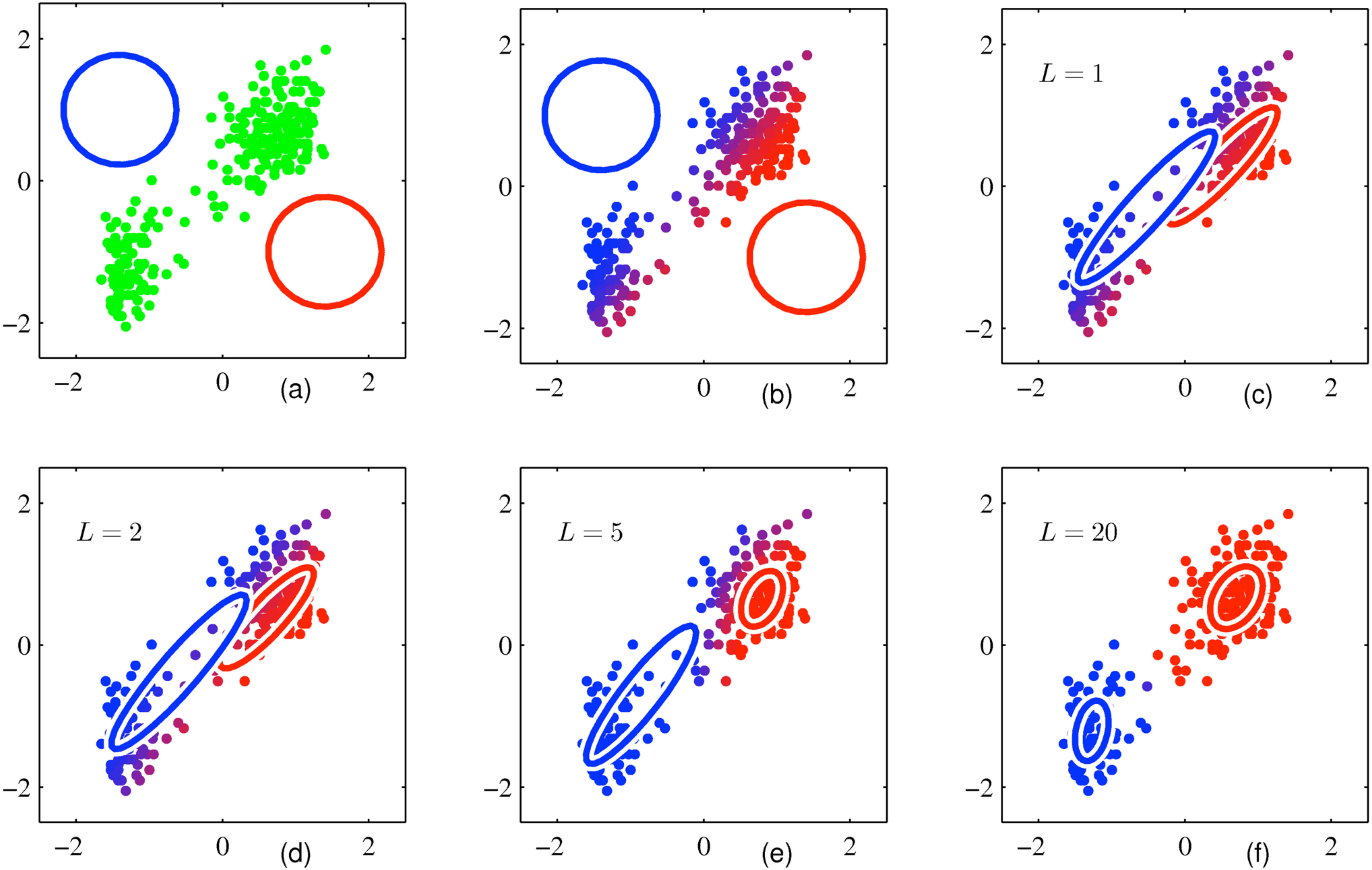
Repetition: 2D Gaussian Mixture Model



Repetition: Mixtures of Gaussians

- Assume that the data consists of K clusters
- The data within each cluster is Gaussian
- For any data point $\mathbf{x}$ we introduce a K-dimensional binary random variable $\mathbf{z}$ so that:

$$p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \underbrace{p(z_k = 1)}_{=:\pi_k} \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

where

$$z_k \in \{0, 1\}, \quad \sum_{k=1}^{K} z_k = 1$$

Repetition: Mixtures of Gaussians

- Assume that the data consists of K clusters
- The data within each cluster is Gaussian
- For any data point $\mathbf{x}$ we introduce a K-dimensional binary random variable $\mathbf{z}$ so that:

$$p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \underbrace{p(z_k = 1)}_{=:\pi_k} \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

- For all data points:

