9. 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 (next week) and variational methods.
- In variational optimization, we seek a tractable distribution q that approximates the posterior.
- Optimization is done using functionals.

Some Basics Beforehand

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. Each has the probability:

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

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

Some Basics Beforehand

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 three times “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:

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

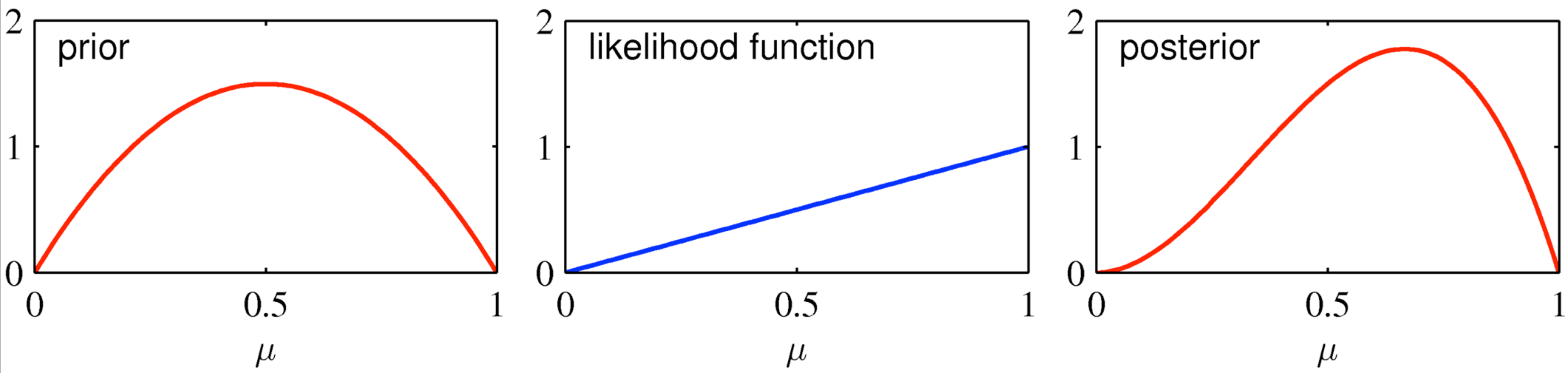
$$l = N - m$$

This gives another beta-distribution:

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

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

A Simple Example



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

- 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:

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

Back to 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: $\mathrm{KL}(q \| p) \neq \mathrm{KL}(p \| q)$

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

This follows from Jensen’s inequality

Mean Field Theory

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

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

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

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:

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

Mean Field Theory

Therefore, the optimal solution in general is

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

In words: the log of the optimal solution for a factor q_i 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

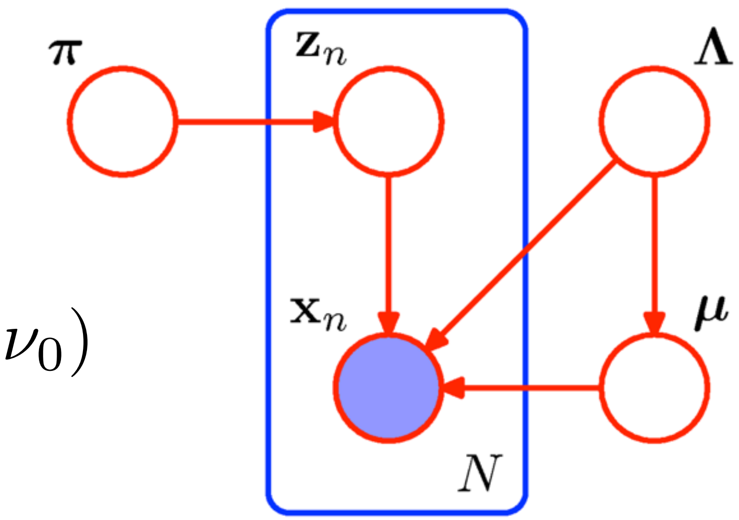
- Again, we have observed data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and latent variables $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Furthermore we have

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

- 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:

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

- From this it can be shown 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 conditional distribution of Z.

