Excurse: Conjugacy

Assume we have a binary random variable $x \in \{0, 1\}$ and we are given a parameter μ, $0 \leq \mu \leq 1$ so that

$$p(x = 1 \mid \mu) = \mu \qquad\qquad p(x = 0 \mid \mu) = 1 - \mu$$

together this gives: $p(x \mid \mu) = \mu^x (1 - \mu)^{1-x}$ **"Bernoulli distribution"**

Now we have a set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of independent binary events. It has the probability:

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

$$= \prod_{x_n=1} \mu^{x_n} (1 - \mu)^{1-x_n} \prod_{x_n=0} \mu^{x_n} (1 - \mu)^{1-x_n}$$

Excurse: Conjugacy

which results in: $p(\mathcal{D} \mid \mu) = \mu^m (1-\mu)^{N-m}$

where m is the number of events where $x_n = 1$.

There exist $\left(\begin{array}{c} N \\ m \end{array}\right)$ possibilities for $\mathcal{D}$, so

"Binomial distribution"

$$p(m \mid N, \mu) = \left(\begin{array}{c} N \\ m \end{array}\right) \mu^m (1-\mu)^{N-m}$$

is the probability that there are m positive events in a set (sequence) of N, where

$$\left(\begin{array}{c} N \\ m \end{array}\right) = \frac{N!}{(n-m)!m!}$$

Maximum Likelihood

To find an optimal parameter μ we can use MLE:

$$\log p(\mathcal{D} \mid \mu) = \sum_{n=1}^{N} \log p(x_n \mid \mu) = \sum_{n=1}^{N} (x_n \log \mu + (1 - x_n) \log(1 - \mu)$$

Maximum Likelihood

To find an optimal parameter μ we can use MLE:

$$\log p(\mathcal{D} \mid \mu) = \sum_{n=1}^{N} \log p(x_n \mid \mu) = \sum_{n=1}^{N} (x_n \log \mu + (1 - x_n) \log(1 - \mu)$$

and we obtain: $\mu = \frac{1}{N} \sum_{n=1}^{N} x_n$ or, equivalently: $\mu = \frac{m}{N}$

Suppose we observe "1" in three trials,
i.e. $x_1 = x_2 = x_3 = 1$. It follows $\mu_{ML} = 1$.

This is an example of extreme overfitting due to the maximum likelihood approach!

Bayesian Inference

To address the problem of overfitting, we define a prior probability for the parameter μ and compute:

$$p(\mu \mid m, N) = Z_p^{-1} p(m \mid \mu, N) p(\mu)$$

Posterior — Normalizer — Likelihood — Prior

Goal: Find a prior distribution so that the posterior has the same functional form as the prior!

Then, the posterior can be used as a new prior when new data is observed.

Such a prior is called **conjugate** to the likelihood.

A Conjugate Prior for the Binomial Dist.

Observation: if prior is proportional to powers of μ $1 - \mu$ then the posterior will be so, too.

A Conjugate Prior for the Binomial Dist.

Observation: if prior is proportional to powers of μ $1-\mu$ then the posterior will be so, too.

Thus, the conjugate prior for the binomial distribution is the **beta-distribution**:

$$p(\mu \mid a, b) = Z_\beta^{-1} \mu^{a-1} (1-\mu)^{b-1} \quad a > 0, b > 0$$

$$Z_\beta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Here, a and b can be interpreted as the assumed prior number of positive and negative events

Obtaining the Posterior

Now we can use the prior and the likelihood:

$$p(\mu \mid m, N, a, b) \propto p(m \mid \mu, N) p(\mu) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

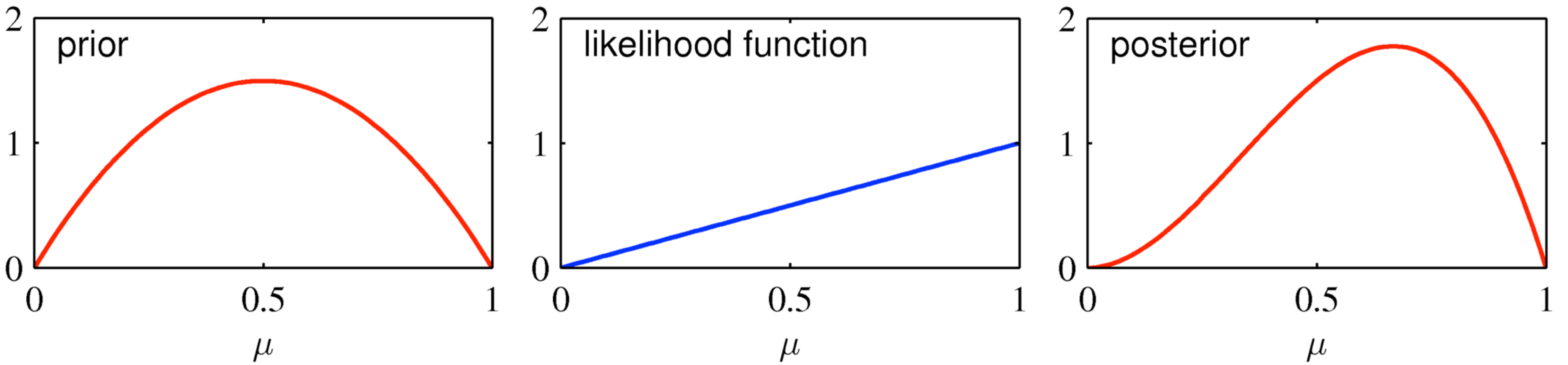
$$l = N - m$$

This gives another beta-distribution:

$$p(\mu \mid m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1}$$

where the **effective number of observations** for $x = 1$ and $x = 0$ has been increased by m and l

