IFT 615 – Intelligence Artificielle

Raisonnement probabiliste temporel

Professeur: Froduald Kabanza
Assistants: D'Jeff Nkashama

Université de Sherbrooke

Sujets couverts

- Types d’inférence probabiliste temporelle
  - Filtrage
  - Prédiction
  - Explication la plus plausible
  - Lissage
- Réseau bayésien dynamique
- Chaîne de Markov
- Chaîne de Markov cachée
- Filtre de particules

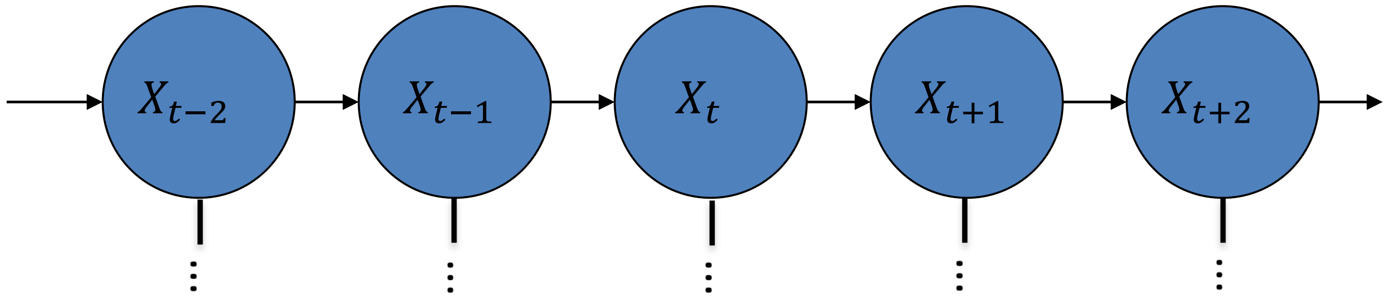
Réseau bayésien dynamique (RBD)

- 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*

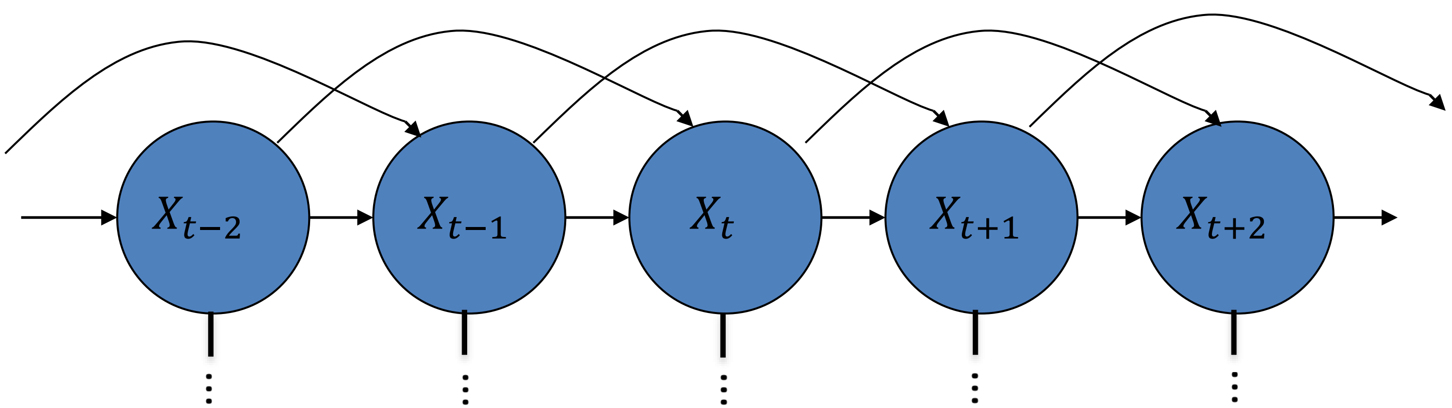
- Le terme dynamique réfère au dynamisme du système qu'on veut modéliser et la structure du réseau

Exemple de RDBs

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

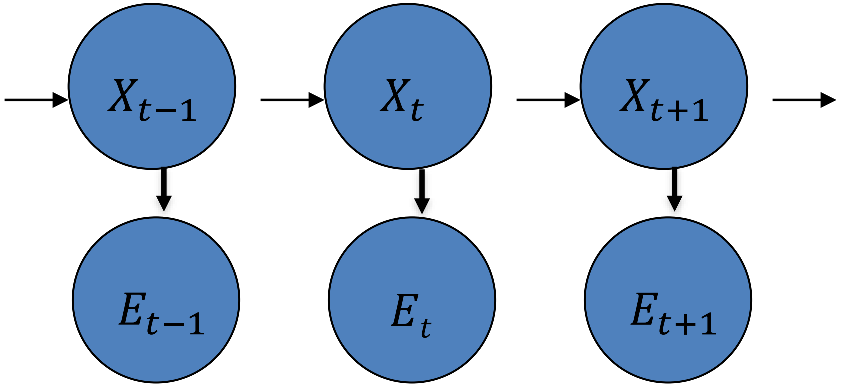


- RBD du second ordre:



Exemple de RDBs

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

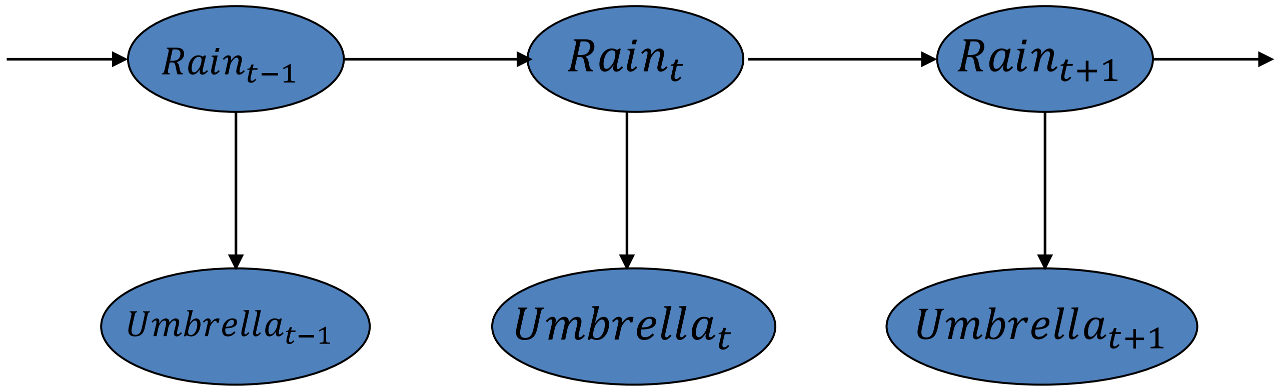


Représentation dans un RBD

- Problème:
  - il faudrait spécifier un grand nombre (même infini) de tables de probabilités conditionnelles, c.-à-d. une pour chaque temps *t*
  - chaque table pourrait impliquer un nombre infini de parents
- Solution:
  1. supposer que les changements dynamiques sont causés par un **processus homogène dans le temps** - 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 des **changements d'états 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 des **capteurs markoviens :** 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

Exemple

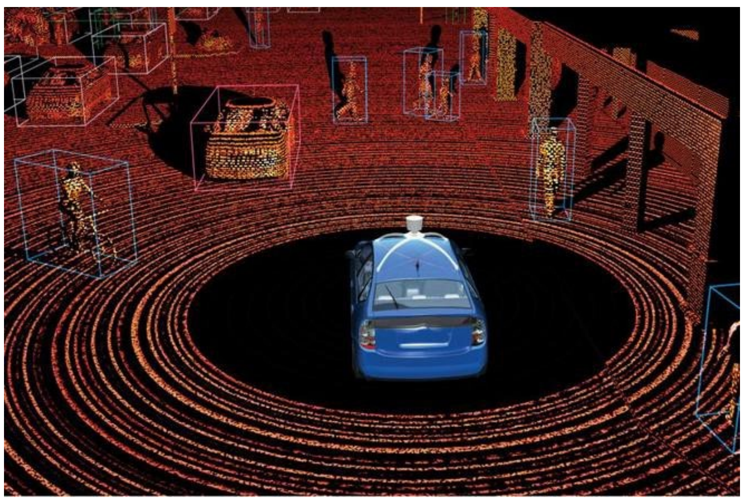
- « Un gardien de sécurité passe un mois dans un édifice sous-terrain, sans sortir. Chaque jour, son directeur arrive avec ou sans parapluie. Le gardien veut inférer la possibilité qu'il ait plu ou non en fonction des séquences d'observation du parapluie. »
- Modélisation:
  - Variables: $X_t = \{R_t\}$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



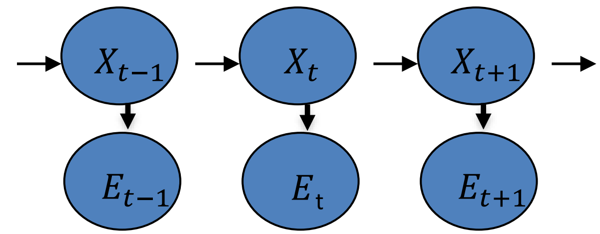
  - Modèle des transitions: $\mathbf{P}(R_t \mid R_{t-1})$. Modèle d'observation: $\mathbf{P}(U_t \mid R_t)$

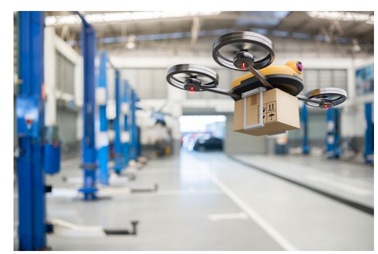
Application - Localisation

- **Modèle (filtre de particule)**
  - E_t sont l'information fournie par les capteurs du robot
  - X_t sont l'information sur la position du robot



