

11. Variational Inference

Motivation

- A major task in probabilistic reasoning is to evaluate the **posterior distribution** $p(Z | X)$ 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**.



Motivation

- A major task in probabilistic reasoning is to evaluate the **posterior distribution** $p(Z | X)$ of a set of latent variables Z given data X (**inference**)
- How to do inference? (we will see later)
because we can't sample from $p(Z | X)$ directly
- Two main approaches:
 - Exact methods
 - Approximate methods
- In variational optimization, we seek a tractable distribution q that **approximates** the posterior.
- Optimization is done using **functionals**.

Careful: Different notation!

In Bishop (and in the following slides)
 **Z are hidden states
and X are observations**

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.



The KL-Divergence

Aim: define a functional that resembles a “difference” between distributions p and q

Idea: use the average additional amount of information:

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

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

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

$$\text{KL}(p\|q) \geq 0 \quad \text{KL}(p\|q) = 0 \Leftrightarrow p \equiv q$$

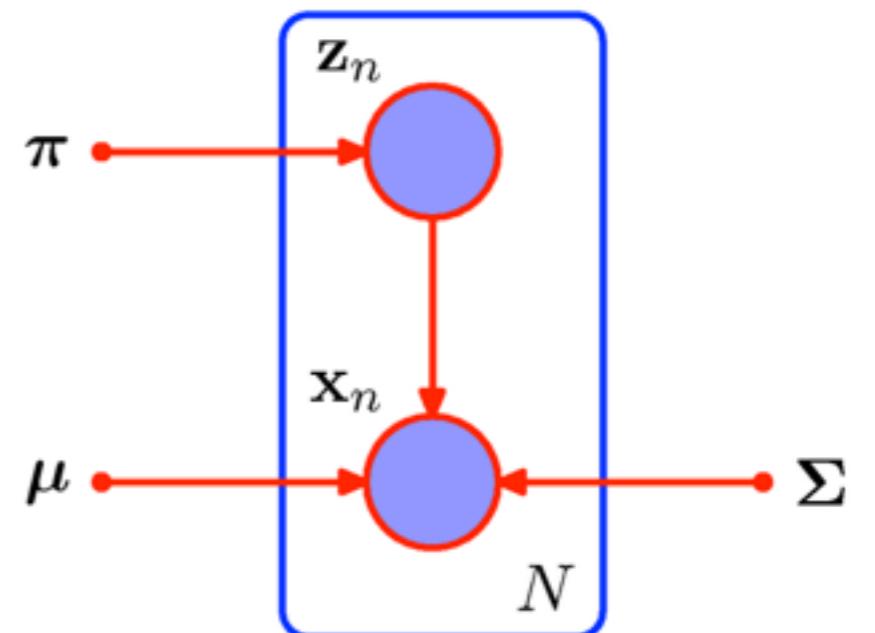
This follows from Jensen's inequality



Example: A Variational Formulation of EM

Assume for a moment that we observe X and the binary latent variables Z . The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^N p(z_n \mid \pi)p(x_n \mid z_n, \mu, \Sigma)$$



Example: A Variational Formulation of EM

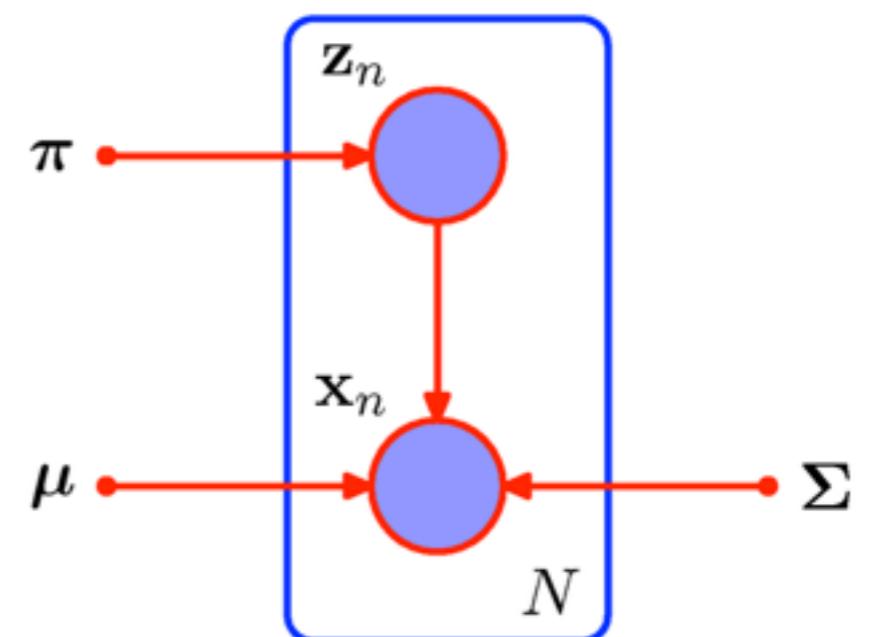
Assume for a moment that we observe X and the binary latent variables Z . The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^N p(\mathbf{z}_n \mid \pi)p(\mathbf{x}_n \mid \mathbf{z}_n, \mu, \Sigma)$$

Remember:
 $z_{nk} \in \{0, 1\}, \sum_{k=1}^K z_{nk} = 1$

where $p(\mathbf{z}_n \mid \pi) = \prod_{k=1}^K \pi_k^{z_{nk}}$ and

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \mu, \Sigma) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$



Example: A Variational Formulation of EM

Assume for a moment that we observe X and the binary latent variables Z . The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^N p(\mathbf{z}_n \mid \pi)p(\mathbf{x}_n \mid \mathbf{z}_n, \mu, \Sigma)$$

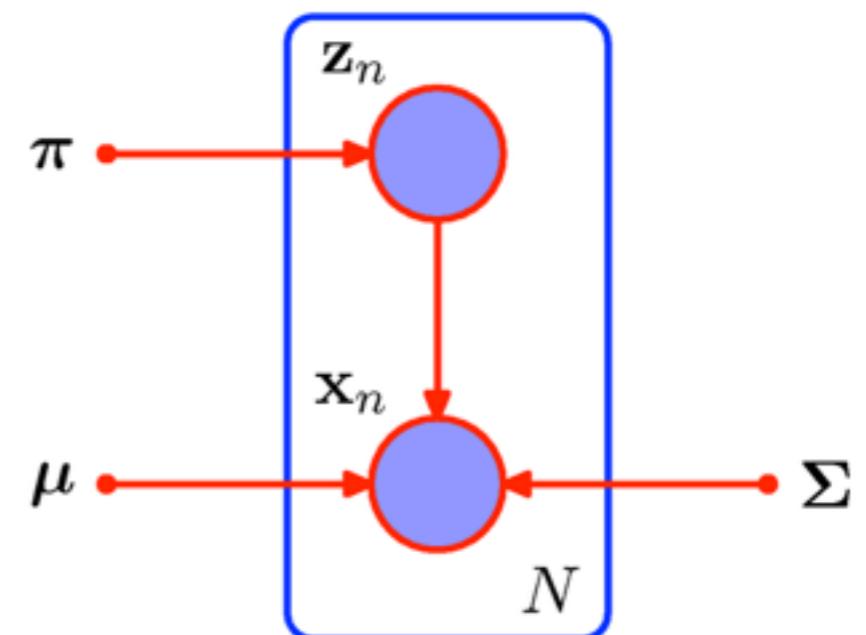
Remember:
 $z_{nk} \in \{0, 1\}, \sum_{k=1}^K z_{nk} = 1$

where $p(\mathbf{z}_n \mid \pi) = \prod_{k=1}^K \pi_k^{z_{nk}}$ and

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \mu, \Sigma) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

which leads to the log-formulation:

$$\log p(X, Z \mid \pi, \mu, \Sigma) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k))$$



