IFT 615 – Intelligence artificielle

Réseaux bayésiens dynamiques

Hugo Larochelle
Département d'informatique
Université de Sherbrooke
http://www.dmi.usherb.ca/~larocheh/cours/ift615.html

Sujets couverts

- C'est quoi un réseau bayésien dynamique (RBD)?
- Exemple d'inférence simple dans un RBD
- Cas particuliers des modèles de Markov cachés

Réseaux bayésiens dynamiques (RBD)

- Comment modéliser des situations dynamiques?
  - les changements dynamiques peuvent être vus **comme une séquence d'états**, chaque état représentant la situation à un instant *t* donné
  - X_t: ensemble des **variables non observables** (**cachées**) décrivant l'état au temps *t*
  - E_t: ensembles de **variables observées** (*evidence*) au temps *t*

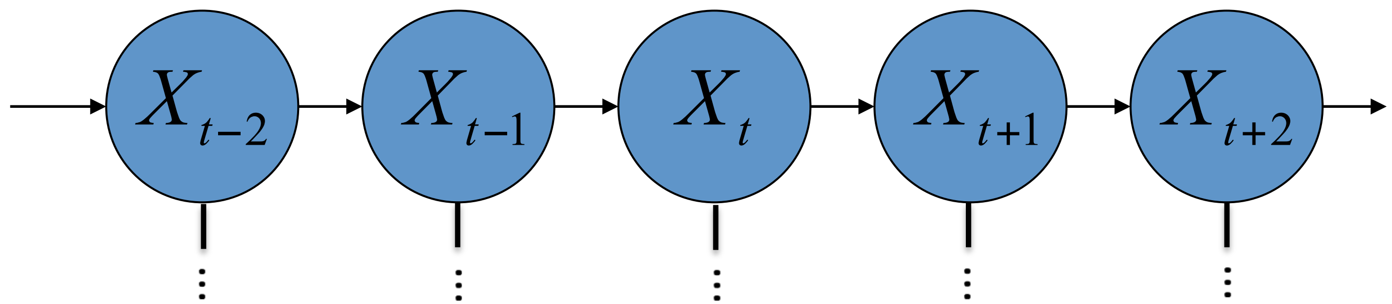
- Le terme dynamique réfère au dynamisme du système qu'on veut modéliser et la structure du réseau

Représentation dans un RBD

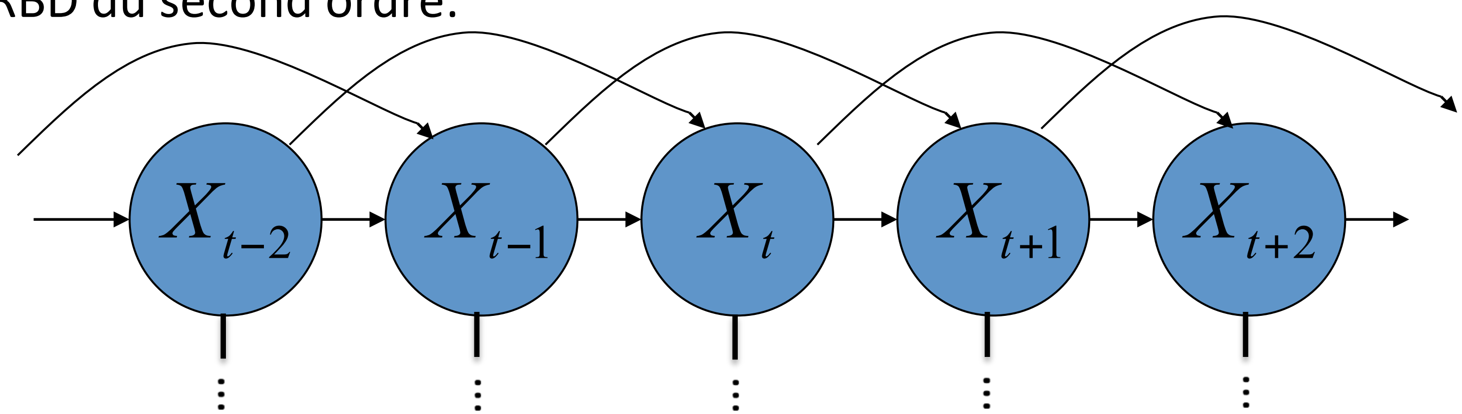
- Problème:
  - il faudrait spécifier un nombre infini de tables de probabilités conditionnelles, c.-à-d. une pour chaque variable, dans chaque état (chaque temps *t*)
  - chaque table pourrait impliquer un nombre infini de parents
- Solution:
  1. supposer que les changements dynamiques sont causés par un **processus stationnaire** - les probabilités ne changent pas dans le temps: $\mathbf{P}(X_t \mid Parent(X_t))$ **est la même** pour tous les *t*
  2. supposer que les changements dynamiques sont causés par un **processus markovien** – l'état courant dépend seulement d'un nombre fini d'états précédents
     - » ex.: processus markoviens **du premier ordre**:
       - – $\mathbf{P}(X_t \mid X_{0:t-1}) = \mathbf{P}(X_t \mid X_{t-1})$ modèle pour les transitions
  3. supposer que l'observation **dépend uniquement de l'état courant**
     - – $\mathbf{P}(E_t \mid X_{0:t}, E_{0:t-1}) = \mathbf{P}(E_t \mid X_t)$ modèle pour les observations/capteurs

Illustration d'un RDB

- Réseau bayesien dynamique (RBD) du premier ordre avec une seule variable *X*, répliquées dans les différents états pour modéliser la dynamique du système:

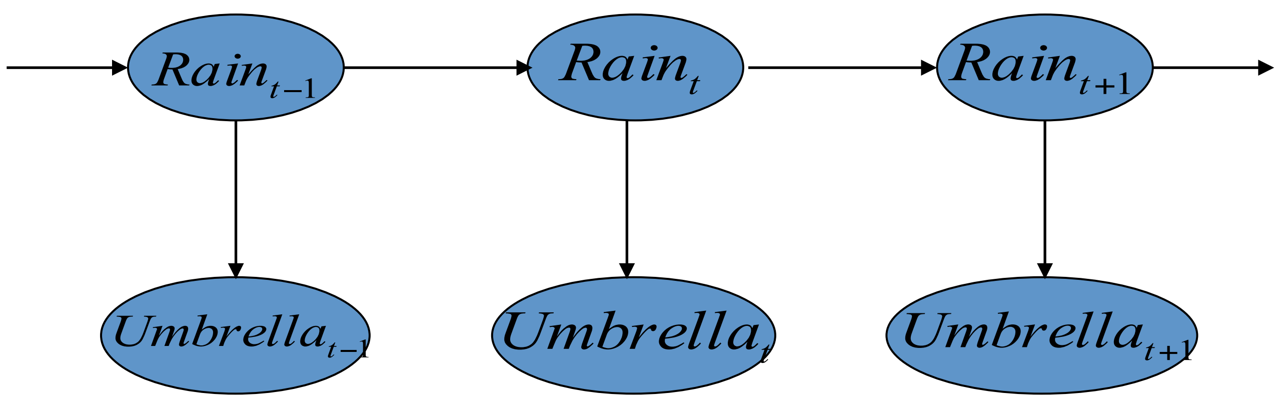


- RBD du second ordre:



Exemple

- « Un gardien de sécurité passe un mois dans un édifice sous-terrain, sans sortir. Chaque jour, son directeur arrive avec ou sans parapluie. Le gardien veut inférer la possibilité qu'il ait plu ou non en fonction des séquences d' observation du parapluie. »
- Modélisation:
  - Variables: $X_t = \{R_t\}$ (pour « *Rain* ») et $E_t = \{U_t\}$ (pour « *Umbrella* »).
  - Dépendances entre les variables (c-.à-d., le RBD):



  - Modèle des transitions: $\mathbf{P}(R_t \mid R_{t-1})$. Modèle d'observation: $\mathbf{P}(U_t \mid R_t)$

RBD

- Comment rendre un RBD plus précis?
  - **augmenter l'ordre du modèle markovien**
    - ex.: $Rain_t$ aurait comme parents, non seulement $Rain_{t-1}$ mais aussi $Rain_{t-2}$ pour un processus markovien du second ordre
    - ceci donnerait des prédictions plus précises
  - **augmenter le nombre de variables d'états**
    - par exemple on pourrait ajouter:
      - une variable $Season_t$ pour tenir compte des statistiques historiques sur les temps de pluie selon les saisons
      - des variables $Temperature_t$, $Humidity_t$ and $Pressure_t$ pour tenir compte de la physique des conditions de pluie
  - **permettre des interactions directes entre la variables d'observation**
    - on pourrait avoir plutôt $\mathbf{P}(E_t \mid X_{0:t}, E_{0:t-1}) = \mathbf{P}(E_t \mid X_t, E_{t-1})$
    - ça peut rendre l'inférence encore plus complexe

Types d'inférence dans un RBD

- **Filtrage** (***filtering***): calcul de l'**état de croyance** (***belief state***), c.-à-d. la distribution a posteriori de la variable cachée la plus récente
  - $\mathbf{P}(X_t | e_{1:t})$
- **Prédiction**: calculer la distribution a posteriori sur un état futur
  - $\mathbf{P}(X_{t+k} | e_{1:t})$ où k > 0
- **Lissage** (***smoothing***): calculer la distribution a posteriori sur un état passé
  - $\mathbf{P}(X_k | e_{1:t})$ où $0 \leq k < t$
- **Explication la plus plausible**: trouver la séquence d'états cachés qui explique le mieux les observations
  - $\underset{x_{1:t}}{\mathrm{argmax}}\ P(x_{1:t} | e_{1:t}) = \underset{x_{1:t}}{\mathrm{argmax}}\ P(x_{1:t}, e_{1:t}) / P(e_{1:t}) = \underset{x_{1:t}}{\mathrm{argmax}}\ P(x_{1:t}, e_{1:t})$
- En plus, on pourrait vouloir faire de l'**apprentissage**: trouver les tables de probabilités conditionnelles telles que **nos données observées sont le plus vraisemblable possible** (à voir plus tard dans le cours)

Filtrage avec RBD

- Calculer l'**état de croyance** (*belief state*) – c-.à-d., la distribution de probabilité a posteriori de l'état courant, étant données les observations jusque là:

$$\mathbf{P}(X_t \mid e_{1:t})$$

- Un agent intelligent a besoin du filtrage pour maintenir à jour son état courant
  - ceci est nécessaire pour prendre des décisions rationnelles (déterminer l'action appropriée étant donné l'état courant)

Filtrage avec RBD

- Étant donnés les résultats du filtrage jusqu'au temps *t*, on peut facilement calculer le filtrage au temps *t+1* à partir des nouvelles observations e_{t+1}
- En appliquant la règle de Bayes et l'hypothèse markovienne, nous arrivons à:

$$\mathbf{P}(X_{t+1} \mid e_{1:t+1}) = \mathbf{P}(X_{t+1} \mid e_{1:t}, e_{t+1})$$

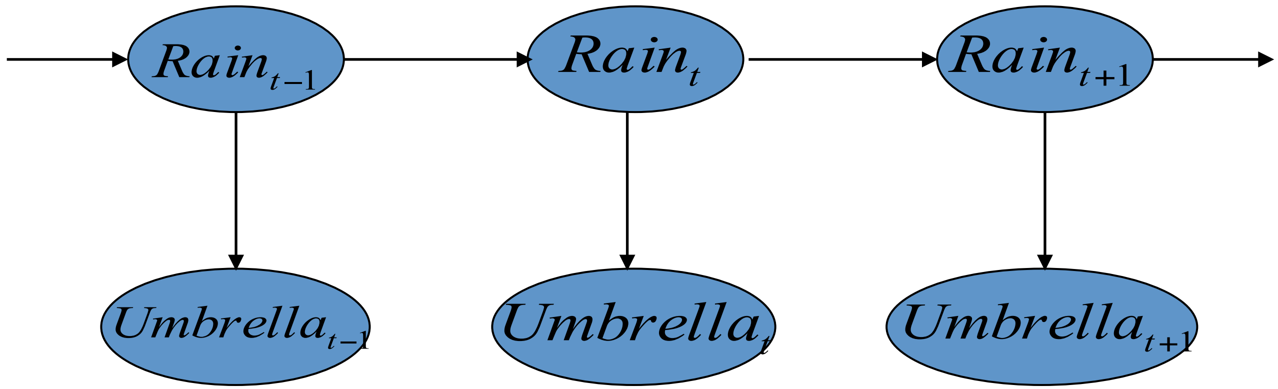
(détails page 572 du manuel de référence)

$$= \alpha \underbrace{P(e_{t+1} \mid X_{t+1})}_{\text{probabilité de la nouvelle observation (disponible dans la table des probabilités)}} \underbrace{\sum_{x_t} \mathbf{P}(X_{t+1} \mid x_t)\, P(x_t \mid e_{1:t})}_{\text{prédiction du prochain état en se basant sur notre état de croyance au temps } t}$$

α: constante de normalisation

Exemple de l'agent de sécurité

- **RBD**:
  - une distribution de **probabilité a priori P**(R_0), par exemple <0.5, 0.5>
  - un **modèle des transition P**($R_t | R_{t-1}$)
  - un **modèle d'observation P**($U_t | R_t$)

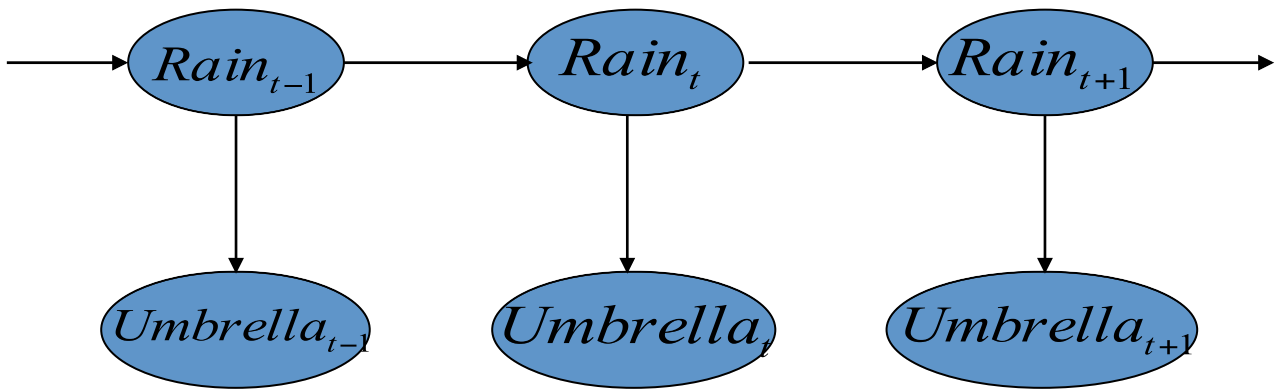


R_{t-1}	P($r_t | R_{t-1}$)
T	0.7
F	0.3

R_t	P($u_t | R_t$)
T	0.9
F	0.2

- **Jour 1**: le parapluie apparait, (U_1=*true* ou u_1)
  - le filtrage de t=0 à t=1 est: **P**(R_1 | u_1) = α $P(u_1 | R_1)$ **P**(R_1)

