EECS 545: Machine Learning

Lecture 20. Hidden Markov Models

Honglak Lee

3/23/2011

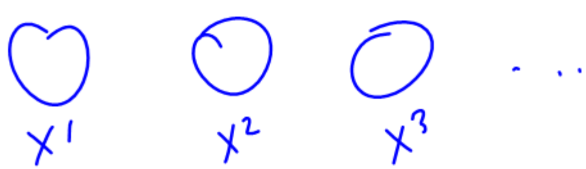
Outline

- Hidden Markov Models

Sequential Data

- Some data has intrinsic sequential structure.
  - Time series: speech, EKGs, stock market, etc.
  - Spatial sequences: DNA, natural language, etc.

- We could treat data points as i.i.d. samples
  - But that's false (they are not i.i.d.), so any conclusions we draw are likely to be wrong.
  - We are ignoring valuable constraints in the data.



Markov Chains

$z^{(i)} \in \{1, \ldots, K\}$

- A Markov chain is a series of random variables $z^{(1)}, \ldots, z^{(M)}$, such that

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

$$= (K-1) \quad K^m \qquad (K-1) \quad K$$

- This is the *Markov property*, and can be summarized as:
  - *The future is independent of the past, given the present*.
- Often used to model temporal evolution.

Markov Models

- If a sequence has the Markov property

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

- then the joint probability distribution

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

- has a simplified form

$$p(x_1,x_2,x_3) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_1,x_2)$$

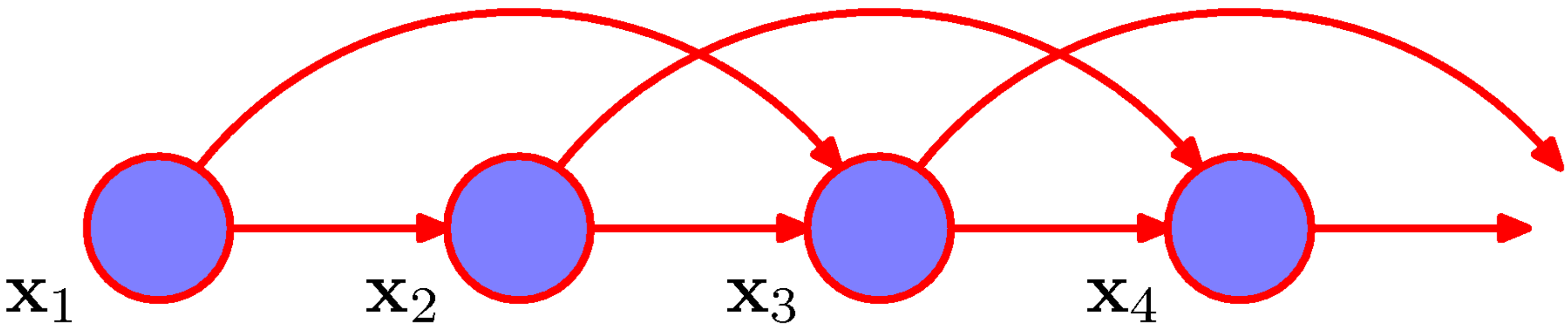
$$p(x_3|x_1,x_2) = p(x_3|x_2)$$

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

$$x_1 \to x_2 \to x_3 \to x_4 \to \cdots \quad x_n$$

Higher-Order Markov Chains

- We can extend the concept of Markov chain to more complex, but still local, kinds of dependency.