The Complete-Data Log-Likelihood

$$\log p(X, Z \mid \pi, \mu, \Sigma) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

- This is called the **complete-data log-likelihood**
- Advantage: solving for the parameters (π_k, μ_k, Σ_k) is much simpler, as the log is inside the sum!
- We could switch the sums and then for every mixture component k only look at the points that are associated with that component.
- This leads to simple closed-form solutions for the parameters
- However: the latent variables Z are not observed!



The Main Idea of EM

Instead of maximizing the joint log-likelihood, we maximize its **expectation** under the latent variable distribution:

$$\mathbb{E}_Z[\log p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)] = \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}_Z[z_{nk}] (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$



The Main Idea of EM

Instead of maximizing the joint log-likelihood, we maximize its **expectation** under the latent variable distribution:

$$\mathbb{E}_Z[\log p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)] = \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}_Z[z_{nk}] (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

where the latent variable distribution per point is:

$$\begin{aligned} p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) &= \frac{p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta}) p(\mathbf{z}_n \mid \boldsymbol{\theta})}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \quad \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{\prod_{l=1}^K (\pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l))^{z_{nl}}}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned}$$



The Main Idea of EM

The expected value of the latent variables is:

$$\mathbb{E}[z_{nk}] = \gamma(z_{nk})$$

plugging in we obtain:

$$\mathbb{E}_Z[\log p(X, Z | \pi, \mu, \Sigma)] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk})(\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k))$$

We compute this iteratively:

1. Initialize $i = 0, (\pi_k^i, \mu_k^i, \Sigma_k^i)$
2. Compute $\mathbb{E}[z_{nk}] = \gamma(z_{nk})$
3. Find parameters $(\pi_k^{i+1}, \mu_k^{i+1}, \Sigma_k^{i+1})$ that maximize this
4. Increase i ; if not converged, goto 2.

Remember:

$$\gamma(z_{nk}) = p(z_{nk} = 1 | \mathbf{x}_n)$$



Why Does This Work?

- We have seen that EM maximizes the **expected complete-data log-likelihood**, but:
- Actually, we need to maximize the log-marginal

$$\log p(X \mid \theta) = \log \sum_Z p(X, Z \mid \theta)$$

- It turns out that the log-marginal is maximized **implicitly!**



A Variational Formulation of EM

- We have seen that EM maximizes the **expected complete-data log-likelihood**, but:
- Actually, we need to maximize the log-marginal

$$\log p(X \mid \theta) = \log \sum_Z p(X, Z \mid \theta)$$

- It turns out that the log-marginal is maximized **implicitly!**

$$\log p(X \mid \theta) = \mathcal{L}(q, \theta) + \text{KL}(q \| p)$$

$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)}$$

$$\text{KL}(q \| p) = - \sum_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)}$$



A Variational Formulation of EM

- Thus: The Log-likelihood consists of two functionals

$$\log p(X \mid \theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

where the first is (proportional to) an **expected complete-data log-likelihood** under a distribution q

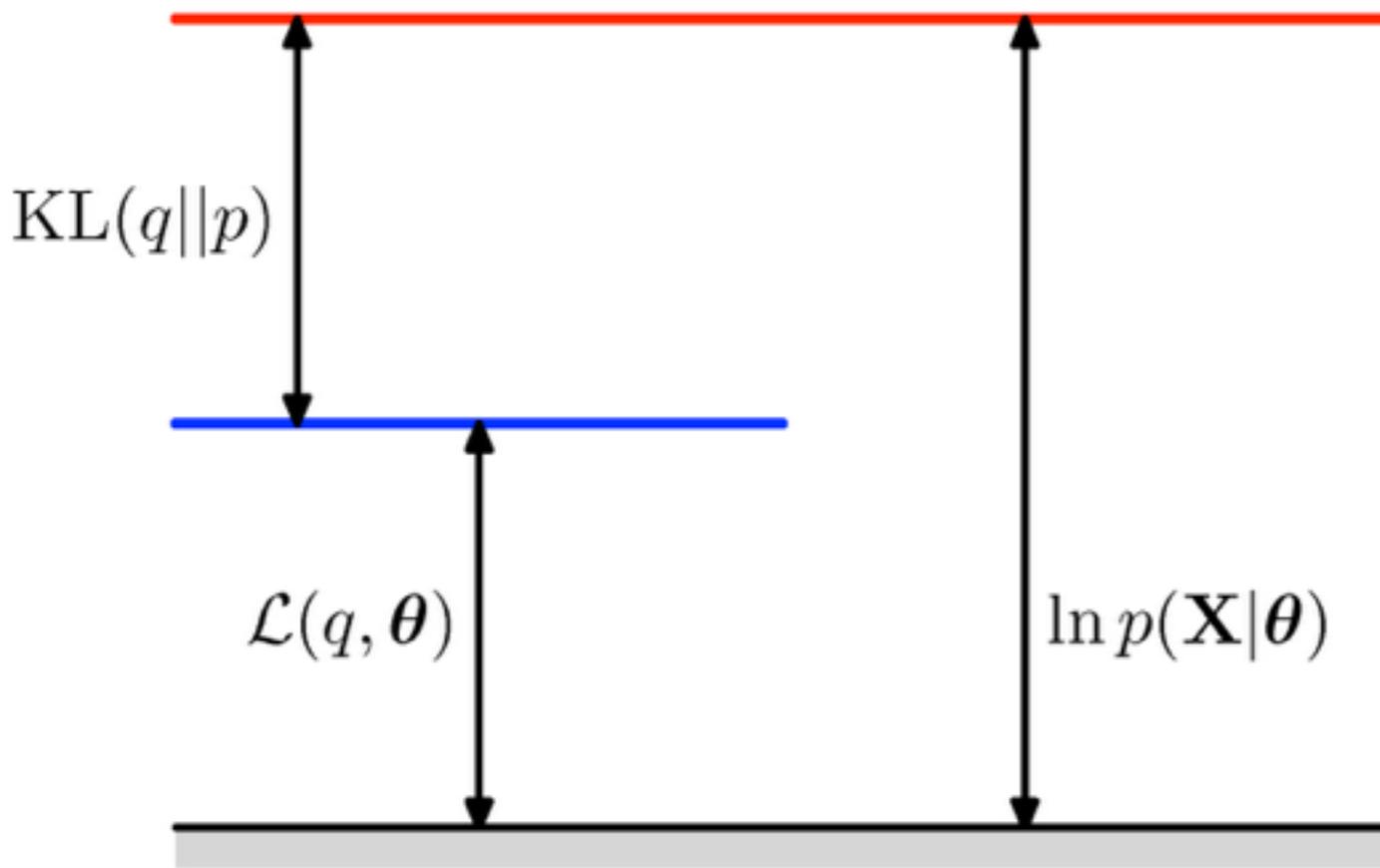
$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)}$$

and the second is the **KL-divergence** between p and q :

$$\text{KL}(q\|p) = - \sum_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)}$$



