Variational Bayes

Richard Xu

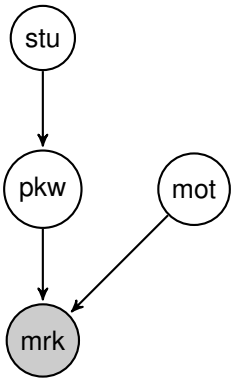
January 9, 2023

1 Topic Summary

1.1 toy example

Let's look at the same "student mark" toy example you saw in both the probabilistic graphical models and MCMC sections:

1. "months of studies" (stu)
2. "prior knowledge" (pkw)
3. "motivation" (mot)
4. "mark obtained" (mrk)



We have three latent variables, (stu), (pkw), (mot) and one observation (mrk), then if we want to perform posterior inference, i.e.,:

$$\Pr\left(\underbrace{\text{stu}, \text{pkw}, \text{mot}}_{\text{latent}} \middle| \underbrace{\text{mrk}}_{\text{observation}}\right) \tag{1}$$

which allows us to compute things such as:

$$\mathbb{E}_{\Pr\left(\text{stu},\text{pkw},\text{mot}|\text{mrk}\right)}[\text{stu}, \text{pkw}, \text{mot}] \tag{2}$$

Unlike Markov Chain Monte Carlo (MCMC), in variational inference, instead of sampling, we propose a set of proposal distributions, $q_{\phi_{\text{stu}}}(\text{stu})$, $q_{\phi_{\text{pkw}}}(\text{pkw})$ and $q_{\phi_{\text{mot}}}(\text{mot})$ which aim to minimize between:

$$q(\text{stu}, \text{pkw}, \text{mot}) = q_{\phi_{\text{stu}}}(\text{stu}) \times q_{\phi_{\text{pkw}}}(\text{ pkw}) \times q_{\phi_{\text{mot}}}(\text{mot}) \tag{3}$$

and

$$\Pr\big(\text{stu}, \text{pkw}, \text{mot}|\text{mrk}\big) \tag{4}$$

Obviously, we need to optimize with respect to the parameters ϕ_{stu}, ϕ_{pkw} and ϕ_{mot}. then after that one can approximate:

$$\mathbb{E}_{\Pr\big(\text{stu},\text{pkw},\text{mot}|\text{mrk}\big)}[\text{stu}, \text{pkw}, \text{mot}] \approx \mathbb{E}_{q\big(\text{stu},\text{pkw},\text{mot}\big)}[\text{stu}, \text{pkw}, \text{mot}] \tag{5}$$

1.2 LDA example

For example, in the LDA example:

$$\begin{aligned} \text{observation:} \quad & \{w_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\} \\ \text{latent variable:} \quad & \{\{\boldsymbol{\beta}_j\}_{j=1}^K, \{\boldsymbol{\theta}_d\}_{d=1}^D, \{z_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\}\} \end{aligned} \tag{6}$$

Therefore, in LDA, the variational inference aims to:

$$\begin{aligned} & p\Big(\{\boldsymbol{\beta}_j\}_{j=1}^K, \{\boldsymbol{\theta}_d\}_{d=1}^D, \{z_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\}\Big|\{w_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\}\Big) \\ & \approx \prod_{j=1}^K q(\boldsymbol{\beta}_j|\boldsymbol{\lambda}_j) \prod_{d=1}^D q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{d=1}^D \prod_{n=1}^N q(z_{d,n}|\boldsymbol{\phi}_{d,n}) \end{aligned} \tag{7}$$

it allows us to approximate:

$$\mathbb{E}_p\Big[\Big\{\{\boldsymbol{\beta}_j\}_{j=1}^K, \{\boldsymbol{\theta}_d\}_{d=1}^D, \{z_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\}\Big\}\Big] \approx \mathbb{E}_q\Big[\Big\{\{\boldsymbol{\beta}_j\}_{j=1}^K, \{\boldsymbol{\theta}_d\}_{d=1}^D, \{z_{d\in\{1\ldots D\},n\in\{1\ldots N\}}\}\Big\}\Big] \tag{8}$$

we just need to optimize with respect to $\{\boldsymbol{\lambda}_j\}, \{\boldsymbol{\gamma}_d\}, \{\boldsymbol{\phi}_{d,n}\}$

1.3 A bit of history ...

This note started in 2010 when I was inspired to help people read Chapter 10 of Bishop [**?**] where I was trying to explain a few things in an oversimplified (hopefully!) way. I revamped it for the class. I also added exponential family distributions and an example on LDA when the model is fully conjugate [**?**]

2 The Variational Bayes Framework

2.1 what is Evidence Lowerbound?

2.2 use Jensen Inequality

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z) \\
&= \log \int_z \frac{p(x,z)}{q_\phi(z|x)} q_\phi(z|x) \\
&= \log \left[\mathbb{E}_{z\sim q_\phi(z|x)} \left(\frac{p(x,z)}{q_\phi(z|x)} \right) \right] \\
&\geq \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{q_\phi(z|x)} \right) \right] \qquad \text{by Jensen's inequality} \\
&= \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log(p(x,z)] - \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log(q_\phi(z|x)] \right. \right. \\
&= \text{ELBO}(q)
\end{aligned} \tag{9}
$$

2.3 simple expansion

$$
\begin{aligned}
\log\left(p(x)\right) &= \log \left(\frac{p(x,z)}{p(z|x)} \right) \\
&= \log\left(p(x,z)\right) - \log\left(p(z|x)\right) \\
&= \left[\log\left(p(x,z)\right) - q_\phi(z)\right] - \left[\log\left(p(z|x)\right) - q_\phi(z)\right] \qquad \because \pm q_\phi(z) \\
&= \log \left(\frac{p(x,z)}{q_\phi(z)} \right) - \log \left(\frac{p(z|x)}{q_\phi(z)} \right)
\end{aligned} \tag{10}
$$

now, let's taking the expectation on both sides, given $q_\phi(z)$:

$$
\begin{aligned}
\log\left(p(x)\right) &= \int q_\phi(z) \log \left(\frac{p(x,z)}{q_\phi(z)} \right) dz - \int q_\phi(z) \log \left(\frac{p(z|x)}{q_\phi(z)} \right) \mathrm{d}z \\
&= \int q_\phi(z) \log \left(\frac{p(x,z)}{q_\phi(z)} \right) dz + \int q_\phi(z) \log \left(\frac{q_\phi(z)}{p(z|x)} \right) \mathrm{d}z \\
&= \text{ELBO}(q) + \mathbb{KL}(q\|p)
\end{aligned} \tag{11}
$$

2.3.1 name to both terms

$$
\begin{aligned}
\text{ELBO}(q) &= \int q_\phi(z) \log \left(\frac{p(x,z)}{q_\phi(z)} \right) \mathrm{d}z \\
\mathbb{KL}(q\|p) &= \int q_\phi(z) \log \left(\frac{p(z|x)}{q_\phi(z)} \right) \mathrm{d}z
\end{aligned}
$$

the question of why we do not minimize $\mathbb{KL}$ term directly? The **key** is that the $\mathbb{KL}$ term contains $p(z|x)$ and ELBO term contains $p(x|z)p(z)$!

since we can choose any $q_\phi(z)$ we'd like, and since we want $\mathbb{KL}(\cdot)$ to be minimized, there it's ideal to make:

$$q_\phi(z) \equiv q_\phi(z|x) \tag{12}$$

i.e., it should also depend on x. Otherwise, it's highly unlikely that the $\mathbb{KL}\big(q\|p(z|x)\big)$ will be minimized:

$$\mathbb{KL}(q\|p) = \int q_\phi(z|x) \log\left(\frac{q_\phi(z|x)}{p(z|x)}\right) \mathrm{d}z \tag{13}$$

We know that $p(x) = \text{ELBO}(q) + \mathbb{KL}(q\|p)$. We consider $\text{ELBO}(q)$ is the lower bound of $p(x)$. Minimizing $\mathbb{KL}(q\|p)$ is the same as maximizing the lower bound $\text{ELBO}(q)$, since the addition of the two becomes $p(x)$

3 The choice of $q(\mathbf{z})$: mean-field approximation

Since any $q(\mathbf{z})$ will work, therefore, we will choose the most simple form. Suppose let's choose $q(\mathbf{z})$, such that:

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(z_i) \tag{14}$$

this is called mean-filed approximation.

$$\begin{aligned}
\text{ELBO}(q) &= \int q_\phi(z) \log\left(\frac{p(x,z)}{q_\phi(z)}\right) \mathrm{d}z \\
&= \int q_\phi(z) \log(p(x,z)) \mathrm{d}z - \int q_\phi(z) \log(q_\phi(z)) \mathrm{d}z \\
&= \underbrace{\int \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}\mathbf{z}}_{\text{part (1)}} - \underbrace{\int \prod_{i=1}^{M} q_i(z_i) \sum_{i=1}^{M} \log\left(q_i(z_i)\right) \mathrm{d}\mathbf{z}}_{\text{part (2)}}
\end{aligned} \tag{15}$$

Since you have the objective function for ELBO(q), a natural approach would be to optimize it repetitively using the parameters associated with each q.

