Book Reading Seminar
Pattern Recognition and Machine Learning
by Christopher M. Bishop
Chapter 13: Sequential Data

Viet H.Q.H.

Advanced Institute for Materials Research (AIMR)
Tohoku University

hqhviet@tohoku.ac.jp

March 28, 2018

Review - Lagrange multiplier

Consider the problem of finding the maximum of a function $f(\boldsymbol{x})$ subject to a constraint relating $\boldsymbol{x}$

$$\boldsymbol{x} = \arg\max_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to } g(\boldsymbol{x}) = 0$$

We define the Lagrangian function given by

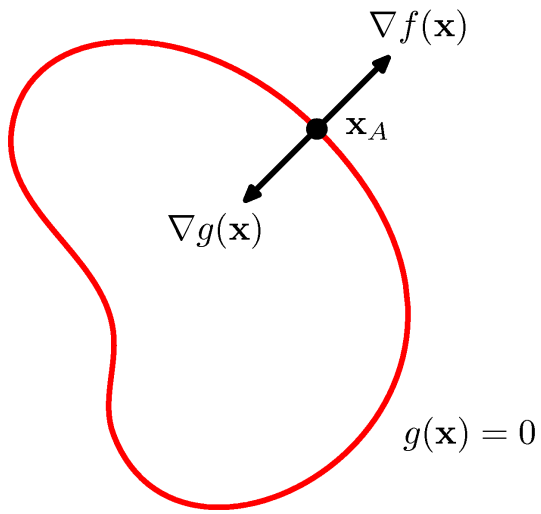
$$L(\boldsymbol{x}, \lambda) \equiv f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}). \quad \text{(E.4)}$$

where $\lambda \neq 0$ is known as a *Lagrange multiplier*. λ can have plus or minus sign.

The extreme value of f subject to the constraint $g(\boldsymbol{x})$ are critical points of the Lagrangian function L. $\boldsymbol{x}$ and λ simultaneously satisfy $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = 0$ and $\nabla_{\lambda} L(\boldsymbol{x}, \lambda) = 0$. This is equivalent to

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = 0$$

$$g(\boldsymbol{x}) = 0$$



Review - Lagrange multiplier

As an example, let's find $(q_1, ..., q_N)$ that maximizes the function

$$f(q_1, q_2, ..., q_N) = \sum_{i=1}^{N} p_i \log(q_i)$$

subject to $q_1 + q_2 + ... + q_N = 1$

The Lagrangian of the function is

$$L(q_1, q_2, \ldots, q_n, \lambda) = \sum_{i=1}^{N} p_i \log(q_i) + \lambda(\sum_{i=1}^{N} q_i - 1)$$

We need to solve $\nabla_{q_1,\ldots,q_n,\lambda} L(q_1, \ldots, q_N, \lambda) = 0$. This is equivalent to solving the linear system

$$\begin{cases} \frac{p_i}{q_i} + \lambda = 0, \quad i = 1, \ldots, N \\ q_1 + q_2 + \cdots + q_N = 1 \end{cases}$$

Review - Lagrange multiplier

$$\begin{cases} \frac{p_i}{q_i} + \lambda = 0, \quad i = 1, \ldots, N, \quad (*) \\ q_1 + q_2 + \cdots + q_N = 1 \quad (**) \end{cases}$$

From (*) we have $p_i = -\lambda q_i$. Summing both side over i we get

$$-\sum_{i=1}^{N} p_i = \lambda \left(\sum_{i=1}^{N} q_i \right).$$

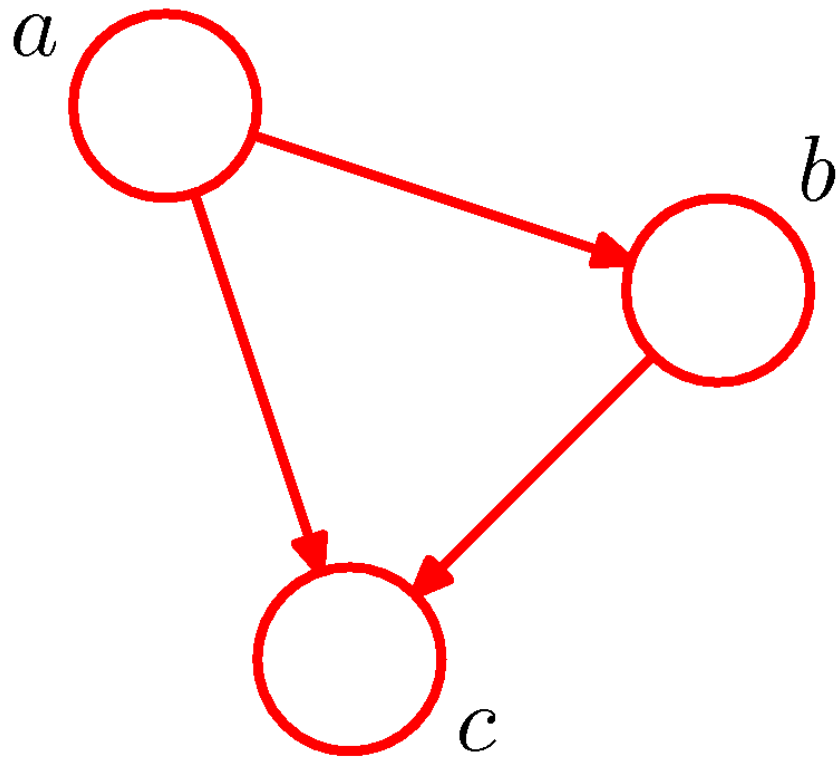
From (**) we obtain $\lambda = -\sum_{i=1}^{N} p_i$

From (*) we get

$$q_i = -\frac{p_i}{\lambda} = \frac{p_i}{\sum_{i=1}^{N} p_i} \quad i = 1, \ldots, N.$$

Review - Graphical models

Advantage of graphical models: Conditional independence can be obtained directly from the graph.



$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$

Review - Graphical models

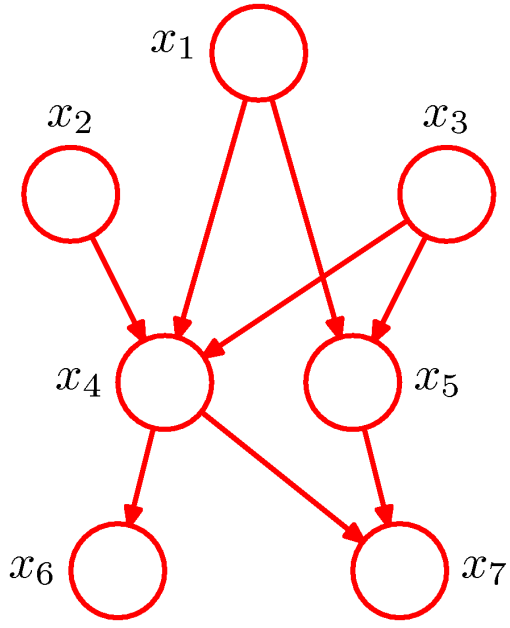
In the graph, conditional distribution is conditioned only on the parents of the corresponding node.

For a graph with K nodes the joint distribution is given by

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|pa_k) \tag{8.5}$$

where $\boldsymbol{x} = \{x_1, x_2, \ldots, x_K\}$ and pa_k denotes the set of parents of x_k.

Review - Graphical models



$$p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) =$$
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Review - Conditional Independence

Suppose that the conditional distribution of a, given b and c, is such that it does not depend on the value of b

$$p(a|b,c) = p(a|c). \tag{8.20}$$

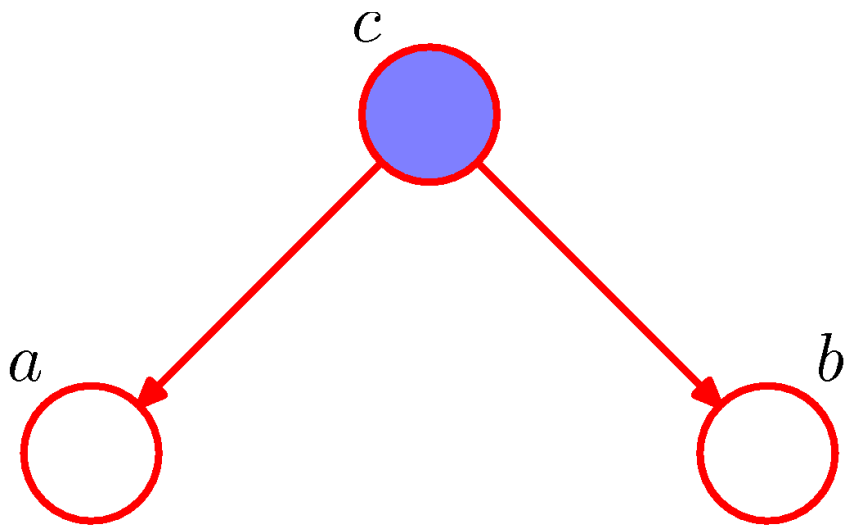
We say that a is conditionally independent of b given c.

This can be expressed in a slightly different way if we consider the joint distribution of a and b conditioned on c

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c). \end{aligned} \tag{8.21}$$

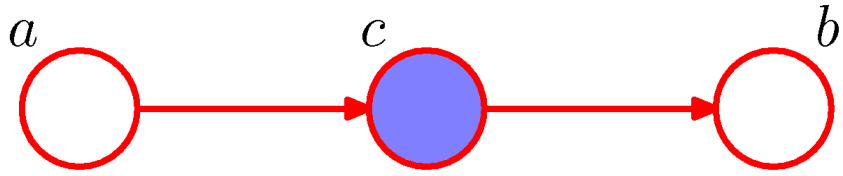
Shorthand notation for conditional independence

$$a \perp\!\!\!\perp b|c \tag{8.22}$$



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$$

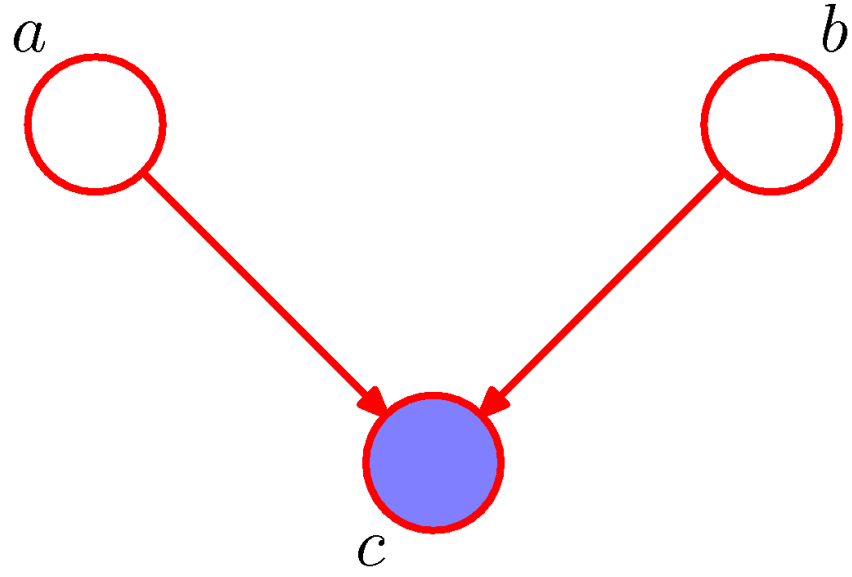
$$a \perp\!\!\!\perp b|c$$



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)}$$

$$= \underbrace{\frac{p(a)p(c|a)}{p(c)}}_{\text{Bayes theorem}} p(b|c) = p(a|c)p(b|c)$$

$$a \perp\!\!\!\perp b|c$$



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

In this case, $p(a,b|c)$ cannot be factorized into the product $p(a|c)p(b|c)$

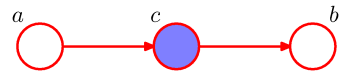
$$a \not\perp b|c$$

Review - D-separation

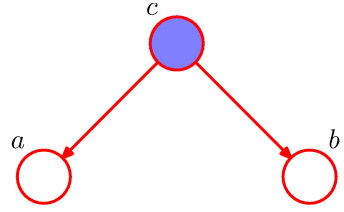
Consider a directed graph in which A, B and C are arbitrary non-intersecting sets of nodes. Node y is a *descendant* of node x if there is a path from x to y in which each step of the path follows the directions of the arrows.

To make sure of whether a statement $A \perp\!\!\!\perp B|C$ is true or not, we consider all possible paths from any node in A to any node in B. A path is said to be *blocked* if it includes a node such that
(a) the arrows on the path meet either *head-to-tail* node or *tail-to-tail* node and the node is in the set C or
(b) the arrows meet *head-to-head* node, and neither the node, nor any of its descendants, is in the set C.

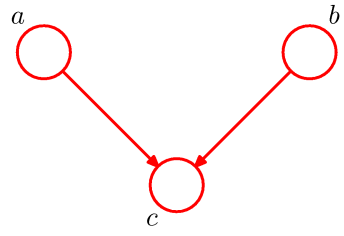
Head-to-tail node



Tail-to-tail node

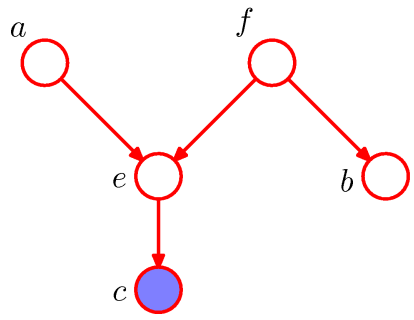


Head-to-head node



If all paths are blocked, A is said to be *d-separated* from B by C.

