathbf{q_t}) = \beta_t(i)} \quad \leftarrow \text{variable backward}$$

$$\boxed{P(o_1 \dots o_t, \ \mathbf{q_t}) = \alpha_t(i)} \quad \leftarrow \text{variable forward}$$

[5] qui est le noeud "entrant" vers $(o_{t+1} \dots o_T)$.

5.2 application des variables backward-forward

On a pour un instant donné t :

$$\begin{aligned} P(o,\ q_t = i\ /\ \lambda) &= P(o_1 \dots o_t,\ q_t,\ o_{t+1} \dots o_T) = P(o_{t+1} \dots o_T\ /\ o_1 \dots o_t,\ q_t)\ P(o_1 \dots o_t,\ q_t) \\ &= \underbrace{P(o_{t+1} \dots o_T\ /\ q_t)}_{\beta_t(i)}\ \underbrace{P(o_1 \dots o_t,\ q_t)}_{\alpha_t(i)} \end{aligned}$$

Donc

$$P(o,\ q_t = i\ /\ \lambda) = \beta_t(i)\ \alpha_t(i) \tag{7}$$

$$\Rightarrow P(o\ /\ \lambda) = \sum_{i=1}^{M} P(o,\ q_t = i\ /\ \lambda) = \sum_{i=1}^{M} \beta_t(i)\ \alpha_t(i) \tag{8}$$

$$\Rightarrow P(q_t = i\ /\ o,\ \lambda) = \frac{P(o,\ q_t = i\ /\ \lambda)}{P(o\ /\ \lambda)} = \frac{\alpha_t(i)\ \beta_t(i)}{\sum_{i=1}^{M} \alpha_t(i)\ \beta_t(i)} \tag{9}$$

5.3 formules à deux états

On étudie de la même façon que précédemment les probabilités suivantes faisant intervenir les états associés à deux instants consécutifs :

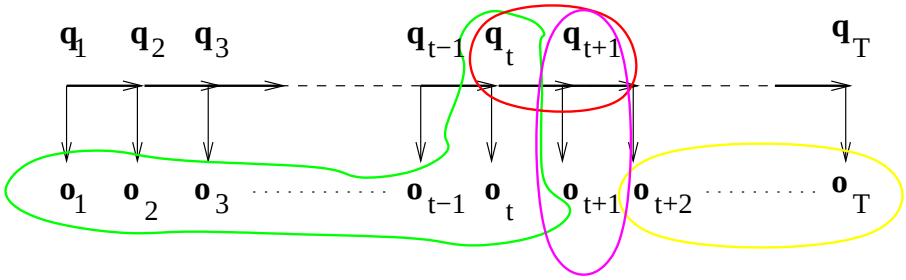


FIG. 2: aspect graphique des HMM : formules à 2 états

Décomposons la probabilité suivante :

$$P(o,\ q_t = i,\ q_{t+1} = j)$$

On a :

$$\begin{aligned} P(o,\ q_t = i,\ q_{t+1} = j) &= P(o_{t+2} \dots o_T\ /\ o_1 \dots o_{t+1},\ q_t,\ q_{t+1})\ P(o_1 \dots o_{t+1},\ q_t,\ q_{t+1}) \\ &= P(o_{t+2} \dots o_T\ /\ q_{t+1})\ P(o_{t+1}\ /\ q_{t+1})\ P(q_{t+1}\ /\ q_t)\ P(o_1 \dots o_t,\ q_t) \\ &= \beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij}\ \alpha_t(i) \end{aligned}$$

Donc

$$P(o,\ q_t = i,\ q_{t+1} = j\ /\ \lambda) \ = \ \beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij}\ \alpha_t(i) \tag{10}$$

$$P(q_t = i,\ q_{t+1} = j\ /\ o, \lambda) \ = \ \frac{P(o,\ q_t = i,\ q_{t+1} = j\ /\ \lambda)}{P(o\ /\ \lambda)} = \frac{\beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij}\ \alpha_t(i)}{\sum_{i=1}^{M} \alpha_t(i)\ \beta_t(i)} \tag{11}$$

• dans (**??**) "sommons" sur toutes les valeurs possibles de j :

$$P(o,\ q_t = i\ /\ \lambda) \ = \ \alpha_t(i)\ \left[\sum_{j=1}^{M} \beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij}\right] = \alpha_t(i)\ \beta_t(i) \quad \text{(d'après (??)) !}$$

$$\Rightarrow \beta_t(i) \ = \ \sum_{j=1}^{M} \beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij}$$

