

Lesson six: Variational Bayes

Richard Yi Da Xu

School of Computing & Communication, UTS

March 9, 2016

Log-likelihood and Evidence Lower Bound (ELOB)

- It is universally true that:

$$\ln(p(X)) = \ln(p(X, Z)) - \ln(p(Z|X))$$

- It's also true (a bit silly) that:

$$\ln(p(X)) = [\ln(p(X, Z)) - \ln(q(Z))] - [\ln(p(Z|X)) - \ln(q(Z))]$$

- The above is so that we can insert an arbitrary pdf $q(Z)$ into, now we get:

$$\ln(p(X)) = \ln\left(\frac{p(X, Z)}{q(Z)}\right) - \ln\left(\frac{p(Z|X)}{q(Z)}\right)$$

- Taking the expectation on both sides, given $q(Z)$:

$$\begin{aligned}\ln(p(X)) &= \int q(Z) \ln\left(\frac{p(X, Z)}{q(Z)}\right) dZ - \int q(Z) \ln\left(\frac{p(Z|X)}{q(Z)}\right) dZ \\ &= \underbrace{\int q(Z) \ln(p(X, Z)) dZ}_{\mathcal{L}(q)} - \underbrace{\int q(Z) \ln(q(Z)) dZ + \left(-\int q(Z) \ln\left(\frac{p(Z|X)}{q(Z)}\right) dZ\right)}_{\mathbb{KL}(q||p)} \\ &= \mathcal{L}(q) + \mathbb{KL}(q||p)\end{aligned}$$

We often see the following alternative derivation:

$$\begin{aligned}\ln(p(X)) &= \log \int_Z p(X, Z) dz \\ &= \log \int_Z p(X, Z) \frac{q(Z)}{q(Z)} dz \\ &= \log \left(\mathbb{E}_q \left[\frac{p(X, Z)}{q(Z)} \right] \right) \\ &\geq \mathbb{E}_q \left[\log \left(\frac{p(X, Z)}{q(Z)} \right) \right] \text{ using Jensen's inequality} \\ &= \mathbb{E}_q [\log(p(X, Z))] - \mathbb{E}_q [\log(q(Z))] \\ &\triangleq \mathcal{L}(q)\end{aligned}$$

It can be proven easily that the “missing” part, i.e., $\ln(p(X)) - \mathcal{L}(q) = \mathbb{KL}(q||p)$.

Maximize Evidence Lower Bound (ELOB)

$$\ln(p(X)) = \mathcal{L}(q) + \mathbb{KL}(q\|p)$$

- ▶ We can give a name to both terms:

Evidence Lower Bound (ELOB): $\mathcal{L}(q) = \int q(Z) \ln(p(X, Z)) dZ - \int q(Z) \ln(q(Z)) dZ$

KL divergence: $\mathbb{KL}(q\|p) = \int q(Z) \ln\left(\frac{p(Z|X)}{q(Z)}\right) dZ$

- ▶ Notice $p(X)$ is fixed with respect to the choice of $q(Z)$. We wanted to choose a $q(Z)$ function that minimize KL divergence, so that $q(Z)$ becomes closer and closer to $p(Z|X)$. Of course, let's see what happens when $q(Z) = p(Z|X)$:

$$\mathbb{KL}(q\|p) = - \int p(Z|X) \ln\left(\frac{p(Z|X)}{p(Z|X)}\right) dZ = 0$$

- ▶ We know that $p(X) = \mathcal{L}(q) + \mathbb{KL}(q\|p)$. Minimizing $\mathbb{KL}(q\|p)$ is the same as maximizing the Evidence Lower Bound $\mathcal{L}(q)$.

- ▶ Arbitrary (i.e., non Exponential Family) distributions
- ▶ No maximizing variational distribution parameters

The choice of $q(Z)$

- Suppose let's choose $q(Z)$, such that:

$$q(Z) = \prod_{i=1}^M q_i(Z_i)$$

- Substitute this choice into Evidence Lower Bound (ELOB):

$$\begin{aligned}\mathcal{L}(q) &= \int q(Z) \ln(p(X, Z)) dZ - \int q(Z) \ln(q(Z)) dZ \\ &= \underbrace{\int \prod_{i=1}^M q_i(Z_i) \ln(p(X, Z)) dZ}_{\text{part (1)}} - \underbrace{\int \prod_{i=1}^M q_i(Z_i) \sum_{i=1}^M \ln(q_i(Z_i)) dZ}_{\text{part (2)}}\end{aligned}$$

Simplification of (Part 1):

$$(\text{Part 1}) = \int \prod_{i=1}^M q_i(Z_i) \ln(p(X, Z)) dZ$$

$$\int_{Z_1} \int_{Z_2} \dots \int_{Z_M} \prod_{i=1}^M q_i(Z_i) \ln(p(X, Z)) dZ_1, dZ_2, \dots, dZ_M$$

- Rearrange the expression by taking a particular $q_j(Z_j)$ out of the integral:

$$(\text{Part 1}) = \int_{Z_j} q_j(Z_j) \left(\int_{Z_{i \neq j}} \dots \int \prod_{i \neq j}^M q_i(Z_i) \ln(p(X, Z)) \prod_{i \neq j}^M dZ_i \right) dZ_j$$

- or, more compactly:

$$(\text{Part 1}) = \int_{Z_j} q_j(Z_j) \left(\int_{Z_{i \neq j}} \dots \int \ln(p(X, Z)) \prod_{i \neq j}^M q_i(Z_i) dZ_i \right) dZ_j$$