Exemple de l'agent de sécurité



- **Jour 2**: le parapluie apparait de nouveau, c.-à-d., U_2=*true*
  - le filtrage de t=1 à t=2 est:

$$\mathbf{P}(R_2 \mid u_1, u_2) = \alpha\, P(u_2 \mid R_2)\, \mathbf{P}(R_2 | u_1)$$
$$= \alpha\, P(u_2 \mid R_2) \sum_{r_1} \mathbf{P}(R_2 | \mathrm{r}_1)\, P(r_1 | u_1)$$

Chaînes de Markov

- Une **chaîne de Markov** (de premier ordre) est un cas particulier de RBD avec **une seule** variable aléatoire **discrète** S_t dans l'état au temps t
- Le domaine de S_t est souvent un symbole (ex.: un caractère, un mot, etc.)
- Une **distribution a priori** (initiale) de probabilités sur les symboles (états) est spécifiée $\mathbf{P}(S_1)$
- Une **matrice de transition** contenant les probabilités conditionnelles $\mathbf{P}(S_{t+1} \mid S_t)$

Illustration

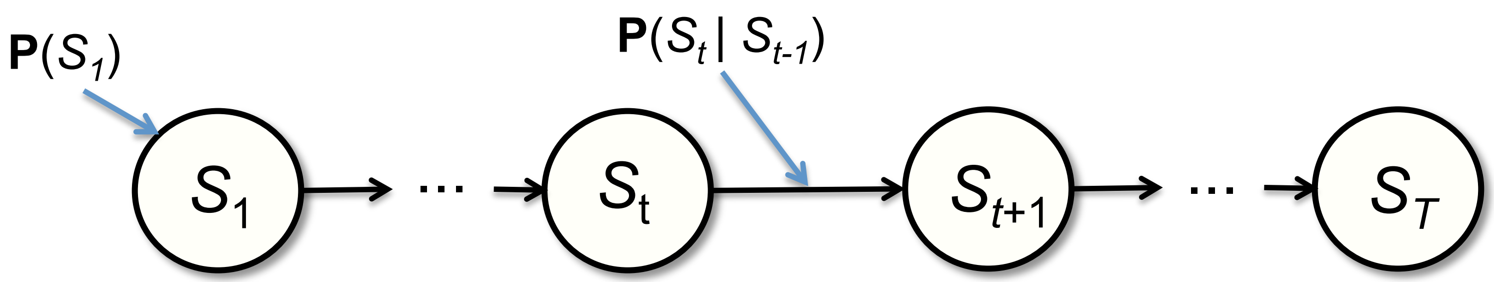


Illustration dans le cas d'une **chaîne finie**

Probabilité de générer une chaîne

produit des probabilités, une pour chaque terme de la séquence

$$\mathbf{P}(S_{1:T}) = \mathbf{P}(S_1) \prod_{t=2}^{T} \mathbf{P}(S_t \mid S_{t-1})$$

une séquence de symboles, allant du temps 1 au temps *T*

distribution initiale de probabilités

distribution de transition

Visualisation d'une chaîne de Markov

Représentation matricielle

Symbole actuel

Prochain symbole

	a	*b*	*c*
a	.7	.3	0
b	.2	.7	.5
c	.1	0	.5

Représentation graphique

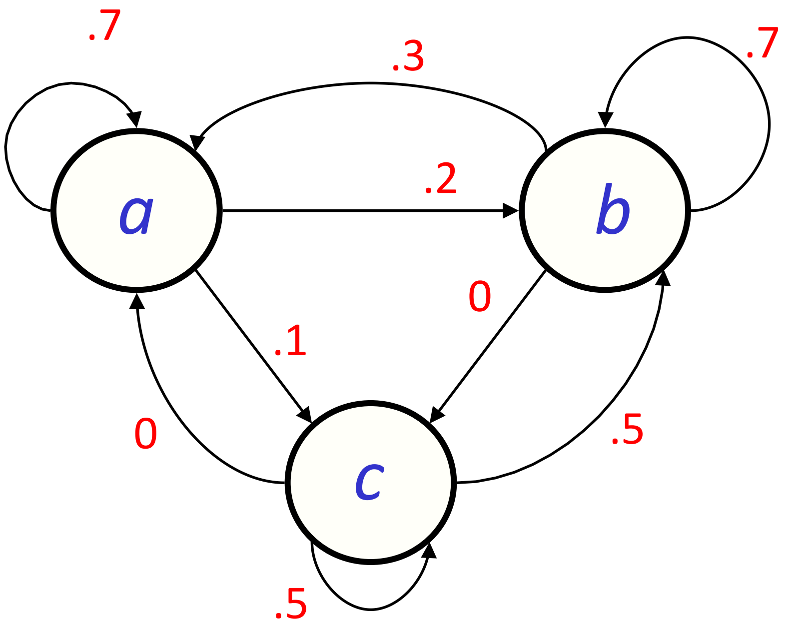


Illustration dans le cas d'une **chaîne infinie** (flux de symboles)

Exemple de chaîne: *ccbbbbaaaaabaabacbabaaa*

Apprendre la table des probabilités conditionnelles

- Observer plusieurs chaînes et définir les probabilités conditionnelles en fonction des fréquences d'occurrence des symboles

$$P(B{=}b\,|\,\mathrm{a}) = \frac{\sum_{\text{chaînes}} \mathrm{freq}(a,b)}{\sum_{\text{chaînes}} \mathrm{freq}(a)}$$

- Pour éviter les problèmes avec zéro occurrences, on utilise plutôt:

$$P(B{=}b\,|\,A{=}a) = \frac{1 + \sum_{\text{chaînes}} \mathrm{freq}(a,b)}{\text{Nb. symboles} + \sum_{\text{chaînes}} \mathrm{freq}(a)}$$

Modèle de Markov caché

- Dans une **modèle de Markov caché** (***hidden Markov model*** ou **HMM**):
  - il y a des **variables cachées** H_t et des **variables d'observation** S_t, toutes les deux discrètes
  - la chaîne de Markov est sur les variables cachées H_t
  - le symbole émis (observé) $S_t=s_t$ dépend uniquement de la variable cachée actuelle H_t