A Simple Example



$p(\mu) = \text{Beta}(\mu \mid a = 2, b = 2)$ $\quad p(m \mid \mu, N) = \text{Bin}(m = 1 \mid N = 1, \mu)$ $\quad p(\mu) = \text{Beta}(\mu \mid a = 3, b = 2)$

- Consider the example m=1, N=1
- The prior is defined by a=2, b=2
- Using Bayesian inference we obtain the posterior that is shifted towards μ =1
- Overfitting can be avoided!

The Same For Multinomial Variables

In the case of K possible states of x we have

$$\mathbf{x} = (x_1, \ldots, x_K) \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \quad \mu_k \geq 0 \quad \sum_{k=1}^{K} \mu_k = 1$$

The likelihood is then a **multinomial** distribution:

$$\mathrm{Mult}(m_1, \ldots, m_K \mid \boldsymbol{\mu}, N) = \binom{N}{m_1, \ldots, m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

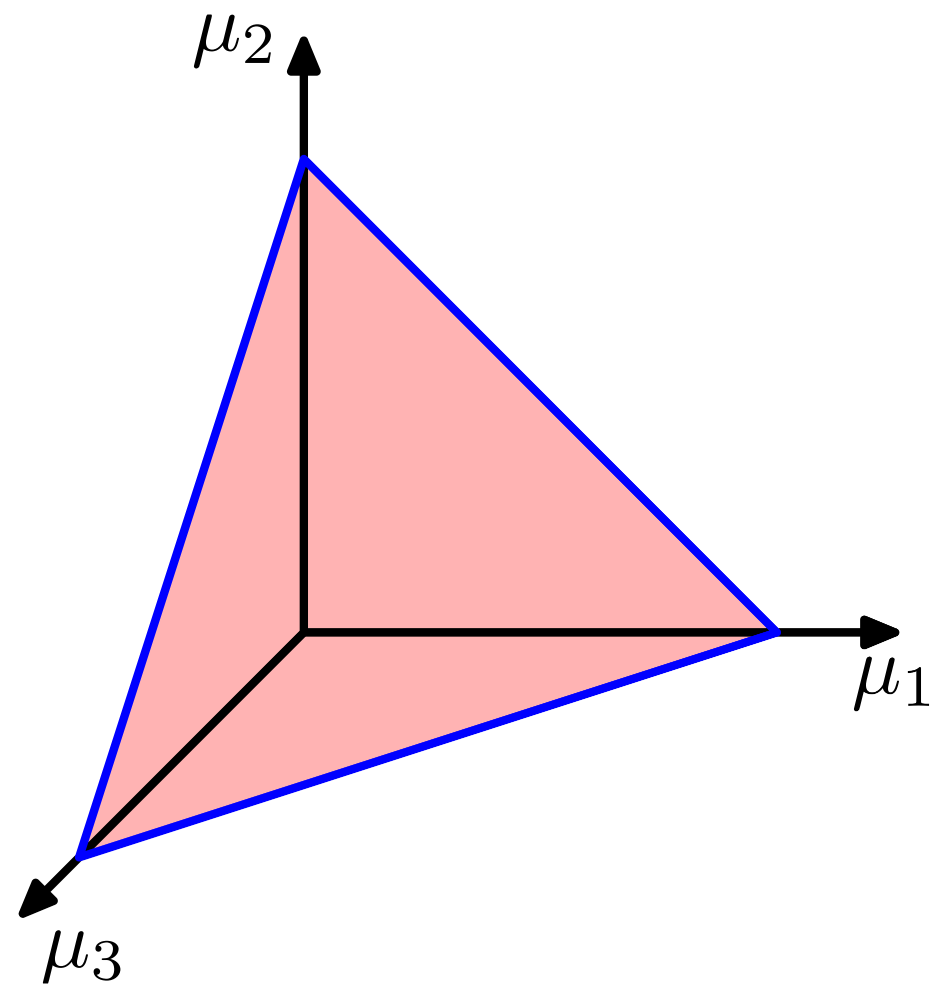
The conjugate prior of that is the **Dirichlet** distribution:

$$\mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

The Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k \qquad 0 \leq \mu_k \leq 1 \qquad \sum_{k=1}^{K} \mu_k = 1$$



- Example with three variables
- The distribution is confined to a simplex (in this case a triangle)

Sampling Methods II

Gibbs Sampling

- Initialize $\{z_i : i = 1, \ldots, M\}$
- For $\tau = 1, \ldots, T$
  - Sample $z_1^{(\tau+1)} \sim p(z_1 \mid z_2^{(\tau)}, \ldots, z_M^{(\tau)})$
  - Sample $z_2^{(\tau+1)} \sim p(z_2 \mid z_1^{(\tau+1)}, \ldots, z_M^{(\tau)})$
  - ...
  - Sample $z_M^{(\tau+1)} \sim p(z_M \mid z_1^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$