3.1 Simplification of (Part 1):

$$\begin{aligned}
\text{(Part 1)} &= \int \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}\mathbf{z} \\
&= \int_{Z_1} \int_{Z_2} \ldots \int_{Z_M} \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}z_1, \mathrm{d}z_2, \ldots \mathrm{d}z_M
\end{aligned} \tag{16}$$

Rearrange the expression by taking a particular $q_j(z_j)$ out of the integral. Note that unlike (Part2), we are not treating any terms to const.:

$$\begin{aligned}
\text{(Part 1)}_{q_j} &\equiv \text{(Part 1)} \\
&= \int_{z_j} q_j(z_j) \Bigg(\int_{Z_{i\neq j}} \cdots \int \prod_{i\neq j}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \prod_{i\neq j}^{M} \mathrm{d}z_i \Bigg) \mathrm{d}z_j \\
&= \int_{z_j} q_j(z_j) \Bigg(\int_{Z_{i\neq j}} \cdots \int \log\left(p(\mathbf{x},\mathbf{z})\right) \prod_{i\neq j}^{M} q_i(z_i) \mathrm{d}z_i \Bigg) \mathrm{d}z_j
\end{aligned} \tag{17}$$

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

$$\text{(Part 1)}_{q_j} = \int_{z_j} q_j(z_j) \left[\mathbb{E}_{i\neq j}\left[\log\left(p(\mathbf{x},\mathbf{z})\right)\right]\right] \mathrm{d}z_j \tag{18}$$

note that one may consider $\log(\tilde{p}_j(\mathbf{x}, \mathbf{z})) \equiv \mathbb{E}_{i \neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big]$. Obviously, note that

$$\begin{aligned} \tilde{p}_j(\mathbf{x}, \mathbf{z}) &\neq p(z_j|\mathbf{x}) \\ &\neq q(z_j|\mathbf{x}) \end{aligned} \tag{19}$$

and we have:

$$\tilde{p}_j(\mathbf{x}, \mathbf{z}) = \exp\left(\mathbb{E}_{i \neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big]\right) \tag{20}$$

3.2 Simplification of (Part 2):

$$\text{(Part 2)} = \int \prod_{i=1}^{M} q_i(z_i) \sum_{i=1}^{M} \log\left(q_i(z_i)\right) \mathrm{d}\mathbf{z} \tag{21}$$

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

$$\text{(Part 2)} = \sum_{i=1}^{M} \left(\int_{z_i} q_1(z_i) \log\left(q_i(z_i)\right) \mathrm{d}z_i \right) \tag{22}$$

For a particular $p_j(z_j)$, the rest of the sum can be treated like a constant, therefore for $p_j(z_j)$ can be written as:

$$\text{(Part 2)}_{q_j} = \int_{z_j} q_i(z_i) \log\left(q_i(z_i)\right) \mathrm{d}z_j + \text{const.} \tag{23}$$

where const. are the term does not involve z_j.

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

write ELBO(q) in terms of q_j, i.e., ELBO(q_j), in which we try to optimize q_j. The rest of the terms would also need to be optimized $\{q_i\}$:

$$\begin{aligned} \text{ELBO}(q_j) &= \text{Part (1)}_{q_j} - \text{Part (2)}_{q_j} \\ &= \int_{z_j} q_j(z_j) \mathbb{E}_{i \neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big] \mathrm{d}z_j - \int_{z_j} q_j(z_j) \log\left(q_j(z_j)\right) \mathrm{d}z_j + \text{const.} \end{aligned} \tag{24}$$

the key to realize is that we do not need to take derivative as one would normally do. All we need is to re-arrange the terms, and to realize it's the KL term, so we can just math the two distributions.

Note that $\mathbb{E}_{i \neq j}\left[\log\left(p(\mathbf{x}, \mathbf{z})\right)\right]$ would be some log probability of z, we name it $\log(\tilde{p}(\mathbf{x}, \mathbf{z}))$, i.e.,:

$$\log(\tilde{p}(\mathbf{x}, \mathbf{z})) = \mathbb{E}_{i \neq j}\left[\log\left(p(\mathbf{x}, \mathbf{z})\right)\right] \tag{25}$$

Or equivalently as:

$$\begin{aligned}\text{ELBO}(q) &= \int_{z_j} q_j(z_j) \log\left[\frac{\tilde{p}(\mathbf{x},\mathbf{z})}{q_i(z_i)}\right] + \text{const.}. \\ &= -\mathbb{KL}\Big(\mathbb{E}_{i\neq j}\left[\log\left(p(\mathbf{x},\mathbf{z})\right)\right] \| q_i(z_i)\Big)\end{aligned} \tag{26}$$

Now **this is the key**: We can maximize $\text{ELBO}(q)$, by minimizing the KL divergence, where we can find approximate and optimal $q_i^*(z_i)$, such that:

$$\begin{aligned}\log\left(q_i^*(z_i)\right) &= \log(\tilde{p}(\mathbf{x},\mathbf{z})) \\ &= \mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x},\mathbf{z})\right)\big] \\ \implies q_i^*(z_i)) &= \exp\left(\mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x},\mathbf{z})\right)\big]\right)\end{aligned} \tag{27}$$

4 Example: Gaussian-Gamma (Conjugate) posterior

4.1 model

4.1.1 likelihood

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

$$\begin{aligned} p(\mathcal{D}|\mu, \tau) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau}{2}(x_i - \mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(\frac{-\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned} \tag{28}$$

4.1.2 prior

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \propto \exp\left(\frac{-\lambda_0\tau}{2}(\mu - \mu_0)^2\right) \\ p(\tau) &= \text{Gamma}(\tau|a_0, b_0) \propto \tau^{a_0 - 1} \exp^{-b_0\tau} \end{aligned} \tag{29}$$

4.1.3 posterior

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

$$\begin{aligned} p(\mu, \tau|\mathcal{D}) &\propto p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau) \\ &= \mathcal{N}(\mu_n, (\lambda_n\tau)^{-1})\text{Gamma}(\tau|a_n, b_n) \end{aligned} \tag{30}$$

where:

$$\begin{aligned} \mu_n &= \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n \\ a_n &= a_0 + n/2 \\ b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\lambda_0 n(\bar{x} - \mu_0)^2}{2(\lambda_0 + n)} \end{aligned} \tag{31}$$

the exact derivation will be omitted and can be found from external sources easily.

