Classic VI often seeks close-formed updates of the parameters of the variational distribution q. So, it requires a simple q. One common simplified assumption often used in classic VI is the mean-field assumption.

Mean-field VI

Mean-field VI assumes $q(z)$ can be factorized into a product of independent components, i.e. it assumes independence amongst latent variables. Formally, we have $q(z) = \prod_j q_j(z_j)$. For example, in MoG, we can assume

$$q(\mu_{1:K}, z_{1:N}) = \prod_{k=1}^{K} p(\mu_k|\tilde{\mu}_k, \tilde{\sigma}_k^2) \prod_{i=1}^{N} q(z_i|\phi_i)$$

where p and q are Gaussian and Multinomial distributions, respectively. This expression is much simpler than the posterior $p(z|x)$ shown in Lecture 1 so sampling from q will be possible. However, this convenience comes with a cost. The naive and unrealistic assumption of independence makes the posterior approximation using the variational distribution imperfect. Take a 2d Gaussian as an example, the joint distribution cannot be approximated correctly by the product of two 1D Gaussians as shown in Fig. 1.

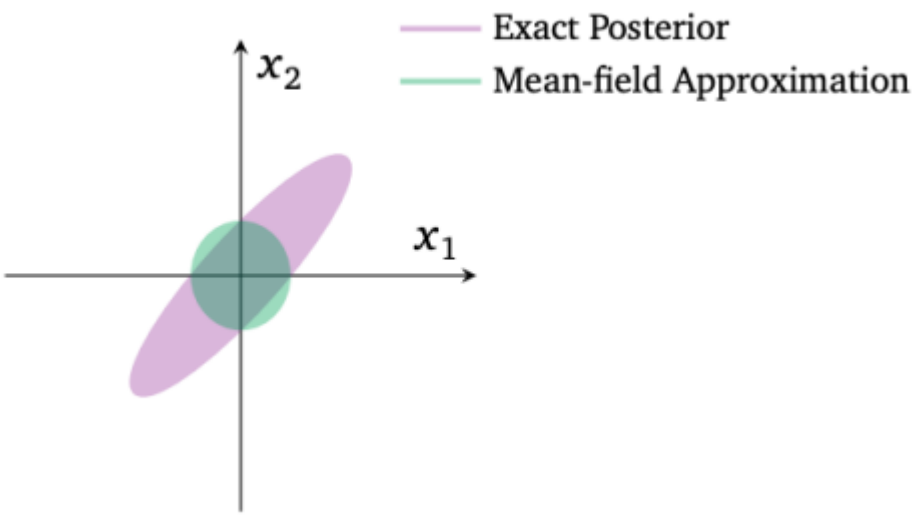

Figure 1: Mean-fied approximation error (Blei et al.,2016)

Take MoG as another example; assigning a data point x_i to a cluster k (i.e. interfering z_i) will depend on the location of the cluster (μ_k). They are not independent as assumed by mmean-fieldVI. That said, the simple form of mean-field distributions is backed by the complexity of the function that parameterises the distribution, especially in high-dimensional cases.

Optimal Mean-field Variational Distribution

One good thing about the simplicity of the mean-field assumption is that we can find the local optimal solution for the inference problem. Recall Lecture 2, we want to maximise the ELBO and we rewrite the objective in this case as

$$\begin{aligned}\mathrm{ELBO}(q) &= \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)] &(1)\\ &= \log p(x) + \mathbb{E}_q[\log p(z|x)] - \sum_{j=1}^{m} \mathbb{E}_{q_j}[\log q_j(z_j)] &(2)\end{aligned}$$

We approach this optimisation problem using *coordinate ascent*, i.e. optimising ELBO w.r.t each $q(z_j)$ one by one. Let's rewrite the ELBO as a function of some q_k by grouping all terms containing q_k as follows,

$$\begin{aligned}\mathcal{L}_k &= \mathbb{E}_q[\log p(z_k|z_{-k},x) - \mathbb{E}_{q_k}[\log q_k(z_k)] + \mathrm{const} &(3)\\ &= \int_{z_k} q_k(z_k)\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k},x)] - \int_{z_k} q_k(z_k)\log q_k(z_k) + \mathrm{const} &(4)\end{aligned}$$

where the const term contains functions of variables not including $q_k(z_k)$. As we want to find $q_k(z_k)$, a function, to maximise $\mathcal{L}_k$, a function of functions (functional), it is a functional optimisation and requires *calculus of variations* to compute the derivatives $\frac{\partial \mathcal{L}_k}{\partial q_j(z_j)}$. Also, the maximiser $q_k^*(z_k)$ must be a valid distribution, hence the functional optimisation reads

$$\begin{aligned}q_k^*(z_k) &= \arg\max_{q_k} \int_{z_k} q_k(z_k)\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k},x)] - \int_{z_k} q_k(z_k)\log q_k(z_k) &(5)\\ s.t &\int q_k^*(z_k)dz_k = 1 &(6)\end{aligned}$$

