Expectation Propagation

José Miguel Hernández-Lobato

Department of Engineering, Cambridge University

April 11, 2013

Introduction

EP can be used to approximate an un-normalized distribution by a simpler parametric distribution, in a similar way as VI.

Also based on the minimization of the KL-divergence, but in its direct way $\mathrm{KL}(p||\mathcal{Q})$ instead of $\mathrm{KL}(\mathcal{Q}||p)$ (the one used by VI).

EP is a generalization of LBP to GM which may contain continuous variables.

The distribution $\mathcal{Q}$ is restricted to belong to a family of probability distributions that is closed under the product operation. This is the exponential family:

$$\mathcal{Q}(\mathbf{z}) = \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{z}) - g(\boldsymbol{\eta})\right) , \qquad g(\boldsymbol{\eta}) = \log \int \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{z})\right) d\mathbf{z}$$

where $\boldsymbol{\eta}$ is a vector of natural parameters of $\mathcal{Q}$, $\mathbf{u}(\mathbf{z})$ are the sufficient statistics and $g(\boldsymbol{\eta})$ is a log normalizer.

Examples of Distributions in the Exponential Family

Gaussian $\mathcal{N}(z|\mu,\sigma^2) = 1/\sqrt{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(z-\mu)^2\right\}$:

$$\boldsymbol{\eta} = (\mu/\sigma^2, -1/(2\sigma^2))^{\mathsf{T}}\,, \qquad \mathbf{u}(z) = (z, z^2)^{\mathsf{T}}\,, \qquad g(\boldsymbol{\eta}) = \frac{1}{2}\log\frac{\pi}{-\eta_2} - \frac{\eta_1^2}{4\eta_2}\,.$$

Multinomial for a single observation $p(\mathbf{z}) = \prod_{k=1}^{M}\mu_k^{z_k}$:

$$\boldsymbol{\eta} = (\log\mu_1, \ldots, \log\mu_M)^{\mathsf{T}}\,, \qquad \mathbf{u}(\mathbf{z}) = \mathbf{z}\,, \qquad g(\boldsymbol{\eta}) = 0\,.$$

Bernoulli $\mathrm{Bern}(z|\mu) = \mu^z(1-\mu)^{1-z}$:

$$\eta = \log\left(\frac{\mu}{1-\mu}\right)\,, \qquad u(z) = z\,, \qquad g(\eta) = \log(1+\exp(\eta))\,.$$

Most of the simplest parametric distributions belong to the exponential family.

KL Divergence Minimization

Consider any distribution $p(\mathbf{z})$ and the KL-divergence between p and $\mathcal{Q}$:

$$\mathrm{KL}(p||\mathcal{Q}) = -\int p(\mathbf{z}) \log \left\{ \frac{\mathcal{Q}(\mathbf{z})}{p(\mathbf{z})} \right\} d\mathbf{z} = g(\boldsymbol{\eta}) - \boldsymbol{\eta}^\mathrm{T} \mathbb{E}_p[\mathbf{u}(\mathbf{z})] + \mathrm{Const}\,.$$

To minimize $\mathrm{KL}(p||\mathcal{Q})$ with respect to the natural parameters $\boldsymbol{\eta}$ we do

$$\frac{\partial \mathrm{KL}(p||\mathcal{Q})}{\partial \boldsymbol{\eta}} = 0 \Longleftrightarrow \frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbb{E}_p[\mathbf{u}(\mathbf{z})]\,, \qquad \frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbb{E}_\mathcal{Q}[\mathbf{u}(\mathbf{z})]\,.$$

Minimizing $\mathrm{KL}(p||\mathcal{Q})$ is equivalent to matching expected sufficient statistics.

If $\mathcal{Q}$ is Gaussian, then we have to match $\mathbb{E}_\mathcal{Q}[\mathbf{z}] = \mathbb{E}_p[\mathbf{z}]$ and $\mathbb{E}_\mathcal{Q}[\mathbf{z}\mathbf{z}^\mathrm{T}] = \mathbb{E}_p[\mathbf{z}\mathbf{z}^\mathrm{T}]$.

This result is systematically exploited in EP to carry out approximate inference.