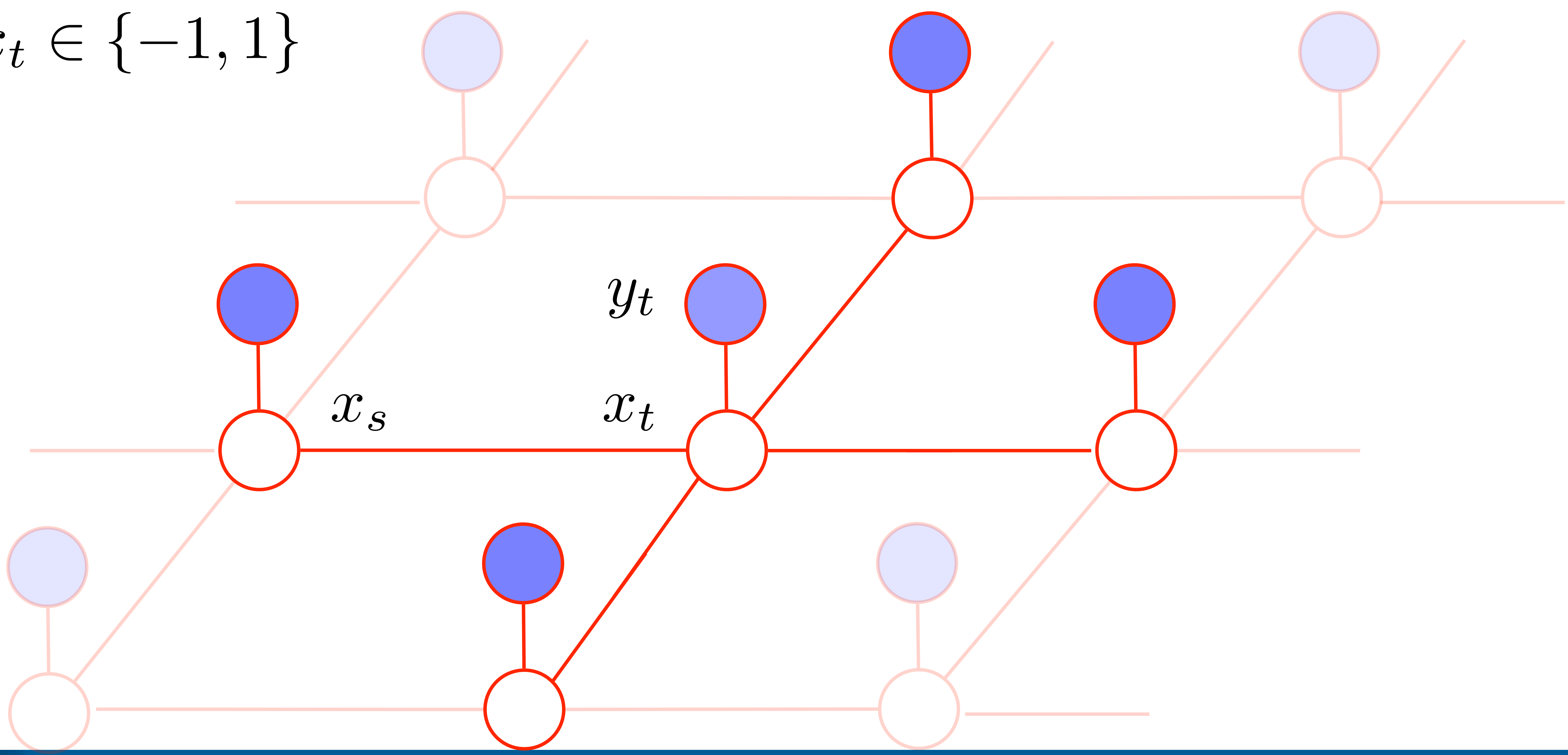
Idea: sample from the full conditional

This can be obtained, e.g. from the Markov blanket in graphical models.

Gibbs Sampling: Example

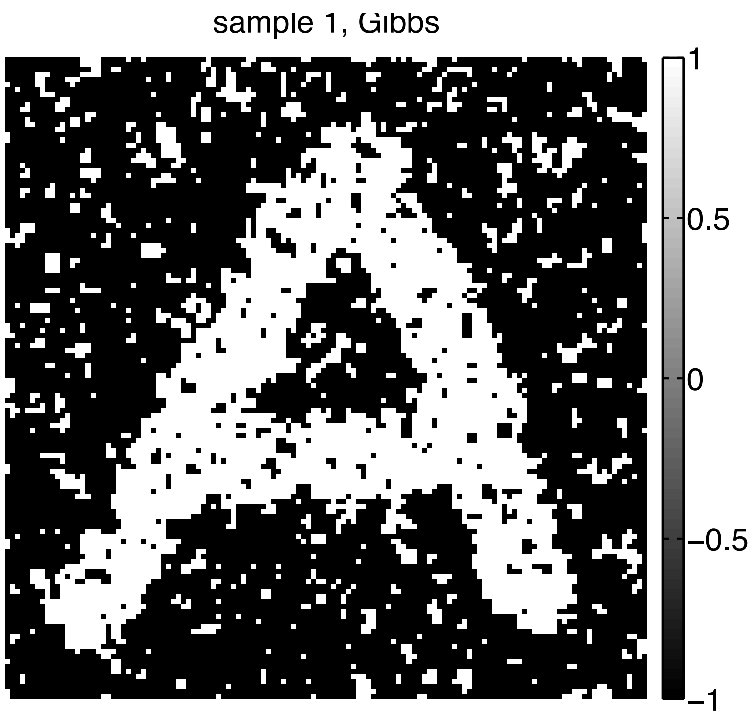
- Use an MRF on a binary image with edge potentials $\psi(x_s, x_t) = \exp(J x_s x_t)$ ("Ising model") and node potentials $\psi(x_t) = \mathcal{N}(y_t \mid x_t, \sigma^2)$

