13. Sequential Data

박찬연

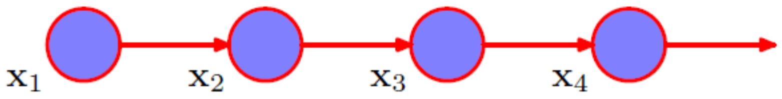
Introduction

- Here we focus on the case where the distribution from which sequential data is generated remains the same (stationary generative distribution)
- Here we consider Markov models, whose future predictions are independent of all but the most recent observations.
- Here we focus on the two state space models.
    - Hidden Markov Model (HMM)
        - Latent variables are discrete
    - Linear Dynamical System (LDS)
        - Latent variables are Gaussian
    - Described by directed graph having a tree structure.
    - Inference can be performed using sum-product algorithm.

13.1 Markov Models

- First-order Markov chain

Figure 13.3 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 $\mathbf{x}_{n-1}$.



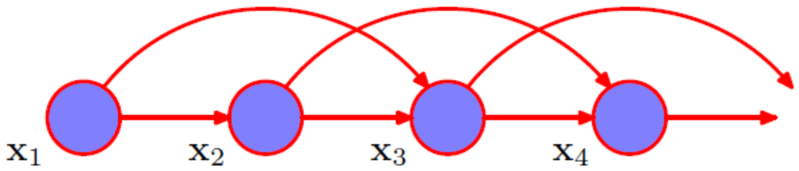
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}). \qquad (13.2)$$

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

13.1 Markov Models

- Higher-order Markov chain
  - Second-order Markov chain

Figure 13.4 A second-order Markov chain, in which the conditional distribution of a particular observation $\mathbf{x}_n$ depends on the values of the two previous observations $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$.

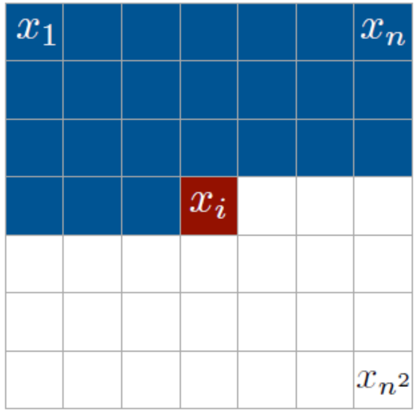


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{n=3}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}). \qquad (13.4)$$

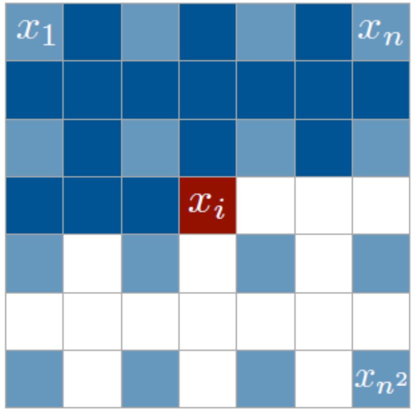
- Increased flexibility requires exponentially larger number of parameters.
  - 1st order Markov chain
    - K discrete variables → $K-1$ parameters for each $\mathbf{x}_{n-1}$ of $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ because $\sum_i p(\mathbf{x}_i|\mathbf{x}_j) = 1$ → total $K(K-1)$ parameters.
  - Mth order Markov Chain
    - $p(\mathbf{x}_n|\mathbf{x}_{n-M}, \ldots, \mathbf{x}_{n-1})$ → total $K^M(K-1)$ parameters.
    - Impractical for large M.

13.1 Markov Models

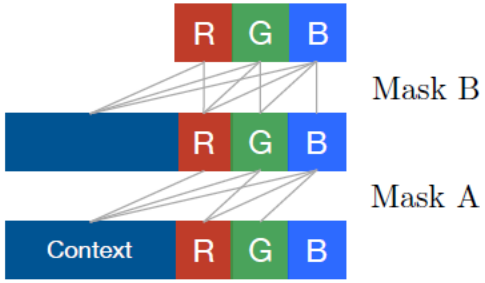
- For continuous variables, can use linear-Gaussian conditional distributions.
  - Each node has a Gaussian distribution whose mean is a linear function of its parents.
  - Known as autoregressive model.
- Or use a parametric model for $p(\mathbf{x}_n|\mathbf{x}_{n-M},\dots,\mathbf{x}_{n-1})$ such as a neural network.
  - PixelRNN



Context



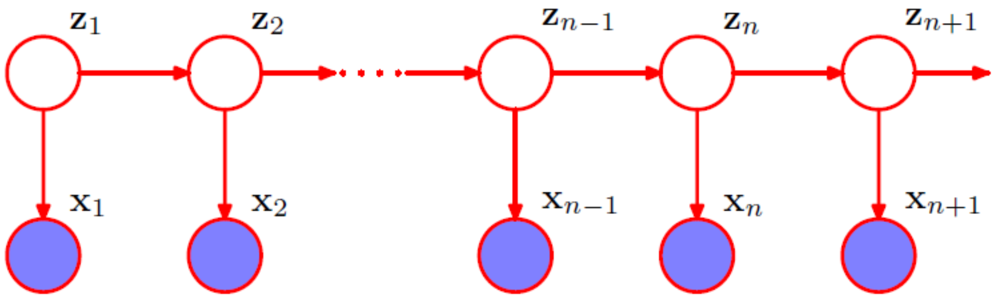
Multi-scale context



13.1 Markov Models

- State space model
  - Markov chain of latent variables.
  - A model for sequences that is not limited by the Markov assumption (that the observation is independent of all previous observations but the M most recent ones) but can be specified using a limited number of parameters.

Figure 13.5 We can represent sequential data using a Markov chain of latent variables, with each observation conditioned on the state of the corresponding latent variable. This important graphical structure forms the foundation both for the hidden Markov model and for linear dynamical systems.



$$\mathbf{z}_{n+1} \perp\!\!\!\perp \mathbf{z}_{n-1} \mid \mathbf{z}_n. \qquad (13.5)$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n). \qquad (13.6)$$

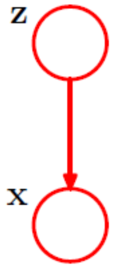
13.1 Markov Models

- Using d-separation, there is always a path connecting $\mathbf{x}_n$ and $\mathbf{x}_m$ via the latent variables.
- Therefore predictions for $\mathbf{x}_n$ depends on all previous observations, and they do not satisfy the Markov property at any order.
- Hidden Markov Model (HMM)
    - Latent variables are discrete
- Linear Dynamical System (LDS)
    - Latent variables are Gaussian

13.2 Hidden Markov Model

- A single time slice of HMM
  - corresponds to a mixture distribution, with component densities given by $p(\mathbf{x}|\mathbf{z})$.
  - can be interpreted as an extension of a mixture model whose choice of mixture component for each observation is not selected independently but depends on the choice of component for the previous observation.

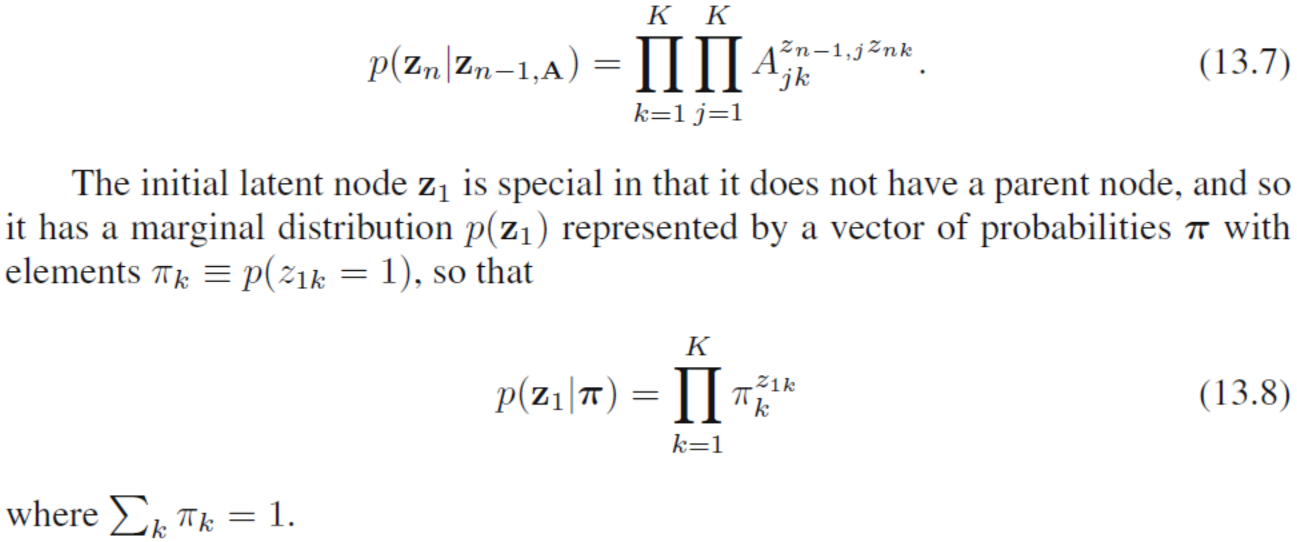
Figure 9.4 Graphical representation of a mixture model, in which the joint distribution is expressed in the form $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.



$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (9.12)$$

13.2 Hidden Markov Model

- The latent variables are the discrete multinomial variables $\mathbf{z}_n$ describing which component of the mixture is responsible for generating the corresponding observation $\mathbf{x}_n$.
- $\mathbf{z}_n$ depends on $\mathbf{z}_{n-1}$ via $p(\mathbf{z}_n|\mathbf{z}_{n-1})$, which corresponds to a matrix $\mathbf{A}$ satisfying
    - $A_{jk} \equiv p(z_{nk} = 1\,|z_{n-1,j} = 1)$, $0 \leq A_{jk} \leq 1$, $\sum_k A_{jk} = 1$,
- $\mathbf{A}$ has $K(K-1)$ parameters.



$$p(\mathbf{z}_n|\mathbf{z}_{n-1,\mathbf{A}}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}. \tag{13.7}$$

The initial latent node $\mathbf{z}_1$ is special in that it does not have a parent node, and so it has a marginal distribution $p(\mathbf{z}_1)$ represented by a vector of probabilities $\boldsymbol{\pi}$ with elements $\pi_k \equiv p(z_{1k} = 1)$, so that

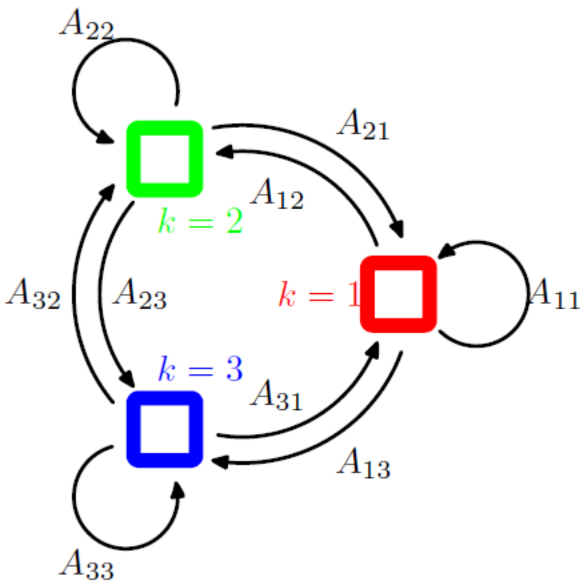
$$p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \tag{13.8}$$

where $\sum_k \pi_k = 1$.