- or, even more meaningfully, it can be put into an expectation function, and since $\prod_{i \neq j}^M q_i(Z_i)$ is a joint probability density

$$(\text{Part 1}) = \int_{Z_j} q_j(Z_j) \left[\mathbb{E}_{i \neq j} \left[\ln(p(X, Z)) \right] \right] dZ_j$$

Simplification of (Part 2):

$$(\text{Part 2}) = \int \prod_{i=1}^M q_i(Z_i) \sum_{i=1}^M \ln(q_i(Z_i)) dZ$$

- Note that the above needs to integrate out all $Z = \{z_1, \dots, z_M\}$, which is quite daunting. However, notice that each term in the sum, $\sum_{i=1}^M \ln(q_i(Z_i))$ involves only a single i , therefore, we are able to simplify the above into the following:

$$(\text{Part 2}) = \sum_{i=1}^M \left(\int q_i(Z_i) \ln(q_i(Z_i)) dZ_i \right)$$

- For a particular $p_j(Z_j)$, the rest of the sum can be treated like a constant, part 2 can be written as:

$$(\text{Part 2}) = \int_{Z_j} q_i(Z_i) \ln(q_i(Z_i)) dZ_j + \text{const.}$$

where const. are the term does not involve Z_j .

Putting Part (1) and Part (2) together:

$$\mathcal{L}(q) = \text{Part (1)} - \text{Part (2)} = \int_{Z_j} q_j(Z_j) \mathbb{E}_{i \neq j} \left[\ln(p(X, Z)) \right] dZ_j - \int_{Z_j} q_j(Z_j) \ln(q_j(Z_j)) dZ_j + \text{const.}$$

- Note that $\mathbb{E}_{i \neq j} [\ln(p(X, Z))]$ would be some $\ln[p(Z_j)]$, we name it $\ln(\tilde{p}_j(X, Z_j))$, i.e.,:

$$\ln(\tilde{p}_j(X, Z_j)) = \mathbb{E}_{i \neq j} [\ln(p(X, Z))]$$

- Or equivalently we can express Evidence Lower Bound (ELOB) in terms of:

$$\mathcal{L}(q_j) = \int_{Z_j} q_j(Z_j) \ln \left[\frac{\tilde{p}_j(X, Z_j)}{q_j(Z_j)} \right] + \text{const.}$$

This is the same as $-\text{KL} \left(\mathbb{E}_{i \neq j} [\ln(p(X, Z))] \parallel q_i(Z_i) \right)$

- **This is the key:** We can maximize ELOB, or $\mathcal{L}(q)$, by minimizing this special KL divergence, where we can find approximate and optimal $q_i^*(Z_i)$, such that:

$$\ln(q_i^*(Z_i)) = \mathbb{E}_{i \neq j} [\ln(p(X, Z))]$$

Example: Gaussian-Gamma Conjugate prior

- ▶ Let $\mathcal{D} = \{x_1, \dots, x_n\}$:

$$\begin{aligned} p(\mathcal{D}|\mu, \tau) &= \prod_{i=1}^n \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau}{2}(x_i - \mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(\frac{-\tau}{2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

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

$$p(\tau) = \text{Gamma}(\tau|a_0, b_0) \propto \tau^{a_0-1} \exp^{-b_0\tau}$$

- ▶ Complete data-likelihood is:

$$p(\mathcal{D}, \mu, \tau) = p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau)$$

Of course, due to conjugacy, the solution can be found exactly:

$$p(\mu, \tau | \mathbf{d}) \propto p(\mathcal{D} | \mu, \tau) p(\mu | \tau) p(\tau) = \mathcal{N}(\mu_n, (\lambda_n \tau)^{-1}) \text{Gamma}(\tau | a_n, b_n)$$

where:

$$\mu_n = \frac{\lambda_0 \mu_0 + n \bar{x}}{\lambda_0 + n}$$

$$\lambda_n = \lambda_0 + n$$

$$a_n = a_0 + n/2$$

$$b_n = b_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\lambda_0 n (\bar{x} - \mu_0)^2}{2(\lambda_0 + n)}$$

However, for demo purpose, we assume $q(\mu, \tau)$:

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

Computing $\ln(q_\mu^*(\mu)) = \mathbb{E}_{q_\tau(\tau)} [\ln(p(\mu, \tau|\mathcal{D}))]$ (1)

$$\begin{aligned}\ln(q_\mu^*(\mu)) &= \mathbb{E}_{q_\tau} [\ln(p(\mu, \tau|\mathcal{D}))] \\&= \mathbb{E}_{q_\tau} [\ln(p(\mathcal{D}|\mu, \tau)) + \ln p(\mu|\tau)] + \text{const.} \quad \text{remove terms do NOT contain } \mu \\&= \mathbb{E}_{q_\tau} \left[\underbrace{-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2}_{\ln(p(\mathcal{D}|\mu, \tau))} + \underbrace{\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2}_{\ln p(\mu|\tau)} \right] + \text{const.} \\&= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \underbrace{\left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]}_{\text{terms contain } \mu \text{ but does not contain } \tau} + \text{const.}\end{aligned}$$

Completing the square for the μ terms:

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 &= n\mu^2 - 2n\mu\bar{x} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \text{const.} \\&= (n + \lambda_0)\mu^2 - 2\mu(n\bar{x} + \lambda_0\mu_0) = (n + \lambda_0) \left(\mu^2 - \frac{2\mu(n\bar{x} + \lambda_0\mu_0)}{(n + \lambda_0)} \right) \\&= (n + \lambda_0) \left(\mu - \frac{(n\bar{x} + \lambda_0\mu_0)}{(n + \lambda_0)} \right)^2 + \text{const.}\end{aligned}$$

