Hidden Markov Model
-- Probabilistic Graphical Model Perspective

Rui Li

Probabilistic Graphical Model

- **PDF Representation**
  - $P(X, Y)$
- **Inference**
  - Given Y'
  - Use $P(X \mid Y')$ to solve problems
- **Learning**
  - Given $\{(x_t, y_t)\}$
  - Fit $P(X \mid Y)$

Inference

- **MAP (Maximum A Posteriori)**

$$X^* = \arg\max_X P(X \mid Y')$$

- **Marginalization**

$$P(x_t \mid Y') \qquad \forall t$$

Henceforth, “inference”== marginalization

Marginalization

$$P(x_t = i \mid Y') = \sum_{X:x_t=i} P(X \mid Y')$$

Computational Complexity:

$x_t \in \{1,2,\ldots,100\}$ and $T = 1000$

$$|P(x_t = i \mid Y')| = 100^{999}$$

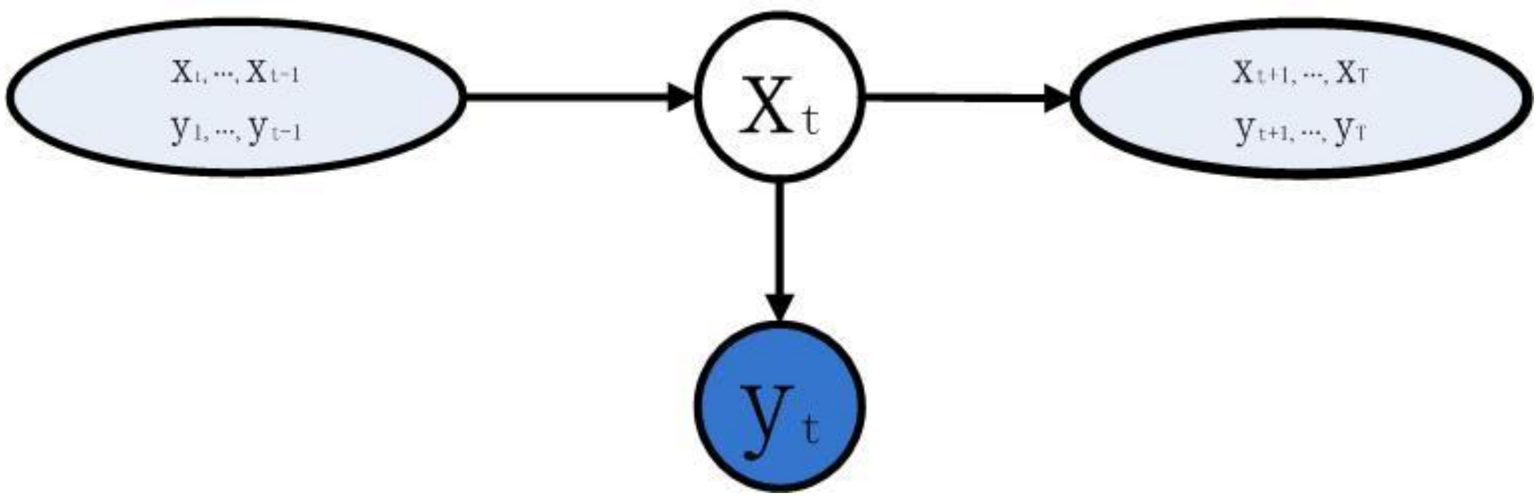
Exponentially many terms

Forward-Backward Algorithm

$$
\begin{aligned}
& P(x_t = i \mid Y) \\
&= P(x_t = i \mid y_1, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_T) \\
&= P(y_1, \ldots, y_{t-1}) P(x_t = i \mid y_1, \ldots, y_{t-1}) P(y_t \mid x_t = i) P(y_{t+1}, \ldots, y_T \mid x_t = i) \\
&= \alpha_t(i) P(y_t \mid x_t = i) \beta_t(i)
\end{aligned}
$$

