Bayesian Methods for Machine Learning

Lecture 5 - Variational inference

Simon Leglaive

CentraleSupélec

Approximate inference

In Bayesian methods, it's all about posterior computation:

$$p(\mathbf{z}|\mathbf{x};\theta) = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{p(\mathbf{x};\theta)} = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{\int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z}}$$

- Easy to compute for conjugate prior.
- Hard for many models were conjugacy does not hold.

In Bayesian methods, it's all about posterior computation:

$$p(\mathbf{z}|\mathbf{x};\theta) = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{p(\mathbf{x};\theta)} = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{\int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z}}$$

- Easy to compute for conjugate prior.
- Hard for many models were conjugacy does not hold.

For example:

$$p(\mathbf{x}|\mathbf{z};\theta) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}),\boldsymbol{\Sigma}_\theta(\mathbf{z})\right),$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are neural networks.

The marginal likelihood $p(\mathbf{x};\theta) = \int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z} = \int \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}),\boldsymbol{\Sigma}_\theta(\mathbf{z})\right)p(\mathbf{z};\theta)d\mathbf{z}$ is intractable due to the non-linearities.

For simplicity of notations, we denote the parameters of the likelihood and prior by θ even though they are generally different.

We need approximate inference techniques when exact posterior computation is infeasible:

- Variational inference (focus of today)
- Markov Chain Monte Carlo

Variational inference

The main idea of variational inference is to cast inference as an optimization problem.

We want to find a variational distribution $q(\mathbf{z}) \in \mathcal{F}$ which approximates the true intractable posterior $p(\mathbf{z}|\mathbf{x})$.

We need to define:

- a measure of fit between $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x};\theta)$, to be minimized,
- a variational family $\mathcal{F}$, which corresponds to the set of acceptable solutions for the variational distribution.

The KL divergence

$$D_{\mathrm{KL}}(q \parallel p) = \mathbb{E}_q[\ln q - \ln p].$$

The KL divergence has the following properties:

- $D_{\mathrm{KL}}(q \parallel p) \geq 0$
- $D_{\mathrm{KL}}(q \parallel p) = 0$ if and only if $q = p$
- $D_{\mathrm{KL}}(q \parallel p) \neq D_{\mathrm{KL}}(p \parallel q)$

- Why do we choose $D_{\mathrm{KL}}(q \parallel p)$ and not $D_{\mathrm{KL}}(p \parallel q)$?

  $D_{\mathrm{KL}}(q \parallel p)$ involves an expectation w.r.t q, while for $D_{\mathrm{KL}}(p \parallel q)$ the expectation is taken w.r.t p, which is intractable (when it is the posterior).

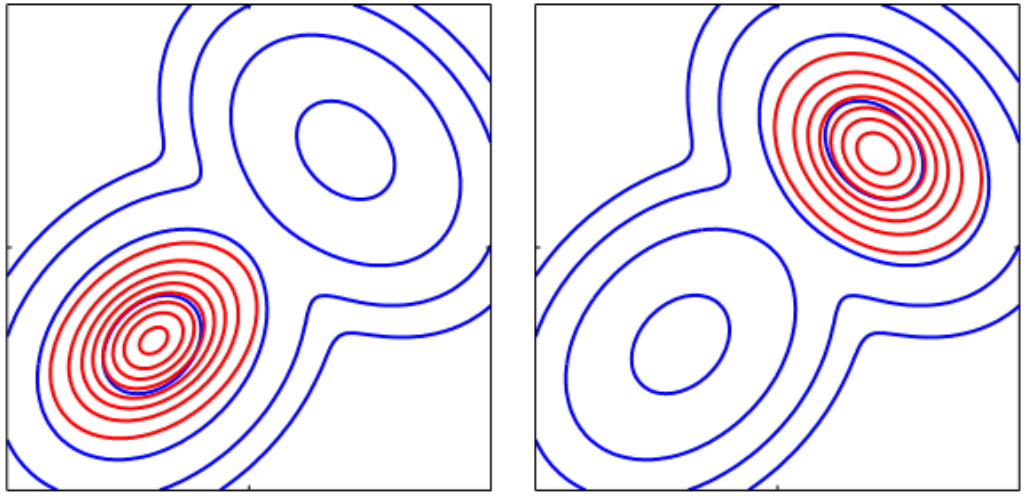
- How does this choice influence the approximation?

$$D_{\mathrm{KL}}(q \parallel p) = \mathbb{E}_q[\ln\left(q/p\right)]$$

This is the reverse KL, which is large when p is close to zero and q is not.