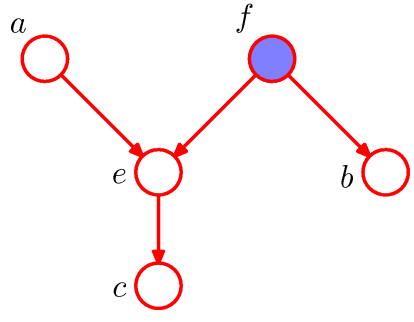
(a) the arrows on the path meet either *head-to-tail* node or *tail-to-tail* node and the node is in the set C



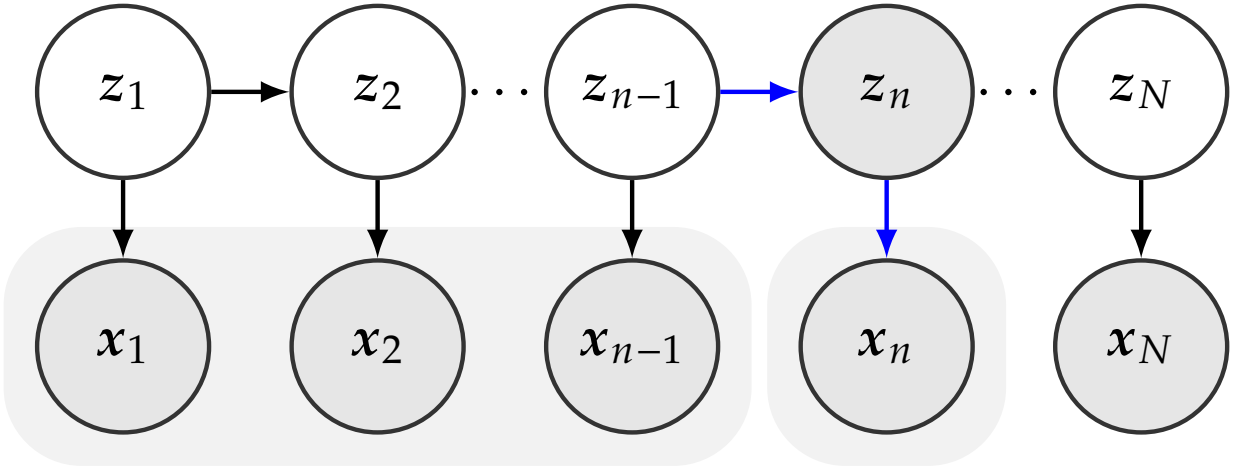
$$A = \{a\}, B = \{b\}, C = \{c\},$$
$$A \not\perp\!\!\!\perp B | C$$

or

(b) the arrows meet *head-to-head* node, and neither the node, nor any of its descendants, is in the set C.



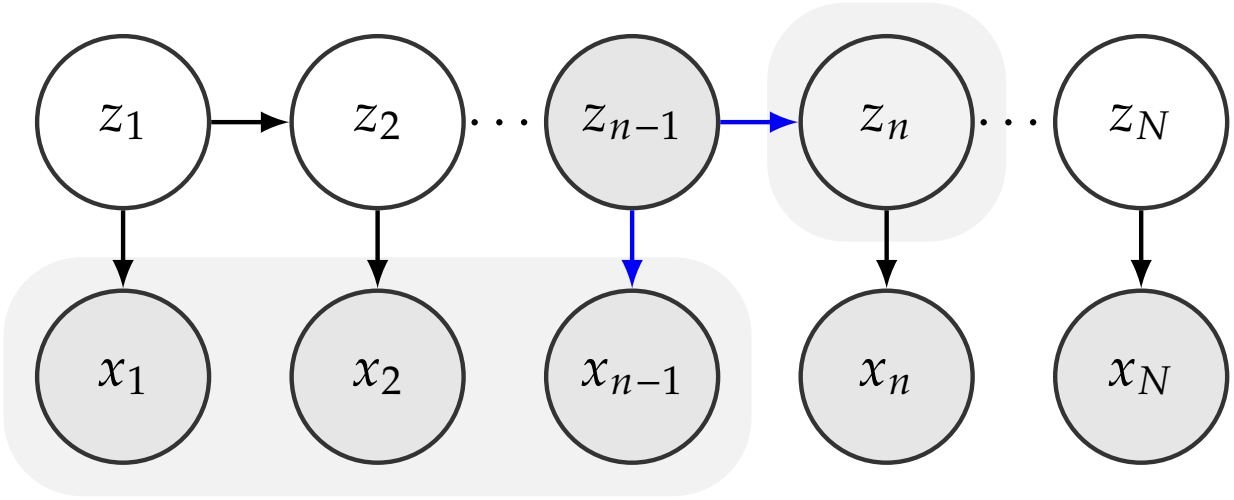
$$A = \{a\}, B = \{b\}, C = \{f\},$$
$$A \perp\!\!\!\perp B | C$$



$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} \perp\!\!\!\perp \boldsymbol{x}_n | \boldsymbol{z}_n$$

$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\}$ is conditionally independent of $\boldsymbol{x}_n$ given $\boldsymbol{z}_n$.

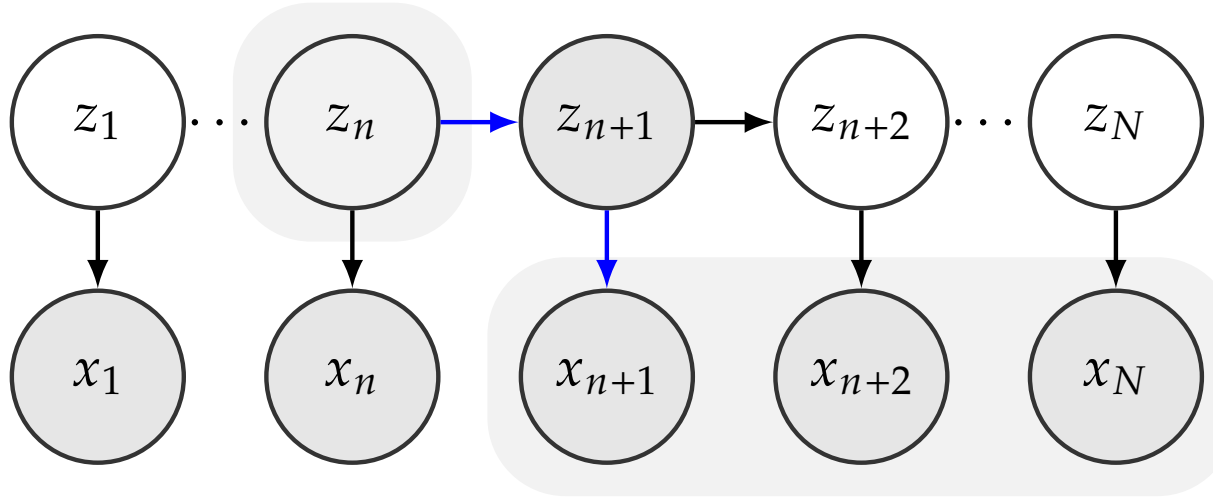
$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{x}_n, \boldsymbol{z}_n) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_n) \tag{13.25}$$



$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} \perp\!\!\!\perp \boldsymbol{z}_n | \boldsymbol{z}_{n-1}$$

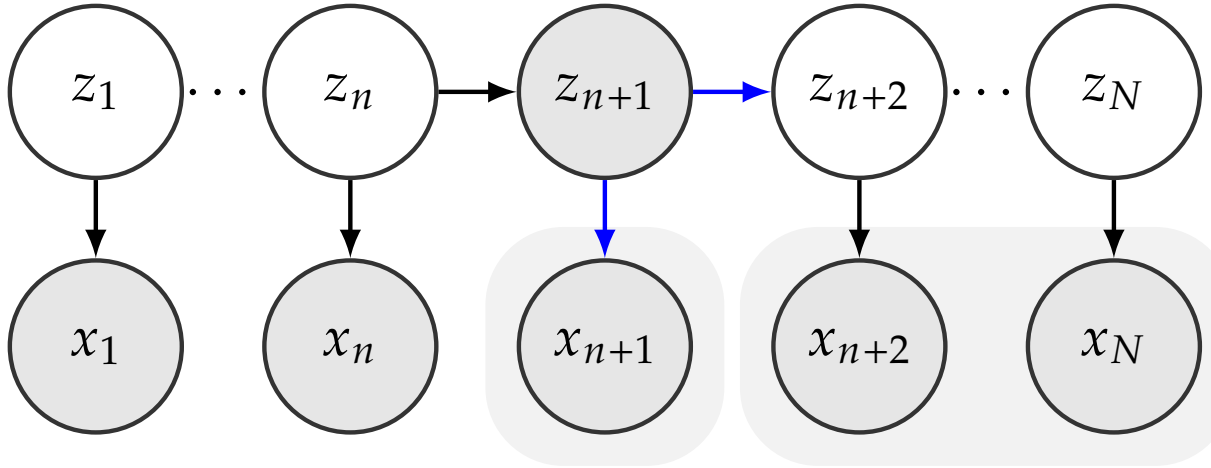
$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\}$ is conditionally independent of $\boldsymbol{z}_n$ given $\boldsymbol{z}_{n-1}$.

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}) \tag{13.26}$$



$$\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\} \perp\!\!\!\perp \boldsymbol{z}_n | \boldsymbol{z}_{n+1}$$

$$p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) = p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) \tag{13.27}$$



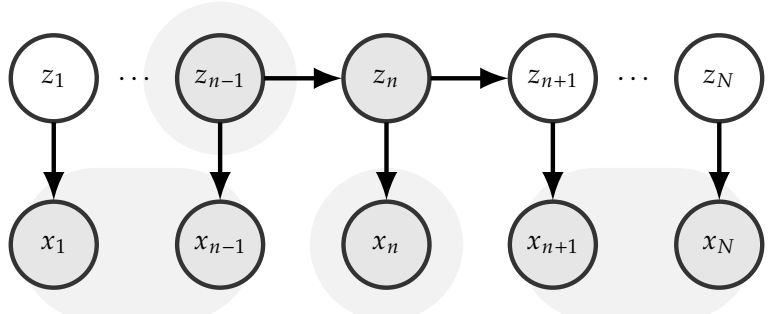
$$\{\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N\} \perp\!\!\!\perp \boldsymbol{x}_{n+1} \big| \boldsymbol{z}_{n+1}$$

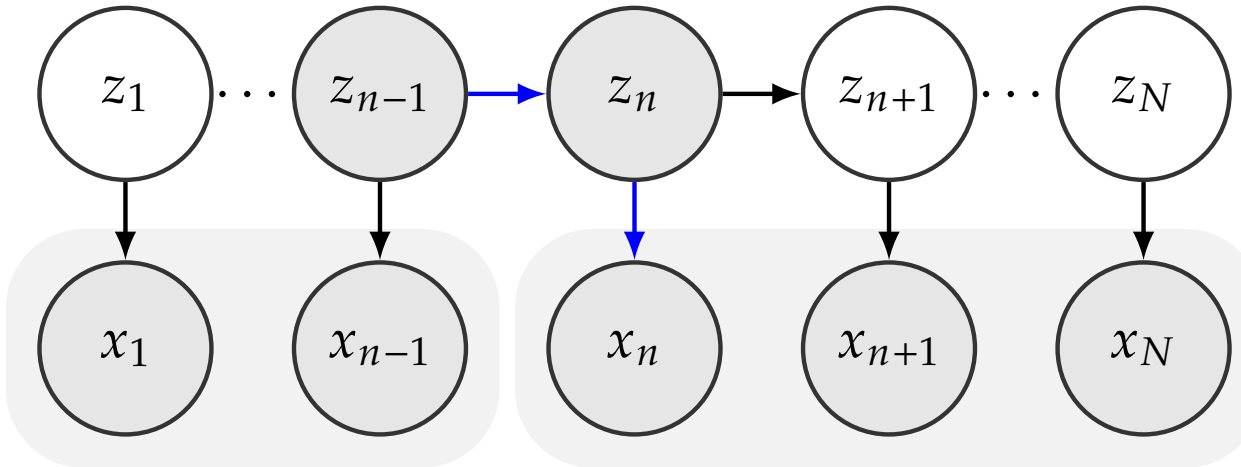
$$p(\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}, \boldsymbol{x}_{n+1}) = p(\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) \tag{13.28}$$

$$
\begin{aligned}
p(\boldsymbol{X}) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}) p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n) \qquad (13.29)
\end{aligned}
$$

We will verify (13.29) using the conditional independence properties:

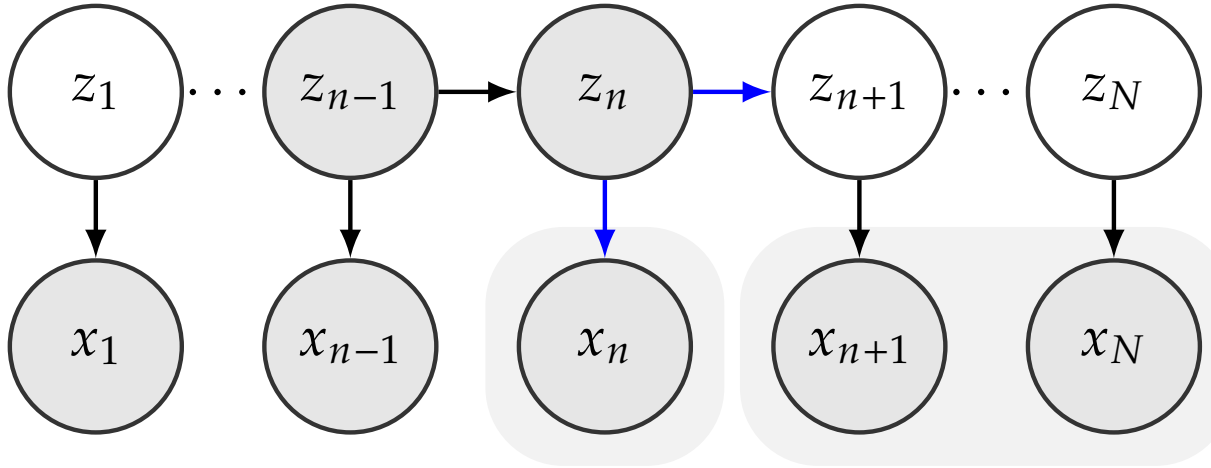
$$
\begin{aligned}
\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} &\perp\!\!\!\perp \{\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, x_N\} \big| \{\boldsymbol{z}_{n-1}, \boldsymbol{z}_n\} \\
\{\boldsymbol{x}_n\} &\perp\!\!\!\perp \{\boldsymbol{x}_{n+1}, \ldots, x_N\} \big| \{\boldsymbol{z}_{n-1}, \boldsymbol{z}_n\} \\
\{\boldsymbol{x}_n\} &\perp\!\!\!\perp \{\boldsymbol{z}_{n-1}\} \big| \{\boldsymbol{z}_n\} \\
\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\} &\perp\!\!\!\perp \{\boldsymbol{z}_{n-1}\} \big| \{\boldsymbol{z}_n\}
\end{aligned}
$$