$x_t \in \{-1, 1\}$

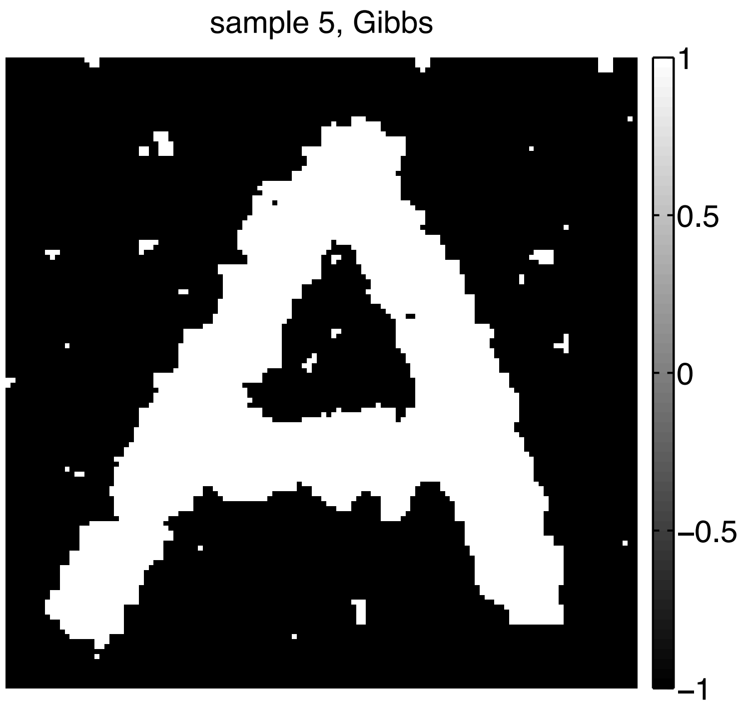


Gibbs Sampling: Example

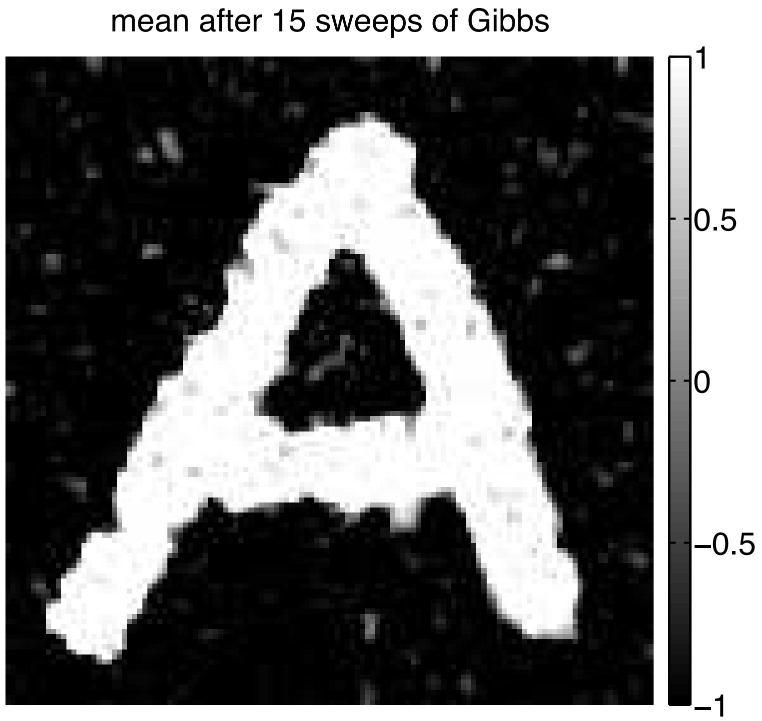
- Use an MRF on a binary image with edge potentials $\psi(x_s, x_t) = \exp(J x_s x_t)$ ("Ising model") and node potentials $\psi(x_t) = \mathcal{N}(y_t \mid x_t, \sigma^2)$
- Sample each pixel in turn