Visualization

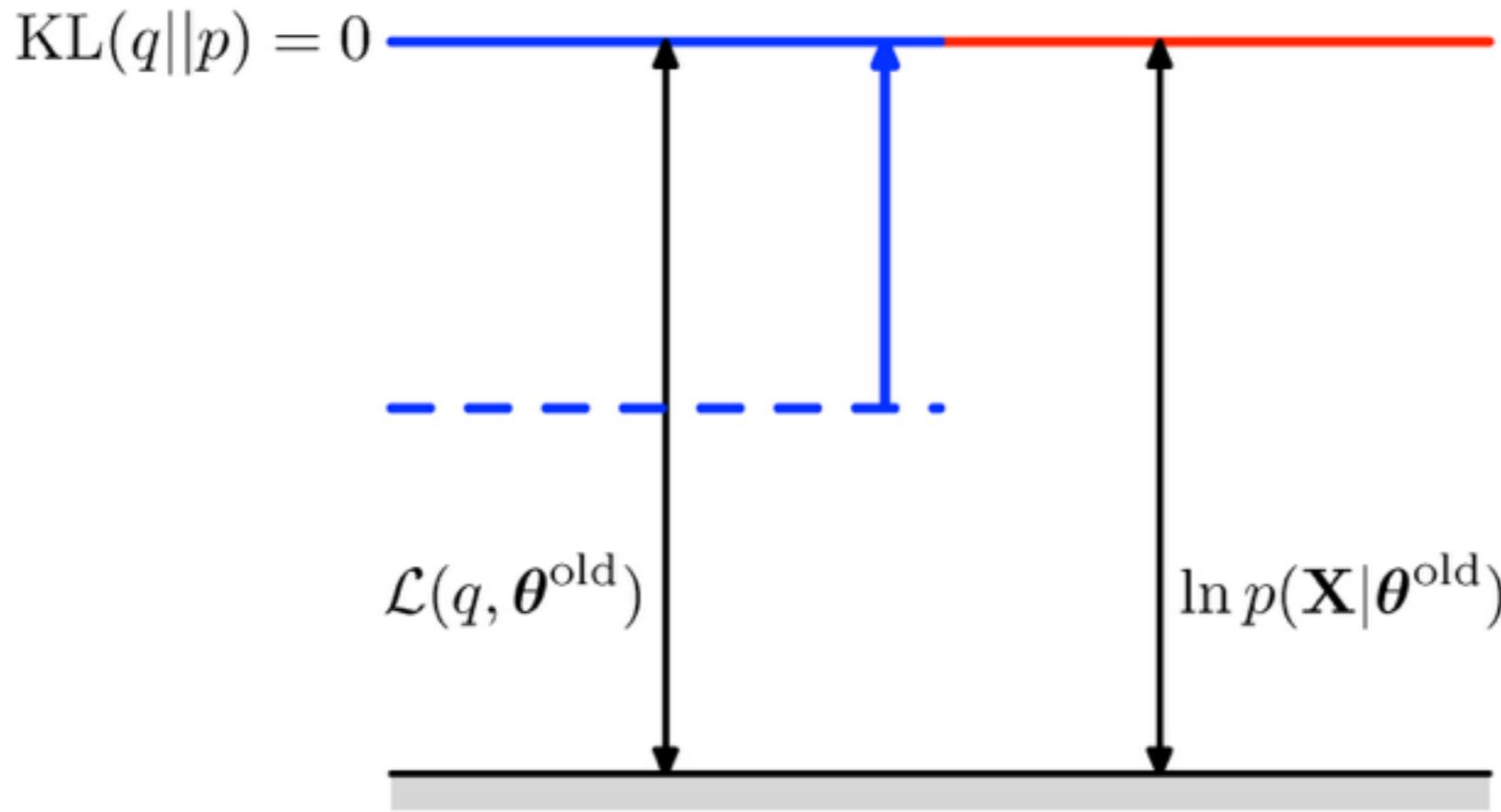


- The KL-divergence is positive or 0
- Thus, the log-likelihood is at least as large as \mathcal{L} or:
- \mathcal{L} is a **lower bound** of the log-likelihood:

$$\log p(X | \theta) \geq \mathcal{L}(q, \theta)$$



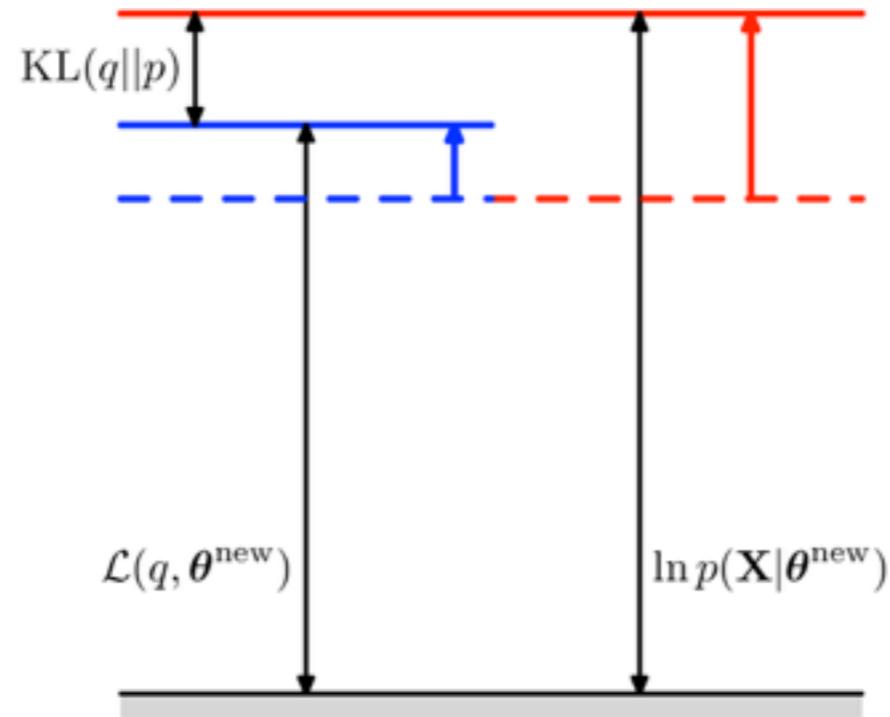
What Happens in the E-Step?



