# Introduction to Machine Learning

# HMM for Sequential Data

SHENG-JYH WANG

NATIONAL YANG MING CHIAO TUNG UNIVERSITY, TAIWAN

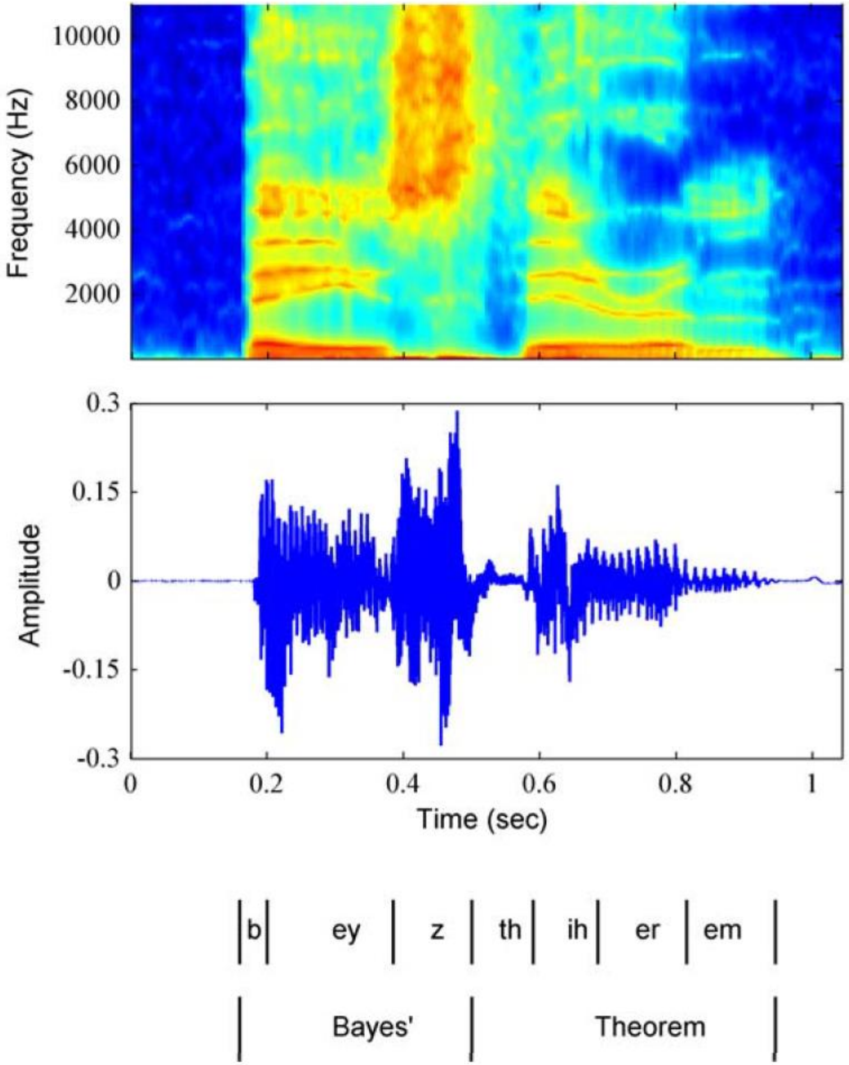
SPRING, 2024

# Markov Model (1/6)

The simplest approach to modeling a sequence of observations is to treat them as independent. This approach would fail to exploit the sequential patterns in the data.

On the other hand, it would be impractical to consider a general dependence of future observations on all previous observations.

⇒ We usually assume future predictions are independent of all but the most recent observations.



# Markov Model (2/6)

**General Rule:**

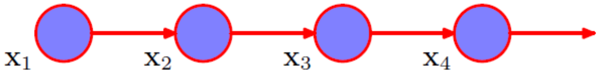
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}).$$

***First-order Markov Chain***:

$$p(\mathbf{x}_n | \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2} p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

A first-order Markov chain of observations $\{\mathbf{x}_n\}$ in which the distribution $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ of a particular observation $\mathbf{x}_n$ is conditioned on the value of the previous observation $\mathbf{x}_{n-1}$.

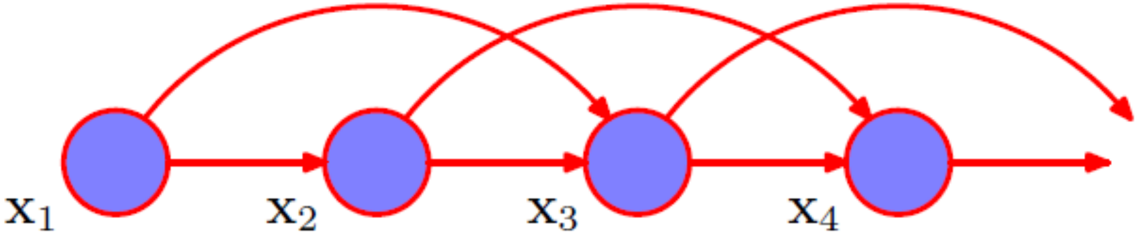


In most applications, the conditional distributions $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ is constrained to be equal.
⇒ *homogeneous* Markov chain.

