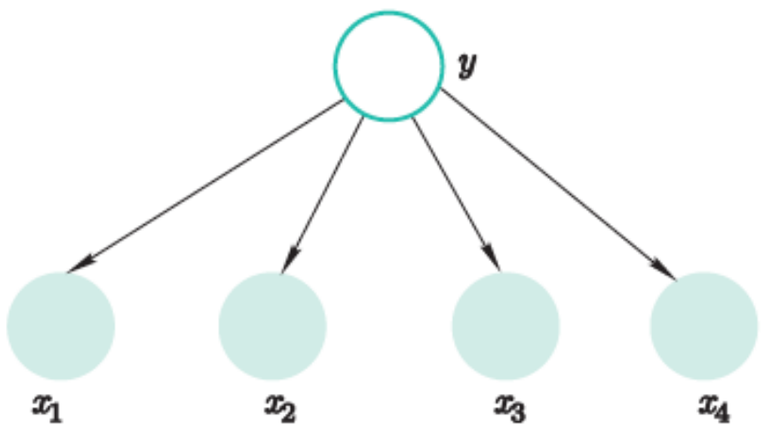
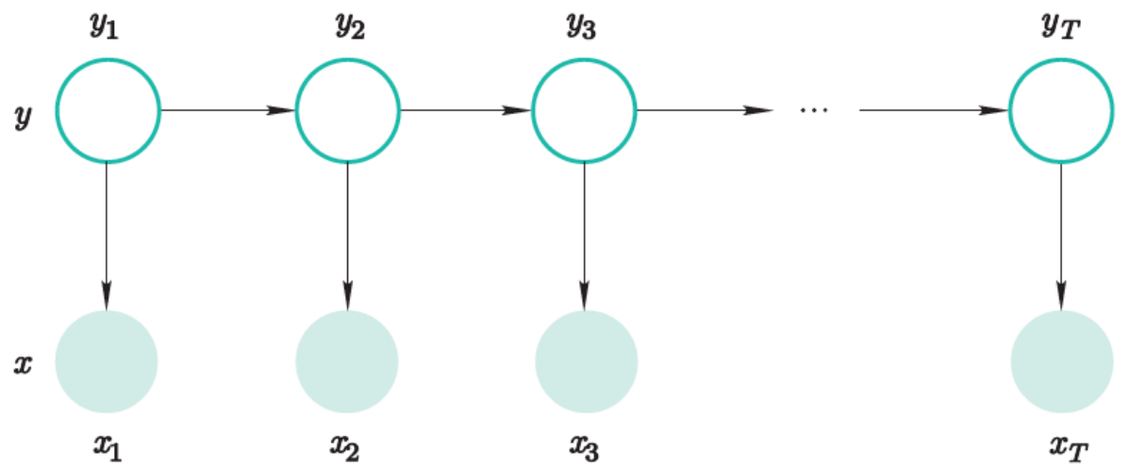
(Probabilistic) Graphical Models





Some of figures in the slides are from Bishop (2016)

Roadmap

- What are graphical models and why should we study them?
- Directed/Undirected graphical models (i.e Bayesian networks and Markov random field)
- Mixture models and EM algorithm
- Hidden Markov Model (HMM) and Conditional Random Field (CRF)
- Approximate inference (variational inference)
- Markov Chain Monte Carlo (MCMC)

The Fundamental Questions in ML

- Representation
  - How to capture/model uncertainty in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?
  - …
- Inference $P(X_i|\mathcal{D})$
  - How do I answer questions/queries based on the model or the data?
- Learning $\mathcal{M} = \arg\max_{\mathcal{M}\in M} L(\mathcal{D};\mathcal{M})$
  - What model is “right” for my data?

- Representation

2^6 configurations?

$$P(X_1, X_2, X_3, X_4, X_5, X_6)$$

  - What is the joint distribution over multiple variables?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

  - Should we really need sum over 2^5 configurations to obtain P(X_i)?

- Learning: where do we get these probabilities?

  - Maximum likelihood estimation

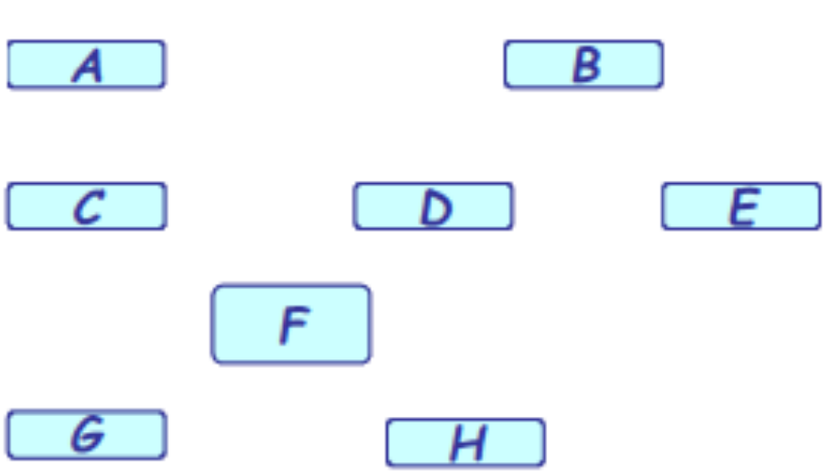
What are graphical models? why should we study them?

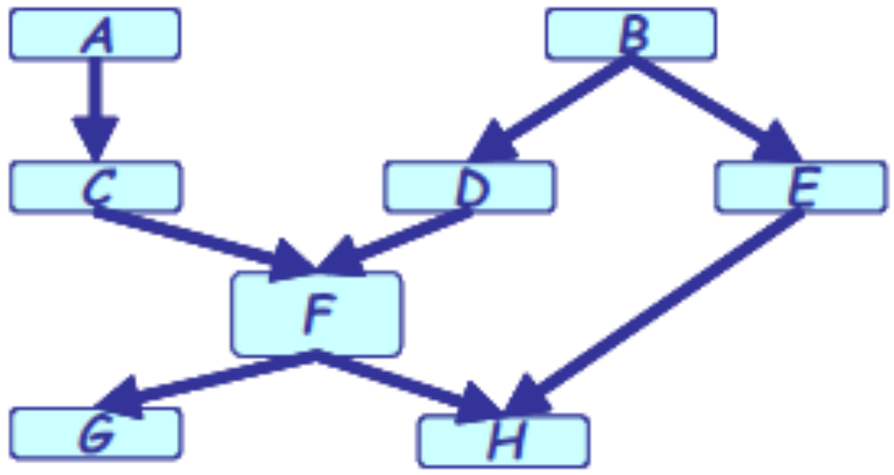
Graphical Models

- Use graph to represent joint probability distribution over multiple random variables
  - Express the dependency (via edges) between variables (nodes)
  - Simple, suitable for interpretability

- It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with **structured semantics**.

- It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables.





$$P(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)$$

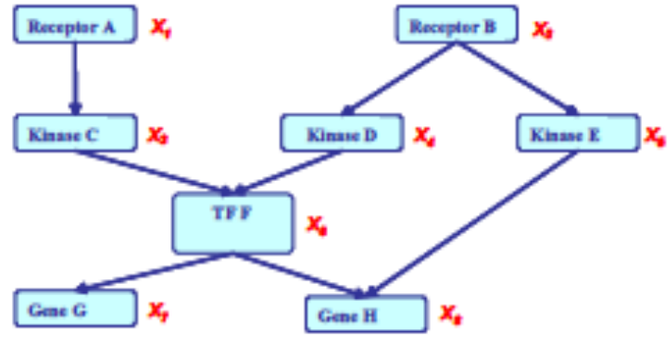
$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3,X_4)P(X_7 \mid X_6)P(X_8 \mid X_5,X_6)$$

Two Types of GMs

- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

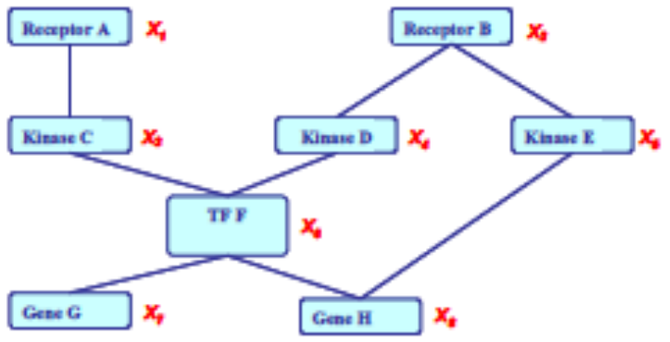
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$



- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Adapted from Prof. Eric Xing

Applications of GMs

- Machine learning and statistics
- Computer vision and graphics
- Natural language processing
- Information retrieval
- Computational biology and bioinformatics
- Finance and economics
- Robotics control
- Other applications related to decision making under uncertainty

Why Graphical Models?

- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.
- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- **Many of the classical multivariate probabilistic systems** studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

By Michael Jordan

Directed Graphical Models

- Also named as Bayesian networks (BN)

- A directed graph whose **nodes** represent the random variables and whose **edges** represent direct influence of one variable on another

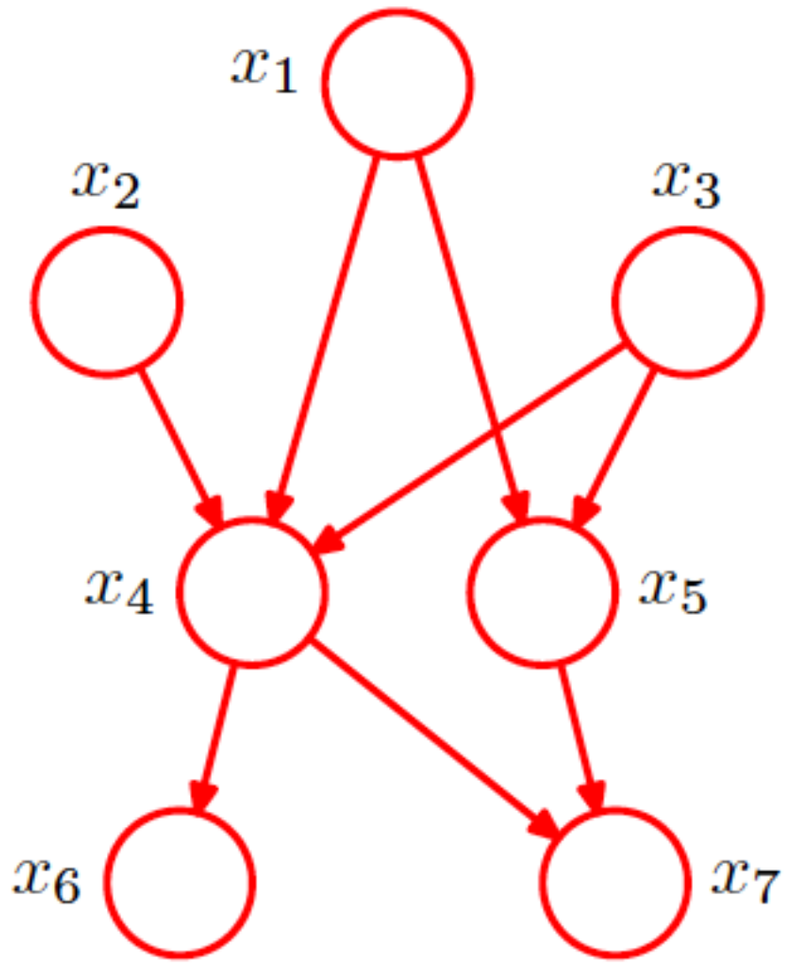
- A data structure that provides the skeleton for **representing a joint distribution** compactly in a **factorized** way

- A compact representation for a set of **conditional independence assumptions** about a distribution

- Encoding a **generative sampling process** executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents.

- Joint distribution can be represented in product of factorized conditional distributions over its parents.

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$



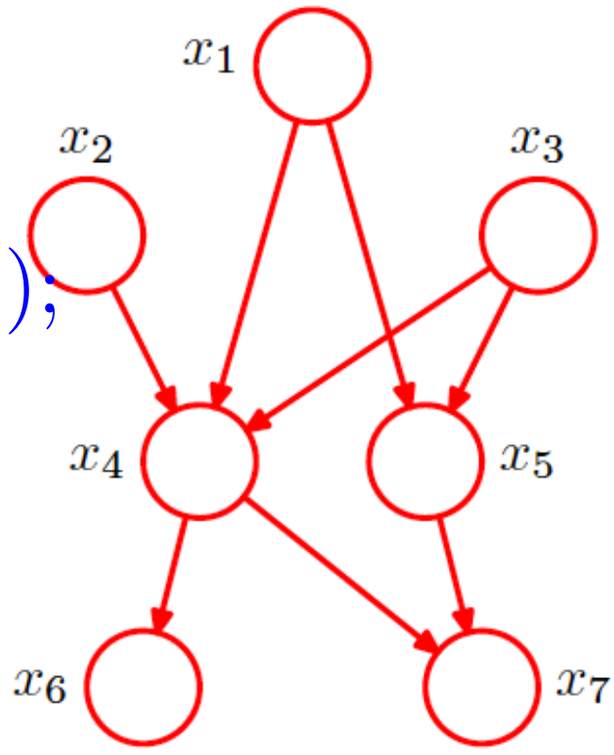
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

- Based on pre-specified ordering of nodes.

- No directed cycles: directed acyclic graphs (DAGs)

- Generative process
  - Aiming to obtain a sample according to the directed graphical models
  - **Ancestral sampling:** starting from lowest-numbering node and draw each variable given its already sampled parents.

1. Sample $x_1 \sim p(x_1)$, $x_2 \sim p(x_1)$, $x_3 \sim p(x_3)$;
2. Sample $x_4 \sim p(x_4|x_1, x_2, x_3)$, $x_5 \sim p(x_5|x_1, x_3)$;
3. Sample $x_6 \sim p(x_4|x_4)$, $x_7 \sim p(x_7|x_4, x_5)$.



Conditional Independence

If $p(a|b,c) = p(a|c)$, then we say that a is conditionally independent of b given c, denoted as $a \perp b|c$.

How to test conditional dependence given a graphical model?