- The log-likelihood is independent of q
- Thus: \mathcal{L} is maximized iff KL divergence is minimal ($=0$)
- This is the case iff $q(Z) = p(Z | X, \boldsymbol{\theta})$



What Happens in the M-Step?



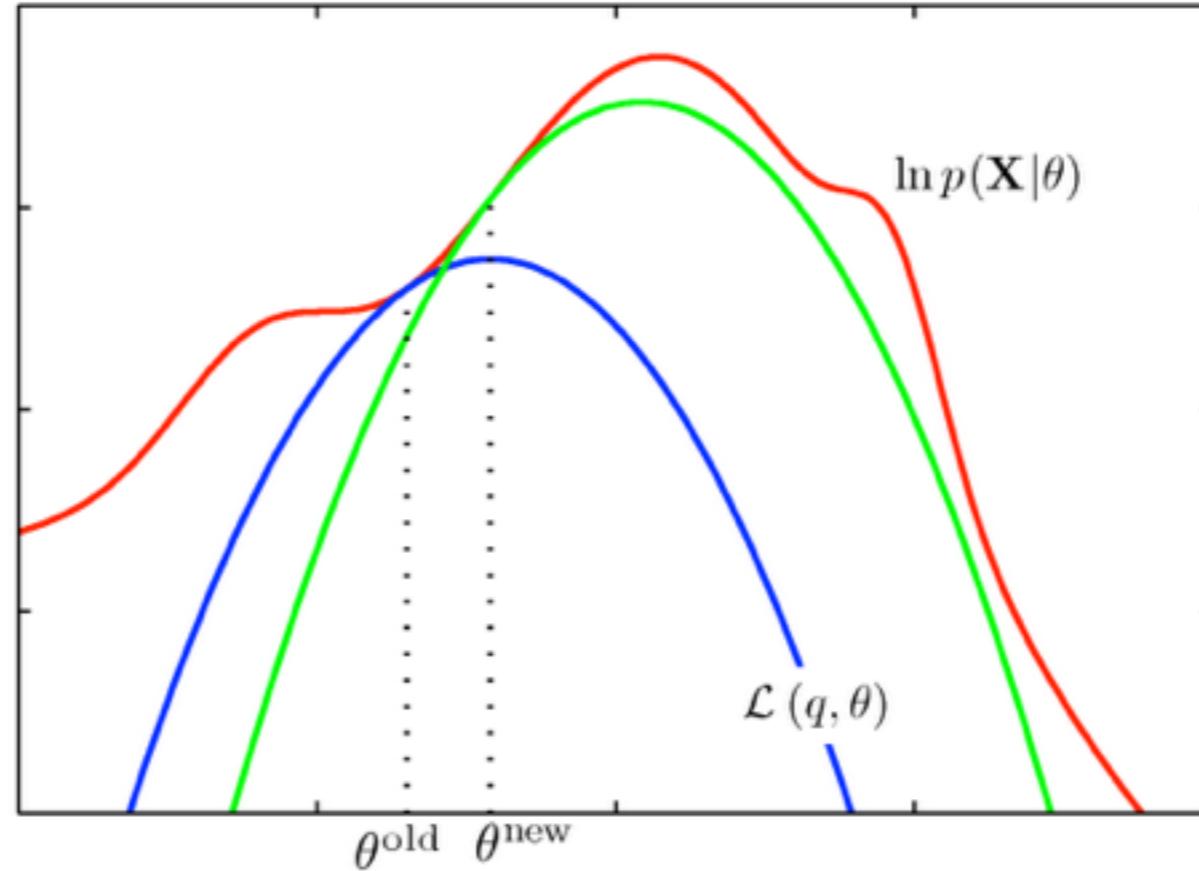
- In the M-step we keep q fixed and find new θ

$$\mathcal{L}(q, \theta) = \sum_Z p(Z | X, \theta^{\text{old}}) \log p(X, Z | \theta) - \sum_Z q(Z) \log q(Z)$$

- We maximize the first term, the second is indep.
- This implicitly makes KL non-zero
- The log-likelihood is maximized even more!



Visualization in Parameter-Space



- In the E-step we compute the concave lower bound for given old parameters θ^{old} (blue curve)
- In the M-step, we maximize this lower bound and obtain new parameters θ^{new}
- This is repeated (green curve) until convergence



Generalizing the Idea

- In EM, we were looking for an optimal distribution q in terms of KL-divergence
- Luckily, we could compute q in closed form
- In general, this is not the case, but we can use an approximation instead: $q(Z) \approx p(Z | X)$
- Idea: make a simplifying assumption on q so that a good approximation can be found
- For example: Consider the case where q can be expressed as a **product** of simpler terms



Factorized Distributions

We can split up q by partitioning Z into disjoint sets and assuming that q factorizes over the sets:

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

Shorthand:
 $q_i \leftarrow q_i(Z_i)$

This is the only assumption about q !

Idea: Optimize $\mathcal{L}(q)$ by optimizing wrt. each of the factors of q in turn. Setting $q_i \leftarrow q_i(Z_i)$ 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 $\mathcal{L}(q) = -\text{KL}(q_j \parallel \tilde{p}(X, Z_j)) + \text{const}$

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

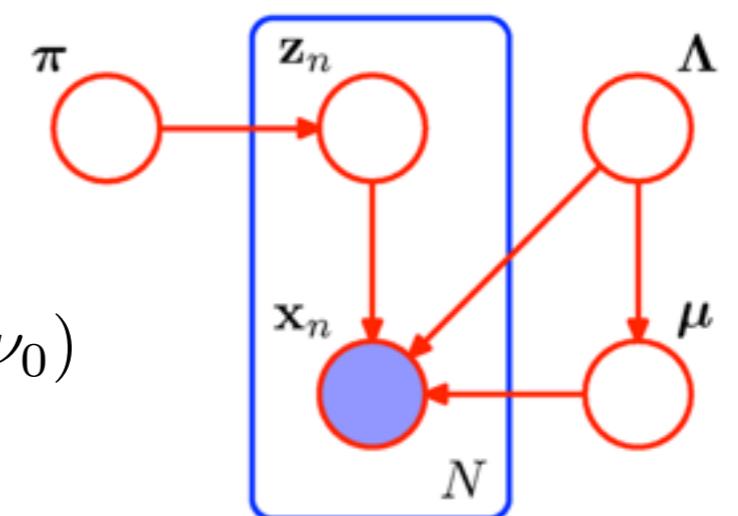
- Again, we have observed data $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and latent variables $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$
- Furthermore we have

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

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

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

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



Variational Mixture of Gaussians

- The joint probability is then:

$$p(X, Z, \pi, \mu, \Lambda) = p(X | Z, \mu, \Lambda)p(Z | \pi)p(\pi)p(\mu | \Lambda)p(\Lambda)$$

- We consider a distribution q so that

$$q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda)$$

- Using our general result:

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

- Plugging in:

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



Variational Mixture of Gaussians

- The joint probability is then:

$$p(X, Z, \pi, \mu, \Lambda) = p(X | Z, \mu, \Lambda)p(Z | \pi)p(\pi)p(\mu | \Lambda)p(\Lambda)$$

- We consider a distribution q so that

$$q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda)$$

- Using our general result:

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

- Plugging in:

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

- 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.