Source: Udacity

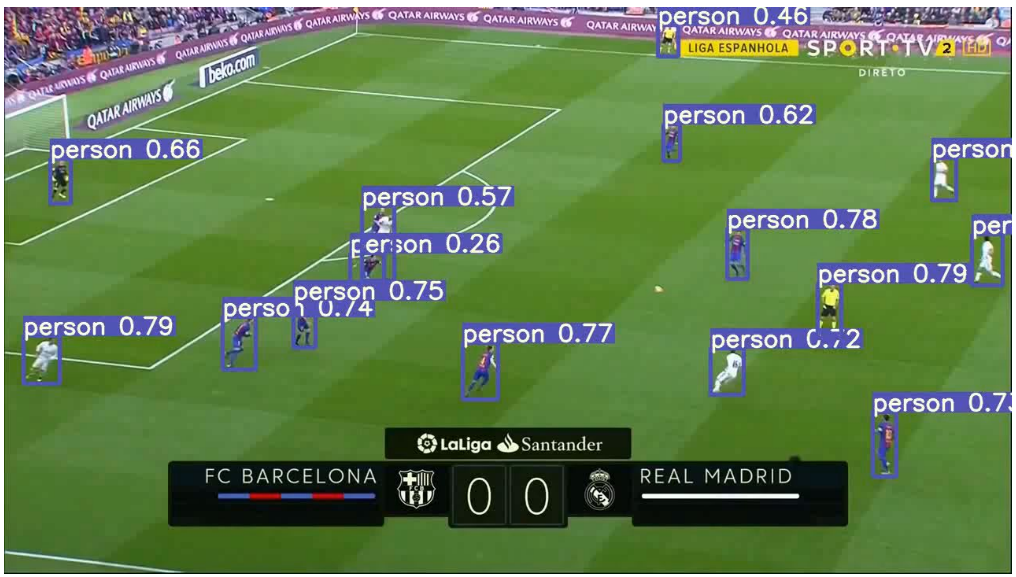


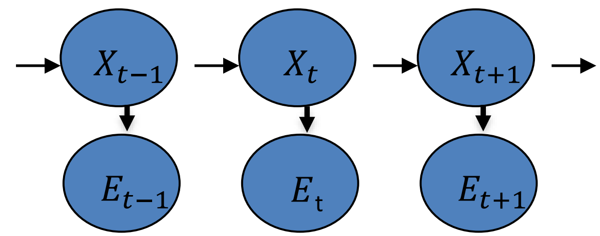


- Très utilisé en robotique et conduite autonome (filtre de particules)

Application – Suivi d’objets (tracking)

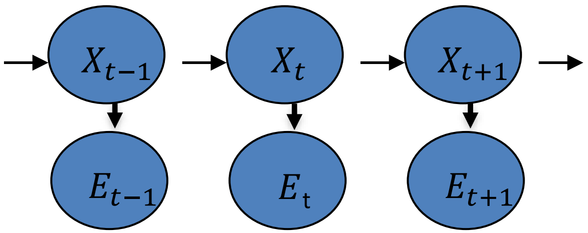
- **Modèle (modèle de Markov caché et filtre de Kalman)**
  - E_t sont les *frames* de la vidéo
  - X_t sont l’information sur la position d’un/des objet(s)

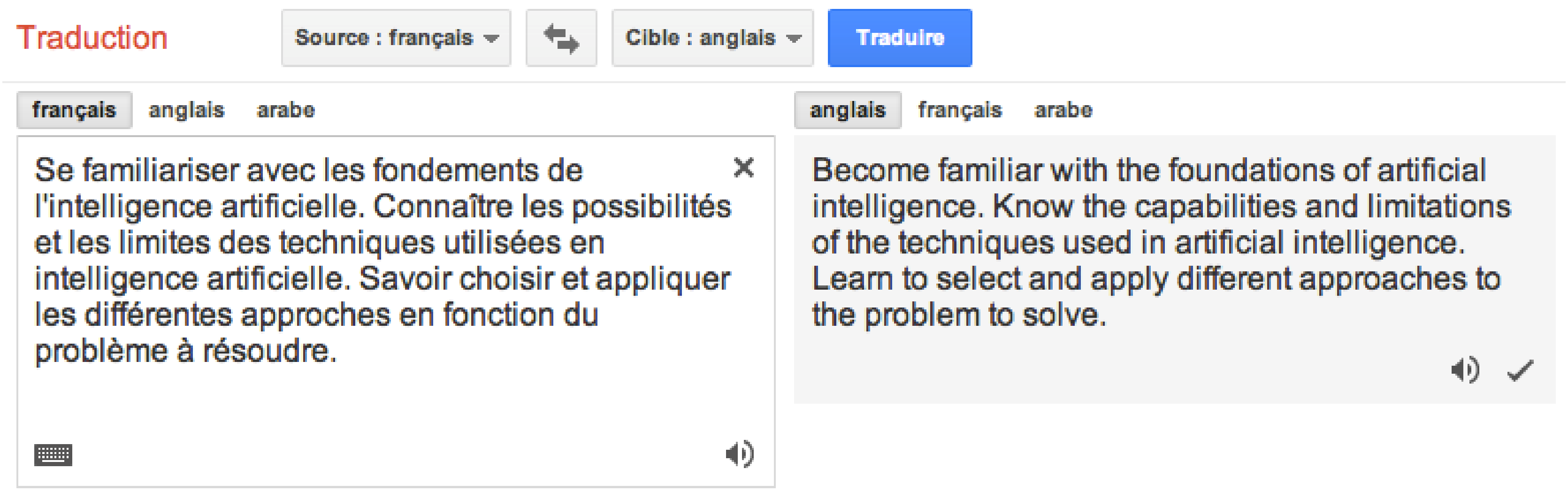




Application – Traduction automatique

- **Modèle:**
  - E_t sont les mots en français
  - X_t sont les mots de la traduction en anglais

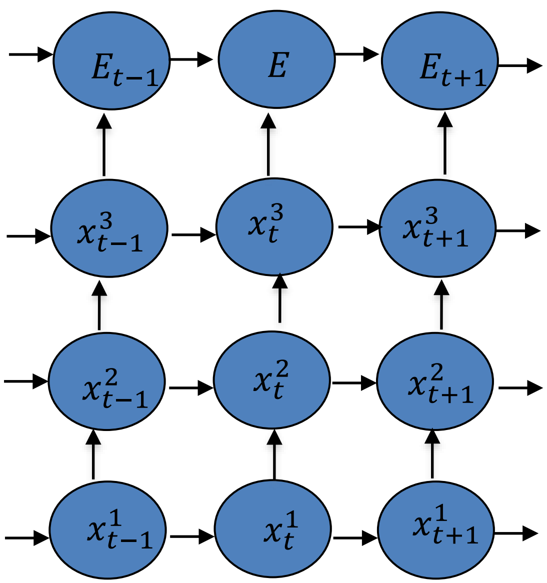


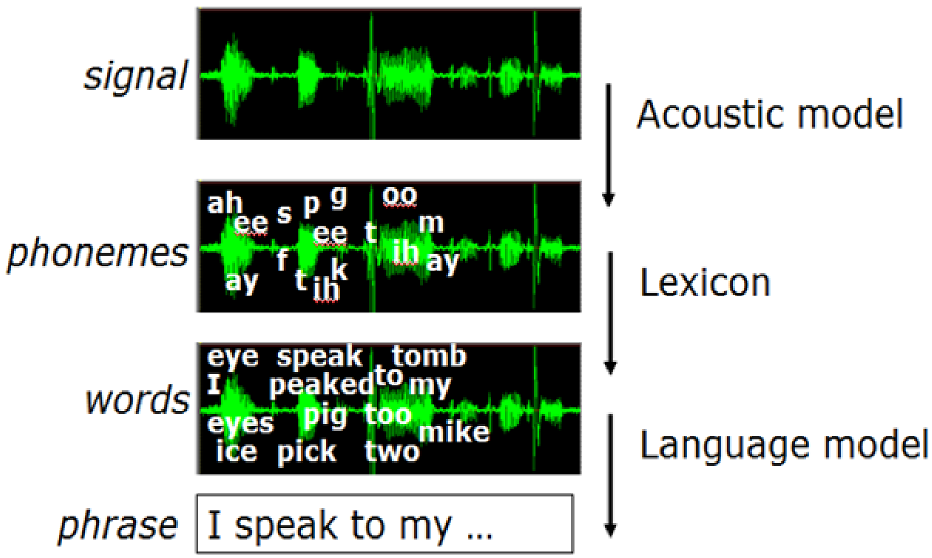


- Les réseaux de neurones ont pris le dessus sur les approches probabilistes.

Application - Reconnaissance vocale

- **Modèle:**
  - E_t sont les éléments du signal sonore
  - X_t sont les phrases et les mots prononcés ainsi que les phonèmes





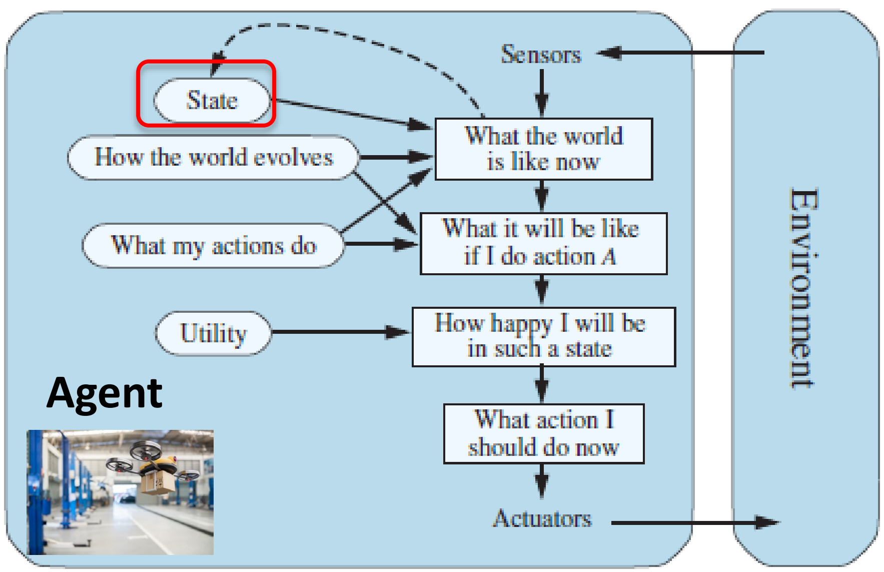
- Les réseaux de neurones ont remplacé les approches probabilistes.

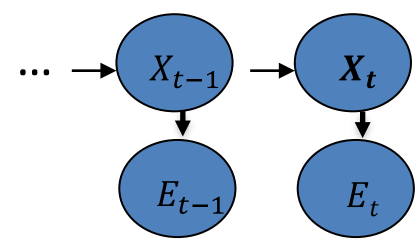
Types d'inférence dans un RBD

- **Filtrage** (***filtering***): calcul de l'**état de croyance** (***belief state***), c.-à-d. la distribution à posteriori de la variable cachée la plus récente

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

  - ex. : quelle est la probabilité qu'il pleuve aujourd'hui ?
  - ex. : quelle est la position du robot (la plus probable) ?