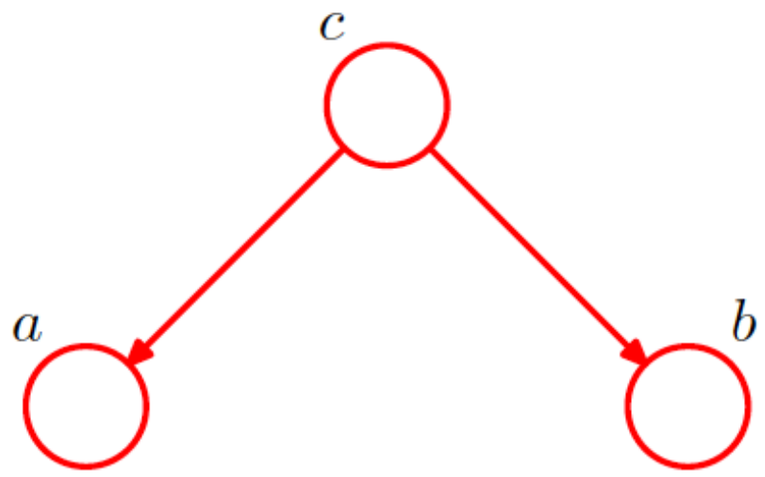
"D-separation" (Pearl, 1988)

Prof. Judea Pearl
ACM Turing Award Laureate, 2011

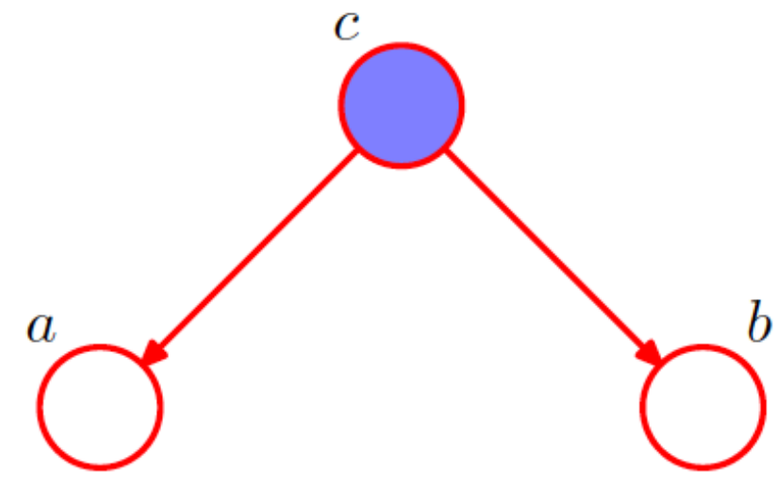
Three Example Graphs

- Example 1:

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



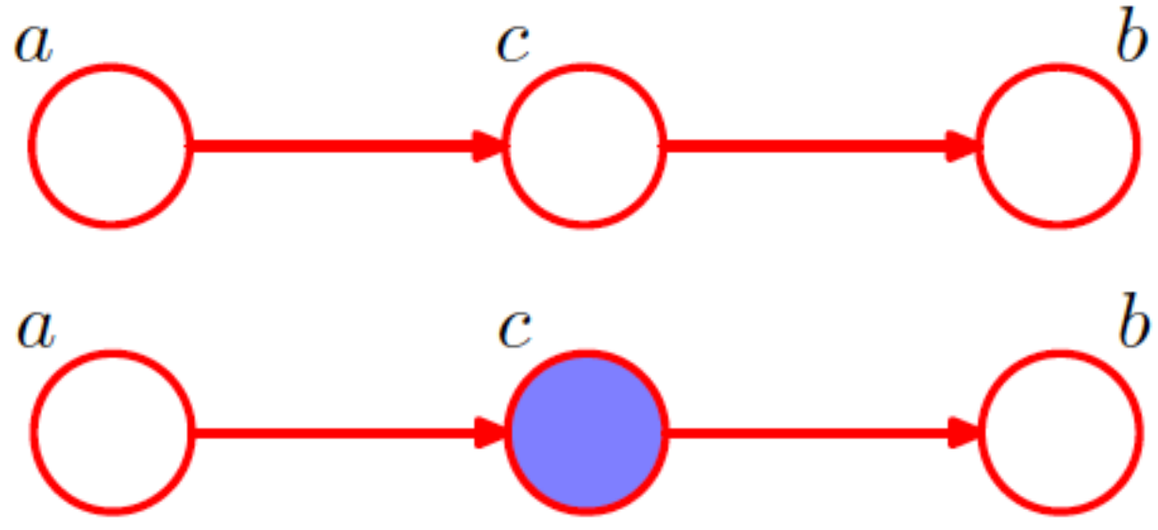
$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \longrightarrow a \not\perp\!\!\!\perp b \mid \emptyset$$



$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned} \longrightarrow a \perp\!\!\!\perp b \mid c$$

tail-to-tail

- Example 2:



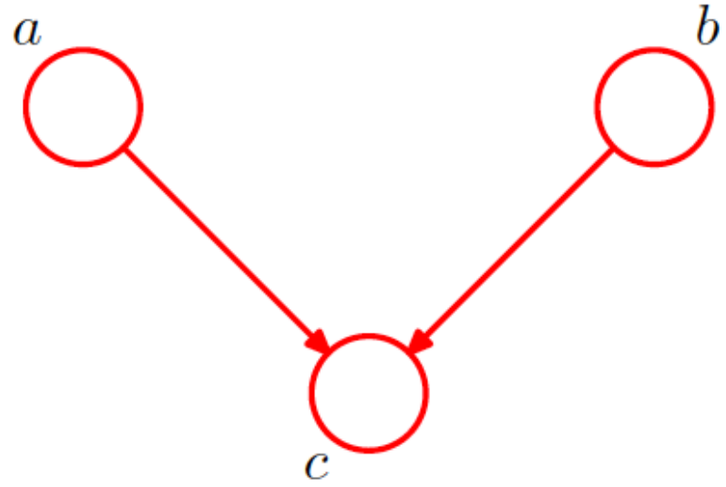
head-to-tail

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

$$p(a,b) = p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a) \longrightarrow a \not\perp\!\!\!\perp b \mid \emptyset$$

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \longrightarrow a \perp\!\!\!\perp b \mid c \\ &= p(a|c)p(b|c) \end{aligned}$$

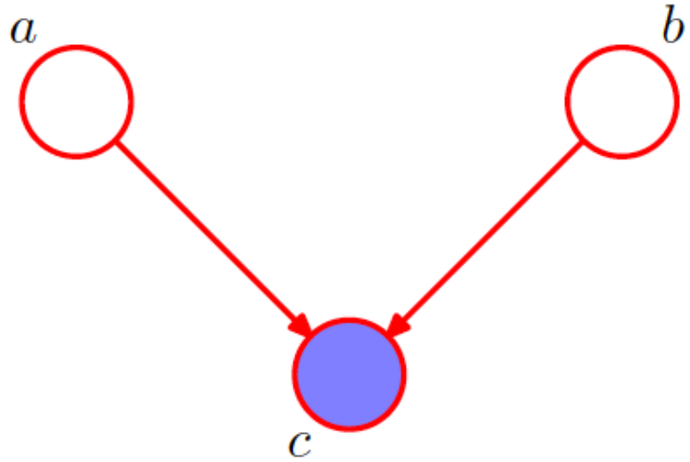
- Example 3:



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

$$p(a,b) = p(a)p(b) \longrightarrow a \perp\!\!\!\perp b \mid \emptyset$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \longrightarrow a \not\perp\!\!\!\perp b \mid c$$



head-to-head

D-separation

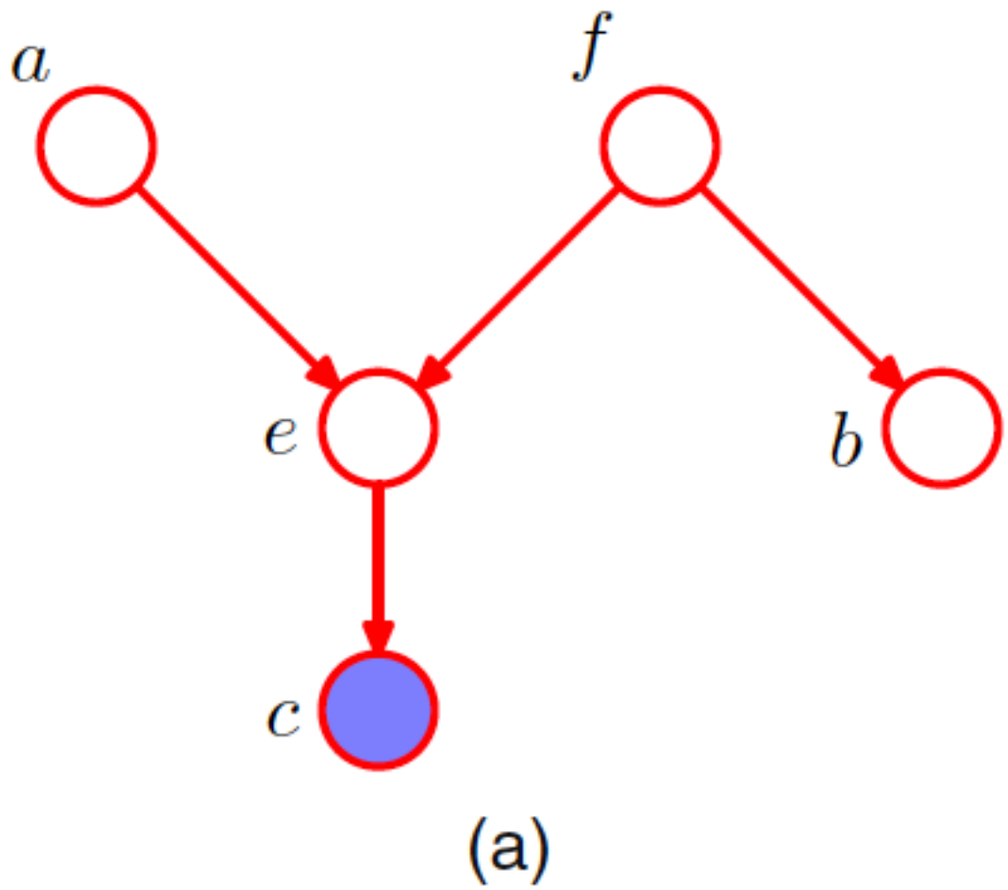
- Consider all paths from set of nodes A to B, we call any path "***blocked***" if they includes a node such that either

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

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

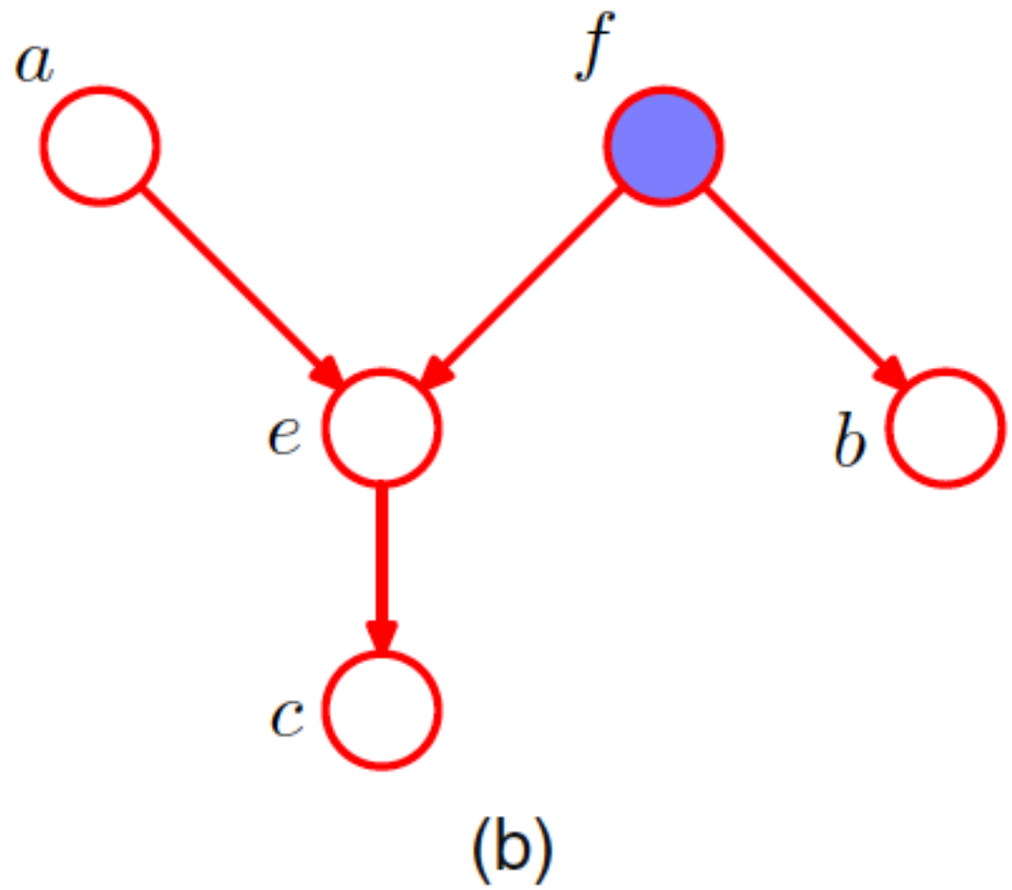
If all paths are blocked, then A is said to be d-separated from B by C, and the joint distribution over all of the variables in the graph will satisfy $A \perp\!\!\!\perp B \mid C$.

Quiz



(a)

a and b are conditionally independent given c?