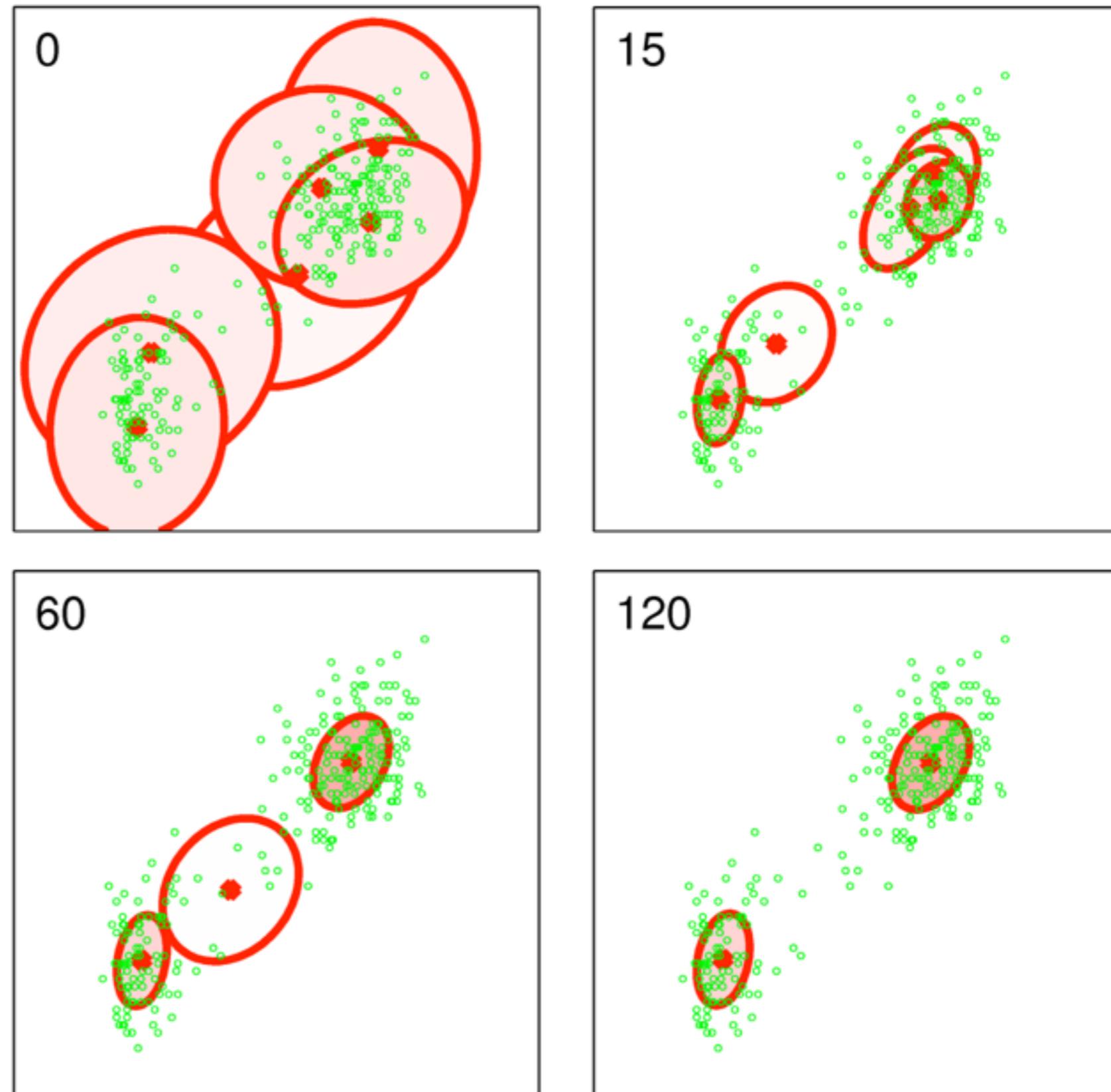
However: the factors q depend on moments computed with respect to the other variables, i.e. the computation has to be done iteratively.

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



Expectation Propagation

Exponential Families

Definition: A probability distribution p over \mathbf{x} is a member of the **exponential family** if it can be expressed as

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}))$$

where $\boldsymbol{\eta}$ are the **natural parameters** and

$$g(\boldsymbol{\eta}) = \left(\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) d\mathbf{x} \right)^{-1}$$

is the normalizer.

h and \mathbf{u} are functions of \mathbf{x} .



Exponential Families

Example: Bernoulli-Distribution with parameter μ

$$\begin{aligned} p(x | \mu) &= \mu^x (1 - \mu)^{1-x} \\ &= \exp(x \ln \mu + (1 - x) \ln(1 - \mu)) \\ &= \exp(x \ln \mu + \ln(1 - \mu) - x \ln(1 - \mu)) \\ &= (1 - \mu) \exp(x \ln \mu - x \ln(1 - \mu)) \\ &= (1 - \mu) \exp\left(x \ln\left(\frac{\mu}{1 - \mu}\right)\right) \end{aligned}$$

Thus, we can say

$$\eta = \ln\left(\frac{\mu}{1 - \mu}\right) \Rightarrow \mu = \frac{1}{1 + \exp(-\eta)} \Rightarrow 1 - \mu = \frac{1}{1 + \exp(\eta)} = g(\eta)$$



MLE for Exponential Families

From:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) d\mathbf{x} = 1$$

we get:

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

$$\Rightarrow -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

which means that $-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$



MLE for Exponential Families

From:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) d\mathbf{x} = 1$$

we get:

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

$$\Rightarrow -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

which means that $-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$

$\Sigma \mathbf{u}(\mathbf{x})$ is called the **sufficient statistics** of p .