Problem: computing $\mathbb{E}_p[\mathbf{u}(\mathbf{z})]$ is intractable!

Factorization of the Joint Distribution

In many GMs (mainly those that assume i.i.d. data) the joint distribution $p(\mathbf{z}, \mathbf{e})$ of the observed variables $\mathbf{e}$ and the latent variables $\mathbf{z}$ factorizes as

$$p(\mathbf{z}, \mathbf{e}) = \prod_i f_i(\mathbf{z}),$$

where each factor f_1 depends on $\mathbf{z}$ or a subset of these variables.

The factors f_i can be produced by a likelihood or a prior for $\mathbf{z}$.

Given $p(\mathbf{z}, \mathbf{e})$, the posterior for $\mathbf{z}$ is obtained after normalizing by $p(\mathbf{e})$:

$$p(\mathbf{z}|\mathbf{e}) = \frac{1}{p(\mathbf{e})} \prod_i f_i(\mathbf{z}), \qquad p(\mathbf{e}) = \int \prod_i f_i(\mathbf{z}) d\mathbf{z},$$

The Approximation to the Joint Distribution

EP approximates $p(\mathbf{z}, \mathbf{e})$ using a product of simpler factors:

$$p(\mathbf{z}, \mathbf{e}) = \prod_i f_i(\mathbf{z}) \approx \prod_i \tilde{f}_i(\mathbf{z}) \, .$$

Each approximate factor $\tilde{f}_i$ approximates the corresponding exact factor f_i.

The $\tilde{f}_i$ are in an exponential family but need not be normalized. For example, the $\tilde{f}_i$ can be unnormalized Gaussians.

Because the exponential family is closed under the product operation, the product of the $\tilde{f}_i(\mathbf{z})$ has a simple form and can be easily normalized:

$$p(\mathbf{z}|\mathbf{e}) = \frac{1}{p(\mathbf{e})} \prod_i f_i(\mathbf{z}) \approx \frac{1}{Z} \prod_i \tilde{f}_i(\mathbf{z}) = \mathcal{Q}(\mathbf{z}) \, ,$$

where $Z = \int \prod_i \tilde{f}_i(\mathbf{z}) d\mathbf{z}$ approximates $p(\mathbf{e})$, the model evidence. Importantly, $\mathcal{Q}$ has the same form as the approximate factors $\tilde{f}_i$.

Updating the Approximate Factors I

How do we adjust the parameters of the approximate factors $\tilde{f}_i$?

Ideally, we would like to minimize $\mathrm{KL}(p||\mathcal{Q})$. However, this involves computing averages with respect to the exact posterior which is intractable. EP minimizes the KL divergence between f_j and $\tilde{f}_j$ in the context of all the other approximate factors $\tilde{f}_i$, $i \neq j$. This ensures that $\tilde{f}_j$ is accurate where $\mathcal{Q}^{\setminus j} = \prod_{i \neq j} \tilde{f}_i$ takes large values.

To refine $\tilde{f}_j$, we first remove it from $\mathcal{Q}$: $\mathcal{Q}^{\setminus j}(\mathbf{z}) \propto \prod_{i \neq j} \tilde{f}_i(\mathbf{z}) = \mathcal{Q}(\mathbf{z}) / \tilde{f}_j(\mathbf{z})$.

We then adjust $\tilde{f}_j$ so that the distributions

$$\mathcal{Q}_{\mathrm{new}}(\mathbf{z}) \propto \tilde{f}_j(\mathbf{z}) \mathcal{Q}^{\setminus j}(\mathbf{z}) \quad \text{and} \quad \hat{p}(\mathbf{z}) = \frac{1}{Z_j} f_j(\mathbf{z}) \mathcal{Q}^{\setminus j}(\mathbf{z}) \,, \quad Z_j = \int f_j(\mathbf{z}) \mathcal{Q}^{\setminus j}(\mathbf{z}) d\mathbf{z} \,,$$

are as close as possible in terms of the KL divergence, where $\mathcal{Q}^{\setminus j}$ is kept fixed.