After 1 sample



After 5 samples

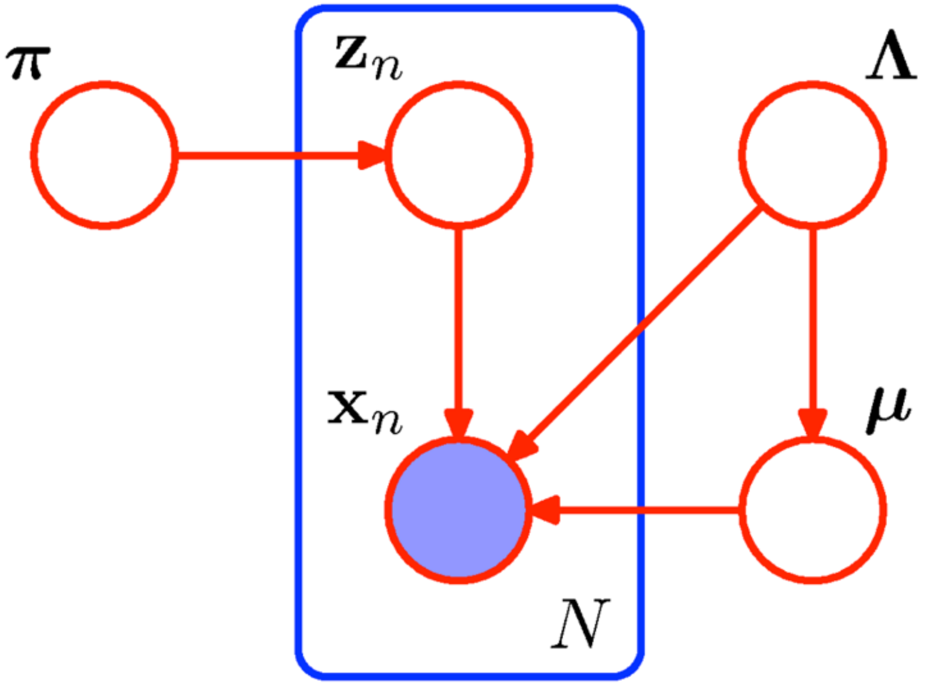


Average after 15 samples

Gibbs Sampling for GMMs

- We start with the full joint distribution:

$$p(X, Z, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\mu}, \Sigma) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) p(\Sigma_k)$$



Gibbs Sampling for GMMs

- We start with the full joint distribution:

$$p(X, Z, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\mu}, \Sigma) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) p(\Sigma_k)$$

- It can be shown that the full conditionals are:

$$p(z_i = k \mid \mathbf{x}_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) \propto \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k)$$

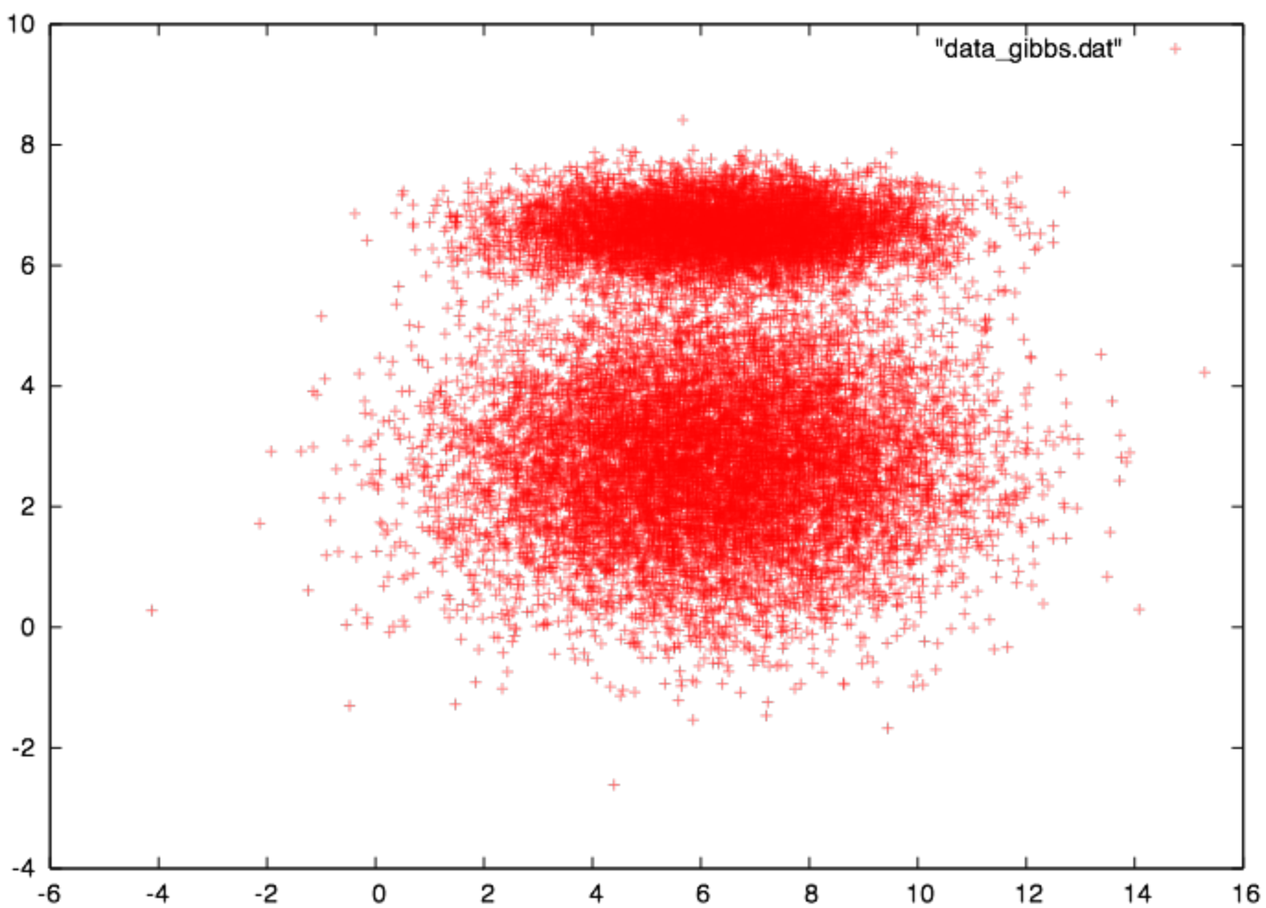
$$p(\boldsymbol{\pi} \mid \mathbf{z}) = \mathrm{Dir}(\{\alpha_k + \sum_{i=1}^{N} z_{ik}\}_{k=1}^{K})$$