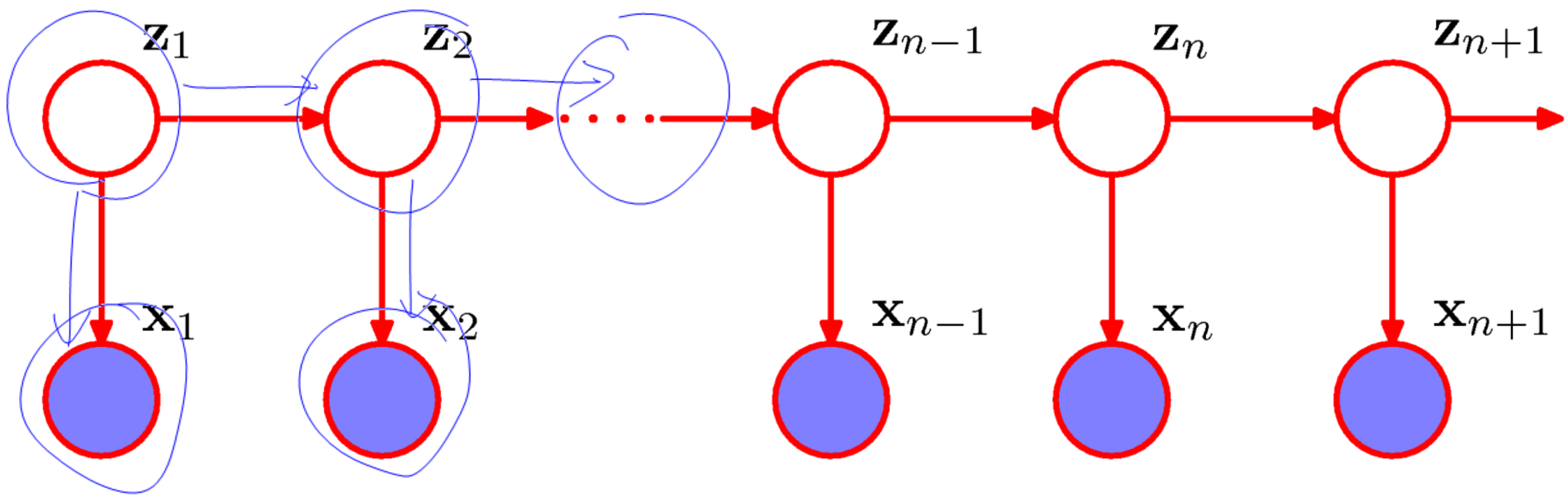


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\underline{\prod_{n=3}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2})}$$

$\prod_i p(x_i | Pa\, x_i)$

Markov chain with latent variable

- For each observation x_n, we assume there is a latent variable z_n, and the z_n form a Markov chain.



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

$p(z_n | z_{n-1})$: transition prob $\quad$ $P(x_n | z_n)$: observation prob

Markov chain with latent variable

- This leads to
  - **Hidden Markov Models**
    - when the latent variable is discrete, and
  - **Linear Dynamical Systems**
    - when the latent variable is Gaussian.

Hidden Markov Models

$$z_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^K \quad \{0,1\}^K$$

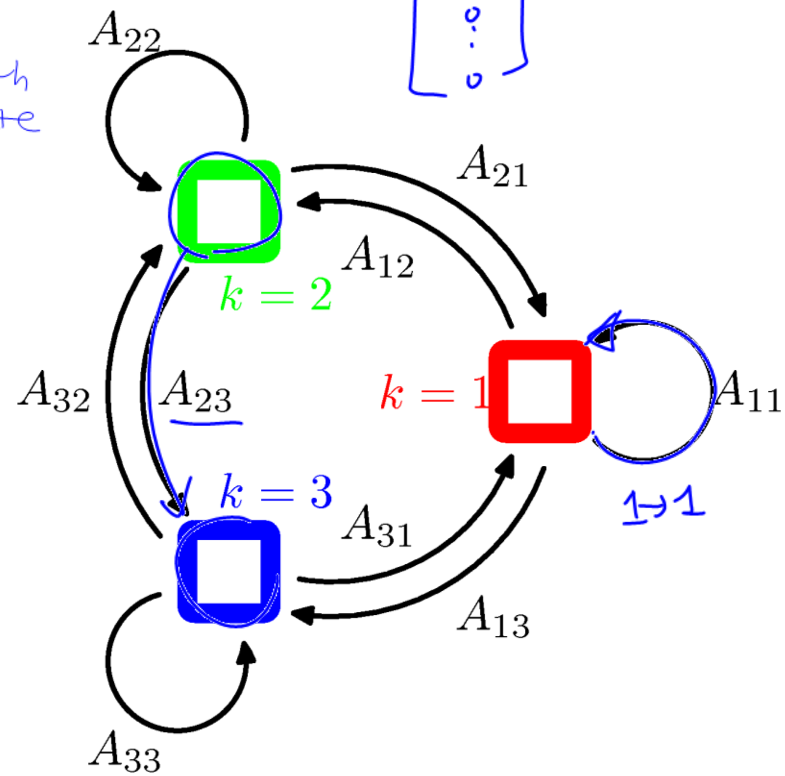
- Use 1-of-*K* coding for values of z_n.

$z_{ni} = 1$ iff z_n is k-th state

- A is the table of transition probabilities.