Updating the Approximate Factors II

First, we minimize $\mathrm{KL}(Z_j^{-1} f_j(\mathbf{z}) \mathcal{Q}^{\backslash j}(\mathbf{z}) || \mathcal{Q}_{\text{new}}(\mathbf{z}))$ with respect to $\mathcal{Q}_{\text{new}}$.

Done by matching expected sufficient statistics between $\mathcal{Q}_{\text{new}}$ and $1/Z_j f_j \mathcal{Q}^{\backslash j}$.
For this, expectations with respect to $1/Z_j f_j \mathcal{Q}^{\backslash j}$ must be tractable.

Then $\tilde{f}_j$ is updated using

$$\tilde{f}_j(\mathbf{z}) = Z_j \frac{\mathcal{Q}_{\text{new}}(\mathbf{z})}{\mathcal{Q}^{\backslash j}(\mathbf{z})}\,, \qquad \text{recall that} \quad \mathcal{Q}_{\text{new}} \propto \tilde{f}_j(\mathbf{z}) \mathcal{Q}^{\backslash j}(\mathbf{z})\,,$$

which ensures that $\tilde{f}_j(\mathbf{z}) \mathcal{Q}^{\backslash j}(\mathbf{z})$ and $f_j(\mathbf{z}) \mathcal{Q}^{\backslash j}(\mathbf{z})$ integrate the same.

Several passes are made trough the factors until they converge.

The model evidence is approximated by the normalizing constant of the product of all the $\tilde{f}_i$.

The Expectation Propagation Algorithm

Computes $\mathcal{Q}$ and an approximation to the model evidence.

1. Initialize $\mathcal{Q}$ and each $\tilde{f}_i$ to be uniform.
2. Repeat until convergence of the $\tilde{f}_i$:
    1. Choose a factor $\tilde{f}_j$ to refine.
    2. Remove $\tilde{f}_j$ from $\mathcal{Q}$ by division $\mathcal{Q}^{\setminus j} = \mathcal{Q}/\tilde{f}_j$.
    3. Compute Z_j and find $\mathcal{Q}_{\text{new}}$ by minimizing $\text{KL}(\hat{p}||\mathcal{Q}_{\text{new}})$.
    4. Compute and store the new factor $\tilde{f}_j = Z_j \mathcal{Q}_{\text{new}}/\mathcal{Q}^{\setminus j}$.
3. Evaluate the approximation to the model evidence:

$$p(\mathbf{e}) \approx Z = \int \prod_i \tilde{f}_j(\mathbf{z}) d\mathbf{z} \,.$$

A simplification is known as assumed density filtering (ADF).

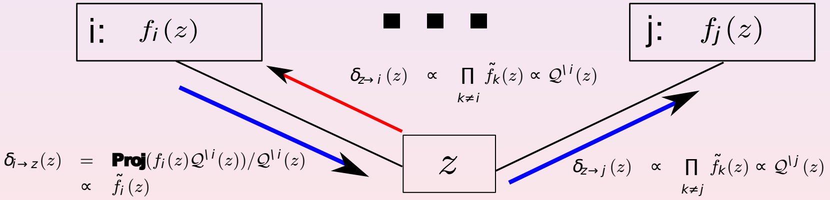
In ADF only one pass is done for each factor (faster but less accurate).

Expectation Propagation as a Message Passing Algorithm

EP is a generalization of LBP with **approximate messages** in a cluster graph, often the Bethe cluster graph. If there is no approximation they are **equivalent**.

In **LBP** the messages are **factors** (the product of factors is **another factor**). **EP** keeps messages consistent by **projecting** to the chosen **exponential family**.

The **approximate messages** sent in EP are the approximate factors $\tilde{f}_i$.



Node i (contains factor f_i) sends a message $\mathcal{Q}^{\setminus j}$ to node j (contains factor f_j) through the **empty node** z. At convergence, the clusters are approximately calibrated and the **product of the messages** in node z give $\mathcal{Q}$. Note the **division** in the computations carried out at node i.