Programmation dynamique pour le filtrage

- Étant donné le résultat du filtrage au temps *t*, on peut calculer le résultat du filtrage au temps *t+1* à partir des nouvelles observations e_{t+1}.

$$\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) = P(X_{t+1}|e_{1:t}, e_{t+1})$$ (Séparer l'évidence en 2)

$$= \alpha P\left(e_{t+1}\middle|X_{t+1}, e_{1:t}\right)P(X_{t+1}|e_{1:t})$$ (Règle de Bayes)
α : constante de normalisation

$$= \alpha P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t})$$ (Hypothèse markovienne du modèle sensoriel)

$$= \alpha P(e_{t+1}|X_{t+1}) \sum_{x_t} P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})$$ Marginalisation

$$= \alpha \underbrace{P(e_{t+1}|X_{t+1})}_{\text{modèle sensoriel}} \sum_{x_t} \underbrace{P(X_{t+1}|x_t)}_{\text{modèle dynamique}} \underbrace{\mathbf{P}(\mathbf{x}_t|\mathbf{e}_{1:t})}_{\text{récursion}}$$ (Hypothèse Markovienne)

Programmation dynamique pour le filtrage

L'équation

$$\boldsymbol{P}(\boldsymbol{X}_{t+1}|\boldsymbol{e}_{1:t+1}) = \alpha \underbrace{P(e_{t+1}|X_{t+1})}_{\text{modèle sensoriel}} \sum_{x_t} \underbrace{P(X_{t+1}|x_t)}_{\text{modèle dynamique}} \underbrace{\boldsymbol{P}(\boldsymbol{x}_t|\boldsymbol{e}_{1:t})}_{\text{récursion}}$$

Peut se réécrire

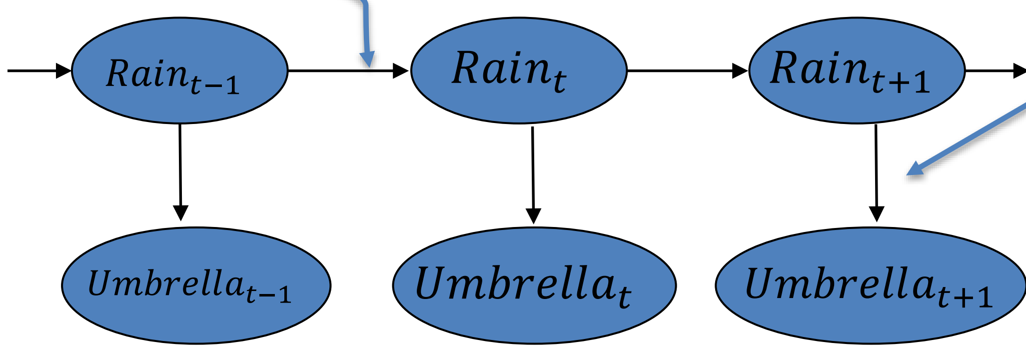
$f_{1:t+1} = \alpha \text{FORWARD}(\mathbf{f}_{1:t}\ \mathbf{e}_{t+1})$

où FORWARD est un algorithme de programmation dynamique implémentant la récursion de l'équation précédente et $f_{1:0} = P(X_0)$.

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($R_t | R_{t-1}$)
V	0.7
F	0.3



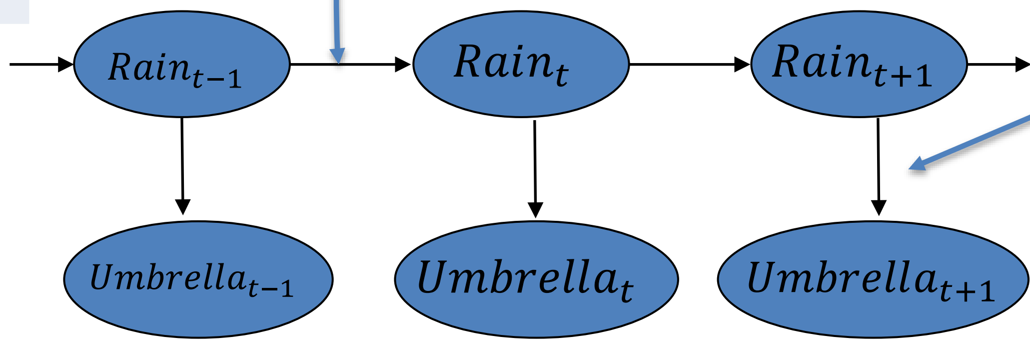
R_t	P($U_t | R_t$)
V	0.9
F	0.2

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

$$= \alpha <0.9, 0.2> <0.5, 0.5>$$
$$= \alpha <0.45, 0.1>$$
$$\approx < 0.818, 0.182>$$

Exemple de l'agent de sécurité

R_{t-1}	$P(R_t \mid R_{t-1})$
V	0.7
F	0.3



R_t	$P(U_t \mid R_t)$
V	0.9
F	0.2

$\boldsymbol{P(R_1 \mid u_1)}$
$\approx$ < 0.818, 0.182>

$$\boldsymbol{P}(\boldsymbol{X_{t+1}}|\boldsymbol{e_{1:t+1}}) = \alpha P(e_{t+1}|X_{t+1}) \sum_{h_t} P(X_{t+1}|x_t)\, \boldsymbol{P}(\boldsymbol{x_t}|\boldsymbol{e_{1:t}})$$

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

$$\boldsymbol{P(R_2 \mid u_1, u_2)} = \alpha\, P(u_2 \mid R_2) \textstyle\sum_{r_1} P(R_2|r_1)\ \boldsymbol{P}(\boldsymbol{r_1}\,|\boldsymbol{u_1})$$

$$= \alpha \text{ <0.9, 0.2> (<0.7, 0.3>*0.818 + <0.7, 0.3>*0.182)}$$

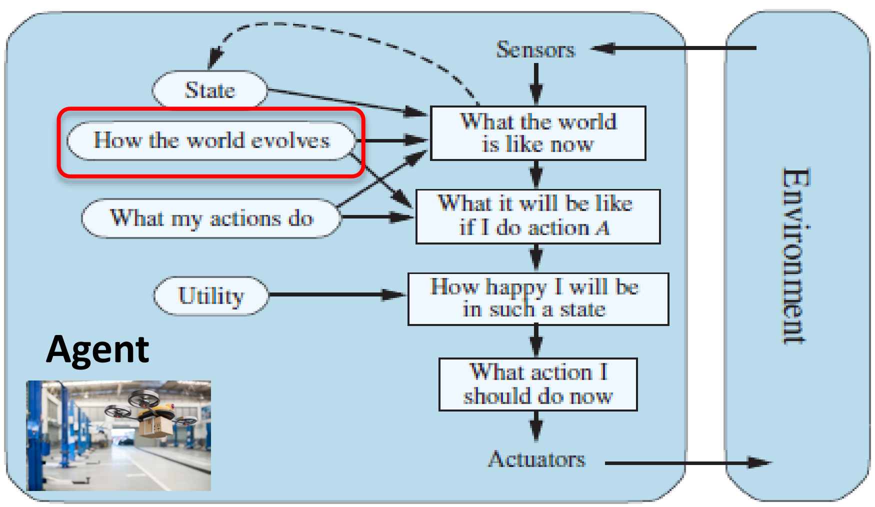
$$= \alpha \text{ <0.565, 0.075>} \approx \text{<0.883, 0.117>}$$

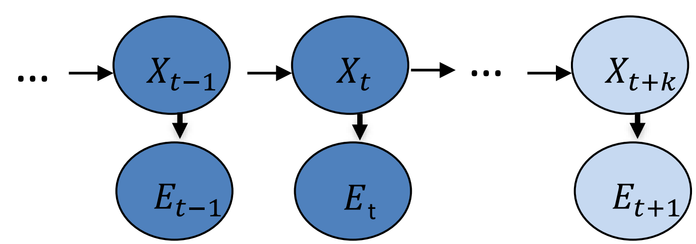
Types d'inférence dans un RBD

- **Prédiction**: calculer la distribution a posteriori sur un état futur

$$\mathbf{P}(X_{t+k} \mid e_{1:t}) \text{ où } k > 0$$

  - ex. : quelle est la probabilité qu'il pleuve dans *k* jours ?
  - ex.: quelle est la mission la plus probable du drone ?
  - ex: quelle est la position probable à un temps future ?
  - ex: quelle est la probabilité qu'une composante tombe en panne





- Les réseaux de neurones ont pris le dessus sur les approches probabilistes.

Types d'inférence dans un RBD

- **Lissage** (***smoothing***): calculer la distribution a posteriori sur un état passé

$$\mathbf{P}(X_k | e_{1:t}) \text{ où } 0 \leq k < t$$

  - ex. : quelle est la probabilité qu'il y ait eu de la pluie hier (k=t-1) ?

- **Explication la plus plausible**: trouver la séquence d'états cachés qui explique le mieux les observations

$$\underset{x_{1:t}}{\text{argmax}}\, P(x_{1:t} | e_{1:t}) = \underset{x_{1:t}}{\text{argmax}}\, P(x_{1:t}, e_{1:t}) \,/\, P(e_{1:t}) = \underset{x_{1:t}}{\text{argmax}}\, P(x_{1:t}, e_{1:t})$$

  - ex. : quelle a été la météo la plus probable pour toutes les t dernières journées ?
  - ex. : quelle est la traduction en anglais d'une phrase donnée en français ?
  - ex. : quelle est la phrase qui a été prononcée ?

Chaînes de Markov

- Une **chaîne de Markov** (de premier ordre) est un cas particulier de RBD
  - avec **une seule** variable aléatoire **discrète** X_t dans l'état au temps t

- Le domaine de X_t est souvent un ensemble de symboles (ex.: un caractère, un mot, etc.)

- Une **distribution a priori** (initiale) de probabilités sur les symboles (états) est spécifiée $\mathbf{P}(X_1)$

- Une **matrice de transition** contenant les probabilités conditionnelles $\mathbf{P}(X_{t+1} \mid X_t)$