Forward message: $\alpha_t(i) = P(x_t = i, y_1, \ldots, y_{t-1})$

Backward message: $\beta_t(i) = P(y_{t+1}, \ldots, y_T \mid x_t = i)$



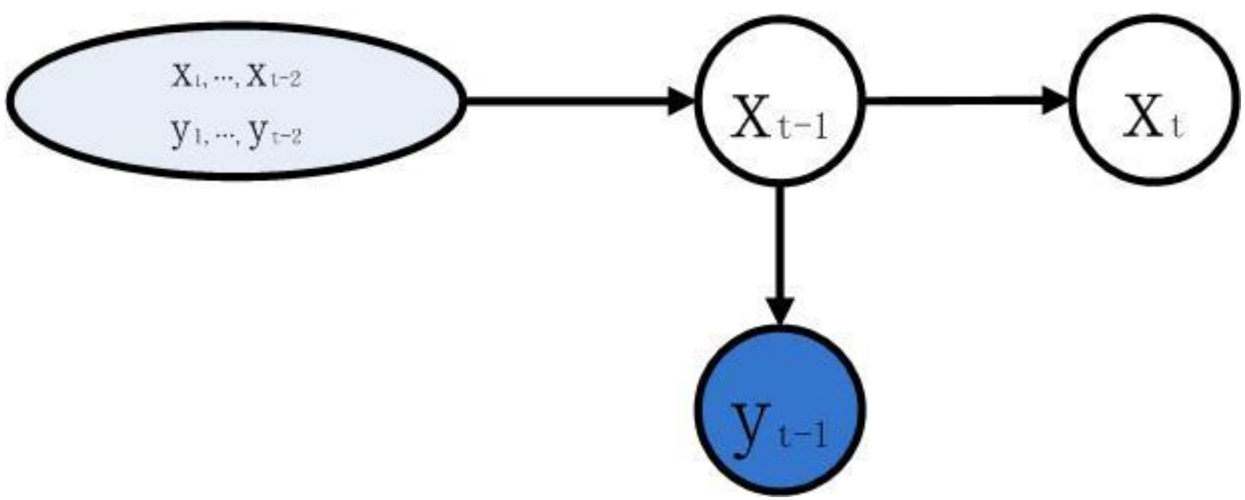
Forward-Backward Algorithm

Forward message: $\alpha_t(i) = P(x_t = i, y_1, ..., y_{t-1})$

$$= \sum_j P(x_t = i, x_{t-1} = j, y_1, ..., y_{t-2}, y_{t-1})$$

$$= \sum_j P(x_{t-1} = j, y_1, ..., y_{t-2}) P(y_{t-1} \mid x_{t-1} = j) P(x_t = i \mid x_{t-1} = j)$$

$$= \sum_j \alpha_{t-1}(j) P(y_{t-1} \mid x_{t-1} = j) P(x_t = i \mid x_{t-1} = j)$$