This form penalizes distributions q that put probability mass where p is small.

However it is ok if q is close to zero while p is not. As a consequence, it may underestimate the support of p, i.e. q may concentrate on a single mode of p.

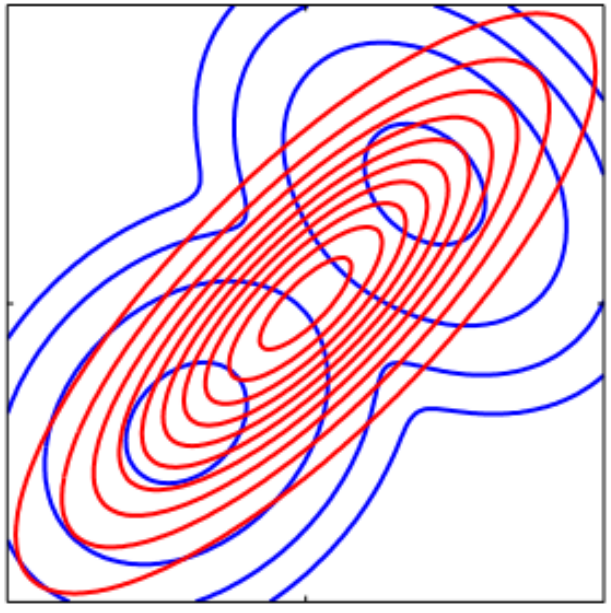


$$D_{\mathrm{KL}}(p \parallel q) = \mathbb{E}_p[\ln\left(p/q\right)]$$

The forward KL is large when q is close to zero and p is not.

This form penalizes distributions q that "would not sufficiently cover" p.

However it is ok if q has probability mass where p is close to zero. As a consequence, it may overestimate the support of p, i.e. q may have probability mass on regions where p does not.