Illustration

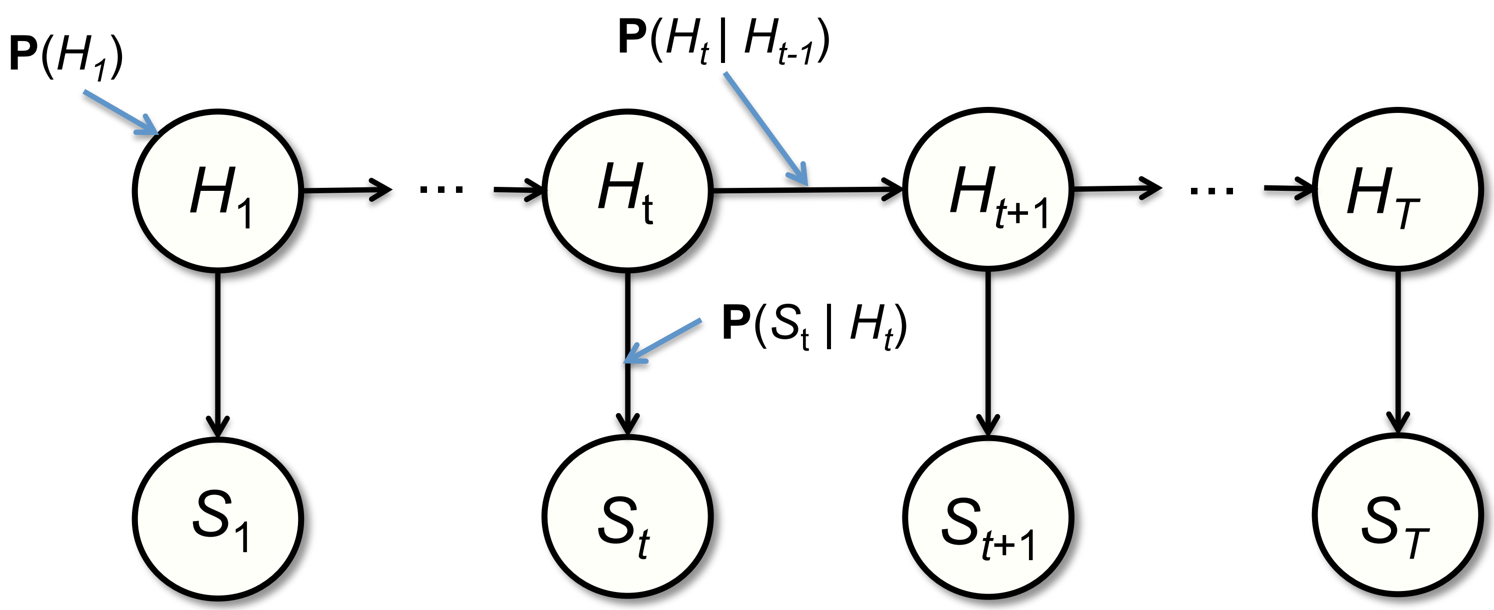


Illustration dans le cas d'une **chaîne finie**

Probabilité de générer une séquence cachée et une séquence visible

distribution initiale de probabilités

distribution de transition

$$\mathbf{P}(S_{1:T}, H_{1:T}) = \mathbf{P}(H_1)\,\mathbf{P}(S_1 | H_1) \prod_{t=2}^{T} \mathbf{P}(H_t \mid H_{t-1})\,\mathbf{P}(S_t \mid H_t)$$

séquence de nœuds cachés et de symboles de sortie

probabilité de sortir le symbole S_t à partir de H_t

Simuler d'un HMM

- Il est facile de générer des observations d'un HMM
  - échantillonner une valeur initiale $H_1 = h_1$ de $\mathbf{P}(H_1)$
  - pour t = 2 jusqu'à T, répéter les deux échantillonnage suivants:
    - » utiliser les probabilités de transition de l'état caché courant pour obtenir un échantillon h_t, sachant l'état caché précédent: $\mathbf{P}(H_t \mid H_{t-1} = h_{t-1})$
    - » utiliser les probabilités de sortie de la variable d'observation étant donné l'état caché courant, pour obtenir le symbole d'observation (émission) s_t: $\mathbf{P}(S_t \mid H_t = h_t)$

- On peut aussi générer la séquence des états cachés d'abord et ensuite générer les observations
  - les variables cachées dépendent uniquement des variables cachées précédentes
  - chaque émission ne dépendra pas des autres émissions

Illustration

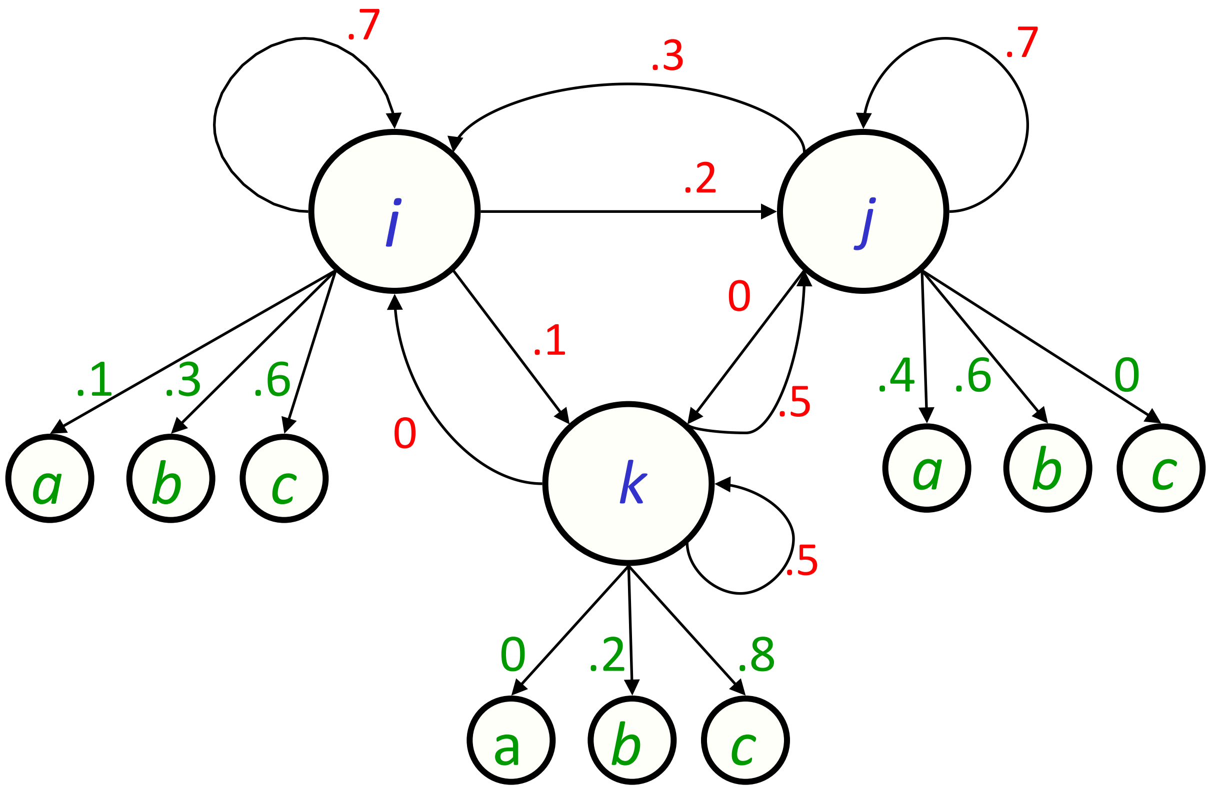


Illustration dans le cas d'une **chaîne infinie**, avec visualisation des valeurs de la variable cachée et la variable d'observation

Chaque **nœud caché** (valeur possible *h* de *H*) a un vecteur de **probabilités de transitions** et un **vecteur de probabilités de sorties (observations)**

