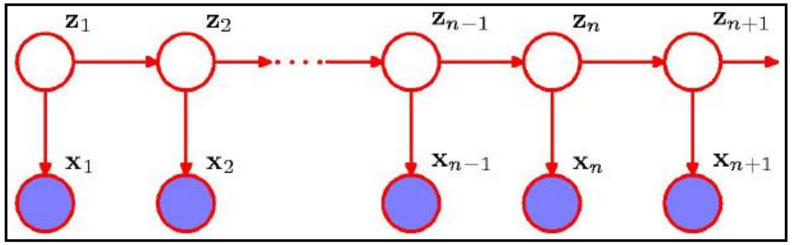
Maximum Likelihood for the HMM

Sargur Srihari
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 $X=\{x_1,..x_N\},\ Z=\{z_1,..z_N\},\ \theta=\{\pi,A,\phi\}$

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

1.Initial Probabilities of first latent variable:

Π 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 x K matrix of transition probabilities A_{ij}

3. Emission Probabilities (Observations conditioned on latent):

ϕ are parameters of conditional distribution $p(\mathrm{x}_k|\mathrm{z}_k)$

- A and π parameters are often initialized uniformly
- Initialization of ϕ depends on form of distribution

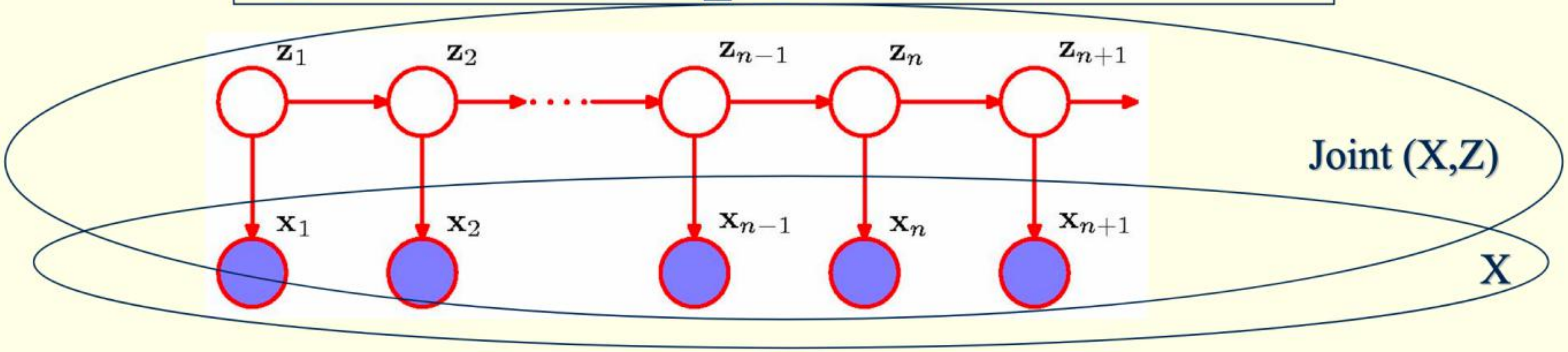
Determining HMM Parameters

- Given data set $X=\{\mathrm{x}_1,..\mathrm{x}_n\}$ we can determine HMM parameters $\theta=\{\pi,A,\phi\}$ using maximum likelihood
- Likelihood function obtained from joint distribution by marginalizing over latent variables $Z=\{\mathrm{z}_1,..\mathrm{z}_n\}$

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

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

$$\boxed{p(X|\theta) = \Sigma_Z\, p(X,Z\,|\,\theta)}$$



Computational Issues for Parameters

- Joint distribution is $p(X|\theta) = \sum_Z p(X,Z|\theta)$
  - where

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

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

- Joint distribution $p(X,Z|\boldsymbol{\theta})$ does not factorize over n, we cannot treat each of the summations over $\mathbf{z}_n$ 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* θ^{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 θ is evaluated