• dans (**??**) on "somme" maintenant sur toutes les valeurs possibles de i :

$$P(o,\ q_{t+1} = j\ /\ \lambda) \ = \ \beta_{t+1}(j)\ b_j(o_{t+1})\ \left[\sum_{i=1}^{M} \alpha_t(i)\ a_{ij}\right] = \alpha_{t+1}(j)\ \beta_{t+1}(j) \quad \text{(d'après (??)) !}$$

$$\Rightarrow \alpha_{t+1}(j) \ = \ b_j(o_{t+1})\ \left[\sum_{i=1}^{M} \alpha_t(i)\ a_{ij}\right]$$

5.4 calcul des variables backward-forward

En récapitulant les résultats précédents on obtient donc les formules de récurrence :

$$\alpha_{t+1}(j) \ = \ b_j(o_{t+1})\ \left[\sum_{i=1}^{M} \alpha_t(i)\ a_{ij}\right] \tag{12}$$

$$\beta_t(i) \ = \ \sum_{j=1}^{M} \beta_{t+1}(j)\ b_j(o_{t+1})\ a_{ij} \tag{13}$$

L'initialisation de ces variables se fait avec les remarques simples suivantes :

$$\alpha_T(i) = P(o, q_T = i) = \alpha_T(i)\ \beta_T(i) \Rightarrow$$

$$\begin{array}{lcl} \beta_T(i) & = & 1 \\ \alpha_1(i) & = & P(o_1,\ q_1 = i) = P(o_1\ /\ q_1 = i)\ P(q_1 = i) = \pi_i\ b_i(o_1) \end{array}$$

6 Estimation des paramètres en "données incomplètes"

Il s'agit du problème le plus difficile de ce chapitre. On ne connaît maintenant que les séquences d'observations (pas les états). Or on veut aussi estimer les paramètres du modèle HMM, et de façon optimale ! Pour utiliser les résultats vus en données complètes, il va falloir effectuer une "statistique" combinatoire sur l'ensemble des séquences d'états cachés possibles (chemins) associés à chaque observation (séquence temporelle). On verra que c'est la loi de l'ensemble des séquences d'états $P(q \ / \ o, \lambda)$ **a posteriori** *ie.* connaissant les séquences d'observations, qui va s'imposer d'un point de vue statistique dans ce qui suit.
Pour toute v.a. U, nous noterons alors son espérance a posteriori par : $\tilde{\mathbb{E}}[U] = \mathbb{E}[U \ / \ o, \lambda]$.
Annonçons maintenant de suite les résultats :

• les paramètres de la chaine de Markov (probabilités) vérifient les équations suivantes :

$$\widehat{\pi_i} = \frac{\tilde{\mathbb{E}}[N_i^{(1)}]}{L} = \frac{\sum_{l=1}^{L} \tilde{\mathbb{E}}[\mathbb{1}_{q_1^{(l)}=i}]}{L} = \frac{\sum_{l=1}^{L} P(q_1^{(l)} = i \ / \ o^{(l)}, \widehat{\lambda})}{L} \tag{14}$$

$$\widehat{a_{ij}} = \frac{\tilde{\mathbb{E}}[N_{ij}]}{\tilde{\mathbb{E}}[N_i{}^*]} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} \tilde{\mathbb{E}}[\mathbb{1}_{q_t^{(l)}=i,\ q_{t+1}^{(l)}=j}]}{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} \tilde{\mathbb{E}}[\ \mathbb{1}_{q_t^{(l)}=i}]} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} P(q_t^{(l)} = i,\ q_{t+1}^{(l)} = j \ / \ o^{(l)}, \widehat{\lambda})}{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}-1} P(q_t^{(l)} = i \ / \ o^{(l)}, \widehat{\lambda})} \tag{15}$$

D'où cela vient-il ? Prenons le cas des probabilités de transition. On a vu que

$$\frac{\partial \log P(o,\ q \ / \ \lambda)}{\partial a_{ij}} = \frac{1}{a_{ij}} \ N_{ij}(q)$$

Ici on indique bien que la quantité N_{ij} dépend de la séquence d'états courante q. Or

$$\begin{aligned} P(o \ / \ \lambda) &= \sum_q P(o,\ q \ / \ \lambda) \\ \Rightarrow \frac{\partial P(o \ / \ \lambda)}{\partial a_{ij}} &= \frac{1}{a_{ij}} \ \sum_q N_{ij}(q) \ P(o,\ q \ / \ \lambda) \\ \Rightarrow \frac{\partial \log P(o \ / \ \lambda)}{\partial a_{ij}} &= \frac{1}{a_{ij}} \ \sum_q N_{ij}(q) \ \frac{P(o,\ q \ / \ \lambda)}{P(o \ / \ \lambda)} = \frac{1}{a_{ij}} \ \tilde{\mathbb{E}}[N_{ij}] \end{aligned}$$

