10. 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 and variational methods.

• In variational optimization, we seek a tractable distribution $q$ that approximates the posterior.

• Optimization is done using functionals.
Variational Inference

In general, variational methods are concerned with mappings that take \textbf{functions} as input.

Example: the entropy of a distribution $p$

$$\mathbb{H}[p] = \int p(x) \log p(x) \, dx$$

Variational optimization aims at finding \textbf{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:

\[ \log p(X) = \mathcal{L}(q) + \text{KL}(q \parallel p) \]

\[ \mathcal{L}(q) = \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ \quad \text{KL}(q) = - \int q(Z) \log \frac{p(Z \mid X)}{q(Z)} dZ \]

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

$$-\int p(x) \log q(x) dx - \left( -\int p(x) \log p(x) dx \right) = -\int p(x) \log \frac{q(x)}{p(x)} dx = \text{KL}(p\|q)$$

This is known as the **Kullback-Leibler divergence**

It has the properties:

$$\text{KL}(q\|p) \neq \text{KL}(p\|q)$$

$$\text{KL}(p\|q) \geq 0$$

$$\text{KL}(p\|q) = 0 \iff p \equiv q$$

This follows from Jensen’s inequality
Factorized Distributions

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

Thus, we have
\[ \mathcal{L}(q) = -\text{KL}(q_j \Vert \tilde{p}(X, Z_j)) + \text{const} \]

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:
\[ \mathbb{E}_{-j} [\log p(X, Z)] = \int \log p(X, Z) \prod_{i \neq j} q_i dZ_{-j} \]
Mean Field Theory

Therefore, the optimal solution in general is

$$\log q^*_j(Z_j) = \mathbb{E}_{-j} \left[ \log p(X, Z) \right] + \text{const}$$

In words: the log of the optimal solution for a factor $q_j$ 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 $X = \{x_1, \ldots, x_N\}$ and latent variables $Z = \{z_1, \ldots, z_N\}$

• Furthermore we have

\[
p(Z \mid \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{nk}^{zn_k} \quad p(X \mid Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n \mid \mu_k, \Lambda^{-1})^{zn_k}
\]

• We introduce priors for all parameters, e.g.

\[
p(\pi) = \text{Dir}(\pi \mid \alpha_0)
\]

\[
p(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k \mid m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid W_0, \nu_0)
\]
Variational Mixture of Gaussians

• The joint probability is then:
\[ p(X, Z, \pi, \mu, \Lambda) = p(X \mid Z, \mu, \Lambda)p(Z \mid \pi)p(\pi)p(\mu \mid \Lambda)p(\Lambda) \]

• We consider a distribution \( q \) so that
\[ q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda) \]

• Using our general result:
\[ \log q^*(Z) = \mathbb{E}_{\pi, \mu, \Lambda}[\log p(X, Z, \pi, \mu, \Lambda)] + \text{const} \]

• Plugging in:
\[ \log q^*(Z) = \mathbb{E}_\pi[\log p(Z \mid \pi)] + \mathbb{E}_{\mu, \Lambda}[\log p(X \mid Z, \mu, \Lambda)] + \text{const} \]
Variational Mixture of Gaussians

• The joint probability is then:
\[ p(X, Z, \pi, \mu, \Lambda) = p(X | Z, \mu, \Lambda)p(Z | \pi)p(\pi)p(\mu | \Lambda)p(\Lambda) \]

• We consider a distribution \( q \) so that
\[ q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda) \]

• Using our general result:
\[ \log q^*(Z) = \mathbb{E}_{\pi, \mu, \Lambda}[\log p(X, Z, \pi, \mu, \Lambda)] + \text{const} \]

• Plugging in:
\[ \log q^*(Z) = \mathbb{E}_\pi[\log p(Z | \pi)] + \mathbb{E}_{\mu, \Lambda}[\log p(X | Z, \mu, \Lambda)] + \text{const} \]

• From this we can show that:
\[ q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r^\frac{z_{nk}}{n_k} \]
Variational Mixture of Gaussians

This means: the optimal solution to the factor $q(Z)$ has the same functional form as the prior 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.
Example: Clustering

- 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
10. Variational Inference: Expectation Propagation
**Excurse: 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

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

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

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

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

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

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

Then we have:

$$g(\eta) \int h(x) \exp(\eta^T u(z)) dx = 1$$

Then we have:

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

$$-\nabla \log g(\eta) = \mathbb{E}_p[u(x)]$$
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 \( \text{KL} \left( \frac{1}{p(\mathcal{D})} \prod_{i=1}^{M} f_i(\theta) \parallel \frac{1}{Z} \prod_{i=1}^{M} \tilde{f}_i(\theta) \right) \)

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

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

using moment matching, including the zeroth order moment:

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

• evaluate the new factor

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

• After convergence, we have

$$p(D) \approx \int \prod_i \tilde{f}_i(\theta)d\theta$$
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 $KL(q||p)$
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)

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

- 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

$$p(\mathcal{D}, \theta) = p(\theta) \prod_{n=1}^{N} p(x_n | \theta)$$

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

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

the factors are (unnormalized) Gaussians:

$$\tilde{f}_0(\theta) = p(\theta) \quad \tilde{f}_n(\theta) = s_n \mathcal{N}(\theta | m_n, \nu_n I)$$
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:

\[
q_{-n}(\theta) = \frac{q(\theta)}{\tilde{f}_n(\theta)}
\]

\( q_{-n}(\theta) = \mathcal{N}(\theta \mid m_{-n}, v_{-n}I) \)

• 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

• 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