Illustration

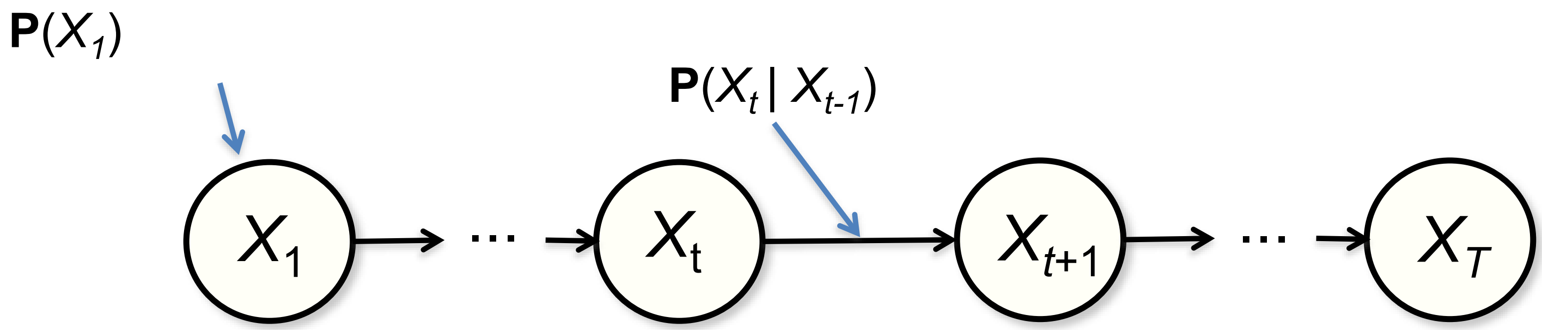


Illustration dans le cas d'une **chaîne finie**

Visualisation d'une chaîne de Markov

Représentation matricielle

X_t

Symbole actuel

X_{t+1} Prochain symbole

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

Représentation graphique

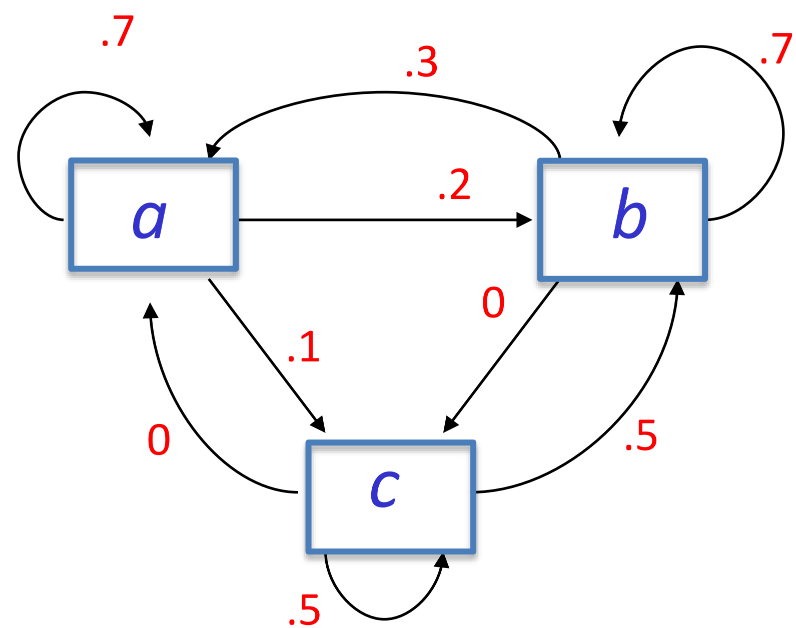
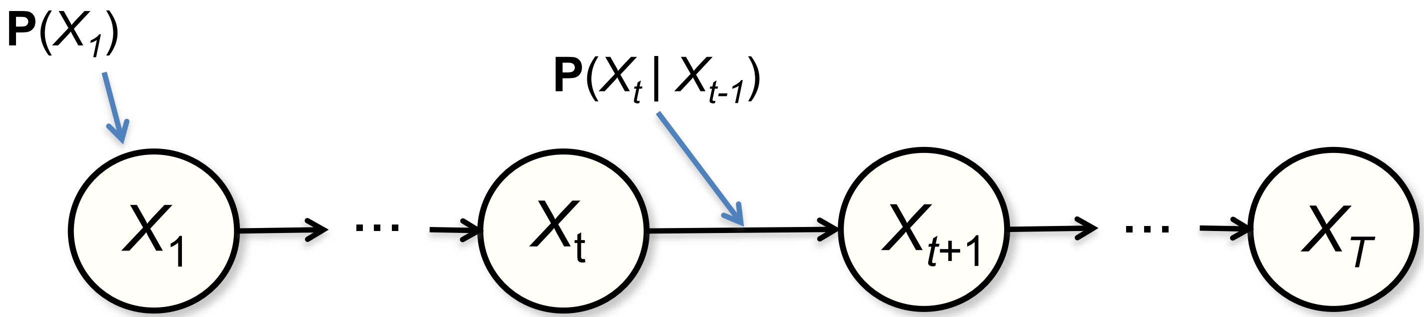


Illustration dans le cas d'une **chaîne infinie** (flux de symboles)

Exemple de chaîne: *ccbbbbaaaaabaabacbabaaa*

Probabilité de générer une séquence d'états



produit des probabilités, une pour chaque terme de la séquence

$$P(X_{1:t}) = P(X_1)\ \Pi_{i=2}^{t}\ P(X_i \mid X_{i-1})$$

distribution initiale de probabilités

une séquence d'états, allant du temps 1 au temps *t*

modèle de transition

Modèle de Markov caché

- Dans un **modèle de Markov caché** (***hidden Markov model*** ou **HMM**):
  - il y a des **variables cachées** $\boldsymbol{X_t}$ et des **variables d'observation** $\boldsymbol{E_t}$, toutes les deux discrètes
  - la chaîne de Markov est sur les variables cachées X_t
  - le symbole observé (émis) $E_t=e_t$ dépend uniquement de la variable cachée actuelle X_t

Illustration

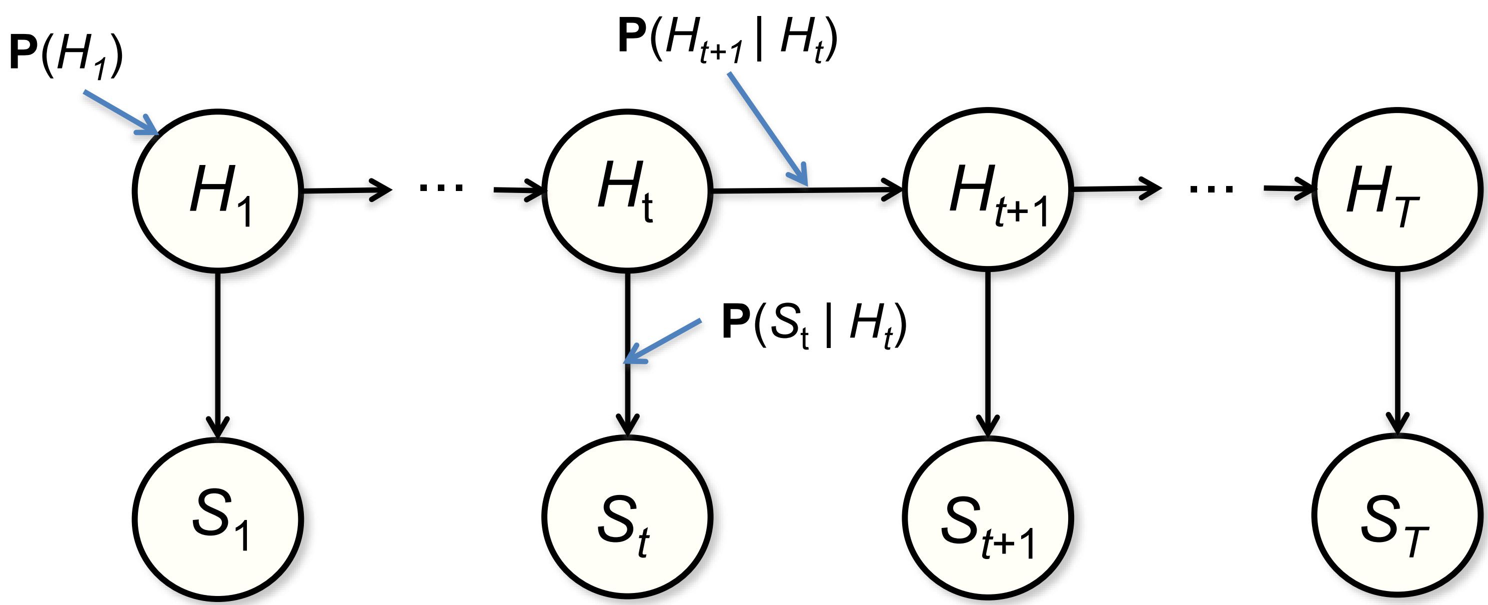
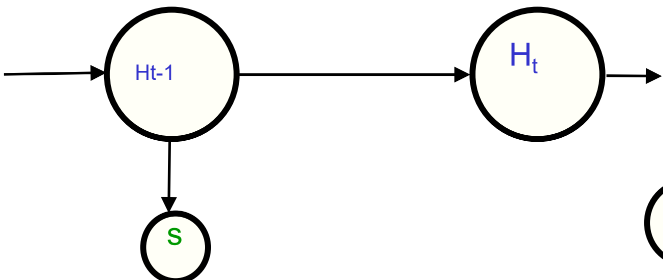


Illustration dans le cas d'une **chaîne finie**

Illustration

H_{t-1}	$P(H_t=i\|H_{t-1})$	$P(H_t=j\|H_{t-1})$	$P(H_t=k\|H_{t-1})$
i	.7	.2	.1
j	.3	.7	0
k	0	.5	.5



H	P(S=a|H)	P(S=b|H)	P(S=c|H)
i	.1	.3	.6
j	.4	.6	0
k	0	.2	.8

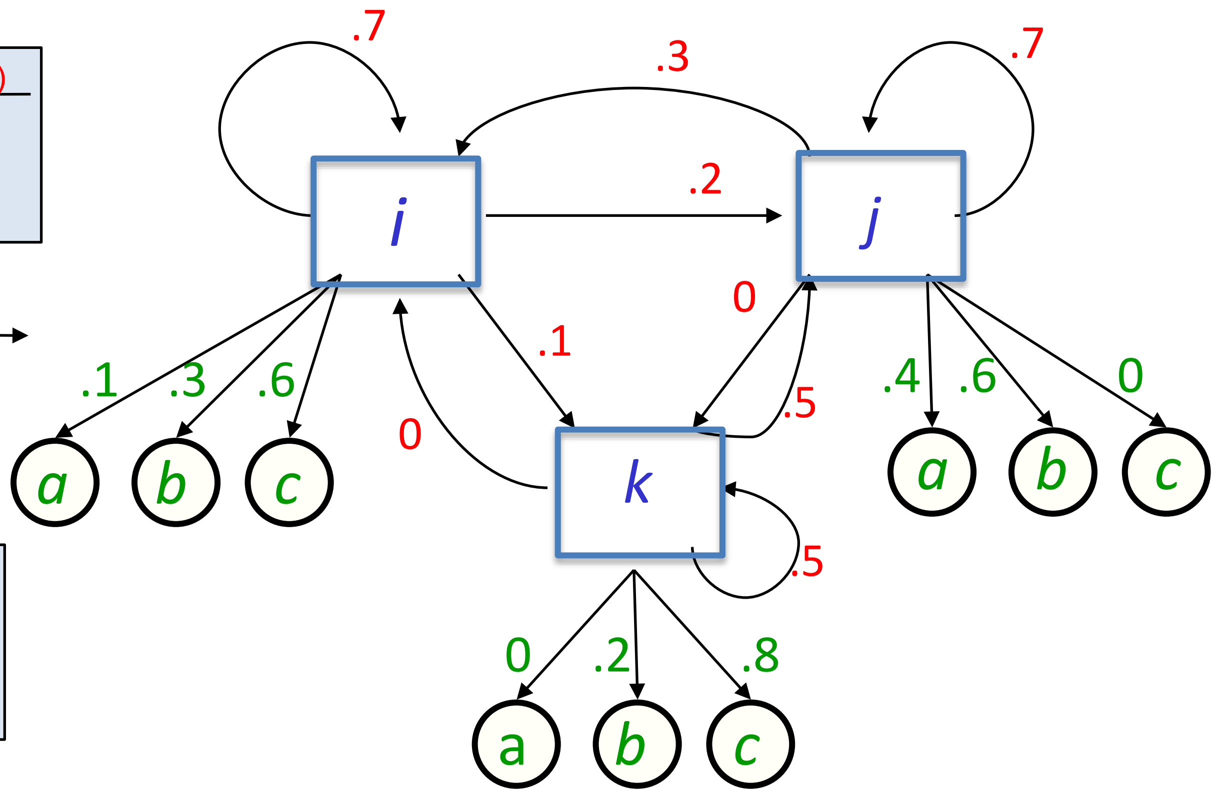
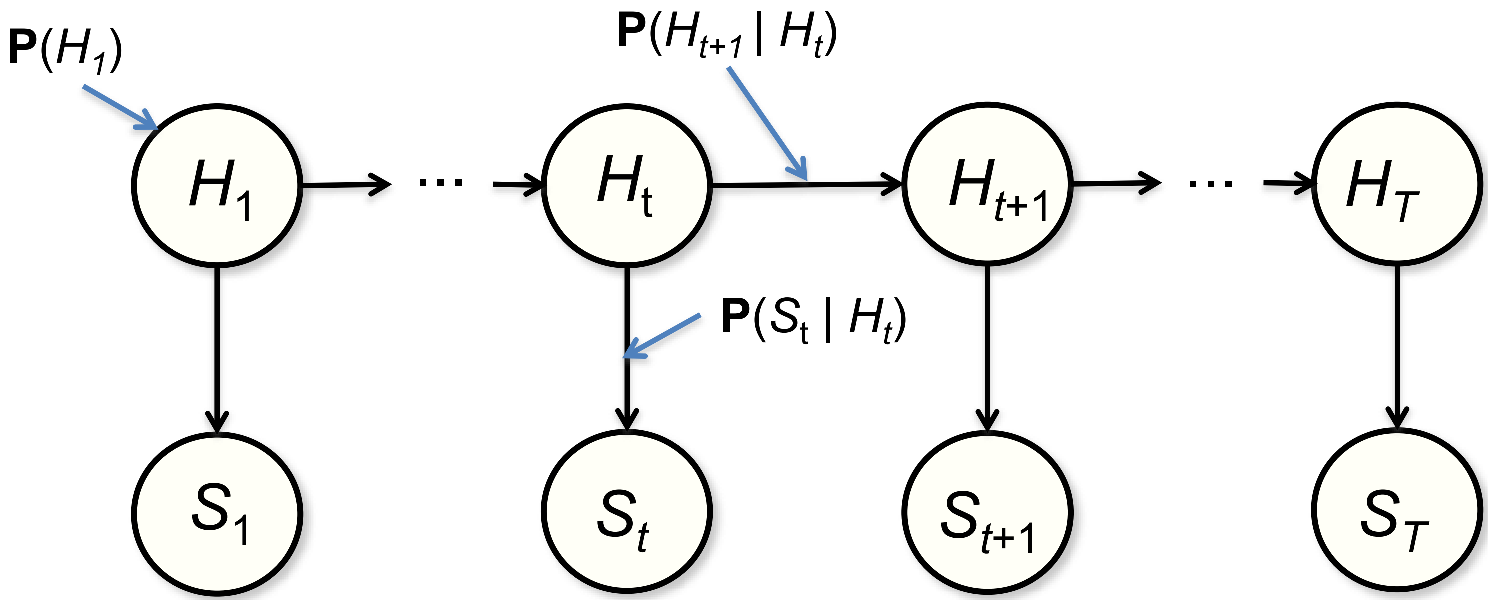