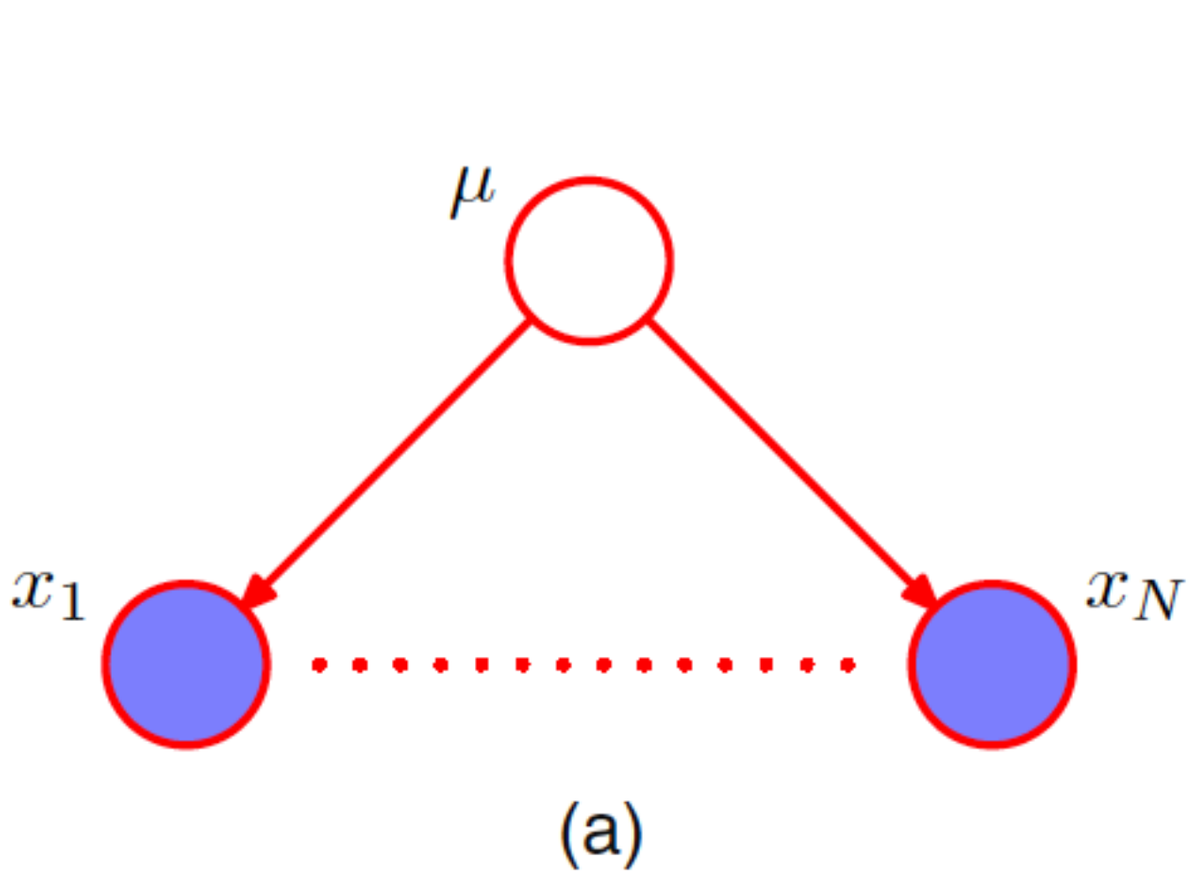


(b)

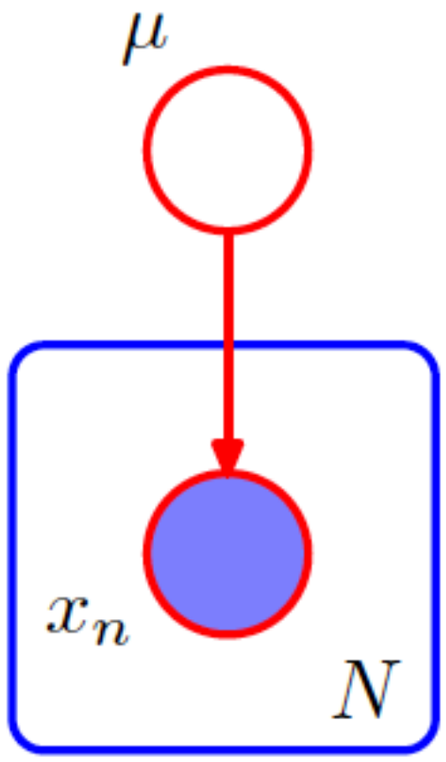
a and b are conditionally independent given f?

- Application of D-separation: **i.i.d Gaussian with mean**

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

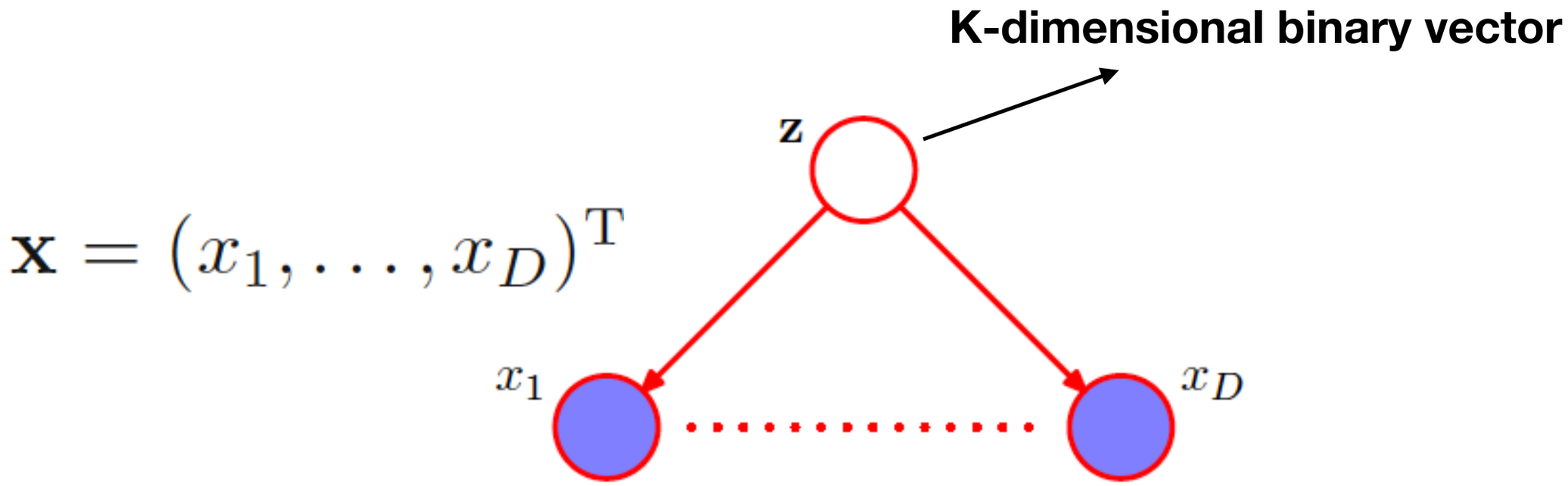


(a)



(b)

- Application of D-separation: **Naive Bayes classifier**



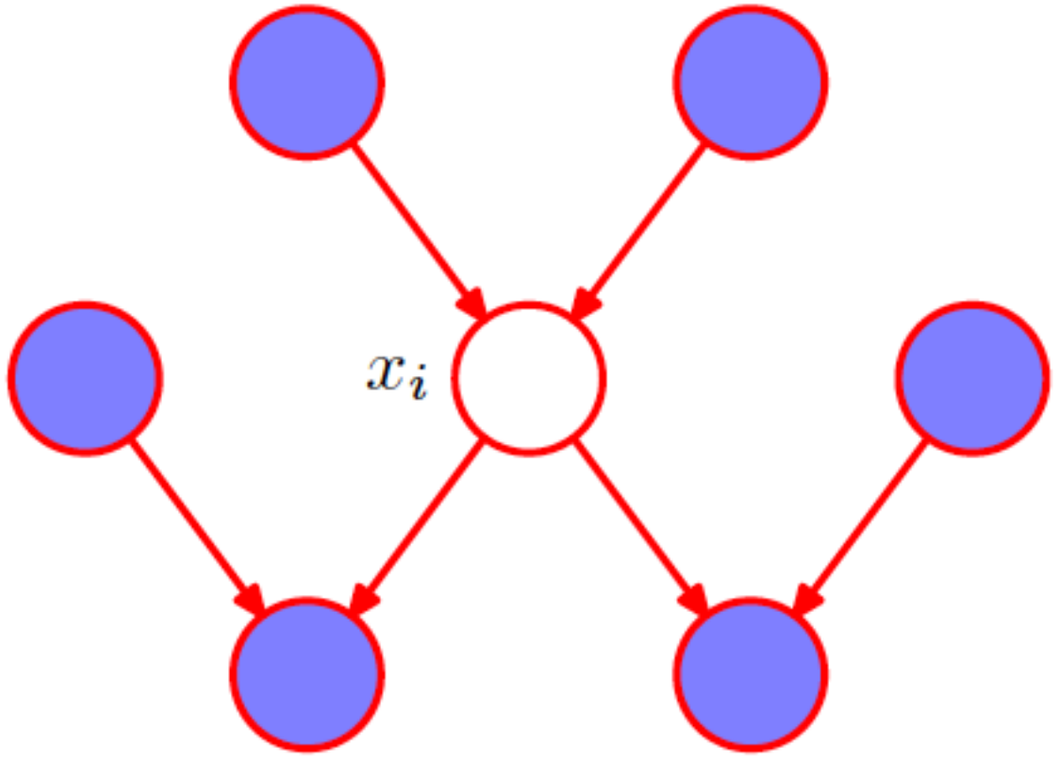
x_i and x_j are conditionally independent given class variable z.

Could you write out the joint distribution?

- Application of D-separation: **Markov Blanket**

Question：what is **the minimal set of nodes** that isolates the xi from the rest of the graph?

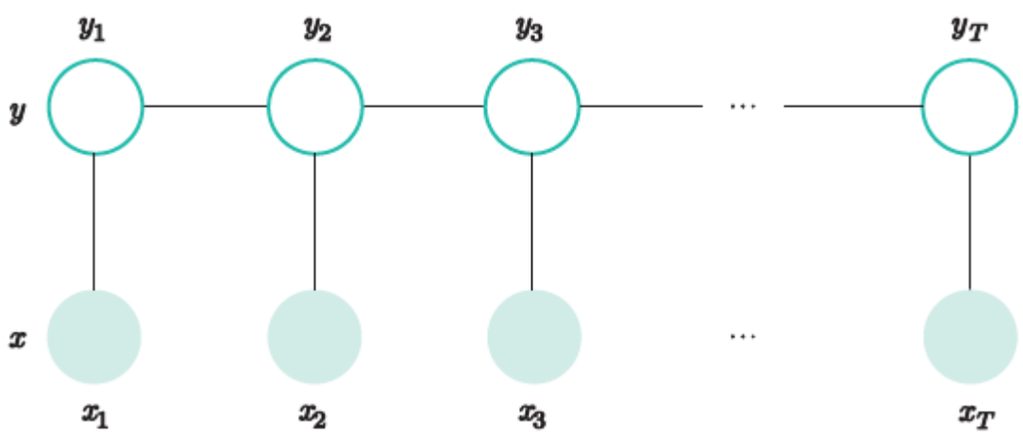
$$p(x_i|MB(x_i)) = p(x_i|x_{\setminus i})$$

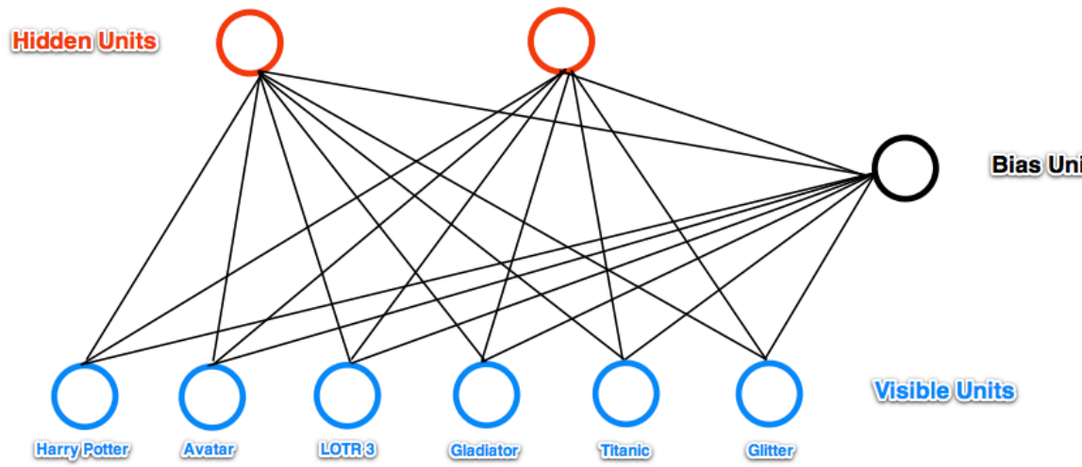


Markov blanket:
parents, children, co-parents

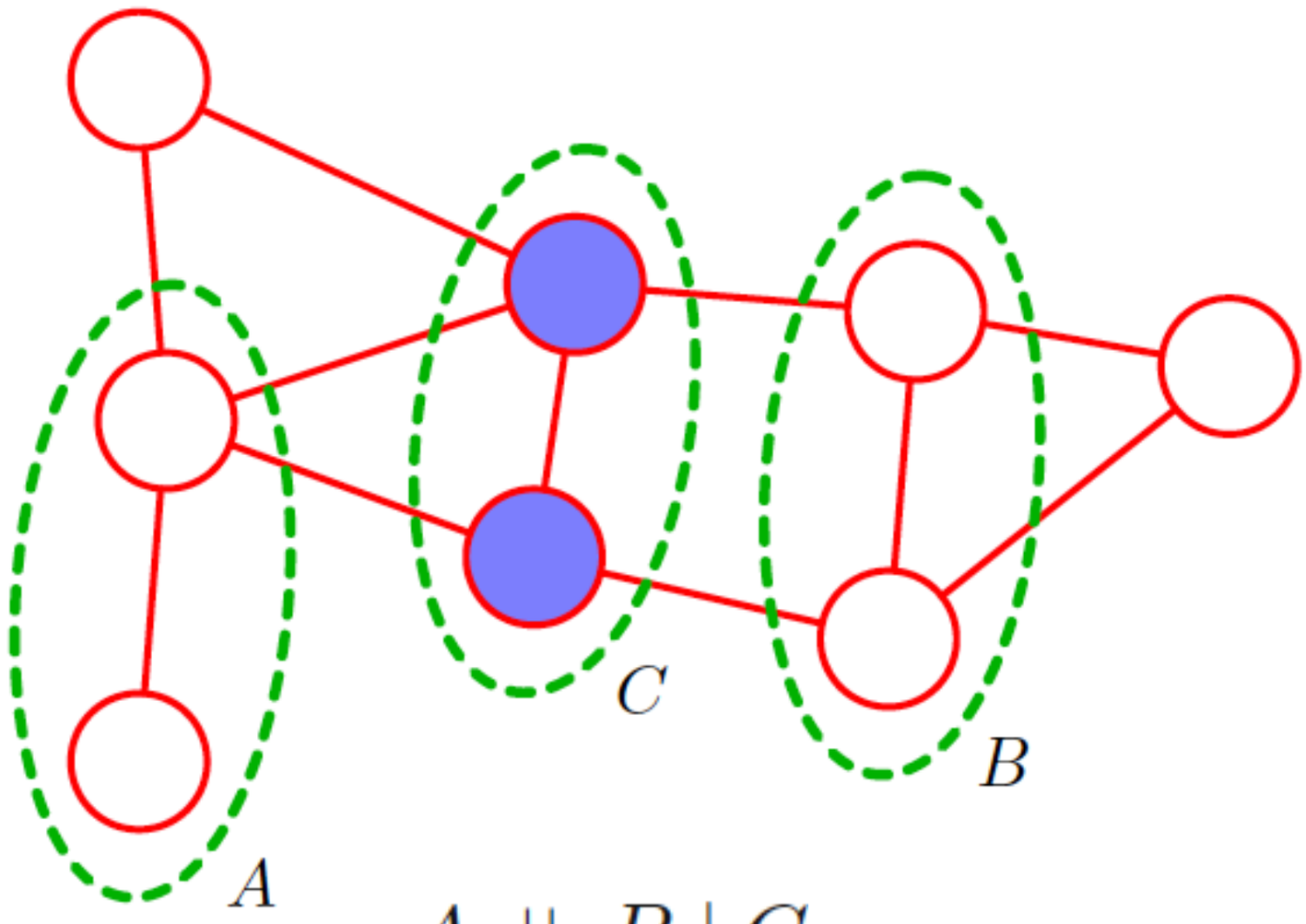
Undirected Graphical Models

- Also called Markov random field (MRF) or Markov network
- No directed links
- Some famous modes UGMs
  - Conditional random field (CRF)
  - Restricted Boltzmann machine (RBM)