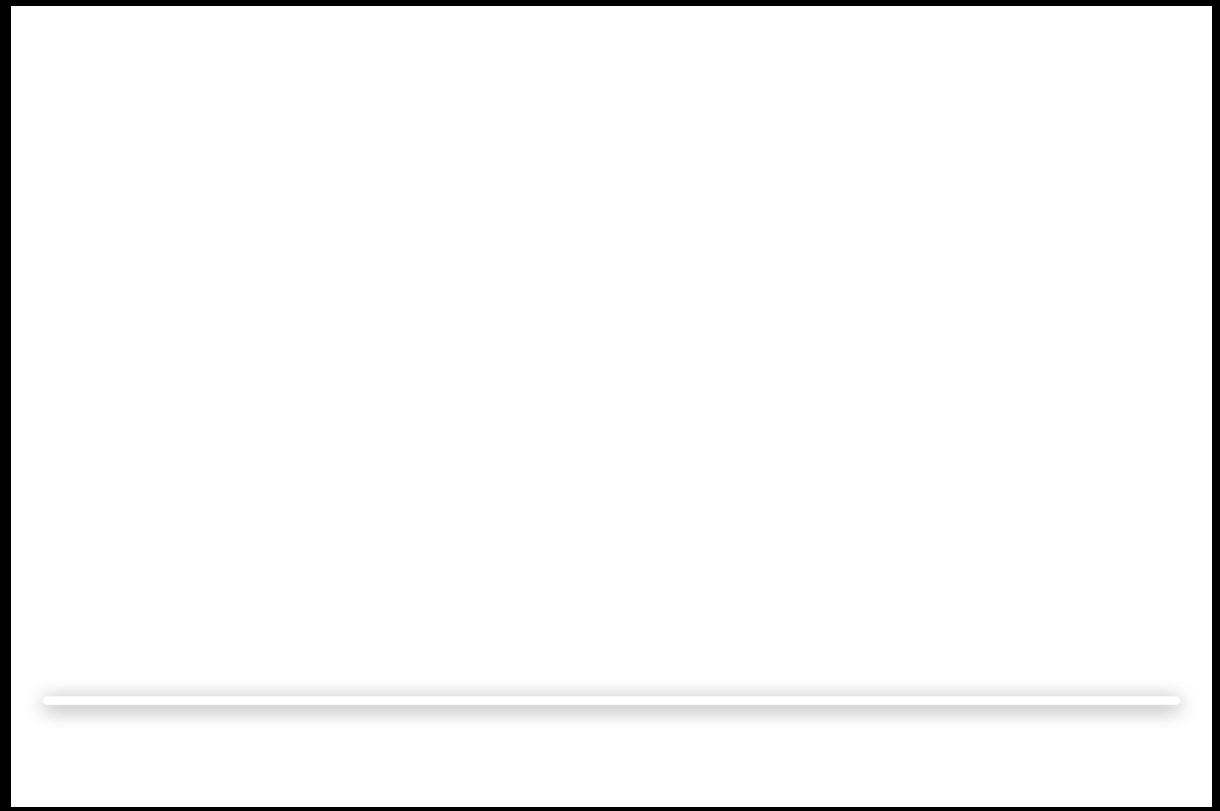


Image credit: https://twitter.com/ari_seff/status/1303741288911638530

The ELBO

So we take the reverse KL divergence as a measure of fit between the variational distribution and the intractable posterior:

$$\begin{aligned} D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) &= \mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z}) - \ln p(\mathbf{z}|\mathbf{x};\theta)] \\ &= \mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z}) - \ln p(\mathbf{x},\mathbf{z};\theta) + \ln p(\mathbf{x};\theta)] \\ &= \ln p(\mathbf{x};\theta) - \mathcal{L}(q(\mathbf{z}),\theta) \end{aligned}$$

where

$$\mathcal{L}(q(\mathbf{z}),\theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q(\mathbf{z})]$$

is the evidence lower bound (ELBO) that we've already encountered in the EM algorithm. It is also called the (negative) variational free energy.

The ELBO can be further decomposed as:

$$\mathcal{L}(q(\mathbf{z}), \theta) = E(q(\mathbf{z}), \theta) + H(q(\mathbf{z})),$$

where $E(q(\mathbf{z}), \theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$ and $H(q(\mathbf{z})) = -\mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z})]$ is the differential entropy of $q(\mathbf{z})$ which does not depend on the model parameters θ.

Variational inference consists in solving the following optimization problem:

$$\min_{q \in \mathcal{F}} \quad D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) \qquad \Leftrightarrow \qquad \max_{q \in \mathcal{F}} \quad \mathcal{L}(q(\mathbf{z}), \theta).$$

Variational inference consists in solving the following optimization problem:

$$\min_{q \in \mathcal{F}} \quad D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) \qquad \Leftrightarrow \qquad \max_{q \in \mathcal{F}} \quad \mathcal{L}(q(\mathbf{z}), \theta).$$

If the variational family $\mathcal{F}$ is not constrained (i.e. it is the set of all pdfs over $\mathbf{z}$), we have:

$$\begin{aligned} q^\star(\mathbf{z}) &= \arg\min_{q \in \mathcal{F}} \quad D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) \\ &= \arg\max_{q \in \mathcal{F}} \quad \mathcal{L}(q(\mathbf{z}), \theta) \\ &= p(\mathbf{z}|\mathbf{x};\theta), \end{aligned}$$

which corresponds to the E-step of the EM algorithm for exact inference...

Variational inference consists in solving the following optimization problem:

$$\min_{q \in \mathcal{F}} \quad D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) \qquad \Leftrightarrow \qquad \max_{q \in \mathcal{F}} \quad \mathcal{L}(q(\mathbf{z}),\theta).$$

If the variational family $\mathcal{F}$ is not constrained (i.e. it is the set of all pdfs over $\mathbf{z}$), we have:

$$\begin{aligned} q^\star(\mathbf{z}) &= \arg\min_{q \in \mathcal{F}} \quad D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)) \\ &= \arg\max_{q \in \mathcal{F}} \quad \mathcal{L}(q(\mathbf{z}),\theta) \\ &= p(\mathbf{z}|\mathbf{x};\theta), \end{aligned}$$

which corresponds to the E-step of the EM algorithm for exact inference...