Expectation Propagation

In mean-field we minimized $\text{KL}(q||p)$. But: we can also minimize $\text{KL}(p||q)$. Assume q is from the **exponential family**:

$$q(\mathbf{z}) = h(\mathbf{z})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{z}))$$

natural parameters

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{z})) d\mathbf{x} = 1$$

normalizer

Then we have:

$$\text{KL}(p||q) = - \int p(\mathbf{z}) \log \frac{h(\mathbf{z})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{z}))}{p(\mathbf{z})} d\mathbf{z}$$



Expectation Propagation

This results in $\text{KL}(p\|q) = -\log g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \mathbb{E}_p[\mathbf{u}(\mathbf{x})] + \text{const}$

We can minimize this with respect to $\boldsymbol{\eta}$

$$-\nabla \log g(\boldsymbol{\eta}) = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$$



Expectation Propagation

This results in $\text{KL}(p\|q) = -\log g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \mathbb{E}_p[\mathbf{u}(\mathbf{x})] + \text{const}$

We can minimize this with respect to $\boldsymbol{\eta}$

$$-\nabla \log g(\boldsymbol{\eta}) = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$$

which is equivalent to

$$\mathbb{E}_q[\mathbf{u}(\mathbf{x})] = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$$

Thus: the KL-divergence is minimal if the exp. sufficient statistics are the same between p and q !

For example, if q is Gaussian: $\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$

Then, mean and covariance of q must be the same as for p (**moment matching**)



Expectation Propagation

Assume we have a factorization $p(\mathcal{D}, \theta) = \prod_{i=1}^M f_i(\theta)$ and we are interested in the posterior:

$$p(\theta | \mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_{i=1}^M f_i(\theta)$$

we use an approximation $q(\theta) = \frac{1}{Z} \prod_{i=1}^M \tilde{f}_i(\theta)$

Aim: minimize $\text{KL} \left(\frac{1}{p(\mathcal{D})} \prod_{i=1}^M f_i(\theta) \middle\| \frac{1}{Z} \prod_{i=1}^M \tilde{f}_i(\theta) \right)$

Idea: optimize each of the approximating factors in turn, assume exponential family



The EP Algorithm

- Given: a joint distribution over data and variables

$$p(\mathcal{D}, \theta) = \prod_{i=1}^M f_i(\theta)$$

- Goal: approximate the posterior $p(\theta | \mathcal{D})$ with q
- Initialize all approximating factors $\tilde{f}_i(\theta)$
- Initialize the posterior approximation $q(\theta) \propto \prod_i \tilde{f}_i(\theta)$
- Do until convergence:
 - choose a factor $\tilde{f}_j(\theta)$
 - remove the factor from q by division: $q^{\\backslash j}(\theta) = \frac{q(\theta)}{\tilde{f}_j(\theta)}$



The EP Algorithm

- find q^{new} that minimizes

$$\text{KL} \left(\frac{f_j(\theta) q^{\setminus j}(\theta)}{Z_j} \middle| q^{\text{new}}(\theta) \right)$$

using moment matching, including the zeroth order moment:

$$Z_j = \int q^{\setminus j}(\theta) f_j(\theta) d\theta$$

- evaluate the new factor

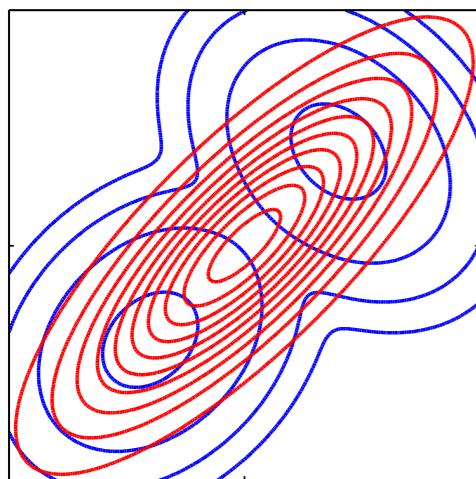
$$\tilde{f}_j(\theta) = Z_j \frac{q^{\text{new}}(\theta)}{q^{\setminus j}(\theta)}$$

- After convergence, we have $p(\mathcal{D}) \approx \int \prod_i \tilde{f}_j(\theta) d\theta$

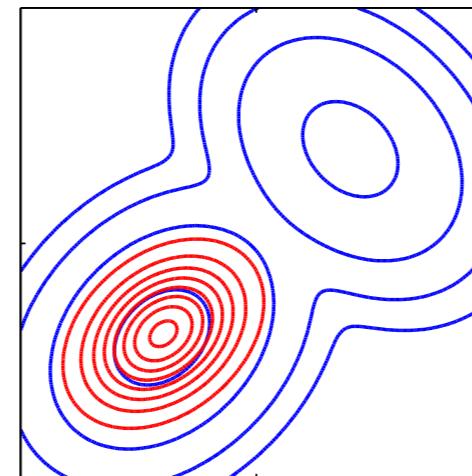


Properties of EP

- There is no guarantee that the iterations will converge
- This is in contrast to variational Bayes, where iterations do not decrease the lower bound
- EP minimizes $KL(p\|q)$ where variational Bayes minimizes $KL(q\|p)$



