keywords and concepts:

■ The Evidence Lower Bound

In variational inference, we specify a family $\mathcal{Q}$ of densities over the latent variables. (the parameter of interest).
Each $q(z) \in \mathcal{Q}$ is a candidate approximation to the exact conditional.
Our goal is to find the best candidate, the one closest in KL divergence to the exact conditional.

Inference now amounts to solving

$$q^*(z) = \arg\min_{q(z) \in \mathcal{Q}} KL(q(z) \| p(z|x))$$

once found, $q^*(\cdot)$ is the best approximation of the conditional, within the family $\mathcal{Q}$. The complexity of the family determines the complexity of this optimization.

However,

$$KL(q(z) \| p(z|x))$$

is not computable.

To see why,

$$KL(q(z) \| p(z|x))$$

$$= E_{p(z|x)}\left(\log \frac{q(z)}{p(z|x)}\right)$$

$$= E_{p(z|x)}\left[\log q(z) - \log p(z|x)\right)$$

$$= E_{p(z|x)}\left[\log q(z) - \log\left\{\frac{p(z,x)}{p(x)}\right\}\right]$$

$$= E_{p(z|x)}\left[\log q(z) - \log p(z,x) + \log p(x)\right]$$

$$= E_{p(z|x)}\left[\log q(z) - \log p(z,x)\right] + \log p(x)$$

But remember that we are choosing $q(z)$ to minimize KL divergence, so $\log p(x)$ doesn't really matter

thus, define

$$ELBO(q) = E[\log p(z,x)] - E[\log q(z)]$$

minimize KL is equivalent to minimize ELBO

Rewrite $ELBO(q)$

$$ELBO(q) = E[\log p(z,x)] - E[\log q(z)]$$

$$= E[\log p(z) \cdot p(x|z)] - E[\log q(z)]$$

$$= E[\log p(x|z)] + E[\log p(z)] - E[\log q(z)]$$

$$= E[\log p(x|z)] - E\left[\log \frac{q(z)}{p(z)}\right]$$

$$= E[\log p(x|z)] - KL[q(z) \| p(z)]$$

and from this expression, given a chosen q.

we encourage to put the parameter z on the place which explains the observed data,

and the second term encourage to choose parameter z which makes the divergence between q, the family and the prior, close.

■ Mean-field variational family

What family $\mathcal{L}$ shall be considered?

Mean-field variational family is a family of distributions such that the latent variables are mutually independent and each governed by distinct factors.

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

Motivating Example:

Bayesian mixture of Gaussians.

Consider a Bayesian mixture of unit-variance univarite Gaussians.

There are K mixture components, corresponding to K Gaussian distributions, with mean $M = (M_1 \cdots M_K)$.
The mean parameters are drawn from a common prior $P(M_K)$, which we assume to be Normal $(0, \sigma^2)$, the prior variance σ^2 is a hyper parameter.

To generate an observation X_i, you first choose a C_i from Categorical $(1/K \cdots 1/K)$ that is, C_i would be an indicator variable telling you X_i should come from a normal $(M_i, 1)$

full model:

$$M_K \sim N(0, \sigma^2)$$

$c_i \sim \text{Categorial}(1/k \cdots 1/k)$

$x_i | c_i, \mu \sim N(c_i^T u, I)$.

the latent variables are (μ, c)

for $\mu = \{\mu_1 \cdots \mu_k\}$

and

$c = \{c_1 \cdots c_k\}$

the posterior distribution for the parameter $z = \{\mu, c\}$ is

$$P(z|x) = \frac{P(z,x)}{P(x)}$$

In this case

$$P(z,x) = P(\mu, c, x)$$

$$= P(x|\mu, c) \cdot P(\mu, c).$$

$$= P(\mu) \prod_{i=1}^{n} P(c_i) P(x_i | c_i, \mu)$$

and

$$P(x) = \int P(\mu) \prod_{i=1}^{n} P(c_i) P(x_i | c_i, \mu) d\mu$$

Yet, this is very difficult to compute

and to be honest, you are not sure if the posterior of μ_i and c_i would be independent or not...

However, the mean-field family only consider using distribution factorizing the latent variables to approximate.

Consider $q(u, c) = \prod_{k=1}^{K} q(u_k, m_k, s_k^2) \cdot \prod_{i=1}^{n} q(c_i; u_i)$

Reason of choice.

① your prior for μ_i is normal, so you choose also normal for the q for μ.

② your c_i is discrete, so, you use a discrete/categorical distribution to serve as the approximation for c_i

③ you use some distributions because you know the posterior distribution for the (μ, c) is desired.

Following the mean-field recipe, each latent variable is governed by its own variational factor (that is, parameters to characterize the variational distribution).

The factors $q(\mu_k; m_k, s_k^2)$ is a Gaussian distribution on the kth mixture component's mean parameter; its mean is m_k and its variance is s_k^2; The factor $q(c_i, \varphi_i)$ is a distribution on the ith observation's mixture assignment; its assignment probabilities are a K-vector φ_i

Thus, the problem becomes: optimize the ELBO by choosing variational parameters! and then the variational parameters fully characterize

your posterior distribution.

↓

thoughts:

Now you cannot directly see the data in your variational distribution, so when the data influences your posterior? your data enter your posterior when you do the optimization, the data shows in your ELBO.