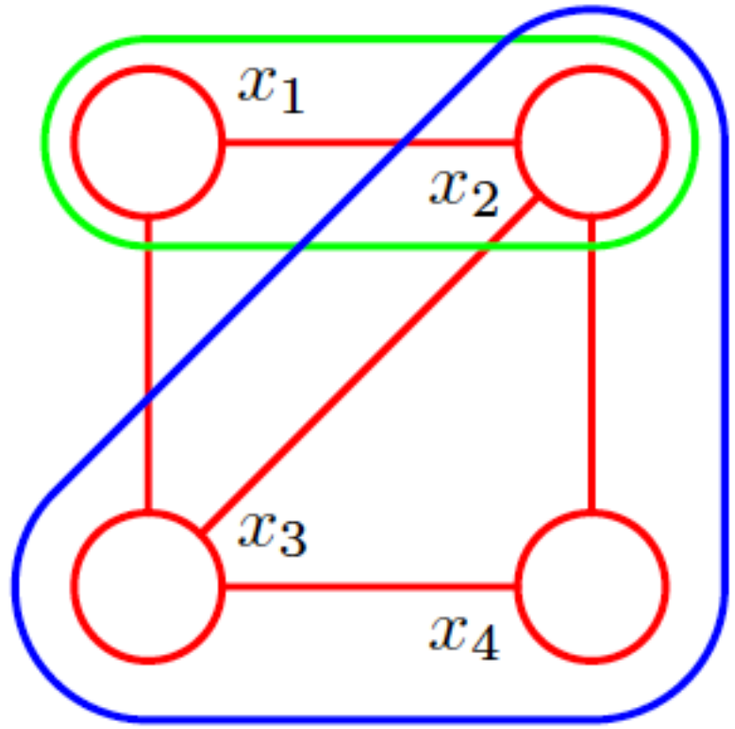


- Simpler conditional independence rule, since we don't have head-to-head node



$$A \perp\!\!\!\perp B \mid C$$

- Factorization properties for UGMs
  - Clique: group of nodes that are linked (green)
  - Maximal clique



$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C)$$

potential function over maximal clique

normalization constant partition function

$$Z = \sum_{\mathbf{x}}\prod_{C}\psi_C(\mathbf{x}_C)$$

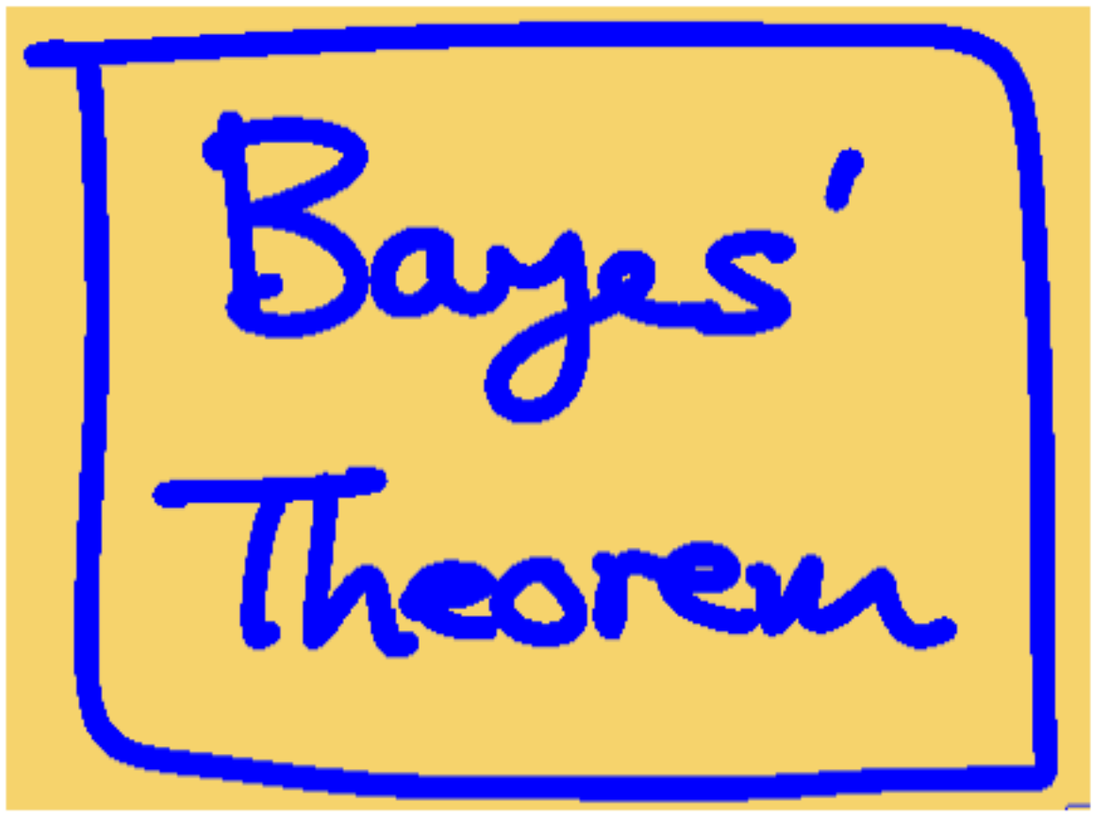
$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

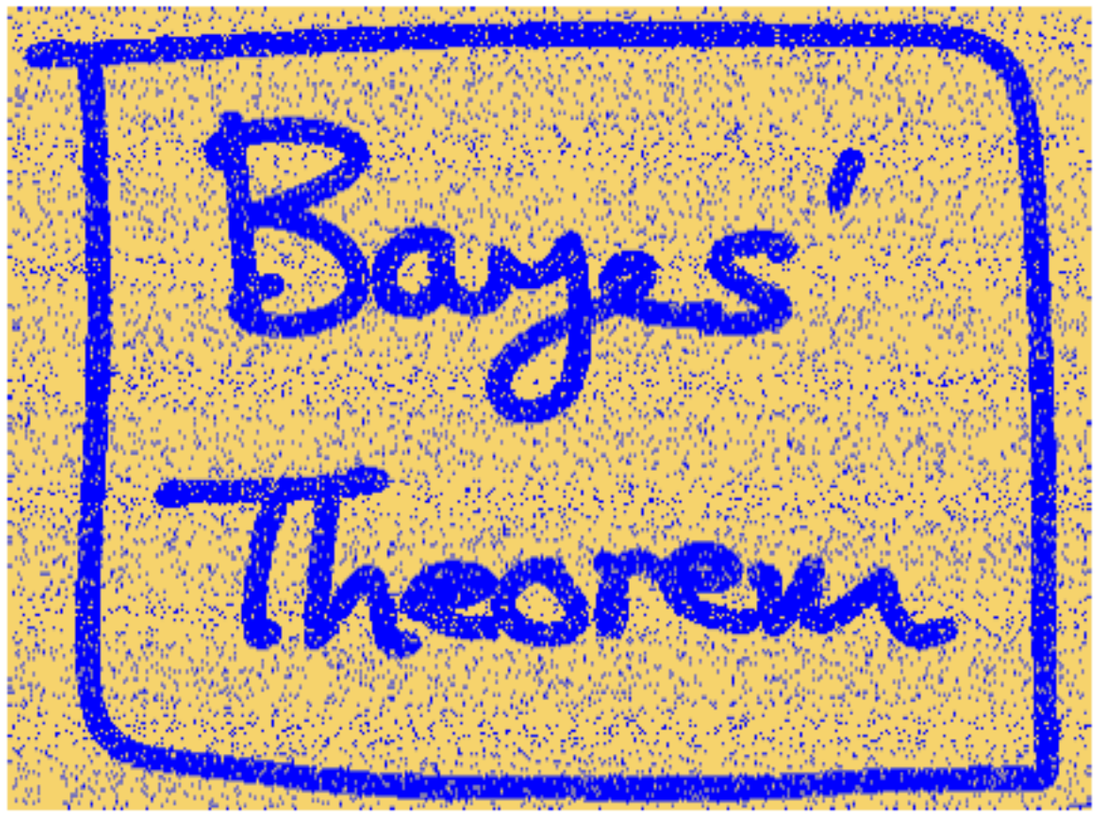
energy function

- An example: **image de-noising**

  - Observed noise binary pixels (by random flipping): $y_i \in \{+1, -1\}$

  - Underlying clean image: $x_i \in \{+1, -1\}$



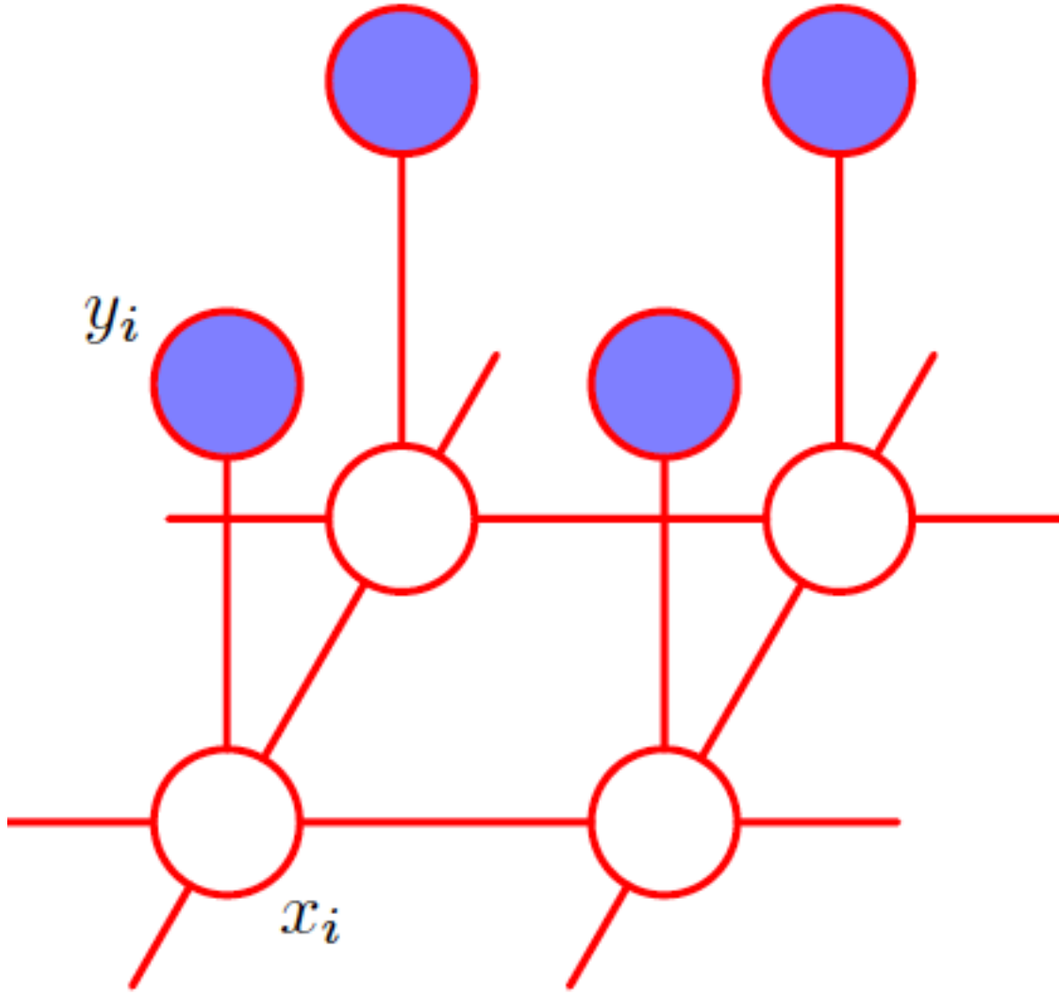


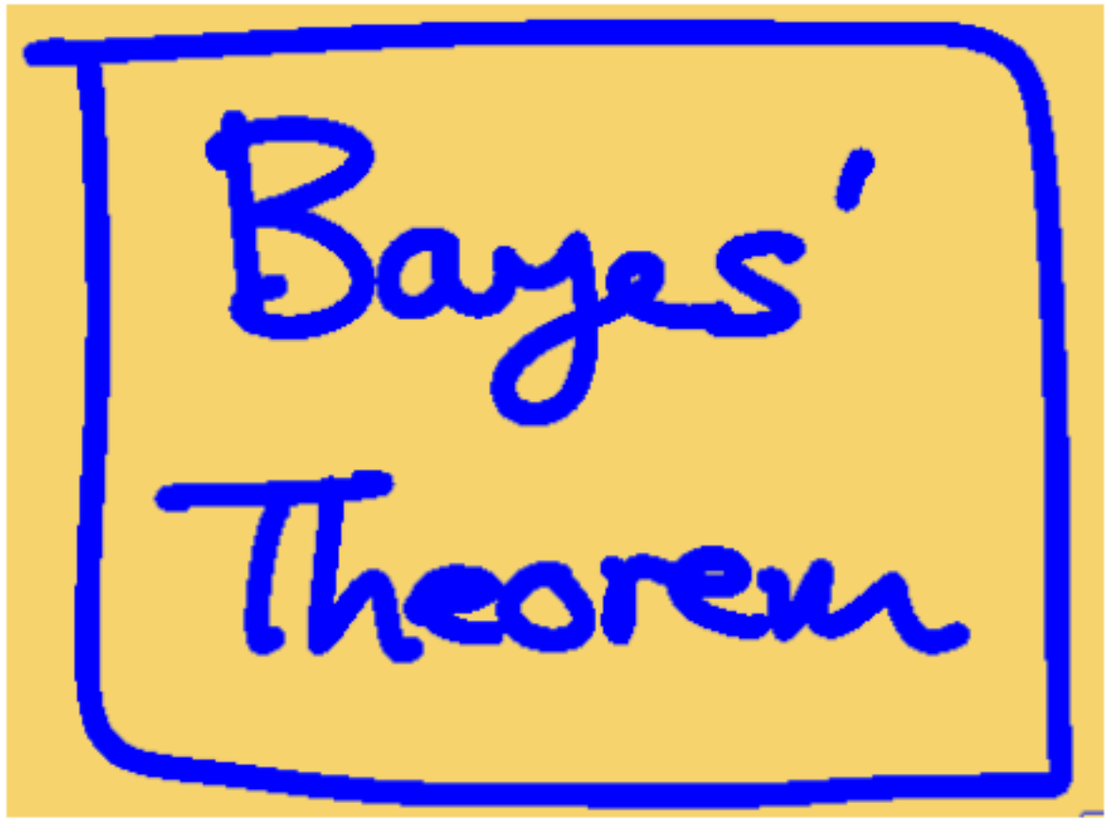
- Modeling the noisy image (Ising model)