# Markov Model (3/6)

***Second-order Markov Chain***:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{n=3}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2})$$



We wish to build a model for sequences that is not limited by the Markov assumption to any order and yet that can be specified using a limited number of free parameters.

# Markov Model (4/6)

⇒ For each observation $\mathbf{x}_n$, we introduce a corresponding latent variable $\mathbf{z}_n$ (which may be of different type or dimensionality to the observed variable). We now assume that it is the latent variables that form a Markov chain, giving rise to the graphical structure known as a *state space model*.

$$\mathbf{z}_{n+1} \perp\!\!\!\perp \mathbf{z}_{n-1} \mid \mathbf{z}_n$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[ \prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n).$$

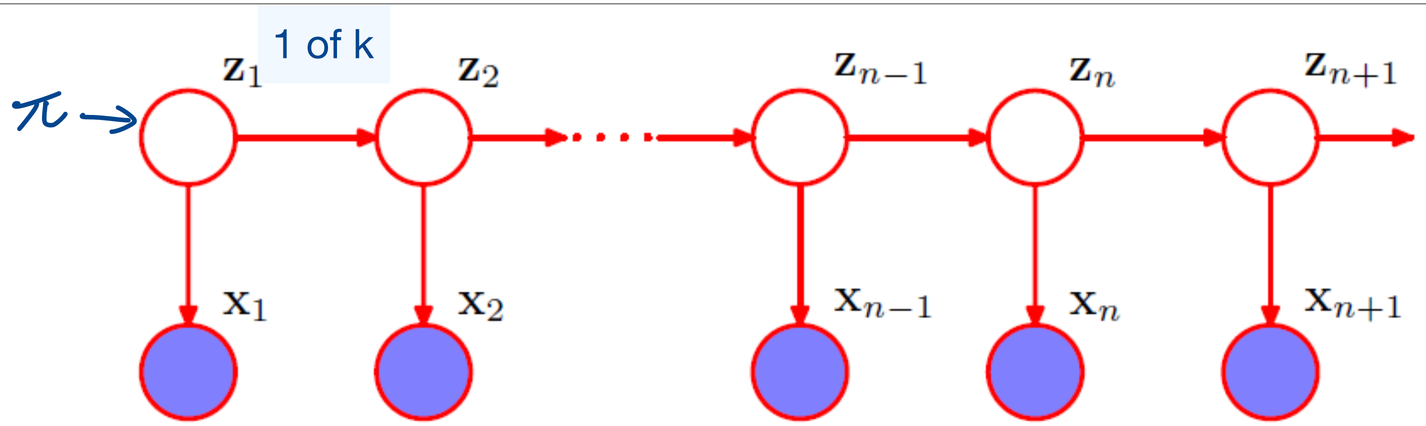
# Markov Model (5/6)

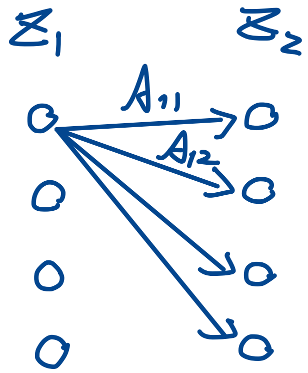
There is always a path connecting any two observed variables $\mathbf{x}_n$ and $\mathbf{x}_m$ via the latent variables.

$\Rightarrow$ The predictive distribution $p(\mathbf{x}_{n+1} | \mathbf{x}_1, . . . , \mathbf{x}_n)$ for observation $\mathbf{x}_{n+1}$ given all previous observations does not exhibit any conditional independence properties.

That is, the observed variables do not satisfy the Markov property at any order.

# Markov Model (6/6)





**Hidden Markov Model (HMM)**: the latent variables are discrete.

**Linear Dynamic System**: both the latent and the observed variables are Gaussian (with a linear-Gaussian dependence of the conditional distributions on their parents).

# Hidden Markov Models (1/30)

- A specific instance of the state space model in which the latent variables are discrete.
- An extension of a mixture model in which the choice of mixture component for each observation is not selected independently but depends on the choice of component for the previous observation.
- Widely used in speech recognition, natural language modelling, on-line handwriting recognition, and for the analysis of biological sequences such as proteins and DNA.

# Hidden Markov Models (2/30)

It is convenient to use a 1-of-*K* coding scheme.