En reprenant les conditions de normalisation des probabilités $\sum_{j=1}^{M} a_{ij} = 1 \ \ \forall i$ sous la forme des multiplicateurs de Lagrange précédents (paragraphe **??**) on arrive facilement à :

$$\widehat{a_{ij}} = \frac{\tilde{\mathbf{E}}[N_{ij}]}{\sum_{j=1}^{M} \tilde{\mathbf{E}}[N_{ij}]} = \frac{\tilde{\mathbf{E}}[N_{ij}]}{\tilde{\mathbf{E}}[N_i{}^*]}$$

On voit (et ceci est très général) que d'un point de vue "mnémotechnique", il suffit de remplacer au numérateur et au dénominateur les quantités déterministes par leurs espérances a posteriori.

• les paramètres des lois d'observations gaussiennes vérifient les équations suivantes :

$$\widehat{\mu_i} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \tilde{\mathbb{E}}[\mathbb{1}_{q_t^{(l)}=i}]}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbb{E}}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})} \tag{16}$$

$$\widehat{(\sigma_i)^2} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} (o_t^{(l)}-\widehat{\mu_i})^2\ \tilde{\mathbb{E}}[\mathbb{1}_{q_t^{(l)}=i}]}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbb{E}}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} (o_t^{(l)}-\widehat{\mu_i})^2\ P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})} \tag{17}$$

On obtient donc un ensemble d'équations (**??**) (**??**) et (**??**) (**??**) auto-cohérentes puisque les paramètres figurent des deux cotés des équations, et non solubles analytiquement. Une méthode itérative extrêmement puissante pour estimer ces paramètres va maintenant être abordée.

7 Algorithme EM

Cet algorithme a pour base les résultats fondamentaux obtenus en données complètes. Pour les lecteurs intéressés, on trouvera un exposé complet de l'algorithme EM dans [**?**], avec en particulier l'application importante à l'estimation de mélanges (de gaussiennes). Son principe général en est le suivant. Chaque étape de l'algorithme comprend deux phases :

- phase E (Expectation) : estimation des statistiques a posteriori courantes $\tilde{\mathbb{E}}^{(\mathbf{n})}[\ U\ ]$
- phase M (Maximisation) : remise à jour des paramètres du modèle suivant le schéma :

$$\lambda^{(\mathbf{n+1})} = \arg\max_{\lambda} \tilde{\mathbb{E}}^{(\mathbf{n})}[\ \log P(q,\ o\ /\ \lambda)\ ] \tag{18}$$

Le résultat fondamental est que la vraisemblance des paramètres croît au cours des itérations. Dans notre cas l'EM va transformer le système d'équations précédentes exactes en un système itératif. Annonçons là aussi de suite les résultats de la phase M (Maximization) :

• en ce qui concerne les paramètres de la chaîne de Markov cachés :

$$\pi_i{}^{(\mathbf{n+1})} = \frac{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_i^{(1)}]}{L} = \frac{\sum_{l=1}^{L} P(q_1^{(l)}=i\ /\ o^{(l)},\lambda^{(\mathbf{n})})}{L} \tag{19}$$

$$a_{ij}{}^{(\mathbf{n+1})} = \frac{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_{ij}]}{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_i{}^*]} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}-1} P(q_t^{(l)}=i,\ q_{t+1}^{(l)}=j\ /\ o^{(l)},\lambda^{(\mathbf{n})})}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}-1} P(q_t^{(l)}=i\ /\ o^{(l)},\lambda^{(\mathbf{n})})} \tag{20}$$

D'où cela vient-il ? Prenons ici encore à titre d'exemple le cas des probabilités de transition.

On forme le lagrangien : $\mathcal{L} = \tilde{\mathbb{E}}^{(\mathbf{n})}[\ \log P(q,\ o\ /\ \lambda)\ ] - \sum_{i=1}^{M} \nu_i \left(\sum_{j=1}^{M} a_{ij} - 1 \right)$.