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

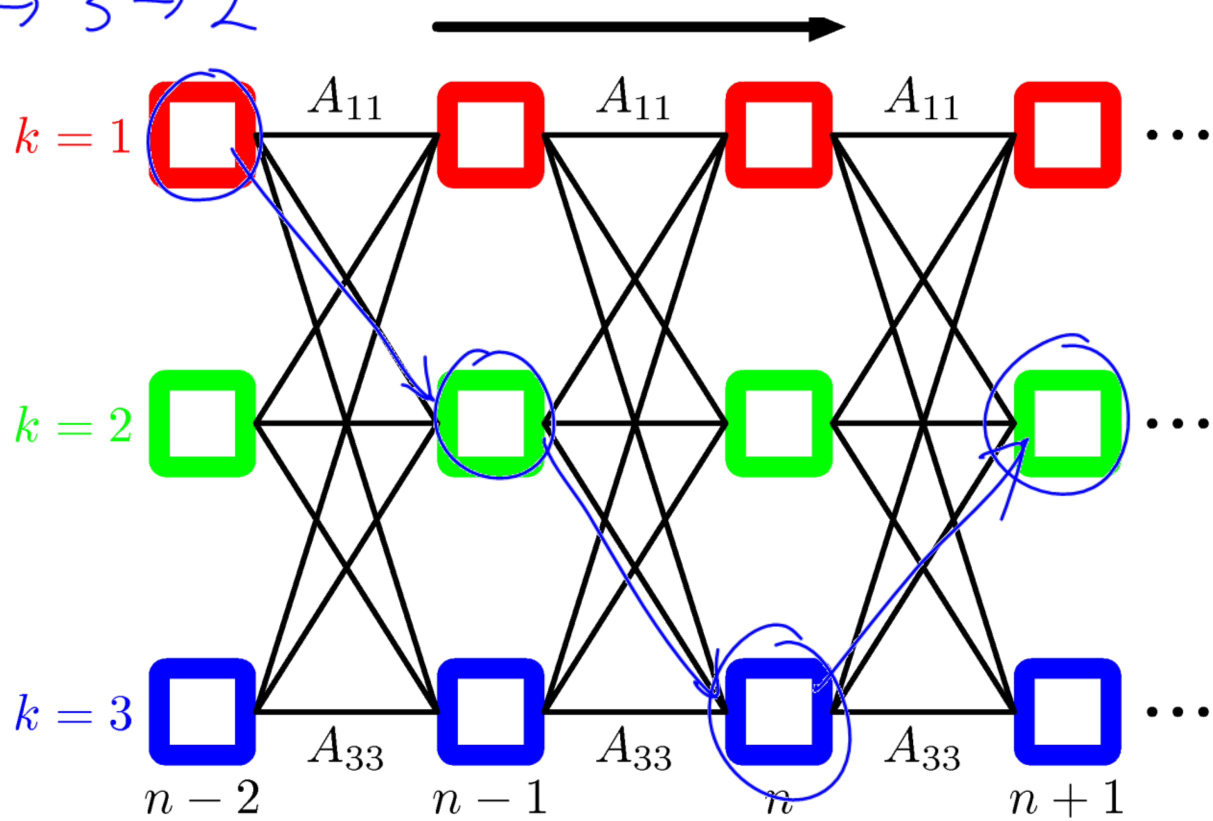
- This is *not* a graph of variables. These are transitions among values of *one* variable.



Hidden Markov Models

- Lattice representation of transition diagram

$1 \to 2 \to 3 \to 2$



Hidden Markov Models

- The prior distribution at the initial state:

$$p(\mathbf{z}_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$$

P("$z_1 = 1$") = π_1

P("$z_1 = 2$") = π_2

$\vdots$

P("$z_1 = k$") = π_K

- The conditional distribution (transition table):

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

= 1

at time n-1, state = j

at time n, state = k

= P("$z_n = k$" | "$z_{n-1} = j$")

- Emission probabilities of observables:

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

= 1

time n, state = k

Hidden Markov Models

- So, the overall joint probability distribution, over both observed and latent variables, is

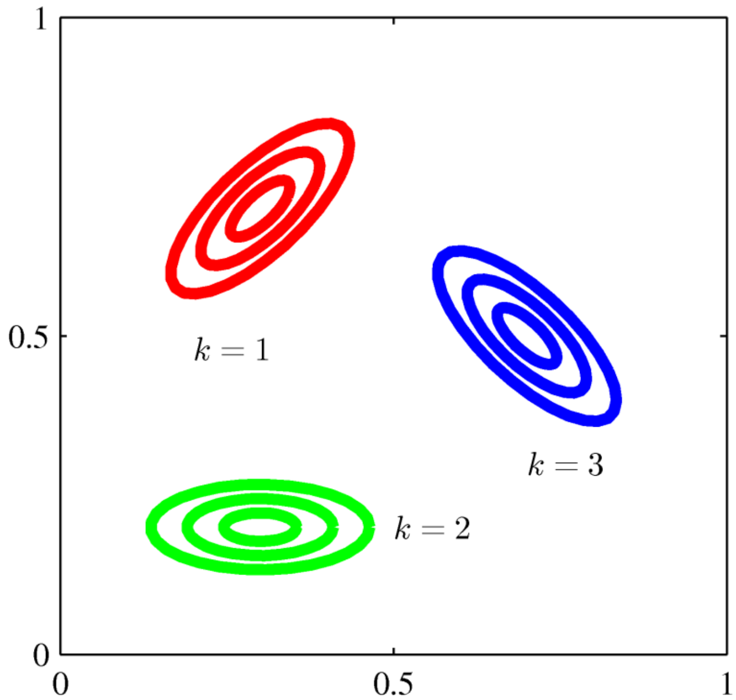
$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{z}_1|\pi) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi)$$