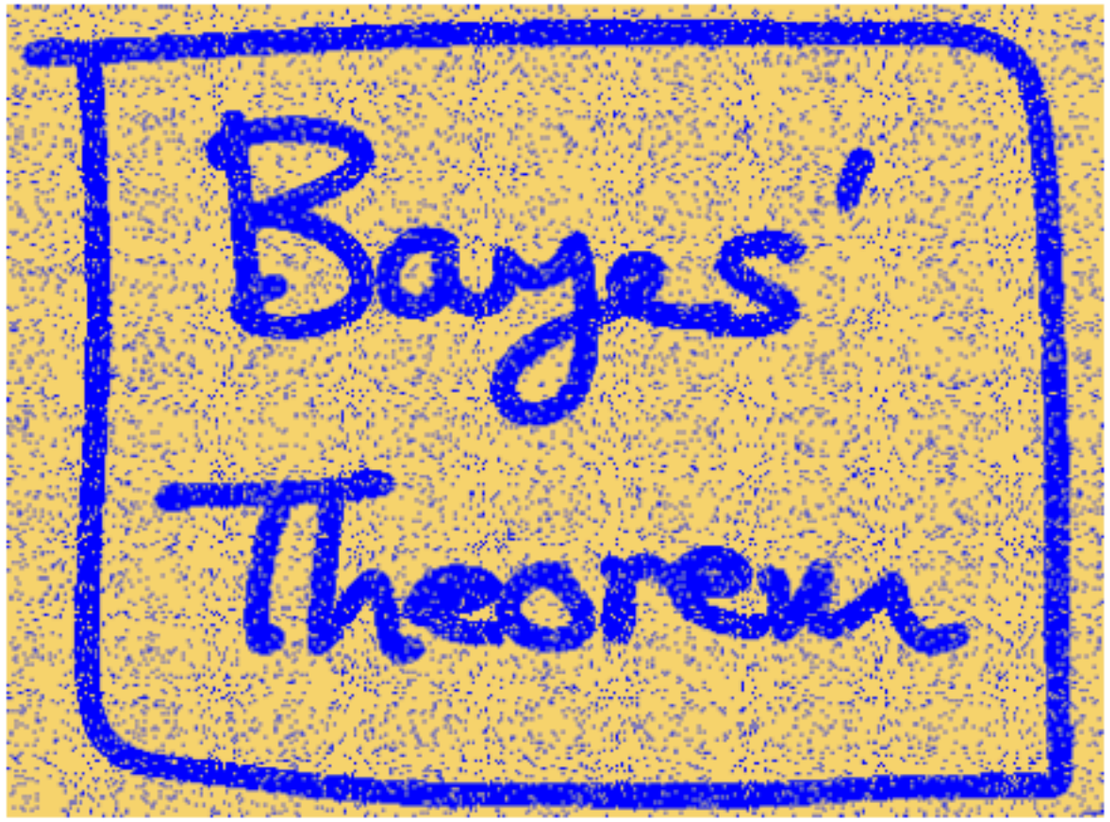
$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

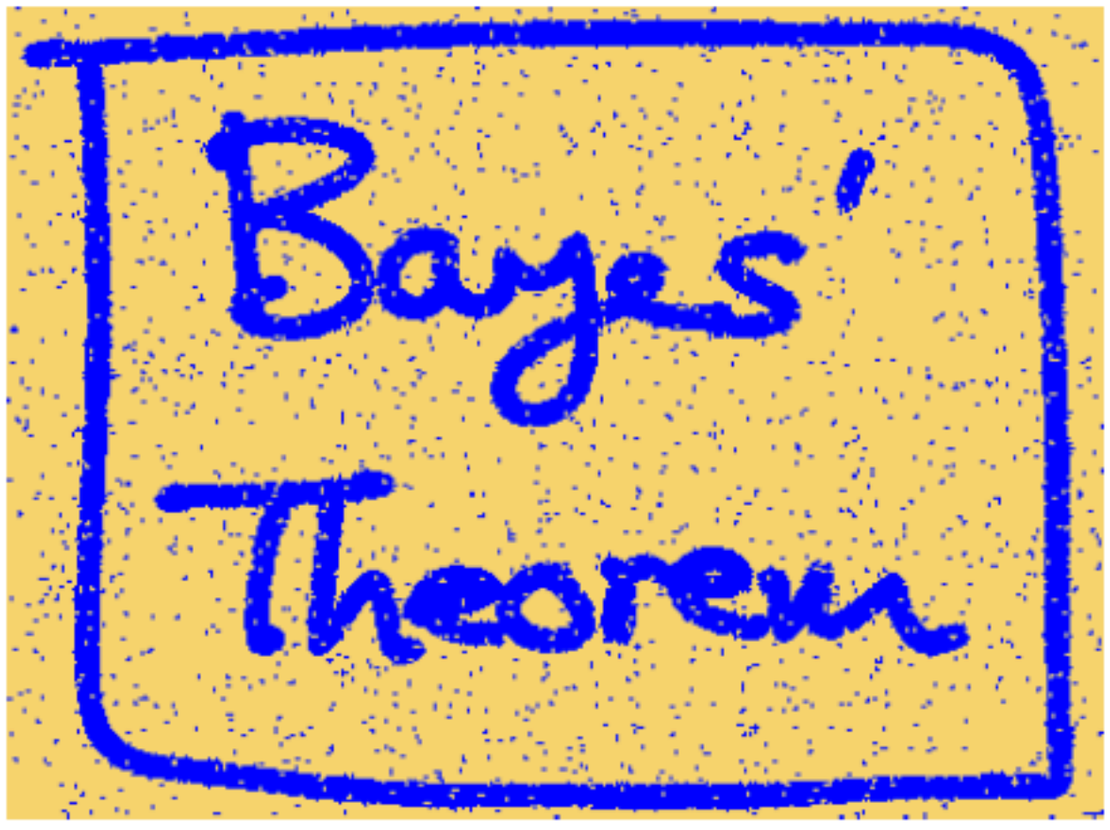
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

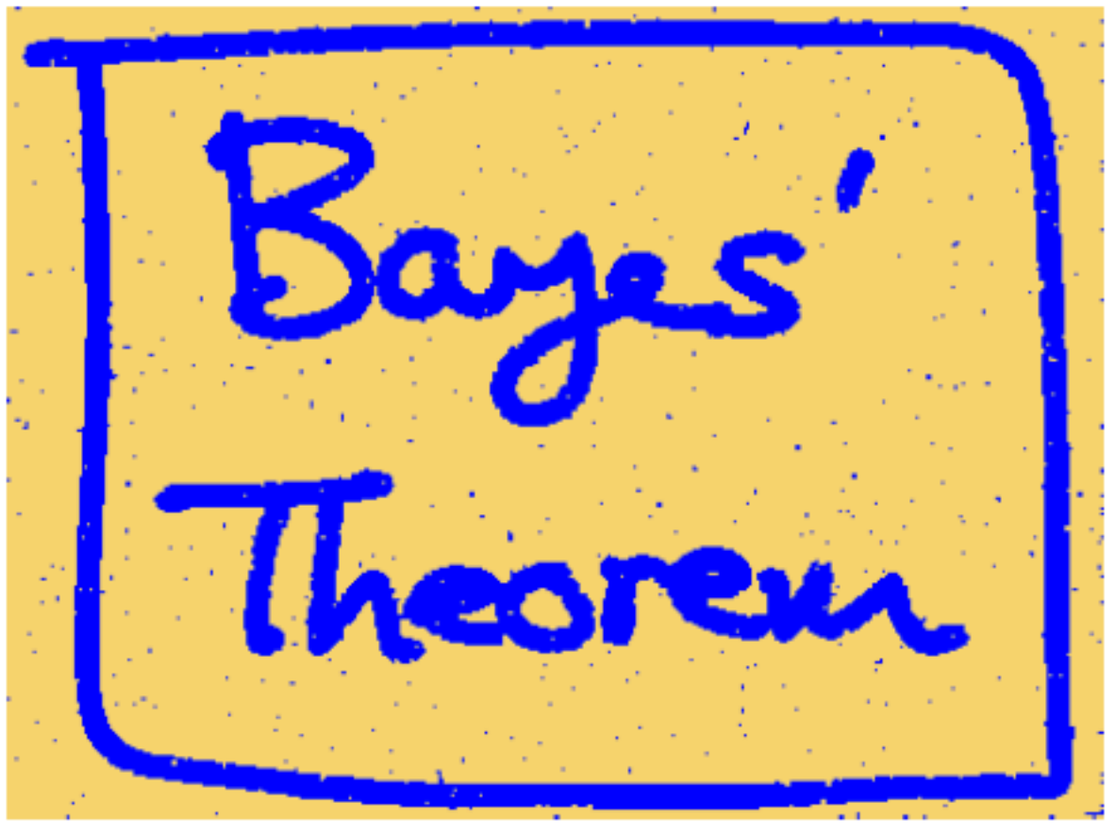
$$\max \log p(\boldsymbol{x}|\boldsymbol{y})$$







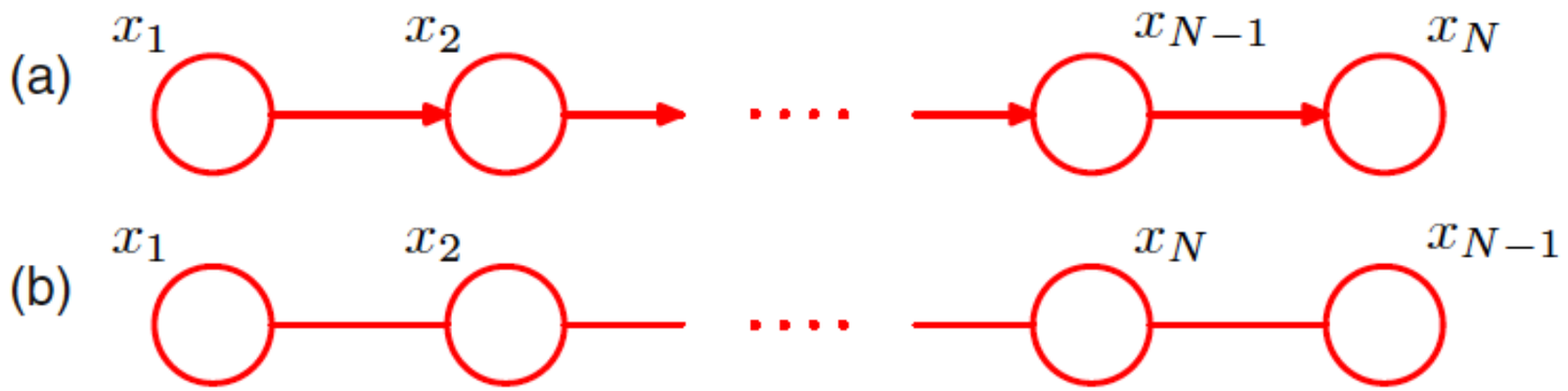




Iterative conditional modes

graph cut

- Relation between directed and undirected graphs
  - Chain graph



$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1})$$

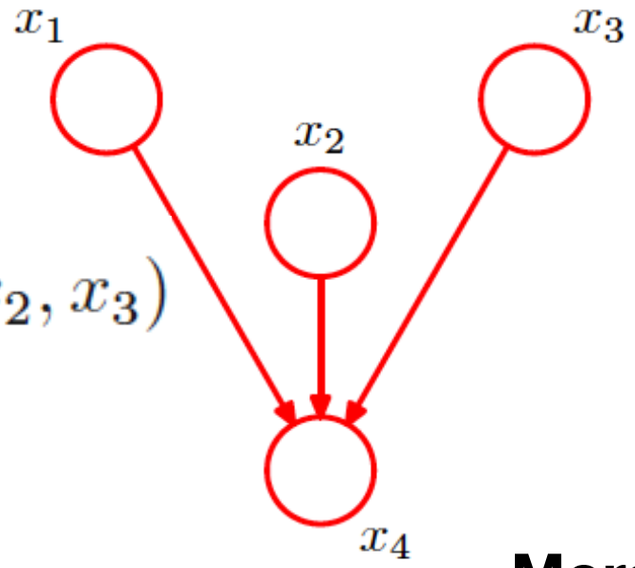
$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\cdots\psi_{N-1,N}(x_{N-1},x_N)$$

$$\begin{aligned}
\psi_{1,2}(x_1,x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2,x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1},x_N) &= p(x_N|x_{N-1})
\end{aligned}$$