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

Rep.: The Complete-Data Log-Likelihood

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

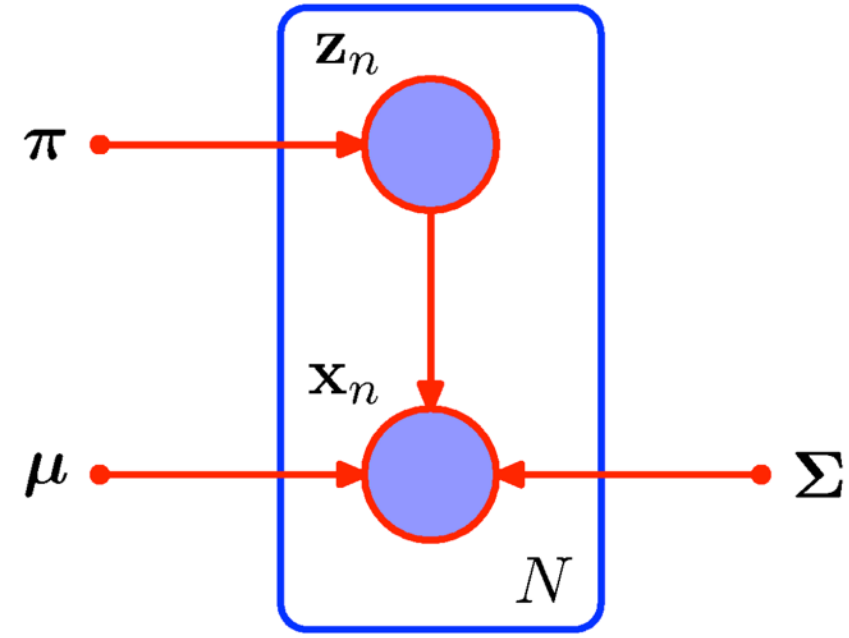
$$p(X, Z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^{N} p(\mathbf{z}_n \mid \boldsymbol{\pi}) p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\mu}, \Sigma)$$

Remember:

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

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

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



which leads to the log-formulation:

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

Recap: The Main Idea of EM

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

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

Recap: The Main Idea of EM

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

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

where the latent variable distribution per point is:

$$p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta}) p(\mathbf{z}_n \mid \boldsymbol{\theta})}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \qquad \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)$$

$$= \frac{\prod_{l=1}^{K} (\pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l))^{z_{nl}}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \Sigma_j)}$$

The Main Idea of EM

The expected value of the latent variables is:

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

Remember:

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

plugging in we obtain:

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

We compute this iteratively:

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

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

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

- 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

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

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

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

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{Z} q(Z) \log \frac{p(X, Z \mid \boldsymbol{\theta})}{q(Z)} \qquad \mathrm{KL}(q \| p) = -\sum_{Z} q(Z) \log \frac{p(Z \mid X, \boldsymbol{\theta})}{q(Z)}$$