13.2 Hidden Markov Model

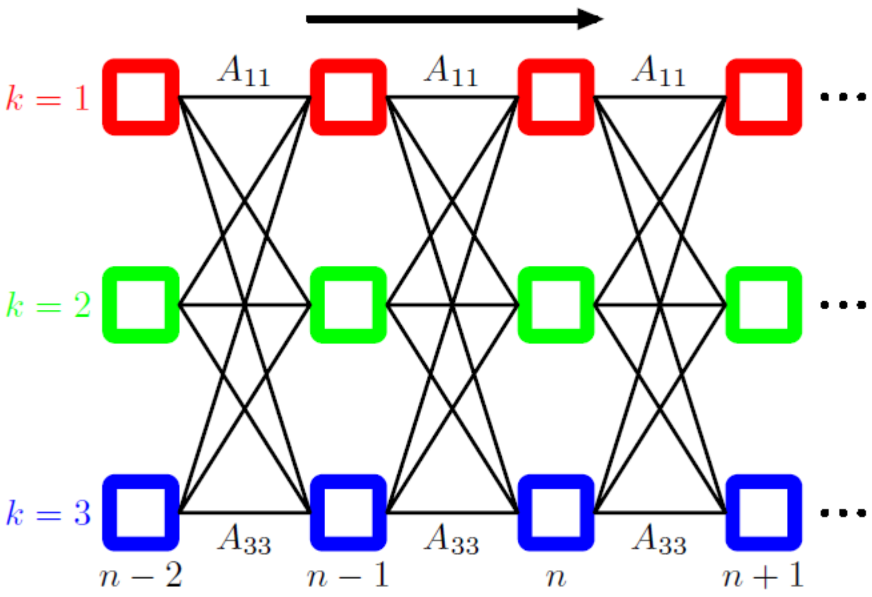
- State transition diagram

Figure 13.6 Transition diagram showing a model whose latent variables have three possible states corresponding to the three boxes. The black lines denote the elements of the transition matrix A_{jk}.



- Lattice / trellis diagram

Figure 13.7 If we unfold the state transition diagram of Figure 13.6 over time, we obtain a lattice, or trellis, representation of the latent states. Each column of this diagram corresponds to one of the latent variables $\mathbf{z}_n$.



13.2 Hidden Markov Model

- Emission probabilities $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi})$
  - Conditional distributions of the observed variables $\mathbf{x}_n$.
  - For continuous $\mathbf{x}_n$, can be for example given by

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{9.11}$$

  - Or by conditional probability tables if $\mathbf{x}_n$ is discrete.
  - For an observed $\mathbf{x}_n$ and given $\boldsymbol{\phi}$, this is a vector of K numbers. Therefore,

$$p(\mathbf{x}_n|\mathbf{z}_n, \phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}. \tag{13.9}$$

13.2 Hidden Markov Model

- Focus on homogeneous models that share the same $\mathbf{A}$ and $\boldsymbol{\phi}$
- Then the joint probability distribution over both latent and observed variables is

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi) \qquad (13.10)$$

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

- The discussion of HMM here is independent of the particular choice of $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi})$, the model is tractable for a wide range of emission distributions including
  - Discrete tables
  - Gaussians
  - Mixture of Gaussians
  - Neural networks

13.2 Hidden Markov Model

- Generative viewpoint
  - Mixture of Gaussians
    1. Choose a component at random with probability given by π_k,
    2. Generate a sample vector $\mathbf{x}$ from the Gaussian component
    3. Repeat 1 & 2 N times.
  - HMM
    1. Choose $\mathbf{z}_1$ according to π_1 then sample $\mathbf{x}_1$ according to $p(\mathbf{x}_1|\mathbf{z}_1, \boldsymbol{\phi})$.
    2. Choose $\mathbf{z}_2$ according to $p(\mathbf{z}_2|\mathbf{z}_1) \equiv \mathbf{A}$ using $\mathbf{z}_1$, then sample $\mathbf{x}_\mathbf{2}$ according to $p(\mathbf{x}_1|\mathbf{z}_2, \boldsymbol{\phi})$
    3. Continue until we sample $\mathbf{x}_\mathrm{N}$.
  - This is an example of ancestral sampling of a directed graphical model, Section 8.1.2.

13.2 Hidden Markov Model

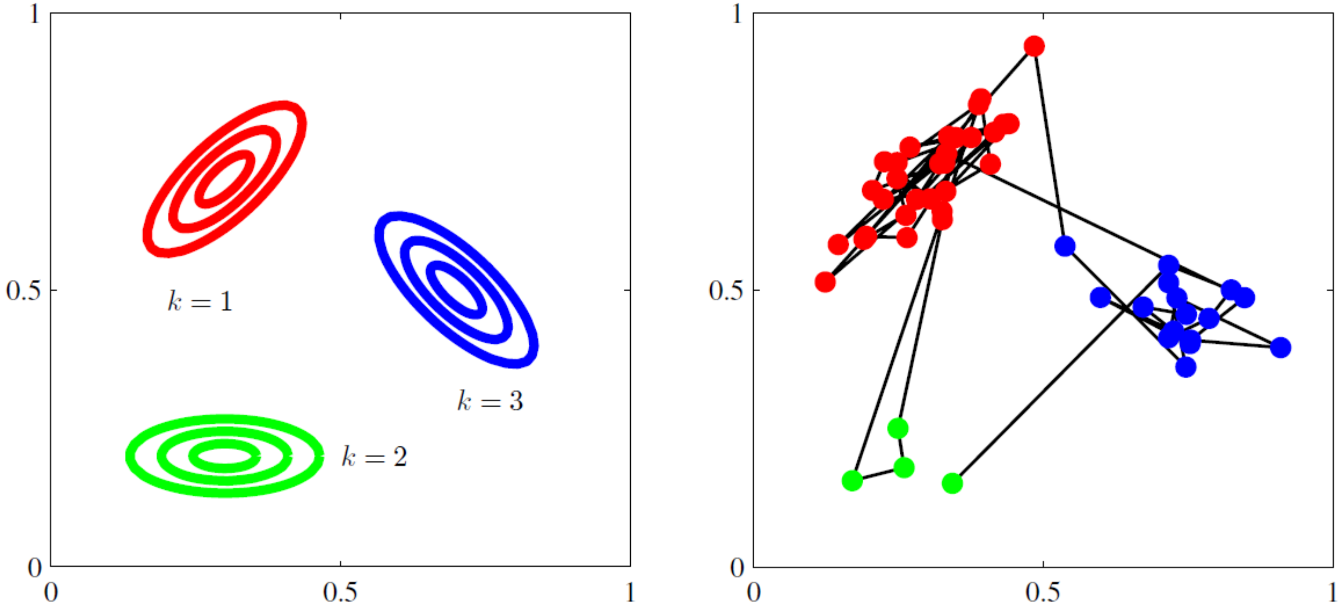


Figure 13.8 Illustration of sampling from a hidden Markov model having a 3-state latent variable $\mathbf{z}$ and a Gaussian emission model $p(\mathbf{x}|\mathbf{z})$ where $\mathbf{x}$ is 2-dimensional. (a) Contours of constant probability density for the emission distributions corresponding to each 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.

13.2 Hidden Markov Model

- Variants of the standard HMM
  - Left-to-right: $A_{jk} = 0$ if $k < j$.

Figure 13.9 Example of the state transition diagram for a 3-state left-to-right hidden Markov model. Note that once a state has been vacated, it cannot later be re-entered.

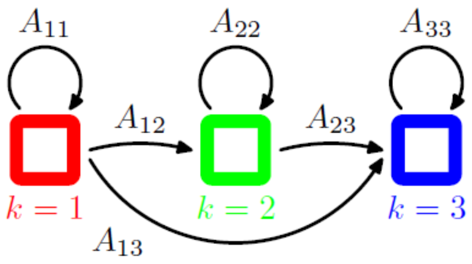
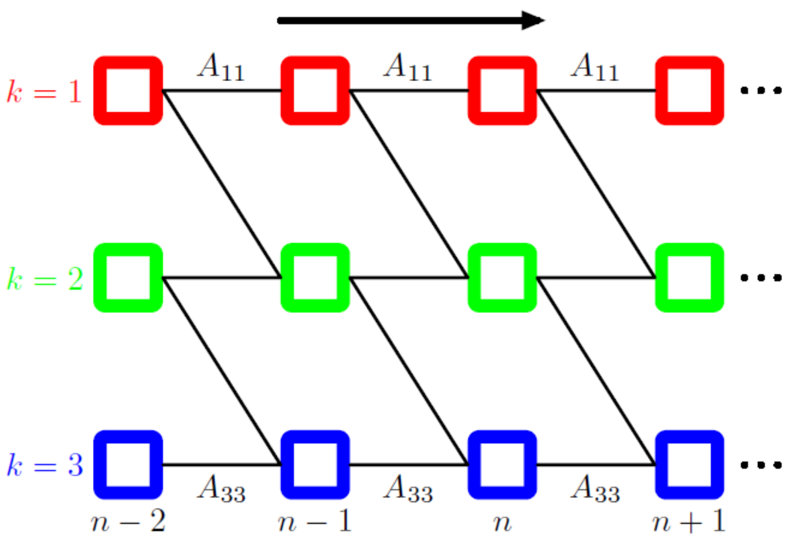


Figure 13.10 Lattice diagram for a 3-state left-to-right HMM in which the state index k is allowed to increase by at most 1 at each transition.



13.2.1 Maximum likelihood for the HMM

- Using observed data $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ determine the parameters of an HMM using ML.
- The likelihood function is obtained by marginalizing the joint distribution (13.10) over the latent variables.

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi) \qquad (13.10)$$

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \qquad (13.11)$$

- Cannot treat the latent variables independently, not cannot do the summation explicitly because the cost is exponential in N.
- Apply a similar technique to Section 8.4.1 to make the cost scale linearly in N.

13.2.1 Maximum likelihood for the HMM

- Use EM algorithm, starts with some initial model parameters θ_{old}.
- E step
  - Take θ_{old} and find the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta_{\text{old}})$ of the latent variables.
  - Use $p(\mathbf{Z}|\mathbf{X}, \theta_{\text{old}})$ to evaluate the expectation of $\log p(\mathbf{X}, \mathbf{Z}|\theta)$ as a function of θ.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \qquad (13.12)$$

13.2.1 Maximum likelihood for the HMM

- E step
  - Take θ_{old} and find the posterior distribution $p(\mathbf{Z}|\mathbf{X},\theta_{\text{old}})$ of the latent variables.
  - Use $p(\mathbf{Z}|\mathbf{X},\theta_{\text{old}})$ to evaluate the expectation of $\log p(\mathbf{X},\mathbf{Z}|\theta)$ as a function of θ.