$$Z = 1$$

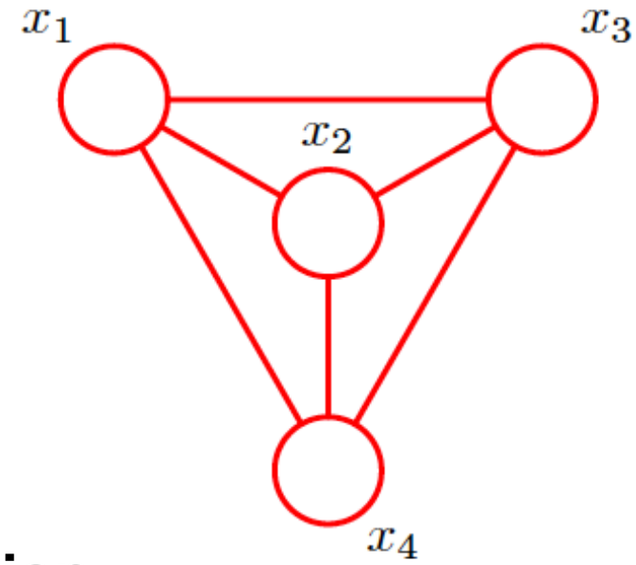
- Head-to-head graph

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$



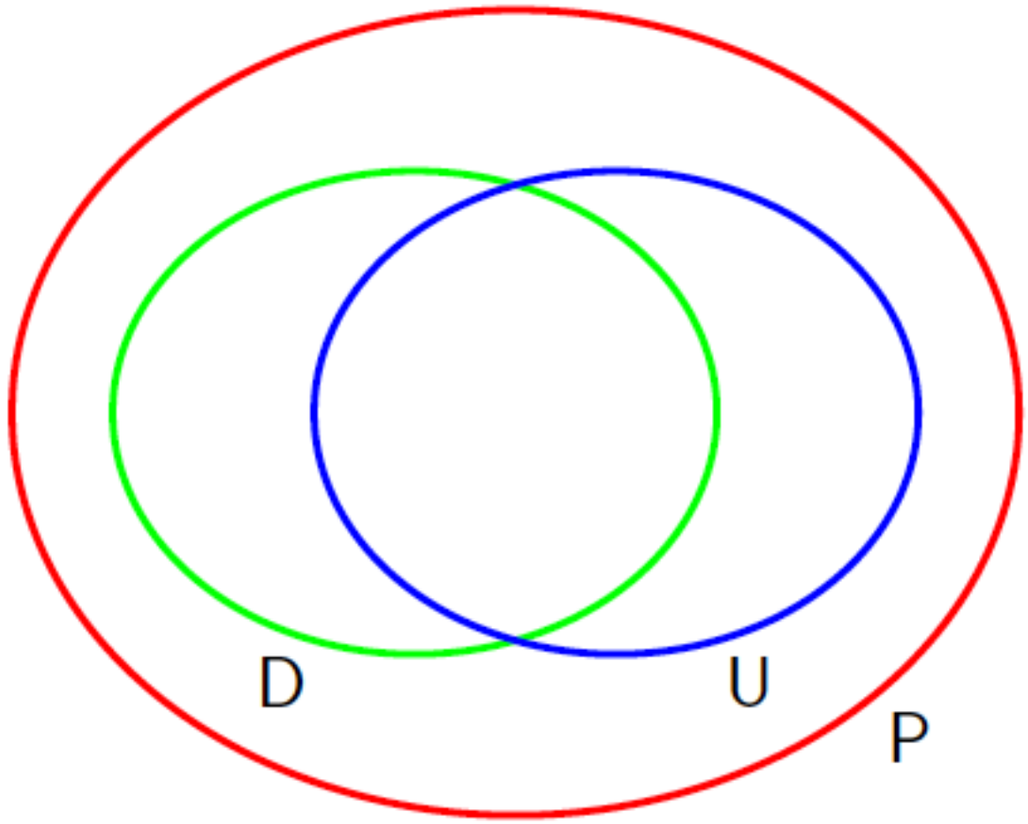
(a)

Moralization



(b)

- Not all directed and undirected graphs can be translated with each other.
  Could you give any examples?



Inference in GMs

- Compute posterior distribution over some nodes given observed variables

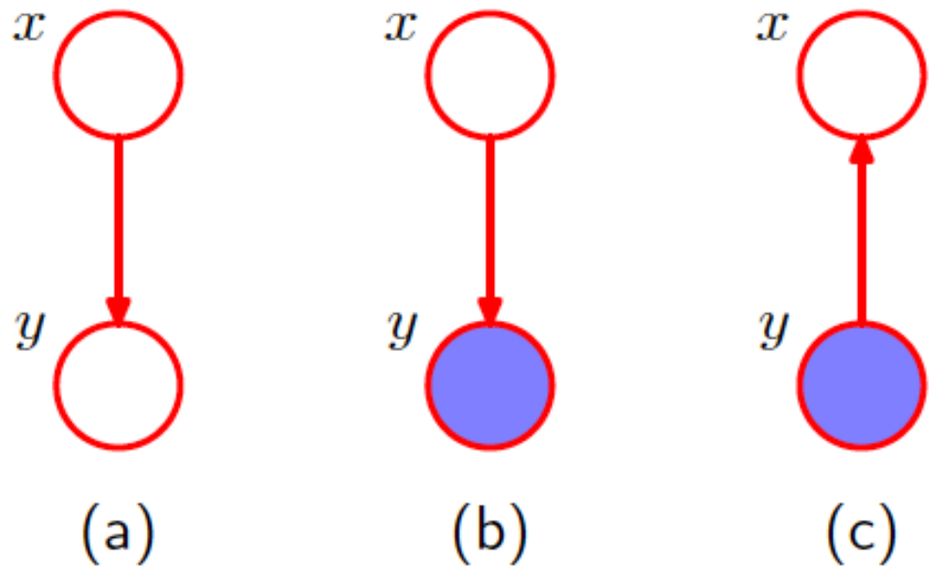
- **Exploit** the graphical structure of the GMs to find efficient inference algorithms

- Propagation of local messages around the graph

- The simplest inference problem: Bayes Theorem

$$p(y) = \sum_{x'} p(y|x')p(x')$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



Inference on a chain

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

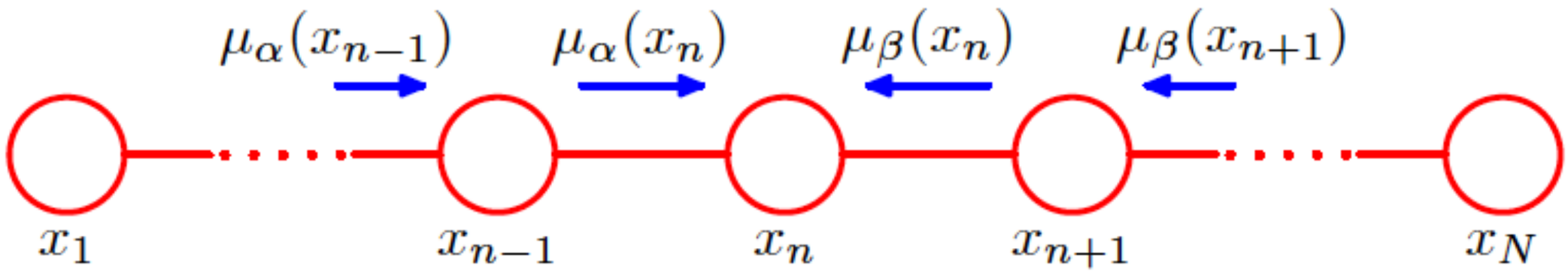
- **Finding the marginal distribution: multiplying two messages**

$$p(x_n) = \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x})$$

$$p(x_n) = \frac{1}{Z}$$

$$\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots\right]}_{\mu_\alpha(x_n)}$$

$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)}.$$



$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n)$$

- Recursively evaluate the messages

$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \left[\sum_{x_{n-2}} \cdots\right]$$
$$= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_\alpha(x_{n-1}).$$

$$\mu_\beta(x_n) = \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n) \left[\sum_{x_{n+2}} \cdots\right]$$
$$= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n)\mu_\beta(x_{n+1}).$$

$$\mu_\alpha(x_2) = \sum_{x_1} \psi_{1,2}(x_1, x_2)$$

- What about the joint distribution?

$$p(x_{n-1}, x_n) = \frac{1}{Z}\mu_\alpha(x_{n-1})\psi_{n-1,n}(x_{n-1}, x_n)\mu_\beta(x_n)$$

- Optional readings (See Bishop (2012) Chap. 8.4.4)
  - Factor graph
  - Sum-product algorithm
  - Max-sum algorithm

Mixture Models and EM Algorithm

Mixture Models

- Latent variable model for modeling generative process of the data

- Simple but powerful representative

  - **Mixture of Gaussians** (MoG) for clustering

- **Expectation-Maximization** Algorithm (EM)

  - Learning mixture models

Recall K-means

- The most popular clustering method
- Clustering with hard cluster assignment
- The idea: consider a cluster as a group of data points whose inter-point distances are small compared with the distances to the points outside the cluster.

- Formalizing the simple idea:

number of clusters

cluster indicator, 0-1

Minimize: $$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

cluster centroid

- Alternatively updating the two variables: cluster indicators and cluster centroids

  - E step (expectation): Update cluster indicators

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

  - M step (maximization): Update cluster centroids

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

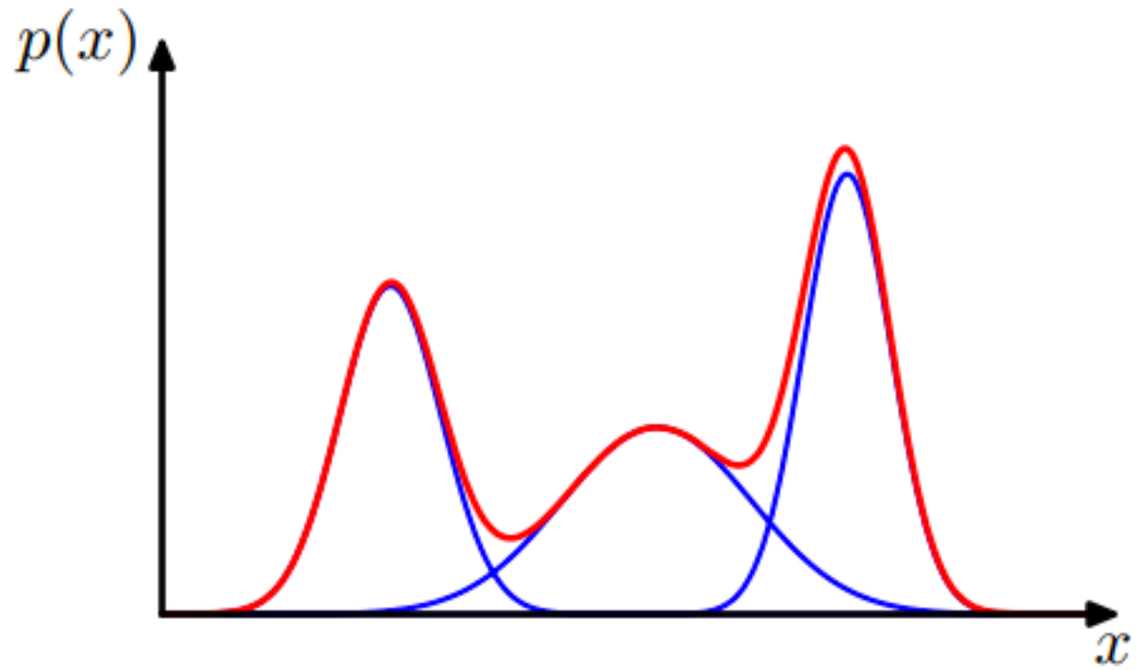
mean of data points assigned to cluster k

- The two steps **converge to local minima** since each iteration lowers the objective function.

Mixture of Gaussian

- The most famous latent variable mode
- Clustering with soft cluster assignment
- A nice model to motivate EM algorithm

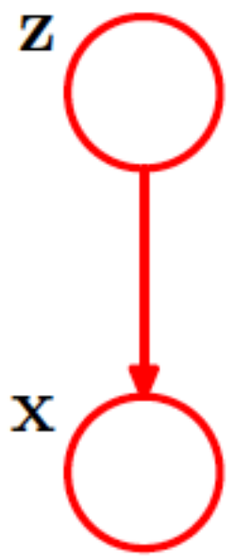
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



- Introduce K-dimensional binary variable z with one-hot representation

$$z_k \in \{0,1\} \text{ and } \sum_k z_k = 1$$

$$p(z_k = 1) = \pi_k \qquad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$



$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

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

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

- The posterior probability of z_k given x: the cluster assignment

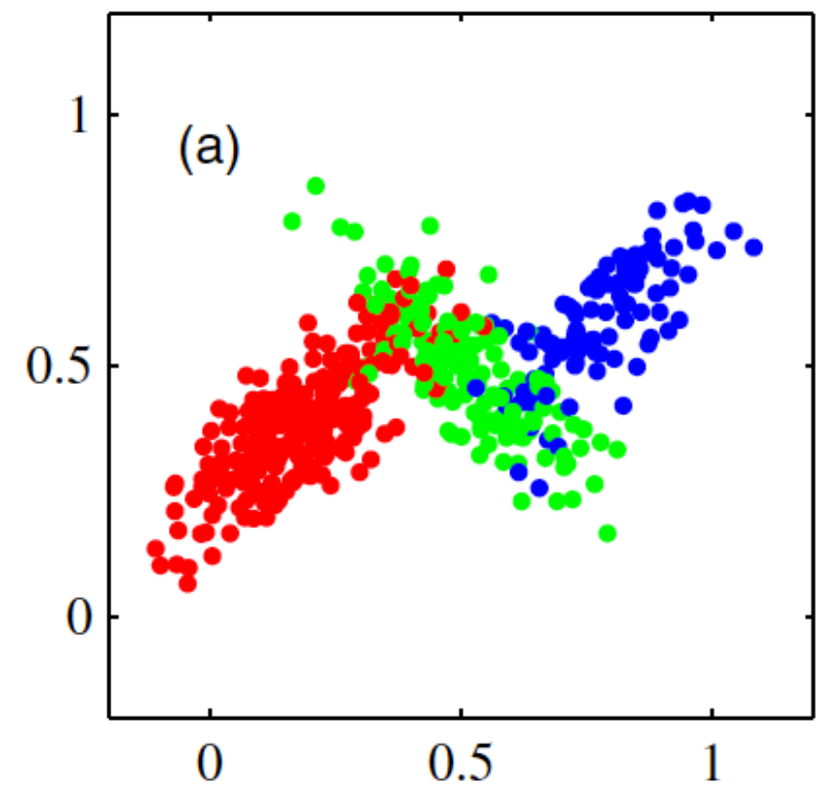
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