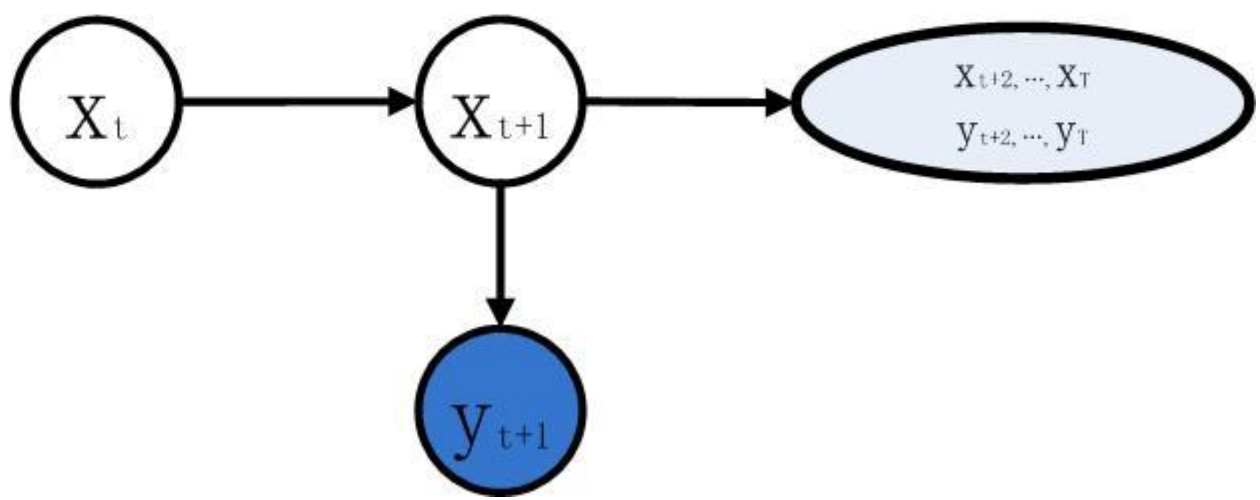
Forward-Backward Algorithm

Backward message: $\beta_t(i) = P(y_{t+1},\ldots,y_T \mid x_t = i)$

$$= \sum_j P(y_{t+1}, y_{t+2}\ldots, y_T, x_{t+1} = j \mid x_t = i)$$

$$= \sum_j P(x_{t+1} = j \mid x_t = i) P(y_{t+1} \mid x_{t+1} = j) P(y_{t+2},\ldots, y_T \mid x_{t+1} = j)$$

$$= \sum_j P(x_{t+1} = j \mid x_t = i) P(y_{t+1} \mid x_{t+1} = j) \beta_{t+1}(j)$$



Forward-Backward Algorithm

$$P(x_t = i \mid Y) = \alpha_t(i) P(y_t \mid x_t = i) \beta_t(i)$$

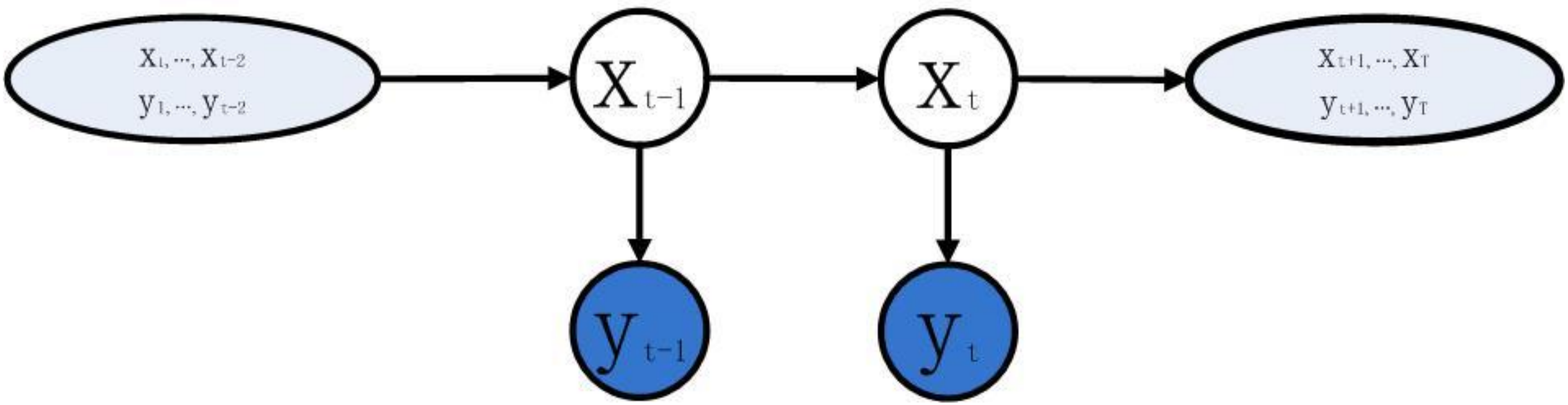
Computational Complexity:

$|\alpha_t(i)| = 100 \times 100$ and $|\beta_t(i)| = 100 \times 100$

$$|P(x_t = i \mid Y)| = 1000 \times 100^2$$

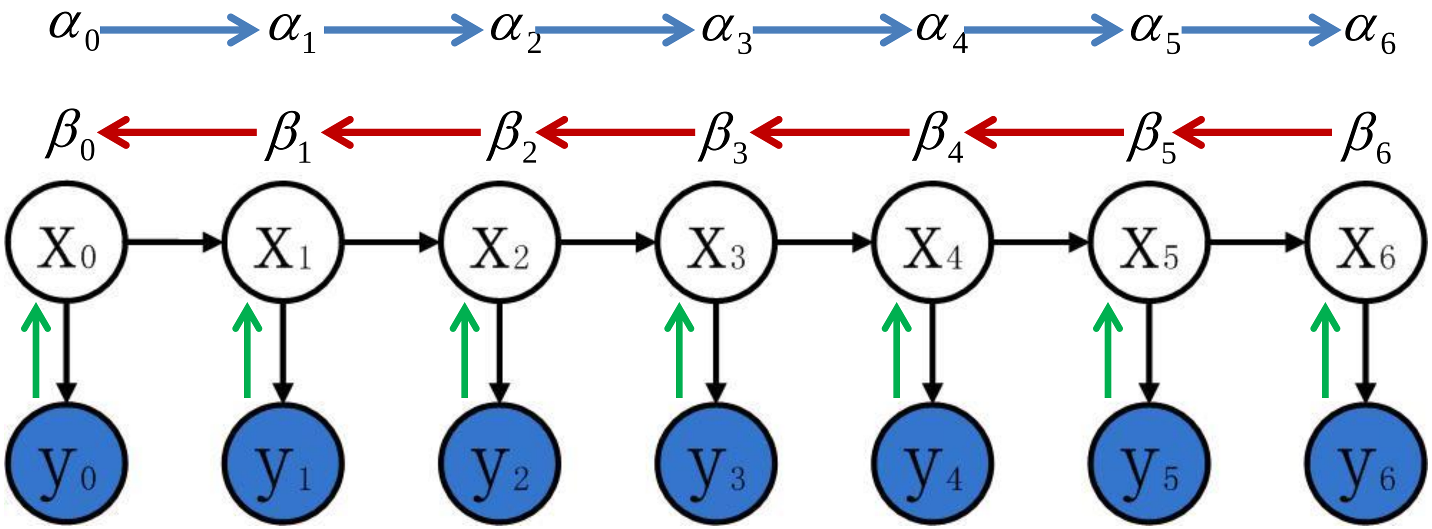
Forward-Backward Algorithm

$$P(x_{t-1} = j, x_t = i \mid Y)$$
$$= \alpha_{t-1}(j) P(y_{t-1} \mid x_{t-1} = j) P(x_t = i \mid x_{t-1} = j) P(y_t \mid x_t = i) \beta_t(i)$$