Computing $\ln(q_\mu^*(\mu)) = \mathbb{E}_{q_\tau(\tau)} [\ln(p(\mu, \tau|\mathcal{D}))]$ (2)

Therefore, we have:

$$\begin{aligned}\ln(q_\mu^*(\mu)) &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2 \right] + \text{const.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau](n + \lambda_0)}{2} \left(\mu - \frac{(n\bar{x} + \lambda_0\mu_0)}{(n + \lambda_0)} \right)^2 + \text{const.} \\ &= \mathcal{N} \left(\frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0}, \mathbb{E}_{q_\tau}[\tau](n + \lambda_0) \right)\end{aligned}$$

Computing $\ln(q_i^*(\tau)) = \mathbb{E}_{q_\mu(\mu)} [\ln(p(\mu, \tau|\mathcal{D}))]$ (1)

$$\begin{aligned}\ln(q_\tau^*(\tau)) &= \mathbb{E}_{q_\mu} [\ln(p(\mu, \tau|\mathcal{D}))] \\ &= \mathbb{E}_{q_\mu} [\ln(p(\mathcal{D}|\mu, \tau)) + \ln p(\mu|\tau) + \ln p(\tau)] + \text{const.} \\ &= \mathbb{E}_{q_\mu} \left[\underbrace{\frac{n}{2} \ln(\tau) - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2}_{\ln(p(\mathcal{D}|\mu, \tau))} - \underbrace{\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2}_{\ln p(\mu|\tau)} + \underbrace{(a_0 - 1) \ln(\tau) - b_0 \tau}_{\ln p(\tau)} \right] + \text{const.}\end{aligned}$$

Bring terms without μ outside of the integral:

$$\begin{aligned}&= \frac{n}{2} \ln(\tau) + (a_0 - 1) \ln(\tau) - b_0 \tau - \frac{\tau}{2} \mathbb{E}_{q_\mu(\mu)} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + \text{const.} \\ &= \left(\underbrace{\frac{n}{2} + a_0 - 1}_{a_n} \right) \ln(\tau) - \tau \left(\underbrace{b_0 + \frac{1}{2} \mathbb{E}_{q_\mu(\mu)} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]}_{b_n} \right) + \text{const.}\end{aligned}$$

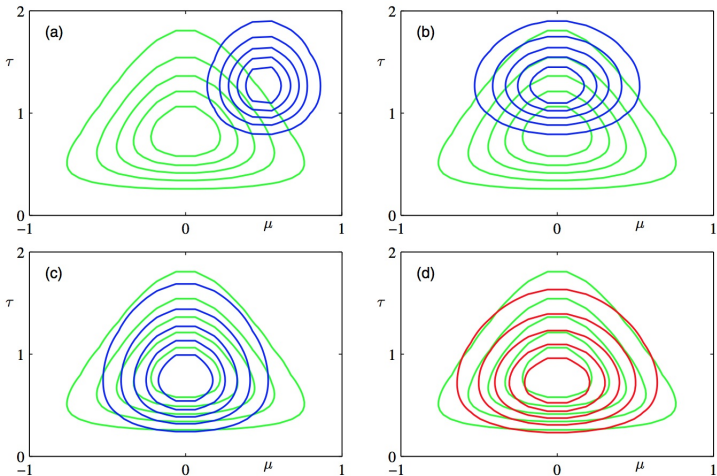
Computing $\ln(q_i^*(\tau)) = \mathbb{E}_{q_\mu(\mu)} [\ln(p(\mu, \tau|\mathcal{D}))]$ (2)

We can rewrite,

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

We will compute $\mathbb{E}_{q_\mu} [\mu]$ and $\mathbb{E}_{q_\mu} [\mu^2]$ since we know of $q_\mu(\mu)$ from previously.

Figure and Demo



Also see Matlab Demo!

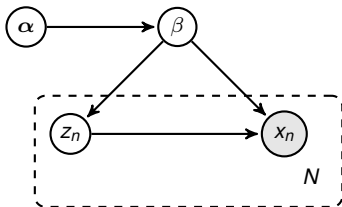
- ▶ Perform direct maximisation over $q(Z)$ using variational calculus

$$\begin{aligned} G(f_x(x), f_y(y)) &= \int_x \int_y f_x(x) f_y(y) \ln p(x, y) dx dy \implies \\ \frac{\delta G(f_x(x), f_y(y))}{\delta f_x(x')} &= \int_x \int_y \frac{\delta f_x(x)}{\delta f_x(x')} f_y(y) \ln p(x, y) dx dy \\ &= \int_x \int_y \delta(x - x') f_y(y) \ln p(x, y) dx dy \\ &= \int_y f_y(y) \ln p(x', y) dy \end{aligned}$$

Let $Z \triangleq (z, \theta)$ $p(X, Z, \theta) = p(X, Z|\theta)p(\theta)$ $q(Z, \theta) = q_z(Z)q_\theta(\theta)$:
Evidence Lower Bound (ELBO) can be written as:

- ▶ Exponential Family distributions
- ▶ Maximizing variational distribution parameters
- ▶ Introduction to Stochastic Variational Inference (SVI)