Expectation Propagation: Considerations

- The minimization of the KL is done by moment matching.
- EP may not converge and the $\tilde{f}_i$ may oscillate forever (same as in LBP).
- Convergence can be improved by damping the EP updates.
- As with loopy BP, the convergence points of EP can be shown to be stationary points of a particular energy function which need not be convex. There can be multiple convergence points of EP.
- It is possible to design convergent versions of EP that directly attempt to optimize the energy function. However, they are much more expensive and most times EP converges successfully.
- No need to replace all the factors in the joint distribution with approximations. For example, if one factor is already in the exponential family, the approximate factor is always the same and exact.
- EP considers global aspects of p by approximately minimizing $\mathrm{KL}(p|\mathcal{Q})$.

EP Example: The Clutter Problem

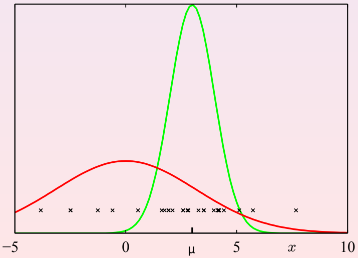
We consider the problem of inferring the mean $\boldsymbol{\mu}$ of a multivariate Gaussian when the Gaussian observations are embedded in background Gaussian clutter.

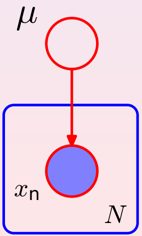
In this problem $\mathbf{z} = \boldsymbol{\mu}$ and $\mathbf{e}$ are the observations $\mathbf{x}$, which are generated from:

$$p(\mathbf{x}|\boldsymbol{\mu}) = (1-w)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{I}) + w\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}a)\,,$$

where $w = 0.5$ is the proportion of clutter and $a = 10$.

The prior for $\boldsymbol{\mu}$ is $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|0, \mathbf{I}b)$ with $b = 100$ (little informative).





Factorization of the Joint Distribution

The joint distribution of $\boldsymbol{\mu}$ and the evidence $\mathbf{e} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is

$$p(\boldsymbol{\mu}, \mathbf{e}) = p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}) = f_0(\boldsymbol{\mu}) \prod_{i=1}^{N} f_i(\boldsymbol{\mu}) \,,$$

a mixture of 2^N terms. Computing $p(\boldsymbol{\mu}|\mathbf{e})$ is intractable for large N.

We choose a parametric form for $\mathcal{Q}$ that belongs to the exponential family:

$$\mathcal{Q}(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, v\mathbf{I}) \,, \qquad \tilde{f}_i(\boldsymbol{\mu}) = \tilde{s}_i \mathcal{N}(\boldsymbol{\mu}|\tilde{\mathbf{m}}_i, \tilde{v}_i \mathbf{I}) \,,$$

with parameters $\mathbf{m}$, $\{\tilde{\mathbf{m}}_i\}_{i=0}^N$, $\{\tilde{s}_i\}_{i=0}^N$, $\{\tilde{v}_i\}_{i=0}^N$ and v.

The $\tilde{f}_i$ are not densities and negative values for $\tilde{v}_i$ are valid.

f_0 can be approximated exactly and the optimal choice for $\tilde{f}_0$ is $\tilde{f}_0 = f_0$.

Once initialized, this term needs not be updated by EP anymore.

Gaussian Identities I

The product and ratio of Gaussians is again Gaussian.

$$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = C\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}, \qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right),$$

$$C = \sqrt{\frac{|\boldsymbol{\Sigma}|}{(2\pi)^d |\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}.$$

$$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) / \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = C\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})^{-1}, \qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right),$$

$$C = \sqrt{\frac{|\boldsymbol{\Sigma}||\boldsymbol{\Sigma}_2|}{(2\pi)^d |\boldsymbol{\Sigma}_1|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}.$$

Gaussian Identities II

Let $f(\mathbf{x})$ be an arbitrary factor of $\mathbf{x}$ and let

$$Z = \int t(\mathbf{x})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\,, \qquad \hat{p}(\mathbf{x}) = \frac{1}{Z} t(\mathbf{x})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\,,$$