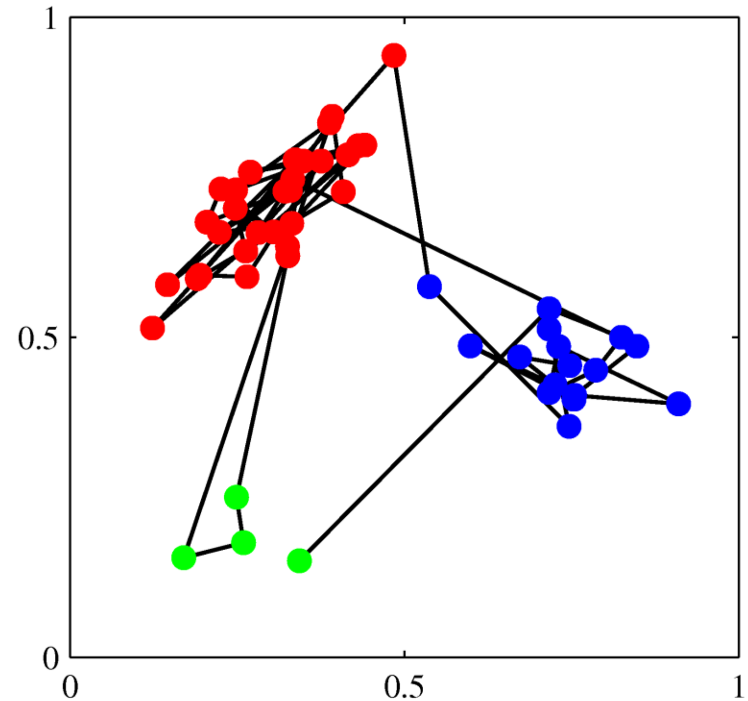
- The parameters are: $\theta = \{\pi, \mathbf{A}, \phi\}$

  – We can use EM to estimate these from data **X**.

Generative sampling from HMM

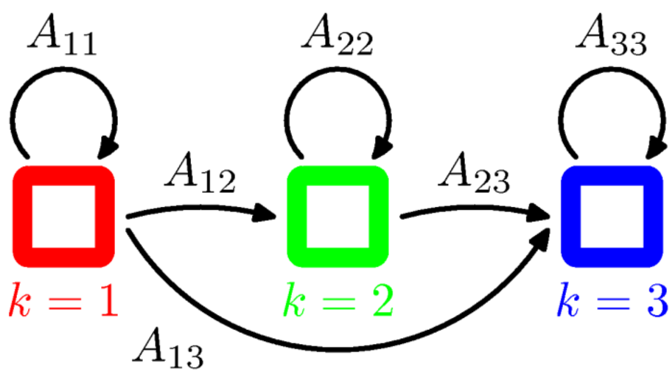
- Transition: 90% of staying in the same state, 5% chance of transition to each other state.