Define the matrix **A**, whose elements are known as *transition probabilities.*

$$A_{jk} \equiv p(z_{nk} = 1 | z_{n-1,j} = 1). \quad 0 \le A_{jk} \le 1$$

$$\textstyle\sum_k A_{jk} = 1$$

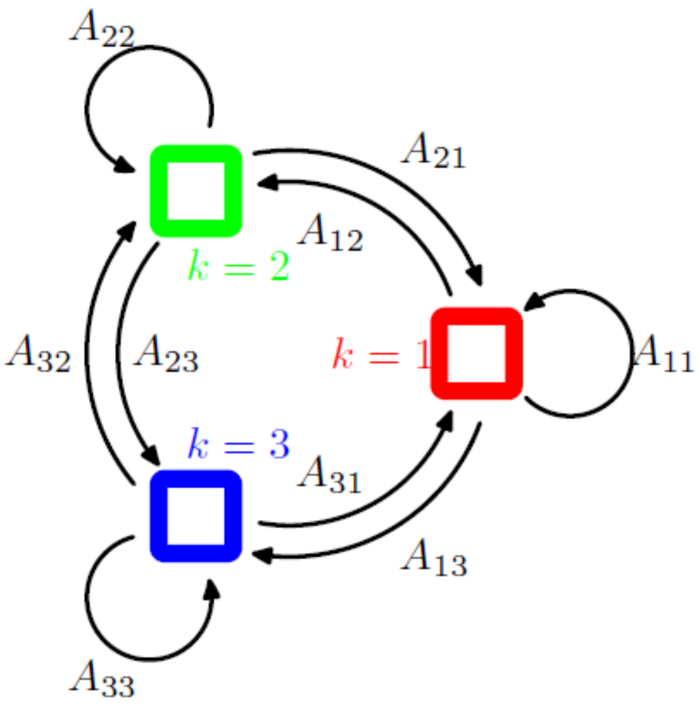
$$p(\mathbf{z}_n | \mathbf{z}_{n-1,\mathbf{A}}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}}$$

The initial latent node $\mathbf{z}_1$ has a marginal distribution $p(\mathbf{z}_1)$ represented by a vector of probabilities $\boldsymbol{\pi}$ with elements $\pi_k \equiv p(z_{1k} = 1)$.

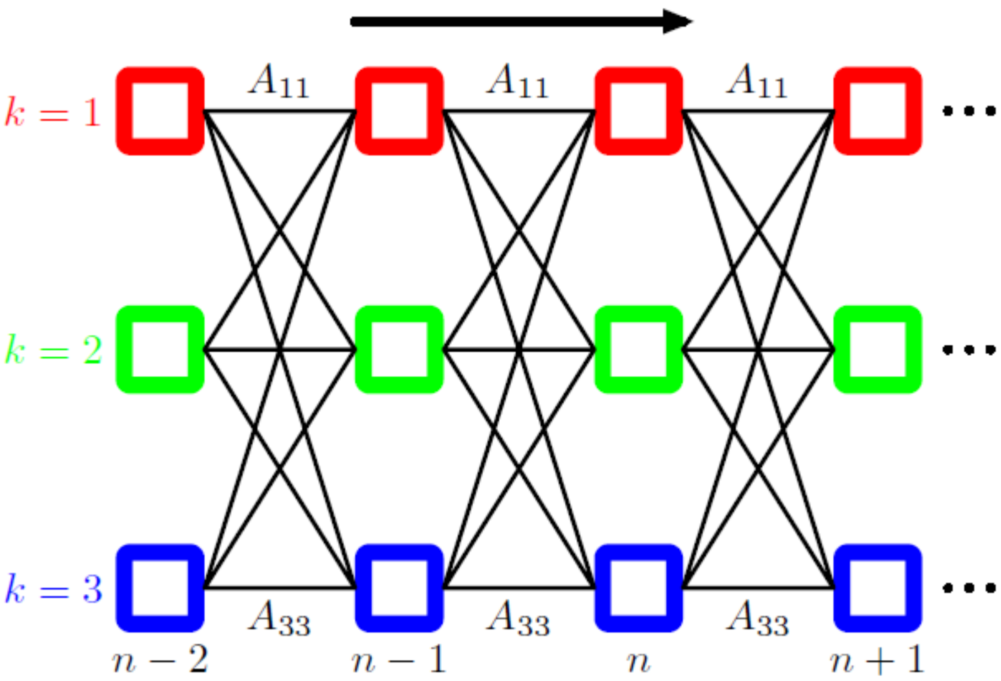
$$\textstyle\sum_k \pi_k = 1$$

# Hidden Markov Models (3/30)

*Transition Diagram*



*Lattice* or *trellis* diagram



# Hidden Markov Models (4/30)

*$p(\mathbf{x}_n|\mathbf{z}_n,\phi)$: emission probabilities*

$$p(\mathbf{x}_n|\mathbf{z}_n,\phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$

*Homogeneous* models: share the same **A** and $\phi$.

Remark: A mixture model for an i.i.d. data set corresponds to the special case in which the parameters $A_{jk}$ are the same for all values of $j$, so that $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ is independent of $\mathbf{z}_{n-1}$.

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right]\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\phi)$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\pi},\mathbf{A},\phi\}$.