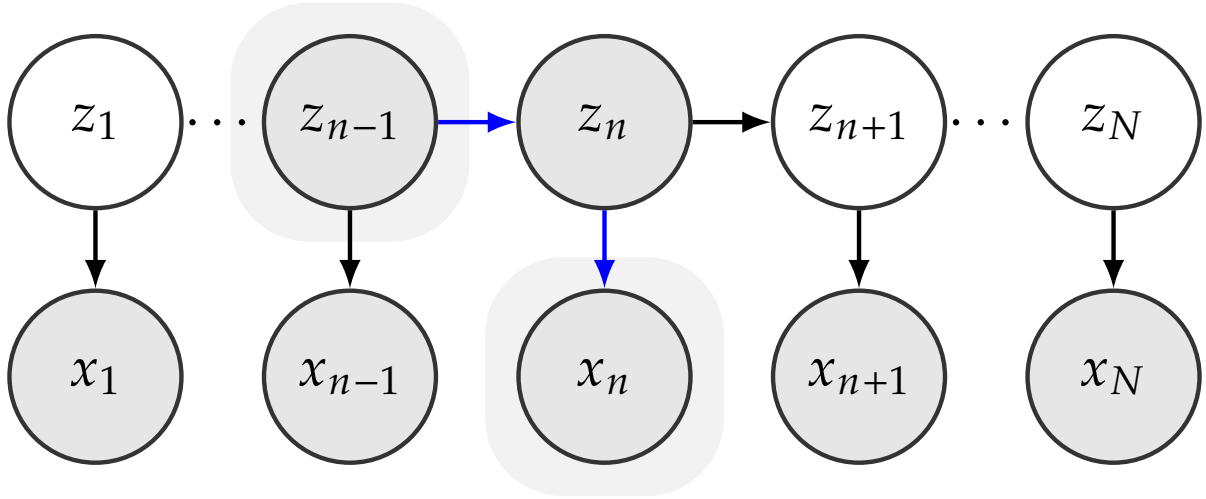
$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} \perp\!\!\!\perp \{\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, x_N\} \big| \{\boldsymbol{z}_{n-1}, \boldsymbol{z}_n\}$$

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



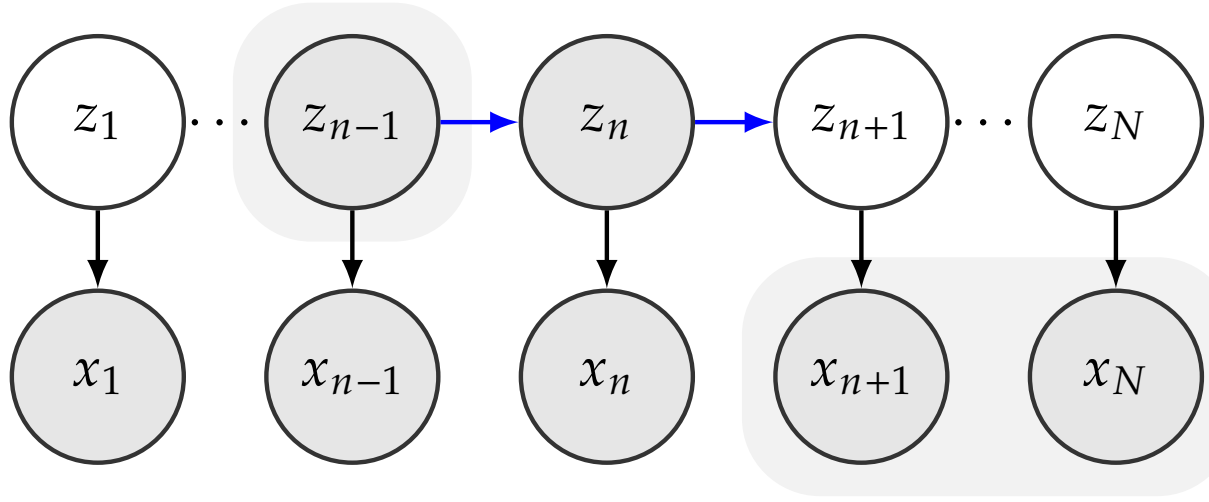
$$\{\boldsymbol{x}_n\} \perp\!\!\!\perp \{\boldsymbol{x}_{n+1}, \ldots, x_N\} \big| \{\boldsymbol{z}_{n-1}, \boldsymbol{z}_n\}$$

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



$$\{\boldsymbol{x}_n\} \perp\!\!\!\perp \{\boldsymbol{z}_{n-1}\} \big| \{\boldsymbol{z}_n\}$$

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



$$\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\} \perp\!\!\!\perp \{\boldsymbol{z}_{n-1}\} \big| \{\boldsymbol{z}_n\}$$

$$\begin{aligned}
&p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) p(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N) | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) p(\boldsymbol{x}_n | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n) \qquad (13.29)
\end{aligned}$$

Review - Complete data set and Incomplete data set

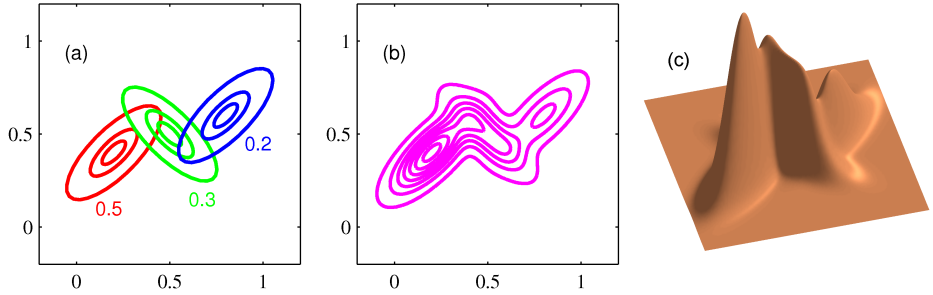
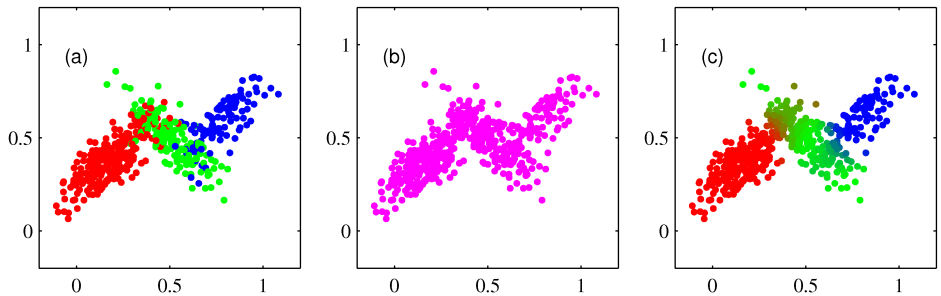
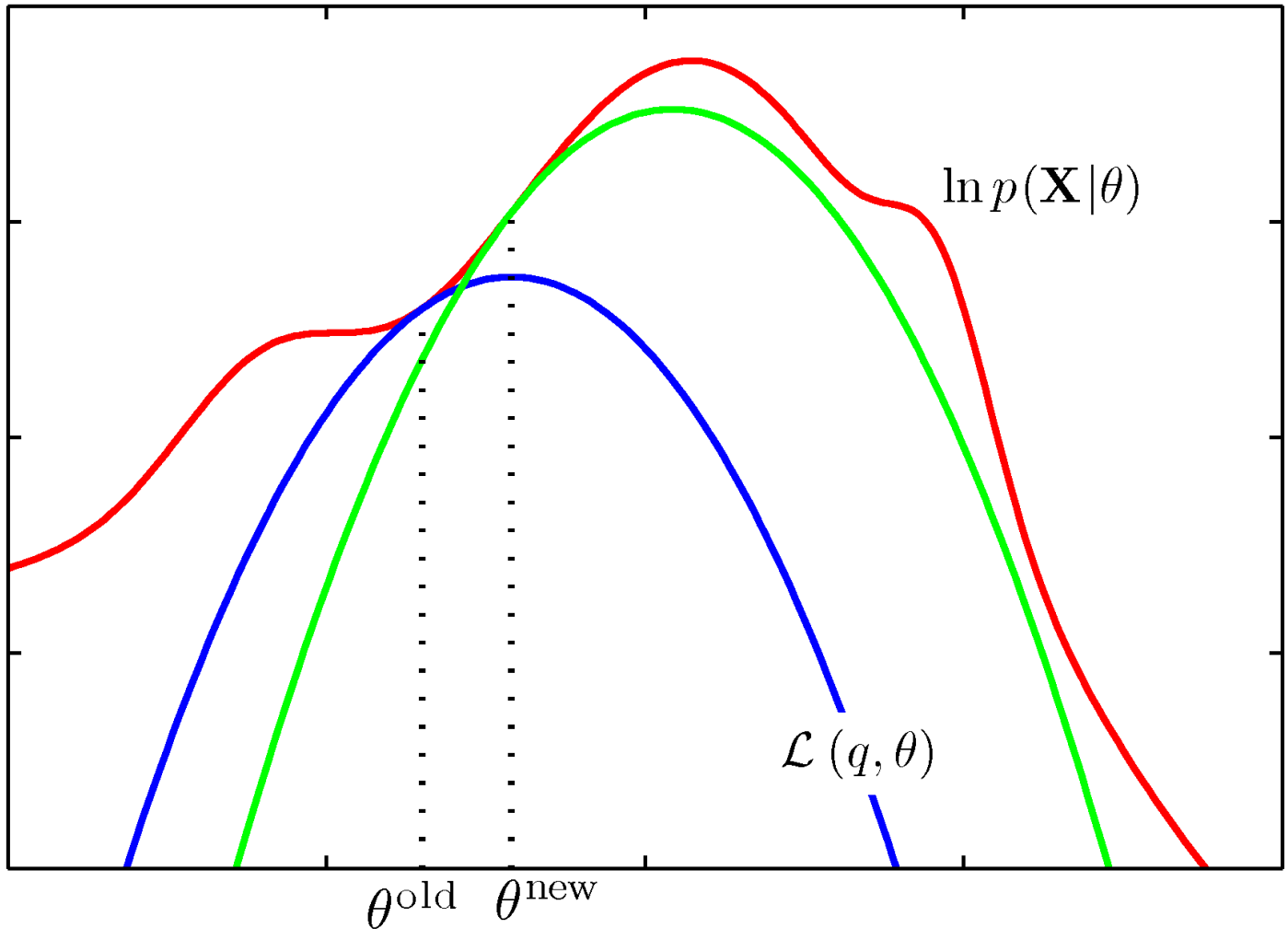


Illustration of a mixture of 3 Gaussians in a 2D space: (a) constant density contours of for each of mixture components; (b) mixture distribution contours ; (c) mixture distribution surface profile.



Example of samples drawn from the mixture distribution.
$\boldsymbol{D}_c = \{(\boldsymbol{x}_n, \boldsymbol{z}_n)\}$ is called the *complete data set* and $D_i = \{\boldsymbol{x}_n\}$ is called *incomplete data set*. $\boldsymbol{x}_n$ is the observed data point and $\boldsymbol{z}_n$ is the cluster label. The data set in (a) is complete, whereas that in (b) is incomplete.

The EM Algorithm in General



The goal of EM algorithm is to find maximum likelihood solutions for models having latent variables.

We denote
- set of all observed data by $\boldsymbol{X}$,
- set of all latent variables by $\boldsymbol{Z}$,
- set of all model parameters by $\boldsymbol{\theta}$ (in case of GMM $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$).

The log likelihood function is given by

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left(\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right) \tag{9.29}$$

We consider the expected value under the posterior distribution of the latent variable.

$$\mathbb{E}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

+ In the E step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{old})$. We use the current parameter values $\boldsymbol{\theta}^{old}$ to find the posterior distribution of latent variables given by $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$. We use this to find the expectation of ln $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$. This expectation, denote by $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ is given by

$$\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln\, p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{9.30}$$

+ In the M step: Determine the parameter $\boldsymbol{\theta}^{new}$ by maximizing the $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ function

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{9.31}$$

The General EM Algorithm

Given joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|)$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize likelihood function $p(\boldsymbol{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{old}$.
2. E step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{old})$.
3. M step: Evaluate $\boldsymbol{\theta}^{new}$ given by

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{9.32}$$

where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{9.33}$$

4. Check for convergence. If the convergence is not satisfied then let

$$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new} \tag{9.34}$$

and return to step 2.

The EM Algorithm in General

- Consider a model with observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$.
- Consider a distribution $q(\boldsymbol{Z})$ defined over latent variables.
- The joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is governed by a set of parameters $\boldsymbol{\theta}$.
- For any choice of $q(\boldsymbol{Z})$ the following decomposition holds.

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(\boldsymbol{q}||\boldsymbol{p}) \tag{9.70}$$

where we have defined

$$\mathcal{L}(q,\theta) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{\boldsymbol{q}(\boldsymbol{Z})} \right] \tag{9.71}$$

$$KL(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(x)} \right\} \tag{9.72}$$

The EM Algorithm in General

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \Big[p(\boldsymbol{X}|\boldsymbol{\theta})\Big] \underbrace{\sum_{\boldsymbol{Z}} q(\boldsymbol{Z})}_{=1} = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \Big[p(\boldsymbol{X}|\boldsymbol{\theta})\Big]$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left[p(\boldsymbol{X}|\boldsymbol{\theta}) \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right]$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left[\overbrace{\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) p(\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}}^{\text{product rule}}\right] + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left[\ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right]$$

$$= \underbrace{\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left[\frac{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]}_{\mathcal{L}(q,\boldsymbol{\theta})} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left[\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

$$= \mathcal{L}(q,\boldsymbol{\theta}) + KL(\boldsymbol{q}||\boldsymbol{p})$$

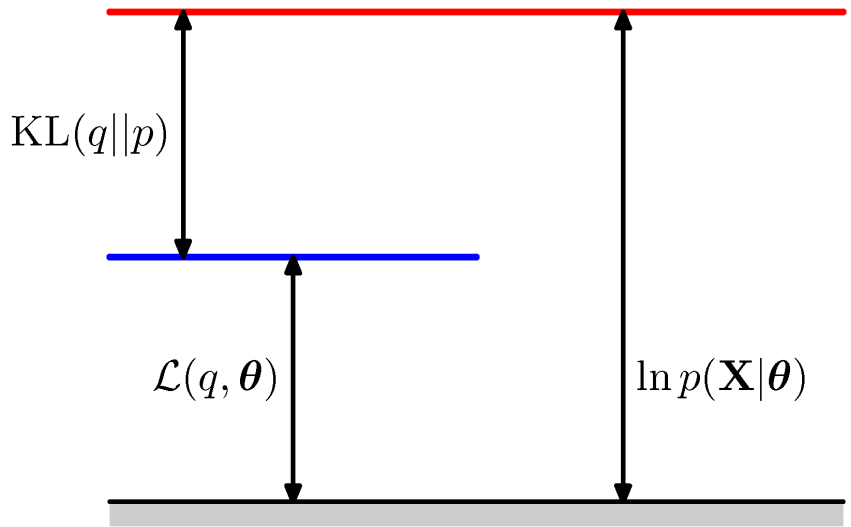
The EM Algorithm in General

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(\boldsymbol{q}||\boldsymbol{p})$$