Il faut bien noter ici que l'espérance a posteriori est prise pour les valeurs de paramètres courants, donc fixés. Comme on sait maintenant presque par coeur (!) que

$$\frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial a_{ij}} = \frac{1}{a_{ij}}\ N_{ij}(q)\ , \quad \text{on en déduit}$$

$$\frac{\partial \mathcal{L}}{\partial a_{ij}} = \tilde{\mathbb{E}}^{(\mathbf{n})}[\ \frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial a_{ij}}\ ] - \nu_i = \frac{1}{a_{ij}}\ \tilde{\mathbb{E}}^{(\mathbf{n})}[N_{ij}] - \nu_i \quad = \quad 0$$

$$\Rightarrow \frac{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_{ij}]}{a_{ij}} = \nu_i = \frac{\sum_{j=1}^{M} \tilde{\mathbb{E}}^{(\mathbf{n})}[N_{ij}]}{\sum_{j=1}^{M} a_{ij}}\ \text{(proportions !)} = \tilde{\mathbb{E}}^{(\mathbf{n})}[N_i^*] \ \Rightarrow\ \widehat{a_{ij}} = \frac{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_{ij}]}{\tilde{\mathbb{E}}^{(\mathbf{n})}[N_i^*]}$$

• pour les paramètres des lois gaussiennes :

$$\mu_i{}^{(\mathbf{n+1})} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \tilde{\mathbb{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbb{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})})}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})})} \tag{21}$$

$$(\sigma_i)^{2(\mathbf{n+1})} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \left(o_t^{(l)} - \mu_i{}^{(\mathbf{n})}\right)^2\ \tilde{\mathbb{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbb{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \left(o_t^{(l)} - \mu_i{}^{(\mathbf{n})}\right)^2\ P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})})}{\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})})} \tag{22}$$

La méthode EM résout donc bien le système (**??** - **??**) sous une forme itérative de type "point-fixe" : $x^{(\mathbf{n+1})} = f(x^{(\mathbf{n})})$ en assurant que la vraisemblance des paramètres augmente à chaque étape. Le calcul des membres de dorite des équations associées nécessite bien de connaître

$$P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})}) \quad \text{et} \quad P(q_t^{(l)} = i,\ q_{t+1}^{(l)} = j\ /\ o^{(l)}, \lambda^{(\mathbf{n})}) \quad \text{(inférence)}$$

à chaque étape, ce qui se fait grâce au calcul des variables backward-forward associées à chaque séquence d'observations $o^{(l)}$ pour les valeurs courantes des paramètres du modèle. Ce calcul étant lui-même mené grâce aux formules de récursion (**??**) et (**??**) . On perçoit bien ici la complexité (au moins computationnelle) de l'ensemble des calculs mis en jeu, bien que les formules associées soient analytiquement exactes [6] dans le cadre de l'algorithme EM et pour le cas des HMMs. Il faut également noter l'importance de l'initialisation des paramètres [7] dans ce type d'algorithme puisque l'optimum atteint est local.

[6] Ceci est complètement différent dans le cadre des Champs de Markov en Traitement d'Images, où la phase d'Estimation de l'EM ne peut en général être menée de façon exacte, et doit se faire par simulation pour les valeurs courantes des paramètres.

[7] Voir remarques dans la conclusion de ce chapitre Section **??**.

8 Recherche de la segmentation optimale

Rappelons ce qui est connu :

- une séquence d'observations o, de durée T.
- un modèle λ.

On cherche à trouver la segmentation optimale, c'est-à-dire une séquence d'états optimale, dans un sens à préciser, associée à la suite d'observations. Nous verrons dans la section suivante l'application pratique de la résolution de ce problème. Deux possibilités s'offrent alors :

a) on peut choisir choisir l'état q_t le plus probable de façon **locale** :
il s'agit de l'estimateur du Maximum de la Marginale a Posteriori (MPM). En effet d'après tout ce que l'on a vu précédemment on peut calculer exactement :

$$P(q_t = i \ / \ o, \ \lambda) = \frac{\alpha_t(i) \ \beta_t(i)}{P(o \ / \ \lambda)} \Rightarrow \boxed{\widehat{q_t} = \arg\max_{i \in E} \alpha_t(i) \ \beta_t(i)}$$

b) on peut choisir un chemin optimal **global** $\widehat{q}$ le plus probable :
il s'agit de l'estimateur du Maximum a Posteriori (MAP). En effet le théorème de Bayes nous permet d'affirmer que :

$$P(q \ / \ o, \ \lambda) = \frac{P(o \ / \ q, \ \lambda) \ P(q \ / \ \lambda)}{P(o \ / \ \lambda)}$$