4.2 mean-field Variational Inference algorithm

we let $q(\mathbf{z})$ to be:

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau) \tag{32}$$

We use Variational Bayes formula:

4.2.1 $\log\left(q_\mu^*(\mu)\right) = \mathbb{E}_{q_\tau(\tau)}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right]$

$$
\begin{aligned}
\log\left(q_\mu^*(\mu)\right) &= \mathbb{E}_{q_\tau}\Big[\log\left(p(\mu,\tau,\mathcal{D})\right)\Big] \\
&= \mathbb{E}_{q_\tau}\Big[\log(p(\mathcal{D}|\mu,\tau)) + \log p(\mu|\tau)\Big] + \text{const.} \qquad \text{leave out terms do NOT contain } \mu \\
&= \mathbb{E}_{q_\tau}\Big[\underbrace{\frac{n}{2}\log(\tau) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}|\mu,\tau))} \underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu|\gamma)}\Big] + \text{const.} \\
&= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.}
\end{aligned}
\tag{33}
$$

Completing the square for the μ terms:

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2 &= n\mu^2 - 2n\mu\bar{x} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \text{const.} \\
&= (n+\lambda_0)\mu^2 - 2\mu(n\bar{\mathbf{x}} + \lambda_0\mu_0) \\
&= (n+\lambda_0)\left(\mu^2 - \frac{2\mu(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right) \\
&= (n+\lambda_0)\left(\mu - \frac{(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right)^2 + \text{const.}
\end{aligned}
\tag{34}
$$

Therefore, we have:

$$
\begin{aligned}
\log\left(q_\mu^*(\mu)\right) &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.} \\
&= -\frac{\mathbb{E}_{q_\tau}[\tau](n+\lambda_0)}{2}\left(\mu - \frac{(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right)^2 + \text{const.} \\
\implies q_\mu^*(\mu) &= \mathcal{N}\left(\frac{n\bar{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0}, \mathbb{E}_{q_\tau}[\tau](n+\lambda_0)\right) \qquad \because -\frac{\tau}{2}(x-\mu)^2
\end{aligned}
\tag{35}
$$

4.3 Computing $\log\left(q_i^*(\tau)\right) = \mathbb{E}_{q_\mu(\mu)}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right]$

$$
\begin{aligned}
\log\left(q_\tau^*(\tau)\right) &= \mathbb{E}_{q_\mu}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right] \\
&= \mathbb{E}_{q_\mu}\left[\log(p(\mathcal{D}|\mu,\tau)) + \log p(\mu|\tau) + \log p(\tau)\right] + \text{const.} \\
&= \mathbb{E}_{q_\mu}\Big[\underbrace{\frac{n}{2}\log(\tau) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}|\mu,\tau))} \underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu|\gamma)} \underbrace{+(a_0-1)\log(\tau) - b_0\tau}_{\log p(\tau)}\Big] + \text{const.}
\end{aligned}
\tag{36}
$$

Bring terms without μ outside of the integral:

$$
\begin{aligned}
&= \frac{n}{2}\log(\tau) + (a_0 - 1)\log(\tau) - b_0\tau - \frac{\tau}{2}\mathbb{E}_{q_\mu(\mu)}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.} \\
&= \Big(\underbrace{\frac{n}{2} + a_0}_{a_n} - 1\Big)\log(\tau) - \tau\underbrace{\Big(b_0 + \frac{1}{2}\mathbb{E}_{q_\mu(\mu)}\Big[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\Big]\Big)}_{b_n} + \text{const.}
\end{aligned} \tag{37}
$$

We can rewrite,

$$
\begin{aligned}
b_n &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] \\
&= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[-2\mu n\bar{x} + n\mu^2 + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu\right] + \sum_{i=1}^{n}(x_i)^2 + \lambda_0\mu_0^2 \\
&= b_0 + \frac{1}{2}\left[(n+\lambda_0)\mathbb{E}_{q_\mu}[\mu^2] - 2\left(n\bar{x} + \lambda_0\mu_0\right)\mathbb{E}_{q_\mu}[\mu] + \sum_{i=1}^{n}(x_i)^2 + \lambda_0\mu_0^2\right]
\end{aligned} \tag{38}
$$

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

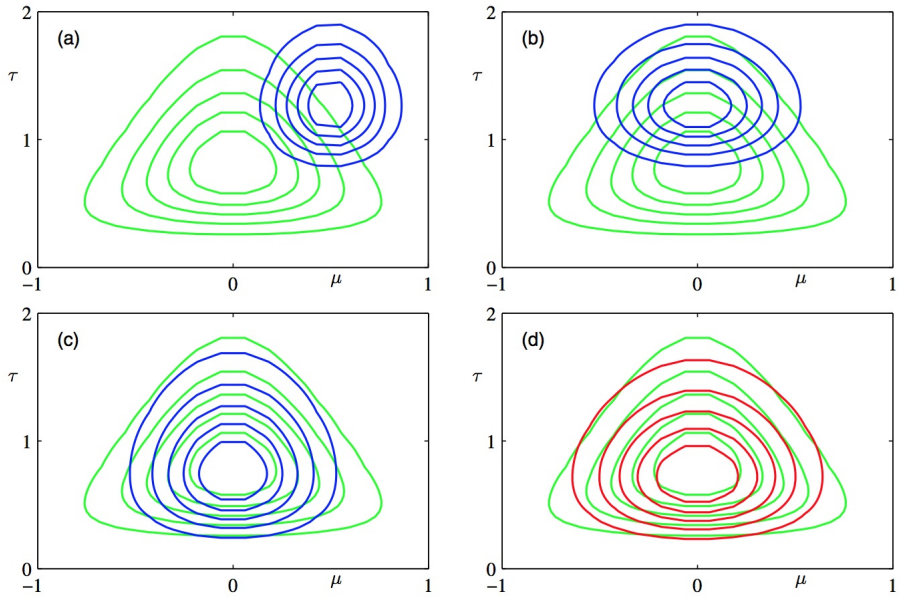


Figure 1: update for Normal Gamma: figure from [**?**]

5 Example of Gaussian Mixture Model Optional for Exam

5.1 The joint density

$$\begin{aligned} p(X, Z, \mu, \Lambda, \pi) &= p(X|Z, \mu, \Lambda, \pi)p(Z|\mu, \Lambda, \pi)p(\mu|\Lambda, \pi)p(\Lambda|\pi)p(\pi) \\ &= p(X|Z, \mu, \Lambda)p(Z|\pi)p(\mu|\Lambda)p(\Lambda)p(\pi) \end{aligned} \tag{39}$$

5.2 Definitions for each probabilities

5.2.1 Definition for $p(Z|\pi)$:

first, is the probability of mixture indices, $Z = \{z_1, ..., z_N\}$, given weights π.

$$\begin{aligned} p(Z|\pi) &= \prod_{i=1}^{N} p(z_n|\pi) \\ &= \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}} \end{aligned} \tag{40}$$

The reason for which $\left(p(z_n|\pi) = \prod_{k=1}^{K} \pi_k^{z_{ik}}\right)$, or $\left(p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}\right)$, is because in Bishop, z is not represented in a scalar form, but rather in a vector of dimension K, which has a single element 1, and the rest are all 0s. For example, instead of using $p(z_n = 2|\pi = [0.2, 0.3, 0.5]) = 0.3$, Bishop uses $p(z_n = [0, 1, 0]|\pi = [0.2, 0.3, 0.5]) = 0.3$. In any case, this refers to the second element of π. Therefore, a more simpler and vocal representation for $p(z|\pi)$ is just the z^{th} value of π.