Because the KL divergence is non-negative $KL(q||p) \geq 0$,

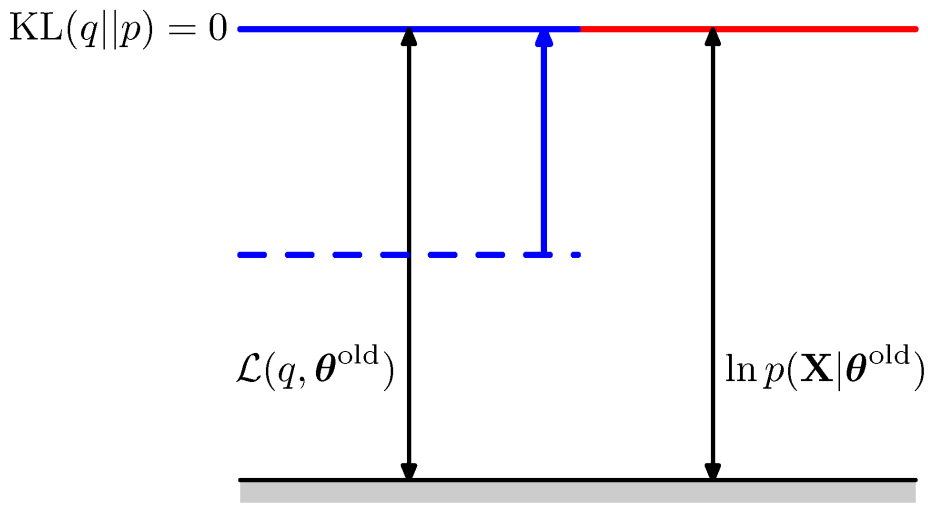
$$\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$$

$\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$



E step

In the E step we maximize $\mathcal{L}$ with respect to $q(\boldsymbol{Z})$. This equivalents to minimizing KL divergence by setting $q(\boldsymbol{Z})$ equal to $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$.



$$KL(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}))}{q(\boldsymbol{Z})} \right\} \tag{9.72}$$

M step

In the M step we maximize $\mathcal{L}$ with respect to θ. If we substitute $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ into (9.71) we see that after the E step, the lower bound takes the form

$$\mathcal{L}(q,\theta) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right] \tag{9.71}$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z})$$

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

$$= \mathcal{Q}(\boldsymbol{\theta}, \theta^{old}) + \text{const} \tag{9.74}$$

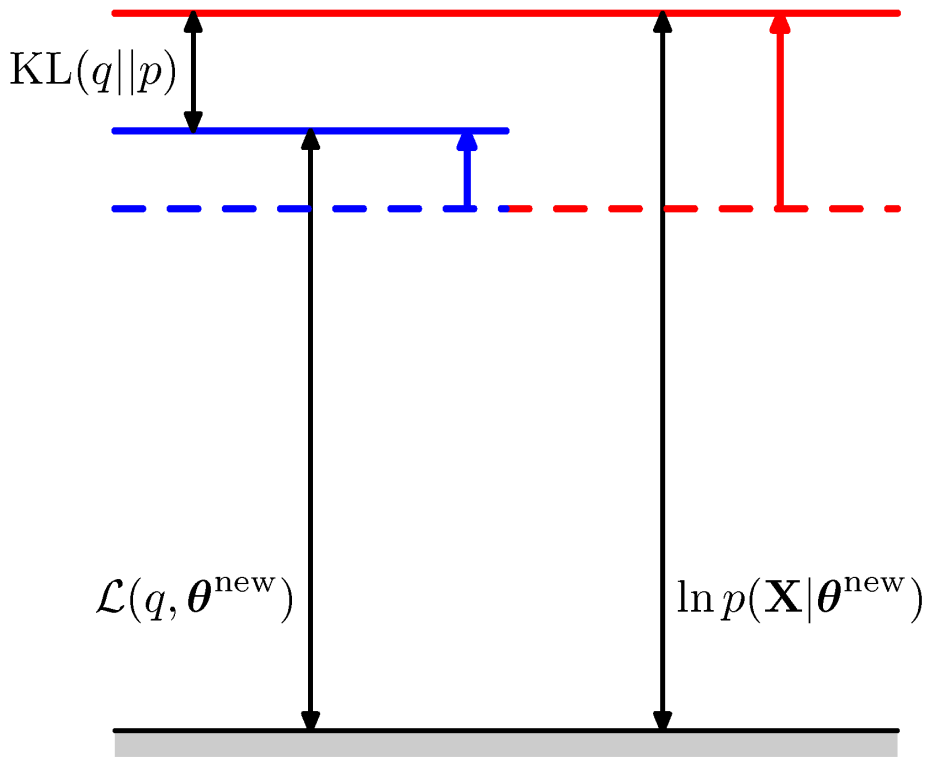
In the M step, the quantity that is being maximized is the expectation of the complete-data log likelihood

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{9.31}$$

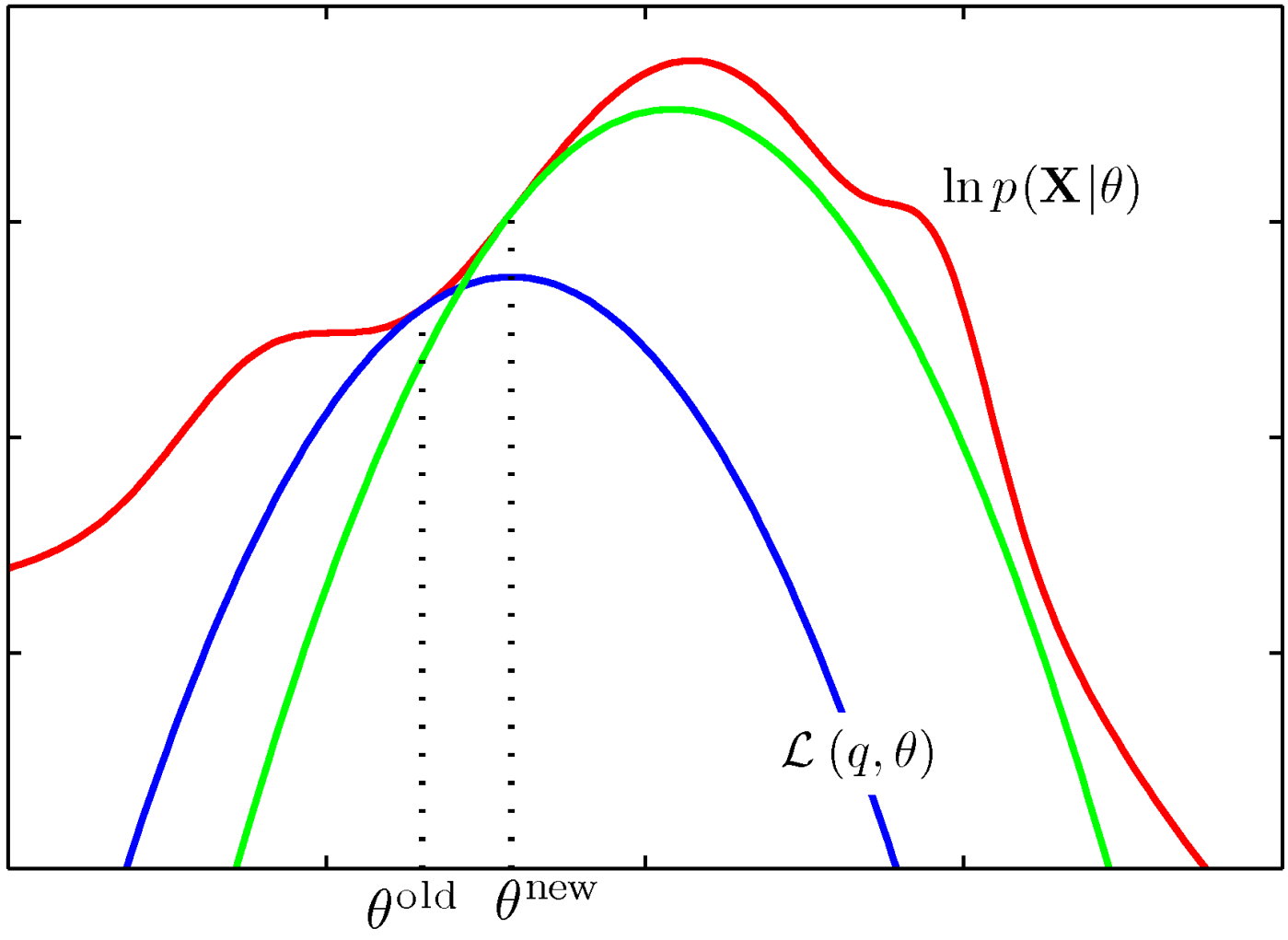
M step

In the M step, the quantity that is being maximized is the expectation of the complete-data log likelihood.

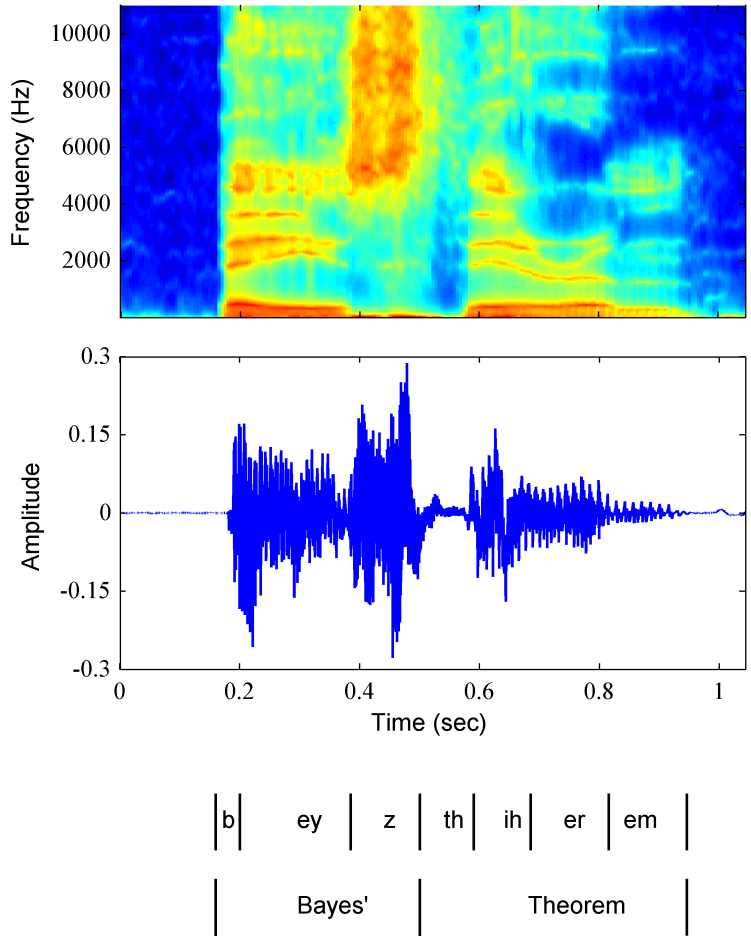
$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{9.31}$$



The EM Algorithm in General



Sequential Data

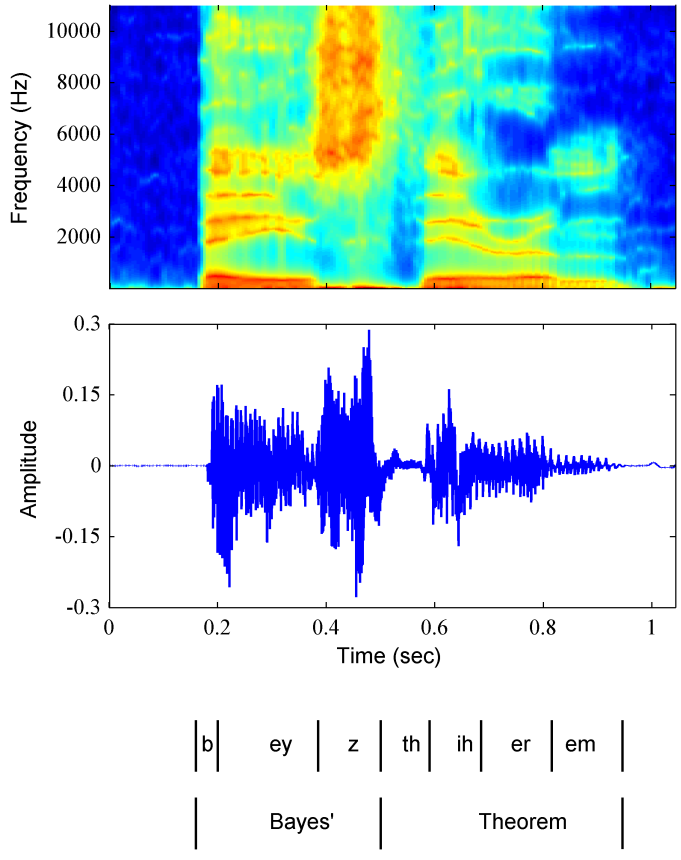


An example of sequential data involving speech data.

Sequential Data

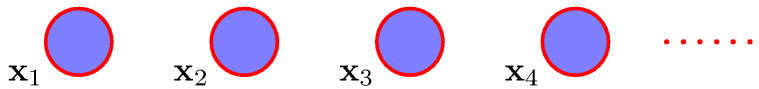
Stationary assumption: The data evolves in time, but the distribution from which it is generated remains the same.

Markov assumption: The future prediction are dependent only upon the most recent observations.



Markov Models

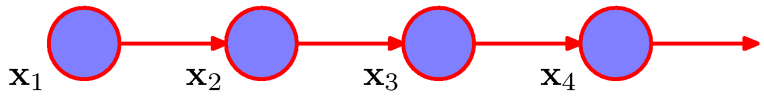
- The easiest way to treat sequential data would be simply to ignore the sequential aspect and treat the observation as i.i.d.



- This approach would fail to exploit the sequential patterns in data, such as correlations between observations that are close in the sequence.
- Consider an instance, that we observe whether on a particular day it rained or not. Weather often exhibits trends that may last for several days. Observing whether or not it rains today is significant help in predicting if it will rain tomorrow.

Markov Models

If in a chain of observations, the n^{th} observation is dependent only upon the $(n-1)^{th}$ observation, we obtain the *first-order Markov chain*



The conditional distribution for observation $\boldsymbol{x}_n$, given all of the observations up to time n, is given by

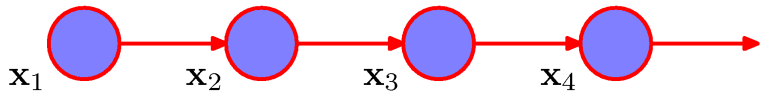
$$p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}). \tag{13.3}$$

If the stationary assumption holds true,

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_2|\boldsymbol{x}_1); \quad n \in \{2, ..., N\}$$

then the model is known as a *homogeneous* Markov chain.