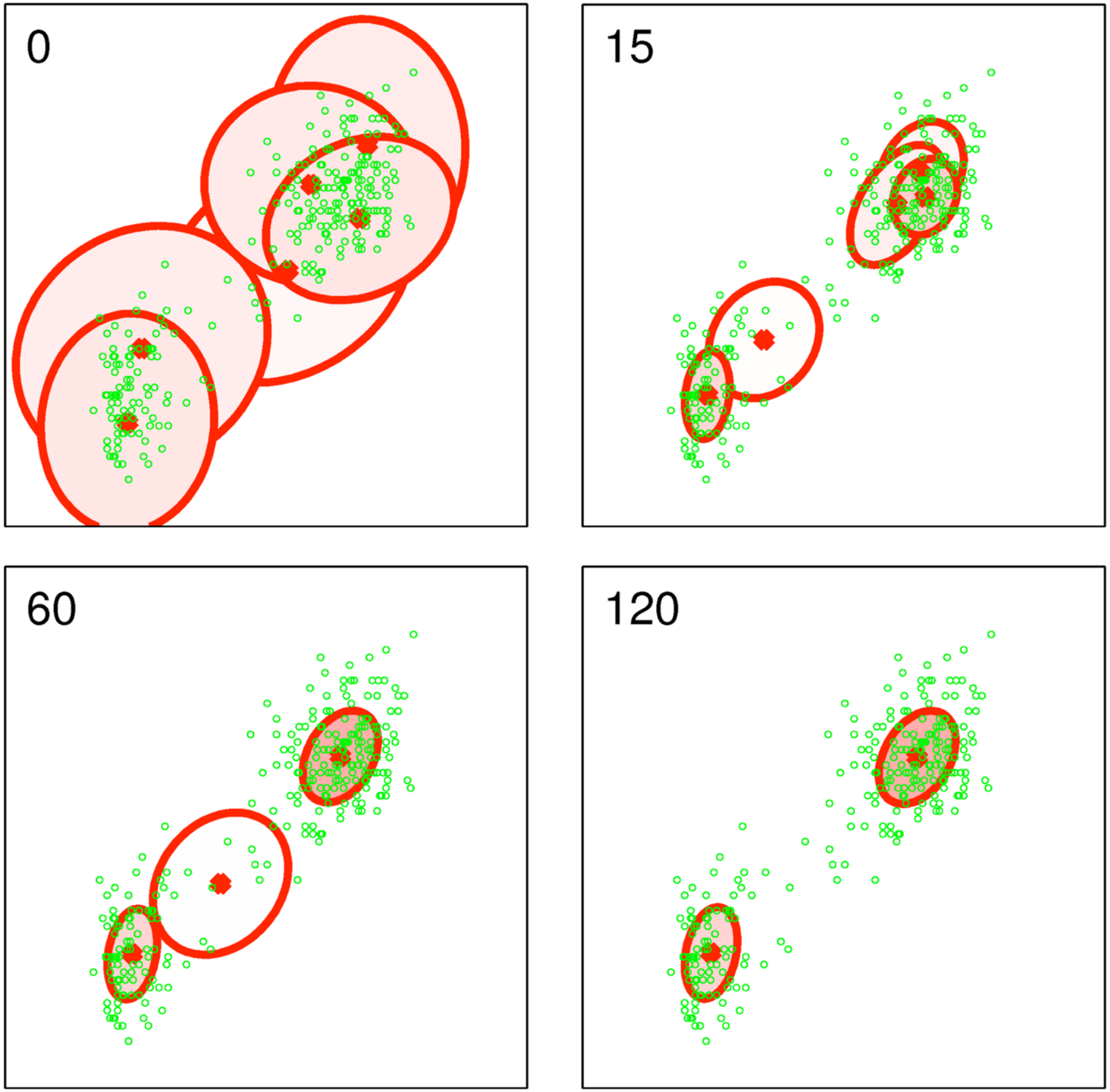
It turns out, this is true for all factors.

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.

The Same Example Again

- 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

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:

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

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 $\boldsymbol{\eta}$

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

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 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 factors in turn.

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 zero-th moment:

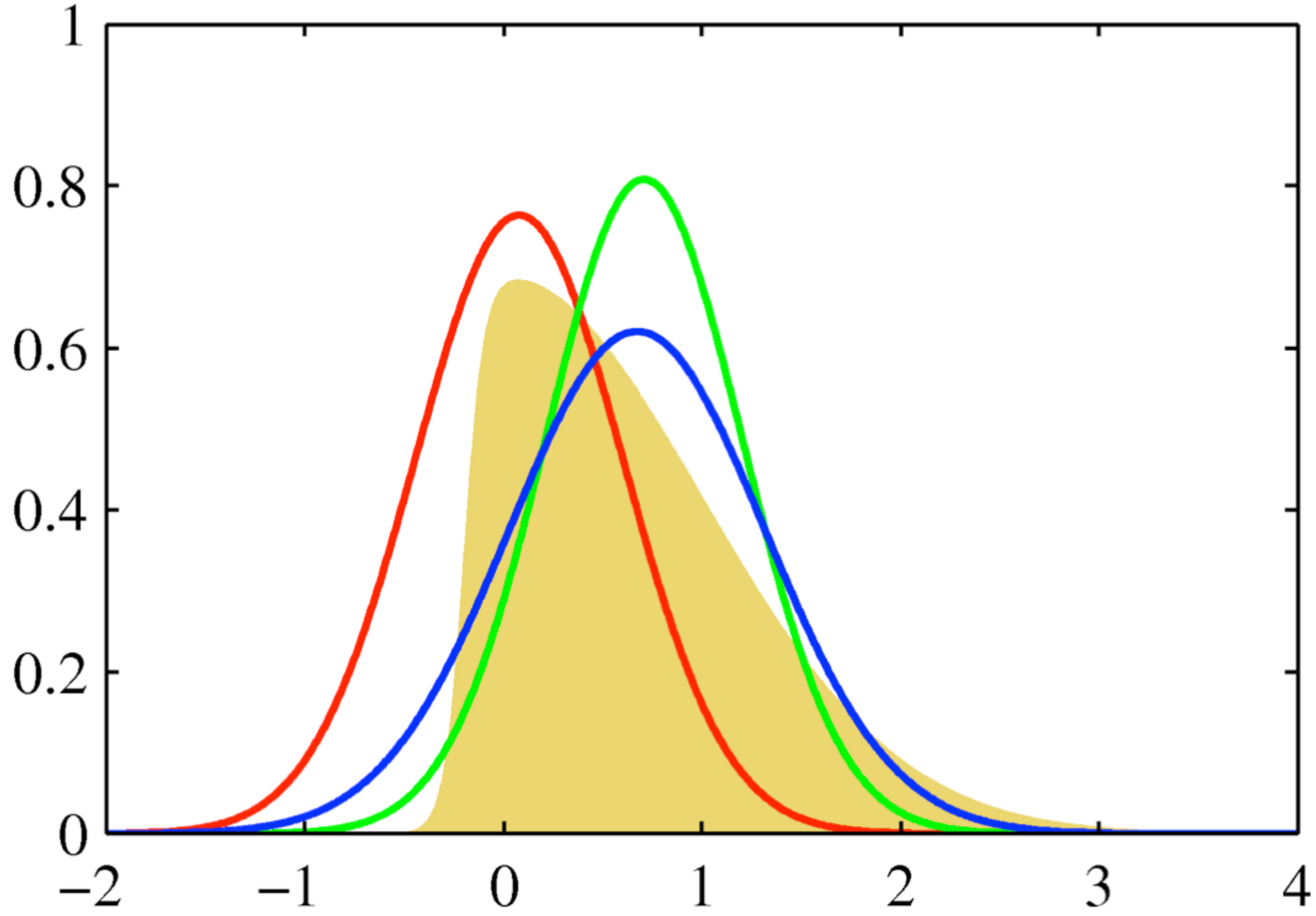
$$Z_j = \int q^{\backslash j}(\boldsymbol{\theta})f_j(\boldsymbol{\theta})d\boldsymbol{\theta}$$

- 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 $p(\mathcal{D}) \approx \int \prod_i \tilde{f}_j(\boldsymbol{\theta})d\boldsymbol{\theta}$