Problem to consider:



$$p(X, Z, \beta | \alpha) = p(\beta | \alpha) \prod_{n=1}^N p(x_n, z_n | \beta)$$
$$p(x_n, z_n | x_{-n}, z_{-n}, \beta, \alpha) = p(x_n, z_n | \beta, \alpha)$$

All posterior is based on Exponential family:

$$p(\beta | X, Z, \alpha) = h(\beta) \exp \left\{ \eta_g(X, Z, \alpha)^T t(\beta) - A_g(\eta_g(X, Z, \alpha)) \right\}$$
$$p(z_{n,j} | x_n, z_{n,-j}, \beta) = h(z_{n,j}) \exp \left\{ \eta_l(x_n, z_{n,-j}, \beta)^T t(z_{n,j}) - A_l(\eta_l(x_n, z_{n,-j}, \beta)) \right\}$$

Let's look at some Important Distributions: **Exponential Family**

Most of the distributions we are going to look at are from **exponential family**
exponential family can be expressed in terms of its natural parameters:

$$h(x) \exp \left(T(x)^T \eta - A(\eta) \right)$$

Think about why is this representation useful?

Always have in mind ask yourself where are the **support** of these distributions, i.e., where $p(X) > 0$?

More about Gaussian 1-d: Natural Parameter Representation

$$\begin{aligned}\mathcal{N}(x; \mu, \sigma^2) &= (2\pi\sigma^2)^{-1/2} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \\&= \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right) \\&= \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right) \\&= \exp\left(\underbrace{\begin{bmatrix} x \\ x^2 \end{bmatrix}}_{T(X)}^T \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right)\end{aligned}$$

$$\blacktriangleright \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \quad A(\eta) = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-2\eta_2)$$

$$\blacktriangleright \eta_2 = -\frac{1}{2\sigma^2} \implies \sigma^2 = -\frac{1}{2\eta_2} \quad \mu = \eta_1 \sigma^2 = \eta_1 \frac{-1}{2\eta_2} = \frac{-\eta_1}{2\eta_2}$$

More about Gaussian 1-d: Natural Parameter Representation (2)

► Reverse is: $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}$

$$\begin{aligned}\tilde{\mathcal{N}}(x, \eta) &= \exp \left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right) \\ &= \exp \left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} - \frac{\left(\frac{-\eta_1}{2\eta_2}\right)^2}{2\left(\frac{-1}{2\eta_2}\right)} - \frac{1}{2} \ln \left(2\pi \left(\frac{-1}{2\eta_2} \right) \right) \right) \\ &= \exp \left(T(x)^T \eta + \frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln \left(\frac{2\pi}{-2\eta_2} \right) \right) \\ &= \exp \left(T(x)^T \eta + \frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \ln(-2\eta_2) - \frac{1}{2} \ln(2\pi) \right)\end{aligned}$$

$$\mathcal{N}_{\text{nat}}(x, \eta) = \exp \left(T(x)^T \eta - \underbrace{\left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-2\eta_2) \right)}_{A(\eta)} - \frac{1}{2} \ln(2\pi) \right)$$

conditional per (x_n, z_n) : $p(\beta|x_n, z_n)$

- Prior:

$$p(\beta) = h(\beta) \exp\{\alpha^T t(\beta) - A_g(\alpha)\} = \underbrace{\exp(-A_g(\alpha))}_{\text{normalization}} h(\beta) \exp\{\alpha^T t(\beta)\}$$

$$\Rightarrow \int_{\beta} h(\beta) \exp\{\alpha^T t(\beta)\} = \exp(A_g(\alpha))$$

Let sufficient statistics $t(\beta) = [\beta, \underbrace{-A_I(\beta)}_{\text{same}}]^T \Rightarrow \alpha = [\alpha_1 \ \alpha_2]^T$:

$$p(\beta) = h(\beta) \exp \left\{ [\alpha_1, \alpha_2]^T [\beta, \underbrace{-A_I(\beta)}_{\text{same}}] - A_g(\alpha) \right\}$$

- Likelihood density per (x_n, z_n) is:

$$p(x_n, z_n|\beta) = h(x_n, z_n) \exp \left\{ t(x_n, z_n)\beta - \underbrace{A_I(\beta)}_{\text{same}} \right\}$$

- Posterior is:

$$\begin{aligned} p(\beta|x_n, z_n, \alpha) &\propto \underbrace{h(\beta) \exp\{\alpha^T t(\beta)\}}_{\text{prior}} \underbrace{\exp\{t(x_n, z_n)\beta - A_I(\beta)\}}_{\text{likelihood}} \\ &= h(\beta) \exp\{(\alpha_1 + t(x_n, z_n))\beta - \alpha_2 A_I(\beta) - A_I(\beta)\} \\ &= h(\beta) \exp\{(\alpha_1 + t(x_n, z_n))\beta - (\alpha_2 + 1)A_I(\beta)\} \\ &= h(\beta) \exp\{[\alpha_1 + t(x_n, z_n) \quad \alpha_2 + 1]^T t(\beta)\} \end{aligned}$$

The complete posterior

- Complete likelihood:

$$p(X, Z|\beta) = \prod_{n=1}^N h(x_n, z_n) \exp\{\beta^T t(x_n, z_n) - A_l(\beta)\} = h(X, Z) \exp\left\{\sum_{n=1}^N \beta^T t(x_n, z_n) - N \times A_l(\beta)\right\}$$

- Complete posterior:

Since: $p(\beta|x_n, z_n, \alpha) \propto h(\beta) \exp\{[\alpha_1 + t(x_n, z_n) \quad \alpha_2 + 1]^T t(\beta)\}$:

$$\implies p(\beta|X, Z, \alpha) \propto h(\beta) \exp\left\{\left[\alpha_1 + \sum_{n=1}^N t(x_n, z_n) \quad \alpha_2 + N\right]^T t(\beta)\right\}$$

When we use the expression:

$$\begin{aligned} p(\beta|X, Z, \alpha) &= h(\beta) \exp\{\eta_g(X, Z, \alpha)^T t(\beta) - A_g(\eta_g(X, Z, \alpha))\} \\ \implies \eta_g(X, Z, \alpha) &= \begin{bmatrix} \alpha_1 + \sum_{n=1}^N t(x_n, z_n) & \alpha_2 + N \end{bmatrix} \\ \implies A_g(\eta_g(X, Z, \alpha)) &= \int_{\beta} h(\beta) \exp\{\eta_g(X, Z, \alpha)^T t(\beta)\} \end{aligned}$$

Example: Posterior of Gaussian mean (1)

To summarise:

$$\begin{aligned} p(\beta|X, Z, \alpha) &\propto h(\beta) \exp \left\{ [\hat{\alpha}_1 \quad \hat{\alpha}_2]^T [\beta \quad -A_I(\beta)] \right\} \\ &= h(\beta) \exp \left\{ \left[\underbrace{\alpha_1 + \sum_{n=1}^N t(x_n, z_n)}_{\hat{\alpha}_1} \quad \underbrace{\alpha_2 + N}_{\hat{\alpha}_2} \right]^T t(\beta) \right\} \end{aligned}$$

- Example: suppose data x_i come from unit variance Gaussian:

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x - \mu)^2 \right\} = \underbrace{\frac{\exp(-x^2/2)}{\sqrt{2\pi}}}_{h(x)} \exp \left\{ \underbrace{\mu}_{\beta} \underbrace{x}_{t(x)} - \underbrace{\frac{\mu^2}{2}}_{A_I(\beta)} \right\}$$

- So we have:

$$\begin{aligned} \beta &= \mu & t(x) &= x \\ A_I(\beta) &= \frac{\beta^2}{2} & h(x) &= \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \end{aligned}$$

Example: Posterior of Gaussian mean (2)

$$p(x|\mu) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \exp\left\{\mu x - \frac{\mu^2}{2}\right\} = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \exp\left\{\beta x + \underbrace{-\frac{\beta^2}{2}}_{A_I(\beta)}\right\}$$

- For the above likelihood, its conjugate prior MUST be in the form of:

$$\begin{aligned} p(\beta|\alpha) &= h(\beta) \exp\left\{\alpha_1 \beta + \alpha_2 \underbrace{(-\beta^2/2)}_{A_I(\beta)} - A_g(\alpha)\right\} \\ &= h(\beta) \exp\left\{\left[\alpha_1 \quad -\frac{\alpha_2}{2}\right]^T \begin{bmatrix} \beta & \beta^2 \end{bmatrix} - A_g(\alpha)\right\} \end{aligned}$$

Example: Posterior of Gaussian mean (3)

- ▶ Since we know,

$$p(\beta|\alpha) = h(\beta) \exp \left\{ [\alpha_1 \quad \alpha_2/2]^T \begin{bmatrix} \beta & \beta^2 \end{bmatrix} - A_g(\alpha) \right\}$$

From our knowledge, a distribution with $t(\beta) = [\beta \quad \beta^2]$ is a Gaussian.

- ▶ Suppose the data come from an exponential family. Every exponential family has a conjugate prior in theory.
- ▶ The natural parameter $\alpha = [\alpha_1, \alpha_2]$ has dimension $\dim(\beta) + 1$.
- ▶ The sufficient statistics of the prior are $[\beta, -A_l(\beta)]$

For exponential family distribution: $\mathbb{E}_q[t(\beta)] = \nabla_\lambda A_g(\lambda)$

Given $q(\beta|\lambda) = h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} = \frac{1}{\exp(A_g(\lambda))} h(\beta) \exp\{\lambda^T t(\beta)\}$

Why $\mathbb{E}_q[t(\beta)] = \nabla_\lambda A_g(\lambda)$

$$\int_{\beta} q(\beta|\lambda) d\beta = \int_{\beta} h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} d\beta = 0$$

$$\implies \nabla_\lambda \left(\int_{\beta} h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} d\beta \right) = 0$$

$$\implies \int_{\beta} \nabla_\lambda \left(h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} \right) d\beta = 0$$

$$\implies \int_{\beta} \left(h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} \right) (t(\beta) - \nabla_\lambda A_g(\lambda)) d\beta = 0$$

$$\implies \int_{\beta} \left(h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} \right) t(\beta) d\beta - \int_{\beta} \left(h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\} \right) \nabla_\lambda A_g(\lambda) d\beta = 0$$

$$\implies \mathbb{E}_q[t(\beta)] - \nabla_\lambda A_g(\lambda) = 0$$

The choice of $q(\beta, Z)$

We choose $q(\beta, Z)$ to decouple β and Z completely:

$$q(\beta, Z) = q(\beta|\lambda) \prod_{n=1}^N \prod_{j=1}^J q(z_{n,j}|\phi_{n,j})$$