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}). \qquad (13.12)$$

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \qquad (13.13)$$

$$\xi(\mathbf{z}_{n-1},\mathbf{z}_n) = p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X},\boldsymbol{\theta}^{\text{old}}). \qquad (13.14)$$

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{nk} \qquad (13.15)$$

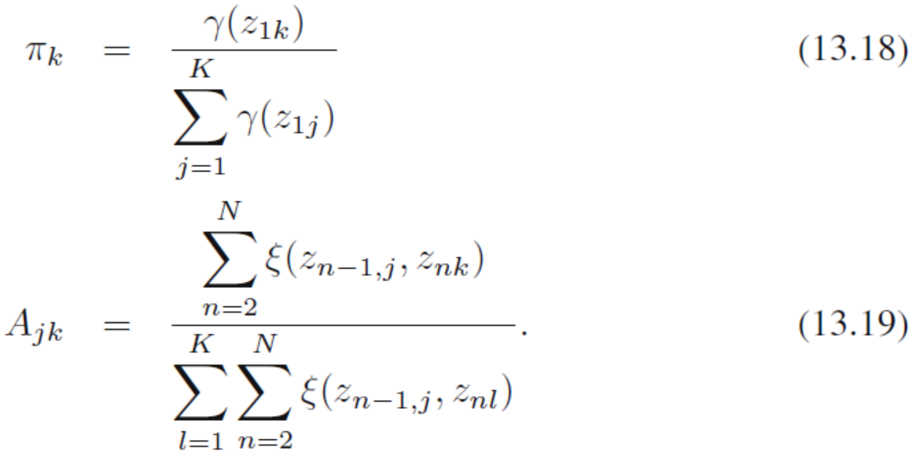
$$\xi(z_{n-1,j},z_{nk}) = \mathbb{E}[z_{n-1,j}z_{nk}] = \sum_{\mathbf{z}} \boxed{\gamma(\mathbf{z}) z_{n-1,j} z_{nk}.} \qquad (13.16)$$

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right] \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\boldsymbol{\phi}) \qquad (13.10)$$

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

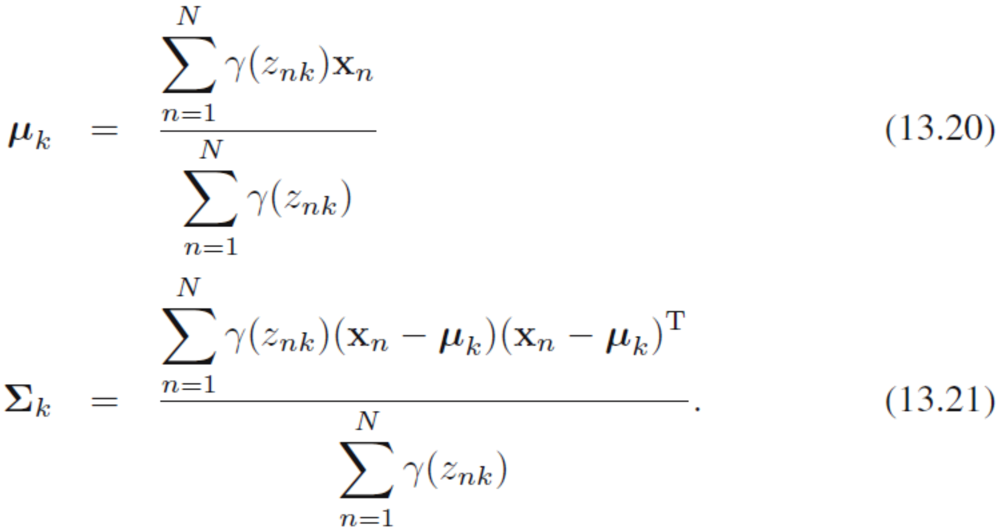
13.2.1 Maximum likelihood for the HMM

- M step
  - Maximize Q$(\theta, \theta_{\text{old}})$ w.r.t $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}\}$.
  - Straightforward for $\boldsymbol{\pi}, \mathbf{A}$ using Lagrange multiplier
  - For $\boldsymbol{\phi}$, we assume that $\boldsymbol{\phi}_k$ are independent and the maximizations of $\gamma(z_{nk}) \log p(\mathbf{x}_n|\boldsymbol{\phi}_k)$ can be done efficiently.
  - For $p(\mathbf{x}_n|\boldsymbol{\phi}_k) = N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,



$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \tag{13.18}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}. \tag{13.19}$$



$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{13.20}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}. \tag{13.21}$$

13.2.2 The forward-backward algorithm

- Seek an efficient procedure for evaluating $\gamma(z_{nk})$ and $\xi(z_{n-1,j}, z_{nk})$.
- The graph for the HMM is a tree, so we can use a two-stage message passing algorithm of Section 8.4., which is known as the forward-backward algorithm in the context of HMM.
- This is independent of the form of the emission density $p(\mathbf{x}|\mathbf{z})$, all we need is the values of $p(\mathbf{x}_n|\mathbf{z}_n)$ for each value of $\mathbf{z}_n$ for every n.

13.2.2 The forward-backward algorithm

- $\gamma(z_{nk})$
  - Using d-separation,

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{X})}. \tag{13.32}$$

$$\gamma(\mathbf{z}_n) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.33}$$

$$\alpha(\mathbf{z}_n) \equiv p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n) \tag{13.34}$$

$$\beta(\mathbf{z}_n) \equiv p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n). \tag{13.35}$$

  - Each of $\gamma(\mathbf{z}_n)$, $\alpha(\mathbf{z}_n)$, $\beta(\mathbf{z}_n)$ is a set of K numbers.

13.2.2 The forward-backward algorithm

- Recursion relation of $\alpha(\mathbf{z}_n)$
  - Using d-separation,

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1}). \qquad (13.36)$$

  - Initial condition

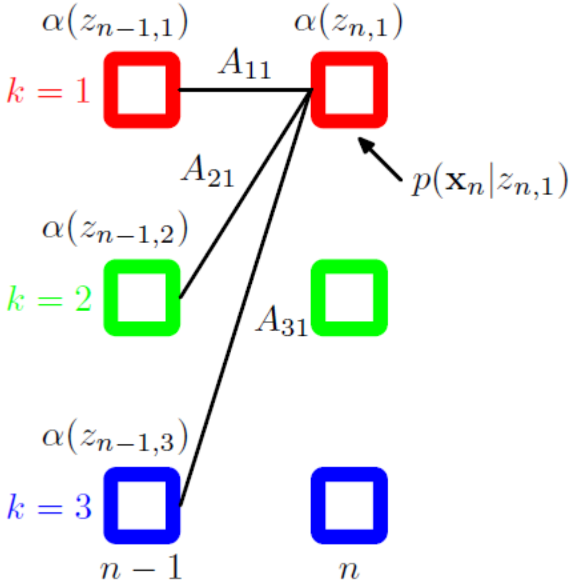
$$\alpha(\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\mathbf{x}_1|\boldsymbol{\phi}_k)\}^{z_{1k}} \qquad (13.37)$$

$$p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \qquad (13.8)$$

$$p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{z_{nk}}. \qquad (13.9)$$

  - Each step $O(K^2)$, overall $O(K^2N)$.

Figure 13.12 Illustration of the forward recursion (13.36) for evaluation of the α variables. In this fragment of the lattice, we see that the quantity $\alpha(z_{n1})$ is obtained by taking the elements $\alpha(z_{n-1,j})$ of $\alpha(\mathbf{z}_{n-1})$ at step $n-1$ and summing them up with weights given by A_{j1}, corresponding to the values of $p(\mathbf{z}_n|\mathbf{z}_{n-1})$, and then multiplying by the data contribution $p(\mathbf{x}_n|z_{n1})$.



13.2.2 The forward-backward algorithm

- Recursion relation of $\beta(\mathbf{z}_n)$

$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n). \qquad (13.38)$$

  - Initial condition: $\beta(\mathbf{z}_N) = 1$.

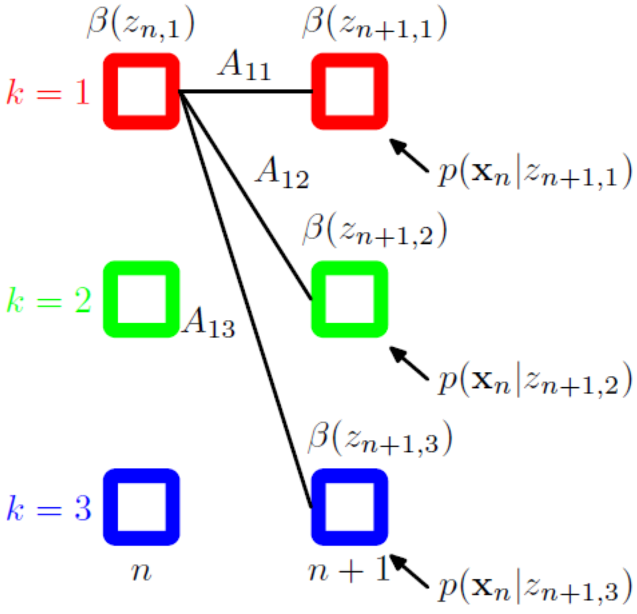
- Likelihood function $p(\mathbf{X})$
  - Not needed for M step but Monitored during the EM optimization.

$$p(\mathbf{X}) = \sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n). \qquad (13.41)$$

$$p(\mathbf{X}) = \sum_{\mathbf{z}_N} \alpha(\mathbf{z}_N). \qquad (13.42)$$

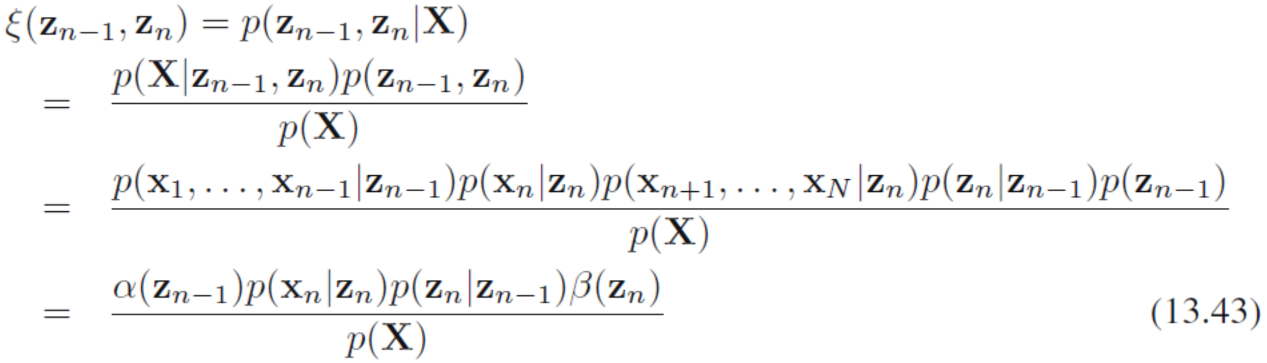
  - Computational cost reduced from exponential to linear in N.

Figure 13.13 Illustration of the backward recursion (13.38) for evaluation of the β variables. In this fragment of the lattice, we see that the quantity $\beta(z_{n1})$ is obtained by taking the components $\beta(z_{n+1,k})$ of $\beta(\mathbf{z}_{n+1})$ at step $n+1$ and summing them up with weights given by the products of A_{1k}, corresponding to the values of $p(\mathbf{z}_{n+1}|\mathbf{z}_n)$ and the corresponding values of the emission density $p(\mathbf{x}_n|z_{n+1,k})$.



13.2.2 The forward-backward algorithm

- $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$
  - Using conditional independence from d-separation, together with $\alpha(\mathbf{z}_n)$ and $\beta(\mathbf{z}_n)$,



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

13.2.2 The forward-backward algorithm

- Summary of EM for HMM
- E step
    1. Make an initial selection of θ_{old} where $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}\}$.
        - $\boldsymbol{\pi}, \mathbf{A}$ often initialized uniformly or randomly from a uniform distribution.
        - $\boldsymbol{\phi}$ according to the form of the distribution, for example in the case of Gaussians, using the means and the covariances of K-means clusters.
    2. Run the forward α recursion and the backward β recursion.
    3. Use the result of step 2 to evaluate $\gamma(z_{nk})$ and $\xi(z_{n-1,j}, z_{nk})$.
    4. Evaluate the likelihood $p(\mathbf{X})$
- M step
    - Maximize $\mathrm{Q}(\theta, \theta_{\mathrm{old}})$ to find a revised set of parameters θ_{new}.
- Alternate E and M steps until some convergence, for example when the change of $p(\mathbf{X})$ is below some threshold.