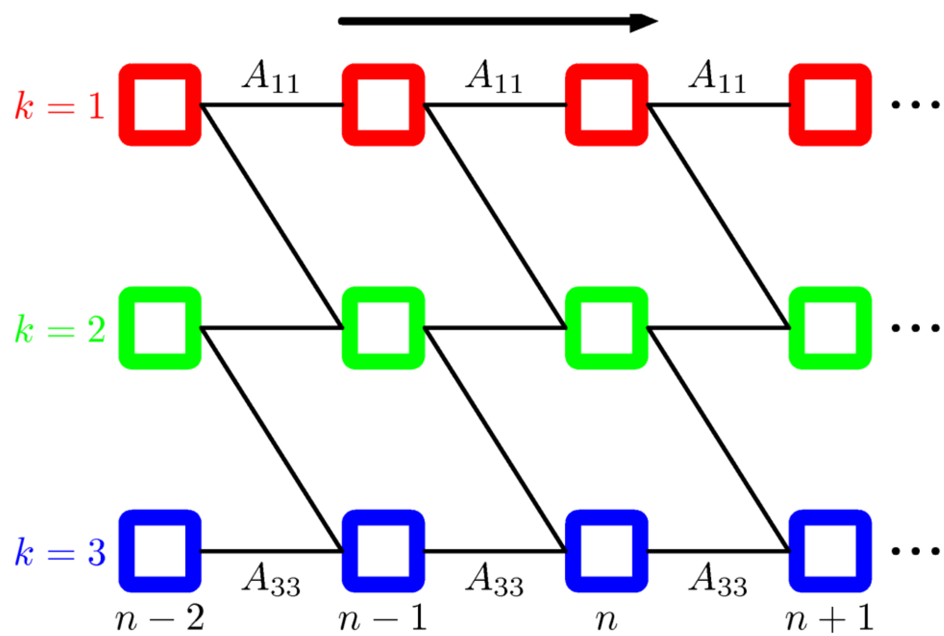




Constraints on HMM transitions

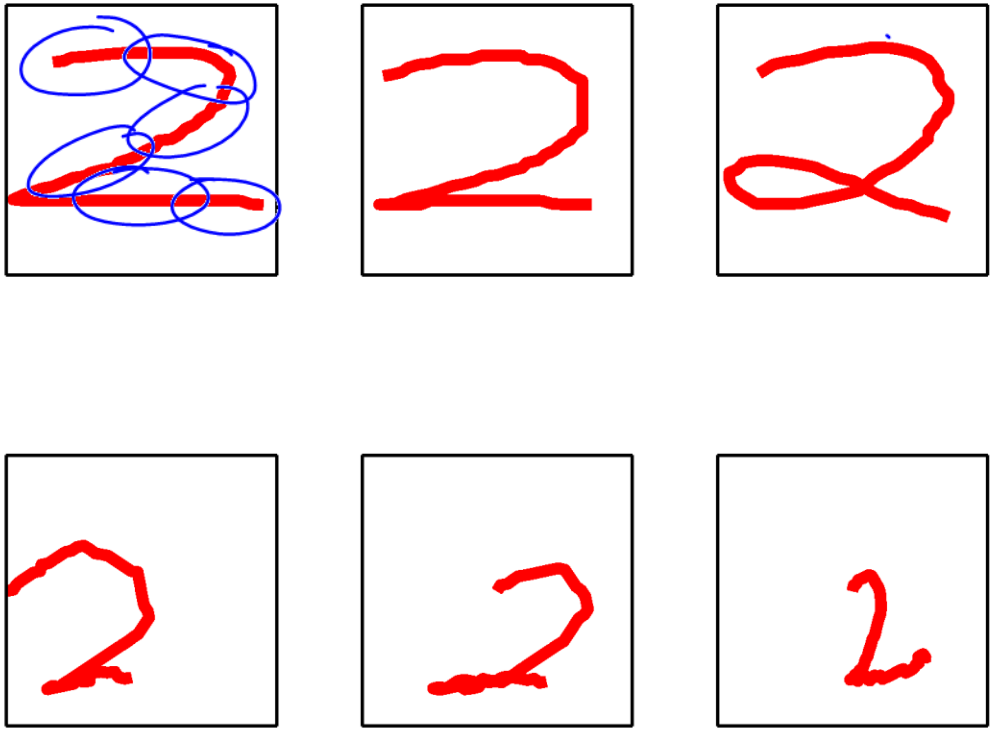
- Left-to-right constraint to describe a temporal process.
- Max change constraint to describe continuity.





HMM for online handwritten digits

- One set of states for the top arc, then a cusp, then a set of states for the base.



Maximum Likelihood for the HMM

- Given a set X of observations, we want to use maximum likelihood to estimate the parameters $\theta = \{\pi, \mathbf{A}, \phi\}$
  – and the latent variables **Z**.

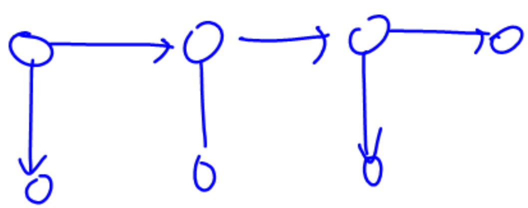
E step: $Q(z) = P_\theta(z | X)$

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

M step: $\max_\theta \sum_{i=1}^{N} Q(z^{(i)}) \ln P(x^{(i)}, z^{(i)} | \theta)$

- Part of applying E-M to this will be evaluating

$$p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$$



$$p(z_n^{=k} | z_{n-1}^{=j}) \propto \frac{\sum_n \text{Count}[z_{n-1,j}, z_{n,k}]}{\sum_n \text{Count}[z_{n-1,j}]}$$

(For x_n: binary) $p(x_n | z_n^{=j}) \propto \frac{\sum_n \text{Count}[z_{n,j}, x_n]}{\sum_n \text{Count}[z_{n,j}]}$

E-M for HMMs

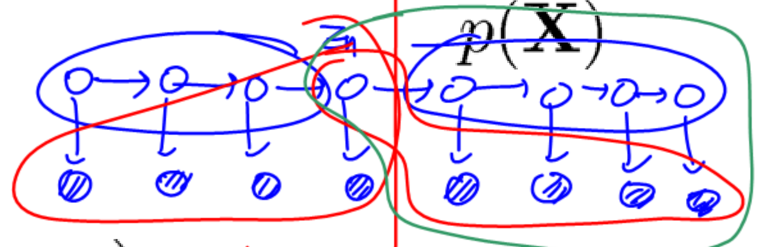
$$\sum_{z_1,\ldots,z_{n-1}} \sum_{z_{n+1},\ldots,z_N} P(X, Z) \;\; = \;\; P(X, z_n)$$

- Part of the E-step is evaluating $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$
- a key term is

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

- where

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



- we define

$$\alpha(\mathbf{z}_n) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$$

$$\beta(\mathbf{z}_n) \equiv p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)$$