- ▶ $q(\beta|\lambda)$ is the SAME distribution type as $p(\beta|X, Z, \alpha)$, they only differ in parameter

$$q(\beta|\lambda) = h(\beta) \exp\{\lambda^T t(\beta) - A_g(\lambda)\}$$

compare with:
$$p(\beta|X, Z, \alpha) = h(\beta) \exp\{\eta_g(X, Z, \alpha)^T t(\beta) - A_g(\eta_g(X, Z, \alpha))\}$$

- ▶ $q(z_{n,j}|\phi_{n,j})$ is the SAME distribution type as $p(z_{n,j}|x_n, z_{n,-j}, \beta)$, they only differ in parameter

$$q(z_{n,j}|\phi_{n,j}) = h(z_{n,j}) \exp\{\phi_{n,j}^T t(z_{n,j}) - A_l(\phi_{n,j})\}$$

compare with:
$$p(z_{n,j}|x_n, z_{n,-j}, \beta) = h(z_{n,j}) \exp\{\eta_l(x_n, z_{n,-j}, \beta)^T t(z_{n,j}) - A_l(\eta_l(x_n, z_{n,-j}, \beta))\}$$

Update for $\mathcal{L}(\lambda)$

We need to maximize the ELBO, i.e.,

$$\mathcal{L}(q) \triangleq \mathbb{E}_q[\log p(X, Z, \beta | \alpha)] - \mathbb{E}_q[\log q(Z, \beta)]$$

Note that q used here is $q(\beta, Z)$ not just $q(\beta | \lambda)$

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{E}_q[\log p(\beta | X, Z, \alpha)] + \mathbb{E}_q[\log p(X, Z)] - \mathbb{E}_q[\log q(\beta)] \\ &= \mathbb{E}_q[\log p(\beta | X, Z, \alpha)] - \mathbb{E}_q[\log q(\beta)] + \text{const.} \\ &= \mathbb{E}_q \left[\log \left(h(\beta) \exp \{ \eta_g(x, z, \alpha)^T t(\beta) - A_g(\eta_g(x, z, \alpha)) \} \right) \right] - \mathbb{E}_q[\log q(\beta)] + \text{const.} \\ &= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_q[\eta_g(x, z, \alpha)^T t(\beta)] - \mathbb{E}_q[\log h(\beta) \exp \{ \lambda^T t(\beta) - A_g(\lambda) \}]}_{\text{}} + \text{const.} \\ &= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \mathbb{E}_{q(\beta|\lambda)}[t(\beta)] - \mathbb{E}_q[\log h(\beta)] - \mathbb{E}_q[\lambda^T t(\beta)] + A_g(\lambda)}_{\text{}} + \text{const.} \\ &= \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \mathbb{E}_q[t(\beta)] - \lambda^T \mathbb{E}_q[t(\beta)] + A_g(\lambda) + \text{const.}\end{aligned}$$

Substitute $\mathbb{E}_q[t(\beta)] = \nabla_\lambda A_g(\lambda)$:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \nabla_\lambda A_g(\lambda) - \lambda^T \nabla_\lambda A_g(\lambda) + A_g(\lambda) + \text{const.}$$

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \nabla_\lambda A_g(\lambda) - \lambda^T \nabla_\lambda A_g(\lambda) + A_g(\lambda) + \text{const.}$$

Maximize $\mathcal{L}(\lambda)$ we get:

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \nabla_\lambda^2 A_g(\lambda) \underbrace{- \nabla_\lambda A_g(\lambda) - \lambda^T \nabla_\lambda^2 A_g(\lambda) + \nabla_\lambda A_g(\lambda)} = 0$$

$$= \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T \nabla_\lambda^2 A_g(\lambda) - \lambda^T \nabla_\lambda^2 A_g(\lambda) = 0$$

$$\implies \nabla_\lambda^2 A_g(\lambda) \left(\mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]^T - \lambda^T \right) = 0$$

$$\lambda = \mathbb{E}_{q(Z|\Phi)}[\eta_g(x, z, \alpha)]$$

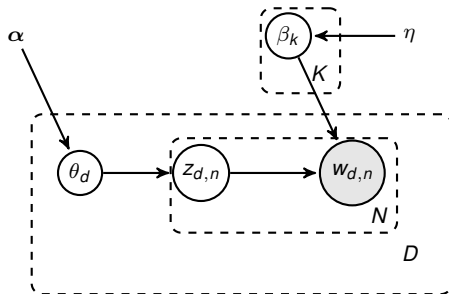
In a very similar fashion to $\mathcal{L}(\lambda)$, we can prove:

$$\nabla_{\phi_{n,j}} \mathcal{L}(\phi_{n,j}) = \nabla_{\phi_{n,j}}^2 \mathbf{A}_I(\phi_{n,j}) \left(\mathbb{E}_{q(\lambda)} [\eta_I(x_n, z_{n,-j}, \beta)]^T - \phi_{n,j}^T \right) = 0$$

$$\phi_{n,j} = \mathbb{E}_{q(\lambda)} [\eta_I(x_n, z_{n,-j}, \beta)]$$

Continue next time for Stochastic Maximization ...