Markov Models



The joint-probability of a sequence of observations is given by

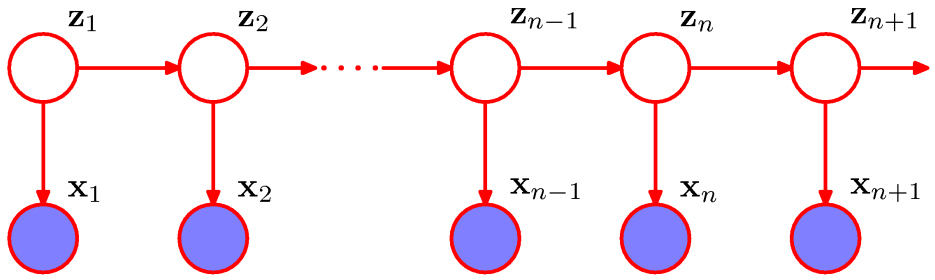
$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \prod_{n=2}^{N} p(\boldsymbol{x}_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}). \tag{13.1}$$

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

(13.1) is true for any joint distributions.
(13.2) is true with the Markov assumption.

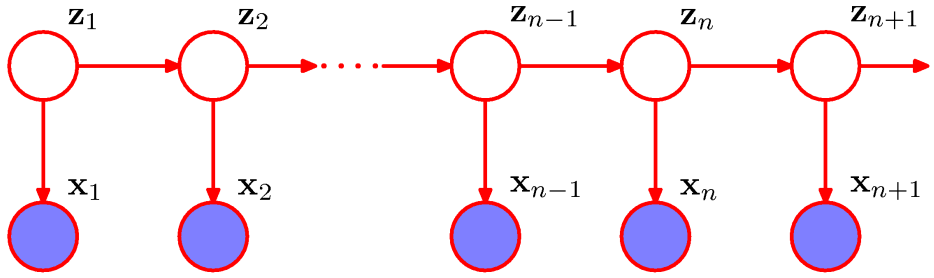
Hidden Markov Models

If each observation is conditioned on the state of latent (hidden) variable, and the latent variables are discrete and form a Markov chain, we have the *hidden Markov model* or *HMM*.



Note that the observed variables do not satisfy the Markov property at any order.

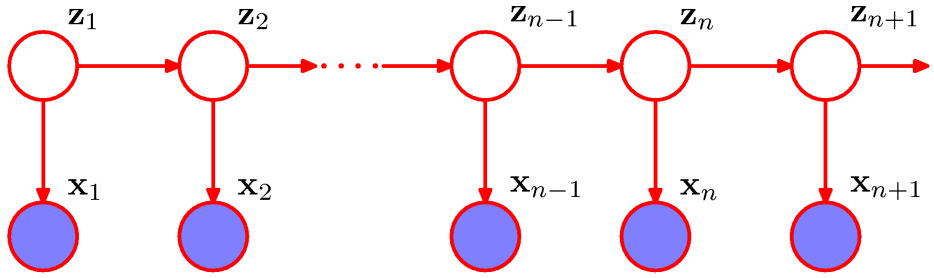
Hidden Markov Models



The joint distribution is given by

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

Parameters in HMM



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

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$$

$$= \underbrace{p(\boldsymbol{z}_1)}_{\text{initial probabilities}} \prod_{n=2}^{N} \overbrace{p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})}^{\text{transition probabilities}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{x}_n|\boldsymbol{z}_n)}_{\text{emission probabilities}} .$$

Parameters in HMM

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$$

$$= \underbrace{p(\boldsymbol{z}_1)}_{\text{initial probabilities}} \prod_{n=2}^{N} \overbrace{p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})}^{\text{transition probabilities}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{x}_n|\boldsymbol{z}_n)}_{\text{emission probabilities}} .$$

Initial probabilities: $p(\boldsymbol{z}_1)$. This parameter can be represented by a vector with K elements where K is the number of states.

Transition probabilities: $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$. This parameter can be represented by a square matrix of size $K \times K$.

Emission probabilities: $p(\boldsymbol{x}_n|\boldsymbol{z}_n)$. For the case of discrete observed variables, this parameter can be represented by a matrix of size $K \times D$ where D is the number of observations.

Transition probabilities

In this book, latent variables are represented by *K-dimensional binary vectors*:

	1	2	...	k	...	K
z_n	0	0	0	1	0	0

$z_{nk} = 1$ when the state of variable z_n is k. The vector z_n is also called *one-hot vector* with the property that $\sum_{k=1}^{K} z_{nk} = 1$.

The probability distribution of z_n depends on the state of the previous latent variable z_{n-1} through a conditional distribution $p(z_n|z_{n-1})$.

	$z_n = 1$	...	$z_n = k$	...	$z_n = K$
$z_{n-1} = 1$	A_{11}	...	A_{1k}	...	A_{1k}
...			...		
$z_{n-1} = k$	A_{j1}	...	A_{jk}	...	A_{jK}
...			...		
$z_{n-1} = K$	A_{K1}	...	A_{Kk}	...	A_{KK}

Transition probabilities

The probability distribution of z_n depends on the state of the previous latent variable z_{n-1} through a conditional distribution $p(z_n|z_{n-1})$.

	$z_n = 1$	...	$z_n = k$	...	$z_n = K$
$z_{n-1} = 1$	A_{11}	...	A_{1k}	...	A_{1k}
...			...		
$z_{n-1} = k$	A_{j1}	...	A_{jk}	...	A_{jK}
...			...		
$z_{n-1} = K$	A_{K1}	...	A_{Kk}	...	A_{KK}

The state of variable z_n at k is denoted by $z_{nk} = 1$.

	$z_{n,1} = 1$	...	$z_{n,k} = 1$	...	$z_{n,K} = 1$
$z_{n-1,1} = 1$	A_{11}	...	A_{1k}	...	A_{1k}
...			...		
$z_{n-1,j} = 1$	A_{j1}	...	A_{jk}	...	A_{jK}
...			...		
$z_{n-1,K} = 1$	A_{K1}	...	A_{Kk}	...	A_{KK}

Transition probabilities

The probability distribution of $\boldsymbol{z}_n$ depends on the state of the previous latent variable $\boldsymbol{z}_{n-1}$ through a conditional distribution $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$.

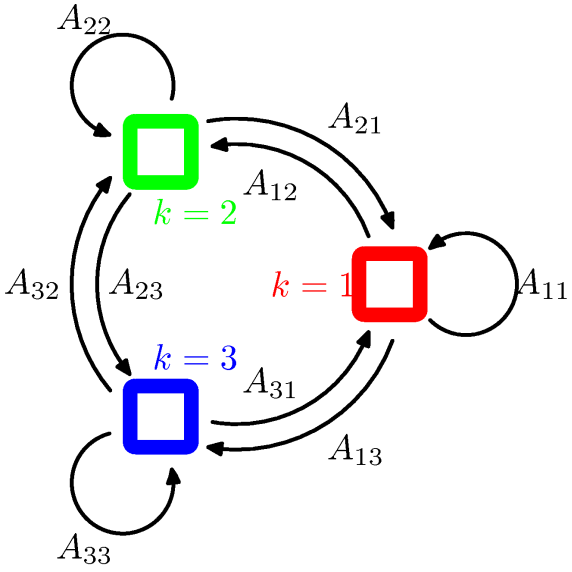
	$z_{n,1}=1$	$\ldots$	$z_{n,k}=1$	$\ldots$	$z_{n,K}=1$
$z_{n-1,1}=1$	A_{11}	$\ldots$	A_{1k}	$\ldots$	A_{1k}
$\ldots$			$\ldots$		
$z_{n-1,j}=1$	A_{j1}	$\ldots$	A_{jk}	$\ldots$	A_{jK}
$\ldots$			$\ldots$		
$z_{n-1,K}=1$	A_{K1}	$\ldots$	A_{Kk}	$\ldots$	A_{KK}

This conditional distribution is defined by a state transition matrix $\boldsymbol{A}$. The value A_{jk} is the transition probability from state j to state state k.

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

Because A_{jk} are probabilities, they satisfy $0 \leq A_{jk} \leq 1$ with $\sum_k A_{jk} = 1$.

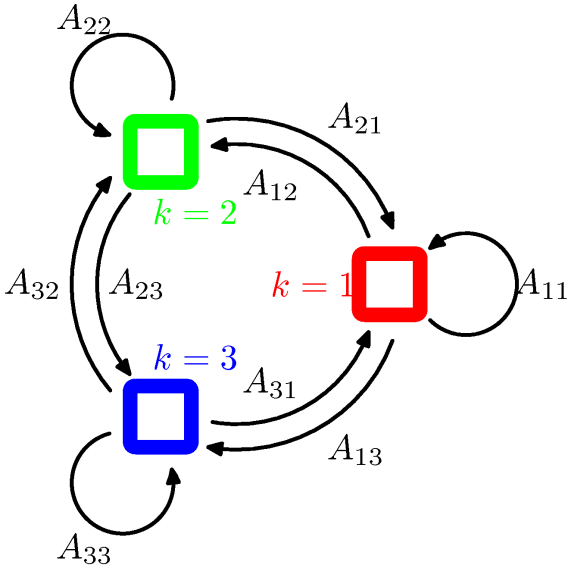
Transition probabilities



	$z_n = 1$ $(1,0,0)$	$z_n = 2$ $(0,1,0)$	$z_n = 3$ $(0,0.1)$
$z_{n-1} = 1$ $(1,0,0)$	A_{11}	A_{12}	A_{13}
$z_{n-1} = 2$ $(0,1,0)$	A_{21}	A_{22}	A_{23}
$z_{n-1} = 3$ $(0,0,1)$	A_{31}	A_{32}	A_{33}

$$A_{11} + A_{12} + A_{13} = 1$$
$$A_{21} + A_{22} + A_{23} = 1$$
$$A_{31} + A_{32} + A_{33} = 1$$

Transition probabilities



	$z_n = 1$ $(1,0,0)$	$z_n = 2$ $(0,1,0)$	$z_n = 3$ $(0,0.1)$
$z_{n-1} = 1$ $(1,0,0)$	A_{11}	A_{12}	A_{13}
$z_{n-1} = 2$ $(0,1,0)$	A_{21}	A_{22}	A_{23}
$z_{n-1} = 3$ $(0,0,1)$	A_{31}	A_{32}	A_{33}

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

$$p(\boldsymbol{z}_n = 3|\boldsymbol{z}_{n-1} = 2,\boldsymbol{A}) = A_{23}$$

Initial variable probabilities

Initial latent node $\boldsymbol{z}_1$ is a node without parent node. The initial variable probabilities $\boldsymbol{z}_1$ is represented by a vector of probabilities $\boldsymbol{\pi}$ with element $\pi_k \equiv p(z_{1k} = 1)$, so that

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

where $\sum_k \pi_k = 1$. For example, $p(\boldsymbol{z}_1 = 2|\boldsymbol{\pi}) = \pi_2$.

	$\boldsymbol{z}_1 = 1$	$\boldsymbol{z}_1 = 2$	...	$\boldsymbol{z}_1 = k$	...	$\boldsymbol{z}_1 = K$
	$(1, 0, ..., 0)$	$(0, 1, ..., 0)$		$(0, ..., 1, ..., 0)$		$(0, ..., 1)$
$p(\boldsymbol{z}_1\|\boldsymbol{\pi})$	π_1	π_2	...	π_k	...	π_K

$$\boldsymbol{z}_1 \in \{1, \ldots, K\}$$

$$\pi_1 + \pi_2 + \cdots + \pi_K = 1$$

Emission probabilities

	$x_n = 1$ $(1, 0, \ldots, 0)$	$x_n = 2$ $(0, 1, \ldots, 0)$	$x_n = i$ $(0, \ldots, 1, \ldots, 0)$	$x_n = D$ (0,0,...,1)
$z_n = 1$ $(1, 0, \ldots, 0)$	μ_{11}	μ_{12}	$\cdots$	μ_{1D}
$z_n = k$ $(0, \ldots, 1, \ldots, 0)$	μ_{k1}	μ_{k2}	μ_{ki}	μ_{kD}
$z_n = K$ $(1, 0, \ldots, 0)$	μ_{K1}	μ_{K2}	$\cdots$	μ_{KD}

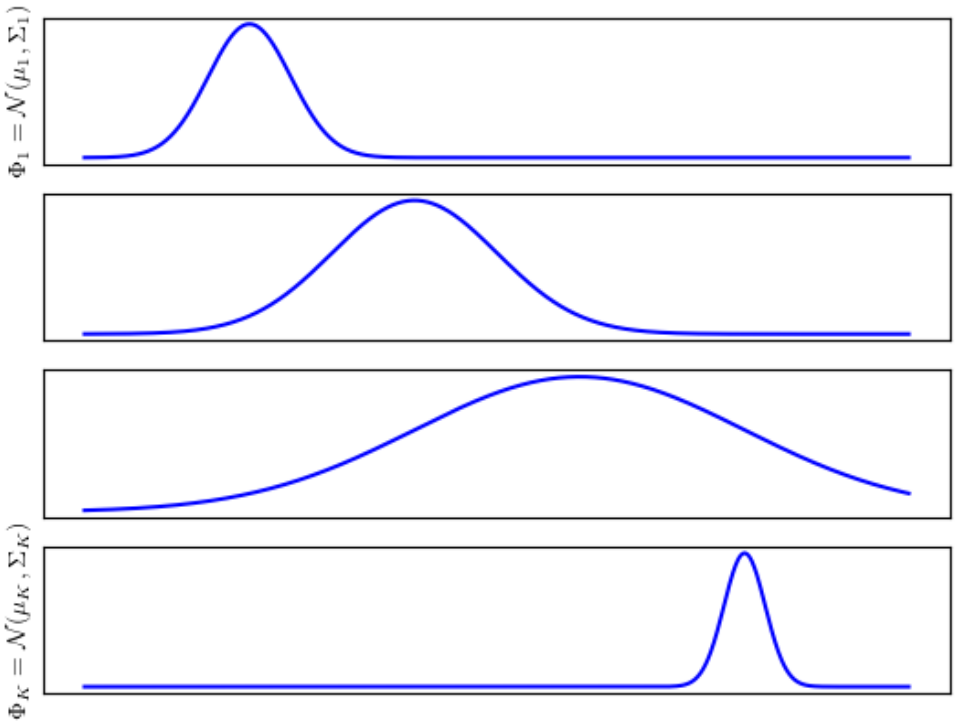
If $z_n \in \{1, 2, \ldots, K\}$, $x_n \in \{1, 2, \ldots, D\}$, the emission probabilities can be represented in the form

