

# Maximum Likelihood for the HMM

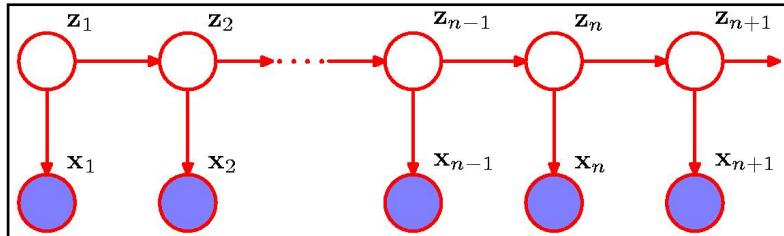
Sargur Srihari

[srihari@cedar.buffalo.edu](mailto:srihari@cedar.buffalo.edu)

# HMM Topics

1. What is an HMM?
2. State-space Representation
3. HMM Parameters
4. Generative View of HMM
5. Determining HMM Parameters Using EM
6. Forward-Backward or  $\alpha-\beta$  algorithm
7. HMM Implementation Issues:
  - a) Length of Sequence
  - b) Predictive Distribution
  - c) Sum-Product Algorithm
  - d) Scaling Factors
  - e) Viterbi Algorithm

# HMM Parameters



$$p(X, Z | \theta) = p(z_1 | \pi) \left[ \prod_{n=2}^N p(z_n | z_{n-1}, A) \right] \prod_{m=1}^N p(x_m | z_m, \phi)$$

where  $X = \{x_1, \dots, x_N\}$ ,  $Z = \{z_1, \dots, z_N\}$ ,  $\theta = \{\pi, A, \phi\}$

We have three sets of HMM parameters:  $\theta = (\pi, A, \phi)$

1. Initial Probabilities of first latent variable:

$\Pi$  is a vector of  $K$  probabilities of the states for latent variable  $z$

2. Transition Probabilities (State-to-state for any latent variable):

$A$  is a  $K \times K$  matrix of transition probabilities  $A_{ij}$

3. Emission Probabilities (Observations conditioned on latent):

$\phi$  are parameters of conditional distribution  $p(x_k | z_k)$

- $A$  and  $\pi$  parameters are often initialized uniformly
- Initialization of  $\phi$  depends on form of distribution

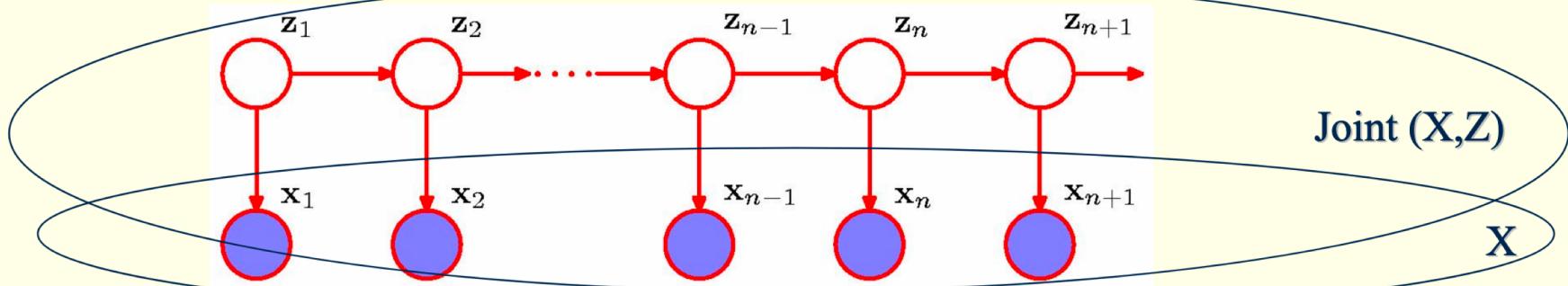
# Determining HMM Parameters

- Given data set  $X = \{x_1, \dots, x_n\}$  we can determine HMM parameters  $\theta = \{\pi, A, \phi\}$  using maximum likelihood
- Likelihood function obtained from joint distribution by marginalizing over latent variables  $Z = \{z_1, \dots, z_n\}$

$$p(X, Z | \theta) = p(z_1 | \pi) \left[ \prod_{n=2}^N p(z_n | z_{n-1}, A) \right] \prod_{m=1}^N p(x_m | z_m, \phi)$$

where  $X = \{x_1, \dots, x_N\}$ ,  $Z = \{z_1, \dots, z_N\}$ ,  $\theta = \{\pi, A, \phi\}$

$$p(X | \theta) = \sum_Z p(X, Z | \theta)$$



# Computational Issues for Parameters

- Joint distribution is  $p(X|\theta) = \sum_Z p(X, Z|\theta)$ 
  - where
$$p(X, Z | \theta) = p(z_1 | \pi) \left[ \prod_{n=2}^N p(z_n | z_{n-1}, A) \right] \prod_{m=1}^N p(x_m | z_m, \phi)$$
where  $X = \{x_1, \dots, x_N\}$ ,  $Z = \{z_1, \dots, z_N\}$ ,  $\theta = \{\pi, A, \phi\}$
- Joint distribution  $p(X, Z | \Theta)$  does not factorize over  $n$ , we cannot treat each of the summations over  $z_n$  independently
- There are  $N$  variables summed over each of which has  $K$  states, so there are  $K^N$  terms
  - No. of terms grows exponentially with length of chain, summing over all paths in trellis