... but our starting hypothesis was "the true posterior is analytically intractable", so we need to induce some constraints on the variational distribution $q(\mathbf{z})$, through the definition of the variational family $\mathcal{F}$.

Our goal is to restrict the family sufficiently such that it comprises only tractable distributions. But at the same time we want the family to be sufficiently rich and flexible such that it can provide a good approximation to the true posterior distribution.

It is important to emphasize that the restriction is imposed purely to achieve tractability, and that subject to this requirement we should use a family of approximating distributions as rich as possible.

Mean-field variational inference

The mean field approximation defines the variational family $\mathcal{F}$ as the set of pdfs that can be factorized as follows:

$$q(\mathbf{z}) = \prod_{i=1}^{L} q_i(z_i),$$

where $\mathbf{z} = \{z_i\}_{i=1}^{L}$.

The mean field approximation assumes that the individual scalar latent variables are independent *a posteriori*, that is for all (i, j) with $i \neq j$,

$$q(z_i, z_j) = q_i(z_i) q_j(z_j),$$

even though this may not hold for the true posterior:

$$p(z_i, z_j | \mathbf{x}; \theta) \neq p(z_i | \mathbf{x}; \theta) p(z_j | \mathbf{x}; \theta).$$

The mean field approximation defines the variational family $\mathcal{F}$ as the set of pdfs that can be factorized as follows:

$$q(\mathbf{z}) = \prod_{i=1}^{L} q_i(z_i),$$

where $\mathbf{z} = \{z_i\}_{i=1}^{L}$.

The mean field approximation assumes that the individual scalar latent variables are independent *a posteriori*, that is for all (i, j) with $i \neq j$,

$$q(z_i, z_j) = q_i(z_i) q_j(z_j),$$

even though this may not hold for the true posterior:

$$p(z_i, z_j | \mathbf{x}; \theta) \neq p(z_i | \mathbf{x}; \theta) p(z_j | \mathbf{x}; \theta).$$

It should be emphasized that we are making no further assumptions about the distribution. In particular, we place no restriction on the functional forms of the individual factors $q_i(z_i)$.

Among all distributions $q(\mathbf{z})$ that factorize as in the mean-field (MF) approximation, we now seek the one that maximizes the ELBO $\mathcal{L}(q(\mathbf{z}), \theta)$.

Let's inject the MF factorization into the definition of the ELBO:

$$
\begin{aligned}
\mathcal{L}(q(\mathbf{z}); \theta) =& \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}, \mathbf{z}; \theta) - \ln q(\mathbf{z})] \\
&= \int \prod_{i=1}^{L} q_i(z_i) \left[\ln p(\mathbf{x}, \mathbf{z}; \theta) - \ln \left(\prod_{i=1}^{L} q_i(z_i) \right) \right] d\mathbf{z} \\
&= \dots \text{ (see derivation details in the supporting document)} \\
&= - D_{\mathrm{KL}}\left(q_j(z_j) \parallel \tilde{p}(\mathbf{x}, z_j; \theta)\right) - \sum_{i \neq j} \mathbb{E}_{q_i(z_i)} \left[\ln q_i(z_i)\right],
\end{aligned}
$$

where $\ln \tilde{p}(\mathbf{x}, z_j; \theta) = \mathbb{E}_{\prod_{i \neq j} q_i(z_i)} \left[\ln p(\mathbf{x}, \mathbf{z}; \theta)\right] + cst.$

The constant ensures that the distribution integrates to one.

We adopt a coordinate ascent approach, where we alternatively maximize $\mathcal{L}(q(\mathbf{z}); \theta)$ with respect to each individual factor $q_j(z_j)$ considering the other ones $\{q_i(z_i)\}_{i \neq j}$ fixed.

We adopt a coordinate ascent approach, where we alternatively maximize $\mathcal{L}(q(\mathbf{z}); \theta)$ with respect to each individual factor $q_j(z_j)$ considering the other ones $\{q_i(z_i)\}_{i \neq j}$ fixed.