# Hidden Markov Models (5/30)

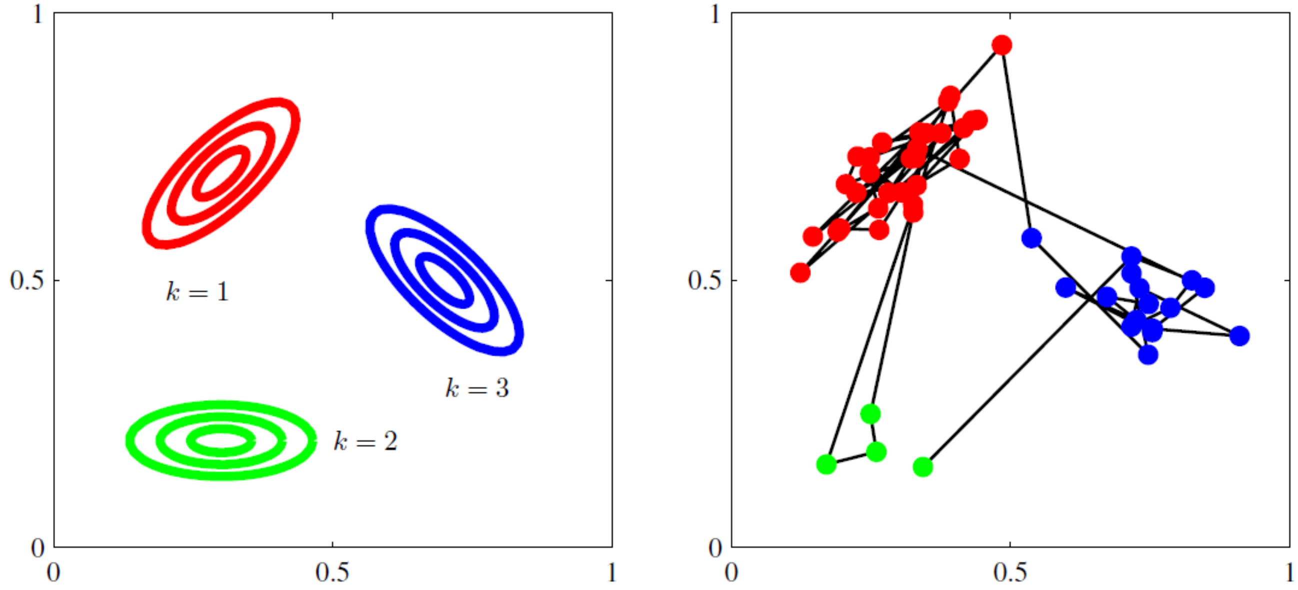


**Figure 13.8** Illustration of sampling from a hidden Markov model having a 3-state latent variable $\mathbf{z}$ and a Gaussian emission model $p(\mathbf{x}|\mathbf{z})$ where $\mathbf{x}$ is 2-dimensional. (a) Contours of constant probability density for the emission distributions corresponding to each of the three states of the latent variable. (b) A sample of 50 points drawn from the hidden Markov model, colour coded according to the component that generated them and with lines connecting the successive observations. Here the transition matrix was fixed so that in any state there is a 5% probability of making a transition to each of the other states, and consequently a 90% probability of remaining in the same state.

# Hidden Markov Models (6/30)

- **Maximum Likelihood for the HMM**

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

- ✓ Because the joint distribution $p(\mathbf{X},\mathbf{Z}|\theta)$ does not factorize over $n$, we cannot simply treat each of the summations over $\mathbf{z}_n$ independently.
- ✓ The number of terms in the summation grows exponentially with the length of the chain.
- ✓ Direct maximization of the likelihood function will lead to complex expressions with no closed-form solutions.

# Hidden Markov Models (7/30)

The EM algorithm starts with some initial model parameters $\theta_{\text{old}}$.

Remark: The **A** and π parameters are often initialized either uniformly or randomly from a uniform distribution.

**E step**: find the posterior distribution of the latent variables $p(\mathbf{Z}|\mathbf{X}, \theta_{\text{old}})$.

use this posterior distribution to evaluate the expectation of the logarithm of the complete-data likelihood function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Here, we define

$$\begin{aligned} \gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\ \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \end{aligned}$$

# Hidden Markov Models (8/30)

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right]\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\phi)$$

where

$$p(\mathbf{z_1}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k{}^{z_{1k}}$$

$$p(\mathbf{z}_n|\mathbf{z}_{n-1,\mathbf{A}}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}$$

$$p(\mathbf{x}_n|\mathbf{z}_n,\phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$

# Hidden Markov Models (9/30)

$$\Longrightarrow \quad p(\mathbf{X},\mathbf{Z}|\theta) = \prod_{k=1}^{K} {\pi_k}^{z_{1k}} \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{n,k}} \prod_{m=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$

$$\ln p(\mathbf{X},\mathbf{Z}|\theta) = \sum_{k=1}^{K} z_{1k} \ln \pi_k + \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{k=1}^{K} z_{n-1,j} z_{n,k} \ln A_{jk} + \sum_{m=1}^{N} \sum_{k=1}^{K} z_{nk} \ln p(\mathbf{x}_n|\phi_k)$$