To solve constrained optimisation, we use Lagrange multiplier and find the local optimum of the following functional

$$I(q_k) = \int_{z_k} q_k(z_k)\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k},x)] - \int_{z_k} \log q_k(z_k) + \lambda(\int_{z_k} q_k(z_k) - 1) \quad (7)$$

Now we make use of *Euler-Lagrange Equation* from calculus of variations, which specifies the condition of local optimum $y(x)$ of $I(y(x)) = \int_x F(x,y,y')$ as follows,

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y} = 0 \quad (8)$$

Translate to notations in Eq. (7), $x \to z_k$, $y \to q_k$,$F \to q_k(z_k)(\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)] - \log q_k(z_k) + \lambda)$. Here, we do not have $q_k'(z_k)$ in F so $\frac{\partial F}{\partial y'} = 0$. The condition in Eq. (8) becomes

$$0 = \frac{\partial F}{\partial y} \tag{9}$$

$$\Leftrightarrow 0 = \mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)] - \log q_k(z_k) - 1 + \lambda \tag{10}$$

Hence, we have the optimal $q_k^*(z_k) = \exp\{\mathbb{E}_{q-k}[\log p(z_k|z_{-k}, x)]\} \times \exp\{-1 + \lambda\}$. We also need to find λ by solving $\frac{\partial I}{\partial \lambda} = 0$ at $q_k^*(z_k)$, which is equivalent to

$$\int_{z_k} q_k^*(z_k) - 1 = 0 \tag{11}$$

$$\Leftrightarrow \int_{z_k} \exp\{\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)]\} \times \exp\{-1 + \lambda\} = 1 \tag{12}$$

$$\Leftrightarrow \exp\{-1 + \lambda\} = \frac{1}{\int_{z_k} \exp\{\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)]\}} \tag{13}$$

Hence,

$$q_k^*(z_k) = \frac{\exp\{\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)]\}}{\int_{z_k} \exp\{\mathbb{E}_{q-k}[\log p(z_k|z_{-k}, x)]\}} \tag{14}$$

$$\propto \exp\{\mathbb{E}_{q_{-k}}[\log p(z_k|z_{-k}, x)]\} \tag{15}$$

$$\propto \exp\{\mathbb{E}_{q_{-k}}[\log p(z, x)]\} \tag{16}$$

As we can see, to infer $q_k^*(z_k)$ we need to either compute the marginal of the log complete conditional $p(z_k|z_{-k}, x)$ or the joint $p(z, x)$ distribution, which can be easy given the mean-field assumption (will see later). The process of derivation requires calculus of variations, thus comes the term "variational". That said, we can simply derive the inference without calculus of variations by rewriting Eq. (3) as

$$
\begin{aligned}
\mathcal{L}_k &= \mathbb{E}_q[\log p(z,x)] - \mathbb{E}_{q_k}[\log q_k(z_k)] + \text{const} && (17)\\
&= \int_{z_k} q_k(z_k)\mathbb{E}_{q_{-k}}[\log p(z,x)] - \int_{z_k} q_k(z_k)\log q_k(z_k) + \text{const} && (18)\\
&= \int_{z_k} q_k(z_k)\log \frac{\exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}/\int_{z_k}\exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}}{q_k(z_k)} + \text{const} && (19)\\
&= -\mathrm{KL}\left(q_k(z_k)\|\frac{\exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}}{\int_{z_k}\exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}}\right) + \text{const} && (20)
\end{aligned}
$$

As KL is non-negative, $\mathcal{L}_k$ is maximised at $q_k^*(z_k) = \exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}/\int_{z_k}\exp\{\mathbb{E}_{q_{-k}}[\log p(z,x)]\}$, which is the same as that of using calculus of variations. Using the above solution, the coordinate ascent algorithm alternates updating each q_k with other q_{-k} fixed until the ELBO converges.