$P(X_1 \ldots X_n, z_n)$ $\quad$ $P(X_{n+1} \ldots X_N | z_n)$

$$\Downarrow$$

$$\sum_{z_{n+1} \ldots z_N} P(X_{n+1}, \ldots, X_N, z_{n+1}, \ldots, z_N | z_n)$$

Forward-Backward Algorithm

- Treat these terms as messages:

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

- Send one forward

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

- And the other backward

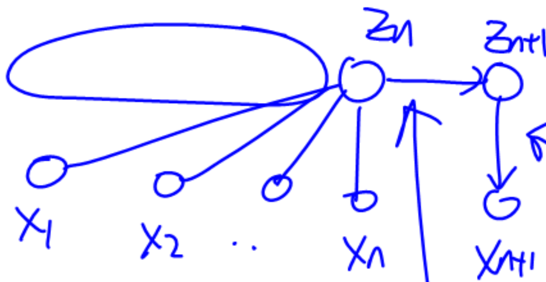
$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1} | \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n)$$

Q. Verify this

$$\alpha(z_{n+1}) = p(x_1 \dots x_{n+1}, z_{n+1})$$

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

$$\alpha(z_1) = p(z_1, x_1)$$
$$= p(x_1|z_1)\,p(z_1)$$

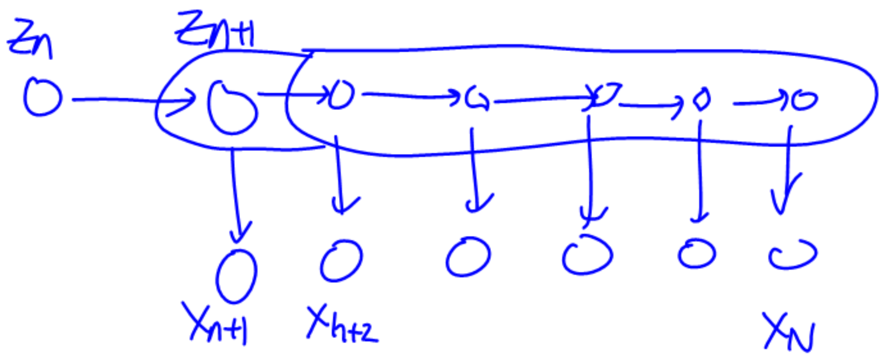


$p(x_{n+1}|z_{n+1})$

$p(z_{n+1}|z_n)$

$$\underbrace{\sum_{z_n} \alpha(z_n)\, p(z_{n+1}|z_n)\, p(x_{n+1}|z_{n+1})} = \sum_{z_n} p(x_1 \dots x_{n+1}, z_n, z_{n+1})$$

$$p(z_{n+1}|z_{n+1}) \sum_{z_n} \alpha(z_n)\, p(z_{n+1}|z_n) = p(x_1 \dots x_{n+1}, z_{n+1})$$
$$= \alpha(z_{n+1})$$

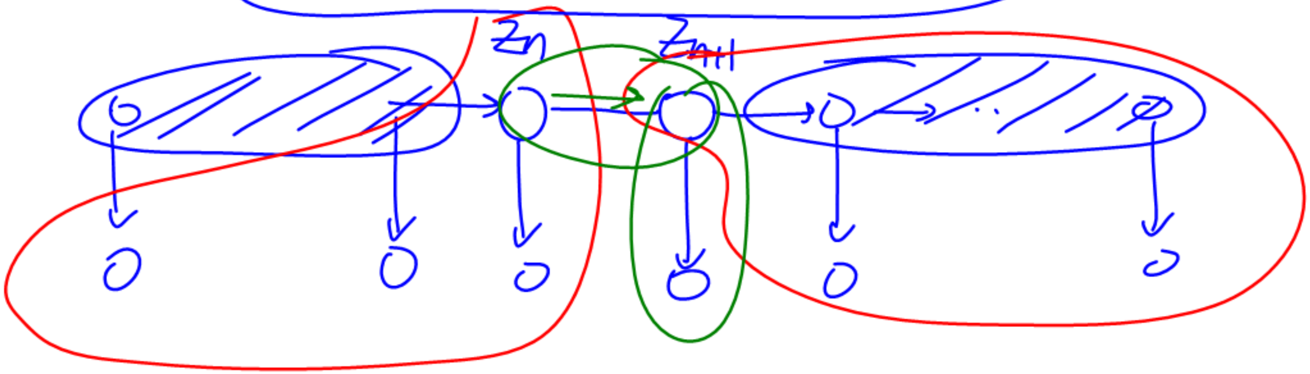


$$\beta(z_n) = p(x_{n+1} \cdots x_N \mid z_n)$$

$$= \sum_{z_{n+1}} p(x_{n+1} \cdots x_N, z_{n+1} \mid z_n)$$

$$= \sum_{z_{n+1}} \underbrace{p(x_{n+2}, \cdots, x_N \mid z_{n+1})}_{\beta(z_{n+1})} \, p(x_{n+1} \mid z_{n+1}) \, p(z_{n+1} \mid z_n)$$