The flaws of mean-field approximation:

The mean-field family is expressive because it can capture any marginal density density of the latent variables. However, it cannot capture correlation between them.

↓ well... because you make them independent by assumption.

Coordinate Ascent Variational Inference

The main objective is to optimize the ELBO in the mean field variational inference, or equivalently, to choose the variational factors that maximize the ELBO.

prerequisite:

Probability chain Rule

$$p(x_n, \dots, x_1) = p(x_n | x_{n-1} \dots x_1) \cdot p(x_{n-1} \dots x_1)$$

thus

$$\begin{aligned} p(x_n, \dots, x_1) = & p(x_n | x_{n-1} \dots x_1) \\ & \cdot p(x_{n-1} | x_{n-2} \dots x_1) \\ & \cdot p(x_{n-2} | x_{n-3} \dots x_1) \\ & \vdots \end{aligned}$$

thus, in general:

$$p\left(\bigcap_{k=1}^{n} x_k\right) = \prod_{k=1}^{n} p\left(x_k \mid \bigcap_{j=1}^{k-1} x_j\right)$$

and Recall

$$ELBO(q) = E[\log p(z,x)] - E[\log q(z)]$$

and now we apply the probability chain rule to the m variation parameter.

(1)

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \cdot p(z_{1:m} \mid x_{1:n})$$

$$= p(x_{1:n}) \cdot \prod_{j=1}^{m} p(z_j \mid z_{1:(j-1)}, x_{1:n})$$

(2)

also, because we assume for mean field family

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \text{ by construction,}$$

thus,

$$E[\log q(z_{1:m})]$$

$$= E[\log \prod_{j=1}^{m} q_j(z_j)]$$

$$= \sum_{j=1}^{m} E\left[\log q_j(z_j)\right]$$

therefore

$$ELBO(q) = E_q\left[\log p(x_{1:n}, z_{1:m})\right] - E\left[\log q(z_{1:m})\right]$$

$$= E_q\left[\log\left(p(x_{1:n}) \prod_{j=1}^{m} p(z_j \mid z_{1:j-1}, x_{1:n})\right)\right]$$

$$- \sum_{j=1}^{m} E_{q_j}\left[\log q(z_j)\right]$$

$$= \log p(x_{1:n}) + \sum_{j=1}^{m} E_q\left[\log p(z_j \mid z_{1:j-1}, x_{1:n})\right]$$

$$- \sum_{j=1}^{m} E_{q_j}\left[\log q_j(z_j)\right]$$

Coordinate ascent method is to optimize the variational factor one by one while holding other variational factor fixed

Since you optimize each variation factor one by one, thus for each update, you only focus on one $q_j(z_j)$

$$\underset{q_j}{\arg\max}\ ELBO(q)$$

$$= \underset{q_j}{\arg\max} \left(\log p(x_{1:n}) + \sum_j^m E_q\left[\log p(z_j \mid z_{1:j-1}, x_{1:n})\right] - \sum_{j=1}^m E_{q_j}\left[\log q(z_j)\right] \right)$$

$$= \underset{q_j}{\arg\max} \left(E_q\left[\log p(z_j \mid z_{\neq j}, x)\right] - E_{q_j}\left[\log q(z_j)\right] \right)$$

$$= \underset{q_j}{\operatorname{argmax}} \int q_j(z_j) \log p(z_j \mid z_{\neq j}, x) \, dz_j$$

$$- \int q(z_j) \log q(z_j) \, dz_j$$

To find argmax, we take derivative with respect to $q(z_j)$

to get

$$\mathcal{L} = \int q(z_j) \log p(z_j \mid z_{\neq j}, x) \, dz_j$$

$$- \int q(z_j) \cdot \log q(z_j) \, dz_j$$

$$\mathcal{L} = \log p(z_j \mid z_{\neq j}, x) \cdot \int q(z_j) \, dz_j$$

$$- \int q(z_j) \log q(z_j) \, dz_j$$

thus (Interchange the order of differentiation and integration)

$$\frac{d\mathcal{L}}{dq(z_j)} = \int \log p(z_j | z_{\neq j}, x)\, dz_j$$

$$- \int \log q(z_j)\, dz_j$$

$$- \int q(z_j) \frac{1}{q(z_j)}\, dz_j$$

$$= E[\log p(z_j | z_{\neq j}, x)]$$

$$- \log q(z_j) - 1 = 0$$

Reorganize.

$$\Rightarrow E[\log p(z_j | z_{\neq j}, x)]$$

$$= \log\left(q(z_j) / e\right)$$

→ take exponential

$$\exp\{E[\log p(z_j | z_{\neq j}, x)]$$

$$= \frac{q(z_j)}{e}$$

thus, the update rule is

$$q^*(z_j) \propto e \cdot \exp\{E[\log p(z_j | z_{\neq j}, x]\}$$

$$\propto \exp[E_{q_{\neq j}}[\log p(z_j, z_{\neq j}, x)]$$

Takeaway message:

- using simpler distribution to approximate posterior
- How to measure the closeness between posterior and proposed distribution?
  - KL divergence and ELBO?
- different ways to see ELBO
- mean field family / flaws
- what is variational parameters?
- the form of coordinate ascent method
- probability chain rule