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{k=1}^{K} \prod_{i=1}^{D} \mu_{ki}^{z_{nk}, x_{ni}} \qquad (13.22^*)$$

$$p(\boldsymbol{x}_n = 2|\boldsymbol{z}_n = 1) = \mu_{12}$$

Emission probabilities

	$\boldsymbol{x}_n$
$z_n = 1$ $(1, \ldots, 0)$	ϕ_1
$z_n = 2$ $(0, \ldots, 0)$	ϕ_2
$z_n = k$ $(0, \ldots, 1, \ldots, 0)$	ϕ_k
$z_n = K$ $(0, \ldots, 1)$	ϕ_K



If $z_n \in \{1, 2, \ldots, K\}$, the elements of $\boldsymbol{x}$ are continuous variables, the emission probabilities can be represented in the form

$$p(\boldsymbol{x}_n | \boldsymbol{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^{K} p(x_n | \phi_k)^{z_{n,k}} \tag{13.9}$$

Parameters in HMM, $\theta = \{\pi, A, \phi\}$

Initial state probabilities $p(z_1)$

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 & \dots & \pi_K \end{bmatrix}^T$$

Transition probabilities $p(z_n|z_{n-1})$

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1K} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & A_{K3} & \dots & A_{KK} \end{bmatrix}$$

Emission probabilities $p(\boldsymbol{x}_n|\boldsymbol{z}_n)$ for the case of discrete observed variables

$$\boldsymbol{\Phi} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \dots & \mu_{1D} \\ \mu_{21} & \mu_{22} & \mu_{23} & \dots & \mu_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{K1} & \mu_{K2} & \mu_{K3} & \dots & \mu_{KD} \end{bmatrix}$$

Joint probability distribution

The joint probbility distribution over both latent and observed variables is given by

$$p(\boldsymbol{x}_1, ..., \boldsymbol{x}_N, \boldsymbol{z}_1, ..., \boldsymbol{z}_N) = p(\boldsymbol{z}_1) \prod_{n=2}^{N} p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n). \quad (13.6)$$

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

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

Maximum likelihood for the HMM

If we have observed a data set $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ we can determine the parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\phi}\}$ of HMM using maximum likelihood.

The likelihood function is obtained from joint distribution by marginalizing over latent variables

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

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

Direct maximization of the likelihood function lead to complex expressions with no closed-form solutions.

EM for the HMM

Initial selection for the model parameters: $\boldsymbol{\theta}^{old}$.
In the E step, we take these parameter values and find posterior distribution of latent variables $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$

We use this posterior distribution to evaluate expectation of the logarithm of the likelihood function $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ which is defined by

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

EM for the HMM

Defining $\gamma(\boldsymbol{z}_n)$ and $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$ as

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

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

For each value of n, we can store $\gamma(\boldsymbol{z}_n)$ by using a vector with K non-negative numbers that sum to unity

$$\gamma(\boldsymbol{z}_n) \equiv \begin{bmatrix} \gamma_1 & \dots & \gamma_K \end{bmatrix}^T$$

$$\sum_{k=1}^{K} \gamma_k = 1.$$

Similarly we can store $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$ by using a matrix with $K \times K$ non-negative numbers that sum to unity

$$\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \equiv \begin{bmatrix} \xi_{11} & \dots & \xi_{1K} \\ \vdots & \ddots & \vdots \\ \xi_{K1} & \dots & \xi_{KK} \end{bmatrix}$$

$$\sum_{j=1}^{K} \sum_{k=1}^{K} \xi_{jk} = 1.$$

EM for the HMM

We also use $\gamma(z_{nk})$ to denote the conditional probability of $z_{nk} = 1$, with a similar use of notation for $\xi(z_{n-1,j}, z_{nk})$. Because the expectation of a binary random variable is just the probability that it takes the value 1, we have

$$\gamma(z_{nk}) = \sum_{k=1}^{K} \gamma(z_{nk}) z_{nk} = \mathbb{E}[z_{nk}] = \sum_{\boldsymbol{z}_n} \gamma(\boldsymbol{z}_n) z_{nk} \tag{13.15}$$

$$\begin{aligned} \xi(z_{n-1,j}, z_{nk}) &= \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) z_{n-1,j} z_{nk} = \mathbb{E}[z_{n-1,j} z_{nk}] \\ &= \sum_{\boldsymbol{z}_{n-1} \boldsymbol{z}_n} \xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n) z_{n-1,j} z_{nk} \end{aligned} \tag{13.16}$$

EM for the HMM

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

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

Substituting $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ given by (13.10) into (13.12) and using the definitions of γ and ξ, we obtain

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln \boldsymbol{A}_{jk}$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\boldsymbol{x}_n|\phi_k) \tag{13.17}$$

EM for HMM

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

Algorithm 1 EM for HMM

repeat
  E step:
    Compute $\gamma(\boldsymbol{z}_n)$.
    Compute $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$.
  M step:
    Compute π_k.
    Compute A_{jk}.
    Compute μ_k.
    Compute Σ_k.
until convergence

EM for the HMM

- The goal of the E step is to evaluate the quantities $\gamma(\boldsymbol{z}_n)$ and $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$ efficiently.
- In the M step, we maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to the parameters $\theta = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\phi}\}$ in which we treat $\gamma(\boldsymbol{z}_n)$ and $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$ as constant. Using Lagrange multipliers, we obtain

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

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

EM for the HMM

For Gaussian emission densities $p(\boldsymbol{x}|\phi_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

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

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

These formulas can be verified by using (2.121) and (2.122) given in the part 2.3.4 (Maximum likelihood for the Gaussian) of Chapter 2.

Evaluating the quantities $\gamma(z_{nk})$

Using Bayes theorem, we have

$$\gamma(z_n) = p(z_n|\mathbf{X}) = \frac{p(\mathbf{X}|z_n)p(z_n)}{p(\mathbf{X})} \tag{13.32}$$

Using conditional independence property (13.24), we obtain

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

Here we have defined

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

$$\beta(z_n) \equiv p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|z_n) \tag{13.35}$$

Evaluating the quantities $\gamma(z_{nk})$

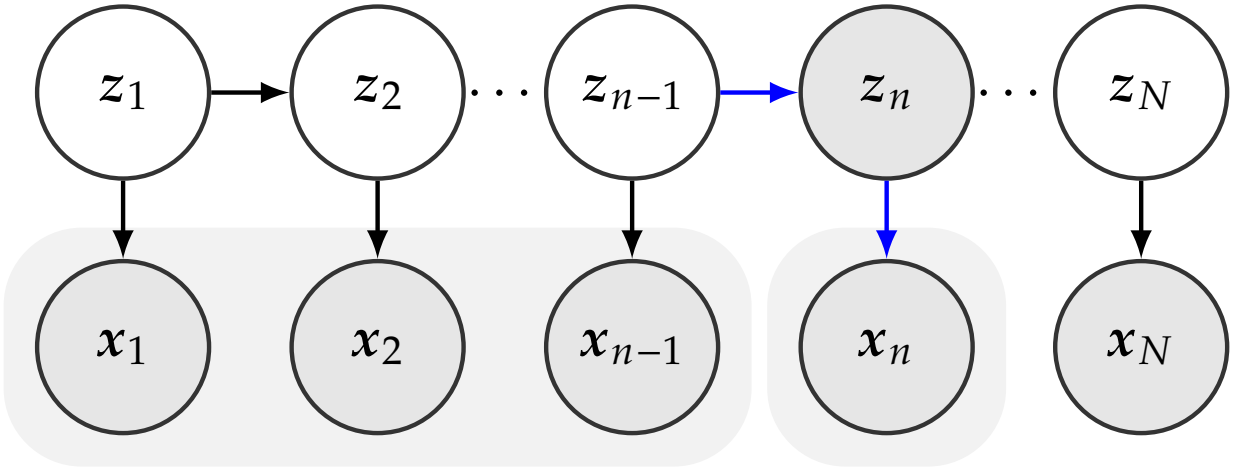
$$\alpha(\boldsymbol{z}_n) \equiv p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n) \tag{13.34}$$

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

- $\alpha(z_n)$ is the joint probability of observing $x_1, ..., x_n$ up to time n and the value of $\boldsymbol{z}_n$.

- $\beta(z_n)$ is the conditional probability of future observation $x_{n+1}, ..., x_N$ from time $n + 1$ up to N given the value of $\boldsymbol{z}_N$.

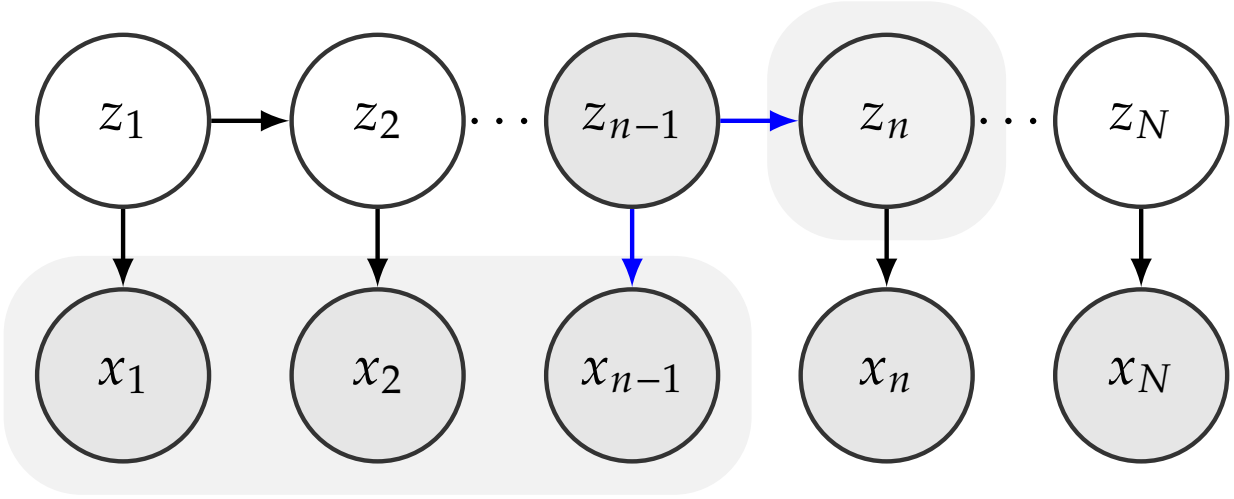
Evaluating the quantities $\alpha(z_{nk})$



$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} \perp\!\!\!\perp \{\boldsymbol{x}_n\} \big| \{\boldsymbol{z}_n\}$$

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n | \boldsymbol{z}_n) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_n) p(\boldsymbol{x}_n | \boldsymbol{z}_n) \tag{13.25}$$

Evaluating the quantities $\alpha(z_{nk})$



$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\} \perp\!\!\!\perp \{\boldsymbol{z}_n\} \,\big|\, \{\boldsymbol{z}_{n-1}\}$$

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{z}_n | \boldsymbol{z}_{n-1}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_n) p(\boldsymbol{z}_n | \boldsymbol{z}_{n-1}) \qquad (13.26)$$

Evaluating the quantities $\alpha(z_{nk})$

$$
\begin{aligned}
\alpha(z_n) &= p(x_1, \ldots, x_n, z_n) \\
&= p(x_1, \ldots, x_n | z_n) p(z_n) \\
&= p(x_n | z_n) p(x_1, \ldots, x_{n-1} | z_n) p(z_n) \\
&= p(x_n | z_n) p(x_1, \ldots, x_{n-1}, z_n) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_{n-1}, z_n) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_n | z_{n-1}) p(z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1} | z_{n-1}) p(z_n | z_{n-1}) p(z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_{n-1}) p(z_n | z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} \alpha(z_{n-1}) p(z_n | z_{n-1})
\end{aligned}
$$

Evaluating the quantities $\alpha(z_{nk})$

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

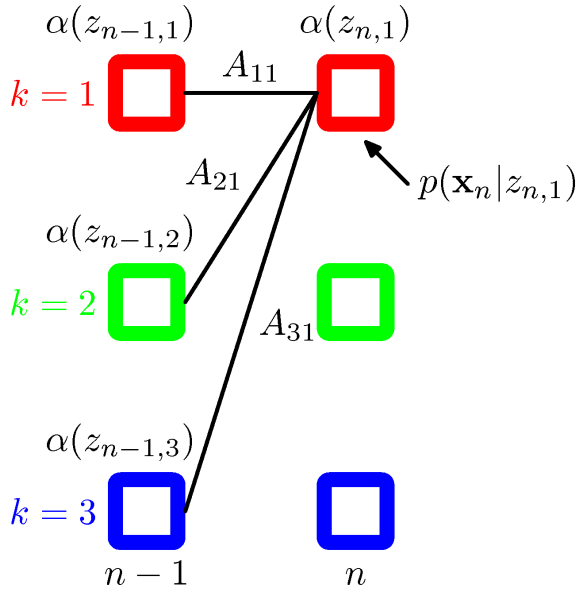
$\alpha(\boldsymbol{z}_n)$ can be evaluated by the recursion expression:

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

with the initial condition

$$\alpha(\boldsymbol{z}_1) = p(\boldsymbol{x}_1, \boldsymbol{z}_1) = p(\boldsymbol{z}_1) p(\boldsymbol{x}_1|\boldsymbol{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{x}_1|\phi_k)\}^{z_1 k} \tag{13.37}$$

Evaluating the quantities $\alpha(z_n)$



$$\begin{aligned}\alpha(\boldsymbol{z}_n) &\equiv p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n) \\ &= p(\boldsymbol{x}_n|\boldsymbol{z}_n) \sum_{\boldsymbol{z}_{n-1}} \alpha(\boldsymbol{z}_{n-1}) p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})\end{aligned} \tag{13.36}$$

Denote $\alpha(\boldsymbol{z}_n) = \begin{bmatrix}\alpha(z_{n1}) & \ldots & \alpha(z_{nK})\end{bmatrix}^T$. Let $\boldsymbol{\mu}_{*,n}$ denote the n^{th} column of the emission matrix $\boldsymbol{\Phi}$: $\boldsymbol{\mu}_{*,n} = \begin{bmatrix}\mu_{1n} & \ldots & \mu_{Kn}\end{bmatrix}^T$. Let the operator $\circ$ denote the element-wise multiplication of two vectors.

