Introduction to Expectation Propagation

Antti Honkela
Helsinki University of Technology
Espoo, Finland

http://www.cis.hut.fi/ahonkela/
Contents

• Approximations and distance measures on distributions
• Limitations of naïve mean field variational Bayes (VB)
• Expectation propagation (EP) and the clutter problem
• Belief networks, loopy belief propagation and EP
• An energy function for EP
Background

• Observations $\mathcal{D}$, model $\mathcal{H}$ with parameters $\theta$

• All information of the parameters is contained in the posterior

\[
p(\theta|\mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D}|\theta, \mathcal{H})p(\theta|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})},
\]

where $p(\mathcal{D}|\mathcal{H}) = \int_{\theta} p(\mathcal{D}|\theta, \mathcal{H})p(\theta|\mathcal{H})d\theta$

• Marginalisation principle:

\[
p(x|\mathcal{D}, \mathcal{H}) = \int_{\theta} p(x|\theta, \mathcal{H})p(\theta|\mathcal{D}, \mathcal{H})d\theta
\]

• How to assess possible approximations $q(\theta)$ of the posterior $p(\theta|\mathcal{D}, \mathcal{H})$?

• How to approximate $p(\mathcal{D}|\mathcal{H})$?
Bayesian analysis of approximations

- Choosing the best approximation is a decision problem
- Bayesian method: specify utility, maximise expected utility
- For approximations \( q(\theta) \in Q \) and “true parameter values” \( \theta \in \Omega \), define a score function \( u : Q \times \Omega \rightarrow \mathbb{R} \)
- Expected utility

\[
\bar{u}(q) = \int u(q, \theta)p(\theta|D)d\theta'
\]
Properties of score functions

- The score function is **proper**, if

\[
\sup \bar{u}(q) = \bar{u}(p(\theta | D))
\]

which is attained only if \( q(\theta) = p(\theta | D) \)

- The score function is **local**, if

\[
u(q, \theta) = u_\theta(q(\theta))
\]
Score functions

Example. The quadratic score function

\[ u(q, \theta) = A \left[ 2q(\theta) - \int q(\theta')^2 d\theta' \right] + B(\theta) \]

corresponding to the expected utility

\[ \bar{u}(q) = -\int (q(\theta) - p(\theta|\mathcal{D}))^2 d\theta \]

is a proper, non-local score function
Bayesian analysis of approximations

**Proposition.** Smooth, proper, local score functions are of the form

\[ u(q, \theta) = A \log q(\theta) + B(\theta), \]

where \( A > 0 \) and \( B(\theta) \) are arbitrary.

**Proof.** We maximise the expected utility

\[ \bar{u}(q) = \int u_\theta(q(\theta)) p(\theta|\mathcal{D}) d\theta \]

subject to constraint \( \int q(\theta) d\theta = 1 \). This is done by finding an extremum of

\[ F(q(\cdot)) = \bar{u}(q) - A \left[ \int q(\theta) d\theta - 1 \right]. \]
Proof contd.
A necessary condition for this follows from the variational principle
\[ \frac{\partial}{\partial \alpha} F(q(\cdot) + \alpha \tau(\cdot)) \bigg|_{\alpha=0} = 0 \]
for any function \( \tau : \Omega \rightarrow \mathbb{R} \). this implies a differential equation
\[ u'(q(\theta)) p(\theta|\mathcal{D}) - A = 0, \]
which should hold for \( q(\theta) = p(\theta|\mathcal{D}) \). The solutions of this are
\[ u(q, \theta) = A \log q(\theta) + B(\theta). \]
Bayesian analysis of approximations

**Theorem.** Differences of expected utilities under smooth, proper, local score functions are given by the (scaled) Kullback–Leibler (KL) divergence

\[ A \cdot D_{KL}(p(\theta|D) \mid q(\theta)) = A \int p(\theta|D) \log \frac{p(\theta|D)}{q(\theta)} d\theta. \]

**Proof.** Evaluate \( \bar{u}(p(\theta|D)) - \bar{u}(q(\theta)) \).
Properties of KL divergence

• In information theory, the KL divergence

\[
D_{KL}(p(\theta|D) \parallel q(\theta)) = \int p(\theta|D) \log \frac{p(\theta|D)}{q(\theta)} d\theta
\]

measures the overhead when using distribution \( q \) to code events following \( p \)

• The choice of \( A \) reflects the choice of unit of measure, essentially the base of the logarithm

• Natural logarithm \( \ln \) yields nats, while \( \log_2 \) gives bits
Exponential families

**Definition** A set of distributions with densities

\[ p(\theta | \xi) = \frac{1}{Z(\xi)} \exp(\xi^T \phi(\theta)) \]

is an exponential family with natural parameters \( \xi \), sufficient statistics \( \phi(\theta) \) and partition function \( Z(\xi) \).