Latent Dirichlet Allocation



- ▶ $\beta_k \sim \text{Dir}(\eta, \dots, \eta)$ for $k \in \{1, \dots, K\}$.
- ▶ For each document d :
 - $\theta \sim \text{Dir}(\alpha, \dots, \alpha)$
 - For each word $w \in \{1, \dots, N\}$:
 - $z_{dn} \sim \text{Mult}(\theta_d)$
 - $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$

Updating $q(z_{d,n}|\phi_{d,n})$: posterior conditional in Exponential Family form

$$\begin{aligned} p(z_{dn} = k | \theta_d, \beta_{1:K}, w_{d,n}) &\propto p(z_{d,n} = k | \theta_d) p(w_{d,n} | z_{d,n} = k, \beta_{1:K}) \\ &= \text{Mult}(\theta_{d,k}) \times \text{Mult}(\beta_{k,w_{d,n}}) \\ &\propto \exp \left(\underbrace{\log(\theta_{d,k}) + \log(\beta_{k,w_{d,n}})}_{\eta_l(\theta_d, \beta_{1:K}, w_{d,n})} \times \underbrace{1}_{t(z_{d,n})} \right) \end{aligned}$$

Updating $q(z_{d,n}|\phi_{d,n})$: optimize $\phi_{d,n}$

- ▶ $q(z_{d,n}) = \text{Mult}(\phi_{d,n})$ or $q(z_{d,n} = k) = \phi_{d,n}^k$ $q(\beta_k) = \text{Dir}(\lambda_k)$ $q(\theta_d) = \text{Dir}(\gamma_d)$
- ▶ Fact about Dirichlet Distribution $\theta \sim \text{Dir}(\gamma_1, \dots, \gamma_K) \implies \mathbb{E}[\log(\theta_k)|\gamma] = \Psi(\gamma_k) - \Psi(\sum_{i=1}^K \gamma_i)$
- ▶ In terms of natural parameter:

$$\begin{aligned}\eta(\phi_{d,n}^k) = \log(\phi_{d,n}^k) &\propto \mathbb{E}_{q(\theta_d)q(\beta_k)} [\eta_l(\theta_d, \beta_{1:K}, w_{d,n})] \\ &= \mathbb{E}_{q(\theta_d)} [\log(\theta_{d,k})] + \mathbb{E}_{q(\beta_k)} [\log(\beta_{k,w_{d,n}})] \\ &= \Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^K \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_v \lambda_{k,v}\right)\end{aligned}$$

- ▶ Change it back to traditional parameter:

$$\begin{aligned}\implies \phi_{d,n}^k &\propto \exp \left[\underbrace{\Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^K \gamma_{d,k}\right)}_{\text{irrelevant in proportionality}} + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_v \lambda_{k,v}\right) \right] \\ &\propto \exp \left[\Psi(\gamma_{d,k}) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_v \lambda_{k,v}\right) \right]\end{aligned}$$

Updating $q(\theta_d|\gamma_d)$: posterior conditional in Exponential Family form

$$\begin{aligned} p(\theta_d|Z_d) &= p(\theta_d|\alpha) \prod_{n=1}^N p(z_{d,n}|\theta_d) = \text{Dir}(\alpha) \times \prod_{n=1}^N \text{Mult}(z_{d,n}|\theta_d) \\ &= \prod_k \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\delta(z_{d,n},k)} \right) \\ &= \exp \left[\log \left(\prod_k \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\delta(z_{d,n},k)} \right) \right) \right] \\ &= \exp \left[\sum_k \log \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\delta(z_{d,n},k)} \right) \right] = \exp \left[\sum_k \left(\log \theta_{d,k}^{\alpha_k-1} + \sum_{n=1}^N \log \left(\theta_{d,k}^{\delta(z_{d,n},k)} \right) \right) \right] \\ &= \exp \left[\sum_k \left((\alpha_k - 1) \log \theta_{d,k} + \sum_{n=1}^N \delta(z_{d,n}, k) \log \theta_{d,k} \right) \right] \\ &= \exp \left[\sum_k \left(\alpha_k - 1 + \sum_{n=1}^N \delta(z_{d,n}, k) \right) \log (\theta_{d,k}) \right] \\ &= \exp \left(\underbrace{[(\alpha_1 - 1 + n_1) \dots (\alpha_K - 1 + n_K)]^T}_{\eta_l(\alpha, Z_d)} \underbrace{[\log(\theta_{d,1}) \dots \log(\theta_{d,K})]}_{t(\theta_d)} \right) \text{ by letting } n_k = \sum_{n=1}^N \delta(z_{d,n}, k) \end{aligned}$$

Updating $q(\theta_d|\gamma_d)$: optimize γ_d

- ▶ $q(z_{d,n}) = \text{Mult}(\phi_{d,n})$ or $q(z_{d,n} = k) = \phi_{d,n}^k$ $q(\beta_k) = \text{Dir}(\lambda_k)$ $q(\theta_d) = \text{Dir}(\gamma_d)$
- ▶ In terms of natural parameter:

$$\begin{aligned}\eta(\gamma_d) &= \mathbb{E}_{q(z_{d,n}|\phi_{d,n})} [\eta_l(\alpha, z_d)] \\ &= \mathbb{E}_{q(z_{d,n}|\phi_{d,n})} [[(\alpha_1 - 1 + n_1) \dots (\alpha_K - 1 + n_K)]] \\ &= \left[\left(\alpha_1 - 1 + \sum_{n=1}^N \delta(z_{d,n}, 1) \phi_{d,n}^1 \right) \dots \left(\alpha_K - 1 + \sum_{n=1}^N \delta(z_{d,n}, K) \phi_{d,n}^K \right) \right]\end{aligned}$$

- ▶ Change it back to traditional parameter:

$$\begin{aligned}\gamma_d &= \left[\left(\alpha_1 + \sum_{n=1}^N \delta(z_{d,n}, 1) \phi_{d,n}^1 \right) \dots \left(\alpha_K + \sum_{n=1}^N \delta(z_{d,n}, K) \phi_{d,n}^K \right) \right] \\ &= \alpha + \sum_{n=1}^N \phi_{d,n}\end{aligned}$$