5.2.2 Definition for $p(X|Z, \mu, \Lambda)$:

$$p(X|Z, \mu, \Lambda) = \prod_{i=1}^{N} p(x_n|z_n, \mu, \Lambda)$$

In normal literatures, such as Bilmes, it is defined as:

$$= \prod_{i=1}^{N} \mathcal{N}(x_n|\mu_{z_n}, \Lambda_{z_n}^{-1}) \tag{41}$$

However, due to the vector representation of Bishop, the above is defined as:

$$= \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}}$$

However, the above two represent the same thing:

5.2.3 Definition for $p(\pi)$:

This is just a straight Dirichlet probability:

$$
\begin{aligned}
p(\pi|\alpha_0) = \mathrm{Dir}(\pi|\alpha_0) &\propto C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_{0k}-1} \\
\implies \log(\pi|\alpha_0) &\propto (\alpha_0 - 1) \sum_{k=1}^{K} \log \pi_k
\end{aligned}
\tag{42}
$$

5.2.4 Definition for $p(\mu|\Lambda)p(\Lambda)$:

This is almost always a Gaussian-Wishart distribution:

$$
\begin{aligned}
p(\mu, \Lambda) &= p(\mu|\Lambda)p(\Lambda) \\
&= \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|W_0, v_0)
\end{aligned}
\tag{43}
$$

5.3 Begin VB of GMM

5.3.1 The expression for $q^*(Z)$:

$$
\begin{aligned}
\log q^*(Z) &= \mathbb{E}_{\pi,\mu,\Lambda}\left[\log p(X, Z, \pi, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\log p(Z|\pi)\right] + \mathbb{E}_{\mu,\Lambda}\left[\log p(X|Z, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}}\right] + \mathbb{E}_{\mu,\Lambda}\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}}\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log \pi_k^{z_{ik}}\right] + \mathbb{E}_{\mu,\Lambda}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}}\right] + \text{const.} \\
&\text{given that:, } (\log a^b = b \log a) : \\
&= \mathbb{E}_{\pi}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log \pi_k\right] + \mathbb{E}_{\mu,\Lambda}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})\right] + \text{const.} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\mathbb{E}_{\pi}[\log \pi_k] + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\mathbb{E}_{\mu,\Lambda}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})] + \text{const.} \\
&\text{Taking the common term to the left, } \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} : \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \left(\mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu,\Lambda}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})]\right) + \text{const.} \\
&\text{Bishop nominates a new term: } \log \rho_{ik} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \left(\log \rho_{ik}\right) + \text{const.}
\end{aligned}
\tag{44}
$$

Let's look at the expression for $\log \rho_{ik}$:

$$
\begin{aligned}
\log \rho_{ik} &= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\log\left(\frac{1}{(2\pi)^{(d/2)}}|\Lambda_k|^{1/2}\exp\left(-\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\log(2\pi)^{\frac{-d}{2}} + \frac{1}{2}\log|\Lambda_k| + \left(-\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\frac{-d}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_k| - \left(\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \frac{-d}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\Lambda_k}\left[\log|\Lambda_k|\right] - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\left[(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right]
\end{aligned}
\tag{45}
$$

Now, since $\log q^*(Z) = \log \rho_{ik}$

$$
\begin{aligned}
\log q^*(Z) &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\log \rho_{ik}\right) + \text{const.} \implies \\
q^*(Z) &= \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\log \rho_{ik}\right) + \text{const.}\right) \\
&= C\prod_{i=1}^{N}\prod_{k=1}^{K}\exp(z_{ik}\left(\log \rho_{ik}\right)) = C\prod_{i=1}^{N}\prod_{k=1}^{K}\exp\left(\log \rho_{ik}^{z_{ik}}\right) = C\prod_{i=1}^{N}\prod_{k=1}^{K}\rho_{ik}^{z_{ik}}
\end{aligned}
\tag{46}
$$

Since $q^*(Z) = \prod_{i=1}^{N} q^*(z_n)$:

$$
q^*(Z) = \prod_{i=1}^{N} C\prod_{k=1}^{K}\rho_{ik}^{z_{ik}} \tag{47}
$$

In a way, $\rho_{ik}^{z_{ik}}$ plays the same role as π in $p(z_n|\pi)$, therefore, $\sum_{k=1}^{K}\pi_k = 1 \implies \sum_{k=1}^{K}\rho_{ik} = 1$:

$$
\begin{aligned}
q^*(Z) &= \prod_{i=1}^{N} q^*(z_i) = \prod_{i=1}^{N}\left(\frac{1}{\sum_{j=1}^{K}\rho_{nj}}\prod_{k=1}^{K}\rho_{ik}^{z_{ik}}\right) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{K}\frac{\rho_{ik}^{z_{ik}}}{\sum_{j=1}^{K}\rho_{nj}} = \prod_{i=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{ik}}
\end{aligned}
\tag{48}
$$

This is a multinomial distribution, therefore, $\mathbb{E}[z_i = k] = r_{ik}$

5.3.2 The expression for $q^*(\pi, \mu, \Lambda)$:

$$
\begin{aligned}
\log q^*(\pi,\mu,\Lambda) &= \mathbb{E}_Z\left[\log p(X, Z, \pi, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_Z\left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \mathbb{E}_Z\left[\log p(\pi)\right] + \mathbb{E}_Z\left[\log p(\mu|\Lambda)\right] + \mathbb{E}_Z\left[\log p(\Lambda)\right] + \text{const.} \\
&= \mathbb{E}_Z\left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \log p(\mu|\Lambda) + \log p(\Lambda) + \text{const.}
\end{aligned}
\tag{49}
$$

Combine the mean and precision together:

$$= \mathbb{E}_Z \left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi) + \log p(\mu,\Lambda) + \text{const.}$$

And since each (μ_k, Λ_k) are independent, therefore:

$$\begin{aligned}
&= \mathbb{E}_Z \left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{ik}}\right] + \mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k,\Lambda_k) + \text{const.} \\
&= \mathbb{E}_Z \left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log(z_{ik})\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})\right] + \mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k,\Lambda_k) + \text{const.} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k,\Lambda_k) + \text{const.} \\
&= \underbrace{\mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi)}_{\log q^*(\pi)} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \sum_{k=1}^{K} \log p(\mu_k,\Lambda_k)}_{\log q^*(\mu,\Lambda)} + \text{const.}
\end{aligned} \tag{50}$$

For the part of $\log q^*(\pi)$:

$$\begin{aligned}
\log q^*(\pi) &= \mathbb{E}_Z \left[\log p(Z|\pi)\right] + \log p(\pi) \\
&= \mathbb{E}_Z \left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}}\right] + \log p(\pi) \\
&= \mathbb{E}_Z \left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log \pi_k\right] + \log p(\pi) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \log \pi_k \mathbb{E}_Z[z_{ik}] + (\alpha_0 - 1)\sum_{k=1}^{K} \log \pi_k + \text{const.} \\
&= \sum_{k=1}^{K} \log \pi_k \sum_{i=1}^{N} r_{i,k} + (\alpha_0 - 1)\sum_{k=1}^{K} \log \pi_k + \text{const.} \\
&= \left(\underbrace{\sum_{i=1}^{N} r_{i,k} + \alpha_0}_{a_n} - 1\right)\sum_{k=1}^{K} \log \pi_k + \text{const.} = \text{DIR}\left(\pi|a_n\right)
\end{aligned} \tag{51}$$

For the part of $\log q^*(\mu,\Lambda)$:

$$\log q^*(\mu,\Lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \sum_{k=1}^{K} \log p(\mu_k,\Lambda_k) \tag{52}$$

We only have the expression for $\mathbb{E}_{q^*(Z)}[Z]$, but not $\mathbb{E}_{q^*(Z)}[\log(Z)]$:

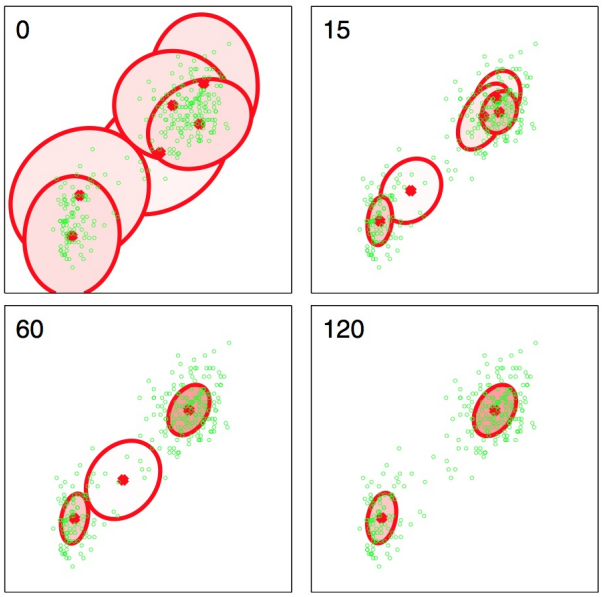


Figure 2: update for Gaussian Mixture Model: figure from [**?**]

6 Exponential Family distributions

6.1 Big picture

Given both the prior and likelihood are exponential family distributions and they form a conjugacy pair, then the variational inference (also mean-field approximation), i.e., $q(\mathbf{z}) = \prod_i q_i(z_i)$ can have the following update formula:

$$\eta_j = \mathbb{E}_{q(\mathbf{z}\setminus z_j|\cdot)}[\eta_{\text{post}}(\mathbf{z} \setminus z_j)] \tag{53}$$

where $\eta_{\text{post}}(\mathbf{z} \setminus z_j)$ is the natural parameter associated with posterior distribution $p(z_j|-)$. Of course it is expressed in terms of all other $\mathbf{z} \setminus z_j$,but z_j as part of its parameter.