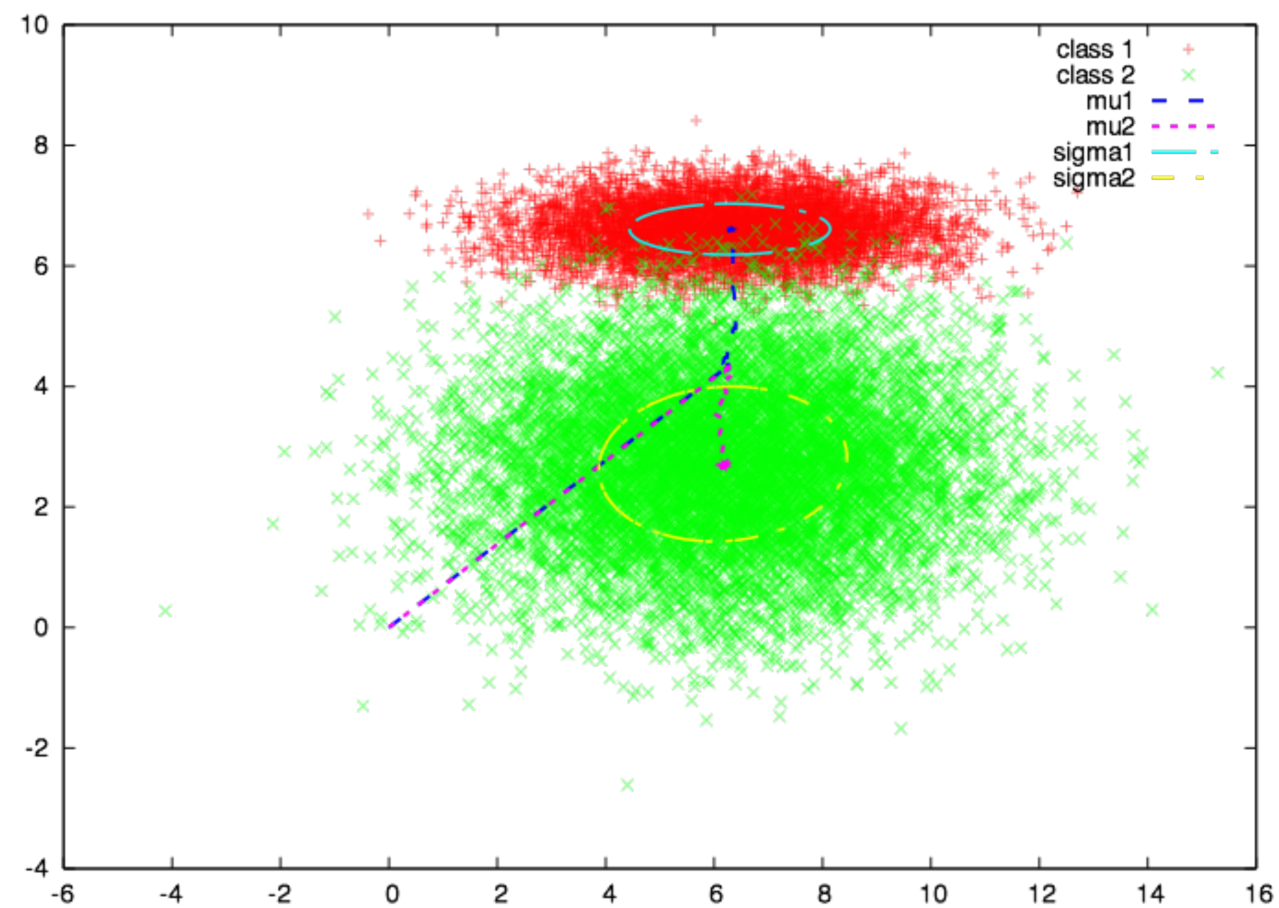
$$p(\boldsymbol{\mu}_k \mid \Sigma_k, Z, X) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, V_k)$$ (linear-Gaussian)

$$p(\Sigma_k \mid \boldsymbol{\mu}_k, Z, X) = \mathcal{IW}(\Sigma_k \mid S_k, \nu_k)$$