Probabilité de générer une séquence visible

- La même séquence de sortie peut être produite par plusieurs séquences cachées différentes
- En fait, il y a un nombre exponentiel de séquences cachées possibles
- Un calcul naïf est donc très inefficace

$$\mathbf{P}(S_{1:T}) = \sum_{h_{1:T}} P(H_{1:T} = h_{1:T})\, \mathbf{P}(S_{1:T} \mid H_{1:T} = h_{1:T})$$

Programmation dynamique pour HMM

- Une façon plus efficace de calculer la probabilité d'une séquence observée $s_{1:T}$
- Idée: utiliser la **programmation dynamique**
  - on définit $\alpha(i,t) = P(S_{1:t}=s_{1:t}, H_t = i)$
  - on note la récursion
    $\alpha(i,t+1) = P(S_{1:t+1}=s_{1:t+1}, H_{t+1} = i)$
    $= \sum_j P(S_{1:t+1}=s_{1:t+1}, H_t = j, H_{t+1} = i)$
    $= P(S_{t+1}= s_{t+1} | H_{t+1} = i) \sum_j P(H_{t+1} = i | H_t = j) P(S_{1:t}=s_{1:t}, H_t = j)$
    $= P(S_{t+1}= s_{t+1} | H_{t+1} = i) \sum_j P(H_{t+1} = i | H_t = j) \alpha(j,t)$
  - on a les valeurs initiales
    $\alpha(i,1) = P(S_1=s_1, H_1 = i) = P(S_1=s_1 | H_1 = i) P(H_1 = i) \quad \forall i$
- Une fois le tableau α calculé, on obtient facilement:
  $P(S_{1:T}=s_{1:T}) = \sum_j P(S_{1:T}=s_{1:T}, H_T = j) = \sum_j \alpha(j,T)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0				
1				

  - initialisation: $\alpha(i,1) = P(S_1{=}s_1, \mathrm{H}_1 = i) = P(S_1{=}s_1 \mid \mathrm{H}_1 = i)\, P(\mathrm{H}_1 = i)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45			
1				

- initialisation: $\alpha(0,1) = P(S_1=0 \mid \mathrm{H}_1 = 0)\, P(\mathrm{H}_1 = 0) = 0.9 \times 0.5 = 0.45$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

  - initialisation: $\alpha(1,1) = P(S_1=0 \mid H_1 = 1)\ P(H_1 = 1) = 0.2 \times 0.5 = 0.1$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

  - récursion (t=1): $\alpha(i,t+1) = P(S_{t+1} = s_{t+1} \mid H_{t+1} = i) \sum_j P(H_{t+1} = i \mid H_t = j)\, \alpha(j,t)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

  - récursion: $\alpha(0,2) = P(S_2 = 1 | H_2 = 0)\ (\ P(H_2 = 0 | H_1 = 0)\ \alpha(0,1) + P(H_2 = 0 | H_1 = 1)\ \alpha(1,1))$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: $S_1=0$, $S_2=1$, $S_3=0$, $S_4=0$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

$\alpha(i,t)$

i \ t	1	2	3	4
0	0.45	0.0175		
1	0.1			

  - récursion: $\alpha(0,2) = 0.1\ (\ 0.3 \times 0.45 + 0.4 \times 0.1) = 0.0175$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175		
1	0.1			

  - récursion: $\alpha(1,2) = P(S_2 = 1 \mid H_2 = 1)\ (\ P(H_2 = 1 \mid H_1 = 0)\ \alpha(0,1) + P(H_2 = 1 \mid H_1 = 1)\ \alpha(1,1))$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: $S_1=0$, $S_2=1$, $S_3=0$, $S_4=0$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175		
1	0.1	0.3		

  - récursion: $\alpha(1,2) = 0.8\ (\ 0.7 \times 0.45 + 0.6 \times 0.1) = 0.3$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175		
1	0.1	0.3		

  - récursion (t=2): $\alpha(i,t+1) = P(S_{t+1}= s_{t+1} \mid H_{t+1} = i) \sum_j P(H_{t+1} = i \mid H_t = j)\, \alpha(j,t)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175	0.112725	
1	0.1	0.3		

  - récursion: $\alpha(0,3)$ = 0.9 (0.3 x 0.0175 + 0.4 x 0.3) = 0.112725

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175	0.112725	0.04427
1	0.1	0.3	0.03845	0.02039

  - on continue d'appliquer la récursion jusqu'à la fin (t=4)...

Filtrage dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

α(i,t)

i \ t	1	2	3	4
0	0.45	0.0175	0.112725	0.04427
1	0.1	0.3	0.03845	0.02039

  - on peut calculer les probabilités de filtrage

$$P(H_4 = 0 \mid S_1{=}0, S_2{=}1, S_3{=}0, S_4{=}0) = \frac{P(H_4 = 0, S_1{=}0, S_2{=}1, S_3{=}0, S_4{=}0)}{\sum_i P(H_4 = \mathrm{i}, S_1{=}0, S_2{=}1, S_3{=}0, S_4{=}0)}$$
$$= \alpha(0,4) \,/\, (\, \alpha(0,4) + \alpha(1,4)\,)$$
$$= 0.04427 \,/\, (0.04427 + 0.02039)$$
$$\approx 0.6847$$

$$P(H_4 = 1 \mid S_1{=}0, S_2{=}1, S_3{=}0, S_4{=}0) = 0.02039 \,/\, (0.04427 + 0.02039)$$
$$\approx 0.3153$$

Programmation dynamique pour HMM

- Le calcul des $\alpha(i,t)$ donne un balayage de gauche à droite
- On peut faire la même chose, mais de droite à gauche
  - on définit $\beta(i,t) = P(S_{t+1:T}=s_{t+1:T} \mid H_t = i)$
  - on note la récursion

    $$\begin{aligned}\beta(i,t\text{-}1) &= P(S_{t:T}=s_{t:T} \mid \mathrm{H}_{t-1} = i) \\ &= \textstyle\sum_j P(S_{t:T}=s_{t:T}, \mathrm{H}_t = j \mid H_{t-1} = i) \\ &= \textstyle\sum_j P(S_t = s_t \mid H_t = j)\, P(\mathrm{H}_t = j \mid H_{t-1} = i)\, P(S_{t+1:T}=s_{t+1:T} \mid H_t = j) \\ &= \textstyle\sum_j P(S_t = s_t \mid H_t = j)\, P(\mathrm{H}_t = j \mid H_{t-1} = i)\, \beta(j,t)\end{aligned}$$
  - on a les valeurs initiales $\beta(i,T) = 1\ \forall i$

- Une fois le tableau β calculé, on obtient facilement:

  $$\begin{aligned}P(S_{1:T}=s_{1:T}) &= \textstyle\sum_j P(S_{1:T}=s_{1:T}, \mathrm{H}_1 = j) \\ &= \textstyle\sum_j P(S_{2:T}=s_{2:T} \mid \mathrm{H}_1 = j)\, P(S_1=s_1 \mid H_1 = j)\, P(H_1 = j) \\ &= \textstyle\sum_j \beta(j,1)\, P(S_1=s_1 \mid H_1 = j)\, P(H_1 = j)\end{aligned}$$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0				
1				

  - initialisation: β(i,4) = 1

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0				1
1				1

  - initialisation: β(i,4) = 1

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0				1
1				1

  - récursion (t=4): $\beta(i,t-1) = \sum_j P(S_t = s_t \mid H_t = j)\, P(H_t = j \mid H_{t-1} = i)\, \beta(j,t)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0				1
1				1