Example



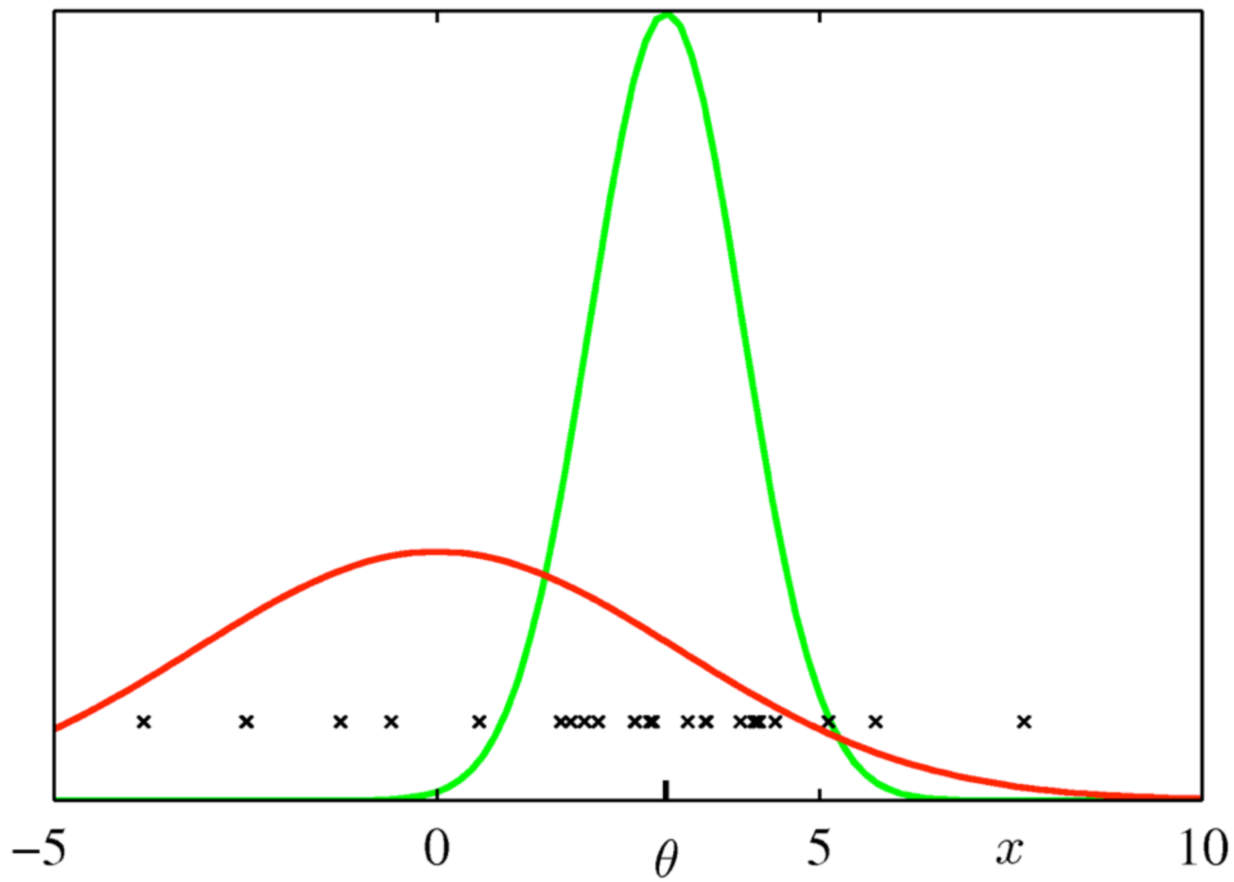
yellow: original distribution

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:

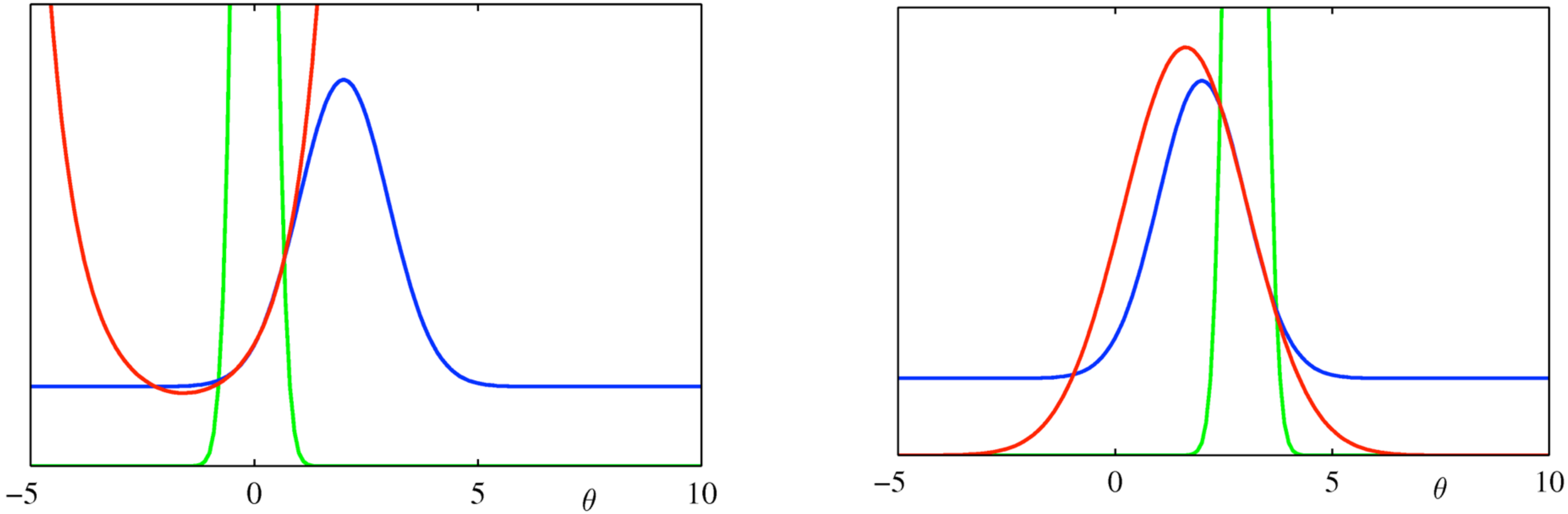
$$q_{-n}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})} \qquad q_{-n}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_{-n}, v_{-n}I)$$

  - 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
- Variational inference for GMMs reduces the risk of overfitting; it is essentially an EM-like algorithm
- Expectation propagation minimizes the reverse KL-divergence of a single factor by moment matching; factors are in the exp. family