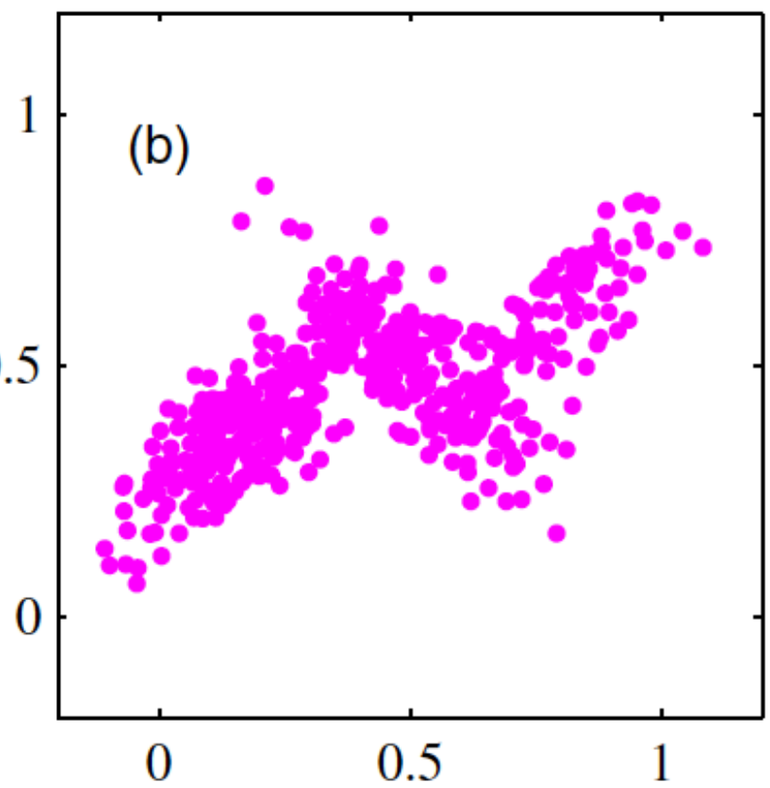
- Modeling the entire generative process of the observed data

  - Ancestral sampling to generate: first sample z_k from prior probability $p(z_k = 1) = \pi_k$
    then sample x according to $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

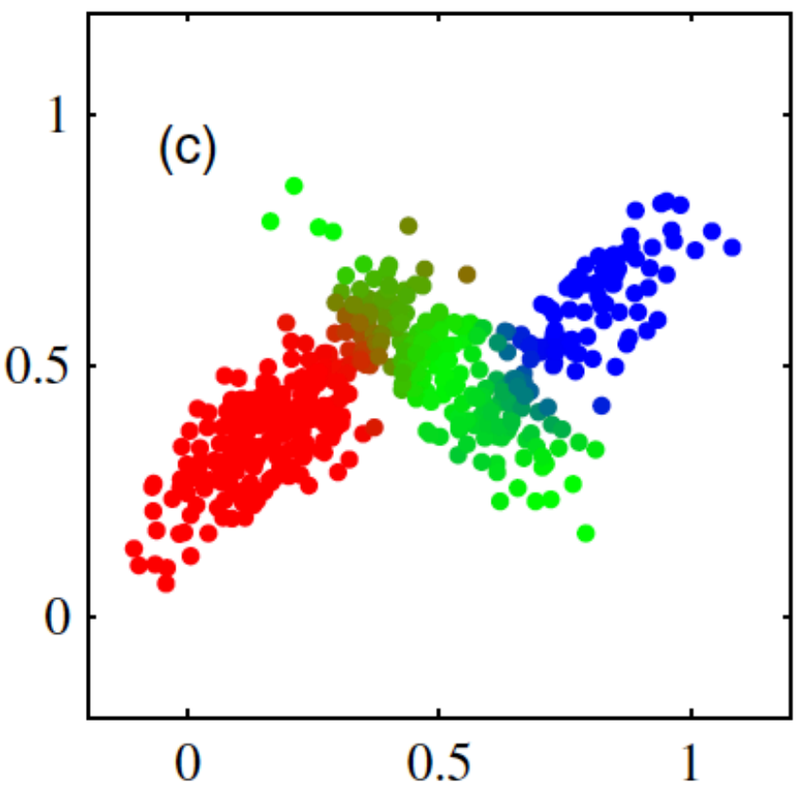
Generating data



$p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$



$p(\mathbf{x})$



$\gamma(z_{nk})$

- Maximum likelihood of Gaussian mixture models

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

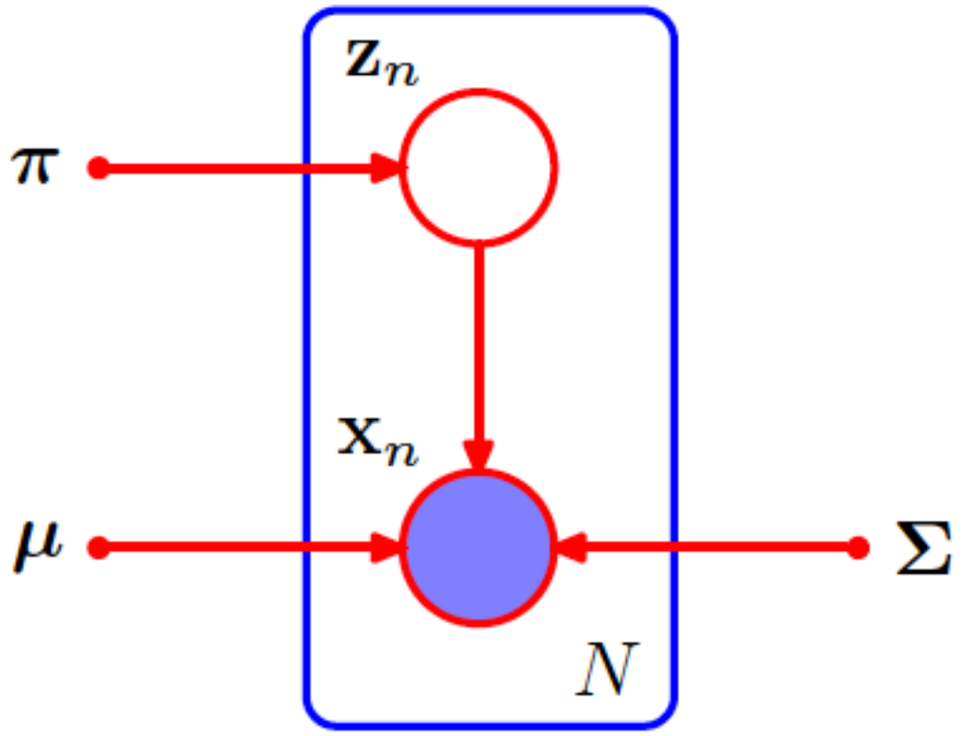
- The issue of maximizing log likelihood of MoGs

- Singularity: “collapse” to single data point

$$\boldsymbol{\mu}_j = \mathbf{x}_n$$

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

$\sigma_j \rightarrow 0 \longrightarrow$ infinite log likelihood



EM for Gaussian Mixtures

- Set the derivative of log likelihood w.r.t. component mean to be zero,

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

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

effective number of samples assigned to cluster k

- Set the derivative of log likelihood w.r.t. component covariance to be zero,

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

- For component mixing proportion,

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

$$\downarrow$$

$$\pi_k = \frac{N_k}{N}$$

- No close-formed solution, but **we can alternatively update the three groups of parameters with some initialization**
  - This is the instance of the **EM** algorithm
  - **E step**: evaluate the posterior probability, i.e. responsibility
  - **M step**: re-estimate the means, covariances, and mixing coefficients
  - The two steps can guarantee convergence to local maxima. Why? (Exercise)
  - In practice, we can use K-means results as the initialization of means, empirical covariance as initialization of covariances, proportion of data points as mixing coefficients.

Alg. EM for Gaussian Mixture

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

3. **M step**. Re-estimate the parameters using the current responsibilities

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

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

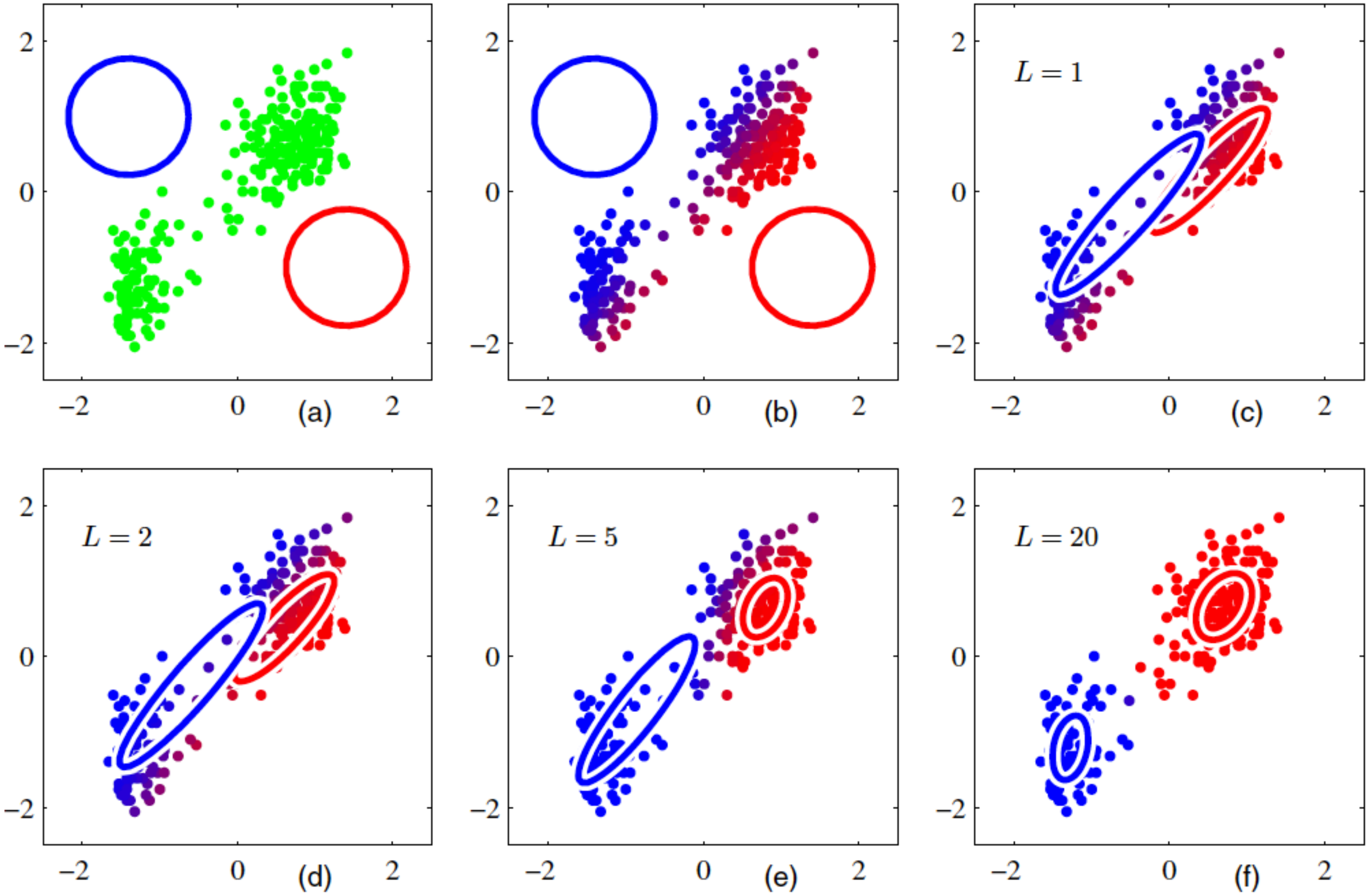
where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}).$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.


2
0
−2
−2
0
2
(a)
(b)
(c)
(d)
(e)
(f)
$L = 1$
$L = 2$
$L = 5$
$L = 20$

A General View over EM

- EM algorithm is particularly suitable for finding maximum likelihood solution for latent variable models.

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

- Expectation step: find the expectation of complete data likelihood w.r.t. posterior of latent variables

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

- Maximization step: maximize the expectation to obtain the new parameter

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$ **tractable**

- Derive maximum likelihood solution of Gaussian mixture under this general view of EM (Exercise)

- EM actually relies on some assumptions

The direction optimization of $p(\boldsymbol{X}|\boldsymbol{\theta})$ is hard, but that optimization of the complete-data likelihood function $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is significantly easier.

