Variational Inference - Expectation Propagation
Exponential Families

**Definition:** A probability distribution $p$ over $x$ is a member of the **exponential family** if it can be expressed as

$$p(x | \eta) = h(x) g(\eta) \exp(\eta^T u(x))$$

where $\eta$ are the **natural parameters** and

$$g(\eta) = \left( \int h(x) \exp(\eta^T u(x)) dx \right)^{-1}$$

is the normalizer.

$h$ and $u$ are functions of $x$. 
Exponential Families

Example: Bernoulli-Distribution with parameter $\mu$

\[
p(x \mid \mu) = \mu^x (1 - \mu)^{1-x}
\]

\[
= \exp(x \ln \mu + (1 - x) \ln(1 - \mu))
\]

\[
= \exp(x \ln \mu + \ln(1 - \mu) - x \ln(1 - \mu))
\]

\[
= (1 - \mu) \exp(x \ln \mu - x \ln(1 - \mu))
\]

\[
= (1 - \mu) \exp \left( x \ln \left( \frac{\mu}{1 - \mu} \right) \right)
\]

Thus, we can say

\[
\eta = \ln \left( \frac{\mu}{1 - \mu} \right) \Rightarrow \mu = \frac{1}{1 + \exp(-\eta)} \Rightarrow 1 - \mu = \frac{1}{1 + \exp(\eta)} = g(\eta)
\]
Exponential Families

Example: Normal-Distribution with parameters $\mu$ and $\sigma$

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

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

$$h(x) = \frac{1}{\sqrt{2\pi}} \quad u(x) = (x, x^2)^T$$
MLE for Exponential Families

From: \[ g(\eta) \int h(x) \exp(\eta^T u(x)) dx = 1 \]

we get:

\[ \nabla g(\eta) \int h(x) \exp(\eta^T u(x)) dx + g(\eta) \int h(x) \exp(\eta^T u(x)) u(x) dx = 0 \]

\[ \Rightarrow -\frac{\nabla g(\eta)}{g(\eta)} = g(\eta) \int h(x) \exp(\eta^T u(x)) u(x) dx = \mathbb{E}[u(x)] \]

which means that \[ -\nabla \ln g(\eta) = \mathbb{E}[u(x)] \]
MLE for Exponential Families

From:
\[
g(\eta) \int h(x) \exp(\eta^T u(x)) dx = 1
\]
we get:
\[
\nabla g(\eta) \int h(x) \exp(\eta^T u(x)) dx + g(\eta) \int h(x) \exp(\eta^T u(x)) u(x) dx = 0
\]

\[
\Rightarrow -\frac{\nabla g(\eta)}{g(\eta)} = g(\eta) \int h(x) \exp(\eta^T u(x)) u(x) dx = E[u(x)]
\]

which means that
\[
-\nabla \ln g(\eta) = E[u(x)]
\]

\(u(x)\) is called the **sufficient statistics** of \(p\).

\(E[u(x)]\) is the vector of **moments**.
Expectation Propagation

In mean-field we minimized $\text{KL}(q\|p)$. But: we can also minimize $\text{KL}(p\|q)$. Assume $q$ is from the exponential family:

$$q(x) = h(x)g(\eta)\exp(\eta^T u(x))$$

Then we have:

$$\text{KL}(p\|q) = -\int p(x) \log \frac{h(x)g(\eta)\exp(\eta^T u(x))}{p(x)} dx$$
Expectation Propagation

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

We can minimize this with respect to \( \eta \)

\[-\nabla \log g(\eta) = \mathbb{E}_p[u(x)]\]
Expectation Propagation

This results in

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

We can minimize this with respect to \( \eta \)

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

which is equivalent to

\[ \mathbb{E}_q[u(x)] = \mathbb{E}_p[u(x)] \]

Thus: the KL-divergence is minimal if the exp. sufficient statistics are the same between \( p \) and \( q \)!

For example, if \( q \) is Gaussian: \( u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \)

Then, mean and covariance of \( q \) must be the same as for \( p \) (moment matching)
Expectation Propagation

Assume we have a factorization \( p(\mathcal{D}, \theta) = \prod_{i=1}^{M} f_i(\theta) \)
and we are interested in the posterior:

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

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

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

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

• Given: a joint distribution over data and variables
  \[ p(\mathcal{D}, \theta) = \prod_{i=1}^{M} f_i(\theta) \]

• Goal: approximate the posterior \( p(\theta \mid \mathcal{D}) \) with \( q \)

• Initialize all approximating factors \( \tilde{f}_i(\theta) \)

• Initialize the posterior approximation \( q(\theta) \propto \prod_i \tilde{f}_i(\theta) \)

• Do until convergence:
  • choose a factor \( \tilde{f}_j(\theta) \)
  • remove the factor from \( q \) by division:
    \[ q \setminus j(\theta) = \frac{q(\theta)}{\tilde{f}_j(\theta)} \]
The EP Algorithm

• find $q^{\text{new}}$ that minimizes

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

using moment matching, including the zeroth order moment:

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

• evaluate the new factor

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

• After convergence, we have

$$p(D) \approx \int \prod_i \tilde{f}_j(\theta) d\theta$$
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 $KL(q\|p)$
yellow: original distribution
red: Laplace approximation
green: global variation
blue: expectation-propagation
Remember: GP Classification

\[ p(f \mid X, y) = \frac{p(y \mid f)p(f \mid X)}{p(y \mid X)} \]