From the previous expression of the ELBO, we have:

$$q_j^\star(z_j) = \underset{q_j(z_j)}{\arg\max} \quad \mathcal{L}(q(\mathbf{z}); \theta) = \underset{q_j(z_j)}{\arg\min} \quad D_{\mathrm{KL}}(q_j(z_j) \parallel \tilde{p}(\mathbf{x}, z_j; \theta)).$$

We adopt a coordinate ascent approach, where we alternatively maximize $\mathcal{L}(q(\mathbf{z}); \theta)$ with respect to each individual factor $q_j(z_j)$ considering the other ones $\{q_i(z_i)\}_{i \neq j}$ fixed.

From the previous expression of the ELBO, we have:

$$q_j^\star(z_j) = \underset{q_j(z_j)}{\arg\max} \quad \mathcal{L}(q(\mathbf{z}); \theta) = \underset{q_j(z_j)}{\arg\min} \quad D_{\mathrm{KL}}(q_j(z_j) \parallel \tilde{p}(\mathbf{x}, z_j; \theta)).$$

The optimal distribution which minimizes the KL divergence is therefore given by:

$$\ln q_j^\star(z_j) = \ln \tilde{p}(\mathbf{x}, z_j; \theta) = \mathbb{E}_{\prod_{i \neq j} q_i(z_i)} [\ln p(\mathbf{x}, \mathbf{z}; \theta)] + cst,$$

The constant can be determined by normalizing $q_j^\star(z_j)$ such that it integrates to one:

$$q_j^\star(z_j) = \frac{\exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}\left[\ln p(\mathbf{x}, \mathbf{z}; \theta)\right]\right)}{\int \exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}\left[\ln p(\mathbf{x}, \mathbf{z}; \theta)\right]\right) dz_j}.$$

However, usually we simply develop $\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}\left[\ln p(\mathbf{x}, \mathbf{z}; \theta)\right]$ and identify the form of a common distribution (e.g. Gaussian, inverse-gamma, etc.)

The optimal distribution $q_j^\star(z_j)$ depends on the other factors $q_i(z_i)$, $i \neq j$, involved in the MF approximation. The solutions for different indices are therefore coupled.

A consistant global solution is obtained iteratively, by first initializing all the factors and then cycling over each individual one to compute the update.

Example: mean-field approximation of the bivariate Gaussian

We consider the problem of approximating a Gaussian distribution using a factorized Gaussian. It will provide useful insights into the types of inaccuracy introduced by the mean-field approximation.

Consider a bivariate Gaussian random vector $\mathbf{z} = [z_1, z_2]^\top$ such that

$$p(\mathbf{z}; \theta) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$

where the parameters θ are assumed to be known and correspond to the mean vector and precision matrix (inverse of the covariance matrix) which are structured as

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix},$$

and $\Lambda_{21} = \Lambda_{12}$ due to the symmetry of the precision matrix.

Suppose now that, under the mean field approximation, we want to find a factorized variational distribution

$$q(\mathbf{z}) = q_1(z_1)q_2(z_2)$$

which approximates

$$p(\mathbf{z}; \theta)$$

using the reverse KL divergence as a measure of discrepancy.

We have seen that:

1. minimizing $D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z};\theta))$ w.r.t $q(\mathbf{z})$ is equivalent to maximizing the ELBO,
2. under the mean field approximation, the optimal factor should satisfy

$$\ln q_j^\star(z_j) = \mathbb{E}_{q_i(z_i)}\left[\ln p(\mathbf{z};\theta)\right] + cst, \qquad j \in \{1,2\}, i \neq j.$$

We now have to develop this expression, ignoring all the terms that do not depend on z_j, because they can be absorbed into the normalization constant.

Let us focus on $q_1^\star(z_1)$, as $q_2^\star(z_2)$ can simply be obtained by symmetry.