Obviously, the corresponding $q(\cdot|-)$ must first exclude z_j.

compare this with the generic update formula:

$$\log\left(q_i^*(z_i)\right) = \mathbb{E}_{i\neq j}\left[\log\left(p(\mathbf{x}, \mathbf{z})\right)\right] \tag{54}$$

using exponential family update formula Eq.(53), the update is directly applied to the parameter.

6.2 Exponential Family

Most of the distributions we are going to look at are from **exponential family**. They are expressed in terms of its natural parameter η:

$$h(x)\exp\left(T(x)^\top \eta - A(\eta)\right) \tag{55}$$

$$\begin{aligned} & \underbrace{\exp(-A(\eta))}_{\text{normalization}} h(x)\exp\{T(x)^\top\eta\} \\ \implies & \exp(-A(\eta))\int_x h(x)\exp\{T(x)^\top\eta\} = 1 \\ \implies & \int_x h(x)\exp\{T(x)^\top\eta\} = \exp(A(\eta)) \end{aligned} \tag{56}$$

6.3 example: 1-d Gaussian

$$\begin{aligned} \mathcal{N}(x;\mu,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \exp\left(-\frac{x^2-2x\mu+\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\left(\begin{bmatrix} x & x^2 \end{bmatrix}\begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \end{aligned} \tag{57}$$

$$
\begin{aligned}
T(\mathbf{x}) &= \begin{bmatrix} x & x^2 \end{bmatrix} \\
\boldsymbol{\eta} &= \begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}
\end{aligned} \tag{58}
$$

1. for η_2:

$$
\eta_2 = -\frac{1}{2\sigma^2} \implies \sigma^2 = -\frac{1}{2\eta_2} \tag{59}
$$

2. for η_1:

$$
\begin{aligned}
\eta_1 = \frac{\mu}{\sigma^2} \implies \mu &= \eta_1 \sigma^2 \\
&= \eta_1 \frac{-1}{2\eta_2} \\
&= \frac{-\eta_1}{2\eta_2}
\end{aligned} \tag{60}
$$

summarize, we have:

$$
\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix} \tag{61}
$$

6.3.1 in natural parameter form

now we can remove μ and σ^2:

$$
\begin{aligned}
\mathcal{N}_{\text{nat}}(x, \boldsymbol{\eta}) &= \exp\left(\begin{bmatrix} x & x^2 \end{bmatrix} \begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix}^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right) \\
&= \exp\left(\begin{bmatrix} x & x^2 \end{bmatrix} \begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix}^\top - \frac{\left(\frac{-\eta_1}{2\eta_2}\right)^2}{2\left(\frac{-1}{2\eta_2}\right)} - \frac{1}{2}\log\left(2\pi\left(\frac{-1}{2\eta_2}\right)\right) \right) \\
&= \exp\left(T(x)^\top \boldsymbol{\eta} + \frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(\frac{2\pi}{-2\eta_2}\right) \right) \\
&= \exp\left(T(x)^\top \boldsymbol{\eta} + \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-2\eta_2) - \frac{1}{2}\log(2\pi) \right)
\end{aligned} \tag{62}
$$

now that the probability is fully in terms of the natural parameter

$$
\mathcal{N}_{\text{nat}}(x, \boldsymbol{\eta}) = \exp\Big(T(x)^\top \boldsymbol{\eta} - \underbrace{\Big(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \Big) - \frac{1}{2}\log(2\pi)}_{A(\boldsymbol{\eta})} \Big) \tag{63}
$$

6.4 conjugate probabilities

conjugacy means that the prior and posterior are of the same **type** of distributions, for example:

$$\underbrace{p_{\eta_{\text{post}}}(\theta|\mathbf{x})}_{\text{same type}} \propto p(\mathbf{x}|\theta)\underbrace{p_{\eta_{\text{prior}}}(\theta)}_{\text{same type}} \tag{64}$$

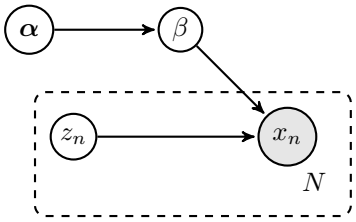
Note that prior and posterior are of the same type, but they are usually not the same distribution, otherwise, the likelihood $p(\mathbf{x}|\theta)$ is not useful at all!

Using exponential family distribution representation, when conjugacy is achieved, it means the prior and posterior have the same sufficient statistics $T(\theta)$ and $h(\theta)$, but different natural parameter, i.e., η_{post} and η_{prior}, and different log-normalizer for both.

It turns out that the posterior inherits $h(\theta)$ from the prior, so we just need to make sure to put the appropriate criteria on the likelihood so that $T(\theta)$ is the same in both the prior and posterior

6.5 What is the criteria for likelihood to pair up with conjugate prior? Optional for Exam

It's always better to have a discussion with a concrete example setup. So we have the following problem setup, described in [**?**], without actually defining what distributions they are using:



the joint density is of the form:

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}|\boldsymbol{\alpha}) = p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(x_n, z_n|\boldsymbol{\beta}) \tag{65}$$

the conditionals are based on Exponential family:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \alpha) &= h(\boldsymbol{\beta}) \exp\left\{T(\boldsymbol{\beta})^\top \eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha))\right\} \\ p(z_{n,j}|x_n, z_{n,-j}, \boldsymbol{\beta}) &= h(z_{n,j}) \exp\left\{T(z_{n,j})\eta_{z_{n,j}}(x_n, z_{n,-j}, \boldsymbol{\beta}) - A_l\big(\eta_{z_{n,j}}(x_n, z_{n,-j}, \boldsymbol{\beta})\big)\right\} \end{aligned} \tag{66}$$

Think about why is this representation useful? Let's have look at a numerical example:

6.5.1 Conjugacy of exponential family distribution

Let's work through a concrete example of posterior $p(\boldsymbol{\beta}|x_n, z_n)$, instead of writing $\boldsymbol{\eta}_\beta$, we write $\boldsymbol{\beta}$ directly:

- **prior**:

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = h(\boldsymbol{\beta}) \exp\{T(\beta)^\top \boldsymbol{\alpha} - A_{\text{pri}}(\boldsymbol{\alpha})\} \tag{67}$$

suppose the sufficient statistics of the **prior** can be written as:

$$\begin{aligned} T(\boldsymbol{\beta}) &= \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix} \\ \implies \boldsymbol{\alpha} &= \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \end{aligned} \tag{68}$$

then the prior itself can be written as:

$$p(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} - A_{\text{pri}}(\alpha) \right\} \tag{69}$$

- **likelihood**:
  and if the likelihood density (x_n, z_n) can be defined as:

$$p(x_n, z_n|\boldsymbol{\beta}) = h(x_n, z_n) \exp\left\{ T(x_n, z_n)^\top \boldsymbol{\beta} - A_l(\boldsymbol{\beta}) \right\} \tag{70}$$

