Variational Inference - 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

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

Exponential Families

Example: Normal-Distribution with parameters μ and σ

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \qquad \mathbf{u}(x) = (x, x^2)^T$$

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

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

$\mathbb{E}[\mathbf{u}(\mathbf{x})]$ is the vector of **moments.**

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

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

normalizer

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

Then we have:

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

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

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

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: $q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta})}$

The EP Algorithm

- find q^{new} that minimizes

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

  using moment matching, including the zeroth order moment:

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

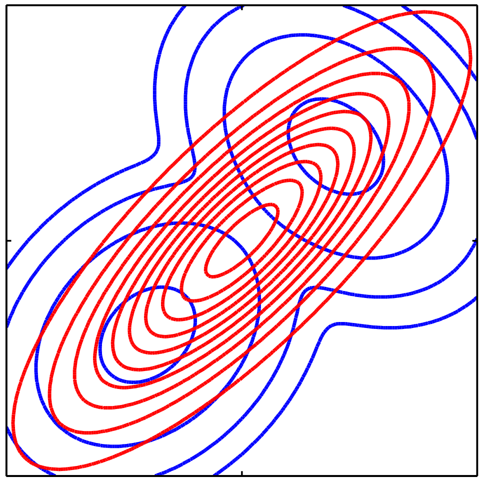
- evaluate the new factor

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

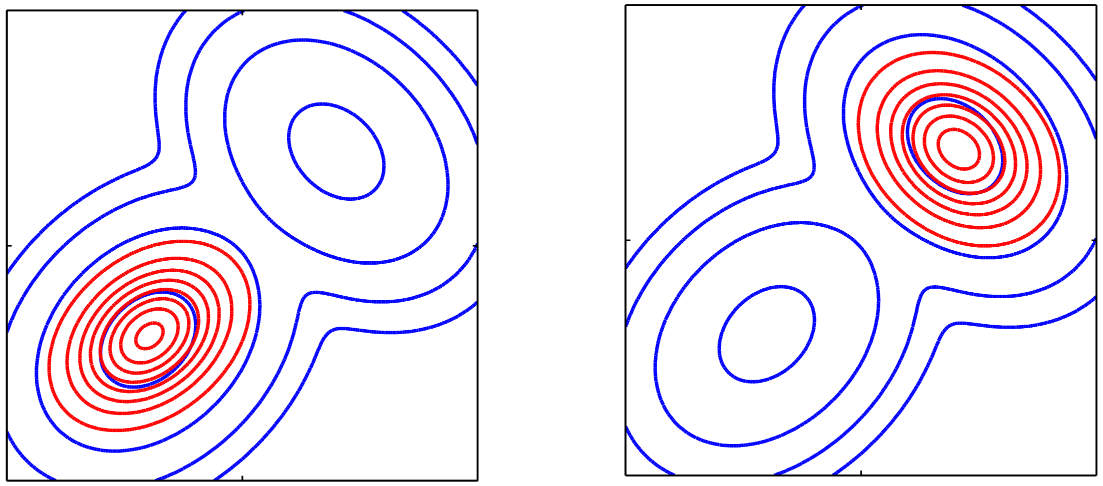
- After convergence, we have $p(\mathcal{D}) \approx \int \prod_i \tilde{f}_j(\boldsymbol{\theta})d\boldsymbol{\theta}$