Illustration dans le cas d'une **chaîne infinie**, avec visualisation des valeurs de la variable cachée et la variable d'observation

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

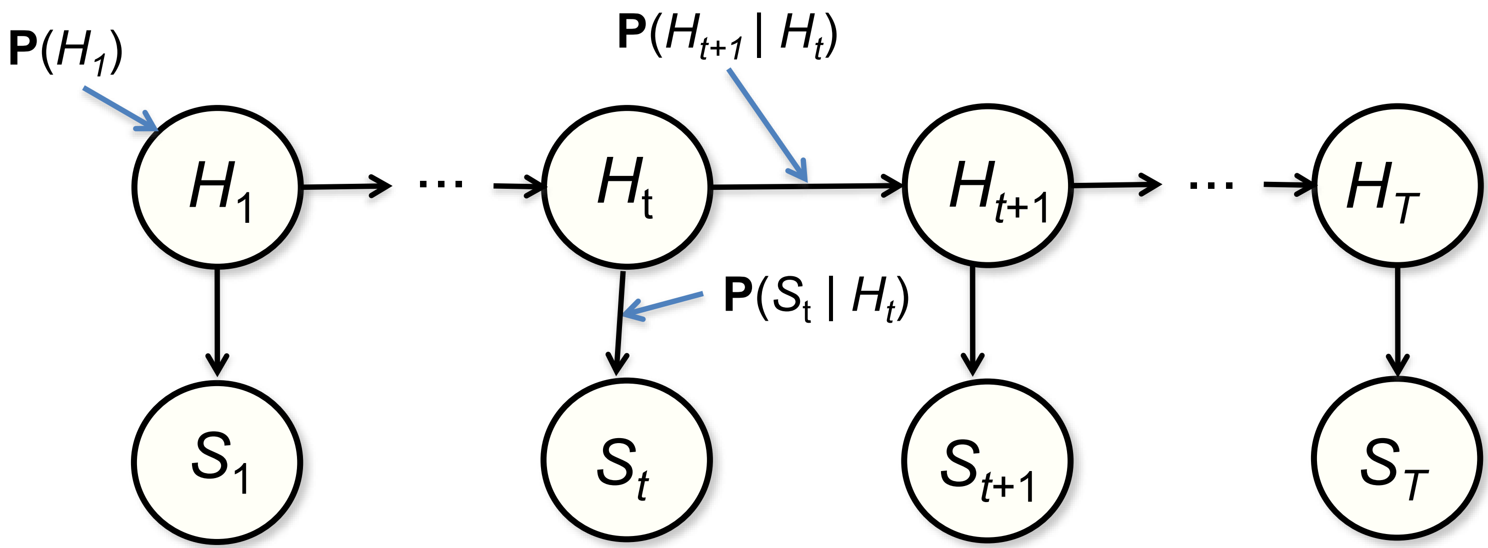
Probabilité de générer une séquence cachée et une séquence visible



La distribution initiale (à priori) $P(H_1)$ est donnée.

$$\underbrace{P(H_{1:t}, E_{1:t})}_{\text{séquence d'états cachés et d'observations}} = P(H_1)\, \Pi_{i=2}^{t} \underbrace{P(H_i \mid H_{i-1})}_{\text{modèle dynamique}}\, \underbrace{P(S_i \mid H_i)}_{\text{modèle d'observation}}$$

Filtrage avec un HMM



- Nous avons déjà vu un algorithme de programmation dynamique pour le filtrage qui itère sur la probabilité conditionnelle $P(H_K = k \mid S_{1:T} = s_{1:T})$

- Nous voyons maintenant une programmation dynamique un peu différente, mais équivalente, qui itère sur la probabilité conjointe $P(S_{1:T} = e_{1:T}, S_K = k)$.

Filtrage avec un HMM

- Par définition de la probabilité conditionnelle et par marginalisation:

$$\boldsymbol{P(H_K = k \mid E_{1:T} = e_{1:T}) = P(E_{1:T} = e_{1:T}, H_K = k) \;\; / \textstyle\sum_i P(E_{1:T} = e_{1:T}, H_K = i)}$$

- On calcule d'abord $P(E_{1:T} = e_{1:T}, H_K = k)$ par **programmation dynamique** :
  - Notons $\alpha(i,t) = P(S_{1:t}=s_{1:t}, H_t = i)$
  - Récursivement, on a:

$$\begin{aligned}
\boldsymbol{\alpha(i,t+1)} &= P(S_{1:t+1}=s_{1:t+1}, H_{t+1} = i) \\
&= \textstyle\sum_j P(S_{1:t+1}=s_{1:t+1}, H_t = j, H_{t+1} = i) \\
&= \textstyle\sum_j P(S_{1:t}=s_{1:t}, H_t = j, H_{t+1} = i, S_{t+1}=s_{t+1}) \\
&= P(S_{t+1}= s_{t+1} \mid H_{t+1} = i) \textstyle\sum_j P(H_{t+1} = i \mid H_t = j)\, P(S_{1:t}=s_{1:t}, H_t = j) \\
&= \underbrace{P(S_{t+1}= s_{t+1} \mid H_{t+1} = i)}_{\text{modèle sensoriel}} \textstyle\sum_j \underbrace{P(H_{t+1} = i \mid H_t = j)}_{\text{modèle dynamique}} \boldsymbol{\alpha(j,t)}
\end{aligned}$$

récursion

  - On a les valeurs initiales

$$\alpha(i,1) = P(S_1=s_1, H_1 = i) = P(S_1=s_1 \mid H_1 = i)\, P(H_1 = i) \quad \forall i$$

- Une fois le tableau α calculé, on peut facilement faire du filtrage:

$$\boldsymbol{P(H_K = k \mid S_{1:T} = s_{1:T}) = \alpha(k,T) \,/ \textstyle\sum_i \alpha(i,T)}$$

Exemple - Programmation dynamique avant

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

Modèle d'observation

	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

Modèle 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: $\alpha(i,1) = P(S_1=s_1, H_1 = i) = P(S_1=s_1 \mid H_1 = i)\, P(H_1 = i)$

Exemple - Programmation dynamique avant pour

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

Modèle d'observation

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

Modèle 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: $\alpha(0,1) = P(S_1=0 \mid H_1 = 0)\, P(H_1 = 0) = 0.9 \times 0.5 = 0.45$

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

Modèle 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: α(0,2) = 0.1 (0.3 x 0.45 + 0.4 x 0.1) = 0.0175

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

Modèle 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: α(1,2) = 0.8 (0.7 x 0.45 + 0.6 x 0.1) = 0.3

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique avant pour un HMM

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

Modèle d'observation

	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

Modèle 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)...

Exemple - Filtrage avec un HMM

- 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

$$P(H_4 = 0 \mid S_1=0, S_2=1, S_3=0, S_4=0) = \frac{P(H_4 = 0, S_1=0, S_2=1, S_3=0, S_4=0)}{\sum_i P(H_4 = \mathrm{i}, S_1=0, S_2=1, S_3=0, S_4=0)}$$

$$= \alpha(0,4) \,/\, (\, \alpha(0,4) + \alpha(1,4)\,)$$

$$= 0.04427 \,/\, (0.04427 + 0.02039)$$

$$\approx 0.6847$$

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

$$\approx 0.3153$$

Probabilité de générer une séquence visible

- Un calcul naïf basée sur la sémantique d'un réseau bayésien serait:

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

- Ce serait inefficace: il y a un nombre exponentiel de séquences cachés possibles (la même séquence de sortie peut être produite par plusieurs séquences cachées différentes).