A Variational Formulation of EM

- Thus: The Log-likelihood consists of two functionals

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

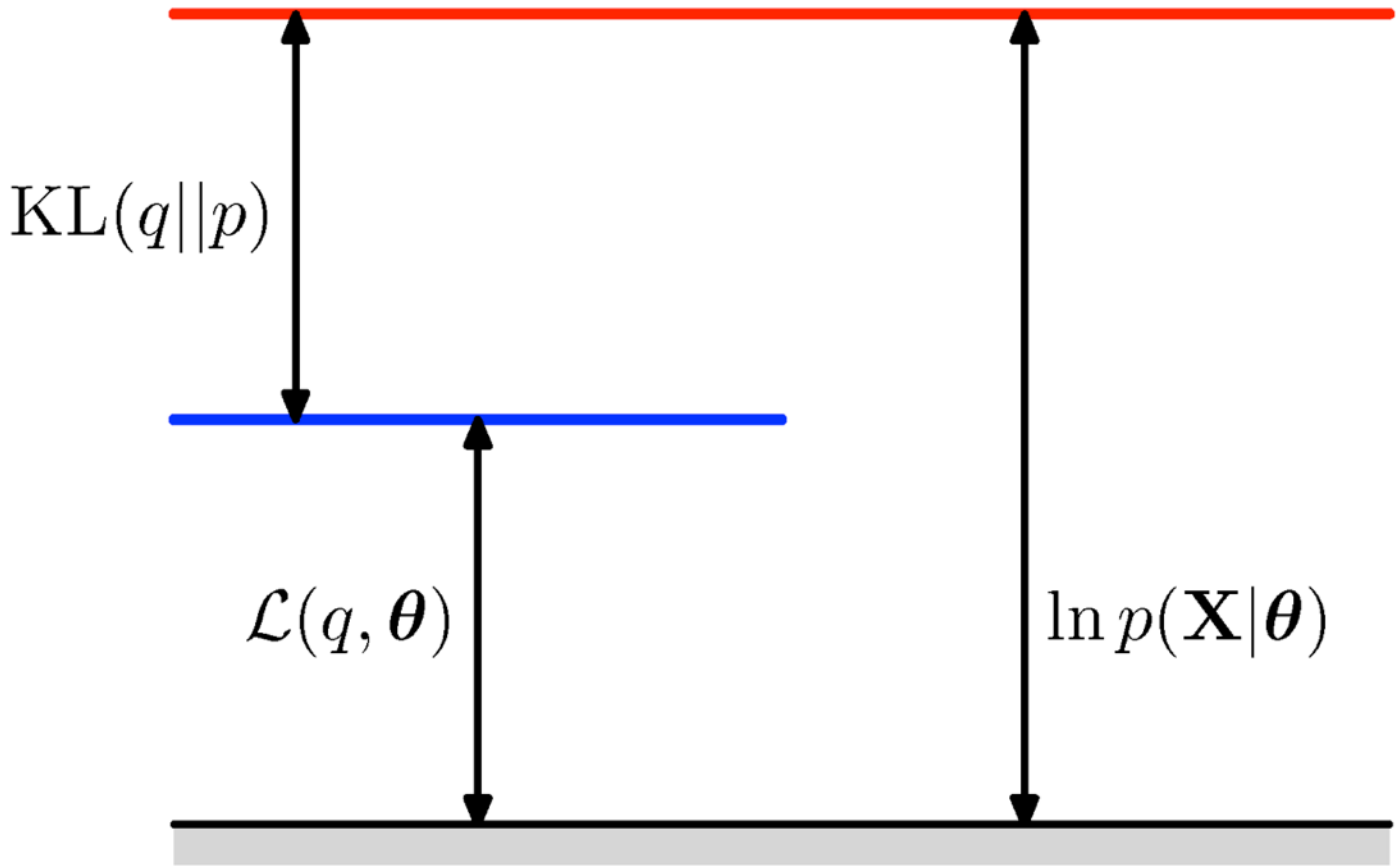
where the first is (proportional to) an **expected complete-data log-likelihood** under a distribution q