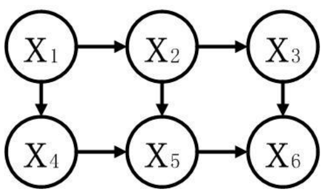
Forward-Backward Algorithm

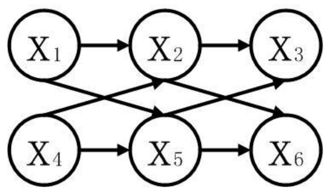
- **Message passing**

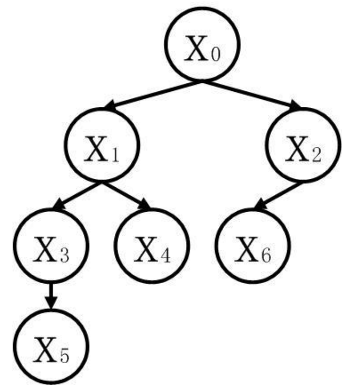


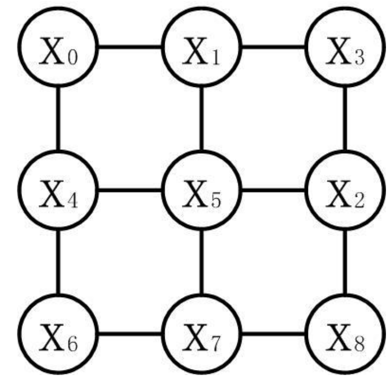
Belief Propagation Algorithm

- **Message passing**







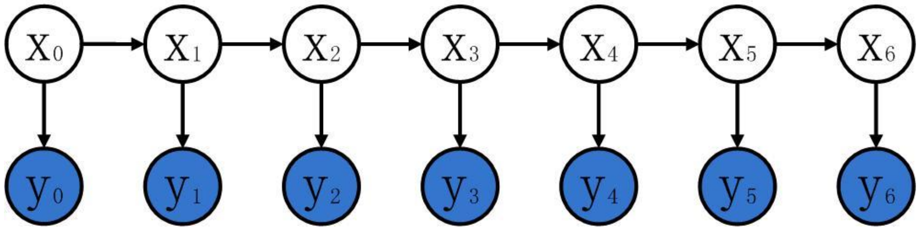


Probabilistic Graphical Model

- **PDF Representation**
  - $P(X,Y)$
- **Inference**
  - Given Y'
  - Use $P(X \mid Y')$ to solve problems
- **Learning**
  - Given $\{(x_t, y_t)\}$
  - Fit $P(X \mid Y)$

Learning

- **Fully observed learning**
  - given $\{(x_t, y_t)\}$
  - fit $P_\theta(X \mid Y)$



- **Partially observed learning**
  - given $\{y_t\}$
  - fit $P_\theta(Y)$

MAP Estimation: $\theta^* = \arg\max_\theta P_\theta(Y)$

Expectation-Maximization Algorithm

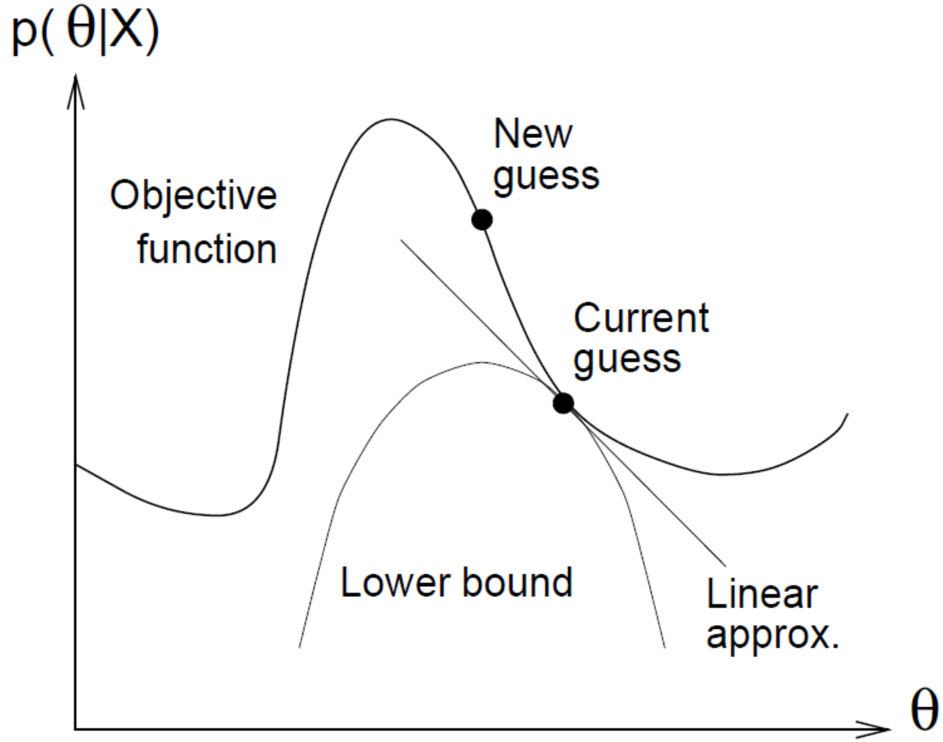
$$\theta^* = \arg\max_{\theta} P_{\theta}(Y)$$

$$= \arg\max_{\theta} P(Y,\theta) = \arg\max_{\theta} \int_X P(Y,X,\theta)dX$$

The basic idea:

- Given a guess for θ , lower-bound $P(Y,\theta)$ with a function $g(\theta, q(X))$
- EM alternates between
  - Optimize $g(\theta, q(X))$ wrt $q(X)$ given θ
  - Maximize $g(\theta, q(X))$ wrt θ given $q(X)$