13.2.2 The forward-backward algorithm

- Number of data points (=total length of the training sequences) should be long enough considering the number of parameters $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}\}$, which are shared at every when considering homogeneous models.
- Calculation of predictive distribution
  - The previous calculations correspond to training, now the inference using the trained model.
  - Again using the conditional independence from the d-separation,

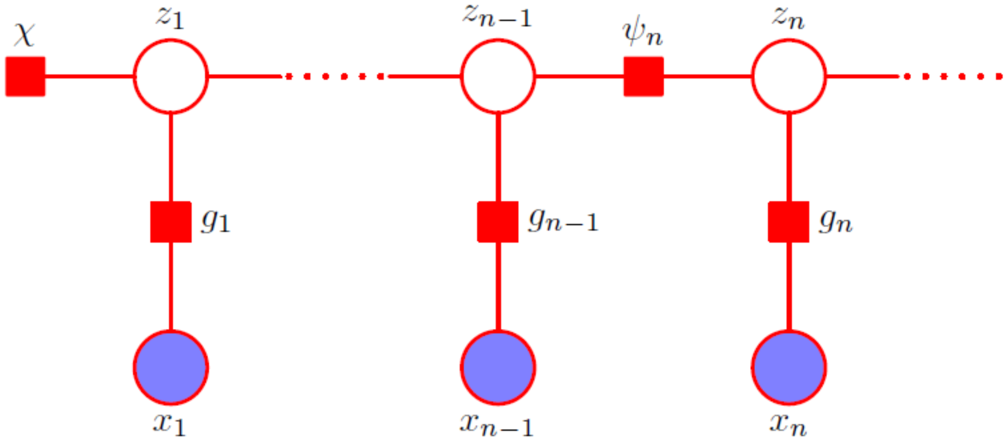
$$p(\mathbf{x}_{N+1}|\mathbf{X}) \;=\; \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}, \mathbf{z}_{N+1}|\mathbf{X}) \;=\; \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}, \mathbf{z}_N|\mathbf{X}) \;=\; \frac{1}{p(\mathbf{X})} \sum_{\mathbf{z}_{N+1}} p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}) \sum_{\mathbf{z}_N} p(\mathbf{z}_{N+1}|\mathbf{z}_N)\alpha(\mathbf{z}_N) \quad (13.44)$$

  - We don’t need β?
  - All data is summarized in the K values of $\alpha(\mathbf{z}_N)$, prediction can be carried forward using a fixed amount of storage.

13.2.3 The sum-product algorithm for the HMM

- The directed graph that represents the HMM is a tree and therefor can find local marginals for the hidden variables using the sum-product algorithm.
- Transform the directed graph into a factor graph.

Figure 13.14 A fragment of the factor graph representation for the hidden Markov model.

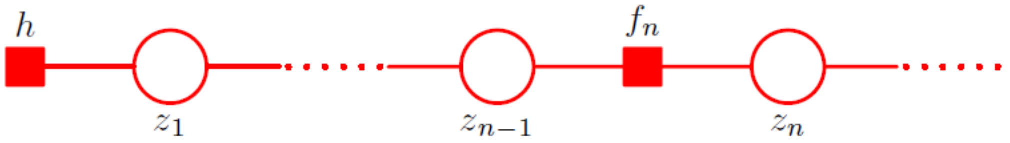


13.2.3 The sum-product algorithm for the HMM

- Simplify the factor graph by absorbing the emission probabilities $p(\mathbf{x}_n|\mathbf{z}_n)$ into the transition probability factors $f(\mathbf{z}_{n-1}, \mathbf{z}_n)$.

$$
\begin{aligned}
h(\mathbf{z}_1) &= p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) && (13.45) \\
f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n). && (13.46)
\end{aligned}
$$

Figure 13.15 A simplified form of factor graph to describe the hidden Markov model.



- First pass messages from the leaf node h to the root node $\mathbf{z}_N$, the final hidden variable.

13.2.3 The sum-product algorithm for the HMM

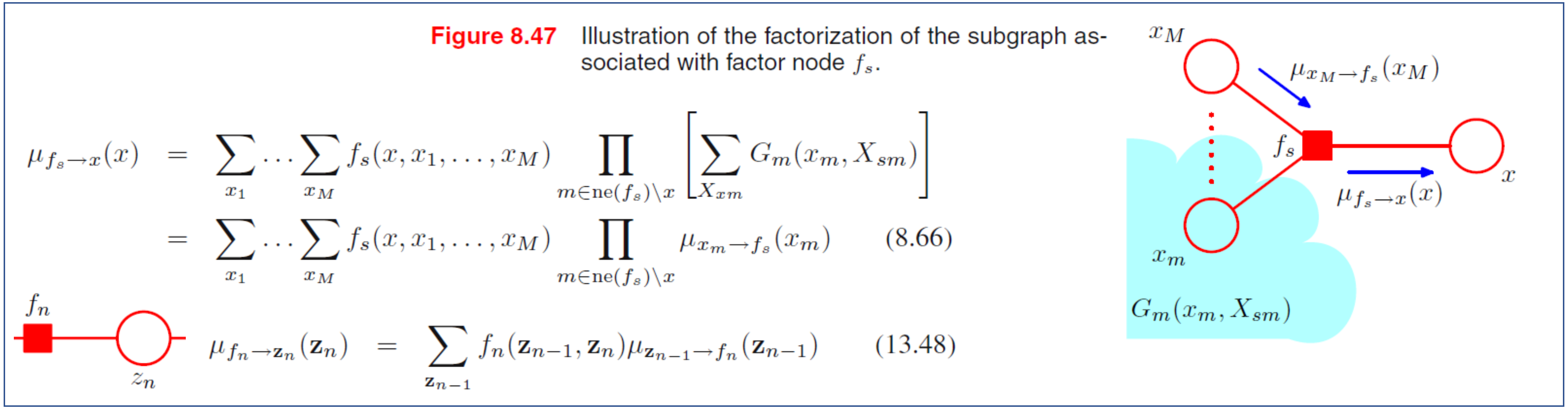


Figure 8.47 Illustration of the factorization of the subgraph associated with factor node f_s.

$$\begin{aligned}\mu_{f_s\to x}(x) &= \sum_{x_1}\cdots\sum_{x_M} f_s(x,x_1,\ldots,x_M)\prod_{m\in\mathrm{ne}(f_s)\setminus x}\left[\sum_{X_{xm}} G_m(x_m,X_{sm})\right]\\ &= \sum_{x_1}\cdots\sum_{x_M} f_s(x,x_1,\ldots,x_M)\prod_{m\in\mathrm{ne}(f_s)\setminus x}\mu_{x_m\to f_s}(x_m) \qquad (8.66)\end{aligned}$$

$$\mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} f_n(\mathbf{z}_{n-1},\mathbf{z}_n)\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1}) \qquad (13.48)$$

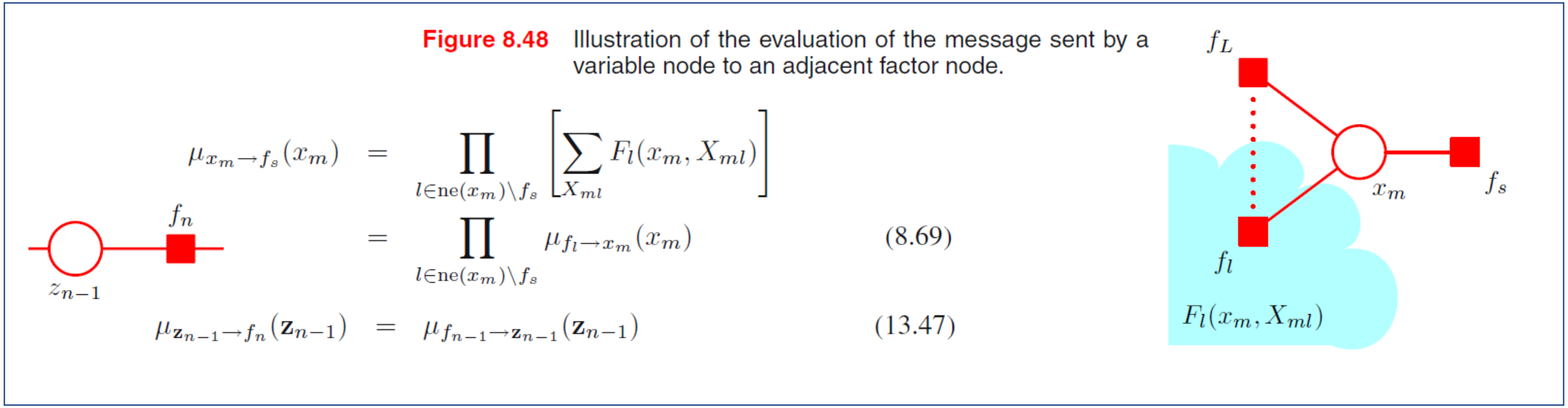
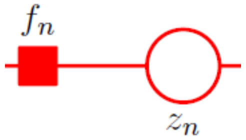


Figure 8.48 Illustration of the evaluation of the message sent by a variable node to an adjacent factor node.

$$\begin{aligned}\mu_{x_m\to f_s}(x_m) &= \prod_{l\in\mathrm{ne}(x_m)\setminus f_s}\left[\sum_{X_{ml}} F_l(x_m,X_{ml})\right]\\ &= \prod_{l\in\mathrm{ne}(x_m)\setminus f_s}\mu_{f_l\to x_m}(x_m) \qquad (8.69)\end{aligned}$$

$$\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1}) = \mu_{f_{n-1}\to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1}) \qquad (13.47)$$

13.2.3 The sum-product algorithm for the HMM



$$\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1}) = \mu_{f_{n-1}\to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1}) \tag{13.47}$$

$$\mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} f_n(\mathbf{z}_{n-1},\mathbf{z}_n)\mu_{\mathbf{z}_{n-1}\to f_n}(\mathbf{z}_{n-1}) \tag{13.48}$$

- Eliminate $\mu_{\mathbf{z}_{n-1}\to f_n}$ from (13.48) using (13.47)
- Get a recursion for the $f \to \mathbf{z}$ message of the form

$$\mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n-1}} f_n(\mathbf{z}_{n-1},\mathbf{z}_n)\mu_{f_{n-1}\to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1}). \tag{13.49}$$

- Then using

$$f_n(\mathbf{z}_{n-1},\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n). \tag{13.46}$$

- And defining

$$\alpha(\mathbf{z}_n) = \mu_{f_n\to \mathbf{z}_n}(\mathbf{z}_n) \tag{13.50}$$

- We get the same recursion relation for α,

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}). \tag{13.36}$$

- And the same initial condition $\alpha(\mathbf{z}_1) = h(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)$

13.2.3 The sum-product algorithm for the HMM

- Next, consider the message propagated from the root node $\mathbf{z}_N$ back to the leaf node h by eliminating the message of the type $\mathbf{z} \to f$

$$\mu_{f_{n+1}\to f_n}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} f_{n+1}(\mathbf{z}_n, \mathbf{z}_{n+1})\mu_{f_{n+2}\to f_{n+1}}(\mathbf{z}_{n+1}) \qquad (13.51)$$