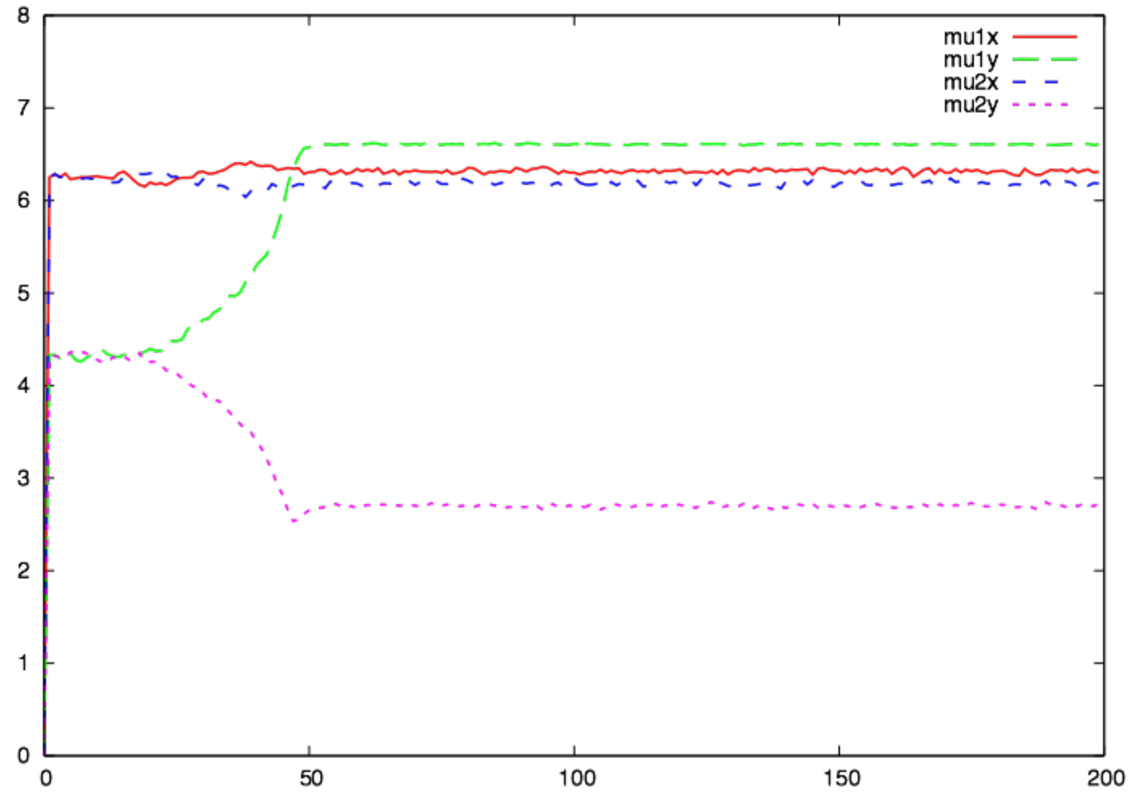
Gibbs Sampling for GMMs

- First, we initialize all variables
- Then we iterate over sampling from each conditional in turn
- In the end, we look at μ_k and Σ_k





How Often Do We Have To Sample?



- Here: after 50 sample rounds the values don't change any more
- In general, the **mixing time** τ_ϵ is related to the **eigen gap** $\gamma = \lambda_1 - \lambda_2$ of the transition matrix:

$$\tau_\epsilon \leq O(\frac{1}{\gamma} \log \frac{n}{\epsilon})$$

Gibbs Sampling is a Special Case of MH

- The proposal distribution in Gibbs sampling is

$$q(\mathbf{x}' \mid \mathbf{x}) = p(x_i' \mid \mathbf{x}_{-i})\mathbb{I}(\mathbf{x}'_{-i} = \mathbf{x}_{-i})$$

- This leads to an acceptance rate of:

$$\alpha = \frac{p(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})} = \frac{p(x_i' \mid \mathbf{x}'_{-i})p(\mathbf{x}'_{-i})p(x_i \mid \mathbf{x}'_{-i})}{p(x_i \mid \mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i' \mid \mathbf{x}_{-i})} = 1$$

- Although the acceptance is 100%, Gibbs sampling does not converge faster, as it only updates one variable at a time.

11. Variational Inference

Motivation

- A major task in probabilistic reasoning is to evaluate the posterior distribution $p(Z \mid X)$ of a set of latent variables Z given data X **(inference)**

However: This is often not tractable, e.g. because the latent space is high-dimensional

- Two different solutions are possible: sampling methods and variational methods.
- In variational optimization, we seek a tractable distribution q that **approximates** the posterior.
- Optimization is done using functionals.

Variational Inference

In general, variational methods are concerned with mappings that take **functions** as input.

Example: the entropy of a distribution p

$$\mathbb{H}[p] = \int p(x) \log p(x) dx$$

"Functional"

Variational optimization aims at finding **functions** that minimize (or maximize) a given functional.

This is mainly used to find approximations to a given function by choosing from a family.

The aim is mostly tractability and simplification.

MLE Revisited

Analogue to the discussion about EM we have:

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q \| p)$$

$$\mathcal{L}(q) = \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ \qquad \mathrm{KL}(q) = -\int q(Z) \log \frac{p(Z \mid X)}{q(Z)} dZ$$

Again, maximizing the lower bound is equivalent to minimizing the KL-divergence.

The maximum is reached when the KL-divergence vanishes, which is the case for $q(Z) = p(Z \mid X)$.

However: Often the true posterior is intractable and we restrict q to a tractable family of dist.

The KL-Divergence

Given: an unknown distribution p

We approximate that with a distribution q

The average additional amount of information is

$$-\int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} - \left(-\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \right) = -\int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \quad = \mathrm{KL}(p \| q)$$

This is known as the **Kullback-Leibler** divergence

It has the properties: $\mathrm{KL}(q \| p) \neq \mathrm{KL}(p \| q)$

$$\mathrm{KL}(p \| q) \geq 0 \qquad\qquad \mathrm{KL}(p \| q) = 0 \Leftrightarrow p \equiv q$$

This follows from Jensen’s inequality

Factorized Distributions

A common way to restrict q is to partition Z into disjoint sets so that q factorizes over the sets:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

This is the only assumption about q!