- then **posterior** condition on a single data point:

$$\begin{aligned} p(\boldsymbol{\beta}|x_n, z_n, \boldsymbol{\alpha}) &\propto \underbrace{h(\boldsymbol{\beta}) \exp\{T(\beta)^\top \boldsymbol{\alpha}\}}\underbrace{\exp\{T(x_n, z_n)^\top \boldsymbol{\beta} - A_l(\boldsymbol{\beta})\}} \\ &= h(\boldsymbol{\beta}) \exp\left\{\boldsymbol{\beta}^\top \alpha_1 - \alpha_2 A_l(\boldsymbol{\beta}) + \boldsymbol{\beta}^\top T(x_n, z_n) - A_l(\boldsymbol{\beta})\right\} \\ &= h(\boldsymbol{\beta}) \exp\left\{\boldsymbol{\beta}^\top (\alpha_1 + T(x_n, z_n)) - \alpha_2 A_l(\boldsymbol{\beta}) - A_l(\boldsymbol{\beta})\right\} \\ &= h(\boldsymbol{\beta}) \exp\left\{\boldsymbol{\beta}^\top (\alpha_1 + T(x_n, z_n)) - (\alpha_2 + 1) A_l(\boldsymbol{\beta})\right\} \\ &= h(\boldsymbol{\beta}) \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \alpha_1 + T(x_n, z_n) \\ \alpha_2 + 1 \end{bmatrix} \right\} \\ &= h(\boldsymbol{\beta}) \exp\left\{ T(\boldsymbol{\beta})^\top \begin{bmatrix} \alpha_1 + T(x_n, z_n) \\ \alpha_2 + 1 \end{bmatrix} \right\} \end{aligned} \tag{71}$$

  notice the posterior "inherited" $h(\boldsymbol{\beta})$ from the prior, so we only need to making sure that the $T(\boldsymbol{\beta})$ are the same for both prior and posterior.

6.5.2 posterior on all data

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}) &\propto h(\boldsymbol{\beta}) \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 \end{bmatrix} \right\} \\ &= h(\boldsymbol{\beta}) \exp\left\{ T(\beta)^\top \begin{bmatrix} \alpha_1 + \sum_{n=1}^N T(x_n, z_n) \\ \alpha_2 + N \end{bmatrix} \right\} \end{aligned} \tag{72}$$

1. **Complete likelihood**

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\beta) &= \prod_{n=1}^{N} h(x_n, z_n) \exp\{\boldsymbol{\beta}^\top T(x_n, z_n) - A_l(\boldsymbol{\beta})\} \\ &= h(\mathbf{x}, \mathbf{z}) \exp\left\{\sum_{n=1}^{N} \boldsymbol{\beta}^\top T(x_n, z_n) - N \times A_l(\boldsymbol{\beta})\right\} \end{aligned} \tag{73}$$

2. **Complete posterior**

now, look at:

$$p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}) \propto h(\boldsymbol{\beta}) \exp\left\{T(\beta)^\top \begin{bmatrix} \alpha_1 + \sum_{n=1}^{N} T(x_n, z_n) \\ \alpha_2 + N \end{bmatrix}\right\} \tag{74}$$

When we use the expression and use use η_{post} instead:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}) &= h(\boldsymbol{\beta}) \exp\left\{\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha})^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha))\right\} \\ \implies \eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}) &= \begin{bmatrix} \alpha_1 + \sum_{n=1}^{N} t(x_n, z_n) \\ \alpha_2 + N \end{bmatrix} \\ \implies \exp\left(A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}))\right) &= \int_{\boldsymbol{\beta}} h(\boldsymbol{\beta}) \exp\left\{\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha})^\top T(\beta)\right\} \mathrm{d}\boldsymbol{\beta} \end{aligned} \tag{75}$$

6.5.3 Example: Posterior of Gaussian mean

suppose data x_i come from unit variance Gaussian. Compare with Section (6.3), we saved one parameter:

$$\begin{aligned} p(x|\mu) &= \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2}(x-\mu)^2 \Big\} \\ &= \underbrace{\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}}_{h(x)} \exp\Big\{ \underbrace{\mu}_{\beta} \underbrace{x}_{T(x)} - \underbrace{\frac{\mu^2}{2}}_{A_l(\beta)} \Big\} \end{aligned} \tag{76}$$

Therefore:

$$\begin{aligned} \beta &= \mu \\ T(x) &= x \\ A_l(\beta) &= \frac{\beta^2}{2} \\ h(x) &= \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \end{aligned} \tag{77}$$

substitute into:

$$p(x|\beta) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \exp\left\{\beta x + \underbrace{-\frac{\beta^2}{2}}_{A_l(\beta)}\right\} \tag{78}$$

6.5.4 criteria for conjugate pair

A **conjugate prior** MUST be:

$$\begin{aligned} p(\beta|\alpha) &= h(\beta) \exp\left\{\alpha_1 \beta + \alpha_2 \underbrace{(-\beta^2/2)}_{A_l(\beta)} - A_g(\alpha)\right\} \\ &= h(\beta) \exp\left\{ \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}^\top \begin{bmatrix} \beta \\ -\frac{\beta_2}{2} \end{bmatrix} - A_g(\alpha)\right\} \end{aligned} \tag{79}$$

Wait, this doesn't look exactly in the form of Eq.(57), i.e.,:

$$\mathcal{N}(x; \mu, \sigma^2) = \exp\left(\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right) \tag{80}$$

We can arrange Eq.(**??**) to look like, but with parameter $\begin{bmatrix}\alpha_1 & -\frac{\alpha_2}{2}\end{bmatrix}^\top$:

$$p(\boldsymbol{\beta}|\alpha) = h(\boldsymbol{\beta}) \exp\left\{ \begin{bmatrix} \alpha_1 \\ -\frac{\alpha_2}{2} \end{bmatrix}^\top \begin{bmatrix} \beta \\ \beta^2 \end{bmatrix} - A_g(\alpha)\right\} \tag{81}$$

From our knowledge, a distribution with sufficient statistics $T(\beta) = \begin{bmatrix}\beta & \beta^2\end{bmatrix}$ is a Gaussian distribution.

Suppose the likelihood is an exponential family distribution. Every exponential family has a conjugate prior in theory. The natural parameter $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2\end{bmatrix}^\top$ has dimension $\dim(\beta) + 1$. The sufficient statistics of the prior are $\begin{bmatrix}\beta & -A_l(\beta)\end{bmatrix}^\top$

6.6 Conjugate exponential family distribution: $\mathbb{E}_q[T(\beta)] = \nabla_\lambda A_g(\lambda)$ Optional for Exam

given that we have:

$$\begin{aligned} q(\beta|\lambda) &= h(\beta) \exp\{\lambda^\top T(\beta) - A_g(\lambda)\} \\ &= \frac{1}{\exp(A_g(\lambda))} h(\beta) \exp\{\lambda^\top T(\beta)\} \end{aligned} \tag{82}$$

why is it that we have:

$$\mathbb{E}_{q(\beta)}[T(\beta)] = \nabla_\lambda A_g(\lambda) \tag{83}$$

$$
\begin{aligned}
&\int_\beta q(\beta|\lambda)\mathrm{d}\beta = \int_\beta h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\mathrm{d}\beta = 1 \\
&\implies \nabla_\lambda \left(\int_\beta h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\mathrm{d}\beta\right) = 0 \\
&\implies \int_\beta \nabla_\lambda \left(h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\right)\mathrm{d}\beta = 0 \\
&\implies \int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}\left(T(\beta) - \nabla_\lambda A_g(\lambda)\right) = 0 \\
&\implies \underbrace{\int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}T(\beta)}_{\mathbb{E}_{q(\beta)}[T(\beta)]} - \underbrace{\int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}}_{=1}\nabla_\lambda A_g(\lambda) = 0 \\
&\implies \mathbb{E}_{q(\beta)}[T(\beta)] - \nabla_\lambda A_g(\lambda) = 0
\end{aligned}
\tag{84}
$$