  - récursion $\beta(0,3) = P(S_4=0|H_4=0)\ P(H_4=0|H_3=0)\ \beta(0,4)$
    $+ P(S_4=0|H_4=1)\ P(H_4=1|H_3=0)\ \beta(1,4)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0			0.41	1
1				1

- récursion β(0,3) = 0.9 x 0.3 x 1 + 0.2 x 0.7 x 1 = 0.41

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0			0.41	1
1				1

  - récursion $\beta(1,3) = P(S_4=0|H_4=0)\, P(H_4=0|H_3=1)\, \beta(0,4) + P(S_4=0|H_4=1)\, P(H_4=1|H_3=1)\, \beta(1,4)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0			0.41	1
1			0.48	1

  - récursion β(1,3) = 0.9 x 0.4 x 1 + 0.2 x 0.6 x 1 = 0.48

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0			0.41	1
1			0.48	1

  - récursion (t=3): $\beta(i,t-1) = \sum_j P(S_t = s_t \mid H_t = j)\, P(H_t = j \mid H_{t-1} = i)\, \beta(j,t)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0			0.41	1
1			0.48	1

- récursion $\beta(0,2) = P(S_3=0|H_3=0)\, P(H_3=0|H_2=0)\, \beta(0,3) + P(S_3=0|H_3=1)\, P(H_3=1|H_2=0)\, \beta(1,3)$

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: $S_1=0$, $S_2=1$, $S_3=0$, $S_4=0$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0		0.1779	0.41	1
1			0.48	1

  - récursion β(0,2) = 0.9 x 0.3 x 0.41 + 0.2 x 0.7 x 0.48 = 0.1779

Programmation dynamique pour HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: $S_1=0$, $S_2=1$, $S_3=0$, $S_4=0$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

β(i,t)

i \ t	1	2	3	4
0	0.120249	0.1779	0.41	1
1	0.105612	0.2052	0.48	1

  - on continue d'appliquer la récursion jusqu'au début (t=1)...

Lissage avec un HMM

- Les tables α(i,t) et β(i,t) peuvent également être utilisées pour faire du lissage
  - $P(H_k = i \mid S_{1:T}=s_{1:T}) = \Upsilon\, P(H_k = i, S_{1:k}=s_{1:k}, S_{k+1:T}=s_{k+1:T})$ (ϒ est la normalisation)
    $= \Upsilon\, P(H_k = i, S_{1:k}=s_{1:k})\, P(S_{k+1:T}=s_{k+1:T} \mid H_k = i)$
    $= \Upsilon\, \alpha(i,k)\, \beta(i,k)$

- On peut également faire du lissage sur deux variables cachées adjacentes
  - $P(H_k = i, H_{k+1}= j \mid S_{1:T}=s_{1:T}) = \Upsilon'\, P(H_k = i, H_{k+1}= j, S_{1:k}=s_{1:k}, S_{k+1:T}=s_{k+1:T})$
    $= \Upsilon'\, P(H_k = i, S_{1:k}=s_{1:k})\, P(H_{k+1}= j \mid H_k = i)\, P(S_{k+1}=s_{k+1} \mid H_{k+1} = j)$
    $P(S_{k+2:T}=s_{k+2:T} \mid H_{k+1} = j)$
    $= \Upsilon'\, \alpha(i,k)\, \beta(j,k+1)\, P(H_{k+1}= j \mid H_k = i)\, P(S_{k+1}=s_{k+1} \mid H_{k+1} = j)$
- À noter que ϒ correspond à une somme sur *i* seulement,
  tandis que ϒ’ est une somme sur *i* **et** *j*

Lissage avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

α(i,t)

i \ t	...	2	...
0	...	0.0175	...
1	...	0.3	...

β(i,t)

i \ t	...	2	...
0	...	0.1779	...
1	...	0.2052	...

  - on peut calculer les probabilités de lissage au temps t=2

$$P(H_2 = 0 \mid S_1=0, S_2=1, S_3=0, S_4=0) = \frac{\alpha(0,2)\ \beta(0,2)}{\sum_i \alpha(i,2)\ \beta(i,2)}$$

$$= \alpha(0,2)\ \beta(0,2) / (\alpha(0,2)\ \beta(0,2) + \alpha(1,2)\ \beta(1,2))$$

$$= 0.0175 \times 0.1779 / (0.0175 \times 0.1779 + 0.3 \times 0.2052)$$

$$\approx 0.04813$$

$$P(H_2 = 1 \mid S_1=0, S_2=1, S_3=0, S_4=0)$$

$$= 0.3 \times 0.2052 / (0.0175 \times 0.1779 + 0.3 \times 0.2052)$$

$$\approx 0.95186$$

Prédiction dans un HMM

- $\alpha(i,t)$ peut être utilisé pour inférer la distribution de prédiction $\mathbf{P}(H_{t+k} | s_{1:t})$
- On utilise également un programme dynamique
  - on définit $\pi(i,k) = P(H_{t+k} = i | S_{1:t} = s_{1:t})$
  - on note la récursion

$$
\begin{aligned}
\pi(i,k+1) &= P(H_{t+k+1} = i | S_{1:t} = s_{1:t}) \\
&= \sum_s \sum_j P(H_{t+k+1} = i, H_{t+k} = j, S_{t+k} = s | S_{1:t} = s_{1:t}) \\
&= \sum_s \sum_j P(S_{t+k} = s | H_{t+k} = j)\, P(H_{t+k+1} = i | H_{t+k} = j)\, P(H_{t+k} = j | S_{1:t} = s_{1:t}) \\
&= \sum_j P(H_{t+k+1} = i | H_{t+k} = j)\, P(H_{t+k} = j | S_{1:t} = s_{1:t}) \sum_s P(S_{t+k} = s | H_{t+k} = j) \\
&= \sum_j P(H_{t+k+1} = i | H_{t+k} = j)\, \pi(j,k)
\end{aligned}
$$

  - on a les valeurs initiales

$$\pi(i,0) = P(H_t = i | s_{1:t}) = \alpha(i,t) / \sum_j \alpha(j,t) \quad \forall i$$

- On pourrait également faire une prédiction de S_{t+k}
  - $P(S_{t+k} = s | S_{1:t} = s_{1:t}) = \sum_j P(S_{t+k} = s | H_{t+k} = j)\, P(H_{t+k} = j | S_{1:t} = s_{1:t})$

$$= \sum_j P(S_{t+k} = s | H_{t+k} = j)\, \pi(j,k)$$

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6\text{=}0 \mid S_1\text{=}0, S_2\text{=}1, S_3\text{=}0, S_4\text{=}0)$

Distribution d'émission

	H_t=0	H_t=1
$P(S_t\text{=}0 \mid H_t)$	0.9	0.2
$P(S_t\text{=}1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t\text{=}0 \mid H_{t-1})$	0.3	0.4
$P(H_t\text{=}1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0			
1			

  - initialisation: $\pi(i,0) = \alpha(i,t) / \sum_j \alpha(j,t)$

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0			
1			

  - initialisation: π(0,0) = α(0,4) / (α(0,4) + α(1,4))

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0, S_2=1, S_3=0, S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466		
1			

- initialisation: π(0,0) = 0.04427 / (0.04427 + 0.02039) = 0.68466

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0, S_2=1, S_3=0, S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466		
1	0.31534		