Then, we have that

$$\mathbb{E}_{\hat{p}}[\mathbf{x}] = \boldsymbol{\mu} + \boldsymbol{\Sigma}\frac{\partial \log Z}{\partial \boldsymbol{\mu}}\,,$$

$$\mathbb{E}_{\hat{p}}[\mathbf{x}\mathbf{x}^\mathsf{T}] - \mathbb{E}_{\hat{p}}[\mathbf{x}]\mathbb{E}_{\hat{p}}[\mathbf{x}]^\mathsf{T} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\left(\frac{\partial \log Z}{\partial \boldsymbol{\mu}}\left(\frac{\partial \log Z}{\partial \boldsymbol{\mu}}\right)^\mathsf{T} - 2\frac{\partial \log Z}{\partial \boldsymbol{\Sigma}}\right)\boldsymbol{\Sigma}\,.$$

These expressions are very useful to find the parameters of $\mathcal{Q}_{\text{new}}$ in EP.

Initialization and Computation of $\mathcal{Q}^{\backslash i}$

The $\tilde{f}_i$ are initialized to be non-informative, $\mathcal{Q}$ is also non-informative:

$$\tilde{s}_i = (2\pi\tilde{v}_i)^{\frac{D}{2}}\,, \quad \tilde{\mathbf{m}}_i = \mathbf{0}\,, \quad \tilde{v}_i \to \infty\,, \quad \mathbf{m} = \mathbf{0}\,, \quad v = b\,, \qquad \text{for } i = 1, \ldots, N.$$

where we have used the Gaussian identities.

After refining $\tilde{f}_0$, $\mathcal{Q}$ is equal to the prior $p(\boldsymbol{\mu})$.

The first step to refine $\tilde{f}_i$ with $i = 1, \ldots, N$, is to compute $\mathcal{Q}^{\backslash i}$ using

$$\mathcal{Q}^{\backslash i}(\boldsymbol{\mu}) \propto \mathcal{Q}(\boldsymbol{\mu})/\tilde{f}_i(\boldsymbol{\mu}) \propto \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}^{\backslash i}, \mathbf{I}v^{\backslash i})\,,$$

where we use the Gaussian identities again to get

$$\mathbf{m}^{\backslash i} = v^{\backslash i}(\mathbf{m}v^{-1} - \tilde{\mathbf{m}}_i\tilde{v}_i^{-1})\,, \qquad (v^{\backslash i})^{-1} = v^{-1} - \tilde{v}_i^{-1}\,.$$

Computation of the New Posterior $\mathcal{Q}_{\text{new}}$

The first step to update $\tilde{f}_i$ is to compute Z_i:

$$Z_i = \int f_i(\boldsymbol{\mu})\mathcal{Q}^{\setminus i}(\boldsymbol{\mu})d\boldsymbol{\mu} = (1-w)\mathcal{N}(\mathbf{x}_i|\mathbf{m}^{\setminus i}, (v^{\setminus i}+1)\mathbf{I}) + w\mathcal{N}(\mathbf{x}_i|\mathbf{0}, a\mathbf{I})\,.$$

which is obtained from the convolution of two Gaussians.

Next, we compute $\mathcal{Q}_{\text{new}}$ by finding the mean and the variance of $f_i\mathcal{Q}^{\setminus i}$:

$$\mathbf{m}_{\text{new}} = \mathbf{m}^{\setminus i} + \rho_i \frac{v^{\setminus i}}{v^{\setminus i}+1}(\mathbf{x}_i - \mathbf{m})\,,$$

$$v_{\text{new}} = v^{\setminus i} - \rho_i \frac{(v^{\setminus i})^2}{v^{\setminus i}+1} + \rho_i(1-\rho_i)\frac{(v^{\setminus i})^2||\mathbf{x}_i - \mathbf{m}^{\setminus i}||^2}{D(v^{\setminus i}+1)^2}\,,$$

where we have used again the Gaussian identities and

$$\rho_i = 1 - \frac{w}{Z_i}\mathcal{N}(\mathbf{x}_i|\mathbf{0}, a\mathbf{I})$$

can be interpreted as the probability of $\mathbf{x}_i$ not being clutter.