$$\begin{bmatrix}\alpha(z_{n,1}) \\ \vdots \\ \alpha(z_{n,K})\end{bmatrix} = \left(\begin{bmatrix}A_{11} & A_{21} & \ldots & A_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1K} & A_{2K} & \ldots & A_{KK}\end{bmatrix}\begin{bmatrix}\alpha(z_{n-1,1}) \\ \vdots \\ \alpha(z_{n-1,K})\end{bmatrix}\right) \circ \begin{bmatrix}\mu_{1n} \\ \vdots \\ \mu_{Kn}\end{bmatrix}$$

$$\alpha(\boldsymbol{z}_n) = \left(\boldsymbol{A}^T \alpha(\boldsymbol{z}_{n-1})\right) \circ \boldsymbol{\mu}_{*,n}$$

Evaluating the quantities $\alpha(\boldsymbol{z}_1)$

$$\alpha(\boldsymbol{z}_n) = \begin{bmatrix} \alpha(z_{n1}) & \ldots & \alpha(z_{nK}) \end{bmatrix}^T$$

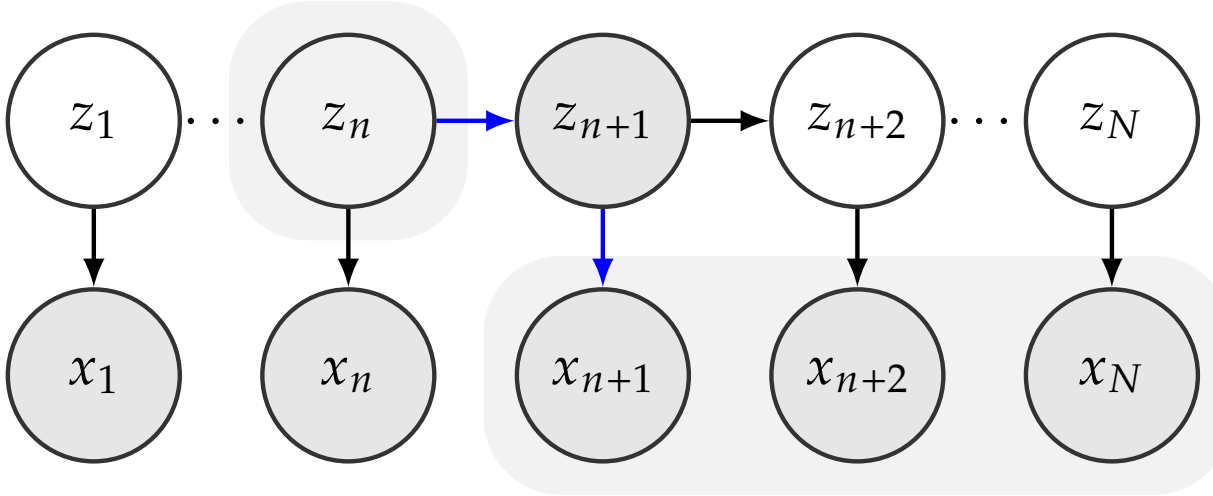
$$\boldsymbol{\mu}_{*,n} = \begin{bmatrix} \mu_{1n} & \ldots & \mu_{Kn} \end{bmatrix}^T$$

$$\alpha(\boldsymbol{z}_1) = p(\boldsymbol{x}_1, \boldsymbol{z}_1) = p(\boldsymbol{z}_1)p(\boldsymbol{x}_1|\boldsymbol{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{x}_1|\phi_k)\}^{z_1 k} \tag{13.37}$$

$$\begin{bmatrix} \alpha(z_{11}) \\ \vdots \\ \alpha(z_{1K}) \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_K \end{bmatrix} \circ \begin{bmatrix} \mu_{11} \\ \vdots \\ \mu_{K1} \end{bmatrix}$$

$$\alpha(\boldsymbol{z}_1) = \boldsymbol{\pi} \circ \boldsymbol{\mu}_{*,1}$$

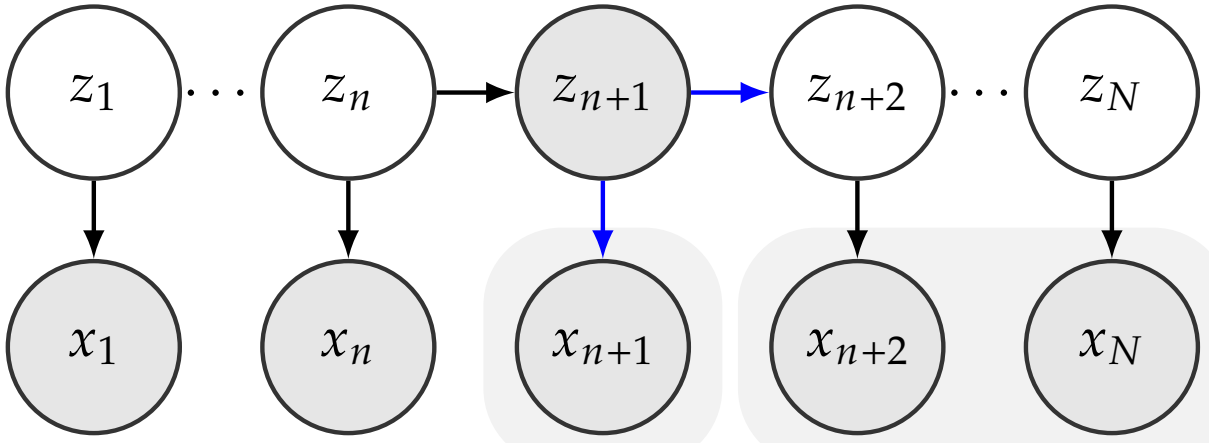
Evaluating the quantities $\beta(\boldsymbol{z}_n)$



$$\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\} \perp\!\!\!\perp \{\boldsymbol{z}_n\} \,|\, \{\boldsymbol{z}_{n+1}\}$$

$$p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) = p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) \tag{13.27}$$

Evaluating the quantities $\beta(\boldsymbol{z}_n)$



$$\{\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N\} \perp\!\!\!\perp \{\boldsymbol{x}_{n+1}\} \,\big|\, \{\boldsymbol{z}_{n+1}\}$$

$$p(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) = p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) \quad (13.28^*)$$

Evaluating the quantities $\beta(z_n)$

$$
\begin{aligned}
\beta(z_n) &\equiv p(x_{n+1}, \ldots, x_N | z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N, z_{n+1} | z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_n, z_{n+1}) p(z_{n+1} | z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_{n+1}) p(z_{n+1} | z_n) \\
&= \sum_{z_{n+1}} p(x_{n+2}, \ldots, x_N | z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n) \\
&= \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n)
\end{aligned}
$$

Evaluating the quantities $\beta(\boldsymbol{z}_n)$

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

$\beta(\boldsymbol{z}_n)$ can be evaluated in terms of $\beta(\boldsymbol{z}_{n+1})$ by the recursion expression:

$$\beta(\boldsymbol{z}_n) = \sum_{\boldsymbol{z}_{n+1}} \beta(\boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) \tag{13.38}$$

with the starting condition

$$\beta(\boldsymbol{z}_N) = 1$$

for all setting of $\boldsymbol{z}_N$. We will verify this in the next slide.

Evaluating the quantities $\beta(\boldsymbol{z}_n)$

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

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

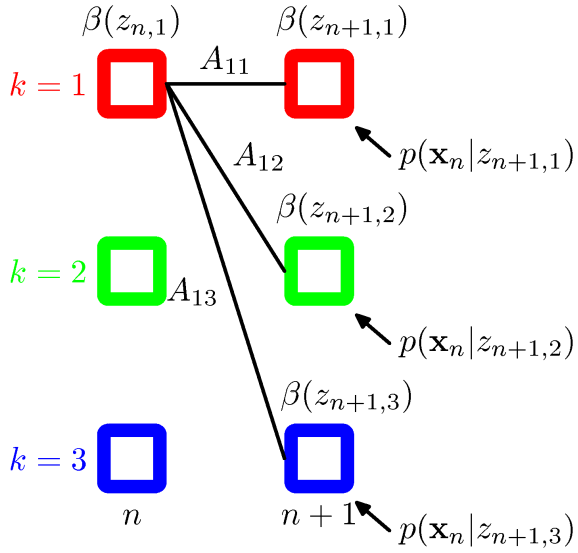
$$\gamma(\boldsymbol{z}_n) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n)p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_n|\boldsymbol{z}_n)}{p(\boldsymbol{X})} = \frac{\alpha(\boldsymbol{z}_n)\beta(\boldsymbol{z}_n)}{p(\boldsymbol{X})} \tag{13.33}$$

By setting $n = N$ in (13.33) we have

$$p(\boldsymbol{z}_N|\boldsymbol{X}) = \frac{p(\boldsymbol{X}, \boldsymbol{z}_N)\beta(\boldsymbol{z}_N)}{p(\boldsymbol{X})} \tag{13.39}$$

(13.39) will be correct if $\beta(\boldsymbol{z}_N) = 1$ for all settings of $\boldsymbol{z}_N$.

Evaluating the quantities $\beta(\boldsymbol{z}_n)$



$$\begin{aligned}\beta(\boldsymbol{z}_n) &\equiv p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{z}_n)\\ &= \sum_{\boldsymbol{z}_{n+1}} \beta(z_{n+1})p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\end{aligned} \tag{13.38}$$

Denote $\beta(\boldsymbol{z}_n) = \begin{bmatrix}\beta(z_{n1}) & \ldots & \beta(z_{nK})\end{bmatrix}^T$. $\mu_{*,n+1}$ is the $(n+1)^{th}$ column of the emission matrix $\boldsymbol{\Phi}$.

$$\begin{bmatrix}\beta(z_{n,1})\\ \vdots \\ \beta(z_{n,K})\end{bmatrix} = \begin{bmatrix}A_{11} & A_{12} & \ldots & A_{1K}\\ \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & \ldots & A_{KK}\end{bmatrix}\left(\begin{bmatrix}\beta(z_{n+1,1})\\ \vdots \\ \beta(z_{n+1,K})\end{bmatrix} \circ \begin{bmatrix}\mu_{1,n+1}\\ \vdots \\ \mu_{K,n+1}\end{bmatrix}\right)$$

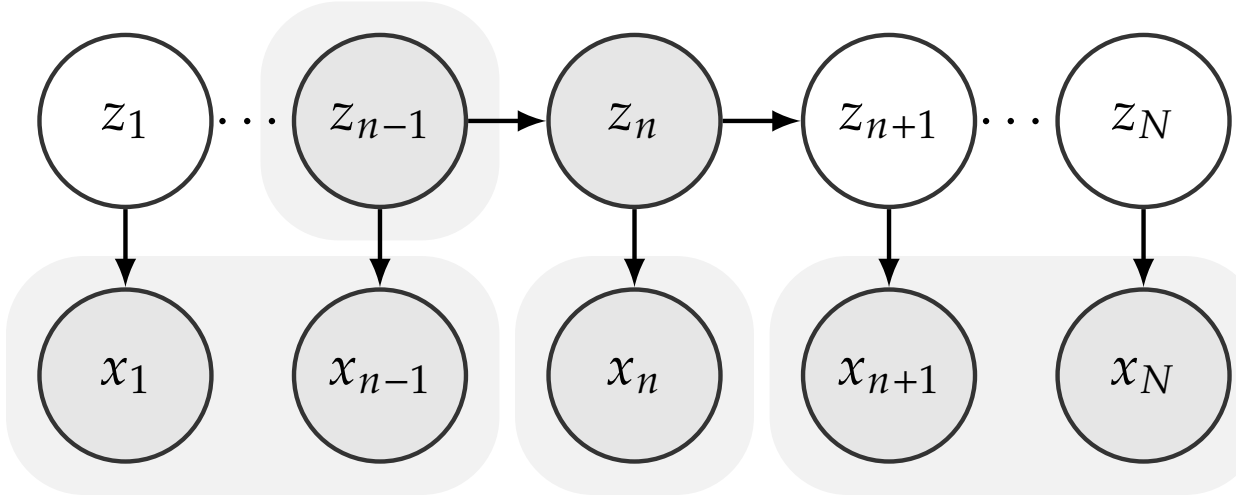
$$\beta(\boldsymbol{z}_n) = \boldsymbol{A}\left[\beta(\boldsymbol{z}_{n+1}) \circ \boldsymbol{\mu}_{*,n}\right]$$

$$\beta(\boldsymbol{z}_N) = \begin{bmatrix}1 & \ldots & 1\end{bmatrix}^T$$

Evaluating $\xi(z_{n-1}, z_n)$

$$
\begin{aligned}
p(\boldsymbol{X}) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}) p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n) \qquad (13.29)
\end{aligned}
$$

$$
p(\boldsymbol{X}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_{n-1}) p(\boldsymbol{x}_n | \boldsymbol{z}_n) \beta(\boldsymbol{z}_n)
$$



Evaluating $\xi(z_{n-1}, z_n)$

Using Bayes theorem

$$
\begin{aligned}
\xi(z_{n-1}, z_N) &\equiv p(z_{n-1}, z_n|\boldsymbol{X}) \\
&= \frac{p(\boldsymbol{X}|z_{n-1}, z_n)p(z_{n-1}, z_n)}{p(\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}|z_{n-1})p(\boldsymbol{x}_n|z_n)\beta(z_n)p(z_{n-1}, z_n)}{p(\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}|z_{n-1})p(\boldsymbol{x}_n|z_n)\beta(z_n)p(z_n|z_{n-1})p(z_{n-1})}{p(\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}|z_{n-1})p(z_{n-1})p(\boldsymbol{x}_n|z_n)\beta(z_n)p(z_n|z_{n-1})}{p(\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, z_{n-1})p(\boldsymbol{x}_n|z_n)\beta(z_n)p(z_n|z_{n-1})}{p(\boldsymbol{X})} \\
&= \frac{\alpha(z_{n-1})p(\boldsymbol{x}_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(\boldsymbol{X})} \qquad (13.43)
\end{aligned}
$$

EM for HMM

Algorithm 2 EM for HMM

repeat
- E step:
  - Compute $\alpha(\boldsymbol{z}_n)$
  - Compute $\beta(\boldsymbol{z}_n)$
  - Compute $\gamma(\boldsymbol{z}_n)$
  - Compute $\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n)$
- M step:
  - Compute π_k
  - Compute A_{jk}
  - Compute μ_k
  - Compute Σ_k

until convergence

Scaling Factor

$\alpha(z_n)$ is the joint probability of observing $x_1, ..., x_n$ which are in the state z_n.

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

The values $\alpha(\boldsymbol{z}_n)$ are obtained from the previous value $\alpha(\boldsymbol{z}_{n-1})$ by multiplying by quantities $p(z_n|z_{n-1})$ and $p(x_n|z_n)$.