The complete-data likelihood is by definition is the joint pdf of the observed and latent variables, what we denoted by $p(\mathbf{x},\mathbf{z};\theta)$ before. In the current example, we only have observed variables and the complete-data likelihood simply corresponds to $p(\mathbf{z};\theta)$.

$$
\begin{aligned}
\ln q_1^\star(z_1) &= \mathbb{E}_{q_2(z_2)}\left[\ln p(\mathbf{z};\theta)\right] + cst \\
&= \mathbb{E}_{q_2(z_2)}\left[\ln \mathcal{N}\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1}\right)\right] + cst \\
&= \mathbb{E}_{q_2(z_2)}\left[-\frac{1}{2}\begin{pmatrix} z_1-\mu_1 \\ z_2-\mu_2 \end{pmatrix}^\top \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}\begin{pmatrix} z_1-\mu_1 \\ z_2-\mu_2 \end{pmatrix}\right] + cst \\
&= \mathbb{E}_{q_2(z_2)}\left[-\frac{1}{2}\Big((z_1-\mu_1)^2\Lambda_{11} + 2(z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\Big)\right] + cst \\
&= \mathbb{E}_{q_2(z_2)}\left[-\frac{1}{2}z_1^2\Lambda_{11} + z_1\Big(\mu_1\Lambda_{11} - \Lambda_{12}(z_2-\mu_2)\Big)\right] + cst \\
&= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\Big(\mu_1\Lambda_{11} - \Lambda_{12}(\mathbb{E}_{q_2(z_2)}\left[z_2\right]-\mu_2)\Big) + cst
\end{aligned}
$$

Let's recall the result:

$$\ln q_1^\star(z_1) = -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \Big(\mu_1 \Lambda_{11} - \Lambda_{12} (\mathbb{E}_{q_2(z_2)}[z_2] - \mu_2) \Big) + cst,$$

This is a quadratic function of z_1, so the optimal distribution is a Gaussian distribution $q_1^\star(z_1) = \mathcal{N}(m_1, \gamma_1^{-1})$.

The mean and precision can be determined by identification:

$$\begin{aligned} \ln q_1^\star(z_1) &= \ln \mathcal{N}(m_1, \gamma_1^{-1}) \\ &= -\frac{1}{2}(z_1 - m_1)^2 \gamma_1 + cst \\ &= -\frac{1}{2} z_1^2 \gamma_1 + z_1 m_1 \gamma_1 + cst, \end{aligned}$$

and we identify:

$$\gamma_1 = \Lambda_{11}, \qquad m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mathbb{E}_{q_2(z_2)}[z_2] - \mu_2).$$

It is important to emphasize that we did not assume that $q_1(z_1)$ is Gaussian. We obtained this result by optimizing the KL divergence under the mean field approximation, which is the only assumption we made.

Note also that we did not compute the normalizing constant for $q_1(z_1)$ explicitely, we simply recognized the form of a known distribution (Gaussian), which implicitely gives us the normalizing constant.

By symmetry we also have:

$$q_2^\star(z_2) = \mathcal{N}(m_2, \gamma_2^{-1}),$$

with

$$\gamma_2 = \Lambda_{22}, \qquad m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}_{q_1(z_1)}[z_1] - \mu_1).$$

To sum up, from an initialization of $q_1^\star(z_1)$ and $q_2^\star(z_2)$ we iterate:

$$q_1^\star(z_1) = \mathcal{N}\Big(\underbrace{\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mathbb{E}_{q_2^\star(z_2)}\left[z_2\right] - \mu_2\right)}_{m_1}, \Lambda_{11}^{-1}\Big),$$

$$q_2^\star(z_2) = \mathcal{N}\Big(\underbrace{\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}_{q_1^\star(z_1)}\left[z_1\right] - \mu_1\right)}_{m_2}, \Lambda_{22}^{-1}\Big),$$

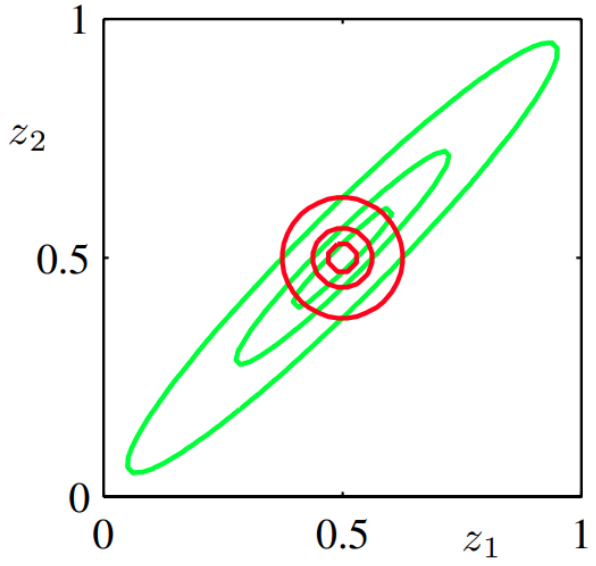
with

$$\mathbb{E}_{q_2^\star(z_2)}\left[z_2\right] = m_2, \qquad \mathbb{E}_{q_1^\star(z_1)}\left[z_1\right] = m_1.$$

It is clear that these solutions are coupled, as $q_1^\star(z_1)$ depends on an expectation computed with respect to $q_2^\star(z_2)$ and vice versa.