# Solution to computational task

- Use conditional independence properties to reorder summations to obtain algorithm that scales linearly with length of chain
- Use Expectation Maximization to maximizing the log-likelihood function in HMMs

# EM for MLE in HMM

1. Start with *initial selection for model parameters*  $\theta^{old}$
2. In E step take these parameter values and find  
*posterior distribution of latent variables*  $p(Z|X, \theta^{old})$

Use this posterior distribution to evaluate  
*expectation of the logarithm of the complete-data likelihood function*  $\ln p(X, Z | \theta)$

Which can be written as

$$Q(\theta, \theta^{old}) = \sum_Z \underline{p(Z | X, \theta^{old})} \ln p(X, Z | \theta)$$

underlined portion independent of  $\theta$  is evaluated

3. In M-Step maximize  $Q$  w.r.t.  $\theta$

# Expansion of Q

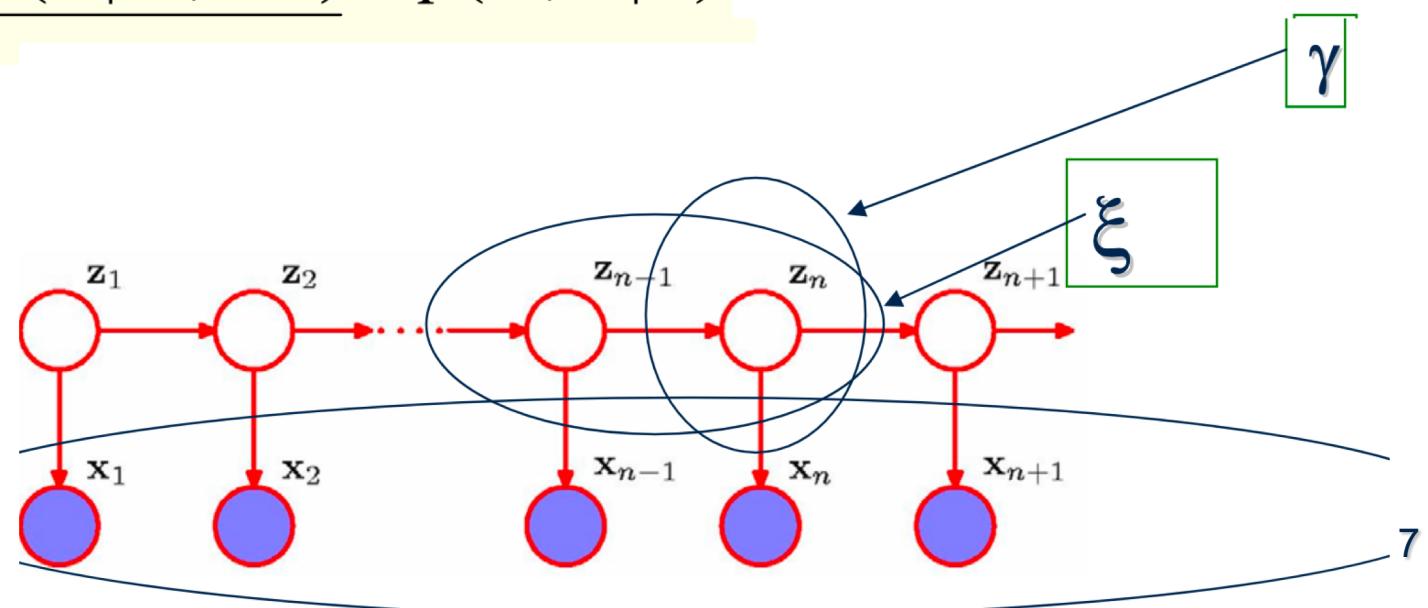
- Introduce notation Gamma and Xi

$\gamma(z_n) = p(z_n | X, \theta^{old})$ : Marginal posterior distribution of latent variable  $z_n$

$\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n | X, \theta^{old})$ : Joint posterior of two successive latent variables

- We will be re-expressing  $Q$  in terms of  $\gamma$  and  $\xi$

$$Q(\theta, \theta^{old}) = \sum_z \underline{p(Z | X, \theta^{old})} \ln p(X, Z | \theta)$$



# Detail of $\gamma$ and $\xi$

For each value of  $n$  we can store

$\gamma(z_n)$  using  $K$  non-negative numbers that sum to unity

$\xi(z_{n-1}, z_n)$  using a  $K \times K$  matrix whose elements also sum to unity