Expectation-Maximization Algorithm



- What's the form of a good lower bound $g(\theta, q(X))$ given θ?
  - –**Expectation Step**
- How to maximize the bound over θ to get the next guess for θ, given $q(X)$?
  - –**Maximization Step**

Expectation Step

Given a guess for θ , lower-bound $P(Y,\theta)$ with a function $g(\theta, q(X))$

$$P(Y,\theta) = \int_X P(Y, X, \theta) dX$$

$$= \int_X [\frac{P(Y, X, \theta)}{q(X)}] \cdot q(X) dX$$

Jensen's inequality

$$\geq \boxed{\prod_X [\frac{P(Y, X, \theta)}{q(X)}]^{q(X)} = g(\theta, q(X))}$$

$$s.t. \quad \int_X q(X) dX = 1$$

Expectation Step

$$G(\theta, q) = \log g(\theta, q(X))$$

$$= \log \prod_X [\frac{P(Y, X, \theta)}{q(X)}]^{q(X)}$$

$$= \int_X \log[\frac{P(Y, X, \theta)}{q(X)}]^{q(X)} dX$$

$$= \int_X q(X) \log[\frac{P(Y, X, \theta)}{q(X)}] dX$$

$$= \boxed{\int_X [q(X) \log P(Y, X, \theta) - q(X) \log q(X)] dX}$$

The inequality is true for any q, but we want the lower bound to touch $P(Y, \theta)$ at the current guess for θ

Expectation Step

So we choose q to maximize:

$$G(\theta,q)=\int_X [q(X)\log P(Y,X,\theta)-q(X)\log q(X)]dX$$

$$s.t.\quad \int_X q(X)dX=1$$

use Lagrange multiplier:

$$\Gamma(\theta,q)=\lambda(1-\int_X q(X)dX)+\int_X [q(X)\log P(Y,X,\theta)-q(X)\log q(X)]dX$$

$$\frac{d\Gamma(\theta,q)}{dq(X)}=-\lambda+\log P(Y,X,\theta)-\log q(X)-1=0$$

$$\boxed{q(X)=\frac{P(Y,X,\theta)}{\int_X P(Y,X,\theta)dX}=P(X\mid Y,\theta)}$$

Expectation Step

Given $q(x) = P(X \mid Y, \theta)$, the bound becomes

$$g(\theta, q) = \prod_X \left[\frac{P(Y, X, \theta)}{P(X \mid Y, \theta)}\right]^{q(X)}$$

$$= \prod_X [P(Y, \theta)]^{q(X)}$$

$$= P(Y, \theta)^{\int_X q(X)dX}$$

$$= P(Y, \theta)$$

So, it touches $P(Y, \theta)$ at the current guess for θ

Expectation-Maximization Algorithm

- **E-step**: find $q(X)$ to get a good bound of $P(Y,\theta)$ at given θ

$$q(X) = P(X \mid Y, \theta)$$

- **M-step:** maximize the bound over θ to get the next guess for θ

$$G(\theta, q) = \int_X [q(X) \log P(Y, X, \theta) - q(X) \log q(X)] dX$$

The relevant term of $G(\theta, q)$ is

$$\boxed{\int_X q(X) \log P(Y, X, \theta) dX = E_{q(X)}[\log P(Y, X, \theta)]}$$

Maximization Step

$$E_{q(X)}[\log P(Y,X,\theta)] = \int_X q(X)\log P(Y,X,\theta)dX$$

- HMM definition: $P(X,Y) = \prod_{t=1}^{T} P(x_t \mid x_{t-1})P(y_t \mid x_t)$

$$\log P(X,Y,\theta) = \log(\prod_{t=1}^{T} P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}})P(y_t \mid x_t;\theta_{x_t}(y_t)))$$

$$= \sum_{t=1}^{T} \log P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}}) + \sum_{t=1}^{T} \log P(y_t \mid x_t;\theta_{x_t}(y_t))$$