• 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)

\[ p(x \mid \theta) = (1 - w)\mathcal{N}(x \mid \theta, I) + w\mathcal{N}(x \mid 0, aI) \]

- The prior is Gaussian:

\[ p(\theta) = \mathcal{N}(\theta \mid 0, bI) \]
The Clutter Problem

The joint distribution for \( \mathcal{D} = (x_1, \ldots, x_N) \) is

\[
p(\mathcal{D}, \theta) = p(\theta) \prod_{n=1}^{N} p(x_n \mid \theta)
\]

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

\[
q(\theta) = \mathcal{N}(\theta \mid m, \nu I) = \tilde{f}_0(\theta) \prod_{n=1}^{N} \tilde{f}_n(\theta)
\]

the factors are (unnormalized) Gaussians:

\[
\tilde{f}_0(\theta) = p(\theta) \quad \tilde{f}_n(\theta) = s_n \mathcal{N}(\theta \mid m_n, \nu_n I)
\]
EP for the Clutter Problem

• First, we initialize $\tilde{f}_n(\theta) = 1$, i.e. $q(\theta) = p(\theta)$

• Iterate:
  • Remove the current estimate of $\tilde{f}_n(\theta)$ from $q$ by division of Gaussians:

$$q_{-n}(\theta) = \frac{q(\theta)}{\tilde{f}_n(\theta)}$$
EP for the Clutter Problem

• First, we initialize \( \tilde{f}_n(\theta) = 1 \), i.e. \( q(\theta) = p(\theta) \)

• Iterate:
  
  • Remove the current estimate of \( \tilde{f}_n(\theta) \) from \( q \) by division of Gaussians:
    \[
    q_{-n}(\theta) = \frac{q(\theta)}{\tilde{f}_n(\theta)} \quad q_{-n}(\theta) = \mathcal{N}(\theta \mid m_{-n}, v_{-n}I)
    \]

  • Compute the normalization constant:
    \[
    Z_n = \int q_{-n}(\theta) f_n(\theta) d\theta
    \]

  • Compute mean and variance of \( q_{\text{new}} = q_{-n}(\theta) f_n(\theta) \)
  
  • Update the factor \( \tilde{f}_n(\theta) = Z_n \frac{q_{\text{new}}(\theta)}{q_{-n}(\theta)} \)
A 1D Example

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

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

Sampling Methods are widely used in Computer Science

• as an **approximation** of a deterministic algorithm
• to represent **uncertainty** without a parametric model
• to obtain higher computational **efficiency** with a small approximation error

Sampling Methods are also often called **Monte Carlo Methods**

Example: Monte-Carlo Integration

• Sample in the bounding box
• Compute fraction of inliers
• Multiply fraction with box size
Non-Parametric Representation

Probability distributions (e.g. a robot’s belief) can be represented:

- **Parametrically**: e.g. using mean and covariance of a Gaussian
- **Non-parametrically**: using a set of hypotheses (samples) drawn from the distribution

Advantage of non-parametric representation:

- No restriction on the type of distribution (e.g. can be multi-modal, non-Gaussian, etc.)
Non-Parametric Representation

The more samples are in an interval, the higher the probability of that interval

But:
How to draw samples from a function/distribution?
Sampling from a Distribution

There are several approaches:

- Probability transformation
  - Uses inverse of the c.d.f (not considered here)
- Rejection Sampling
- Importance Sampling
- Markov Chain Monte Carlo
Rejection Sampling

1. Simplification:
   - Assume $p(z) < 1$ for all $z$
   - Sample $z$ uniformly
   - Sample $c$ from $[0, 1]$

• If $f(z) > c$:
  - keep the sample

otherwise:
  - reject the sample
Rejection Sampling

2. General case:
Assume we can evaluate $p(z) = \frac{1}{Z_p} \tilde{p}(z)$ (unnormalized)

- Find proposal distribution $q$
  - Easy to sample from $q$
- Find $k$ with $kq(z) \geq \tilde{p}(z)$
- Sample from $q$
- Sample uniformly from $[0,kq(z_0)]$
- Reject if $u_0 > \tilde{p}(z_0)$

But: Rejection sampling is inefficient.
Importance Sampling

• **Idea:** assign an *importance weight* $w$ to each sample

• With the importance weights, we can account for the “differences between $p$ and $q$ ”

$$w(x) = \frac{p(x)}{q(x)}$$

• $p$ is called **target**

• $q$ is called **proposal** (as before)
Importance Sampling

- **Explanation:** The prob. of falling in an interval $A$ is the **area** under $p$
- This is equal to the expectation of the **indicator function** $I(x \in A)$

$$E_p[I(z \in A)] = \int p(z)I(z \in A)dz$$
Importance Sampling

- **Explanation:** The prob. of falling in an interval $A$ is the area under $p$
- This is equal to the expectation of the indicator function $I(x \in A)$

$$E_p[I(z \in A)] = \int p(z)I(z \in A)dz$$

$$= \int \frac{p(z)}{q(z)}q(z)I(z \in A)dz = E_q[w(z)I(z \in A)]$$

**Requirement:** $p(x) > 0 \Rightarrow q(x) > 0$

Approximation with samples drawn from $q$:

$$E_q[w(z)I(z \in A)] \approx \frac{1}{L} \sum_{l=1}^{L} w(z_l)I(z_l \in A)$$