3. In M-Step maximize Q w.r.t. θ

Expansion of Q

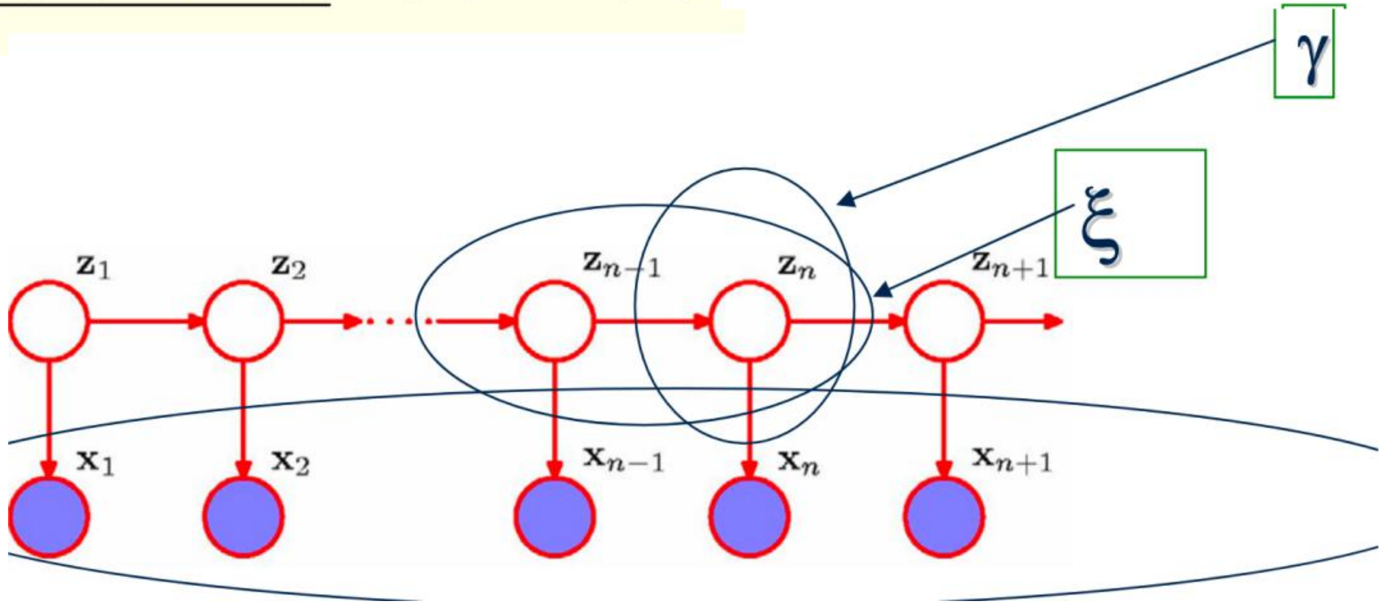
- Introduce notation Gamma and Xi

  $\gamma(z_n) = p(z_n|X,\theta^{old})$: 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 γ and ξ

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



Detail of γ and ξ

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 x 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

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

- Substitute

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

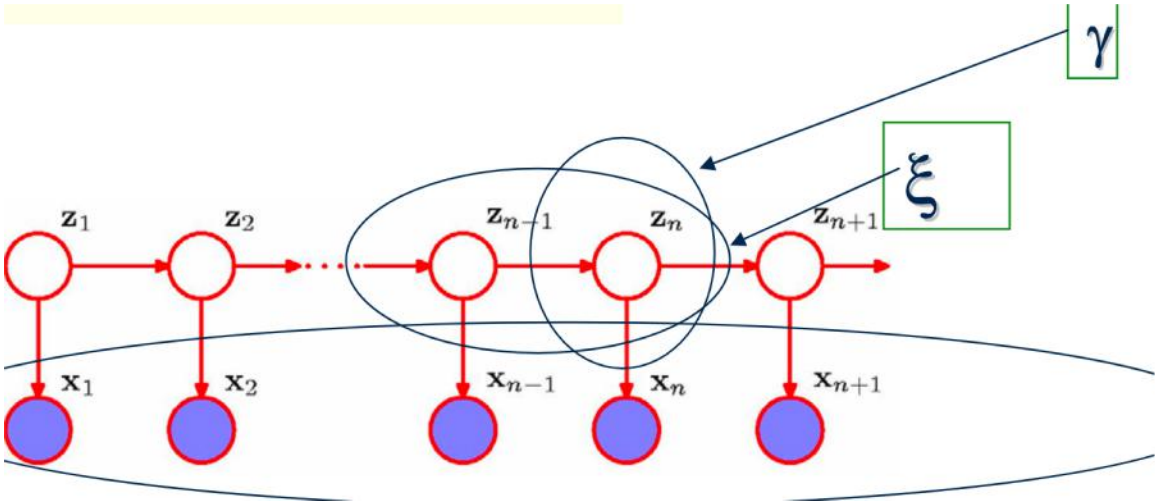
- And use definitions of γ and ξ to get:

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

E-Step

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

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. π and A
  - easily achieved (using Lagrangian multipliers)

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

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

M-Step for Gaussian Emission

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

$$p(\mathbf{x}|\phi_k) \sim N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

- Solution:

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

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

M-Step for Multinomial Observed

- Conditional Distribution of Observations have the form

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

- 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