# Hidden Markov Models (10/30)

$$\Longrightarrow \quad Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n | \boldsymbol{\phi}_k). \qquad (13.17)$$

where

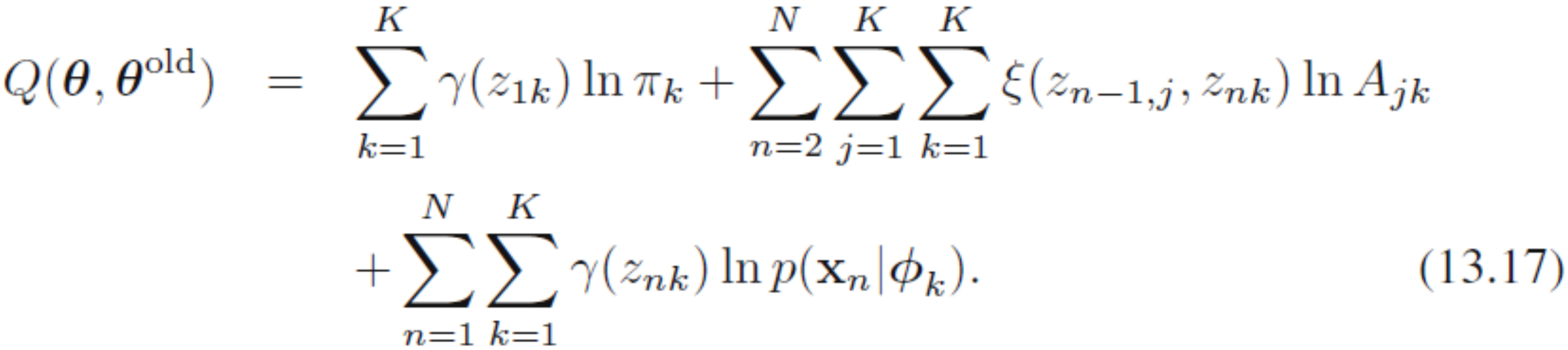
$$\gamma(z_{nk}) = E[z_{nk}] = \sum_{\mathbf{z_n}} \gamma(\mathbf{z}_n) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = E[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}_{n-1}, \mathbf{z_n}} \xi(\mathbf{z}_{n-1}, \mathbf{z_n}) z_{n-1,j} z_{nk}$$

# Hidden Markov Models (11/30)

Here, we denote

- $\gamma(z_{nk})$ as the marginal posterior probability of $z_{nk}$ = 1, and
- $\xi(z_{n-1,j}, z_{nk})$ as the joint posterior distribution of $z_{n-1,j}$ = 1 and $z_{nk}$ = 1.



$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n | \boldsymbol{\phi}_k). \qquad (13.17)$$

Remark: The goal of the E step will be to evaluate the quantities $\gamma(\mathbf{z}_n)$ and $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ efficiently.

# Hidden Markov Models (12/30)

**M step**: we maximize $Q(\theta,\theta_{\text{old}})$ with respect to the parameters $\theta$ **=**$\{\pi,\mathbf{A},\phi\}$ in which we treat $\gamma(\mathbf{z}_n)$ and $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ as constant.

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K}\gamma(z_{1j})} \qquad A_{jk} = \frac{\sum_{n=2}^{N}\xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K}\sum_{n=2}^{N}\xi(z_{n-1,j}, z_{nl})}.$$

The maximization of $Q(\theta,\theta_{\text{old}})$ with respect to $\phi_k$ is the same as that in the case of a standard mixture distribution for i.i.d. data.

# Hidden Markov Models (13/30)

Example: Gaussian emission densities

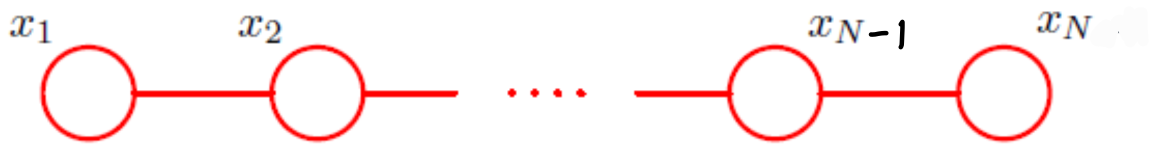
$$p(\mathbf{x}|\phi_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}.$$

# Hidden Markov Models (14/30)

- **Inference on a Chain**



$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N).$$

To find the marginal distribution $p(x_n)$ for a specific node $x_n$,

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{n+1}}\cdots\sum_{x_N} p(\mathbf{x}).$$

# Hidden Markov Models (15/30)

$$p(x_n) = \frac{1}{Z}$$

$$\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots \right]}_{\mu_\alpha(x_n)}$$

$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots \right]}_{\mu_\beta(x_n)}. \qquad (8.52)$$

# Hidden Markov Models (16/30)

We now give a powerful interpretation of this calculation in terms of the passing of local *messages* around on the graph.