Examples: Gaussian, gamma, multinomial, Dirichlet, ...  

**Theorem** For exponential families,

\[ \nabla_\xi \log Z(\xi) = \langle \phi(\theta) \rangle. \]
Properties of the KL divergence

**Theorem.** Given an approximation in an exponential family

\[ q(\theta|\xi) = \frac{1}{Z(\xi)} \exp \left( \xi^T \phi(\theta) \right), \]

the KL divergence \( D_{KL}(p(\theta|\mathcal{D}) \ |\ | q(\theta|\xi)) \) is minimized when

\[ \langle \phi(\theta) \rangle_{p(\theta|\mathcal{D})} = \langle \phi(\theta) \rangle_{q(\theta|\xi)}. \]
**Proof.** Consider

\[
f(\xi) = D_{KL}(p(\theta|\mathcal{D}) \parallel q(\theta|\xi)) = \langle \log p \rangle_p + \langle \log Z(\xi) \rangle_p - \langle \xi^T \phi(\theta) \rangle_p
\]

\[
= \langle \log p \rangle_p + \log Z(\xi) - \xi^T \langle \phi(\theta) \rangle_p.
\]

Zeroing the gradient yields the desired condition, because for exponential families

\[
\nabla_\xi \log Z(\xi) = \langle \phi(\theta) \rangle.
\]

The minimality of the extremum can be checked using the second derivatives.
Properties of the KL divergence

- In VB, the reverse of KL divergence is used:

\[ D_{KL}(q(\theta) \mid \mid p(\theta|\mathcal{D})) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|\mathcal{D})} d\theta. \]

- Having large \( q(\theta) \) with very small \( p(\theta|\mathcal{D}) \) causes large values of the divergence.

- Hence the VB approximation will be contained in the true distribution.
Limitations of naïve mean field variational Bayes

- The marginal likelihoods and especially rankings evaluated by VB are often quite reliable.
- The estimates of the marginals may not be as good, variances can be underestimated.
- Sometimes a simpler mode of solution may be preferred because of inadequate approximation.
Analysis of variational Bayesian ICA
(A. Ilin & H. Valpola)

• Consider the ICA model

\[ x = As + n \]

• Gaussian noise \( n \sim \mathcal{N}(0, \Sigma_x) \)

• Non-Gaussian source prior \( p(s) = \prod p(s_i) \)

• These yield non-diagonal posterior covariance for \( s \):

\[ \Sigma_{s|D}^{-1} \propto \Sigma_s^{-1} + A^T \Sigma_x^{-1} A \]
Limitations of variational Bayes

The form of the true posterior $p(s(t) \mid A, x(t))$

The cost of the posterior and source model misfit

Illustration of the trade-offs between the ICA and PCA solutions.
Limitations of variational Bayes

\[ \nu = 0.6 \quad \nu = 0.7 \]

\[ \nu = 0.9 \quad \nu = 1 \]

VB solutions to ICA problem as a function of non-Gaussianity of the sources
Expectation propagation

- An approximate inference method proposed by Thomas Minka in 2001
- Suitable for approximating product forms

\[ \prod_{i=0}^{N} t_i(\theta) \approx \prod_{i=0}^{N} \tilde{t}_i(\theta) \]

- Iterative refinement of the terms \( \tilde{t}_i(\theta) \)
Expectation propagation

The parameter posterior is

\[
p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})}p(\theta) \prod_{i=1}^{N} p(x_i|\theta)
\]

As a function of \( \theta \), this can be written as

\[
p(\theta) \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=0}^{N} t_i(\theta)
\]

where \( t_0(\theta) = p(\theta) \) and \( t_i(\theta) = p(x_i|\theta) \)
• Now approximate each term separately to get

\[ q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta) \]

• Fit the approximation by finding

\[
\min_{\tilde{t}_i(\theta)} D_{KL}(t_i(\theta) \prod_{j \neq i} \tilde{t}_j(\theta) \parallel \tilde{t}_i(\theta) \prod_{j \neq i} \tilde{t}_j(\theta))
\]
Expectation propagation algorithm