Update of the Approximate Factor $\tilde{f}_i$

$\tilde{f}_i$ is updated to be equal to $Z_i \mathcal{Q}_{\text{new}}/\mathcal{Q}^{\setminus i}$:

$$
\begin{aligned}
(\tilde{v}_i)^{-1} &= (v_{\text{new}})^{-1} - (v^{\setminus i})^{-1}\,, \\
\tilde{\mathbf{m}}_i &= \tilde{v}_i \left(v_{\text{new}}^{-1}\mathbf{m}_{\text{new}} - (v^{\setminus i})^{-1}\mathbf{m}^{\setminus i} \right)\,, \\
\tilde{s}_i &= \frac{Z_i}{\mathcal{N}(\tilde{\mathbf{m}}_i | \mathbf{m}^{\setminus i}, (\tilde{v}_i + v^{\setminus i})\mathbf{I})}\,,
\end{aligned}
$$

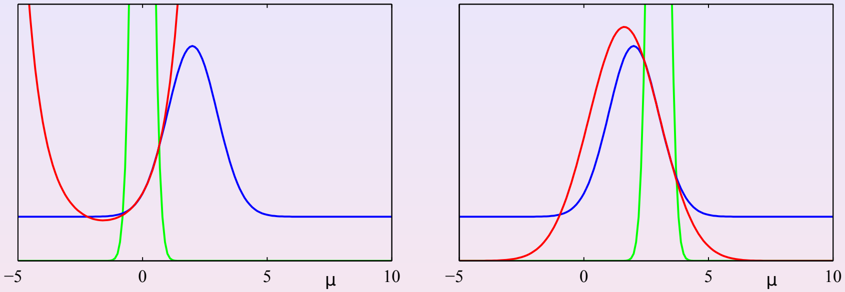
where we used the Gaussian identities.

At convergence we evaluate the approximation of the marginal likelihood:

$$
p(\mathbf{e}) \approx \int \prod_{i=0}^{N} \tilde{f}_i(\boldsymbol{\mu}) d\boldsymbol{\mu} = (2\pi v_{\text{new}})^{D/2} \exp(B/2) \prod_{i=0}^{N} \left[\tilde{s}_i (2\pi \tilde{v}_i)^{-D/2} \right]\,,
$$

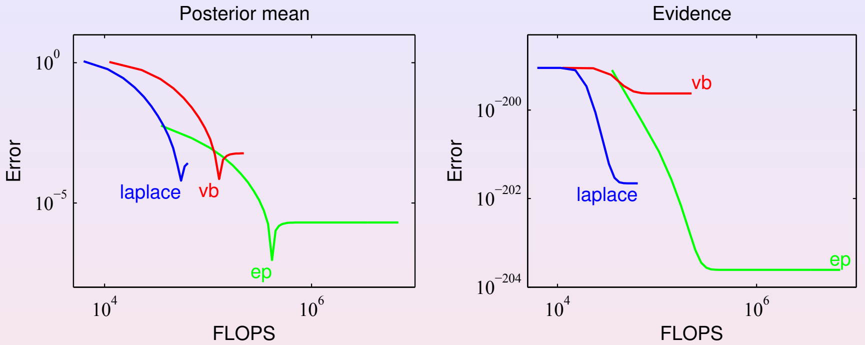
where $B = \mathbf{m}_{\text{new}}^{\mathsf{T}} v_{\text{new}}^{-1} \mathbf{m}_{\text{new}} - \sum_{i=0}^{N} \tilde{\mathbf{m}}^{\mathsf{T}} (\tilde{v}_i)^{-1} \tilde{\mathbf{m}}$ and we have used the Gaussian identities.

EP Example: Computed Approximations of f_i



Approximation of specific factors f_i when $D = 1$. Exact factor $f_i(\mu)$ is shown in blue (a Gaussian plus a constant), approximate factor $\tilde{f}_i(\mu)$ is shown in red, and $\mathcal{Q}^{\setminus i}(\mu)$ in green. The Gaussian approximation is accurate in regions of high posterior probability as estimated by $\mathcal{Q}^{\setminus i}$.

EP Example: Comparison with VI and Laplace



Comparison of EP with Laplace's method and Variational Inference (mean field) on the clutter problem. Accuracy is measured in absolute difference from the true mean and the true integral. Cost is measured in FLOPS (floating point operations).