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 $KL(q\|p)$

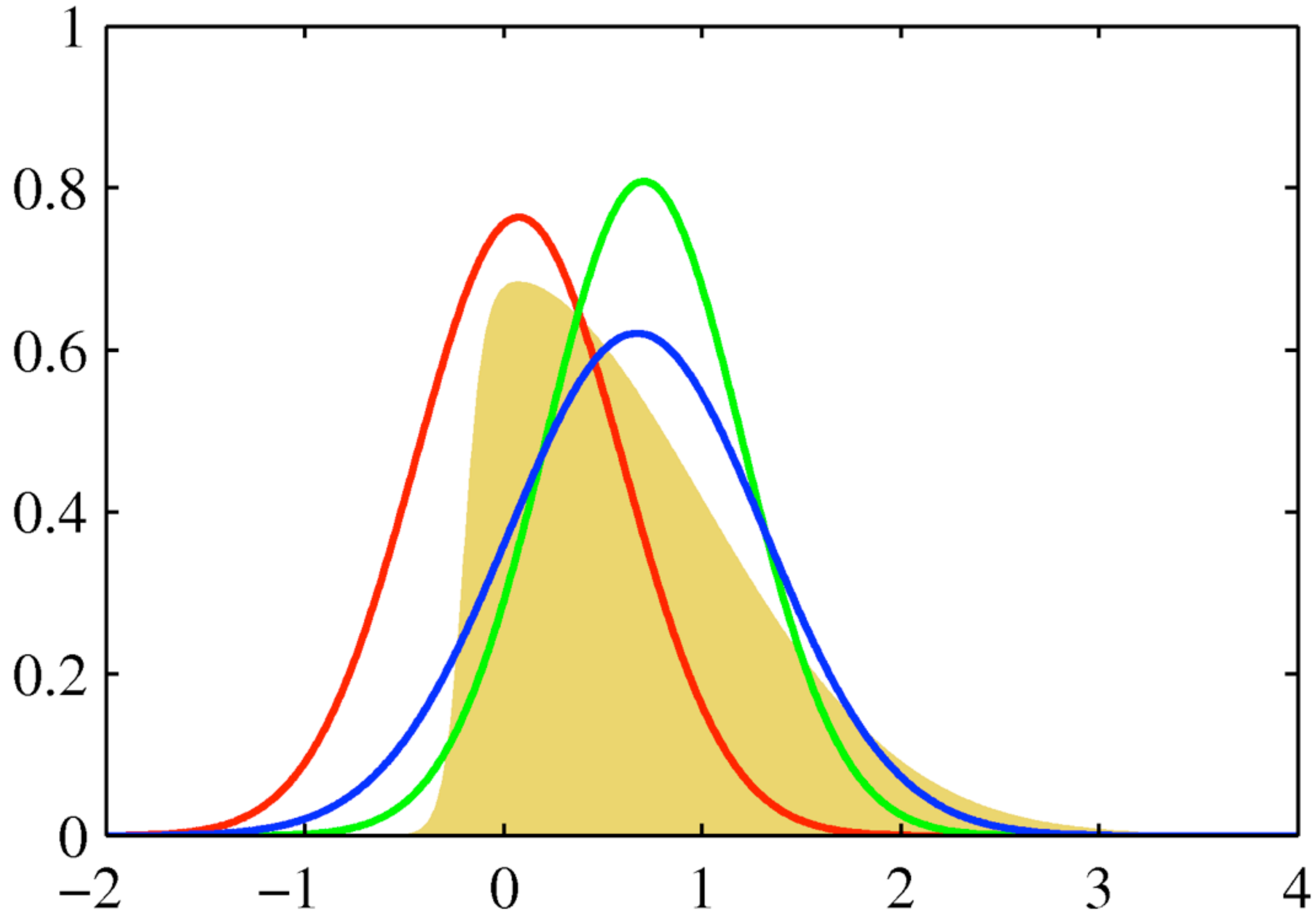


$KL(p\|q)$



$KL(q\|p)$

Example



yellow: original distribution

red: Laplace approximation

green: global variation

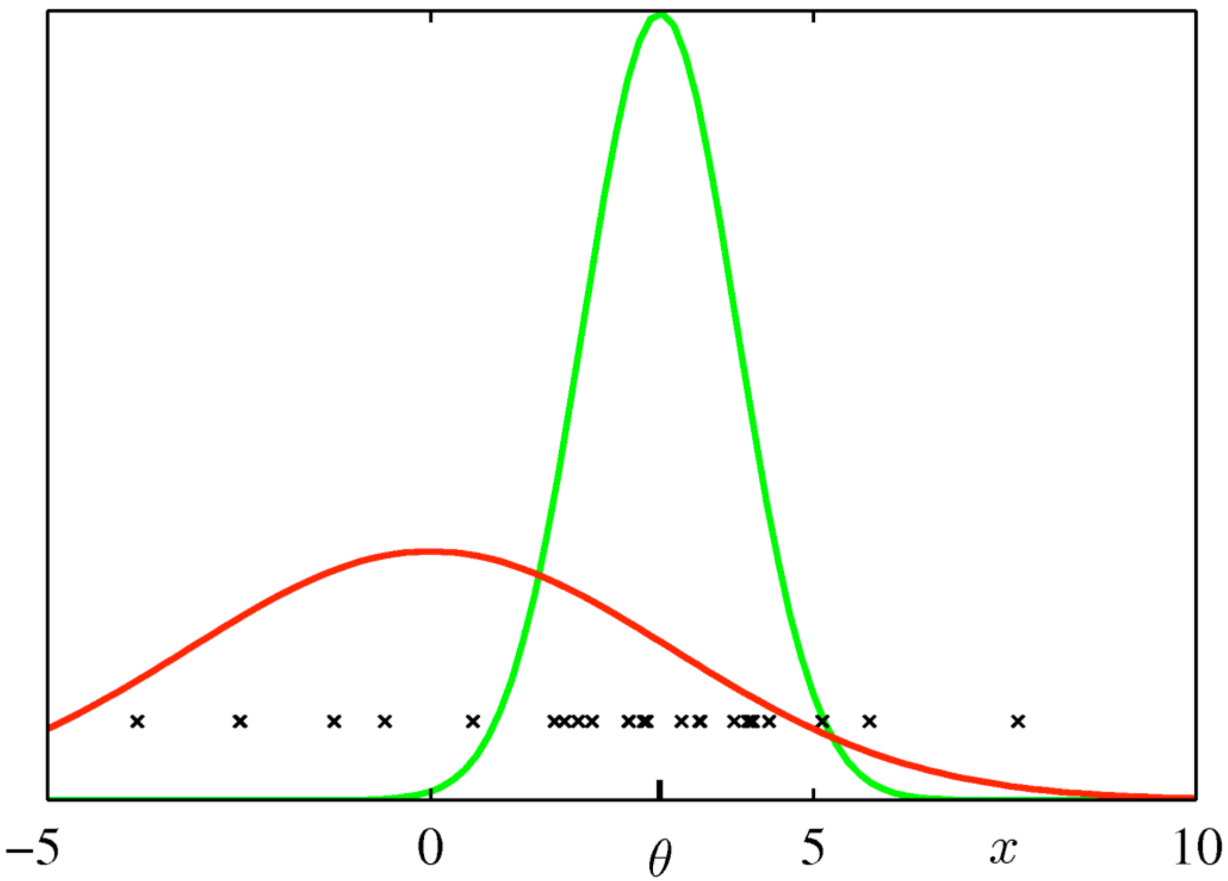
blue: expectation-propagation

Remember: GP Classification

$$p(\mathbf{f} \mid X, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid X)}{p(\mathbf{y} \mid X)}$$

- The likelihood term is not a Gaussian!
- This means, we can not compute the posterior in closed form.
- There are several different solutions in the literature, e.g.:
  - Laplace approximation
  - **Expectation Propagation**
  - Variational methods

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:

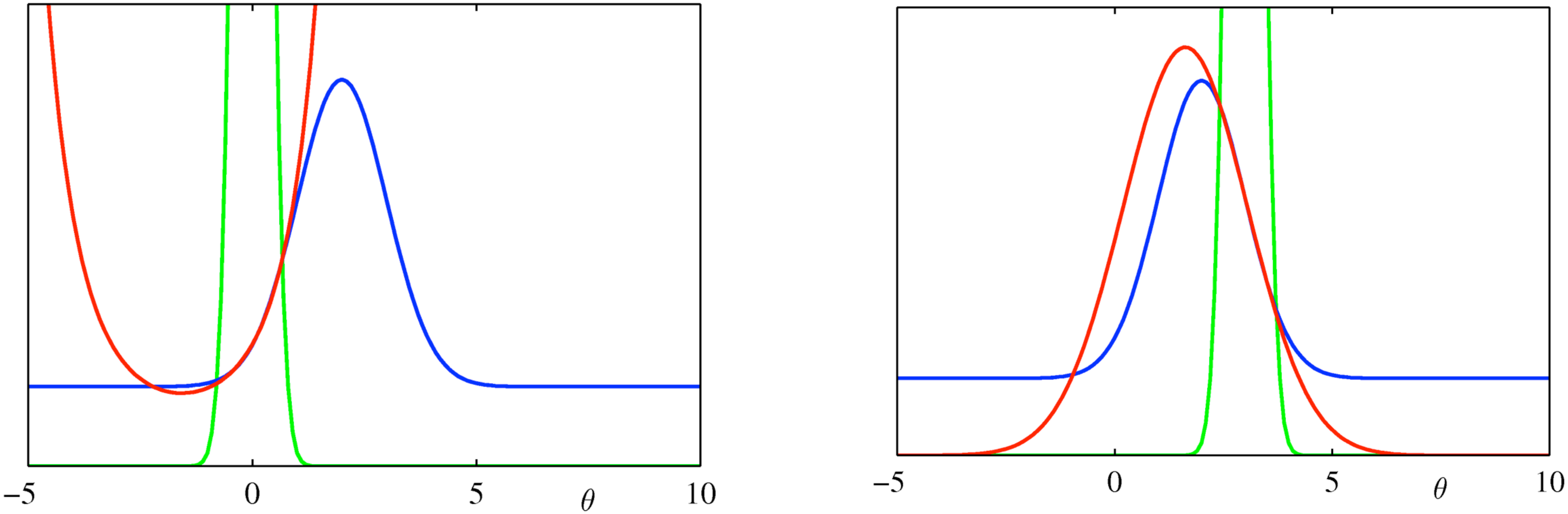
$$q_{-n}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})} \qquad q_{-n}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_{-n}, v_{-n} I)$$

  - Compute the normalization constant:

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

  - Compute mean and variance of $q^{\text{new}} \approx q_{-n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})$
  - 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