Input $t_0(\theta), \ldots, t_N(\theta)$

Initialise $\tilde{t}_0(\theta) = t_0(\theta)$, $\tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta)$

repeat

  for $i = 0, \ldots, N$ do
  
  Deletion: $q_{i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$
  
  Projection: $\tilde{t}_{i}^{\text{new}}(\theta) \leftarrow \arg \min_{\tilde{t}_i(\theta)} D_{KL}(t_i(\theta)q_{i}(\theta) \| \tilde{t}_i(\theta)q_{i}(\theta))$
  
  Inclusion: $q(\theta) \leftarrow \tilde{t}_{i}^{\text{new}}(\theta)q_{i}(\theta)$
  
  end for

until convergence
Expectation propagation algorithm (2)

Input $t_0(\theta), \ldots, t_N(\theta)$

Initialise $\tilde{t}_0(\theta) = t_0(\theta), \tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta)$

repeat
  for $i = 0, \ldots, N$ do
    Deletion: $q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$
    Inclusion: $q(\theta) \leftarrow \arg \min_{q(\theta)} D_{KL}(t_i(\theta)q_{\setminus i}(\theta) \parallel q(\theta))$
    Update: $\tilde{t}_{i}^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$
  end for

until convergence
The clutter problem

Consider a simple Gaussian mixture for $\mathcal{D} = \{x_i\}_{i=1}^N$

$$p(x|\theta) = w\mathcal{N}(x; \theta, I) + (1 - w)\mathcal{N}(x; 0, 10I)$$

$$p(\theta) = \mathcal{N}(\theta; 0, 100I).$$

A suitable exponential family for this is formed by

$$\mathcal{N}(x; m, vI) = \mathcal{N}(x; \xi)$$

with sufficient statistics $\phi(x) = (x, x^T x)$, natural parameters $\xi = (v^{-1}m, -\frac{1}{2}v^{-1})$ and normalisation $Z(\xi) = (2\pi v)^{d/2} \exp(\frac{1}{2v}m^T m)$. 

**Expectation propagation algorithm**

Input $t_0(\theta), \ldots, t_N(\theta)$

Initialise $\tilde{t}_0(\theta) = t_0(\theta), \tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta)$

repeat

for $i = 0, \ldots, N$ do

Deletion: $q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$

Inclusion: $q(\theta) \leftarrow \arg\min_{q(\theta)} D_{KL}(t_i(\theta) q_{\setminus i}(\theta) \parallel q(\theta))$

Update: $\tilde{t}_i^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$

end for

until convergence
EP for the clutter problem (1):
Initialisation

For the clutter problem, we have

\[ t_0(\theta) = p(\theta) \]
\[ t_i(\theta) = p(x_i|\theta), \quad i = 1, \ldots, N. \]

The approximation is of the form

\[ \tilde{t}_0(\theta) = t_0(\theta) = p(\theta) \]
\[ \tilde{t}_i(\theta) = s_i \exp(\xi_i^T \phi(\theta)), \quad i = 1, \ldots, N, \]
\[ q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta) = sN(\theta; \xi) \]

Now initialise \( \xi_i = 0 \) for \( i = 1, \ldots, N. \)
**Expectation propagation algorithm**

Input $t_0(\theta), \ldots, t_N(\theta)$

Initialise $\tilde{t}_0(\theta) = t_0(\theta), \tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta)$

repeat
  for $i = 0, \ldots, N$ do
    Deletion: $q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$
    Inclusion: $q(\theta) \leftarrow \arg \min_{q(\theta)} D_{KL}(t_i(\theta)q_{\setminus i}(\theta) \parallel q(\theta))$
    Update: $\tilde{t}_i^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$
  end for
until convergence
EP for the clutter problem (2): Deletion

When working with natural parameters, the deletion operation

\[ q_i(\theta) \propto \frac{q(\theta)}{t_i(\theta)} \]

is trivial to implement with

\[ \xi_i = \xi - \xi_i. \]
Expectation propagation algorithm

Input $t_0(\theta), \ldots, t_N(\theta)$
Initialise $\tilde{t}_0(\theta) = t_0(\theta), \tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^N \tilde{t}_i(\theta)$
repeat
  for $i = 0, \ldots, N$ do
    Deletion: $q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$
    Inclusion: $q(\theta) \leftarrow \arg\min_{q(\theta)} D_{KL}(t_i(\theta)q_{\setminus i}(\theta) \parallel q(\theta))$
    Update: $\tilde{t}_i^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$
  end for
until convergence
EP for the clutter problem (3): Inclusion