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

Because these probabilities are often significantly less than unity, the values of $\alpha(\boldsymbol{z}_n)$ can go to zero very quickly and exceed the precision of double precision floating points (1.175494351e-38).

Scaling Factor

$$\alpha(\boldsymbol{z}_n) \equiv p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n) \qquad = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) p(\boldsymbol{z}_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n)$$

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

Because $\hat{\alpha}(\boldsymbol{z}_n)$ is a probability distribution over K variables for any value of n, we have $\sum_{k=1}^{K} z_{nk} = 1$.
We expect $\hat{\alpha}(\boldsymbol{z}_n)$ to be *well behaved numerically* because not all values z_{nk} can be arbitrarily small.

Scaling Factor

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

Introducing scaling factors

$$\begin{aligned} c_1 &= p(x_1) \\ c_n &= p(x_n|x_1, \ldots, x_{n-1}) \end{aligned} \tag{13.56}$$

we have

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_n|x_1, \ldots, x_{n-1})p(x_1, \ldots, x_{n-1}) \\ &= c_n c_{n-1} \ldots c_1 \\ &= \prod_{m=1}^{n} c_m \end{aligned} \tag{13.57}$$

Scaling Factor

$$p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_1, \ldots, x_{n-1})$$
$$= \prod_{m=1}^{n} c_m \tag{13.57}$$

$$\hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)} = \frac{\alpha(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$
$$= \frac{\alpha(\boldsymbol{z}_n)}{\prod_{m=1}^{n} c_m}$$

$$\alpha(\boldsymbol{z}_n) = \left(\prod_{m=1}^{n} c_m\right) \hat{\alpha}(\boldsymbol{z}_n) \tag{13.58}$$

Scaling Factor

$$\alpha(\boldsymbol{z}_n) = \left(\prod_{m=1}^{n} c_m\right) \hat{\alpha}(\boldsymbol{z}_n) \tag{13.58}$$

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

From (13.58) and (13.36) we obtain

$$\left(\prod_{m=1}^{n} c_m\right) \hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n) \sum_{\boldsymbol{z}_{n-1}} \left(\prod_{m=1}^{n-1} c_m\right) \hat{\alpha}(\boldsymbol{z}_{n-1}) p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$$

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

Scaling Factor

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

Denoting $\tilde{\alpha}(\boldsymbol{z}_n) = c_n \hat{\alpha}(\boldsymbol{z}_n)$ we have

$$\sum_{k=1}^{K} \tilde{\alpha}(\boldsymbol{z}_n) = \sum_{k=1}^{K} c_n \hat{\alpha}(\boldsymbol{z}_n) = c_n \sum_{k=1}^{K} \hat{\alpha}(\boldsymbol{z}_n) = c_n$$

Scalling Factor

At each stage we calculate $\tilde{\alpha}(z_n), c_n, \alpha(z_n)$ as follows

$$\tilde{\alpha}(z_n) = p(x_n|z_n) \sum_{z_{n-1}} \hat{\alpha}(z_{n-1}) p(z_n|z_{n-1})$$

$$c_n = \sum_{k=1}^{K} c_n \hat{\alpha}(z_n)$$

c_1 is calculated as follows

$$c_1 = p(x_1) = \sum_{z_1} p(x_1, z_1) = \sum_{z_1} p(x_1|z_1) p(z_1) = \sum_{k=1}^{K} \pi_k p(x_1|\phi_k)$$

Scaling Factor

$\beta(\boldsymbol{z}_n)$ is the conditional probability of future observation $x_{n+1}, ..., x_N$ given the state z_n.

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

The values $\beta(\boldsymbol{z}_n)$ are obtained from the previous value $\beta(\boldsymbol{z}_{n+1})$

$$\beta(\boldsymbol{z}_n) = \sum_{\boldsymbol{z}_{n+1}} \beta(z_{n+1}) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) \tag{13.38}$$

Scaling Factor

We define the normalized version of β given by

$$\hat{\beta}(\boldsymbol{z}_N) = 1,$$

$$\hat{\beta}(\boldsymbol{z}_n) = \frac{p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_n)}{p(\boldsymbol{x}_{n+1}, \ldots, x_N | x_1, \ldots, x_n)}$$

Similarly we have

$$\beta(\boldsymbol{z}_n) = \left(\prod_{m=n+1}^{N} c_m \right) \hat{\beta}(\boldsymbol{z}_n) \tag{13.60}$$

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

Scaling Factor

At each stage we calculate $\tilde{\beta}(z_n)$ and $\hat{\beta}(z_n)$ as follows

$$\tilde{\beta}(z_n) = \sum_{z_{n+1}} \hat{\beta}(z_{n+1}) p(\boldsymbol{x}_{n+1}|z_{n+1}) p(z_{n+1}|z_n)$$

$$\hat{\beta}(z_n) = \tilde{\beta}(z_n)/c_{n+1}$$

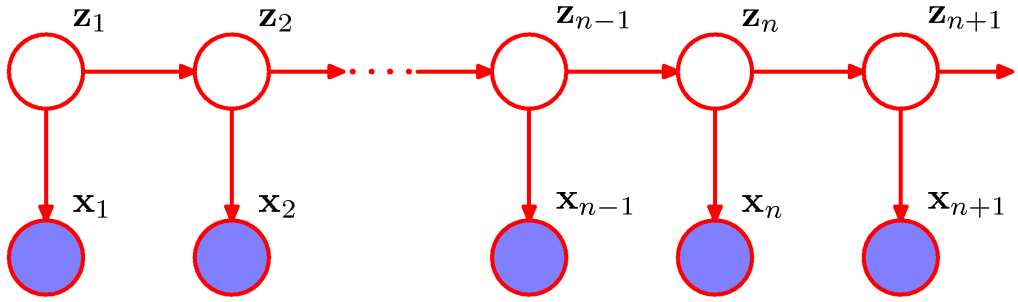
Here c_n were previously computed in the α phase.

$\gamma(z_n)$ and $\xi(z_{n-1}, z_n)$ can be computed by

$$\gamma(z_n) = \frac{\hat{\alpha}(z_n)\hat{\beta}(z_n)}{p(\boldsymbol{X})} \tag{13.64}$$

$$\xi(z_{n-1}, z_n) = c_n^{-1} \hat{\alpha}(z_{n-1}) p(\boldsymbol{x}_n|z_n) p(z_n|z_{n-1}) \hat{\beta}(z_n) \tag{13.65}$$

Linear Dynamical Systems (LDS)



Transition distribution

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

Emission distribution

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

Initial state distribution

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

The parameters of the model: $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{\Gamma}, \boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{P}_0\}$

E-Step in LDS

$$\hat{\alpha}(\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n) \tag{18.84}$$

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

Analogous to HMM, the recursion equation of $\hat{\alpha}(\boldsymbol{z}_n)$ have the form

$$c_n\hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n)\int\hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} \tag{13.85}$$

E-Step in LDS

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

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

$$\hat{\alpha}(\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n) \tag{18.84}$$

$$c_n\hat{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n)\int \hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} \tag{13.85}$$

Using (13.75), (13.76), (13.84) we can rewrite (13.85) as

$$\begin{aligned} &c_n\mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{V}_n) \\ &\quad = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{C}\boldsymbol{z}_n, \boldsymbol{\Sigma})\int \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\boldsymbol{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \boldsymbol{V}_{n-1})d\boldsymbol{z}_{n-1} \end{aligned} \tag{13.86}$$

E-Step in LDS

$$c_n \mathcal{N}(z_n|\mu_n, V_n)$$
$$= \mathcal{N}(x_n|Cz_n, \Sigma) \int \mathcal{N}(z_n|Az_{n-1}, \Gamma)\mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})dz_{n-1} \tag{13.86}$$

Making use of the results (2.115) and (2.116) in Chapter 2 we have

$$\mu_n = A\mu_{n-1} + K_n(x_n - CA\mu_{n-1}) \tag{13.89}$$

$$V_n = (I - K_n C)P_{n-1} \tag{13.90}$$

$$c_n = \mathcal{N}(x_n|CA\mu_{n-1}, CP_{n-1}C^T + \Sigma) \tag{13.91}$$

Here, we have defined $P_{n-1} = AV_{n-1}A^T + \Gamma$ (13.88)

E-Step in LDS - Compute $\gamma(z_n)$

In the LDS, it is usual to formulate the backward recursion in terms of $\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n)$. Because $\gamma(z_n)$ must be Gaussian, we can write

$$\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n) = \mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n) \tag{13.98}$$

$$c_{n+1}\hat{\beta}(z_n) = \sum_{z_{n+1}} \hat{\beta}(z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n) \tag{13.62}$$

From (13.62) we can write $\hat{\beta}(z_n)$ in the form

$$c_{n+1}\hat{\beta}(z_n) = \int \hat{\beta}(z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n)dz_{n+1} \tag{13.99}$$

E-Step in LDS - Compute $\gamma(z_n)$

$$\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n) = \mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n) \tag{13.98}$$

$$\hat{\alpha}(z_n) = \mathcal{N}(z_n|\mu_n, V_n) \tag{18.84}$$

$$\mu_n = A\mu_{n-1} + K_n(x_n - CA\mu_{n-1}) \tag{13.89}$$

$$V_n = (I - K_n C)P_{n-1} \tag{13.90}$$

$$c_n = \mathcal{N}(x_n|CA\mu_{n-1}, CP_{n-1}C^T + \Sigma) \tag{13.91}$$

$$c_{n+1}\hat{\beta}(z_n) = \int \hat{\beta}(z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n)dz_{n+1} \tag{13.99}$$

E-Step in LDS - Compute $\gamma(\boldsymbol{z}_n)$

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

After some manipulation we obtain

$$\hat{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_n + \boldsymbol{J}_n(\hat{\boldsymbol{\mu}}_{n+1} - \boldsymbol{A}\boldsymbol{\mu}_n) \tag{13.100}$$

$$\hat{\boldsymbol{V}}_n = \boldsymbol{V}_n + \boldsymbol{J}_n(\hat{\boldsymbol{V}}_{n+1} - \boldsymbol{P}_n)\boldsymbol{J}_n^T \tag{13.101}$$

where we have defined

$$\boldsymbol{J}_n = \boldsymbol{V}_n\boldsymbol{A}^T(\boldsymbol{P}_n)^{-1} \tag{13.102}$$

E-Step in LDS - Compute $\xi(z_n, z_{n-1})$

$$\begin{aligned}\xi(z_{n-1}, z_n) &= c_n^{-1}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n) \\ &= \frac{\mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})\mathcal{N}(z_n|Az_{n-1}, \Gamma)\mathcal{N}(x_n|Cz_n, \Sigma)\mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n)}{c_n\hat{\alpha}(z_n)}\end{aligned} \tag{13.103}$$

Here we have used

$$\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n) = \mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n) \tag{13.98}$$

$$p(z_n|z_{n-1}) = \mathcal{N}(z_n|Az_{n-1}, \Gamma) \tag{13.75}$$

$$p(x_n|z_n) = \mathcal{N}(x_n|Cz_n, \Sigma) \tag{13.76}$$

E-Step in LDS - Compute $\xi(z_n, z_{n-1})$

$$\begin{aligned}\xi(z_{n-1}, z_n) &= c_n^{-1}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n) \\ &= \frac{\mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})\mathcal{N}(z_n|Az_{n-1}, \Gamma)\mathcal{N}(x_n|Cz_n, \Sigma)\mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n)}{c_n\hat{\alpha}(z_n)}\end{aligned} \tag{13.103}$$

After some manipulation we see that $\xi(z_n, z_{n-1})$ is a Gaussian with mean vector and covariance matrix given by

$$\begin{bmatrix}\mathbb{E}(z_{n-1}) \\ \mathbb{E}(z_n)\end{bmatrix} = \begin{bmatrix}\hat{\mu}_{n-1} \\ \hat{\mu}_n\end{bmatrix}$$

$$\begin{bmatrix}\text{cov}[z_{n-1}, z_{n-1}] & \text{cov}[z_{n-1}, z_n] \\ \text{cov}[z_n, z_{n-1}] & \text{cov}[z_n, z_n]\end{bmatrix} = \begin{bmatrix}\hat{V}_{n-1} & \hat{J}_{n-1}\hat{V}_n \\ \hat{V}_n\hat{J}_{n-1}^T & \hat{V}_n\end{bmatrix}$$

M-Step in LDS

The parameters of the model $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{\Gamma}, \boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{P}_0\}$ can be computed as follows:

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

$$\begin{aligned}\boldsymbol{\Gamma}^{new} = \frac{1}{N-1}\sum_{n=2}^{N}\Big\{ &\mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_n^T] - \boldsymbol{A}^{new}\,\mathbb{E}[\boldsymbol{z}_{n-1} \boldsymbol{z}_n^T] \\ &- \mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_{n-1}^T](\boldsymbol{A}^{new})^T + \boldsymbol{A}^{new}\,\mathbb{E}[\boldsymbol{z}_{n-1} \boldsymbol{z}_{n-1}^T](\boldsymbol{A}^{new})^T\Big\}\end{aligned} \tag{13.114}$$

M-Step in LDS

The parameters of the model $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{\Gamma}, \boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{P}_0\}$ can be computed as follows:

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

$$\begin{aligned} \boldsymbol{\Sigma}^{new} = \frac{1}{N} \sum_{n=1}^{N} \Big\{ & \boldsymbol{x}_n \boldsymbol{x}_n^T - \boldsymbol{C}^{new} \, \mathbb{E}[\boldsymbol{z}_n] \boldsymbol{x}_n^T \\ & - \boldsymbol{x}_n \, \mathbb{E}[\boldsymbol{z}_n^T] (\boldsymbol{C}^{new})^T + \boldsymbol{C}^{new} \, \mathbb{E}[\boldsymbol{z}_n \boldsymbol{z}_n^T] (\boldsymbol{C}^{new})^T \Big\} \end{aligned} \tag{13.116}$$

$$\boldsymbol{\mu}_0^{new} = \mathbb{E}[\boldsymbol{z}_1] \tag{13.110}$$

$$\boldsymbol{P}_0^{new} = \mathbb{E}[\boldsymbol{z}_1 \boldsymbol{z}_1^T] \, \mathbb{E}[\boldsymbol{z}_1] \, \mathbb{E}[\boldsymbol{z}_1^T] \tag{13.111}$$

EM in LDS

To determine parameters θ, the update of the posterior distribution at E-step and the update of the parameter at M-step are repeated until the expected complete data log likelihood $Q(\theta, \theta^{old})$ converges.