Idea: Optimize $\mathcal{L}(q)$ by optimizing wrt. each of the factors of q in turn. Setting $q_i(Z_i) = q_i$ we have

$$\mathcal{L}(q) = \int \prod_i q_i \left(\log p(X, Z) - \sum_i \log q_i \right) dZ$$

Mean Field Theory

This results in:

$$\mathcal{L}(q) = \int q_j \log \tilde{p}(X, Z_j) dZ_j - \int q_j \log q_j dZ_j + \text{const}$$

where

$$\log \tilde{p}(X, Z_j) = \mathbb{E}_{-j} \; [\log p(X, Z)] + \text{const}$$

Thus, we have $\mathcal{L}(q) = -\text{KL}(q_j \| \tilde{p}(X, Z_j)) + \text{const}$

I.e., maximizing the lower bound is equivalent to minimizing the KL-divergence of a single factor and a distribution that can be expressed in terms of an expectation:

$$\mathbb{E}_{-j} \; [\log p(X, Z)] = \int \log p(X, Z) \prod_{i \neq j} q_i dZ_{-j}$$

Mean Field Theory

Therefore, the optimal solution in general is

$$\log q_j^*(Z_j) = \mathbb{E}_{-j} [\log p(X, Z)] + \text{const}$$

In words: the log of the optimal solution for a factor q_j is obtained by taking the expectation with respect to **all other** factors of the log-joint probability of all observed and unobserved variables

The constant term is the normalizer and can be computed by taking the exponential and marginalizing over Z_j

This is not always necessary.

Variational Mixture of Gaussians

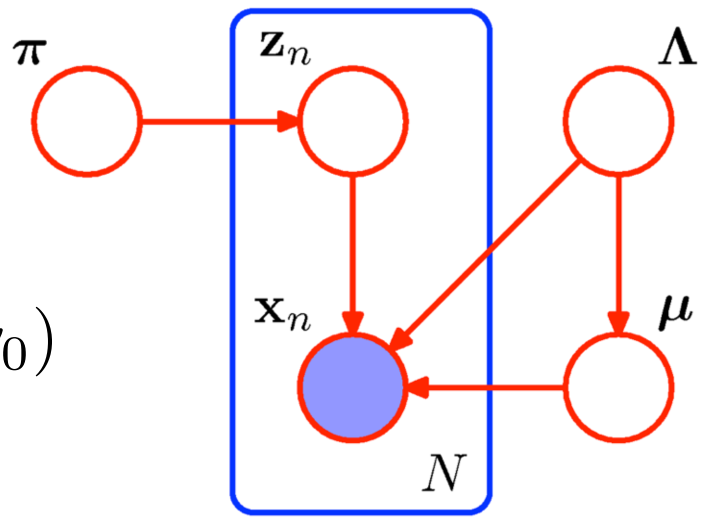
- Again, we have observed data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and latent variables $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Furthermore we have

$$p(Z \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \qquad p(X \mid Z, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Lambda^{-1})^{z_{nk}}$$

- We introduce priors for all parameters, e.g.

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0)$$

$$p(\boldsymbol{\mu}, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid W_0, \nu_0)$$



Variational Mixture of Gaussians

- The joint probability is then:

$$p(X, Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = p(X \mid Z, \boldsymbol{\mu}, \Lambda) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} \mid \Lambda) p(\Lambda)$$

- We consider a distribution q so that

$$q(Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = q(Z) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$$

- Using our general result:

$$\log q^*(Z) = \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda}[\log p(X, Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)] + \text{const}$$

- Plugging in:

$$\log q^*(Z) = \mathbb{E}_{\boldsymbol{\pi}}[\log p(Z \mid \boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu}, \Lambda}[\log p(X \mid Z, \boldsymbol{\mu}, \Lambda)] + \text{const}$$

Variational Mixture of Gaussians

- The joint probability is then:

$$p(X, Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = p(X \mid Z, \boldsymbol{\mu}, \Lambda)p(Z \mid \boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu} \mid \Lambda)p(\Lambda)$$

- We consider a distribution q so that

$$q(Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = q(Z)q(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$$

- Using our general result:

$$\log q^*(Z) = \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda}[\log p(X, Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)] + \text{const}$$

- Plugging in:

$$\log q^*(Z) = \mathbb{E}_{\boldsymbol{\pi}}[\log p(Z \mid \boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu}, \Lambda}[\log p(X \mid Z, \boldsymbol{\mu}, \Lambda)] + \text{const}$$

- From this we can show that: $$q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

Variational Mixture of Gaussians

This means: the optimal solution to the factor $q(Z)$ has the same functional form as the prior of Z.

It turns out, this is true for all factors.

However: the factors q depend on moments computed with respect to the other variables, i.e. the computation has to be done iteratively.

This results again in an EM-style algorithm, with the difference, that here we use conjugate priors for all parameters. This reduces overfitting.

Example: Clustering

- 6 Gaussians
- After convergence, only two components left
- Complexity is traded off with data fitting
- This behaviour depends on a parameter of the Dirichlet prior