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

$\mu_\alpha(x_n)$: a forward message along the chain from node $x_{n-1}$ to node $x_n$.
$\mu_\beta(x_n)$: a backward message along the chain from node $x_{n+1}$ to node $x_n$.

Initially,

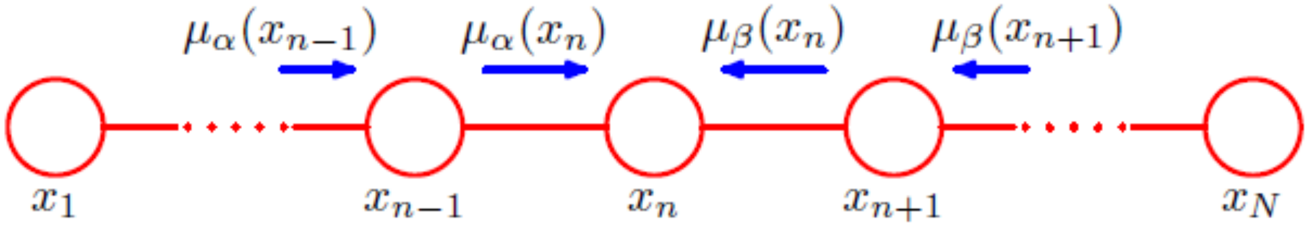
$$\mu_\alpha(x_2) = \sum_{x_1} \psi_{1,2}(x_1, x_2)$$

$$\begin{aligned}\mu_\alpha(x_n) &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\left[\sum_{x_{n-2}} \cdots\right] \\ &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_\alpha(x_{n-1}).\end{aligned}$$

# Hidden Markov Models (17/30)

Similarly,

$$\begin{aligned}\mu_\beta(x_n) &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n)\left[\sum_{x_{n+2}} \cdots\right] \\ &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n)\mu_\beta(x_{n+1}).\end{aligned}$$



# Hidden Markov Models (18/30)

Suppose we wish to evaluate the marginals $p(x_n)$ for every node $n$ ∈{1, . . . , N} in the chain.

- ✓ We first launch a message $\mu_\beta(x_{N-1})$ starting from node $x_N$ and propagate corresponding messages all the way back to node $x_1$.
- ✓ Similarly, we launch a message $\mu_\alpha(x_2)$ starting from node $x_1$ and propagate the corresponding messages all the way forward to node $x_N$.
- ✓ If we have stored all of the intermediate messages along the way, then any node can evaluate its marginal simply by applying

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

# Hidden Markov Models (19/30)

Remarks:

1. If some of the nodes in the graph are observed, then the corresponding variables are simply clamped to their observed values and there is no summation.
2. To calculate the joint distribution $p(x_{n-1}, x_n)$ for two neighboring nodes on the chain, we have

$$p(x_{n-1}, x_n) = \frac{1}{Z}\mu_\alpha(x_{n-1})\psi_{n-1,n}(x_{n-1}, x_n)\mu_\beta(x_n).$$

# Hidden Markov Models (20/30)

- **The Forward-Backward Algorithm**

  An efficient procedure for evaluating $\gamma(\mathbf{z}_n)$ and $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$.

***The Alpha-Beta Algorithm***

$$\begin{aligned}
p(\mathbf{X}|\mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{z}_n) \\
&\quad p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n) \\
p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_n) \\
p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1}) \\
p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n, \mathbf{z}_{n+1}) &= p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1}) \\
p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1}, \mathbf{x}_{n+1}) &= p(\mathbf{x}_{n+2}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1}) \\
p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1}) \\
&\quad p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n) \\
p(\mathbf{x}_{N+1}|\mathbf{X}, \mathbf{z}_{N+1}) &= p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \\
p(\mathbf{z}_{N+1}|\mathbf{z}_N, \mathbf{X}) &= p(\mathbf{z}_{N+1}|\mathbf{z}_N)
\end{aligned}$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

# Hidden Markov Models (21/30)

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$\gamma(\mathbf{z}_n) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$\begin{aligned}\alpha(\mathbf{z}_n) &\equiv p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)\\ \beta(\mathbf{z}_n) &\equiv p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n).\end{aligned}$$

$\alpha(\mathbf{z}_n)$: the joint probability of observing all of the given data up to time $n$ and the value of $\mathbf{z}_n$.

$\beta(\mathbf{z}_n)$: the conditional probability of all future data from time $n + 1$ up to $N$ given the value of $\mathbf{z}_n$.

$\alpha(z_{nk})$: the value of $\alpha(\mathbf{z}_n)$ when $z_{nk} = 1$.

$\beta(z_{nk})$: the value of $\beta(\mathbf{z}_n)$ when $z_{nk} = 1$.