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

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

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

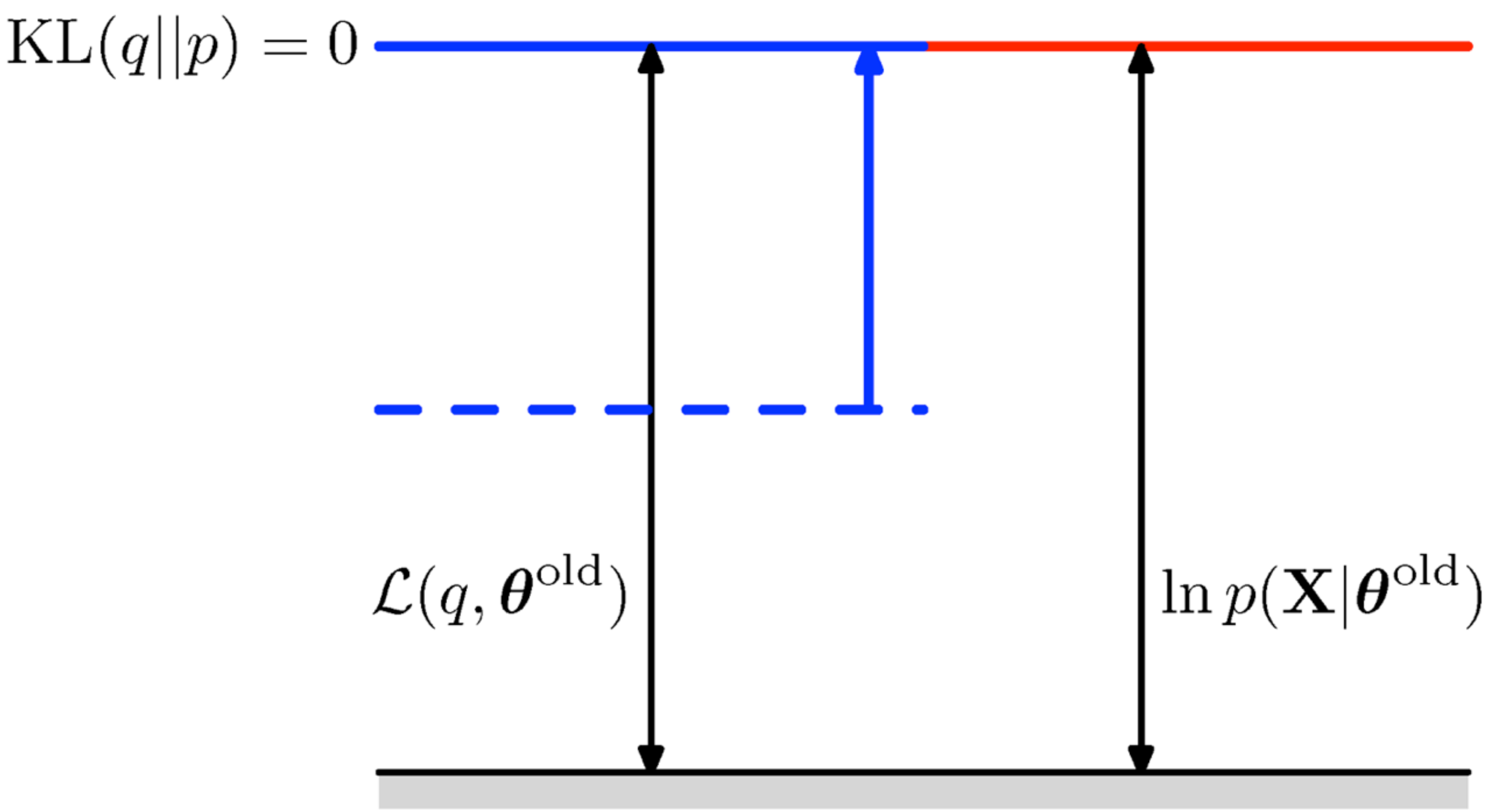
Visualization



- 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:

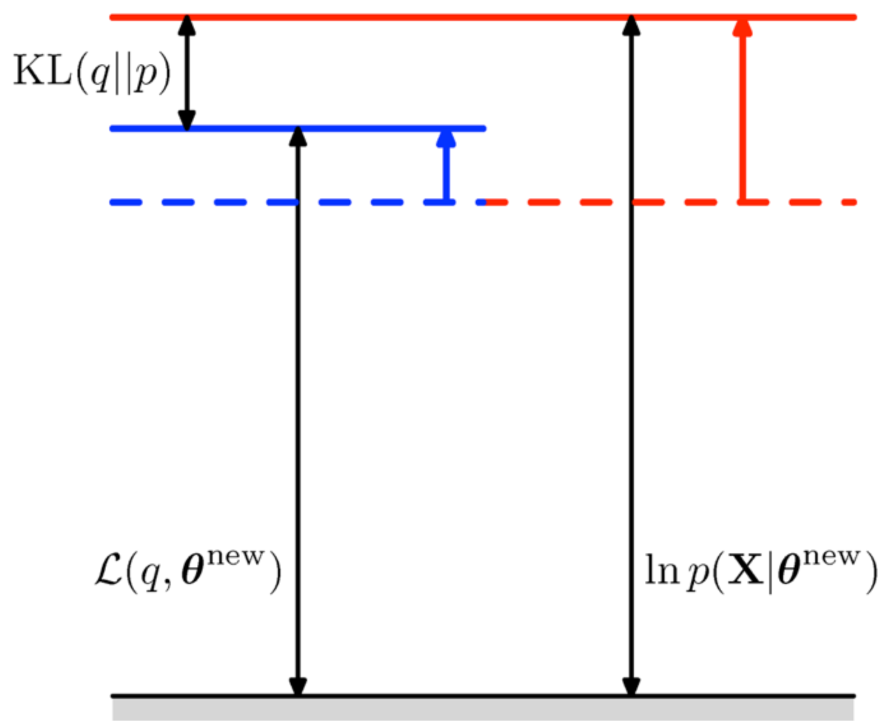
$$\log p(X \mid \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

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 $q(Z) = p(Z \mid X, \boldsymbol{\theta})$

What Happens in the M-Step?