6.6.1 The choice of $q(\beta, \mathbf{z})$

We choose $q(\beta, \mathbf{z})$ to decouple β and $\mathbf{z}$ completely:

$$
q(\beta, \mathbf{z}) = q(\beta|\lambda)\prod_{n=1}^{N}\prod_{j=1}^{J} q(z_{n,j}|\phi_{n,j}) \tag{85}
$$

- $q(\beta|\lambda)$ is the SAME distribution type as $p(\beta|\mathbf{x}, \mathbf{z}, \alpha)$, they only differ in parameter. This means they have the same sufficient statistics $T(\beta)$:

$$
\begin{aligned}
q(\beta|\lambda) &= h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\} \\
\text{compare with:}\quad p(\beta|\mathbf{x}, \mathbf{z}, \alpha) &= h(\beta)\exp\left\{\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha)^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha))\right\}
\end{aligned}
\tag{86}
$$

- $q(z_{n,j}|\phi_{n,j})$ is the SAME distribution type as $p(z_{n,j}|x_n, z_{n,-j}, \beta)$, they only differ in parameter. This means they have the same sufficient statistics $T(z_{n,j})$:

$$
\begin{aligned}
q(z_{n,j}|\phi_{n,j}) &= h(z_{n,j})\exp\left\{\phi_{n,j}^\top T(z_{n,j}) - A_l(\phi_{n,j})\right\} \\
\text{compare with:}\quad p(z_{n,j}|x_n, z_{n,-j}, \beta) &= h(z_{n,j})\exp\left\{\eta_l(x_n, z_{n,-j}, \beta)^\top T(z_{n,j}) - A_l(\eta_l(x_n, z_{n,-j}, \beta))\right\}
\end{aligned}
\tag{87}
$$

6.7 Proof for for ELBO(λ) for $q(\beta|\lambda)$ Optional for Exam

this section shows the proof for the update formula used in Eq.(53), i.e., $\eta_j = \mathbb{E}_{q(\mathbf{z}\setminus z_j|\cdot)}[\eta_{\text{post}}(\mathbf{z}\setminus z_j)]$, we will do so using an example from the setting described in this section.

Our goal is to maximize the ELBO, i.e.,

$$
\text{ELBO}(q) \triangleq \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}, \beta|\alpha)] - \mathbb{E}_{q(\beta,\mathbf{z})}[\log q(\mathbf{z}, \beta)] \tag{88}
$$

Note that q used here is $q(\beta, \mathbf{z})$ not just $q(\beta|\lambda)$

$$\begin{aligned}
\text{ELBO}(\lambda) &= \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\beta|\mathbf{x},\mathbf{z},\alpha)] + \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\mathbf{x},\mathbf{z})] - \mathbb{E}_{q(\beta,\mathbf{z})}[\log q(\beta)] \\
&= \mathbb{E}_q\big[\log p(\beta|\mathbf{x},\mathbf{z},\alpha)\big] - \mathbb{E}_q\big[\log q(\beta)\big] + \text{const.} \\
&= \mathbb{E}_q\left[\log\left(h(\beta)\exp\left\{\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha))\right\}\right)\right] - \mathbb{E}_q[\log q(\beta)] + \text{const.} \\
&= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_q[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)^\top T(\beta)]} - \mathbb{E}_q[\log h(\beta)\exp\{\lambda^\top T(\beta) - A_{\text{pri}}(\lambda)\}] + \text{const.} \\
&= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)]^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)]} - \mathbb{E}_q[\log h(\beta)] - \mathbb{E}_q[\lambda^\top T(\beta)] + A_{\text{pri}}(\lambda) + \text{const.} \\
&= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(x,z,\alpha)]^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)] - \lambda^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)] + A_{\text{pri}}(\lambda) + \text{const.} \qquad \because A_{\text{pri}}(\lambda) \text{ contains } \lambda
\end{aligned} \tag{89}$$

Substitute $\mathbb{E}_{q(\beta|\lambda)}[T(\beta)] = \nabla_\lambda A_{\text{pri}}(\lambda)$:

$$\text{ELBO}(\lambda) = \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(x,z,\alpha)]^\top \nabla_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla_\lambda A_{\text{pri}}(\lambda) + A_{\text{pri}}(\lambda) + \text{const.} \tag{90}$$

Maximize $\text{ELBO}(\lambda)$ we get:

$$\begin{aligned}
\nabla_\lambda \text{ELBO}(\lambda) &= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_g(\mathbf{x},\mathbf{z},\alpha)]^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) - \nabla_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) + \nabla_\lambda A_{\text{pri}}(\lambda) = 0 \\
&= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_g(\mathbf{x},\mathbf{z},\alpha)]^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) = 0 \\
&\implies \nabla^2_\lambda A_{\text{pri}}(\lambda)\big(\mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)]^\top - \lambda^\top\big) = 0
\end{aligned} \tag{91}$$

$$\lambda = \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)] \tag{92}$$

in words, when we try to update λ for $q(\beta|\lambda)$, it find the corresponding posterior $p(\beta|\mathbf{x},\mathbf{z},\alpha)$, and its natural parameter $\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)$, then computes the expectation with all the $q(\cdot)$ that its natural parameter has random variable for.

6.7.1 Update for ELBO$(\phi_{n,j})$ for $q(z_{n,j}|\phi_{n,j})$

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

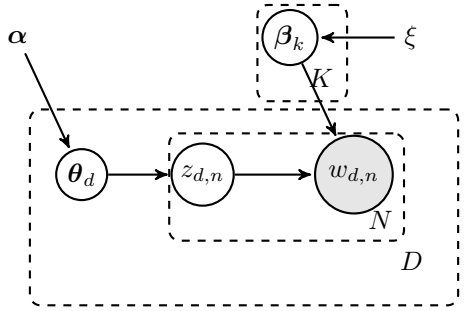
$$\nabla_{\phi_{n,j}} \text{ELBO}(\phi_{n,j}) = \nabla^2_{\phi_{n,j}} A_l(\phi_{n,j})\big(\mathbb{E}_{q(\lambda)}[\eta_l(x_n, z_{n,-j}, \beta)]^\top - \phi_{n,j}^\top\big) = 0 \tag{93}$$

$$\phi_{n,j} = \mathbb{E}_{q(\lambda)}\big[\eta_l(x_n, z_{n,-j}, \beta)\big] \tag{94}$$

in words, when we try to update $\phi_{n,j}$ for $q(z_{n,j}|\phi_{n,j})$, it find the corresponding posterior $p(z_{n,j}|x_n, z_{n,-j})$, and its natural parameter $\eta_l(x_n, z_{n,-j})$, then computes the expectation with all the $q(\cdot)$ that its natural parameter has random variable for.

7 Latent Dirichlet Allocation

let's visit Latent Dirichlet Allocation again [**?**]:



- $\boldsymbol{\beta}_k \sim \text{Dir}(\xi, \ldots, \xi)$ for $k \in \{1, ..., K\}$.
- For each document d:
  $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha, \ldots, \alpha)$
  For each word $w \in \{1, ..., N\}$:
  $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$
  $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$

7.1 define corresponding $q(\cdot)$

1. $q(z_{d,n})$

$$\begin{aligned} q(z_{d,n}) &= \text{Mult}(\boldsymbol{\phi}_{d,n}) \\ \implies q(z_{d,n} = k) &= \phi_{d,n}^k \end{aligned} \tag{95}$$

2. $q(\boldsymbol{\beta}_k)$

$$q(\boldsymbol{\beta}_k) = \text{Dir}(\boldsymbol{\lambda}_k) \tag{96}$$

3. $q(\boldsymbol{\theta}_d)$

$$q(\boldsymbol{\theta}_d) = \text{Dir}(\boldsymbol{\gamma}_d) \tag{97}$$

7.1.1 Facts about Dirichlet Distribution

$$\begin{aligned} &\theta \sim \text{Dir}(\gamma_1, \ldots \gamma_K) \\ \implies &\mathbb{E}[\log(\theta_k)|\gamma] = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right) \quad \text{for component } k \end{aligned} \tag{98}$$

where:

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln\big(\Gamma(x)\big) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{99}$$