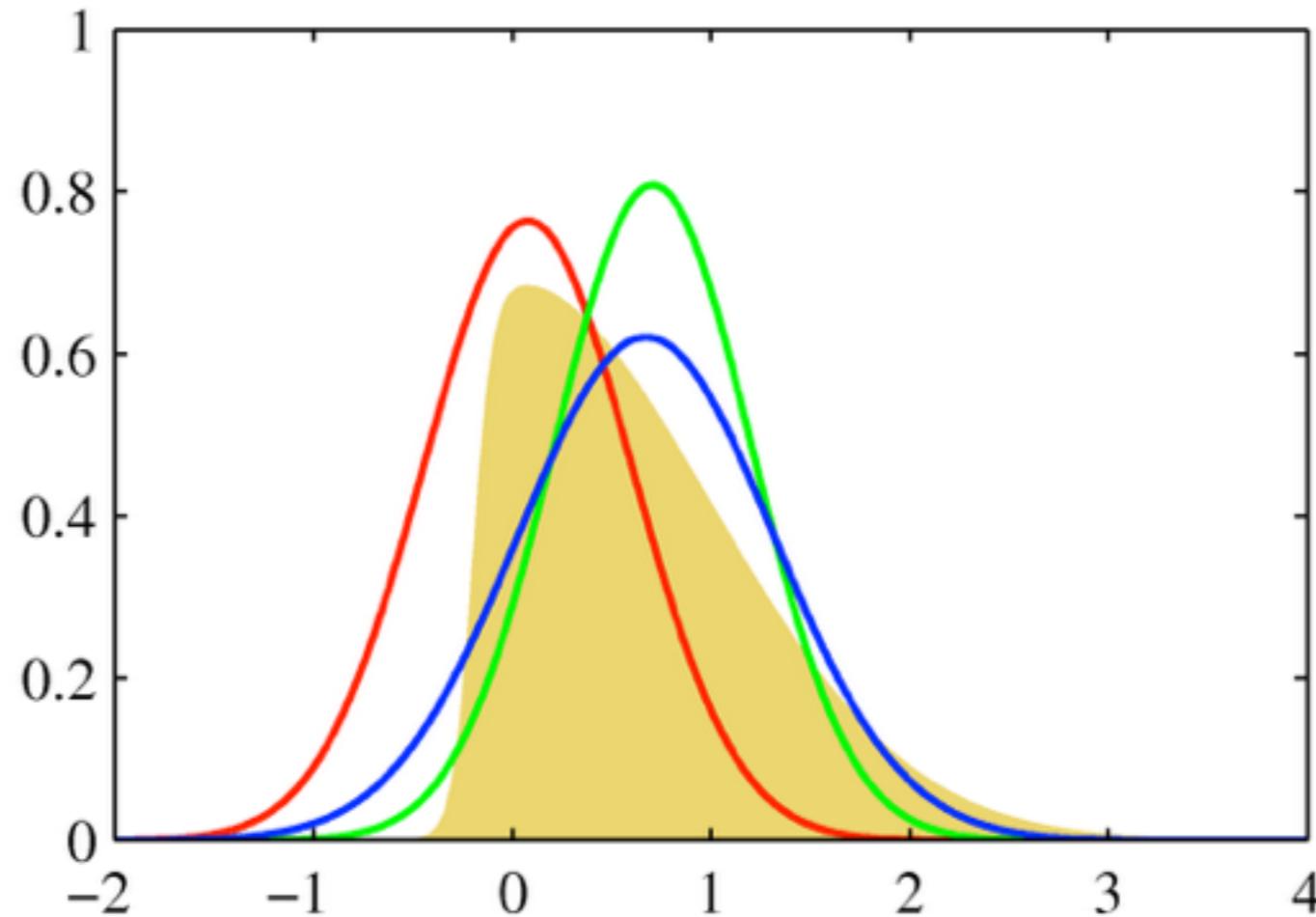
$KL(p\|q)$



$KL(q\|p)$



Example



yellow: original distribution

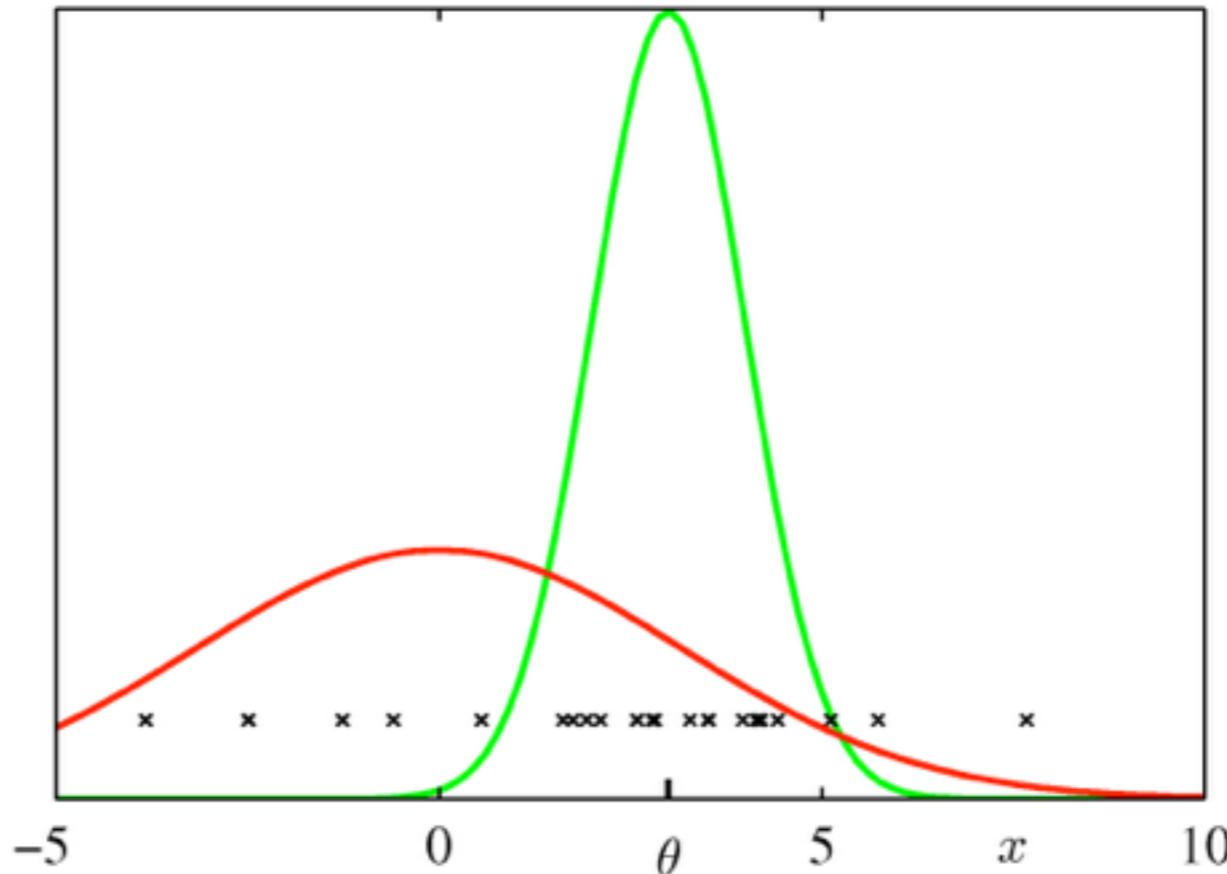
red: Laplace approximation

green: global variation

blue: expectation-propagation



The Clutter Problem



- Aim: fit a multivariate Gaussian into data in the presence of background clutter (also Gaussian)

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = (1 - w)\mathcal{N}(\mathbf{x} \mid \boldsymbol{\theta}, I) + w\mathcal{N}(\mathbf{x} \mid \mathbf{0}, aI)$$

- The prior is Gaussian:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{0}, bI)$$



The Clutter Problem

The joint distribution for $\mathcal{D} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ is

$$p(\mathcal{D}, \theta) = p(\theta) \prod_{n=1}^N p(\mathbf{x}_n \mid \theta)$$

this is a mixture of 2^N Gaussians! This is intractable for large N . Instead, we approximate it using a spherical Gaussian:

$$q(\theta) = \mathcal{N}(\theta \mid \mathbf{m}, vI) = \tilde{f}_0(\theta) \prod_{n=1}^N \tilde{f}_n(\theta)$$

the factors are (unnormalized) Gaussians:

$$\tilde{f}_0(\theta) = p(\theta) \quad \tilde{f}_n(\theta) = s_n \mathcal{N}(\theta \mid \mathbf{m}_n, v_n I)$$



EP for the Clutter Problem

- First, we initialize $\tilde{f}_n(\theta) = 1$, i.e. $q(\theta) = p(\theta)$
- Iterate:
 - Remove the current estimate of $\tilde{f}_n(\theta)$ from q by division of Gaussians:

$$q_{-n}(\theta) = \frac{q(\theta)}{\tilde{f}_n(\theta)}$$



EP for the Clutter Problem

- First, we initialize $\tilde{f}_n(\theta) = 1$, i.e. $q(\theta) = p(\theta)$
- Iterate:
 - Remove the current estimate of $\tilde{f}_n(\theta)$ from q by division of Gaussians:

$$q_{-n}(\theta) = \frac{q(\theta)}{\tilde{f}_n(\theta)} \quad q_{-n}(\theta) = \mathcal{N}(\theta \mid \mathbf{m}_{-n}, v_{-n} I)$$