- Another perspective on EM

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

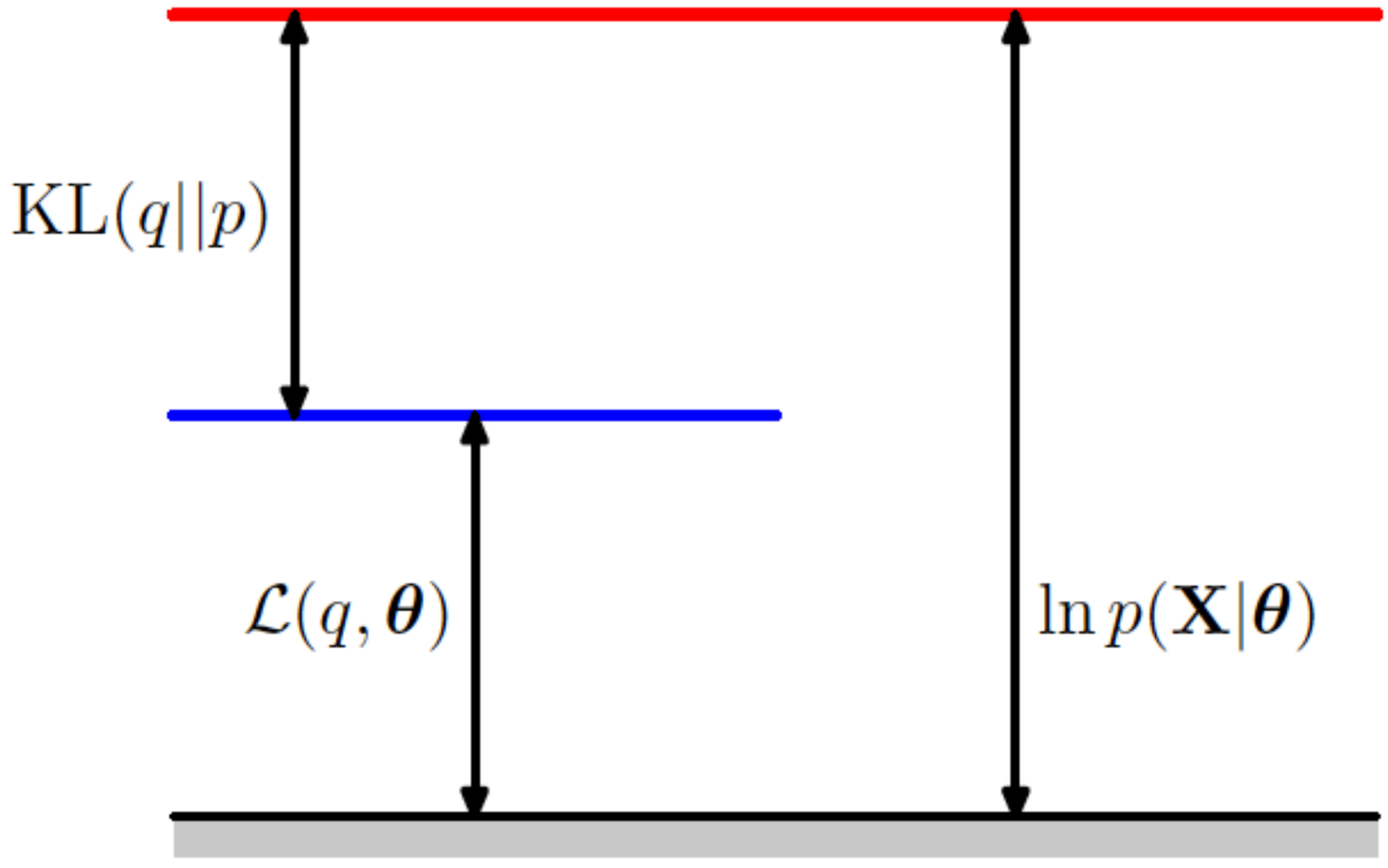
A functional w.r.t. q(Z)

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

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

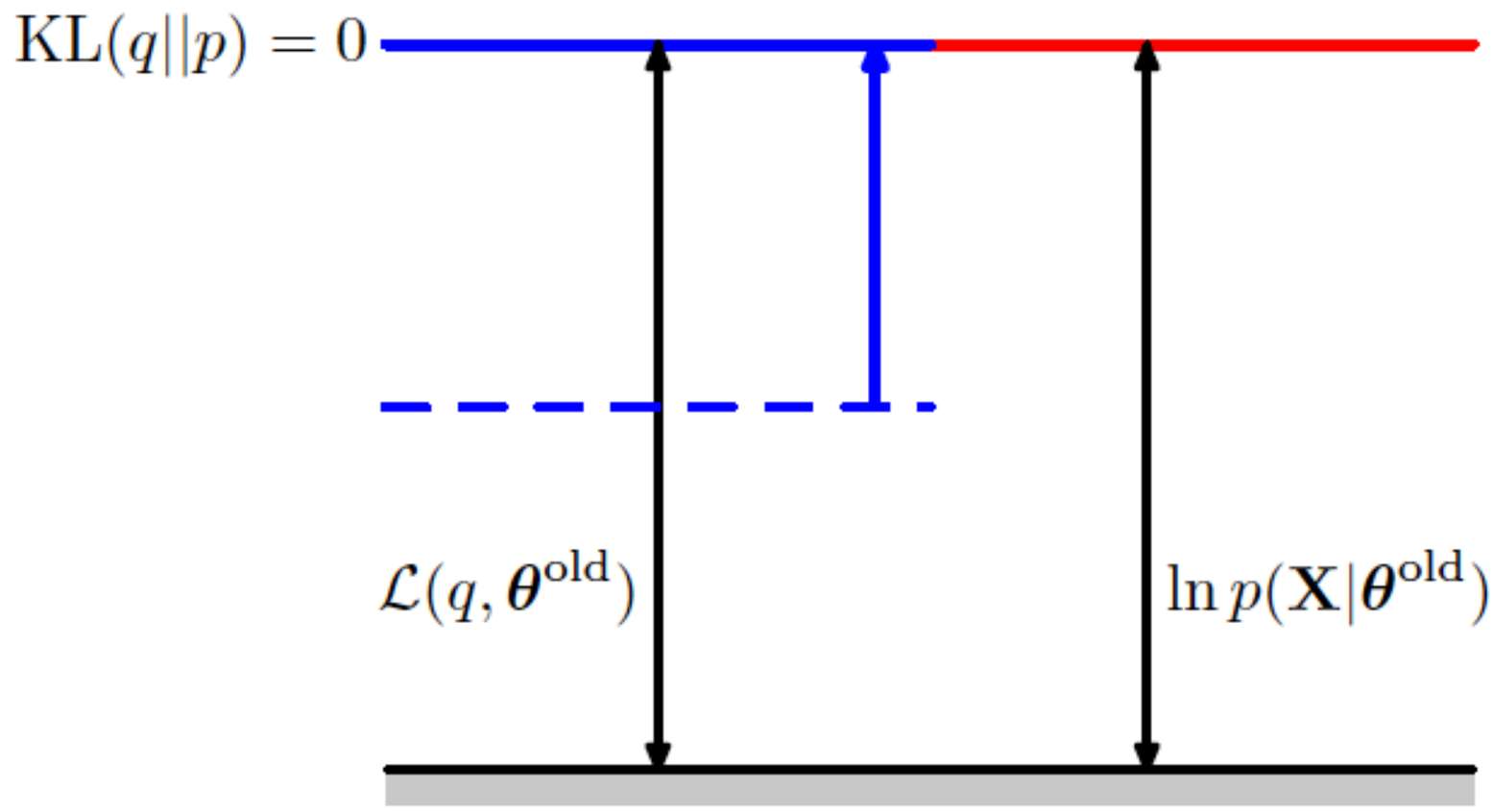
$$\mathcal{L}(q, \boldsymbol{\theta}) \leqslant \ln p(\mathbf{X}|\boldsymbol{\theta})$$

$\mathrm{KL}(q\|p) \geqslant 0$, with equality if, and only if, $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$


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

- E step

The lower bound $L(q, \boldsymbol{\theta}^{\text{old}})$ is maximized w.r.t. $q(\boldsymbol{Z})$ while fixing $\boldsymbol{\theta}^{\text{old}}$, where the maximization is achieved when $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$.

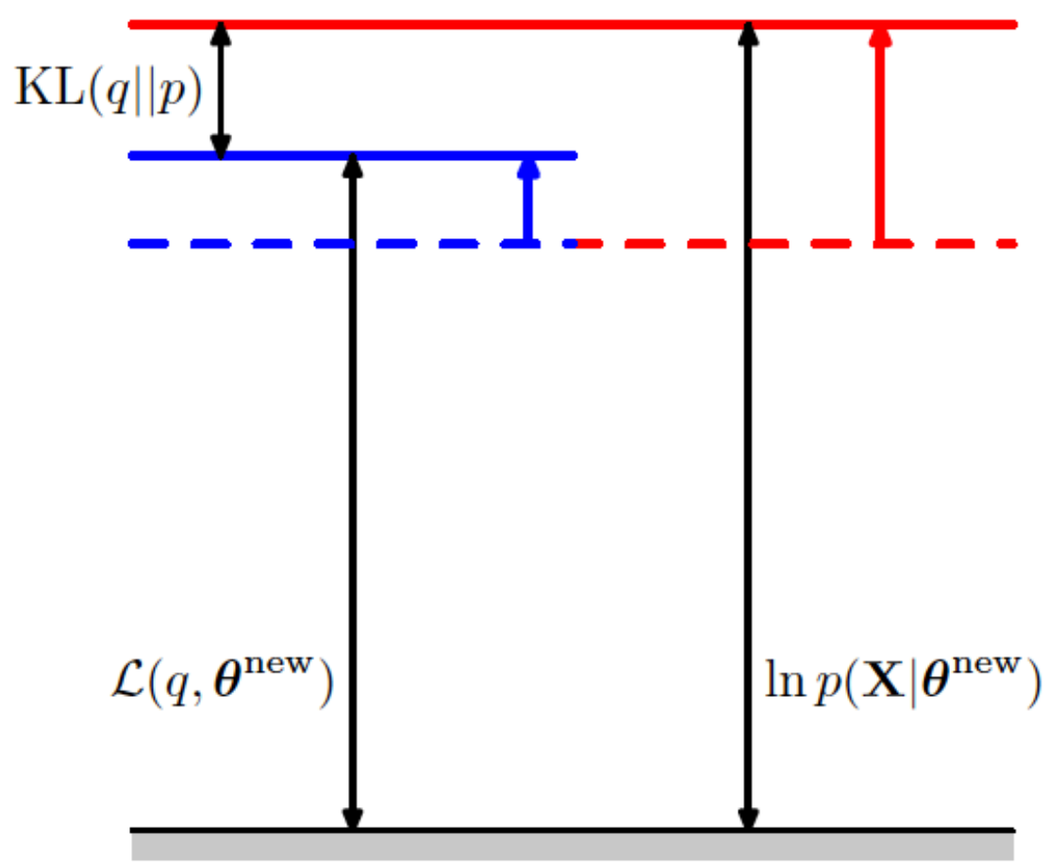


- M step

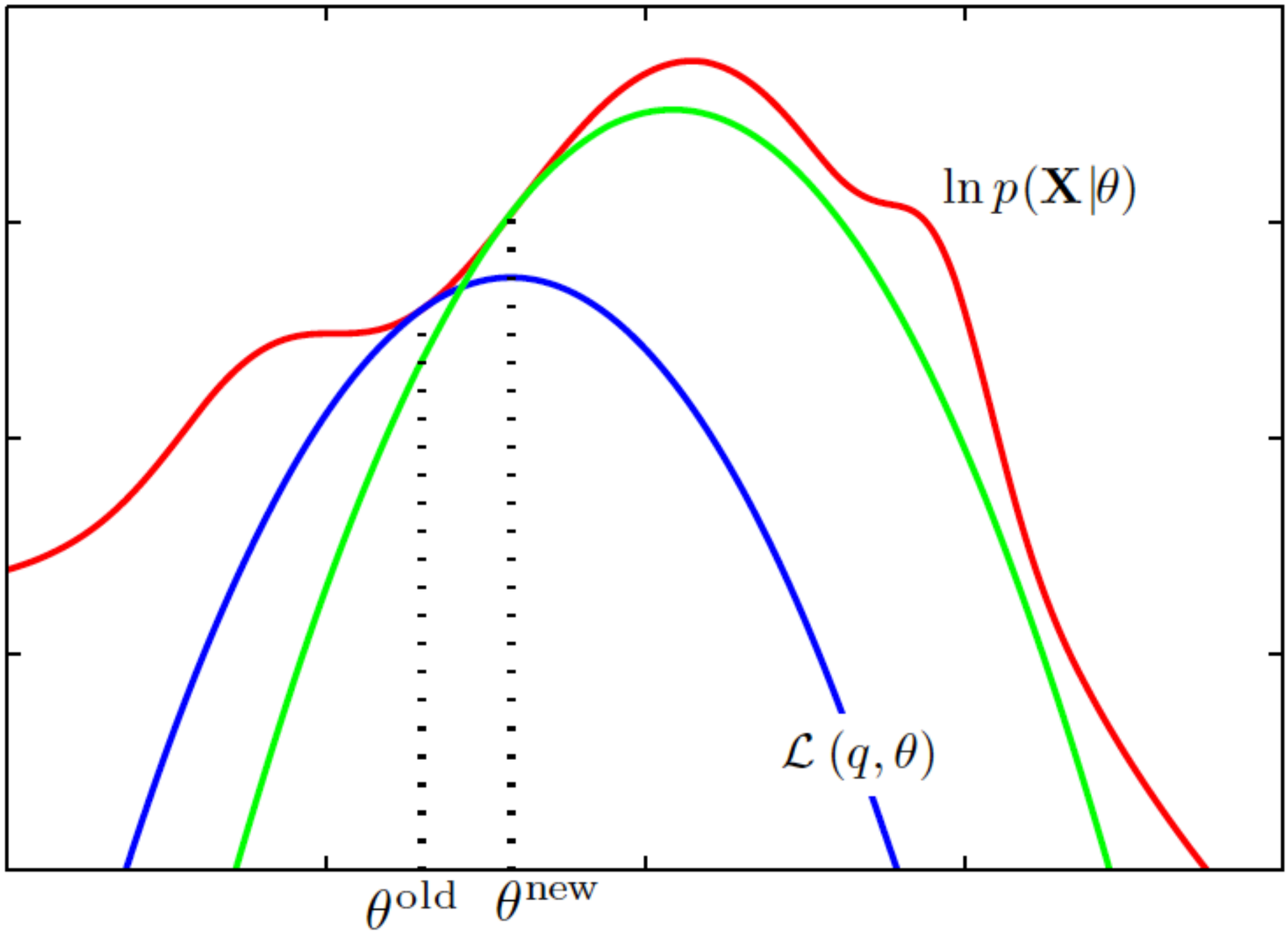
The distribution $q(\boldsymbol{Z})$ is held fixed and the lower bound $L(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$.

After E step

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{const}\end{aligned}$$



The EM


$\ln p(\mathbf{X}|\theta)$
$\mathcal{L}(q,\theta)$
θ^{old}
θ^{new}

- Two extension of EM
  - What if M-step does not have a close-form solution?
    - Perform some nonlinear optimization steps to increase the value of expected complete log likelihood
  - Stochastic EM (Neal and Hinton 1999)
    - Stochastically update one data point to reduce computational complexity

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \right) \left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)$$

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})$$

Prof. Radford Neal

Prof. Geoffrey Hinton

Exercise

- Show why EM algorithm can converge to local maxima for Gaussian mixtures

- Implement EM for Gaussian mixture.