- Les tableaux α(i,t) nous donne une façon plus efficace (O(n)).
- Rappelons-nous que $\alpha(i,t) = P(S_{1:t}=s_{1:t}, H_t = i)$
- Ainsi:

$$\boldsymbol{P(S_{1:T}=s_{1:T})} = \sum_j P(S_{1:\mathrm{T}}=s_{1:\mathrm{T}}, \mathrm{H}_T = j) = \boldsymbol{\sum_{\mathrm{j}} \alpha(\mathrm{j},\mathrm{T})}$$

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
  - $\underset{\text{mots}}{\text{argmax}}\ P(\textit{mots} \mid \textit{signal}) = \underset{\text{mots}}{\text{argmax}}\ \alpha\ P(\textit{mots}, \textit{signal})$

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**

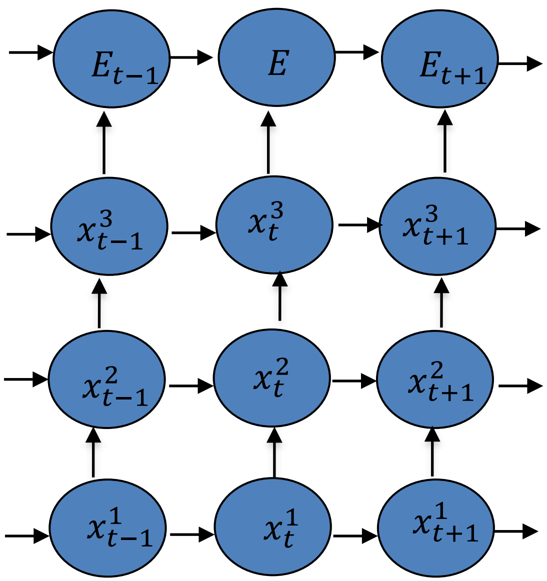
- Chaîne cachée: les mots
- Chaîne observée: le signal

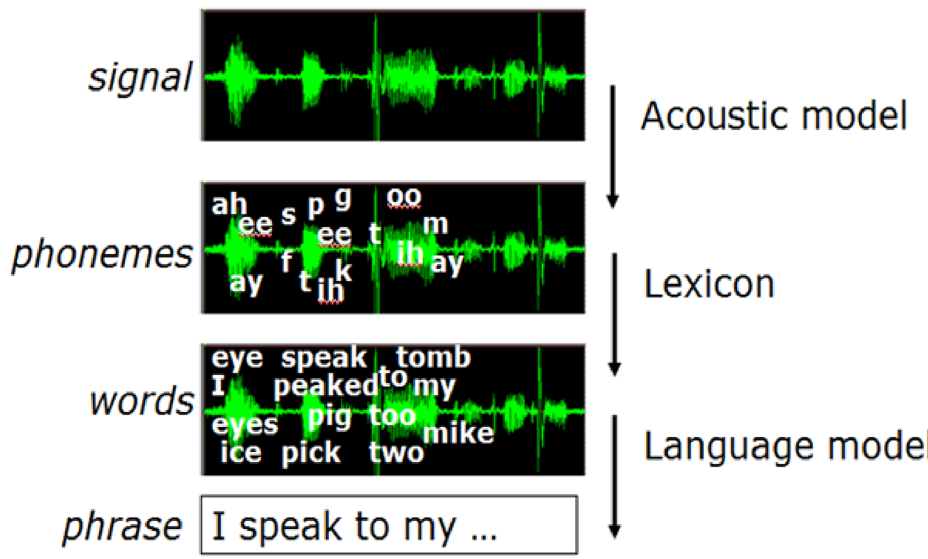
Phones et phonèmes

- 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* »)

Application - Reconnaissance vocale

- **Modèle:**
  - E_t sont les éléments du signal sonore
  - X_t sont les phrases et les mots prononcés ainsi que les phonèmes





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
⋮	⋮	⋮	⋮	⋮	⋮

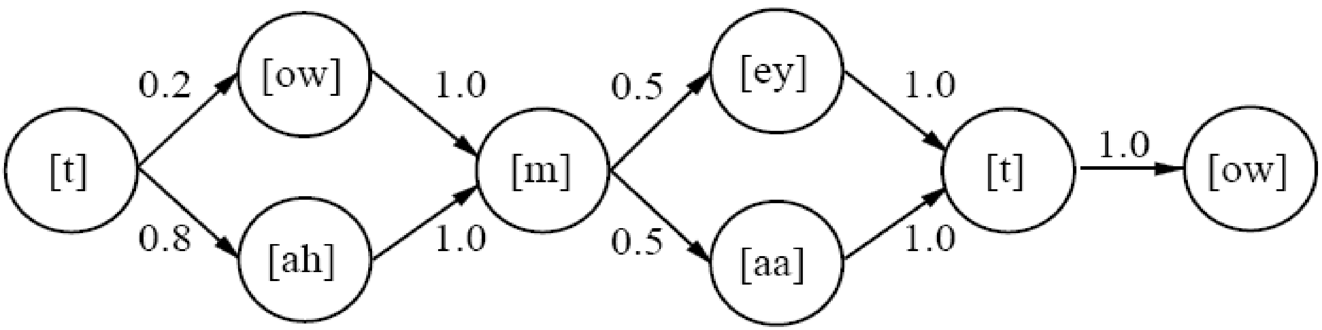
Modèle acoustique

- 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 apprises automatiquement d'un corpus
- Les probabilités sont aussi apprises



Apprendre les tables des probabilités conditionnelles

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

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

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

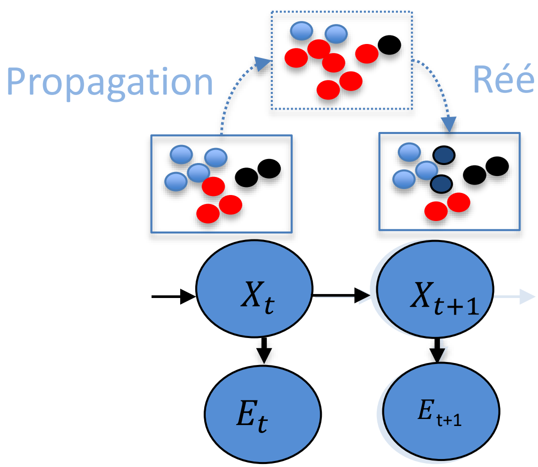
$$P(B{=}b \mid A{=}a) = \frac{1 + \sum_{\text{chaînes}} \text{freq}(a,b)}{\text{Nb. symboles} + \sum_{\text{chaînes}} \text{freq}(a)}$$

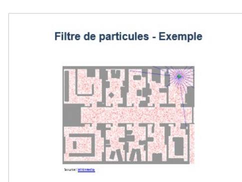
Conversation continue

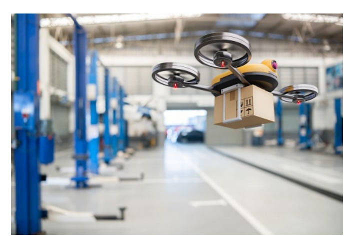
- Dans une conversation continue on doit tenir compte de la corrélation des mots adjacents et non juste la reconnaissance d’un mot isolé.
- Aujourd’hui, les HMM ont été dépassés par les réseaux de neurones pour la reconnaissance vocale.

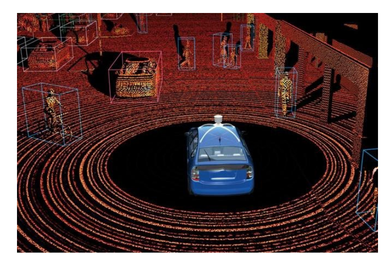
Filtre de particules

- **Un filtre de particule** est un filtrage par inférence approximative
  - Au temps t, on a un incertitude sur la valeur de la variable X
  - On maintient une population **N** des valeurs (états) probables. Chaque valeur probable est représentée par une particule. On a donc N particules.
  - À chaque temps *t*, on met à jour la population de particules, en échantillonnant la population courante, tenant compte du modèle dynamique (**propagation par échantillonnage**) et du modèle de transition (**rééchantillonnage pondérée**)
- Utilisé notamment pour le *tracking* (estimer la position d'un objet mobile).









Source: [Udacity]

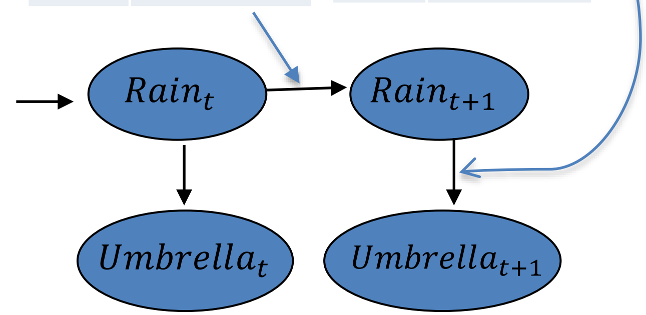
Filtre de particules - Algorithme

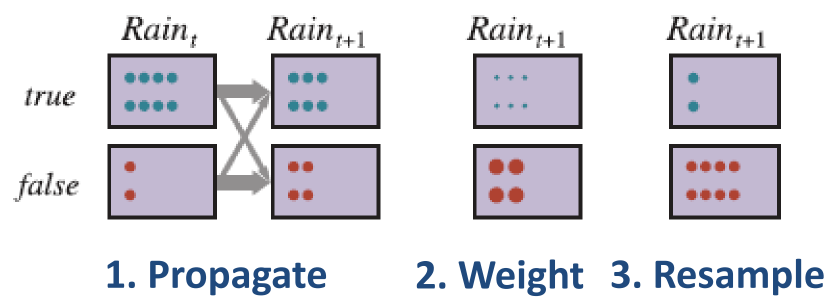
- Générer une population de N particules selon $\boldsymbol{P}(X_0)$
- À chaque transition $X_t \rightarrow X_{t+1}$:

  1. **Propager** la population en échantillonnant la prochaine valeur x_{t+1} étant donné la valeur courante x_t, selon le modèle de transition $\boldsymbol{P}(X_{t+1} \mid X_t)$
  2. Étant donné la nouvelle observation e_{t+1}, **pondérer** chaque particule x_{t+1} par $\boldsymbol{P}(e_{t+1} \mid x_{t+1})$
     - c.-à-d., pondérer chaque particule par la vraisemblance qu'elle accorde à la nouvelle observation
  3. **Rééchantillonner** (avec replacement) la population pour générer N nouveaux particules
     - La probabilité qu'un échantillon particulier soit choisi est proportionnelle à son poids.

R_{t-1}	$P(R_t \mid R_{t-1})$
V	0.7
F	0.3

R_t	$P(U_t \mid R_t)$
V	0.9
F	0.2





Au temps *t+1*,
$\neg$ *Umbrella* est observé

Filtre de particules - Algorithme

function PARTICLE-FILTERING($\mathbf{e}$, N, dbn) **returns** a set of samples for the next time step
  inputs: $\mathbf{e}$, the new incoming evidence
    N, the number of samples to be maintained
    dbn, a DBN defined by $\mathbf{P}(\mathbf{X}_0)$, $\mathbf{P}(\mathbf{X}_1 \mid \mathbf{X}_0)$, and $\mathbf{P}(\mathbf{E}_1 \mid \mathbf{X}_1)$
  persistent: S, a vector of samples of size N, initially generated from $\mathbf{P}(\mathbf{X}_0)$
  local variables: W, a vector of weights of size N

  for i = 1 to N **do**
    $S[i] \leftarrow$ sample from $\mathbf{P}(\mathbf{X}_1 \mid \mathbf{X}_0 = S[i])$ // *step 1*
    $W[i] \leftarrow \mathbf{P}(\mathbf{e} \mid \mathbf{X}_1 = S[i])$ // *step 2*
  $S \leftarrow$ WEIGHTED-SAMPLE-WITH-REPLACEMENT(N, S, W)   // *step 3*
  return S

Consistence de l'algorithme

On peut prouver (par induction) que l'algorithme est consistent, c.-à.d-., que lorsque le nombre d'échantillons *N* tend vers l'infini, il se rapproche de la probabilité de filtrage.

