11. 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**.
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.

Careful: Different notation!

In Bishop (and in the following slides)

\( Z \) are hidden states

and \( X \) are observations
Variational Inference

In general, variational methods are concerned with mappings that take functions as input.

Example: the entropy of a distribution $p$

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

“Functional”

Variational optimization aims at finding 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.
The KL-Divergence

**Aim:** define a functional that resembles a “difference” between distributions $p$ and $q$

**Idea:** use the average additional amount of information:

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

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
Example: A Variational Formulation of EM

Assume for a moment that we observe $X$ and the binary latent variables $Z$. The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} p(z_n \mid \pi)p(x_n \mid z_n, \mu, \Sigma)$$
**Example: A Variational Formulation of EM**

Assume for a moment that we observe $X$ and the binary latent variables $Z$. The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} p(z_n \mid \pi)p(x_n \mid z_n, \mu, \Sigma)$$

where

$$p(z_n \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

and

$$p(x_n \mid z_n, \mu, \Sigma) = \prod_{k=1}^{K} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

Remember:

$$z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1$$
**Example: A Variational Formulation of EM**

Assume for a moment that we observe $X$ and the binary latent variables $Z$. The likelihood is then:

$$p(X, Z \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} p(z_n \mid \pi)p(x_n \mid z_n, \mu, \Sigma)$$

where

$$p(z_n \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

and

$$p(x_n \mid z_n, \mu, \Sigma) = \prod_{k=1}^{K} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

which leads to the log-formulation:

$$\log p(X, Z \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}(\log \pi_k + \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k))$$
The Complete-Data Log-Likelihood

\[
\log p(X, Z \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left( \log \pi_k + \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)
\]

• This is called the **complete-data log-likelihood**

• Advantage: solving for the parameters \((\pi_k, \mu_k, \Sigma_k)\) is much simpler, as the log is inside the sum!

• We could switch the sums and then for every mixture component \(k\) only look at the points that are associated with that component.

• This leads to simple closed-form solutions for the parameters

• However: the latent variables \(Z\) are not observed!
The Main Idea of EM

Instead of maximizing the joint log-likelihood, we maximize its **expectation** under the latent variable distribution:

\[
\mathbb{E}_Z[\log p(X, Z | \pi, \mu, \Sigma)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_Z[z_{nk}](\log \pi_k + \log \mathcal{N}(x_n | \mu_k, \Sigma_k))
\]
The Main Idea of EM

Instead of maximizing the joint log-likelihood, we maximize its **expectation** under the latent variable distribution:

$$
\mathbb{E}_Z[\log p(X, Z | \pi, \mu, \Sigma)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_Z[z_{nk}](\log \pi_k + \log \mathcal{N}(x_n | \mu_k, \Sigma_k))
$$

where the latent variable distribution per point is:

$$
p(z_n | x_n, \theta) = \frac{p(x_n | z_n, \theta)p(z_n | \theta)}{p(x_n | \theta)}
\quad \theta = (\pi, \mu, \Sigma)
$$

$$
= \frac{\prod_{l=1}^{K} (\pi_l \mathcal{N}(x_n | \mu_l, \Sigma_l))^{z_{nl}}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}
$$
The Main Idea of EM

The expected value of the latent variables is:

\[ \mathbb{E}[z_{nk}] = \gamma(z_{nk}) \]

plugging in we obtain:

\[ \mathbb{E}_Z[\log p(X, Z | \pi, \mu, \Sigma)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})(\log \pi_k + \log \mathcal{N}(x_n | \mu_k, \Sigma_k)) \]

We compute this iteratively:

1. Initialize \( i = 0, (\pi_k^i, \mu_k^i, \Sigma_k^i) \)
2. Compute \( \mathbb{E}[z_{nk}] = \gamma(z_{nk}) \)
3. Find parameters \((\pi_k^{i+1}, \mu_k^{i+1}, \Sigma_k^{i+1})\) that maximize this
4. Increase \( i; \) if not converged, goto 2.
Why Does This Work?

• We have seen that EM maximizes the expected complete-data log-likelihood, but:

• Actually, we need to maximize the log-marginal

\[
\log p(X \mid \theta) = \log \sum_Z p(X, Z \mid \theta)
\]

• It turns out that the log-marginal is maximized implicitly!
A Variational Formulation of EM

- We have seen that EM maximizes the **expected complete-data log-likelihood**, but:
- Actually, we need to maximize the log-marginal

\[
\log p(X \mid \theta) = \log \sum_Z p(X, Z \mid \theta)
\]

- It turns out that the log-marginal is maximized implicitly!

\[
\log p(X \mid \theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)
\]

\[
\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \quad \text{KL}(q\|p) = -\sum_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)}
\]
A Variational Formulation of EM

• Thus: The Log-likelihood consists of two functionals

\[
\log p(X \mid \theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)
\]

where the first is (proportional to) an expected complete-data log-likelihood under a distribution \(q\)

\[
\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)}
\]

and the second is the KL-divergence between \(p\) and \(q\):

\[
\text{KL}(q\|p) = -\sum_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)}
\]
The KL-divergence is positive or 0
Thus, the log-likelihood is at least as large as $\mathcal{L}$ or:
$\mathcal{L}$ is a **lower bound** (ELBO) of the log-likelihood (evidence):

$$\log p(X \mid \theta) \geq \mathcal{L}(q, \theta)$$
What Happens in the E-Step?

- The log-likelihood is independent of $q$
- Thus: $\mathcal{L}$ is maximized iff KL divergence is minimal ($=0$)
- This is the case iff $q(Z) = p(Z | X, \theta)$
What Happens in the M-Step?

- In the M-step we keep $q$ fixed and find new $\theta$
  \[
  \mathcal{L}(q, \theta) = \sum_Z p(Z \mid X, \theta^{\text{old}}) \log p(X, Z \mid \theta) - \sum_Z q(Z) \log q(Z)
  \]
- We maximize the first term, the second is indep.
- This implicitly makes KL non-zero
- The log-likelihood is maximized even more!
• In the E-step we compute the concave lower bound for given old parameters $\theta^{old}$ (blue curve).
• In the M-step, we maximize this lower bound and obtain new parameters $\theta^{new}$.
• This is repeated (green curve) until convergence.
VI in General

Analogue to the discussion about EM we have:

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

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

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.
Generalizing the Idea

- In EM, we were looking for an optimal distribution $q$ in terms of KL-divergence.
- Luckily, we could compute $q$ in closed form.
- In general, this is not the case, but we can use an approximation instead: $q(Z) \approx p(Z \mid X)$.
- Idea: make a simplifying assumption on $q$ so that a good approximation can be found.
- For example: Consider the case where $q$ can be expressed as a product of simpler terms.
Factorized Distributions

We can split up $q$ by partitioning $Z$ into disjoint sets and assuming 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 $q_i \leftarrow q_i(Z_i)$ 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

$$\mathcal{L}(q) = -\text{KL}(q_j \| \tilde{p}(X, Z_j)) + \text{const}$$

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

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

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

\[
p(\pi) = \text{Dir}(\pi | \alpha_0)\\
p(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, v_0)
\]
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:
  \[ p(X, Z, \pi, \mu, \Lambda) = p(X \mid Z, \mu, \Lambda)p(Z \mid \pi)p(\pi)p(\mu \mid \Lambda)p(\Lambda) \]

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

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

- Plugging in:
  \[ \log q^*(Z) = E_{\pi}[\log p(Z \mid \pi)] + E_{\mu, \Lambda}[\log p(X \mid Z, \mu, \Lambda)] + \text{const} \]

- From this we can show that:
  \[ q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \]
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