$$\boxed{P(z_{n-1,j}\, z_{n,k} \mid X)}$$



$P(x_1 \dots x_n, z_n)$ $\qquad$ $P(x_{n+2}, \dots, x_N \mid z_{n+1})$

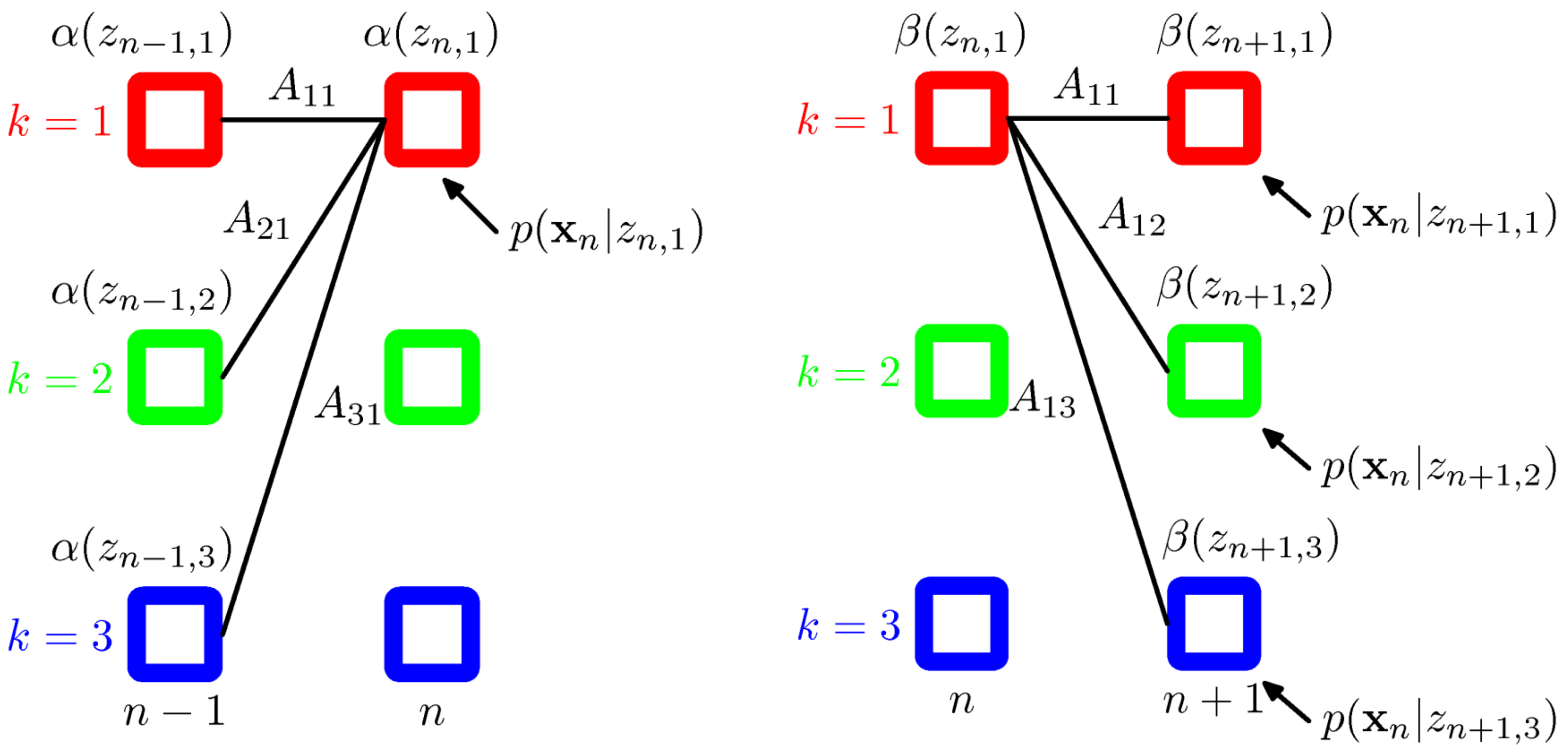
$=$

$\alpha(z_n)$

$$\propto \alpha(z_n)\, P(z_{n+1} \mid z_n)\, P(x_{n+1} \mid z_{n+1})\, \beta(z_{n+1})$$

Forward-Backward Algorithm

- Forward and Backward computations



E-M for HMMs

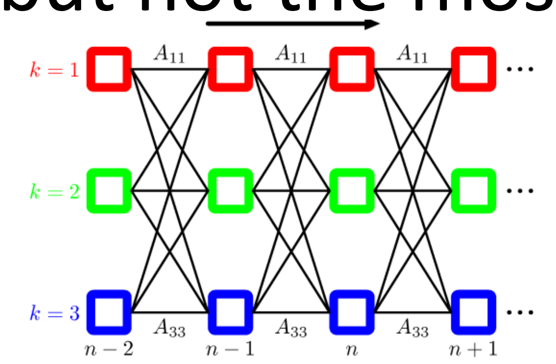
- The E-Step estimates the latent variables

$$p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$$

- The M-Step updates the parameters

$$\theta = \{\pi, \mathbf{A}, \phi\}$$

- After convergence, we have the maximum likelihood values of all these, but not the most likely path