- initialisation: π(1,0) = 0.02039 / (0.04427 + 0.02039) = 0.31534

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6\!=\!0 \mid S_1\!=\!0, S_2\!=\!1, S_3\!=\!0, S_4\!=\!0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466		
1	0.31534		

  - récursion (k=0): $\pi(i,k+1) = \sum_j P(H_{t+k+1} = i \mid H_{t+k} = j)\ \pi(j,k)$

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466		
1	0.31534		

  - récursion (k=0): $\pi(0, 1) = P(H_5 = 0 \mid H_4 = 0)\, \pi(0,0) + P(H_5 = 0 \mid H_4 = 1)\, \pi(1,0)$

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466	0.33154	
1	0.31534		

- récursion (k=0): π(0, 1) = 0.3 x 0.68466 + 0.4 x 0.31534 = 0.33154

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466	0.33154	
1	0.31534		

  - récursion (k=0): $\pi(1, 1) = P(H_5 = 1 | H_4 = 0)\ \pi(0,0) + P(H_5 = 1 | H_4 = 1)\ \pi(1,0)$

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6\!=\!0 \mid S_1\!=\!0, S_2\!=\!1, S_3\!=\!0, S_4\!=\!0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466	0.33154	
1	0.31534	0.66846	

  - récursion (k=0): π(0, 1) = 0.7 x 0.68466 + 0.6 x 0.31534 = 0.66846

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466	0.33154	0.36685
1	0.31534	0.66846	0.63315

  - on continue d'appliquer la récursion jusqu'à la fin (k=2)...

Prédiction dans un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - on souhaite calculer la distribution de prédiction $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0)$

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α(i,t)

i \ t	...	4
0	...	0.04427
1	...	0.02039

π(i,k)

i \ k	0	1	2
0	0.68466	0.33154	0.36685
1	0.31534	0.66846	0.63315

  - $P(H_6=0 \mid S_1=0,S_2=1,S_3=0,S_4=0) = \pi(0,2) = 0.36685$

Explication la plus plausible avec un HMM

- On peut également éviter une énumération exponentielle
  - exemple avec T=3

    $\max_{h^*_{1:3}} P(h^*_1)\, P(s_1|h^*_1)\, P(h^*_2|h^*_1)\, P(s_2|h^*_2)\, P(h^*_3|h^*_2)\, P(s_3|h^*_3)$

    $= \max_{h^*_3} P(s_3|h^*_3) \max_{h^*_2} P(s_2|h^*_2)\, P(h^*_3|h^*_2) \max_{h^*_1} P(h^*_2|h^*_1)\, P(h^*_1)\, P(s_1|h^*_1)$

- Solution: **programmation dynamique**, avec un **max au lieu de la somme**
  - on définit $\alpha^*(\mathrm{i,t}) = P(S_{1:\mathrm{t}}=s_{1:\mathrm{t}}, H_{1:t-1} = h^*_{1:t-1}, H_t = i)$
  - on note la récursion

    $\alpha^*(\mathrm{i,t+1}) = \max_j P(S_{1:\mathrm{t+1}}=s_{1:\mathrm{t+1}}, H_{1:t-1} = h^*_{1:t-1}, H_t = j, H_{t+1} = i)$

    $= \max_j P(S_{\mathrm{t+1}}=s_{\mathrm{t+1}} \mid H_{t+1} = i)\, P(H_{t+1} = i \mid H_t = j)\, P(S_{1:\mathrm{t}}=s_{1:\mathrm{t}}, H_{1:t-1} = h^*_{1:t-1}, H_t = j)$

    $= P(S_{\mathrm{t+1}}=s_{\mathrm{t+1}} \mid H_{t+1} = i) \max_j P(H_{t+1} = i \mid H_t = j)\, \alpha^*(\mathrm{j,t})$
  - on a les valeurs initiales: $\alpha^*(\mathrm{i},1) = P(S_1=s_1| H_1 = i)\, P(H_1 = i) \quad \forall i$
- On a alors que $P(S_{1:\mathrm{T}}=s_{1:\mathrm{T}}, H_{1:T} = h^*_{1:T}) = \max_{\mathrm{j}} \alpha^*(\mathrm{j,T})$
- On retrouve $h^*_{1:T}$ à partir de tous les $\mathrm{argmax}_{\mathrm{j}}$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t$=0 | $H_t)$	0.9	0.2
$P(S_t$=1 | $H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t$=0 | $H_{t-1})$	0.3	0.4
$P(H_t$=1 | $H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0				
1				

  - initialisation: α*(i,1) = $P(S_1{=}s_1, H_1 = i) = P(S_1{=}s_1 \mid H_1 = i)\, P(H_1 = i)$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45			
1				

- initialisation: $\alpha^*(0,1) = P(S_1=0 \mid \mathrm{H}_1 = 0)\, P(\mathrm{H}_1 = 0) = 0.9 \times 0.5 = 0.45$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

  - initialisation: $\alpha^*(1,1) = P(S_1=0 \mid \mathrm{H}_1 = 1)\, P(\mathrm{H}_1 = 1) = 0.2 \times 0.5 = 0.1$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

- récursion (t=1): $\alpha^*(i,t+1) = P(S_{t+1}=s_{t+1} \mid H_{t+1} = i) \max_j P(H_{t+1} = i \mid H_t = j)\, \alpha^*(j,t)$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45			
1	0.1			

  - récursion: $\alpha^*(0,2) = P(S_2{=}1 \mid H_2{=}0) \max\{P(H_2 = 0 \mid H_1 = 0)\ \alpha^*(0,1),\ P(H_2 = 0 \mid H_1 = 1)\ \alpha^*(1,1)\}$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135		
1	0.1			

  - récursion: $\alpha^*(0,2) = 0.1 \max\{ \underline{0.3 \times 0.45}, 0.4 \times 0.1\} = 0.0135$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t\text{=0} \mid H_t)$	0.9	0.2
$P(S_t\text{=1} \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t\text{=0} \mid H_{t-1})$	0.3	0.4
$P(H_t\text{=1} \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135		
1	0.1			

  - récursion: α*(1,2) = $P(S_2\text{=1}|H_2\text{=1})$ max{$P(H_2 = 1|H_1 = 0)$ α*(0,1), $P(H_2 = 1|H_1 = 1)$ α*(1,1)}

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135		
1	0.1	0.252		

  - récursion: α*(1,2) = 0.8 max{ <u>0.7 x 0.45</u>, 0.6 x 0.1} = 0.252

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135		
1	0.1	0.252		

- récursion (t=2): α*(i,t+1) = $P(S_{t+1}=s_{t+1} \mid H_{t+1} = i)$ max$_j$ $P(H_{t+1} = i \mid H_t = j)$ α*(j,t)

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: $S_1=0$, $S_2=1$, $S_3=0$, $S_4=0$

Distribution d'émission

	$H_t=0$	$H_t=1$
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	$H_{t-1}=0$	$H_{t-1}=1$
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	$H_1=0$	$H_1=1$
$P(H_1)$	0.5	0.5

$\alpha^*(i,t)$

i \ t	1	2	3	4
0	0.45	0.0135	0.09072	
1	0.1	0.252		

- récursion: $\alpha^*(0,3) = 0.9 \max\{ 0.3 \times 0.0135, \underline{0.4 \times 0.252}\} = 0.09072$

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135	0.09072	0.02449
1	0.1	0.252	0.03024	0.01270

- on continue d'appliquer la récursion jusqu'à la fin (t=4)...