Exponential Family as Complete Conditional

To make the computation of the optimal q^* manageable, we can further assume the distribution form of the complete condition $p(z_k|z_{-k},x)$ as exponential family. This form is actually very generic and can express a wide range of distributions. To be specific,

$$
\begin{aligned}
p(z_k|z_{-k},x) &= h(z_k)\exp\{\eta_k(z_{-k},x)^\top t(z_k) - a(\eta_k(z_{-k},x))\} && (21)\\
\Rightarrow q_k^*(z_k) &= h(z_k)\exp\{\mathbb{E}_{q_{-k}}[\eta_k(z_{-k},x)^\top t(z_k)] + \text{const} && (22)\\
\Rightarrow q_k^*(z_k) &\propto h(z_k)\exp\{\mathbb{E}_{q_{-k}}[\eta_k(z_{-k},x)^\top t(z_k)] && (23)
\end{aligned}
$$

The optimal q^* turns out to be in the exponential family as well and its natural parameter $\eta_k^* = \mathbb{E}_{q_{-k}}[\eta_k(z_{-k},x)]$. Therefore, if we can define the parameters for the complete condition, we can find the exponential form of q^*.

Example: Mean-field Variational MoG

We conclude this lecture with a derivation of MoG's variational inference (check Lecture 1 if you forget MoG). According to the mean-field assumption, we have the simplified variational distribution for the latent variables:

$$
q(\mu_{1:K}, z_{1:N}) = \prod_{k=1}^{K} q(\mu_k|\tilde{\mu}_k,\tilde{\sigma}_k^2)\prod_{i=1}^{N} q(z_i|\phi_i) \quad (24)
$$

where $\{\tilde{\mu}_k, \tilde{\sigma}_k\}_{k=1}^K, \{\phi_i\}_{i=1}^N$ are variational parameters needed to be found. Let's start with computing the ELBO in this case as

$$
\begin{aligned}
\mathrm{ELBO}(q) &= \mathbb{E}_q[\log p(x_{1:N}, \mu_{1:K}, z_{1:N})] - \mathbb{E}_q[\log q(\mu_{1:K}, z_{1:N}))] && (25)\\
&= \mathbb{E}_q[\log p(x_{1:N}|\mu_{1:K}, z_{1:N})] + \mathbb{E}_q[\log p(\mu_{1:K})) + \mathbb{E}_q[\log p(z_{1:N})] && (26)\\
&- \mathbb{E}_q[\log q(\mu_{1:K})] - \mathbb{E}_q[\log q(z_{1:N})] && (27)\\
&= \sum_{k=1}^K \mathbb{E}_q[\log p(\mu_k)] + \sum_{i=1}^N \mathbb{E}_q[\log p(z_i)] + \sum_{i=1}^N \mathbb{E}_q[\log p(x_i|\mu_{1:K}, z_i)] && (28)\\
&- \sum_{i=1}^K \mathbb{E}_q[\log q(\mu_k)] - \sum_{i=1}^N \mathbb{E}_q[\log q(z_i)] && (29)
\end{aligned}
$$

Each term has a closed-form computation. For example, the cross-entropy ($\int\limits_q q(x) \log p(x) dx$) between 2 Gaussians (terms 1) reads $\frac{1}{2}\log 2\pi\sigma^2 + \frac{\tilde{\sigma} + (\tilde{\mu} - \mu)^2}{2\sigma^2}$. The cross-entropy between 2 Categorical (terms 2) reads $\sum_{k=1}^K \phi_{i,k} \log p(\pi_k)$. The cross-entropy between the variational and the conditional is the most complicated, which reads:

$$
\begin{aligned}
\mathbb{E}_q[\log p(x_i|\mu_{1:K}, z_i)] &= \int\limits_{q(z_i)} \int\limits_{q(\mu_{i:K})} q(\mu_{1:K}) q(z_i) \log p(x_i|\mu_{1:K}, z_i) && (30)\\
&= \sum_{k=1}^K \phi_{i,k} \int\limits_{q(\mu_{i:K})} \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^K} e^{\frac{\sum_{k=1}^K \tilde{\mu}_k^2}{2\tilde{\sigma}^2}} \log \frac{1}{\sqrt{2e}} e^{-\frac{(x_i - \mu_k)^2}{2}} && (31)
\end{aligned}
$$

The integral is computable and so the ELBO.