- $q(X) = P(X \mid Y,\theta^{old}) = \dfrac{P(X,Y,\theta^{old})}{P(Y,\theta^{old})} = cP(X,Y,\theta^{old})$

$$E_{q(X)}[\log P(Y,X,\theta)]$$

$$= \boxed{\int_X P(X,Y,\theta^{old})[\sum_{t=1}^{T} \log P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}}) + \sum_{t=1}^{T} \log P(y_t \mid x_t;\theta_{x_t}(y_t))]dX}$$

Maximization Step

The first term:

$$\int_X P(X,Y,\theta^{old})\sum_{t=1}^{T}\log P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}})dX$$

$$=\sum_i\sum_j\sum_{t=1}^{T}P(x_t=i,x_{t-1}=j,Y,\theta^{old})\log P(x_t=i \mid x_{t-1}=j;\theta_{i,j})$$

use a Lagrange multiplier

$$\frac{d}{d\theta_{i,j}}[\sum_i\sum_j\sum_{t=1}^{T}P(x_t=i,x_{t-1}=j,Y,\theta^{old})\log P(x_t=i \mid x_{t-1}=j;\theta_{i,j})-\lambda(\sum_i\theta_{ij}-1)]=0$$

$$\lambda=\frac{\sum_{t=1}^{T}P(x_t=i,x_{t-1}=j,Y,\theta^{old})}{\sum_{t=1}^{T}\log P(x_t=i \mid x_{t-1}=j;\theta_{i,j})}$$

$$\sum_i\theta_{i,j}=\sum_i\frac{\sum_{t=1}^{T}P(x_t=i,x_{t-1}=j,Y,\theta^{old})}{\lambda}=1$$

$$\boxed{\theta_{i,j}=\frac{\sum_{t=1}^{T}P(x_t=i,x_{t-1}=j,Y,\theta^{old})}{\sum_{t=1}^{T}P(x_{t-1}=j,Y,\theta^{old})}}$$

Maximization Step

The second term:

$$\int_X P(X,Y,\theta^{old}) \sum_{t=1}^{T} \log P(y_t \mid x_t ;\theta_{x_t}(y_t))dX$$

$$= \sum_i \sum_{t=1}^{T} P(x_t = i, Y, \theta^{old}) \log P(y_t \mid x_t = i ;\theta_i(y_t))$$

use a Lagrange multiplier

$$\frac{d}{d\theta_i(y_t = v)}[\sum_i \sum_{t=1}^{T} P(x_t = i, Y, \theta^{old}) \log P(y_t = v \mid x_t = i ;\theta_i(v)) - \lambda(\sum_v \theta_i(v) - 1)] = 0$$

$$\boxed{\theta_i(v) = \frac{\sum_{t=1}^{T} P(x_t = i, Y, \theta^{old})\delta_{y_t,v}}{\sum_{t=1}^{T} P(x_t = i, Y, \theta^{old})}}$$

Baum-Welch Algorithm

initialization of α_0 , β_0 , $p(x_t)$,and $p(x_{t-1}, x_t)$

repeat

for $t = 1 : T$

$$\beta_t(i) = \sum_j P(x_{t+1} = j \mid x_t = i) P(y_{t+1} \mid x_{t+1} = j) \beta_{t+1}(j)$$

$$\alpha_t(i) = \sum_j \alpha_{t-1}(j) P(y_{t-1} \mid x_{t-1} = j) P(x_t = i \mid x_{t-1} = j)$$

$$P(x_t = i \mid Y) = \alpha_t(i) P(y_t \mid x_t = i) \beta_t(i)$$

$$P(x_{t-1} = j, x_t = i \mid Y) = \alpha_{t-1}(j) P(y_{t-1} \mid x_{t-1} = j) P(x_t = i \mid x_{t-1} = j) P(y_t \mid x_t = i) \beta_t(i)$$

end

$$\theta_{i,j} = \frac{\sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old})}{\sum_{t=1}^{T} P(x_{t-1} = j, Y, \theta^{old})} \qquad \theta_i(v) = \frac{\sum_{t=1}^{T} P(x_t = i, Y, \theta^{old}) \delta_{y_t, v}}{\sum_{t=1}^{T} P(x_t = i, Y, \theta^{old})}$$

until converge

Thank you!