# Hidden Markov Models (22/30)

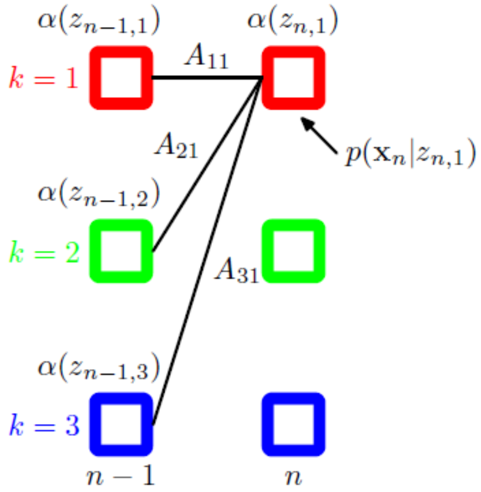
$$\begin{aligned}
\alpha(\mathbf{z}_n) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_n) p(\mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{z}_{n-1}) \\
&= p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1})
\end{aligned}$$

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1}).$$

# Hidden Markov Models (23/30)

To start this recursion, we need an initial condition that is given by

$$\alpha(\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\mathbf{x}_1|\phi_k)\}^{z_{1k}}$$

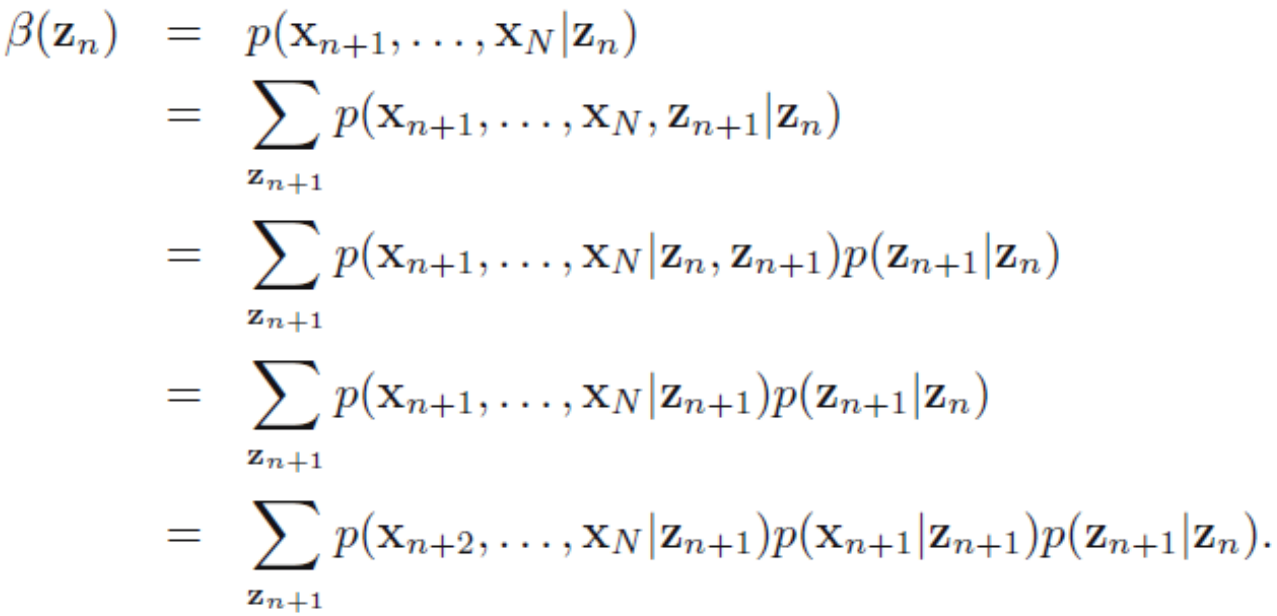


Remark: There are $K$ terms in the summation, and the right-hand side has to be evaluated for each of the $K$ values of $\mathbf{z}_n$.

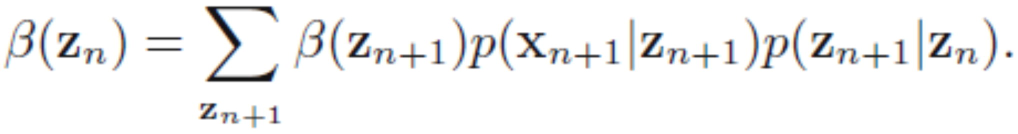
$\Rightarrow$ The cost for each step of the $\alpha$ recursion is $O(K^2)$ and the overall cost for the whole chain is of $O(K^2N)$.

# Hidden Markov Models (24/30)

Similarly, we have



$$\begin{aligned}\beta(\mathbf{z}_n) &= p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N,\mathbf{z}_{n+1}|\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n,\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+2},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n).\end{aligned}$$