- Using

$$f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) \quad = \quad p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n). \qquad (13.46)$$

- And defining

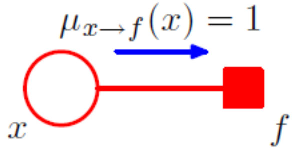
$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n). \qquad (13.38)$$

- We obtain the same recursion relation for β,

$$\beta(\mathbf{z}_n) = \mu_{f_{n+1}\to \mathbf{z}_n}(\mathbf{z}_n) \qquad (13.52)$$

- And the same initial condition from (8.70),

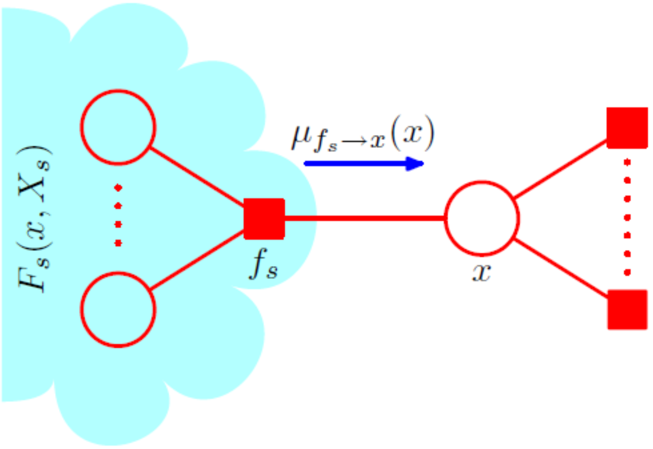
$$\mu_{\mathbf{z}_N\to f_N}(\mathbf{z}_N) = 1 \qquad \mu_{x\to f}(x) = 1 \qquad (8.70)$$



13.2.3 The sum-product algorithm for the HMM

- Computation of $\gamma(\mathbf{z}_n)$

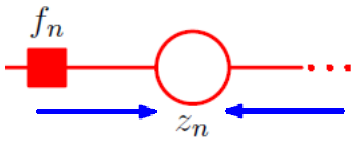
Figure 8.46 A fragment of a factor graph illustrating the evaluation of the marginal $p(x)$.



$$\begin{aligned} p(x) &= \prod_{s \in \mathrm{ne}(x)} \left[\sum_{X_s} F_s(x, X_s) \right] \\ &= \prod_{s \in \mathrm{ne}(x)} \mu_{f_s \to x}(x). \end{aligned} \tag{8.63}$$

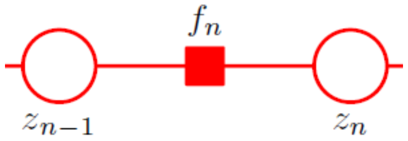
$$\mu_{f_s \to x}(x) \equiv \sum_{X_s} F_s(x, X_s) \tag{8.64}$$

$$p(\mathbf{z}_n, \mathbf{X}) = \mu_{f_n \to \mathbf{z}_n}(\mathbf{z}_n) \mu_{f_{n+1} \to \mathbf{z}_n}(\mathbf{z}_n) = \alpha(\mathbf{z}_n)\beta(\mathbf{z}_n). \tag{13.53}$$



$$\gamma(\mathbf{z}_n) = \frac{p(\mathbf{z}_n, \mathbf{X})}{p(\mathbf{X})} = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.54}$$

13.2.3 The sum-product algorithm for the HMM



- Computation of $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$

$$p(\mathbf{x}_s) = f_s(\mathbf{x}_s) \prod_{i \in \mathrm{ne}(f_s)} \mu_{x_i \to f_s}(x_i) \qquad (8.72)$$

- $p(\mathbf{z}_{n-1}, \mathbf{z}_n, \mathbf{X}) = f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) \cdot \mu_{\mathbf{z}_{n-1} \to f_n}(\mathbf{z}_{n-1}) \cdot \mu_{\mathbf{z}_n \to f_n}(\mathbf{z}_n)$
- Using (8.69),
  - $\mu_{\mathbf{z}_{n-1} \to f_n}(\mathbf{z}_{n-1}) = \mu_{f_{n-1} \to \mathbf{z}_{n-1}}(\mathbf{z}_{n-1})$
  - $\mu_{\mathbf{z}_n \to f_n}(\mathbf{z}_n) = \mu_{f_{n+1} \to \mathbf{z}_n}(\mathbf{z}_n)$

$$\begin{aligned} \mu_{x_m \to f_s}(x_m) &= \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \left[\sum_{X_{ml}} F_l(x_m, X_{ml}) \right] \\ &= \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m) \qquad (8.69) \end{aligned}$$

- Using the definitions

$$f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n). \qquad (13.46)$$

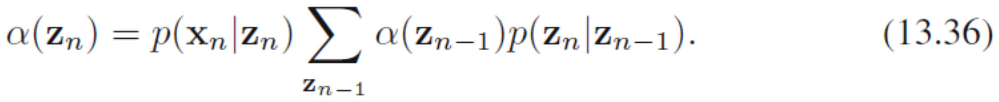
$$\alpha(\mathbf{z}_n) = \mu_{f_n \to \mathbf{z}_n}(\mathbf{z}_n) \qquad (13.50)$$

$$\beta(\mathbf{z}_n) = \mu_{f_{n+1} \to \mathbf{z}_n}(\mathbf{z}_n) \qquad (13.52)$$

- We get $p(\mathbf{z}_{n-1}, \mathbf{z}_n, \mathbf{X}) = p(\mathbf{z}_n | \mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n)\, \alpha(\mathbf{z}_{n-1}) \beta(\mathbf{z}_n)$. Then

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}) \beta(\mathbf{z}_n)}{p(\mathbf{X})} \qquad (13.43)$$

13.2.4 Scaling factors



$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1}). \qquad (13.36)$$

- At each step, $\alpha(\mathbf{z}_n)$ is obtained from $\alpha(\mathbf{z}_{n-1})$ by multiplying by probabilities that are often significantly less than 1, so α can go to zero exponentially quickly.
- Similar problem happens to RNN in the form of vanishing gradients.

13.2.4 Scaling factors

- Use rescaled versions of $\alpha(\mathbf{z}_n)$ of order 1.

$$\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\alpha(\mathbf{z}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)} \qquad (13.55)$$

- Well-behaved numerically because it is a probability distribution over K variables. By introducing scaling factors c_n

$$c_n = p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}). \qquad (13.56)$$

- We get the recursion equation

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}). \qquad (13.59)$$

- c_n is calculated as a normalization constant of the RHS of (13.59) at every stage of the forward message passing phase.

13.2.4 Scaling factors

- Rescaled versions of $\beta(\mathbf{z}_n)$ of order 1.

$$\widehat{\beta}(\mathbf{z}_n) = \frac{p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n)}{p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_n)}. \tag{13.61}$$

$$c_{n+1}\widehat{\beta}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \widehat{\beta}(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n). \tag{13.62}$$

- Use c_n that were previously computed in the α phase.
- From (13.57) likelihood function is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{m=1}^{n} c_m \tag{13.57}$$

$$p(\mathbf{X}) = \prod_{n=1}^{N} c_n. \tag{13.63}$$

- Required marginals are

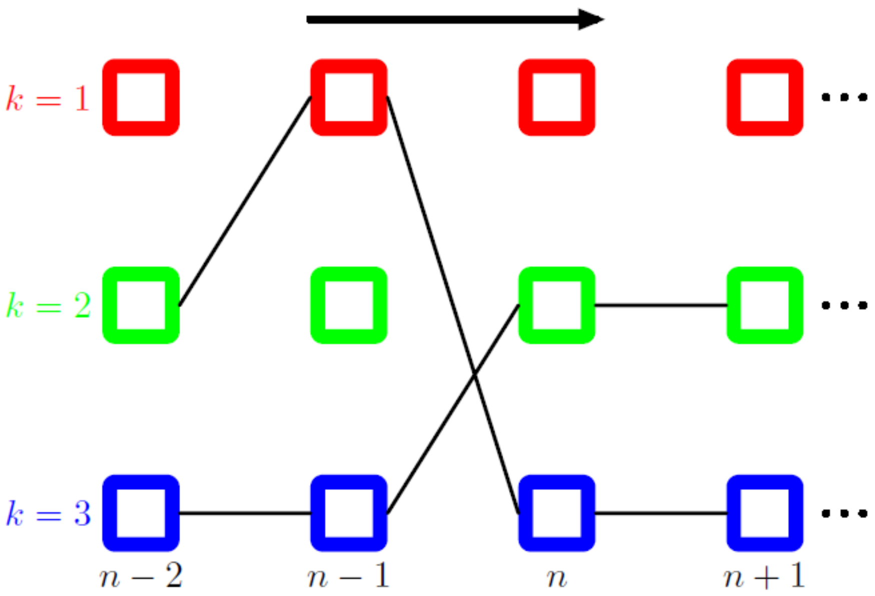
$$\begin{aligned} \gamma(\mathbf{z}_n) &= \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n) && (13.64) \\ \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= c_n \widehat{\alpha}(\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) p(\mathbf{z}_n|\mathbf{z}_{-1}) \widehat{\beta}(\mathbf{z}_n). && (13.65) \end{aligned}$$

13.2.5 The Viterbi algorithm

- $\mathbf{z}_n$ of an HMM have some meaningful interpretation.
- Finding the most probable sequence of $\mathbf{z}_n$ for a given sequence of $\mathbf{x}_n$
  - Because the graph is a directed tree, can be solved using the max-sum algorithm of Section 8.4.5
  - In the context of HMMs this is known as the Viterbi algorithm.
- The problem of finding the most probable sequence of latent states is not the same as that of finding the set of states that are individually the most probable.

13.2.5 The Viterbi algorithm

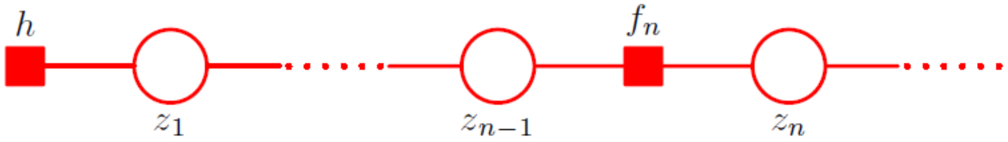
Figure 13.16 A fragment of the HMM lattice showing two possible paths. The Viterbi algorithm efficiently determines the most probable path from amongst the exponentially many possibilities. For any given path, the corresponding probability is given by the product of the elements of the transition matrix A_{jk}, corresponding to the probabilities $p(\mathbf{z}_{n+1}|\mathbf{z}_n)$ for each segment of the path, along with the emission densities $p(\mathbf{x}_n|k)$ associated with each node on the path.



- The number of possible paths grows exponentially with the length of the chain.
- The Viterbi algorithm searches the space of paths efficiently such that the cost grows linearly with the length of the chain.