More precisely, we iterate the updates of the variational parameters, i.e., the parameters of the variational distributions.

After convergence, we can compare the resulting approximation $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$ with the original distribution $p(\mathbf{z}; \theta)$.



Green: original distribution, red: mean field approximation

- it captures the mean correctly,
- the variance is underestimated (due to the choice of the reverse KL),
- the elongated shape is missing (by construction of the mean field approximation).

Image credit: Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Exercise

1D Gaussian with latent mean and variance

Problem

Consider a dataset $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ of i.i.d realizations of a univariate Gaussian random variable $x \sim \mathcal{N}(\mu, \tau^{-1})$.

The mean μ and precision τ are modeled as latent random variables. We are interested in infering their posterior distribution, given the observations $\mathbf{x}$.

Generative model

- Gaussian likelihood:

$$p(\mathbf{x}|\mu,\tau) = \prod_{i=1}^{N} p(x_i|\mu,\tau) = \prod_{i=1}^{N} \mathcal{N}(x_i;\mu,\tau^{-1}) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right).$$

- Gaussian prior for the mean (conjugate):

$$p(\mu|\tau) = \mathcal{N}\left(\mu;\mu_0,(\lambda_0\tau)^{-1}\right) = \left(\frac{\lambda_0\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2\right).$$

- Gamma prior for the precision (conjugate):

$$p(\tau) = \mathcal{G}\left(\tau;a_0,b_0\right) = \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{(a_0-1)}\exp(-b_0\tau),$$

  where $\Gamma(\cdot)$ is the Gamma function.

True posterior (homework)

For this simple problem where the priors are conjugate for the likelihood, the posterior distribution can be found exactly, and it also takes the form of a Gaussian-gamma distribution.

$$p(\mu, \tau|\mathbf{x}) = p(\mu|\mathbf{x}, \tau)p(\tau|\mathbf{x}),$$

with

$$p(\mu|\mathbf{x}, \tau) = \mathcal{N}\left(\mu; \mu_\star, \lambda_\star^{-1}\right), \qquad p(\tau|\mathbf{x}) = \mathcal{G}\left(\tau; \alpha, \beta\right),$$

- $\mu_\star = \dfrac{N\tau}{N\tau + \lambda_0\tau}\bar{x} + \dfrac{\lambda_0\tau}{N\tau + \lambda_0\tau}\mu_0, \qquad \bar{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$
- $\lambda_\star = N\tau + \lambda_0\tau$
- $\alpha = a_0 + \dfrac{N}{2}$
- $\beta = b_0 + \dfrac{1}{2}\displaystyle\sum_{i=1}^{N}(x_i - \bar{x})^2 + \dfrac{\lambda_0 N}{2(\lambda_0 + N)}(\bar{x} - \mu_0)^2$

The proof is quite involved, especially for $p(\tau|\mathbf{x})$, but it's a very good exercise. You can check this document for some hints.

For practice purposes, we will consider an approximate posterior distribution using the mean field approximation:

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau).$$

Note that as shown in the previous slide, the true posterior does not factorize like this.