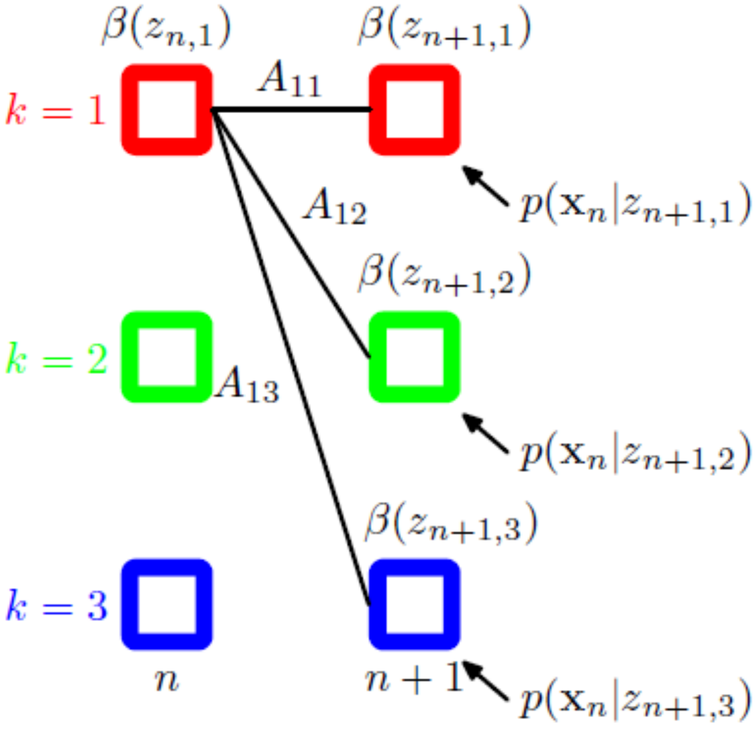
$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n).$$

# Hidden Markov Models (25/30)

Initially, we have

$$p(\mathbf{z}_N|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{z}_N)\beta(\mathbf{z}_N)}{p(\mathbf{X})}$$

with $\beta(\mathbf{z}_N) = 1$ for all settings of $\mathbf{z}_N$.



# Hidden Markov Models (26/30)

**M step**:

$$\mu_k = \frac{\sum_{n=1}^{n} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{n} \gamma(z_{nk})} = \frac{\sum_{n=1}^{n} \alpha(z_{nk})\beta(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{n} \alpha(z_{nk})\beta(z_{nk})}.$$

The quantity *p*(**X**) will cancel out. However, the quantity *p*(**X**) represents the likelihood function whose value we typically wish to monitor during the EM optimization.

$$p(\mathbf{X}) = \sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n).$$

We can evaluate the likelihood function by computing this sum, for any convenient choice of *n*.

# Hidden Markov Models (27/30)

If we only want to evaluate the likelihood function, we can run the α recursion from the start to the end of the chain, and then use this result for *n* = *N.*

$$p(\mathbf{X}) = \sum_{\mathbf{z}_N} \alpha(\mathbf{z}_N).$$

Similarly,

$$\begin{aligned}
\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n)p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1})}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\beta(\mathbf{z}_n)}{p(\mathbf{X})} \qquad (13.43)
\end{aligned}$$

We can calculate the $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ directly by using the results of the α and β recursions.

# Hidden Markov Models (28/30)

**Prediction Distribution**

To predict $\mathbf{x}_{N+1}$ based on the observed data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

$$\begin{aligned}
p(\mathbf{x}_{N+1}|\mathbf{X}) &= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}, \mathbf{z}_{N+1}|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) p(\mathbf{z}_{N+1}|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}, \mathbf{z}_N|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) p(\mathbf{z}_N|\mathbf{X}) \\
&= \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) \frac{p(\mathbf{z}_N, \mathbf{X})}{p(\mathbf{X})} \\
&= \frac{1}{p(\mathbf{X})} \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N) \alpha(\mathbf{z}_N)
\end{aligned}$$

# Hidden Markov Models (29/30)

**Scaling Factor**

At each step, the new value $\alpha(z_n)$ is obtained from the previous value $\alpha(z_{n-1})$ by multiplying by $p(z_n|z_{n-1})$ and $p(x_n|z_n)$.

$\Rightarrow$ The values of $\alpha(z_n)$ can go to zero exponentially quickly.

$\alpha(z_n) = p(x_1, . . . , x_n, z_n)$

We define

$$\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\alpha(\mathbf{z}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

$$c_n = p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}).$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{m=1}^{n} c_m$$

$$\alpha(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n)p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left(\prod_{m=1}^{n} c_m\right)\widehat{\alpha}(\mathbf{z}_n).$$

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}}\widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}).$$

# Hidden Markov Models (30/30)

Similarly, we can define

$$\beta(\mathbf{z}_n) = \left(\prod_{m=n+1}^{N} c_m\right)\widehat{\beta}(\mathbf{z}_n)$$

$$\widehat{\beta}(\mathbf{z}_n) = \frac{p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_n)}.$$

$$c_{n+1}\widehat{\beta}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \widehat{\beta}(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n).$$

The likelihood function can be found using

$$p(\mathbf{X}) = \prod_{n=1}^{N} c_n.$$

$$\begin{aligned}\gamma(\mathbf{z}_n) &= \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n)\\ \xi(\mathbf{z}_{n-1},\mathbf{z}_n) &= c_n\widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{-1})\widehat{\beta}(\mathbf{z}_n).\end{aligned}$$