13.2.5 The Viterbi algorithm

Figure 13.15 A simplified form of factor graph to describe the hidden Markov model.



$$\mu_{f\to x}(x) = \max_{x_1,\ldots,x_M}\left[\ln f(x,x_1,\ldots,x_M) + \sum_{m\in\mathrm{ne}(f_s)\backslash x}\mu_{x_m\to f}(x_m)\right] \quad (8.93)$$

$$\mu_{x\to f}(x) = \sum_{l\in\mathrm{ne}(x)\backslash f}\mu_{f_l\to x}(x). \quad (8.94)$$

$$\mu_{f_{n+1}\to\mathbf{z}_{n+1}}(\mathbf{z}_{n+1}) = \max_{\mathbf{z}_n}\left\{\ln f_{n+1}(\mathbf{z}_n,\mathbf{z}_{n+1}) + \mu_{\mathbf{z}_n\to f_{n+1}}(\mathbf{z}_n)\right\}. \quad (13.67)$$

$$\mu_{\mathbf{z}_n\to f_{n+1}}(\mathbf{z}_n) = \mu_{f_n\to\mathbf{z}_n}(\mathbf{z}_n) \quad (13.66)$$

- Applying (13.66) and (13.46) to (13.67), we have

$$f_n(\mathbf{z}_{n-1},\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n). \quad (13.46)$$

$$\omega(\mathbf{z}_{n+1}) = \ln p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) + \max_{\mathbf{z}_n}\left\{\ln p(\mathbf{x}_{+1}|\mathbf{z}_n) + \omega(\mathbf{z}_n)\right\} \quad (13.68)$$

$$\omega(\mathbf{z}_n) \equiv \mu_{f_n\to\mathbf{z}_n}(\mathbf{z}_n)$$

- Initialization

$$\mu_{x\to f}(x) = 0 \quad (8.95)$$

$$\mu_{f\to x}(x) = \ln f(x) \quad (8.96)$$

$$h(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) \quad (13.45)$$

$$\omega(\mathbf{z}_1) = \ln p(\mathbf{z}_1) + \ln p(\mathbf{x}_1|\mathbf{z}_1). \quad (13.69)$$

13.2.5 The Viterbi algorithm

- $\omega(\mathbf{z}_n)$ have the probabilistic interpretation

$$\omega(\mathbf{z}_n) = \max_{\mathbf{z}_1,\ldots,\mathbf{z}_{n-1}} p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n). \qquad (13.70)$$

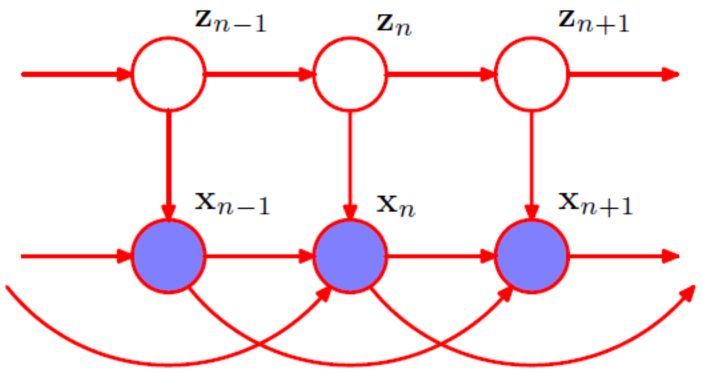
- Once we have completed the final maximization over $\mathbf{z}_N$, we obtain the value of $p(\mathbf{X},\mathbf{Z})$ corresponding to the most probable path.
- Now we want to find the corresponding sequence of $\mathbf{z}_n$.
- To do this, simply use back-tracking procedure discussed in Section 8.4.5.
- Intuitively, the Viterbi algorithm keeps only the most probable path for each state at every stage. When we reach the final step N we find the most probable path, then we trace the path back through the lattice.

13.2.6 Extensions of the hidden Markov model

- Determining the parameters of HMMs using discriminative rather than ML by optimizing the cross-entropy.
- The probability of a given HMM to spend T steps in a state decays exponentially in T, which is unrealistic for many applications.
  - Set all diagonal A_{kk} to zero, and associate each state with a probability distribution $p(T|k)$ of possible duration times.

13.2.6 Extensions of the hidden Markov model

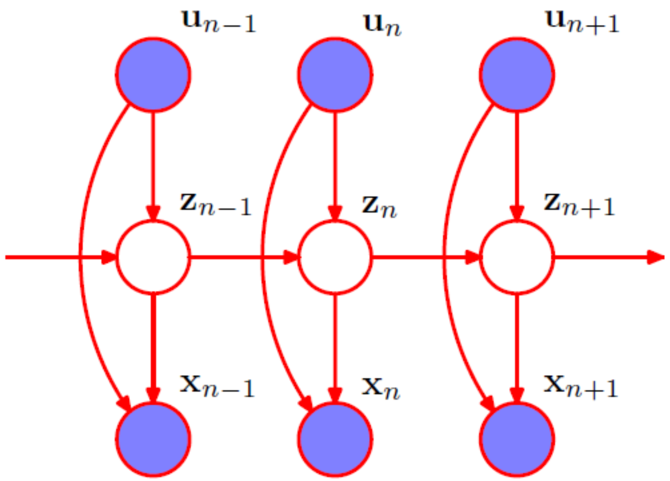
Figure 13.17 Section of an autoregressive hidden Markov model, in which the distribution of the observation $\mathbf{x}_n$ depends on a subset of the previous observations as well as on the hidden state $\mathbf{z}_n$. In this example, the distribution of $\mathbf{x}_n$ depends on the two previous observations $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$.



- HMM is poor at capturing long-range correlations between observed variables.
- Autoregressive HMM
- Number of additional links must be limited to avoid an excessive number of free parameters.
- Can use d-separation to see that, like HMM, $z_{n-1} \perp\!\!\!\perp z_{n+1} | z_n$
- Therefore can use a forward-backward recursion in the E step of the EM.

13.2.6 Extensions of the hidden Markov model

Figure 13.18 Example of an input-output hidden Markov model. In this case, both the emission probabilities and the transition probabilities depend on the values of a sequence of observations $\mathbf{u}_1, \ldots, \mathbf{u}_N$.



- Input-output HMM
- Extends the HMM framework to the domain of supervised learning for sequential data.
- Using d-separation, can show that the Markov property $z_{n-1} \perp\!\!\!\perp z_{n+1} | z_n$ holds.

13.2.6 Extensions of the hidden Markov model

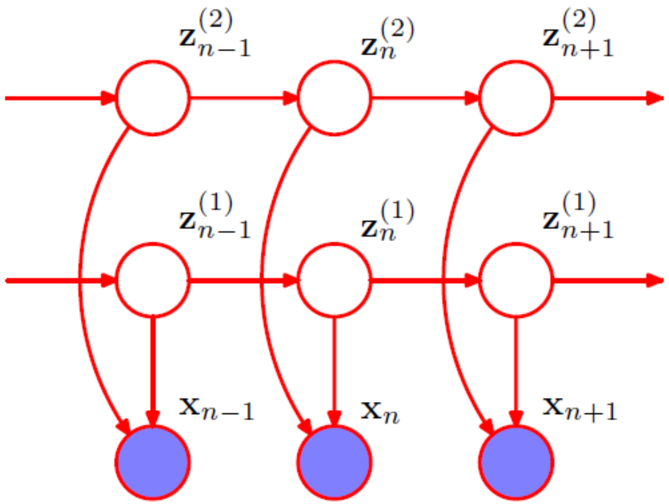


Figure 13.19 A factorial hidden Markov model comprising two Markov chains of latent variables. For continuous observed variables $\mathbf{x}$, one possible choice of emission model is a linear-Gaussian density in which the mean of the Gaussian is a linear combination of the states of the corresponding latent variables.

- Factorial HMM
- Multiple independent Markov chains of latent variables.
- Primary disadvantage in the additional complexity of training
  - due to dependencies between the latent chains and
  - due to the lack of the Markov property for the individual chains.

13.3 Linear Dynamical Systems

- The same graphical model as the HMM.
- Continuous latent variables
    - The summations of the sum-product algorithm becomes integral.
    - The general form of the inference algorithm will be the same as for the HMM.
- Key requirement for efficient inference: linear in the chain length.
    - This requires that the factors in (13.59) not to be complex at each stage.

$$\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n) = \frac{\alpha(\mathbf{z}_n)}{p(\mathbf{x}_1,\ldots,\mathbf{x}_n)} \tag{13.55}$$

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} \widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}). \tag{13.59}$$

- Therefore we need to consider distributions from the exponential family.

13.3 Linear Dynamical Systems

- We consider a linear-Gaussian state space model.
  - $\mathbf{z}_n$ and $\mathbf{x}_n$ are multivariate Gaussian distributions whose means are linear functions of the states of their parents in the graph.
  - A directed graph of linear-Gaussian units is equivalent to a joint Gaussian distribution over all of the variables.
- By contrast, if $p(\mathbf{x}_n|\mathbf{z}_n)$ is a mixture of K Gaussians, $\hat{\alpha}(\mathbf{z}_n)$ will be a mixture of K^{n-1} Gaussians, and exact inference will not be practical.
- HMM ↔ extensions of the mixture models of Chapter 9
- LDS ↔ generalization of the continuous latent variable models of Chapter 12 such as probabilistic PCA and factor analysis.

13.3 Linear Dynamical Systems

- The model is represented by a tree-structured directed graph, therefore inference problems can be solved using the sum-product algorithm.
  - Forward recursions: the Kalman filter equations.
  - Backward recursions: the Kalman smoother equations or RTS equations.
- The sequence of individually most probable latent variable values is the same as the most probable latent sequence.
  - Because the joint distributions over all variables, as well as all marginals and conditionals, will be Gaussian.
  - There is no need to consider the analogue of the Viterbi algorithm.

13.3 Linear Dynamical Systems

- Transition and emission distributions

$$\begin{aligned} p(\mathbf{z}_n|\mathbf{z}_{n-1}) &= \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}) && (13.75) \\ p(\mathbf{x}_n|\mathbf{z}_n) &= \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}). && (13.76) \end{aligned}$$

- Initial latent variable

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0). \qquad (13.77)$$

$$\begin{aligned} \mathbf{z}_n &= \mathbf{A}\mathbf{z}_{n-1} + \mathbf{w}_n && (13.78) \\ \mathbf{x}_n &= \mathbf{C}\mathbf{z}_n + \mathbf{v}_n && (13.79) \\ \mathbf{z}_1 &= \boldsymbol{\mu}_0 + \mathbf{u} && (13.80) \end{aligned}$$

$$\begin{aligned} \mathbf{w} &\sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\Gamma}) && (13.81) \\ \mathbf{v} &\sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \boldsymbol{\Sigma}) && (13.82) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{V}_0). && (13.83) \end{aligned}$$

- The parameters of the model, $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Gamma}, \mathbf{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \mathbf{V}_0\}$, can be determined using maximum likelihood through the EM algorithm.

13.3.1 Inference in LDS

- E step: using the sum-product algorithm, solve the inference problem of
  - determining $\hat{\alpha}(\mathbf{z}_n)$ and $\gamma(\mathbf{z}_n) = \hat{\alpha}(\mathbf{z}_n)\,\hat{\beta}(\mathbf{z}_n)$, and
  - making predictions of $\mathbf{z}_n$ and $\mathbf{x}_n$ conditioned on $\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}$.
- In principle can solve the inference problem from the first principle, sum-product algorithm provide a more efficient way.
- Have the same factor graphs as the HMM, the inference algorithms the same except that the summations over latent variables are replaced by integrations.

13.3.1 Inference in LDS

- The forward equations
  - $\mathbf{z}_N$ as the root, propagate message from the leaf node $h(\mathbf{z}_1)$ to the root.
  - Initial message is Gaussian. $p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0).$ (13.77)
  - All subsequence messages are Gaussian.