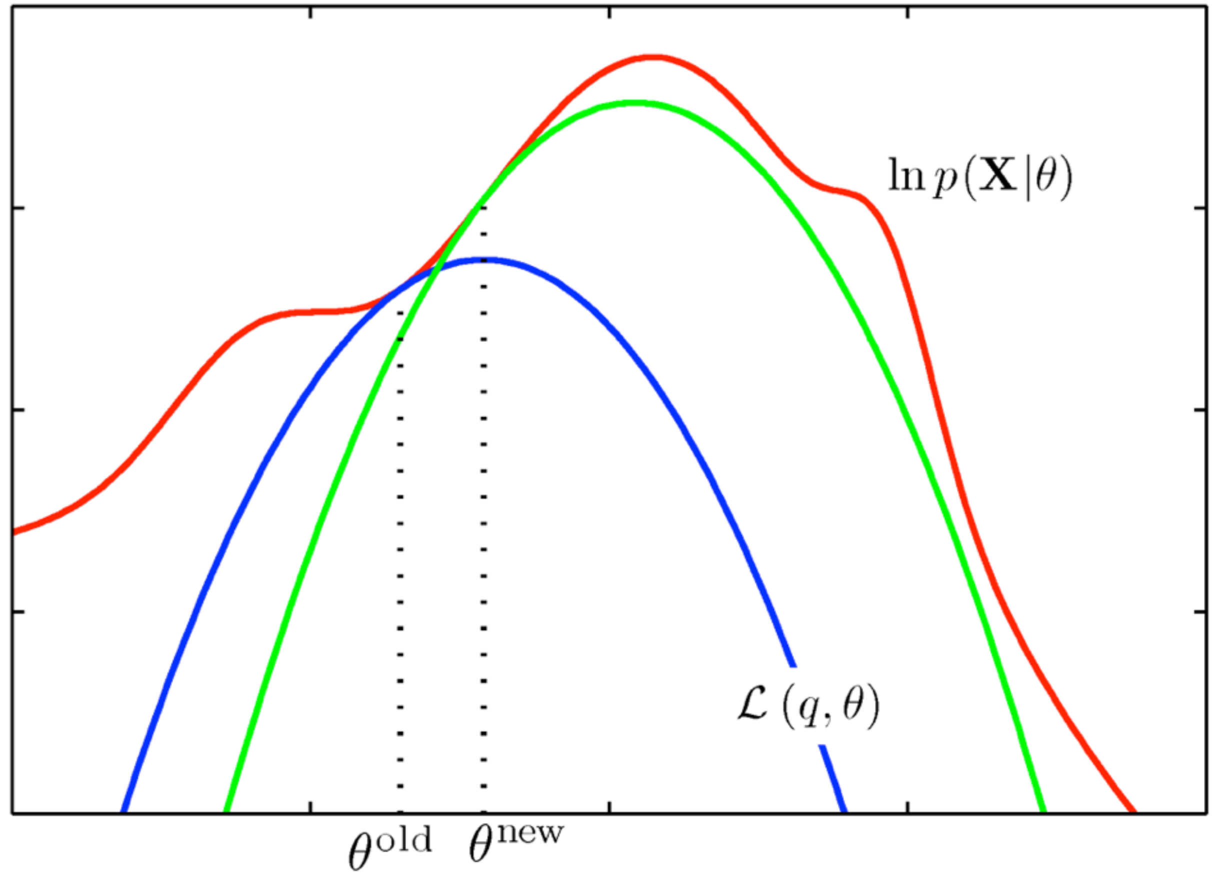


- In the M-step we keep q fixed and find new θ

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

- 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: $q(Z) \approx p(Z \mid X)$
- 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 $\mathcal{L}(q) = -\mathrm{KL}(q_j \| \tilde{p}(X, Z_j)) + \text{const}$

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

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

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.

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

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

Thus, we can say

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

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:

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

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

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 $\mathrm{KL}(q\|p)$. But: we can also minimize $\mathrm{KL}(p\|q)$. Assume q is from the **exponential family**:

natural parameters

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

normalizer

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

Then we have:

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

Expectation Propagation

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

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

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

and we are interested in the posterior:

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

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

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

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

The EP Algorithm

- Given: a joint distribution over data and variables

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

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

The EP Algorithm

- find q^{new} that minimizes

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

  using moment matching, including the zeroth order moment:

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

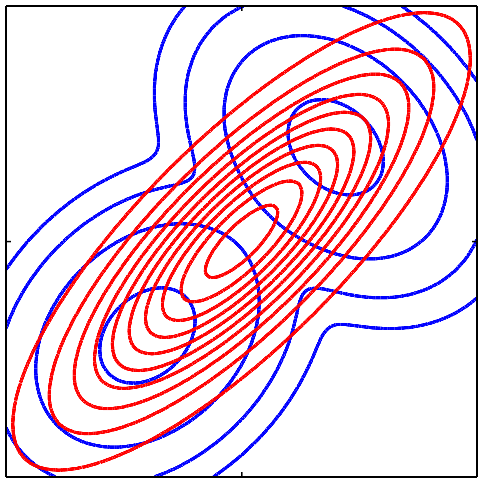
- evaluate the new factor

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

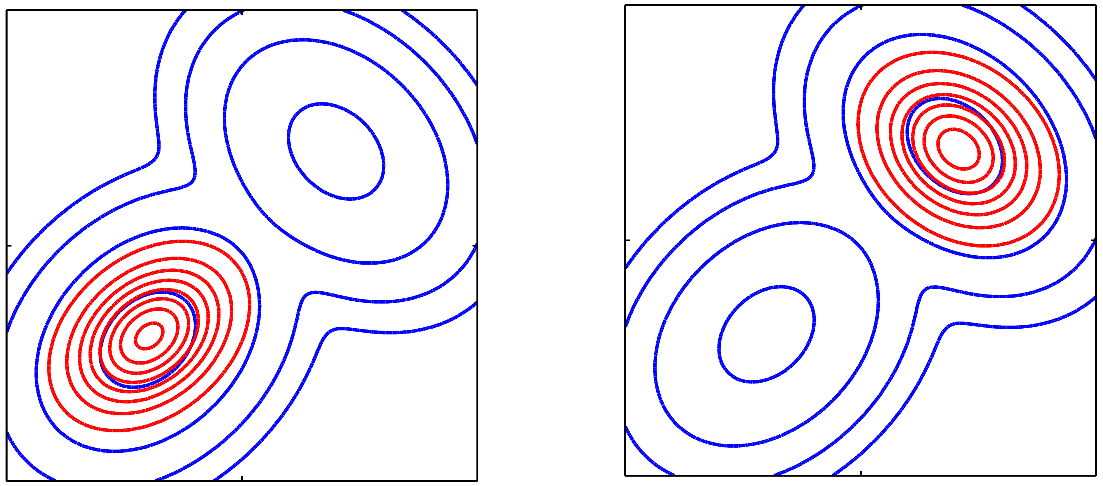
- After convergence, we have $p(\mathcal{D}) \approx \int \prod_i \tilde{f}_j(\boldsymbol{\theta}) d\boldsymbol{\theta}$