Exercise

Using the variational inference recipe, show that the optimal factors $q_\mu^\star(\mu)$ and $q_\tau^\star(\tau)$ are given by:

$$q_\mu^\star(\mu) = \mathcal{N}(\mu; \mu_N, \lambda_N^{-1}), \qquad q_\tau^\star(\tau) = \mathcal{G}(\tau; a_N, b_N),$$

where

$$\lambda_N = \mathbb{E}_{q_\tau(\tau)}[\tau](\lambda_0 + N),$$

$$\mu_N = \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N},$$

$$a_N = a_0 + (N+1)/2,$$

$$\begin{aligned} b_N &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu(\mu)}\left[\sum_{i=1}^{N}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] \\ &= b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 + \mathbb{E}_{q_\mu(\mu)}[\mu^2](\lambda_0 + N) - 2\mathbb{E}_{q_\mu(\mu)}[\mu](\lambda_0\mu_0 + N\bar{x}) + \lambda_0\mu_0^2\right). \end{aligned}$$

Using the properties of the Gaussian and Gamma distributions, the required expectations are given by

$$\mathbb{E}_{q_\tau(\tau)}[\tau] = a_N / b_N,$$

$$\mathbb{E}_{q_\mu(\mu)}[\mu] = \mu_N,$$

$$\mathbb{E}_{q_\mu(\mu)}[\mu^2] = \mu_N^2 + \lambda_N^{-1}.$$


jupyter

Variational EM algorithm

So far, we assumed that all the deterministic parameters of our model (likelihod and priors) are known.

What if they are not?

Generative model with latent variables (reminder)

Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ denote the observed and latent random variables, respectively.

Developing a probabilistic model consists in defining the joint distribution of the observed and latent variables, also called complete-data likelihood:

$$p(\mathbf{x}, \mathbf{z}; \theta) = p(\mathbf{x}|\mathbf{z}; \theta)p(\mathbf{z}; \theta),$$

where θ is a set of unknown deterministic parameters.

to simplify notations we use θ to denote both the parameters of the prior and likelihood, but these two distributions usually depend on disjoint sets of parameters.

Maximum marginal likelihood estimation of the model parameters (reminder)

$$\hat{\theta} = \arg\max_{\theta} \; p(\mathbf{x};\theta) = \arg\max_{\theta} \int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z}.$$

Quite often, directly solving the optimization problem associated with this ML estimation procedure is difficult, if not impossible when the marginal likelihood cannot be computed analytically.

We have seen in a previous lecture that in this case, we can leverage the fact that we have latent variables to derive an expectation-maximization (EM) algorithm to estimate the model parameters.

EM algorithm (reminder)

The EM algorithm is an iterative algorithm which alternates between optimizing the ELBO

$$\mathcal{L}(q(\mathbf{z}), \theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}, \mathbf{z}; \theta) - \ln q(\mathbf{z})]$$

with respect to $q(\mathbf{z}) \in \mathcal{F}$ in the E-Step and with repspect to θ in the M-step.

We first initialize $\theta^\star$, then we iterate:

- E-Step: $q^\star(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{F}}{\arg\max}\, \mathcal{L}(q(\mathbf{z}), \theta^\star)$
- M-Step: $\theta^\star = \underset{\theta}{\arg\max}\, \mathcal{L}(q^\star(\mathbf{z}), \theta)$

When the family $\mathcal{F}$ is unconstrained, the solution of the E-Step is given by the posterior distribution:

$$q^\star(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \theta^\star).$$

But what if this posterior is intractable?

When the family $\mathcal{F}$ is unconstrained, the solution of the E-Step is given by the posterior distribution:

$$q^\star(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \theta^\star).$$

But what if this posterior is intractable?

We have to constrain the family $\mathcal{F}$, typically with the mean field approximation.

Variational EM algorithm with the mean field approximation

Let the family $\mathcal{F}$ denote the set of probability density functions that can be factorized as:

$$q(\mathbf{z}) = \prod_{i=1}^{L} q_i(z_i), \qquad \mathbf{z} = \{z_i\}_{i=1}^{L}.$$

Given an initialization $\theta^\star$, the variational EM (VEM) algorithm consists in iterating:

- E-Step: $q^\star(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathcal{F}}{\arg\max}\, \mathcal{L}(q(\mathbf{z}), \theta^\star)$
- M-Step: $\theta^\star = \underset{\theta}{\arg\max}\, \mathcal{L}(q^\star(\mathbf{z}), \theta)$

We have seen that the solution of the E-Step consists in cyclically computing for $j = 1, ..., L$:

$$\ln q_j^\star(z_j) = \mathbb{E}_{\prod_{i\neq j} q_i(z_i)} \left[\ln p(\mathbf{x}, \mathbf{z}; \theta^\star)\right] + cst.$$