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\int \widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})\,\mathrm{d}\mathbf{z}_{n-1}. \tag{13.85}$$

$$\widehat{\alpha}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n). \tag{13.84}$$

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}) \tag{13.75}$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}). \tag{13.76}$$

$$c_1\widehat{\alpha}(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1). \tag{13.93}$$

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}} \widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1}). \tag{13.59}$$

$$\boldsymbol{\mu}_n = \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}) \tag{13.89}$$

$$\mathbf{V}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1} \tag{13.90}$$

$$c_n = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}). \tag{13.91}$$

$$\mathbf{P}_{n-1} = \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma}. \tag{13.88}$$

$$\mathbf{K}_n = \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}. \tag{13.92}$$

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \mathbf{K}_1(\mathbf{x}_1 - \mathbf{C}\boldsymbol{\mu}_0) \tag{13.94}$$

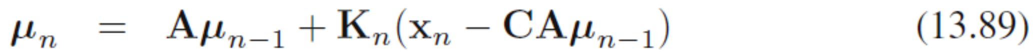
$$\mathbf{V}_1 = (\mathbf{I} - \mathbf{K}_1\mathbf{C})\mathbf{V}_0 \tag{13.95}$$

$$c_1 = \mathcal{N}(\mathbf{x}_1|\mathbf{C}\boldsymbol{\mu}_0, \mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}) \tag{13.96}$$

$$\mathbf{K}_1 = \mathbf{V}_0\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{V}_0\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}. \tag{13.97}$$

13.3.1 Inference in LDS

- The forward equations
  - The likelihood function is $p(\mathbf{x}) = \prod c_n$.
- Interpretation of (13.89)



$$\boldsymbol{\mu}_n = \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}) \qquad (13.89)$$

  - Take the predicted mean of $\mathbf{z}_n$, which is $\mathbf{A}\boldsymbol{\mu}_{n-1}$,
  - Then add a correction proportional to the error $\mathbf{x}_n - \mathbf{CAz}_{n-1}$
    - between the predicted observation $\mathbf{CAz}_{n-1}$ and the actual observation $\mathbf{x}_n$.
  - The coefficient of the correction is the Kalman gain matrix $\mathbf{K}_n$.
- We can view the Kalman filter as a process of making successive predictions and then correcting the predictions in the light of the new observations.
- Observation noise is small → $\boldsymbol{\Sigma} = \mathbf{0}$ → $\mathbf{K}_n$= $\mathbf{C}_n^{-1}$ → $\mathbf{V}_n = \mathbf{0}, \boldsymbol{\mu}_n = \mathbf{C}_n^{-1}\,\mathbf{x}_n$ → $\mathbf{z}_n$ depends only on $\mathbf{x}_n$.
- Latent variable evolves slowly → $\mathbf{z}_n$ is the average of $\mathbf{x}_n$.

13.3.1 Inference in LDS

- Backward recursion
  - Finding $\gamma(\mathbf{z}_n)$, the posterior marginal for a node $\mathbf{z}_n$, given all $\mathbf{x}_1, \cdots, \mathbf{x}_N$.
  - $\gamma(\mathbf{z}_n)$ must be Gaussian. $\gamma(\mathbf{z}_n) = \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n).$ (13.98)

$$c_{n+1}\widehat{\beta}(\mathbf{z}_n) = \int \widehat{\beta}(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n)\,\mathrm{d}\mathbf{z}_{n+1}. \tag{13.99}$$

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}) \tag{13.75}$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}). \tag{13.76}$$

$$\boldsymbol{\mu}_n = \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}) \tag{13.89}$$

$$\mathbf{V}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1} \tag{13.90}$$

$$c_n = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}). \tag{13.91}$$

$$\widehat{\alpha}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n). \tag{13.84}$$

$$c_{n+1}\widehat{\beta}(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \widehat{\beta}(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n). \tag{13.62}$$

$$\gamma(\mathbf{z}_n) = \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n). \tag{13.98}$$

$$\widehat{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_n + \mathbf{J}_n\left(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{A}\boldsymbol{\mu}_N\right) \tag{13.100}$$

$$\widehat{\mathbf{V}}_n = \mathbf{V}_n + \mathbf{J}_n\left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\right)\mathbf{J}_n^{\mathrm{T}} \tag{13.101}$$

$$\mathbf{J}_n = \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\mathbf{P}_n\right)^{-1} \tag{13.102}$$

13.3.1 Inference in LDS

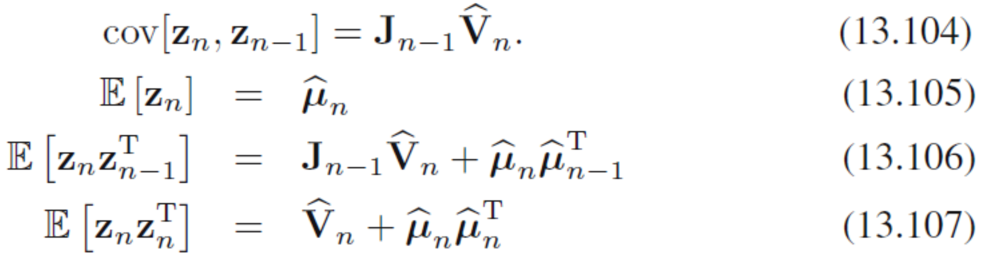
- Pairwise posterior marginals $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$

$$\begin{aligned}\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= (c_n)^{-1}\,\widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{-1})\widehat{\beta}(\mathbf{z}_n) \\ &= \frac{\mathcal{N}(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1})\mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n)}{c_n\widehat{\alpha}(\mathbf{z}_n)}.\end{aligned} \tag{13.103}$$

  - This becomes a Gaussian with
    - Mean given with $\gamma(\mathbf{z}_{n-1})$ and $\gamma(\mathbf{z}_n)$,
    - Covariance between $\mathbf{z}_n$ and $\mathbf{z}_{n-1}$ given by $\operatorname{cov}[\mathbf{z}_n, \mathbf{z}_{n-1}] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n.$ (13.104)

13.3.2 Learning in LDS

- Determination of $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Gamma}, \mathbf{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \mathbf{V}_0\}$ using maximum likelihood and the EM algorithm.
- E step
  - For $\boldsymbol{\theta}^{\text{old}}$, run the inference to determine $\hat{\alpha}(\mathbf{z}_n), \gamma(\mathbf{z}_n), \xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$, in particular,



$$\mathrm{cov}[\mathbf{z}_n, \mathbf{z}_{n-1}] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n. \quad (13.104)$$

$$\mathbb{E}\left[\mathbf{z}_n\right] = \widehat{\boldsymbol{\mu}}_n \quad (13.105)$$

$$\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n + \widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_{n-1}^{\mathrm{T}} \quad (13.106)$$

$$\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right] = \widehat{\mathbf{V}}_n + \widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^{\mathrm{T}} \quad (13.107)$$

  - These are used in the M step.

13.3.2 Learning in LDS

- E step
  - The complete-data log likelihood function from (13.6)

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n). \qquad (13.6)$$

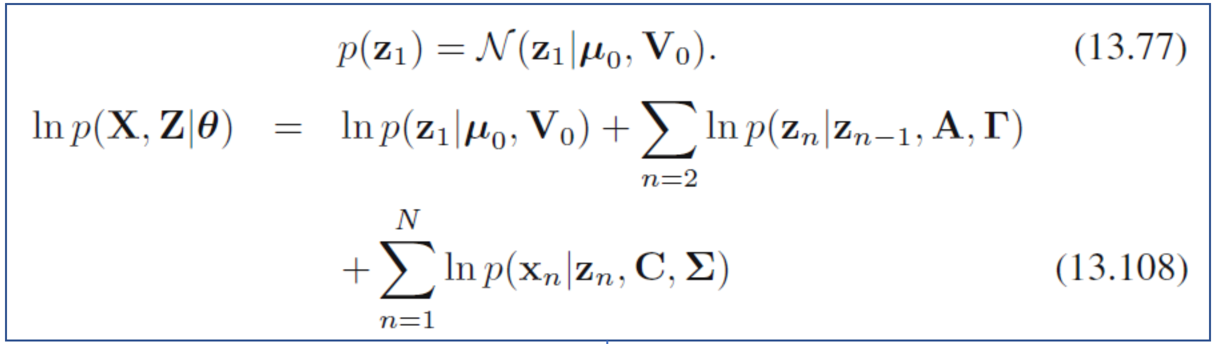
$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) &= \ln p(\mathbf{z}_1 | \boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2} \ln p(\mathbf{z}_n | \mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) \\ &\quad + \sum_{n=1}^{N} \ln p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma}) \qquad (13.108) \end{aligned}$$

  - The expectation of the complete-data log likelihood function with respect to $p(\mathbf{Z}|\mathbf{X}, \theta_{\mathrm{old}})$ defines

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}} \left[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) \right]. \qquad (13.109)$$

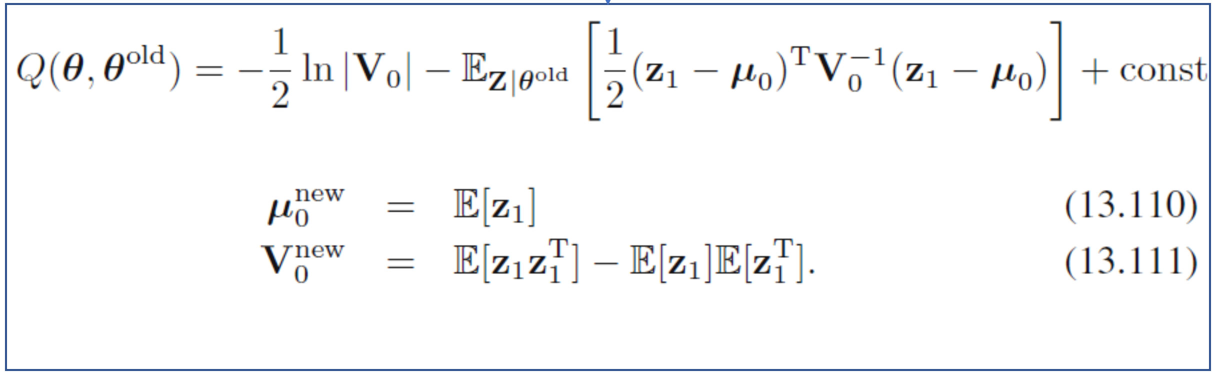
13.3.2 Learning in LDS

- M step
  - $Q(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{\text{old}})$ maximized with respect to $\boldsymbol{\mu}_0, \mathbf{V}_0$.