Now, *the optimal variational distribution over* z_i is

$$
\begin{aligned}
q^*(z_i) &\propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\mu_{1:K}, z_{1:N}, x_{1:N})]\} && (32)\\
&\propto \exp\{\log p(z_i) + \mathbb{E}_{q_{-i}}[\log p(x_i|\mu_{1:K}, z_i)] + \mathrm{const}\} && (33)\\
&\propto \exp\{\log p(z_i) + \mathbb{E}_{q_{-i}}[\log \prod_{k=1}^K p(x_i|\mu_k)^{z_{i,k}}]\} && (34)\\
\Rightarrow q^*(z_i = k) &\propto \exp\{\log \pi_k + \mathbb{E}_{\mu_k}[\log p(x_i|\mu_k)]\} && (35)\\
q^*(z_i = k) &\propto \exp\{\log \pi_k + x_i \mathbb{E}_{\mu_k}[\mu_k] - \mathbb{E}_{\mu_k}[\mu_k^2]/2\} && (36)\\
q^*(z_i = k) &\propto \exp\{\log \pi_k + x_i \tilde{\mu}_k - (\tilde{\mu}_k^2 + \tilde{\sigma}_k^2)/2\} && (37)\\
\Rightarrow \phi_{i,k} &\propto \exp\{\log \pi_k + x_i \tilde{\mu}_k - (\tilde{\mu}_k^2 + \tilde{\sigma}_k^2)/2\} && (38)
\end{aligned}
$$

Then we normalise over all K, resulting in

$$\phi_{i,k} = \frac{\exp\{\log \pi_k + x_i \tilde{\mu}_k - (\tilde{\mu}_k^2 + \tilde{\sigma}_k^2)/2\}}{\sum_{k=1}^K \exp\{\log \pi_k + x_i \tilde{\mu}_k - (\tilde{\mu}_k^2 + \tilde{\sigma}_k^2)/2\}} \tag{39}$$

Then, comes the *optimal variational distribution over* μ_k is

$$\begin{aligned}
q^*(\mu_k) &\propto \exp\{\mathbb{E}_{q_{-k}}[\log p(\mu_{1:K}, z_{1:N}, x_{1:N})]\} && (40)\\
&\propto \exp\{\log p(\mu_k) + \mathbb{E}_{q_{-k}}[\log p(x_{1:N}|\mu_{1:K}, z_{1:N})] + \text{const}\} && (41)\\
&\propto \exp\{\log p(\mu_k) + \sum_{i=1}^N \mathbb{E}_{q_{-k}}[\log \prod_{k=1}^K p(x_i|\mu_k)^{z_{i,k}}]\} && (42)\\
&\propto \exp\{\log p(\mu_k) + \sum_{i=1}^N \mathbb{E}_{q_{-k}}[z_{i,k} \log p(x_i|\mu_k)]\} + \text{const} && (43)\\
&\propto \exp\{\log p(\mu_k) + \sum_{i=1}^N \mathbb{E}_{q_{z_k}}[z_{i,k}] \mathbb{E}_{q_{-\mu_k}}[\log p(x_i|\mu_k)]\} && (44)\\
&\propto \exp\{\log p(\mu_k) + \sum_{i=1}^N \phi_{i,k} \log p(x_i|\mu_k)\} && (45)\\
&\propto \exp\{-\mu_k^2/2\tau^2 - \sum_{i=1}^N \phi_{i,k}(x_i - \mu_k)^2/2\sigma^2\} && (46)\\
&\propto \exp\{-1/2(\sum_{i=1}^N \phi_{i,k}/\sigma^2 + \tau^2)\mu_k^2 + \sum_{i=1}^N x_i \phi_{i,k}/\sigma^2 \mu_k + \text{const}\} && (47)
\end{aligned}$$

This expression is basically a Gaussian $q^*(\mu_k) \sim \mathcal{N}(\mu_k; \tilde{\mu}_k, \tilde{\sigma}_k^2)$ with:

$$\tilde{\mu}_k = \frac{\sum_{i=1}^N x_i \phi_{i,k}}{(\sum_{i=1}^N \phi_{i,k}/\sigma^2 + \tau^2)\sigma^2} \tag{48}$$

$$\tilde{\sigma}_k^2 = \frac{1}{(\sum_{i=1}^N \phi_{i,k}/\sigma^2 + \tau^2)\sigma^2} \tag{49}$$

Exercises

1. Draw the graphical model of the joint $p(z, x)$ in MoG under mean-field assumption.
2. Prove that Eq. (15) and Eq. (16) are equivalent
3. Compute the integral in Eq. (31)