The inclusion operation:

\[
q(\theta) \leftarrow \arg \min_{q(\theta)} D_{KL}(t_i(\theta)q_{\setminus i}(\theta) \parallel q(\theta))
\]

requires matching sufficient statistics of

\[
t_i(\theta)q_{\setminus i}(\theta) = (w\mathcal{N}(x_i; \theta, I) + (1 - w)\mathcal{N}(x_i; 0, 10I))\mathcal{N}(\theta; \xi_{\setminus i})
\]

\[
= \left( w\mathcal{N} \left( \theta; \left( x_i, -\frac{1}{2} \right) \right) \right) + (1 - w)\mathcal{N}(x_i; 0, 10I) \mathcal{N}(\theta; \xi_{\setminus i})
\]

\[
= w \frac{Z(\xi^+)}{Z ((x_i, -\frac{1}{2})) Z(\xi_{\setminus i})} \mathcal{N}(\theta; \xi^+) + (1 - w)\mathcal{N}(x_i; 0, 10I)\mathcal{N}(\theta; \xi_{\setminus i})
\]

\[
\propto rn(\theta; \xi^+) + (1 - r)\mathcal{N}(\theta; \xi_{\setminus i}),
\]

where \( \xi^+ = \xi_{\setminus i} + (x_i, -\frac{1}{2}) \)
We wish to match the sufficient statistics of the Gaussian mixture

\[ t_i(\theta)q_i(\theta) \propto rN(\theta; \xi^+) + (1 - r)N(\theta; \xi_i). \]

These are simply

\[ \mathbf{m} = rm^+ + (1 - r)m_i \]

\[ v + \mathbf{m}^T \mathbf{m} = r \left( v^+ + (m^+)^T m^+ \right) + (1 - r) \left( v_i + m_i^T m_i \right) \]
Expectation propagation algorithm

Input $t_0(\theta), \ldots, t_N(\theta)$
Initialise $\tilde{t}_0(\theta) = t_0(\theta), \tilde{t}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_{i=0}^{N} \tilde{t}_i(\theta)$

repeat

for $i = 0, \ldots, N$ do

Deletion: $q_{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{t}_i(\theta)} = \prod_{j \neq i} \tilde{t}_j(\theta)$

Inclusion: $q(\theta) \leftarrow \arg \min_{q(\theta)} D_{KL}(t_i(\theta)q_{\setminus i}(\theta) \parallel q(\theta))$

Update: $\tilde{t}_i^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$

end for

until convergence
EP for the clutter problem (4): Update

When working with natural parameters, the update operation

$$\tilde{t}_i^{\text{new}}(\theta) \leftarrow \frac{q(\theta)}{q_{\setminus i}(\theta)}$$

is again trivial with

$$\xi_i = \xi - \xi_{\setminus i}.$$
Marginal likelihood by EP

The EP algorithm may be extended to evaluate the marginal likelihood $p(D|H)$.

To do this, we include a scale on $\tilde{t}_i(\theta)$ and through them for $q(\theta)$:

$$\tilde{t}_i(\theta) = Z_i \frac{q^*(\theta)}{q_{\setminus i}(\theta)},$$

where $q^*(\theta)$ is a normalised version of $q(\theta)$ and $Z_i = \int q_{\setminus i}(\theta) t_i(\theta) d\theta$.

Finally we approximate

$$p(D|H) \approx \int q(\theta) d\theta = \int \prod_i \tilde{t}_i(\theta) d\theta.$$
Marginal likelihood for the clutter problem

For the clutter problem

\[ \tilde{t}_i(\theta) = Z_i \frac{q^*(\theta)}{q_i(\theta)} \]

implies

\[ s_i = Z_i \frac{Z(\xi_{\setminus i})}{Z(\xi)} \]

\[ Z_i = w \frac{Z(\xi^+)}{Z((x_i, -\frac{1}{2})) Z(\xi_{\setminus i})} + (1 - w)N(x_i; 0, 10I). \]