Q. Derive the update rule for M-step

Estep: $Q(z_1) \propto \alpha(z_1)\,\beta(z_1)$

Mstep: $\underset{\pi_k}{\overset{\parallel}{P(z_1=k)}} = \dfrac{\text{Counts}(z_1=k)}{\sum_{k'} \text{counts}(z_1=k')} = \dfrac{\alpha(z_1=k)\;\beta(z_1=k)}{\sum_j \alpha(z_1=j)\;\beta(z_1=j)}$

$P(x_n \mid z_n=k) \Rightarrow$

$\parallel$

$\mathcal{N}(x_n \mid \mu_k, \Sigma_k)$

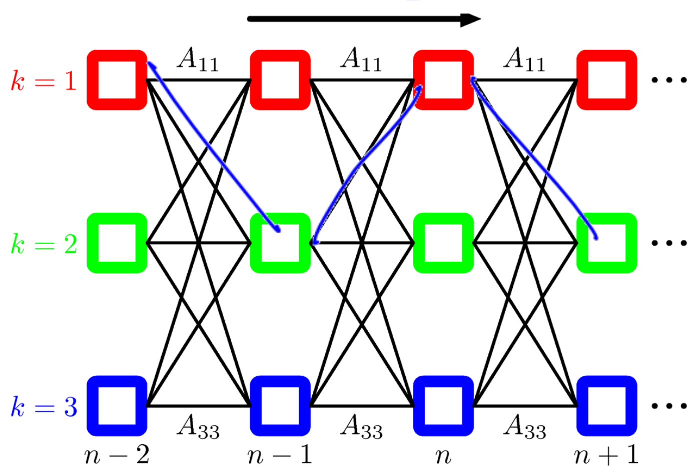
$$\mu_k = \frac{\sum_{n=1}^{N} x_n \, Q(z_n=k)}{\sum_{n=1}^{N} Q(z_n=k)}$$

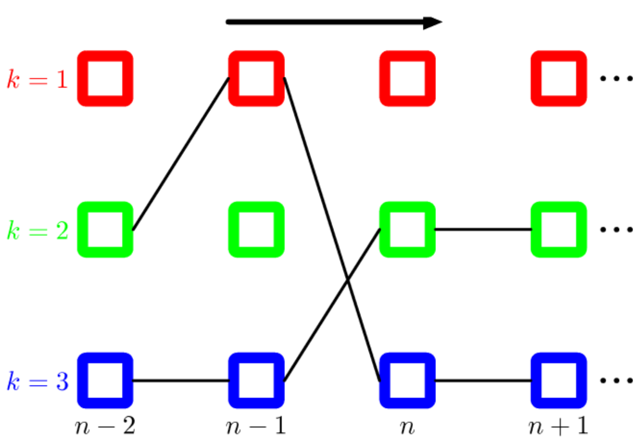
$$\Sigma_k = \frac{\sum_{n=1}^{N} (x_n-\mu_k)(x_n-\mu_k)\; Q(z_n=k)}{\sum_{n=1}^{N} Q(z_n=k)}$$

"state specific mean / covariance"

The Viterbi Algorithm

- This assumes that we have the HMM model including its parameters $\theta = \{\pi, \mathbf{A}, \phi\}$
- We are given the sequence X of observations and we want the most likely sequence Z of states. $\max_{Z} p(Z|X)$





The Viterbi Algorithm

- For each state in z_n, keep track of
  - the probability of reaching that state,
  - the most likely path for reaching that state, and
  - the probability of that path (the Viterbi path).
- This can be updated to z_{n+1} in K^2 time.
  - Multiply by the emission probability of $\mathbf{x}_n$,
  - and all possible transition probabilities.

Next

- **Reinforcement Learning**
  - Four lectures from the Sutton & Barto book
    - The RL problem and the MDP solution approach
    - Finding optimal policies: DP and MC
    - Finding optimal policies: temporal differences
    - Generalization and function approximation

Other Material on Learning

- **Mitchell, *Machine Learning*, 1997.**
- **From Russell & Norvig, *Artificial Intelligence: A Modern Approach*, second edition, 2003**
  - Ch.18: Decision trees, ensemble methods, and computational learning theory.
  - Ch.19: Learning and prior knowledge.
  - Ch.20: Statistical learning (cf. Bishop)
  - Ch.21: Reinforcement learning (cf. Sutton & Barto)
- **From Bishop, PRML.**
  - Ch.14: Combining models (including boosting)