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135	0.09072	0.02449
1	0.1	0.252	0.03024	0.01270

  - on trouve le maximum à la dernière colonne...

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t\text{=0} \mid H_t)$	0.9	0.2
$P(S_t\text{=1} \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t\text{=0} \mid H_{t-1})$	0.3	0.4
$P(H_t\text{=1} \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135	0.09072	0.02449
1	0.1	0.252	0.03024	0.01270

- ... puis on retrouve le chemin associé aux maxima précédents

Explication la plus plausible avec un HMM

- Exemple: décoder un message binaire avec canal bruité (T=4)
  - message observé: S_1=0, S_2=1, S_3=0, S_4=0

Distribution d'émission

	H_t=0	H_t=1
$P(S_t=0 \mid H_t)$	0.9	0.2
$P(S_t=1 \mid H_t)$	0.1	0.8

Distribution de transition

	H_{t-1}=0	H_{t-1}=1
$P(H_t=0 \mid H_{t-1})$	0.3	0.4
$P(H_t=1 \mid H_{t-1})$	0.7	0.6

Distribution initiale

	H_1=0	H_1=1
$P(H_1)$	0.5	0.5

α*(i,t)

i \ t	1	2	3	4
0	0.45	0.0135	0.09072	0.02449
1	0.1	0.252	0.03024	0.01270
	H_1=0	H_2=1	H_3=0	H_4=0

  - ce chemin nous donne la séquence des H_t la plus probable

Au delà du HMM

- **Filtre de Kalman:** cas où les variables d'observation et cachées ne sont pas discrètes mais sont plutôt réelles
  - voir livre de référence, section 15.4

- **État caché avec structure complexe:** cas où il n'est pas possible de faire une sommation exacte sur toutes les configurations de l'état caché
  - on doit alors approximer l'inférence
  - **filtre particulaire** (*particle filter*): inférence approximative basée sur l'échantillonnage, où on maintien une population stochastique de configurations (particules) de l'état caché
  - à chaque temps *t*, on met à jour notre population de particules en tenant compte des nouvelles observations
  - voir livre de référence, section 15.5.3

Application: reconnaissance vocale

- La reconnaissance vocale est difficile:
  - bruit ambiant ou introduit par la digitalisation
  - variations dans la prononciation
  - différents mots ont la même prononciation

- Problème: Quelle est la séquence de mots la plus vraisemblable étant donné un signal sonore ?

- Réponse: Choisir la séquence de mots la plus probable a posteriori
  - $\underset{\text{mots}}{\operatorname{argmax}}\ P(\textit{mots} \mid \textit{signal}) = \underset{\text{mots}}{\operatorname{argmax}}\ \alpha\ P(\textit{mots}, \textit{signal})$

Modèle acoustique et modèle du langage

- En utilisant la règle de Bayes
  - $P(mots \mid signal) = \alpha\, P(signal \mid mots)\, P(mots)$

- On peut donc décomposer le problème en deux:
  - **P**(*Signal* | *Mots*): **modèle acoustique**
  - **P**(*Mots*): **modèle de langage** (plus de détails à venir dans le cours...)

- Chaîne cachée: les mots
- Chaîne observée: le signal

Phones et phonèmes

- Des travaux dans le domaine de phonologie ont montré que toutes les langues humaines sont basées sur un sous-ensemble d'environ 100 sons, appelés **phones**, communs à toutes les langues
- Les phones découlent de l'articulation des lèvres, des dents, de la langue, des cordes vocales et du flux de l'air
- Intuitivement, un phone est un son qui correspond à une seule consonne ou une seule voyelle
- Mais c'est plus subtil ! Des combinaisons de consonnes comme « th » ou « ng » en anglais ont chacun leur phone
- Un **phonème** est la plus petite unité de son distinctive que l'on puisse isoler par segmentation dans un mot
- Un phonème sera associé à un ou plusieurs phones qui peuvent être inter-changés sans changer la compréhension d'un mot
  - phonème /k/: phones [k] (« *cat* », « *kit* ») et [k^h] (« *school* », « *skill* »)

Phones: exemple

- Phones pour l'anglais américain:

[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[ey]	bet	[d]	debt	[s]	set
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[er]	Bert	[l]	let	[w]	wet
[ix]	roses	[ng]	sing	[en]	button
⋮	⋮	⋮	⋮	⋮	⋮

Modèle acoustique

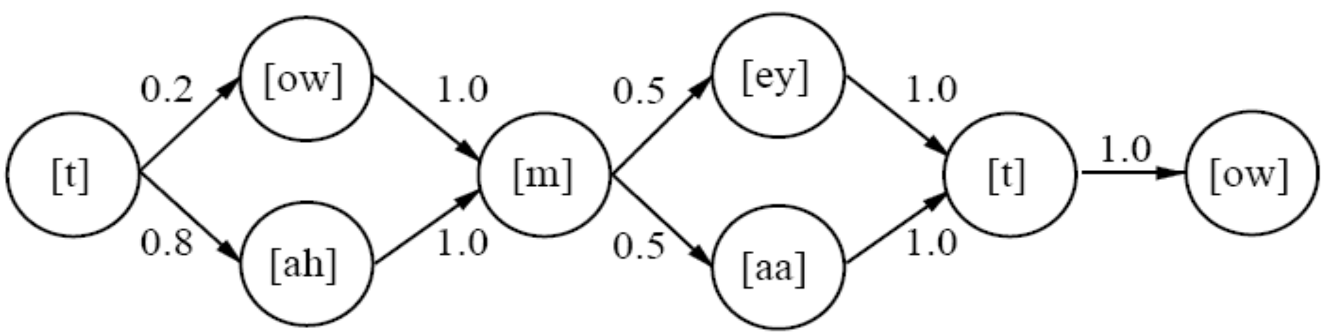
- Rappel:
  - **P**(*Mots* | *Signal*) = α **P**(*Signal* | *Mots*) **P**(*Mots*)
    - » **P**(*Signal* | *Mots*): modèle acoustique
    - » **P**(*Mots*): modèle de langage

- L'existence des phones permet de diviser le modèle acoustique en deux autres parties:
  - **modèle de prononciation**: spécifie, pour chaque mot, une distribution de probabilités sur une séquence de phones
    - » par exemple, « *ceiling* » est parfois prononcé [s iy l ih ng], ou [s iy l ix ng], ou encore [s iy l en]
    - » le phone est une variable cachée, le signal est la variable observée

  - **modèle phonique**: le modèle phonique $P(e_t|x_t)$ donne la probabilité que le signal échantillonné soit observé au temps t si le phone est x_t

Exemple de modèle de prononciation

- Modèle de prononciation
  - P([towmeytow]| « tomato») = P([towmaatow]| « tomato») = 0.1
  - P([tahmeytow]| « tomato») = P([tahmaatow]| « tomato») = 0.4
- Les transitions sont spécifiées manuellement
- Les probabilités sont apprises



Applications

- Reconnaissance vocale
  - CMU Sphinx (publique): http://cmusphinx.sourceforge.net/html/download.php
  - Dragon Naturally Speaking (commercial)
  - IBM ViaVoice (commercial)

- Reconnaissance de caractères
  - variables observées: pixels d'une image
  - variables cachées: mots écrits

- Suivi d'objets dans une vidéo (*object tracking*)
  - variables observées: pixels d'une image, dans le *frame t*
  - variables cachées: la position x_t,y_t de l'objet dans le *frame*