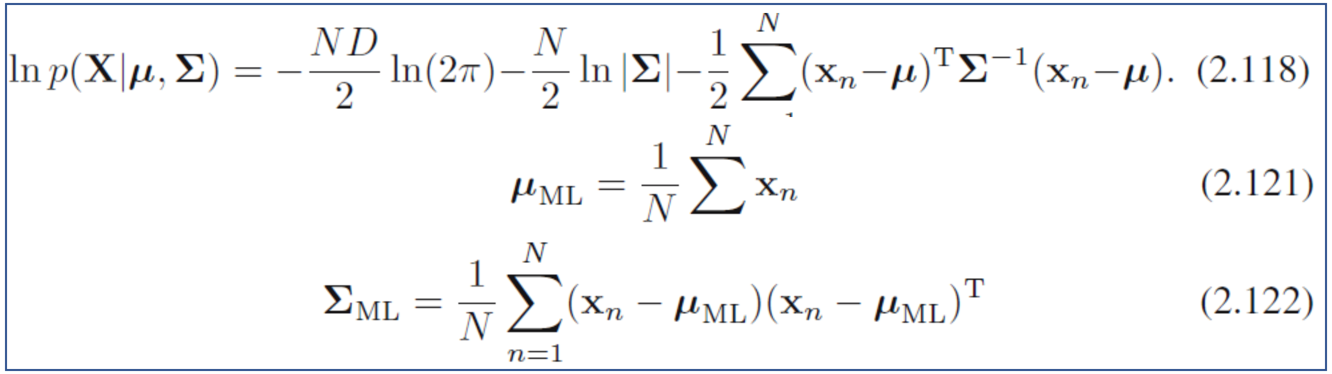
$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0). \tag{13.77}$$

$$\begin{aligned}\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) \\ &\quad + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma}) \end{aligned} \tag{13.108}$$



$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = -\frac{1}{2}\ln|\mathbf{V}_0| - \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}}\left[\frac{1}{2}(\mathbf{z}_1 - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_0)\right] + \text{const}$$

$$\begin{aligned} \boldsymbol{\mu}_0^{\text{new}} &= \mathbb{E}[\mathbf{z}_1] & (13.110) \\ \mathbf{V}_0^{\text{new}} &= \mathbb{E}[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}}] - \mathbb{E}[\mathbf{z}_1]\mathbb{E}[\mathbf{z}_1^{\mathrm{T}}]. & (13.111) \end{aligned}$$



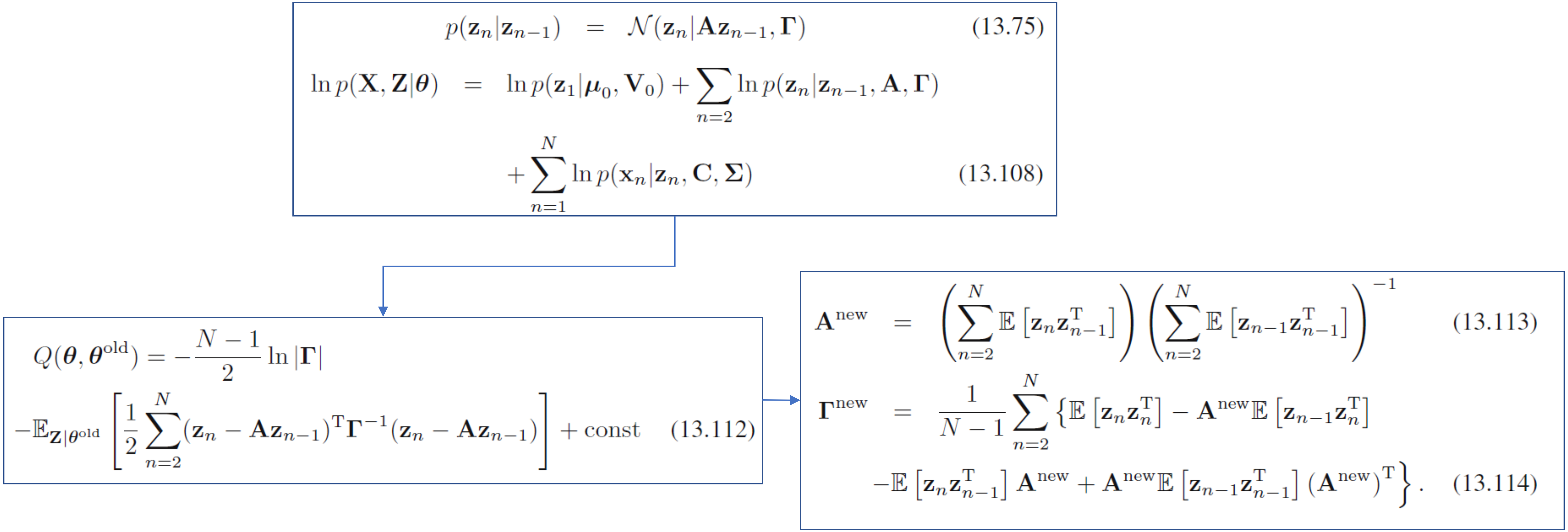
$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.118}$$

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N}\sum^{N}\mathbf{x}_n \tag{2.121}$$

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\mathrm{T}} \tag{2.122}$$

13.3.2 Learning in LDS

- M step
  - $Q(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{\text{old}})$ maximized with respect to $\mathbf{A}, \boldsymbol{\Gamma}$.



$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}) \qquad (13.75)$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma}) \qquad (13.108)$$

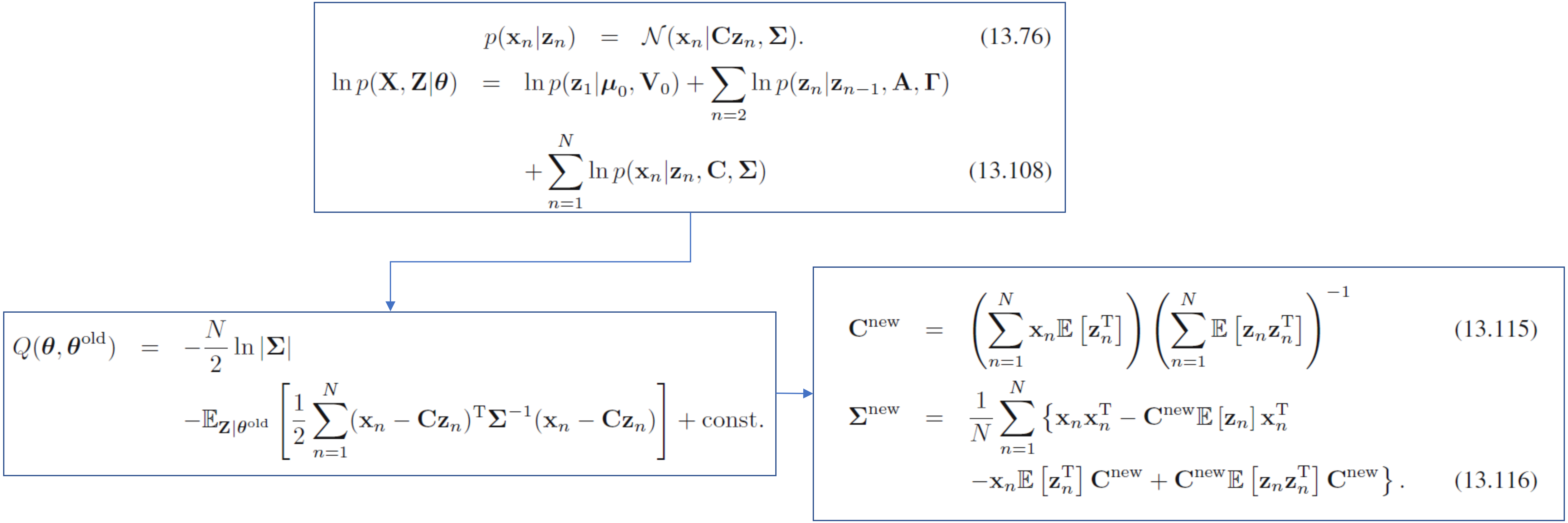
$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = -\frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right] + \text{const} \qquad (13.112)$$

$$\mathbf{A}^{\text{new}} = \left(\sum_{n=2}^{N}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right]\right)\left(\sum_{n=2}^{N}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right]\right)^{-1} \qquad (13.113)$$

$$\boldsymbol{\Gamma}^{\text{new}} = \frac{1}{N-1}\sum_{n=2}^{N}\left\{\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right] - \mathbf{A}^{\text{new}}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}}\right] - \mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right]\mathbf{A}^{\text{new}} + \mathbf{A}^{\text{new}}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right](\mathbf{A}^{\text{new}})^{\mathrm{T}}\right\}. \qquad (13.114)$$

13.3.2 Learning in LDS

- M step
  - $Q(\boldsymbol{\theta}\,,\boldsymbol{\theta}^{\text{old}})$ maximized with respect to $\mathbf{C}, \boldsymbol{\Sigma}$.



$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}). \tag{13.76}$$

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) \\ &\quad + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma}) \end{aligned} \tag{13.108}$$

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= -\frac{N}{2} \ln |\boldsymbol{\Sigma}| \\ &\quad -\mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}} \left[\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n) \right] + \text{const.} \end{aligned}$$

$$\mathbf{C}^{\text{new}} = \left(\sum_{n=1}^{N} \mathbf{x}_n \mathbb{E}\left[\mathbf{z}_n^{\mathrm{T}}\right] \right) \left(\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^{\mathrm{T}}\right] \right)^{-1} \tag{13.115}$$

$$\begin{aligned} \boldsymbol{\Sigma}^{\text{new}} &= \frac{1}{N} \sum_{n=1}^{N} \left\{ \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} - \mathbf{C}^{\text{new}} \mathbb{E}\left[\mathbf{z}_n\right] \mathbf{x}_n^{\mathrm{T}} \right. \\ &\quad \left. - \mathbf{x}_n \mathbb{E}\left[\mathbf{z}_n^{\mathrm{T}}\right] \mathbf{C}^{\text{new}} + \mathbf{C}^{\text{new}} \mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^{\mathrm{T}}\right] \mathbf{C}^{\text{new}} \right\}. \end{aligned} \tag{13.116}$$

13.3.3 Extensions of LDS

- Marginal distributions of the observed variables of LDS is simply a Gaussian, which represents a significant limitation.
  - Use a Gaussian mixture for $p(\mathbf{z}_1)$, then (13.93), (13.85) leads to a mixture of Gaussians over $\mathbf{z}_n$, so this is still tractable.

$$c_1\widehat{\alpha}(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1). \tag{13.93}$$

$$c_n\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\int \widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})\,\mathrm{d}\mathbf{z}_{n-1}. \tag{13.85}$$

  - If the emission density $p(\mathbf{x}_n|\mathbf{z}_n)$ is a mixture of Gaussians, the number of components grows exponentially with the length of the chain, therefore intractable.
  - More generally, if $p(\mathbf{x}_n|\mathbf{z}_n)$ or $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ not in the exponential family, inference becomes intractable, and we either make deterministic approximations (Ch. 10) or use sampling methods (Sec. 11.1.5, 13.3.4).
- Expand the graphical representation.

13.3.4 Particle filters

- Sampling-importance-resampling
- Two stages
  - At time step n, $p(\mathbf{z}_n|\mathbf{X}_n)$ represented as samples $\{\mathbf{z}_n^{(l)}\}$ with weights $\{w_n^{(l)}\}$

$$w_n^{(l)} = \frac{p(\mathbf{x}_n|\mathbf{z}_n^{(l)})}{\sum_{m=1}^{L} p(\mathbf{x}_n|\mathbf{z}_n^{(m)})} \qquad (13.118)$$

  - This can be viewed as a mixture representation of the form (13.119)

$$p(\mathbf{z}_{n+1}|\mathbf{X}_n) = \sum_l w_n^{(l)} p(\mathbf{z}_{n+1}|\mathbf{z}_n^{(l)}) \qquad (13.119)$$

  - Next, first draw L samples from (13.119), then for each sample use the new observation $\mathbf{x}_{n+1}$ to evaluate $w_{n+1}^{(l)} \propto p\left(\mathbf{x}_{n+1}\middle|\mathbf{z}_{n+1}^{(l)}\right)$.