- Using notation

$\gamma(z_{nk})$  denotes conditional probability of  $z_{nk} = 1$

Similar notation for  $\xi(z_{n-1,j}, z_{nk})$

- Because the expectation of a binary random variable is the probability that it takes value 1

$$\gamma(z_{nk}) = E[z_{nk}] = \sum_z \gamma(z) z_{nk}$$

$$\xi(z_{n-1,j}, z_{nk}) = E[z_{n-1,j} z_{nk}] = \sum_z \gamma(z) z_{n-1,j} z_{nk}$$

# Expansion of $Q$

- We begin with

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

- Substitute

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

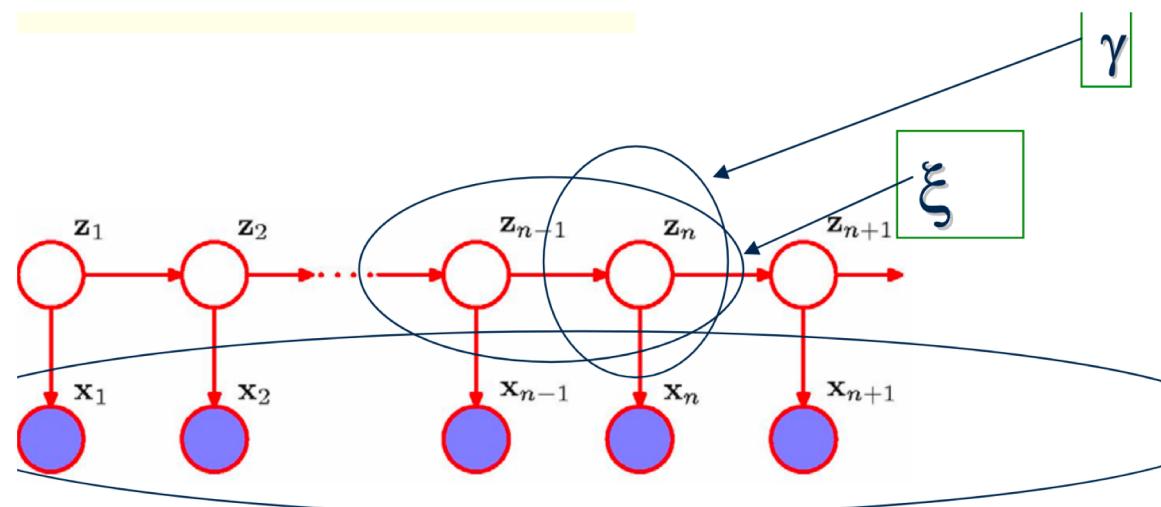
- And use definitions of  $\gamma$  and  $\xi$  to get:

$$\begin{aligned} Q(\theta, \theta^{old}) &= \sum_{k=1}^K \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^N \sum_{j=1}^K \sum_{k=1}^K \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} \\ &\quad + \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \ln p(x_n | \phi_k) \end{aligned}$$

# E-Step

$$Q(\theta, \theta^{old}) = \sum_{k=1}^K \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^N \sum_{j=1}^K \sum_{k=1}^K \xi(z_{n-1,j}, z_{nk}) \ln A_{jk}$$
$$+ \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \ln p(x_n | \phi_k)$$

Goal of E step is to evaluate  $\gamma(z_n)$  and  $\xi(z_{n-1}, z_n)$  efficiently (Forward-Backward Algorithm)



# M-Step

- Maximize  $Q(\theta, \theta^{old})$  with respect to parameters  $\theta = \{\pi, A, \phi\}$ 
  - Treat  $\gamma(z_n)$  and  $\xi(z_{n-1}, z_n)$  as constant
- Maximization w.r.t.  $\pi$  and  $A$ 
  - easily achieved (using Lagrangian multipliers)

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^K \gamma(z_{1j})}$$

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

- Maximization w.r.t.  $\phi_k$ 
  - Only last term of  $Q$  depends on  $\phi_k \rightarrow \boxed{\sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \ln p(x_n | \phi_k)}$
  - Same form as in mixture distribution for i.i.d.

# M-Step for Gaussian Emission

- Maximization of  $Q(\theta, \theta^{old})$  wrt  $\phi_k$
- Gaussian Emission Densities

$$p(\mathbf{x}|\phi_k) \sim N(\mathbf{x}|\mu_k, \Sigma_k)$$

- Solution:

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

$$\Sigma_k = \frac{\sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^N \gamma(z_{nk})}$$

# M-Step for Multinomial Observed

- Conditional Distribution of Observations have the form

$$p(\mathbf{x} \mid \mathbf{z}) = \prod_{i=1}^D \prod_{k=1}^K \mu_{ik}^{x_i z_k}$$

- M-Step equations:

$$\mu_{ik} = \frac{\sum_{n=1}^N \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^N \gamma(z_{nk})}$$

- Analogous result holds for Bernoulli observed variables