- Cas de base: On suppose que la population initiale est consistante avec $P(X0)$.

Notons $N(x_t | e_{1:t})$, le nombre de particules ayant la valeur x_t après avoir observé $e_{1:t}$.

- Cas inductif: Supposons que $N(x_t | e_{1:t})$ / N = $P(x_t | e_{1:t})$

  Montrons que $N(x_{t+1} | e_{1:t+1})$ / N = $P(x_{t+1} | e_{1:t+1})$

Consistance de l'algorithme

- Se rappeler que $N(x_t|e_{1:t})$ est le nombre de particules ayant la valeur x_t après avoir observé $e_{1:t}$.
- On suppose que $N(x_t|e_{1:t})$ / N = $P(x_t|e_{1:t})$ -- Notre hypothèse inductive
- On veut montrer que $N(x_{t+1}|e_{1:t+1})$ / N = $P(x_{t+1}|e_{1:t+1})$
  - Partant de $N(x_t|e_{1:t})$, l'étape de propagation va donner

    $$N(x_{t+1}|e_{1:t}) = \Sigma_{xt}\, P(x_{t+1}|x_t)\, N(x_t|e_{1:t})$$

  - Après avoir observé e_{t+1} l'étape de pondération donne une population de particules dont le poids total est

    $$W(x_{t+1}|e_{1:t}) = P(e_{t+1}|x_{t+1})\, N(x_{t+1}|e_{1:t})$$

  - Après l'étape de rééchantillonnage, vu que chaque échantillon est choisi proportionnellement à son poids, le poids total des échantillons ayant la valeur x_{t+1} est proportionnel à leur poids total avant le rééchantillonnage, c.-à-d.,

    $$N(x_{t+1}|e_{1:t+1})\, /\mathrm{N} = \alpha\, W(x_{t+1}|e_{1:t})$$

  -

Consistance de l'algorithme, suite

- Cas inductif: Supposons que $N(x_t|e_{1:t})$ / N = $P(x_t|e_{1:t})$

  Montrons que $N(x_{t+1}|e_{1:t+1})$ / N = $P(x_{t+1}|e_{1:t+1})$

  - Partant de $N(x_t|e_{1:t})$, l'étape de propagation va donner

    $N(x_{t+1}|e_{1:t}) = \Sigma_{xt}\, \mathrm{P}(\mathrm{x}_{t+1}|\mathrm{x}_t)\, N(x_t|e_{1:t})$

  - Après avoir observé e_{t+1} l'étape de pondération donne que le poids total des particules est

    $W(x_{t+1}|e_{1:t}) = \mathrm{P}(\mathrm{e}_{t+1}|\mathrm{x}_{t+1})\, N(x_{t+1}|e_{1:t})$

  - Après l'étape de rééchantillonnage, vu que chaque échantillon est choisi proportionnellement à son poids, le poids total des échantillons ayant la valeur x_{t+1} est proportionnel à leur poids total avant le rééchantillonnage, c.-à-d.,

    $N(x_{t+1}|e_{1:t+1})$ /N = $\alpha\, W(x_{t+1}|e_{1:t})$

Consistance de l'algorithme, suite …

- Cas inductif : En supposant que $N(x_t|e_{1:t})$ / N = $P(x_t|e_{1:t})$

  On veut montrer $N(x_{t+1}|e_{1:t+1})$ / N = $P(x_{t+1}|e_{1:t+1})$

  - On était rendu à $N(x_{t+1}|e_{1:t+1})$ /N = $\alpha\, W(x_{t+1}|e_{1:t})$

$N(x_{t+1}|e_{1:t+1})$ /N = $\alpha\, W(x_{t+1}|e_{1:t})$

$= \alpha\, P(e_{t+1}/x_{t+1})\, N(x_{t+1}|e_{1:t})$ -- selon la définition de W

$= \alpha\, P(e_{t+1}|x_{t+1})\, \Sigma_{xt}\, P(\mathrm{x}_{t+1}|\mathrm{x}_t)\, N(x_t|e_{1:t})$

selon l'équation de $N(x_{t+1}|e_{1:t})$ à la slide précédente

$= \alpha^*N^*P(\mathrm{e}_{t+1}|\mathrm{x}_{t+1})\, \Sigma_{xt}\, P(x_{t+1}|\mathrm{x}_t)\, P(x_t|e_{1:t})$

Par hypothèse inductive : $N(x_t|e_{1:t})$ / N = $P(x_t|e_{1:t})$

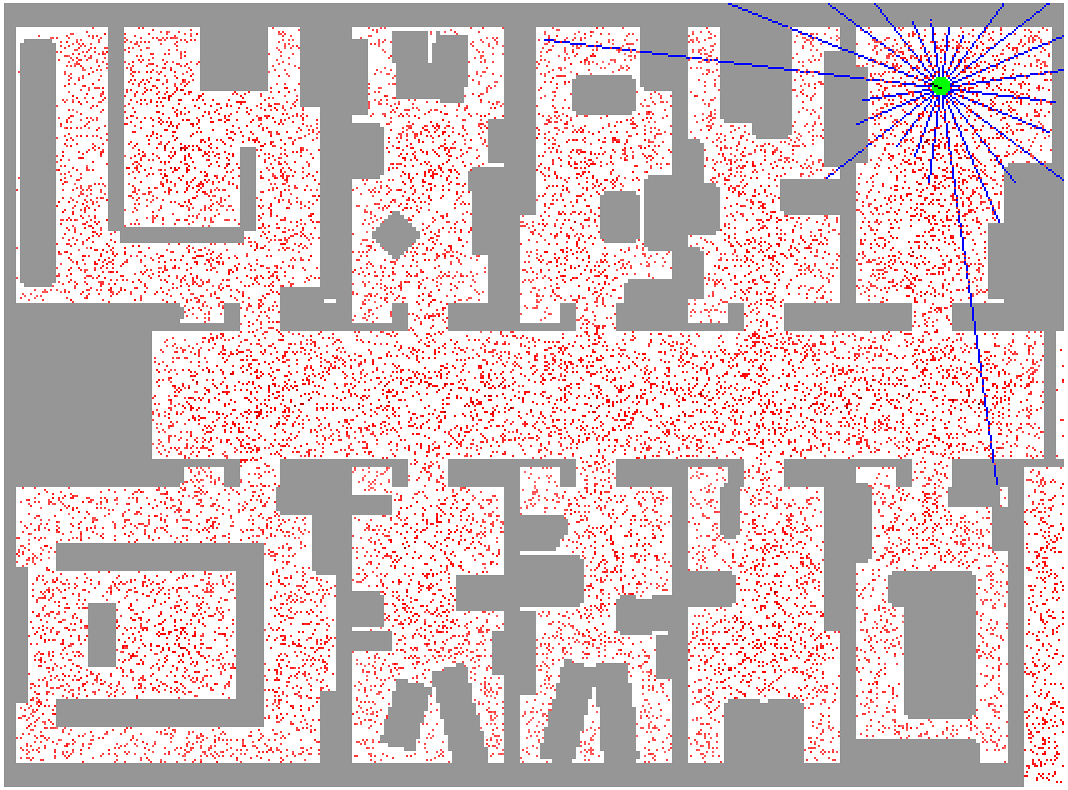
$= \alpha' P(e_{t+1}|\mathrm{x}_{t+1})\, \Sigma_{xt}\, \mathrm{P}(x_{t+1}|\mathrm{x}_t)\, P(x_t|e_{1:t})$ – en posant α'= α*N

$= \mathrm{P}(x_{t+1}|e_{t+1})$

Selon l'équation de programmation dynamique pour le filtrage

  - CQFD

Filtre de particules - Exemple



Source: Wikimedia

Filtre de Kalman

- **Dans un HMM,** la variable cachée est discrète et change dynamiquement ses valeurs en suivant une chaîne de Markov et sa distribution de probabilité conditionnelle (modèle dynamique)

- Un **Filtre de Kalman** est la version continue d'un HMM. La variable est continue, avec une distribution gaussienne.

- La méthode est utilisée pour le suivi d'objets (*tracking*).

- Pas couverte dans ce cours. Voir section 14.4.

Résumé

- Un **réseau bayésien dynamique** (RBD) permet de tenir compte de la nature séquentielle d'un environnement

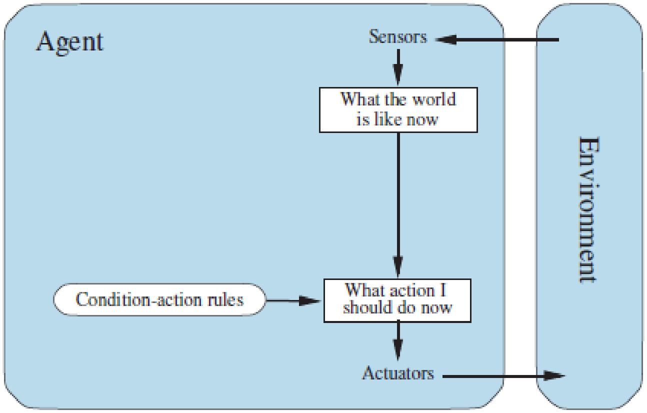
- Un **modèle de Markov caché** (HMM) est un cas particulier de RBD avec
  - une seule variable cachée $X_\mathrm{t} = \{H_t\}$ et une seule variable observée $E_\mathrm{t} = \{S_t\}$
  - les variables H_t et S_t sont discrètes

- Il existe des procédures de programmation dynamique efficaces dans un HMM pour faire de l'inférence : **filtrage, prédiction**, **lissage**, **explication la plus plausible.** Le cours couvre seulement le filtrage.

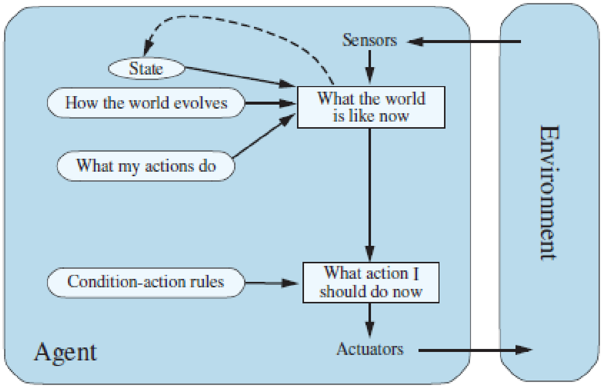
- Un filtre de particule est une méthode approximative de filtrage pour un RBD. Très utilisée pour le suivi des objets en mouvement.