And globally

\[ p(D|\mathcal{H}) \approx \int \prod_i \tilde{t}_i(\theta) d\theta = \frac{Z(\xi)}{Z(\xi_0)} \prod_{i=1}^N s_i \]
EP for belief networks

- A probabilistic model may be represented as a directed graph corresponding to a factorisation of the joint distribution

\[ p(x) = \prod_{x_i \in x} p(x_i | \text{parents}(x_i)) \]

- Derive an EP algorithm using the term factorisation

\[ t_i(x) = p(x_i | \text{parents}(x_i)) \]

and a factorial posterior approximation

\[ q(x) = \prod_{k} q_k(x_k) \]
• For each term $t_i(x)$ the factorisation implies a factorial approximation

$$
\tilde{t}_i(x) = \prod_{k \in \{i, \text{pa}(i)\}} \tilde{t}_{ik}(x_k)
$$

• Equivalently, for each factor $q_k(x_k)$, this corresponds to a regular EP approximation

$$
q_k(x_k) = \prod_{i \in \{i, \text{ch}(i)\}} \tilde{t}_{ik}(x_k),
$$
EP for belief networks

Input $t_1(x), \ldots, t_N(x)$
Initialise $\tilde{t}_{ik}(x_k) = 1$, $q_k(x_k) = \prod_i \tilde{t}_{ik}(x_k)$
repeat
  for $i = 1, \ldots, N$ do
    for all $k$ do
      Deletion: $q_{\setminus i,k}(x_k) \propto \frac{q_k(x_k)}{\tilde{t}_{ik}(x_k)} = \prod_{j \neq i} \tilde{t}_{jk}(x_k)$
    end for
  for all $k$ do
    Projection: $\tilde{t}_{ik}^{\text{new}}(x_k) \leftarrow \sum_{x \setminus x_k} t_i(x) \prod_{j \neq k} q_{\setminus i,j}(x_j)$
    Inclusion: $q_k(x_k) \leftarrow \tilde{t}_{ik}^{\text{new}}(x_k) q_{\setminus i,k}(x_k)$
  end for
end for
until convergence
EP for belief networks (T. Minka)

\[ \tilde{t}_{ik}(x_k) = \sum_{x_i, x_j} p(x_i | x_k, x_j)q_i(x_i)q_j(x_j) \]
EP for belief networks (T. Minka)

\[ \tilde{t}_{ii}(x_i) = \sum_{x_k, x_j} p(x_i | x_k, x_j) q_k(x_k) q_j(x_j) \]
EP for belief networks

- The presented EP algorithm is equivalent to a well-known method called (loopy) belief propagation
- For tree structured graphs, it converges in one pass to yield correct marginals
- For general graphs there are no guarantees and it may even diverge
EP for belief networks

- The EP formulation allows simple generalisation to more accurate approximations
- Use fewer more complicated terms $t_i(x)$
- Factorisation $q(x) = \prod_k q_k(x_k)$ over nodes can still be assumed to only evaluate the marginals
An energy function for EP

- Assume an approximation in an exponential family $\exp(\lambda^T \phi(\theta))$
- With an exact prior,
  
  $$q(\theta) \propto p(\theta) \exp(\nu^T \phi(\theta))$$

  and

  $$q_{\setminus i}(\theta) = p(\theta) \exp(\lambda_i^T \phi(\theta))$$

- Let $N$ be the number of terms $t_i(\theta)$
• Now, EP fixed points correspond to stationary points of the objective

\[ \min_{\nu} \max_{\lambda} \ (N - 1) \log \int p(\theta) \exp(\nu^T \phi(\theta)) d\theta \]

\[ - \sum_{i=1}^{N} \log \int t_i(\theta) p(\theta) \exp(\lambda_i^T \phi(\theta)) d\theta \]

such that \((N - 1)\nu_j = \sum_i \lambda_{ij}\).

• Note: non-convex optimisation problem

• Also other formulations for the energy function
Summary

- Kullback–Leibler divergence $D_{KL}(p(\theta|\mathcal{D}) || q(\theta))$ is a reasonable measure of goodness of approximation.

- EP uses this in a tractable manner to optimise

$$D_{KL}(t_i(\theta)q_{\setminus i}(\theta) || \tilde{t}_i(\theta)q_{\setminus i}(\theta))$$

- Provides good approximations of marginals and marginal likelihood.

- Alternative interpretation to existing belief net algorithms.

- Algorithm may not converge ($\rightarrow$ explicitly minimise the energy?)