Updating $q(\beta_k|\lambda_k)$ posterior conditional in Exponential Family form:

$$\begin{aligned} p(\beta_k|Z, W) &= p(\beta_k|\eta) \prod_{d=1}^D \prod_{n=1}^N p(w_{d,n}|\beta_k)^{\delta(z_{d,n},k)} = \text{Dir}(\eta) \times \prod_{d=1}^D \prod_{n=1}^N \beta_k^{w_{d,n}\delta(z_{d,n},k)} \\ &\propto \exp \left(\underbrace{\left(\eta - 1 + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \delta(z_{d,n}, k) \right)}_{\eta_l(\eta, Z, W)} \times \underbrace{\log(\beta_k)}_{t(\beta_k)} \right) \end{aligned}$$

Updating $q(\beta_k|\lambda_k)$ optimize λ_k

- ▶ $q(z_{d,n}) = \text{Mult}(\phi_{d,n})$ or $q(z_{d,n} = k) = \phi_{d,n}^k$ $q(\beta_k) = \text{Dir}(\lambda_k)$ $q(\theta_d) = \text{Dir}(\gamma_d)$
- ▶ In terms of natural parameter:

$$\begin{aligned}\eta(\lambda_k) &= \mathbb{E}_{\prod_{d=1}^D \prod_{n=1}^N q(z_{d,n})} [\eta_l(\eta, Z, W)] \\ &= \mathbb{E}_{\prod_{d=1}^D \prod_{n=1}^N q(z_{d,n})} \left[\eta - 1 + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \delta(z_{d,n}, k) \right] \\ &= \eta - 1 + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \phi_{d,n}^k\end{aligned}$$

- ▶ Change it back to traditional parameter:

$$\lambda_k = \eta + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \phi_{d,n}^k$$

Collapsed Variational Inference

$$q(z_{d,n}) = \text{Mult}(\phi_{d,n}) \text{ or } q(z_{d,n} = k) = \phi_{d,n}^k \quad q(\beta_k) = \text{Dir}(\lambda_k) \quad q(\theta_d) = \text{Dir}(\gamma_d)$$

$$\begin{aligned} \Rightarrow q(Z, \theta_1 \dots \theta_D, \beta_1 \dots \beta_K) &= \left(\prod_{d=1}^{d=D} \prod_{n=1}^N q(z_{d,n} | \phi_{d,n}) \right) \prod_{d=1}^D q(\theta_d | \gamma_d) \prod_{k=1}^K q(\theta_k | \lambda_k) \\ \text{now change to: } &= \underbrace{\left(\prod_{d=1}^{d=D} \prod_{n=1}^N q(z_{d,n} | \phi_{d,n}) \right)}_{q(Z)} q(\Theta, \beta | Z) \end{aligned}$$

Maximize ELOB, it becomes: (remove X for clarity)

Let $U = \{\Theta, \beta\}$:

$$\begin{aligned} \mathcal{L}(q) &\triangleq \mathbb{E}_{q(U,Z)} [\log p(Z, U)] - \mathbb{E}_{q(U,Z)} [\log q(Z, U)] \\ &= \mathbb{E}_{q(U,Z)} [\log p(Z, U)] - \mathbb{E}_{q(U,Z)} [\log q(U|Z) - \log q(Z)] \\ &= \mathbb{E}_{q(Z)} \left(\mathbb{E}_{q(U|Z)} [\log p(Z, U)] \right) - \mathbb{E}_{q(Z)} \left(\mathbb{E}_{q(U|Z)} [\log q(U|Z)] \right) - \mathbb{E}_{q(Z,U)} [\log q(Z)] \\ &= \mathbb{E}_{q(Z)} \left(\underbrace{\mathbb{E}_{q(U|Z)} ([\log p(Z, U)] - [\log q(U|Z)])}_{\mathcal{L}(q(U|Z))} \right) - \mathbb{E}_{q(Z)} [\log q(Z)] \end{aligned}$$

Think this as treating Z as X .

Collapsed Variational Inference (2)

(removed X for clarity)

$$\begin{aligned}\arg \max_{q(U|Z)} (\mathcal{L}(q)) &= \arg \max_{q(U|Z)} \left[\mathbb{E}_{q(Z)} \left(\underbrace{\mathbb{E}_{q(U|Z)} ([\log p_X(Z, U)] - [\log q(U|Z)])}_{\mathcal{L}(q(U|Z))} \right) - \mathbb{E}_{q(Z)} [\log q(Z)] \right] \\&= \mathbb{E}_{q(Z)} \left(\underbrace{\arg \max_{q(U|Z)} [\mathbb{E}_{q(U|Z)} ([\log p(Z, U)] - [\log q(U|Z)])]}_{\mathcal{L}(q(U|Z))} \right) - \mathbb{E}_{q(Z)} [\log q(Z)] \\&= \mathbb{E}_{q(Z)} \underbrace{[p(Z)]}_{\mathcal{L}(q(U|Z))} - \mathbb{E}_{q(Z)} [\log q(Z)]\end{aligned}$$
$$\arg \max_{q(U|Z)} [\mathbb{E}_{q(U|Z)} ([\log p(Z, U)] - [\log q(U|Z)])] = p(Z)$$

maximum occur when $q(U|Z) = p(U|Z) \implies \mathbb{KL}(q(U|Z) \| p(U|Z)) = 0$