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 $KL(q\|p)$

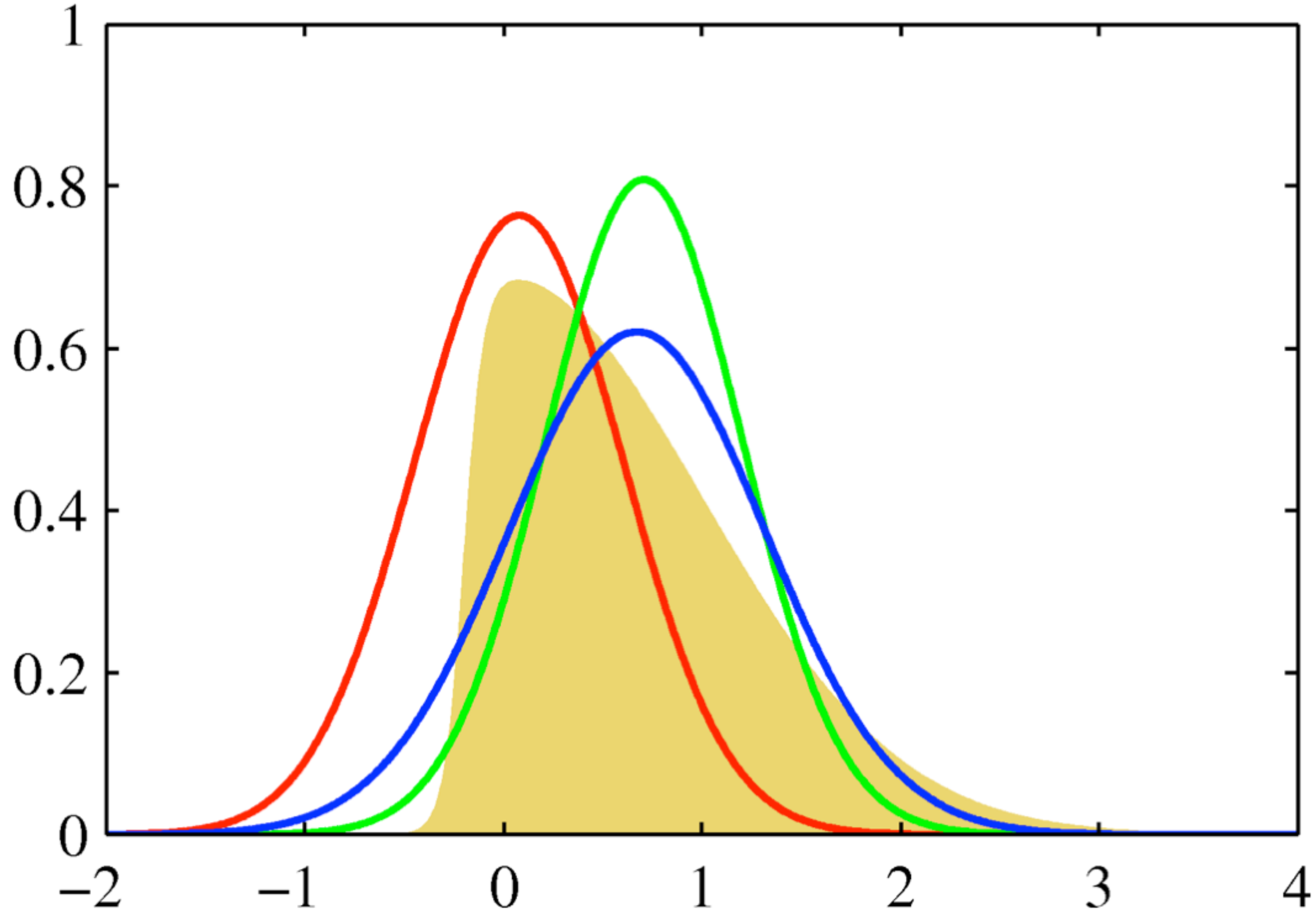


$KL(p\|q)$



$KL(q\|p)$

Example



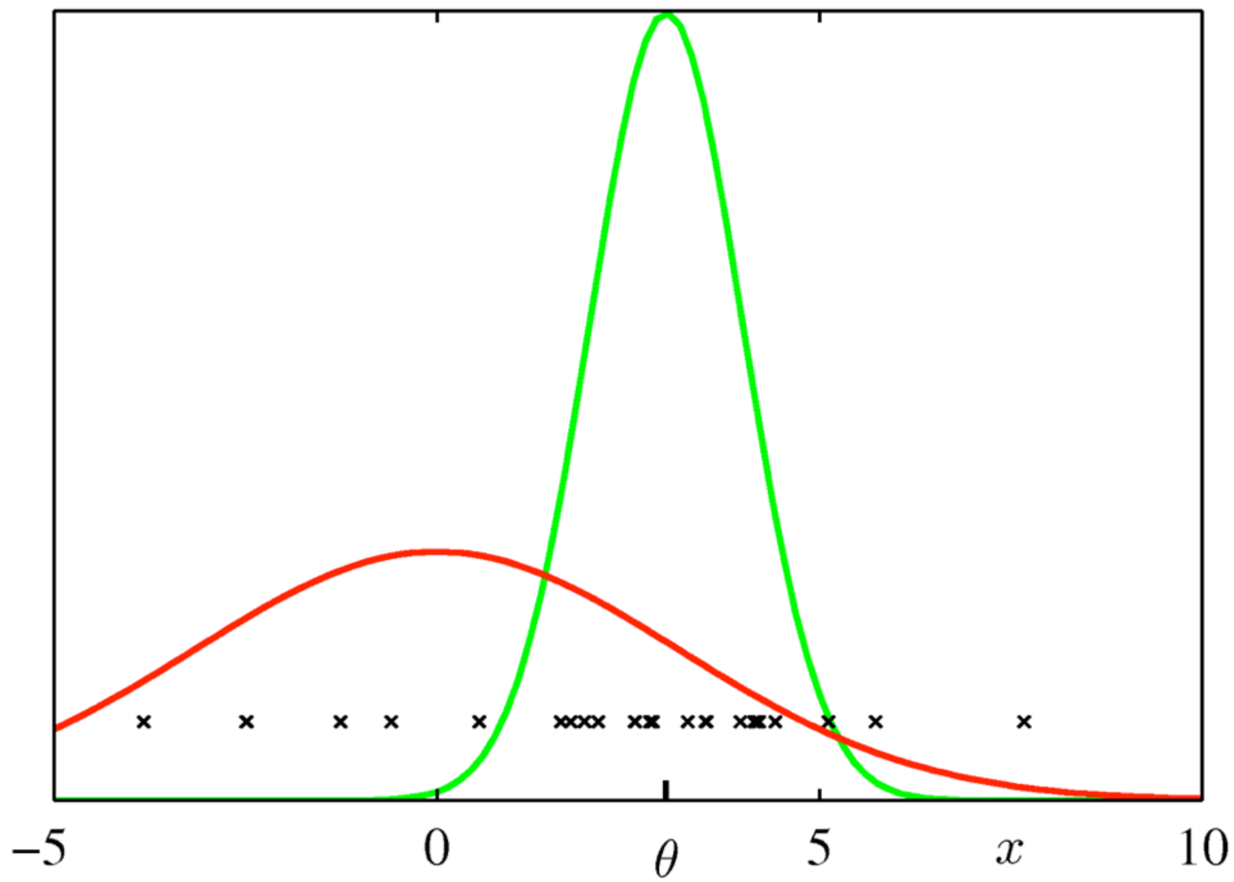
yellow: original distribution

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 $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is

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

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

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

the factors are (unnormalized) Gaussians:

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

EP for the Clutter Problem

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

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

EP for the Clutter Problem

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

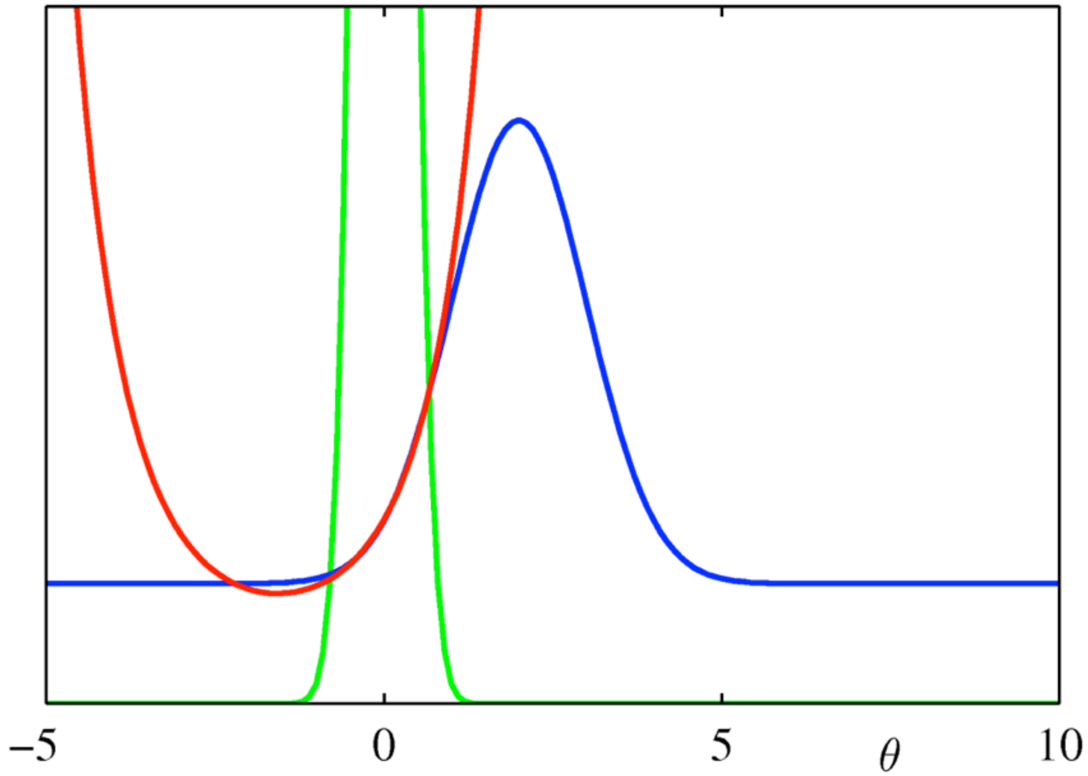
$$q_{-n}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})} \qquad q_{-n}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_{-n}, v_{-n} I)$$

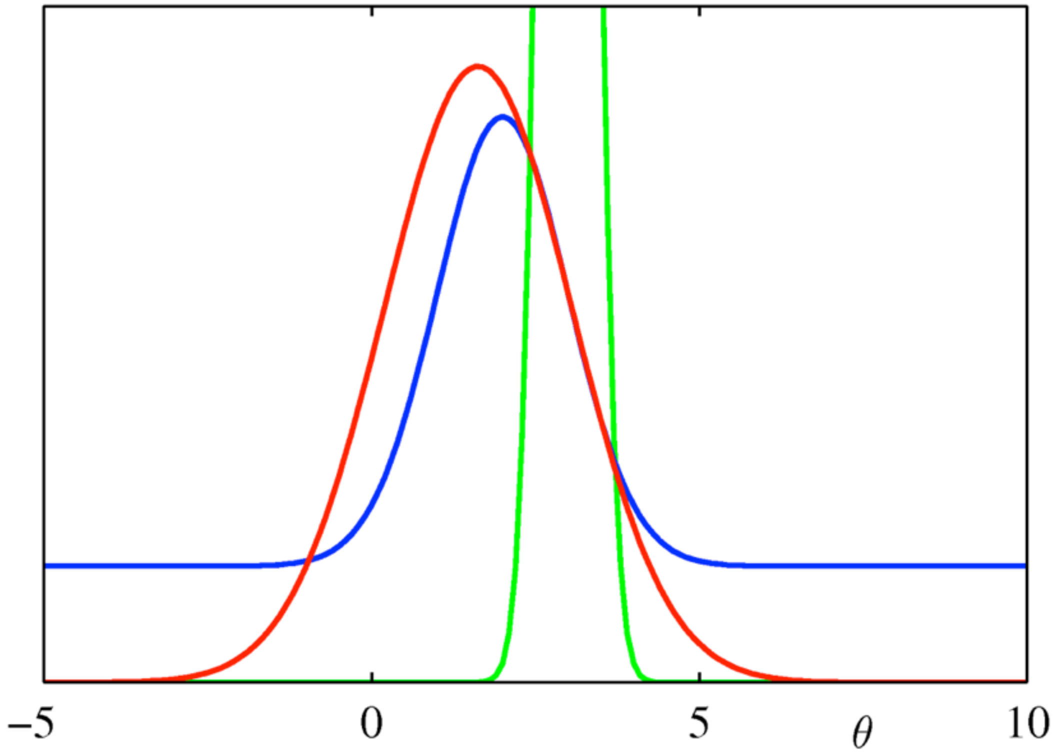
  - Compute the normalization constant:

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

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

A 1D Example





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

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

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