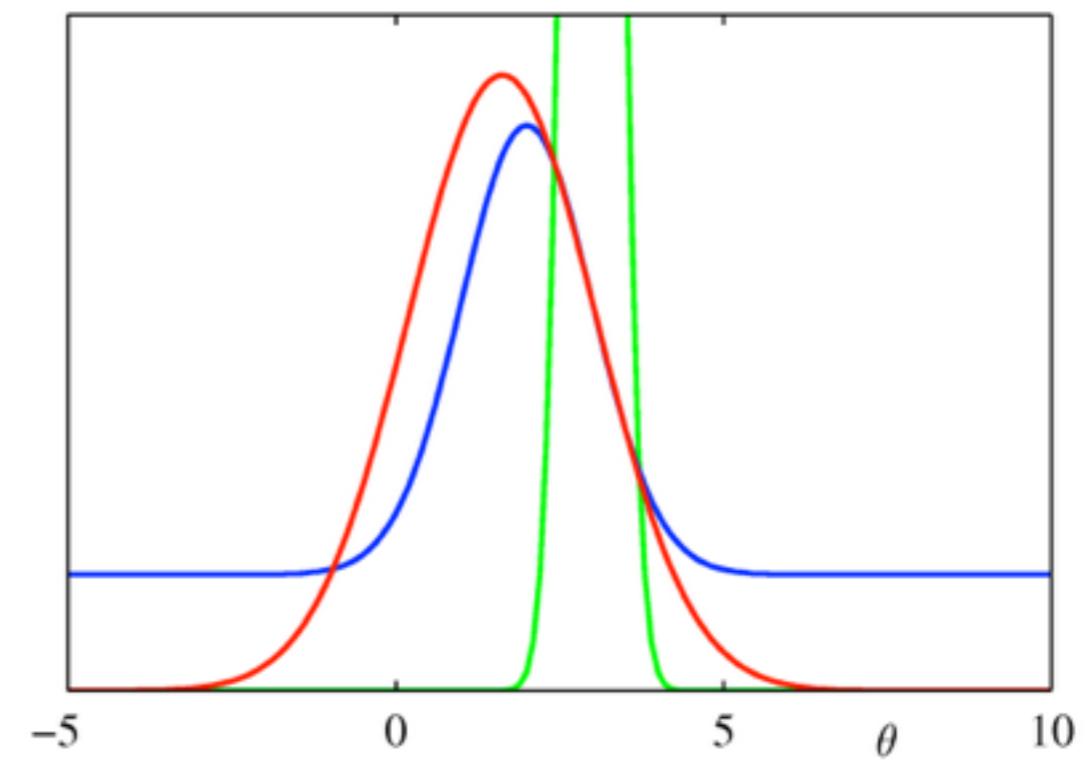
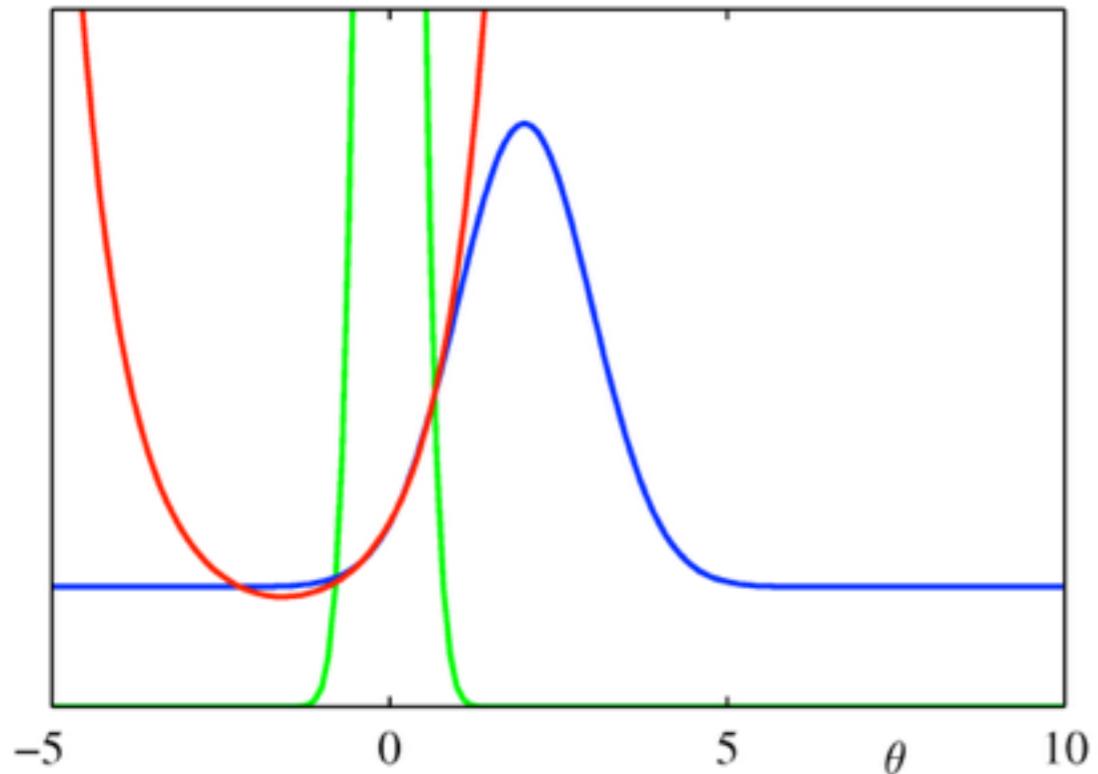
- Compute the normalization constant:

$$Z_n = \int q_{-n}(\theta) f_n(\theta) d\theta$$

- Compute mean and variance of $q^{\text{new}} = q_{-n}(\theta) f_n(\theta)$
- Update the factor $\tilde{f}_n(\theta) = Z_n \frac{q^{\text{new}}(\theta)}{q_{-n}(\theta)}$



A 1D Example



- blue: true factor $f_n(\theta)$
- red: approximate factor $\tilde{f}_n(\theta)$
- green: cavity distribution $q_{-n}(\theta)$

The form of $q_{-n}(\theta)$ controls the range over which $\tilde{f}_n(\theta)$ will be a good approximation of $f_n(\theta)$



Summary

- **Variational Inference** uses approximation of functions so that the KL-divergence is minimal
- In **mean-field** theory, factors are optimized sequentially by taking the expectation over all other variables
- **Expectation propagation** minimizes the reverse KL-divergence of a single factor by moment matching; factors are in the exp. family