Raisonnement probabiliste pour quel type d'agents?

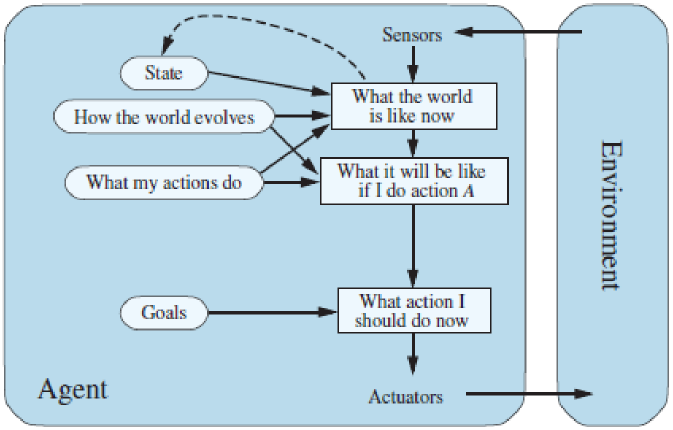
Simple reflex



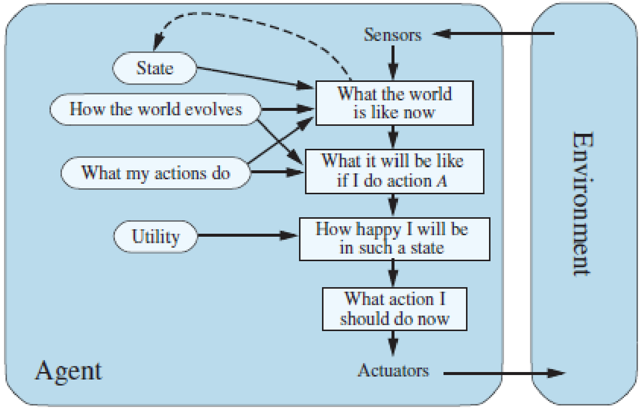
Model-based reflex



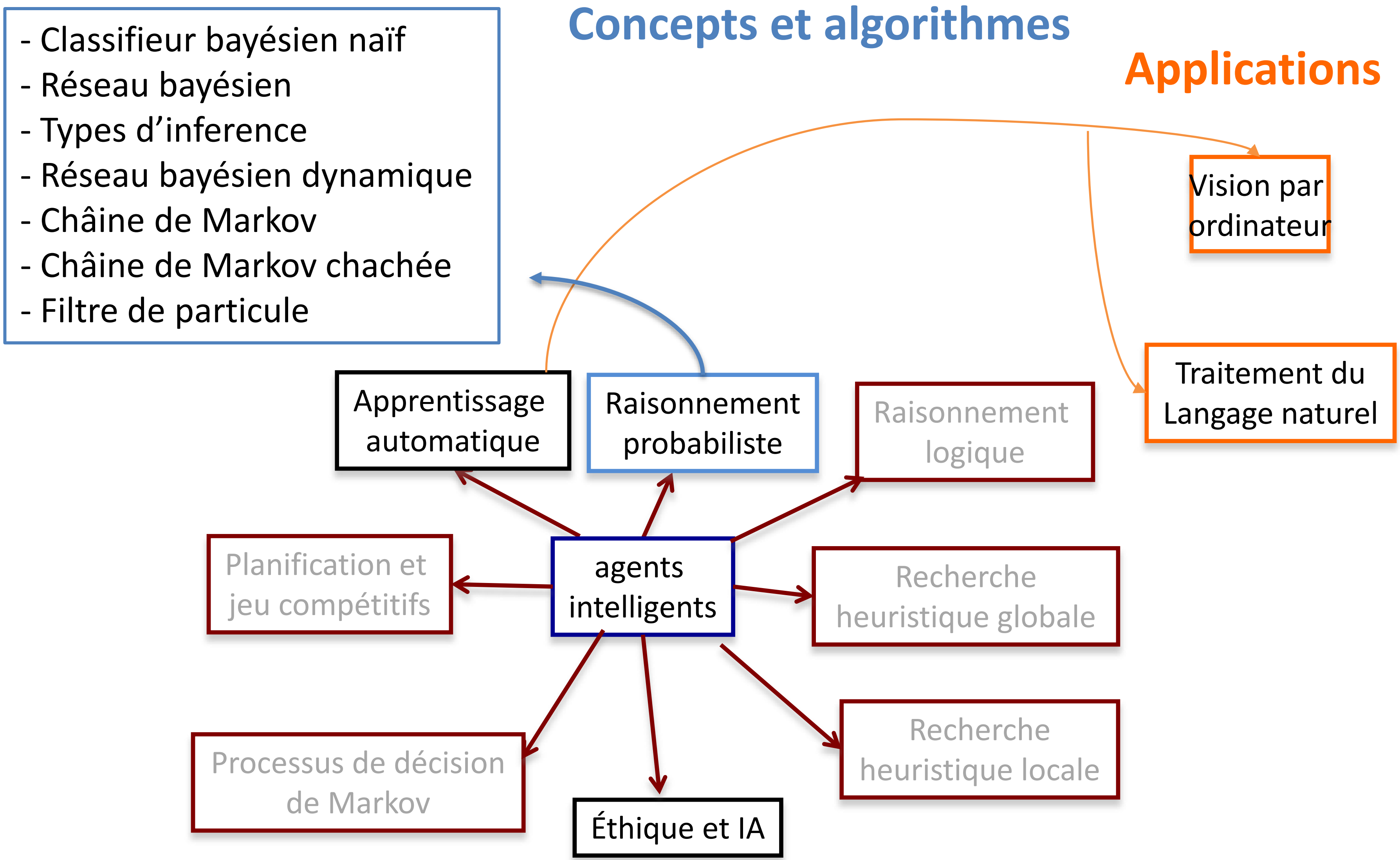
Goal-based



Utiliy-based



Sujets couverts par le cours



Vous devriez être capable de...

- Distinguer les différents types d'inférence probabiliste temporelle:
  - filtrage
  - prédiction
  - lissage
  - explication la plus plausible
- Décrire ce qu'est un modèle de Markov caché
  - Définir et calculer le tableau α par la programmation dynamique avant
  - Appliquer le tableau α pour le filtrage et le calcul d'une séquence visible
- Décrire et appliquer un filtre de particules
- Note: TP #3 porte sur ces sujets.

LA PARTIE SUIVANTE N'EST PAS COUVERTE PAR LES EXAMENS

Prédiction avec 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
  - Notons $\pi(i,k) = P(H_{t+k} = i | S_{1:t}=s_{1:t})$
  - Récursivement:

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

  - On a les valeurs initiales

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

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

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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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			

- **Voir la suite dans les diapos en ligne (cachés)**

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

	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

Modèle 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))

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

	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

Modèle 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): $\pi(0, 1) = P(H_5 = 0 \mid H_4 = 0)\ \pi(0,0) + P(H_5 = 0 \mid H_4 = 1)\ \pi(1,0)$

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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): $\pi(0, 1) = 0.3 \times 0.68466 + 0.4 \times 0.31534 = 0.33154$

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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): $\pi(1, 1) = P(H_5 = 1 \mid H_4 = 0)\ \pi(0,0) + P(H_5 = 1 \mid H_4 = 1)\ \pi(1,0)$

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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

Exemple - Prédiction avec un HMM

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

Modèle d'observation

	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

Modèle 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)...

Exemple - Prédiction avec un HMM

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

Modèle d'observation

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

Modèle 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

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

Lissage pour un HMM

- Pour faire le lissage, en plus du tableau des α(i,t) généré par un balayage de gauche à droite, nous avons besoin d'un tableau analogue, β(i,t), généré de de droite à gauche
  - Notons $\beta(i,t) = P(S_{t+1:T}=s_{t+1:T} \mid H_t = i)$
  - Récursivement on a:
    $\beta(i,t\text{-}1) = P(S_{t:T}=s_{t:T} \mid H_{t-1} = i)$
    $= \sum_j P(S_{t:T}=s_{t:T}, H_t = j \mid H_{t-1} = i) = \sum_j P(S_{t+1:T}=s_{t+1:T}, S_t=s_t, H_t = j \mid H_{t-1} = i)$
    $= \sum_j P(S_t= s_t \mid H_t = j)\, P(H_t = j \mid H_{t-1} = i)\, P(S_{t+1:T}=s_{t+1:T} \mid H_t = j)$
    $= \sum_j P(S_t= s_t \mid H_t = j)\, P(H_t = j \mid H_{t-1} = i)\, \beta(j,t)$
  - On a les valeurs initiales $\beta(i,T) = 1\ \forall i$

- Remarquez le tableau β calculé donne aussi une autre façon de calculer la probabilité d'une séquence de sortie:
  $P(S_{1:T}=s_{1:T}) = \sum_j P(S_{1:T}=s_{1:T}, H_1 = j)$
  $= \sum_j P(S_{2:T}=s_{2:T} \mid H_1 = j)\, P(S_1=s_1 \mid H_1 = j)\, P(H_1 = j)$
  $= \sum_j \beta(j,1)\, P(S_1=s_1 \mid H_1 = j)\, P(H_1 = j)$

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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(1,3) = $P(S_4=0 \mid H_4=0)\ P(H_4=0 \mid H_3=1)\ \beta(0,4)$
  $+\ P(S_4=0 \mid H_4=1)\ P(H_4=1 \mid H_3=1)\ \beta(1,4)$

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

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

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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Programmation dynamique à rebours pour un HMM

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

Modèle d'observation

	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

Modèle 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*

Exemple - Lissage avec un HMM

- 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

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

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

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

$$\approx 0.04813$$

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

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

$$\approx 0.95186$$

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**
  - Notons $\alpha^*(i,t) = P(S_{1:t}=s_{1:t}, H_{1:t-1} = h^*_{1:t-1}, H_t = i)$
  - Recursivement, on a:

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

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

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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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: $\alpha^*(i,1) = P(S_1=s_1, H_1 = i) = P(S_1=s_1 \mid H_1 = i)\, P(H_1 = i)$

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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: $\alpha^*(0,1) = P(S_1=0 \mid H_1 = 0)\, P(H_1 = 0) = 0.9 \times 0.5 = 0.45$

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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: α*(0,2) = 0.1 max{ 0.3 x 0.45,0.4 x 0.1} = 0.0135

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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: α*(1,2) = $P(S_2=1 | H_2=1)$ max{$P(H_2 = 1 | H_1 = 0)$ α*(0,1), $P(H_2 = 1 | H_1 = 1)$ α*(1,1)}

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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)

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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	
1	0.1	0.252		

- récursion: α*(0,3) = 0.9 max{ 0.3 x 0.0135, 0.4 x 0.252} = 0.09072

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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	0.02449
1	0.1	0.252	0.03024	0.01270

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

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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**...

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

	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

Modèle 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

- ... puis on **retrouve le chemin** qui a mené là

Exemple - Explication la plus plausible avec un HMM

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

Modèle d'observation

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

Modèle 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	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

Simuler 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 observation (émission) ne dépendra pas des autres