La segmentation optimale au sens du MAP est donc telle que

$$\boxed{\widehat{q} = \arg\max_q P(q \ / \ o, \ \lambda) = \arg\max_q P(o \ / \ q, \ \lambda) \ P(q \ / \ \lambda) = \arg\max_q P(o, \ q \ / \ \lambda)}$$

$$= \arg\max_q \pi_{q_1} \ b_{q_1}(o_1) \ a_{q_1 q_2} \ b_{q_2}(o_2) \ldots a_{q_{T-1} q_T} \ b_{q_T}(o_T)$$

Le logarithme de la quantité à maximiser

$$\log P(o \ / \ q, \ \lambda) \ P(q \ / \ \lambda)$$

s'écrit donc comme somme de termes :

- à une seule variable $\Phi(q_t) = \log b_{q_t}(o_t)$ associés à l'attache aux données (loi d'observation).
- à deux variables consécutives $\Psi(q_t, q_{t+1}) = \log a_{q_t q_{t+1}}$ associés à la chaîne sous-jacente (probabilités de transition).

$\Rightarrow$ c'est donc ici que la programmation dynamique s'impose comme méthode particulièrement adaptée à la tache de segmentation proposée.

9 Récapitulation pour les HMM

Procédons pour conclure à une récapitulation de l'approche statistique dans la modélisation et l'utilisation des HMMs, en rapport avec les trois problèmes envisagés plus haut Section **??**.

9.1 apprentissage d'un modèle (λ_u)

Rappelons que nous avons décrit la méthode EM qui consiste dans les étapes suivantes :

1) initialisation des paramètres du modèle $[\ \pi_i, a_{ij}, \mu_i, \sigma_i\ ]^{(\mathbf{0})}$:
- en ce qui concerne la chaîne cachée on introduit alors souvent un instant initial "virtuel" $t = 0$ et un état initial "déterministe" $i = 0$ ($\pi_0 = 1$), en imposant :

- soit un modèle ergodique : $a_{01} = \dfrac{1}{M}$ uniforme [8] , π_i et a_{ij} également uniformes.
- ou un modèle gauche-droite : $a_{01} = 1$, de sorte que l'état à $t = 1$ est $i = 1$. Pour les valeurs des états courants on ne retient que les coefficients : a_{ii} et $a_{i\ i+1} = 1 - a_{ii}$.

- en ce qui concerne la loi d'observation supposée gaussienne, on choisit souvent des moyennes μ_i équi-réparties dans un domaine admissible d'observations, et des variances $(\sigma_i)^2$ en conséquence.

2) à l'itération (n) : on utilise les variables backward-forward $\alpha_t(i)$ et $\beta_t(i)$ qui doivent être calculées pour chaque séquence d'apprentissage $o^{(l)}$.

$\Rightarrow$ On a donc résolu ici le problème (3) .

9.2 reconnaissance : observation

On dispose d'une séquence d'obervations $o^{(r)}$ dite "de test". On veut calculer le "score" du modèle λ_u pour cette observation. On procéde pour cela en calculant la vraisemblance du modèle pour la séquence d'obervations de test :
a) soit de façon exacte d'après (**??**) :

$$P(o^{(r)}\ /\ \lambda_u) = \sum_{i \in M_u} \alpha_t(i)\ \beta_t(i) \qquad \text{(avec } t = 1 \text{ ou } T^{(r)} \text{ très souvent)}$$

grâce au calcul encore une fois des variables backward-forward $\alpha_t(i)$ et $\beta_t(i)$ associées à l'observation $o^{(r)}$ et au modèle λ_u.

$\Rightarrow$ On a donc utilisé ici la résolution du problème (1) .
b) soit par l'approximation de Viterbi : on ne retient dans l'expression de la vraisemblance cherchée que la séquence d'états cachés optimale *a posteriori*

$$\begin{aligned} P(o^{(r)}\ /\ \lambda_u) &= \sum_q P(o^{(r)},\ q\ /\ \lambda_u) \approx P(o^{(r)},\ \widehat{q}\ /\ \lambda_u) \\ \text{avec } \widehat{q} &= \arg\max_q P(o^{(r)},\ q\ /\ \lambda) \quad \text{(segmentation optimale MAP)} \end{aligned}$$

On emploie donc pour cela la programmation dynamique vue dans la section précédente, puis l'on calcule ensuite facilement $P(o^{(r)},\ \widehat{q}\ /\ \lambda_u) = P(o^{(r)}\ /\ \widehat{q},\ \lambda_u)\ P(\widehat{q},\ \lambda_u)$
(ou les logarithmes de chacun de ces termes) pour la séquence optimale $\widehat{q}$ ainsi trouvée.

$\Rightarrow$ On a donc utilisé ici la résolution du problème (2) .

[8] ou a_{0i} si on part de l'état i au temps $t = 1$ de façon certaine.