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

7.2.1 find natural parameter of posterior $p(z_{dn} = k|\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n})$

$$\begin{aligned} p(z_{d,n} = k|\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n}) &\propto p(z_{d,n} = k|\boldsymbol{\theta}_d) p(w_{d,n}|z_{d,n} = k, \boldsymbol{\beta}_{1:K}) \\ &= \theta_{d,k} \times \beta_{k,w_{d,n}} \\ &\propto \exp\left(\underbrace{\log(\theta_{d,k}) + \log(\beta_{k,w_{d,n}})}_{\eta_l(\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n})} \times \underbrace{1}_{T(z_{d,n})} \right) \end{aligned} \tag{100}$$

for the last line, we substitute the exponential family form of multinomial distribution. Expressing it using "normal " multinomial distribution, then its parameter is:

$$p(z_{d,n}|\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n}) = \text{Mult}\big(\theta_{d,1} \times \beta_{1,w_{d,n}}, \ldots, \theta_{d,K} \times \beta_{K,w_{d,n}}\big) \tag{101}$$

diagrammatically, $\boldsymbol{\beta}_{1:K}$ is represented as a matrix of:

$$\boldsymbol{\beta}_{1:K} \equiv \begin{bmatrix} -- & \boldsymbol{\beta}_1 & -- \\ \vdots & \vdots & \vdots \\ -- & \boldsymbol{\beta}_k & -- \end{bmatrix} \equiv \begin{bmatrix} \beta_{1,w_1} & \ldots & \beta_{1,w_{|\mathcal{V}|}} \\ \vdots & \vdots & \vdots \\ \beta_{K,w_1} & \ldots & \beta_{K,w_{|\mathcal{V}|}} \end{bmatrix} \tag{102}$$

7.2.2 optimize $\phi_{d,n}$

apply the update formula, in which we need the natural parameter for $p(z_{d,n}|\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n})$ in the exception:

$$\begin{aligned} \eta(\phi^k_{d,n}) = \log(\phi^k_{d,n}) &\propto \mathbb{E}_{q(\boldsymbol{\theta}_d)q(\boldsymbol{\beta}_k)} \left[\eta_l\left(\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n}\right)\right] \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K})} \left[\log(\theta_{d,k})\right] + \mathbb{E}_{q(\boldsymbol{\beta}_k)} \left[\log(\beta_{k,w_{d,n}})\right] \\ &= \Psi(\gamma_{d,k}) - \Psi\Big(\sum_{k=1}^{K} \gamma_{d,k}\Big) + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big) \end{aligned} \tag{103}$$

compare this with Eq.(53), i.e., $\eta_j = \mathbb{E}_{q(\mathbf{z} \setminus z_j)}[\eta_{\text{post}}(\mathbf{z} \setminus z_j)]$, you can see easily that:

$$\mathbf{z} \setminus z_j \equiv \{\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}\} \tag{104}$$

to obtain $\boldsymbol{\phi}_{d,n}$:

$$\begin{aligned} \implies \phi^k_{d,n} &\propto \exp\left[\Psi(\gamma_{d,k}) - \Psi\Big(\sum_{k=1}^{K} \gamma_{d,k}\Big) + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big)\right] \\ &\propto \exp\left[\Psi(\gamma_{d,k}) + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big)\right] \quad \because \sum_{k=1}^{K} \gamma_{d,k} \text{ has same value irrespective of } k \end{aligned} \tag{105}$$

7.3 Updating $q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)$: optimize $\boldsymbol{\gamma}_d$

7.3.1 find natural parameter of posterior $p(\boldsymbol{\theta}_d|\mathbf{z}_d)$

$$
\begin{aligned}
p(\boldsymbol{\theta}_d|\mathbf{z}_d) &= p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_{d,n}|\boldsymbol{\theta}_d) = \text{Dir}(\boldsymbol{\alpha}) \times \prod_{n=1}^{N} \text{Mult}(z_{d,n}|\boldsymbol{\theta}_d) \\
&= \prod_k \left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \\
&= \exp\left[\log\left(\prod_k \left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp\left[\sum_k \log\left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right] \\
&= \exp\left[\sum_k \left(\log \theta_{d,k}^{\alpha_k - 1} + \sum_{n=1}^{N} \log\left(\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp\left[\sum_k \left((\alpha_k - 1) \log \theta_{d,k} + \sum_{n=1}^{N} \mathbb{1}(z_{d,n} = k) \log \theta_{d,k} \right) \right] \\
&= \exp\left[\sum_k \left(\alpha_k - 1 + \sum_{n=1}^{N} \mathbb{1}(z_{d,n} = k) \right) \log\left(\theta_{d,k}\right) \right] \\
&= \exp\left(\underbrace{\begin{bmatrix} (\alpha_1 - 1 + n_1) \\ \dots \\ (\alpha_K - 1 + n_K) \end{bmatrix}^{\top}}_{\eta_l(\alpha, z_d)} \underbrace{\begin{bmatrix} \log(\theta_{d,1}) \\ \dots \\ \log(\theta_{d,K}) \end{bmatrix}}_{T(\boldsymbol{\theta}_d)} \right) \qquad \text{by letting } n_k = \sum_{n=1}^{N} \mathbb{1}(z_{d,n} = k) \\
&= \text{Dir}(\alpha_1 + n_1, \dots, \alpha_K + n_K)
\end{aligned}
\tag{106}
$$

It turns out that the Dirichlet distribution is the only distribution with exactly the same "normal" and natural parameters.

7.3.2 optimize $\boldsymbol{\gamma}_d$

let's use $q(\eta(\boldsymbol{\gamma}_d)) = \text{Dir}((\eta(\boldsymbol{\gamma}_d))$ to approximate $p(\boldsymbol{\theta}_d|\mathbf{z}_d)$ (another Dirichlet distribution), we apply for the exponential family update formula Eq.(53):

$$
\begin{aligned}
\eta(\boldsymbol{\gamma}_d) &= \mathbb{E}_{q(z_{d,n}|\boldsymbol{\phi}_{d,n})} \left[\eta_l \left(\alpha, z_d \right) \right] \\
&= \mathbb{E}_{q(z_{d,n}|\boldsymbol{\phi}_{d,n})} \left[(\alpha_1 - 1 + n_1) \quad \dots \quad (\alpha_K - 1 + n_K) \right] \quad \text{substitute Eq.(??)}
\end{aligned}
\tag{107}
$$

compare this with Eq.(53), i.e., $\eta_j = \mathbb{E}_{q(\mathbf{z}\setminus z_j)}[\eta_{\text{post}}(\mathbf{z} \setminus z_j)]$, you can see easily that:

$$
\mathbf{z} \setminus z_j \equiv \{z_{d,n}\} \tag{108}
$$

how to compute $\mathbb{E}_{q(z_{d,n}|\boldsymbol{\phi}_{d,n})}\big[n_1)\big]$? Firstly, let's recognize that $\boldsymbol{\phi}_{d,n}$ itself is a probability vector:

$$\begin{aligned}\boldsymbol{\phi}_{d,n} &= \begin{bmatrix}\boldsymbol{\phi}_{d,n}^1 & \dots & \boldsymbol{\phi}_{d,n}^K\end{bmatrix} \\ &= \begin{bmatrix}q(z_{d,n}=1) & \dots & q(z_{d,n}=K)\end{bmatrix}\end{aligned} \tag{109}$$

since:

$$\begin{aligned}\mathbb{E}\big[\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=k)\big] &= \sum_{n=1}^{N}\mathbb{E}[\mathbb{1}(z_{d,n}=k)] \\ &= \sum_{n=1}^{N} q(z_{d,n}=k) \\ &= \sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}^k\end{aligned} \tag{110}$$

continue with Eq.(**??**), we have:

$$\eta(\boldsymbol{\gamma}_d) = \begin{bmatrix}\alpha_1 - 1 + \sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}^k & \dots & \alpha_K - 1 + \sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}^k\end{bmatrix} \tag{111}$$

to obtain the Dirichlet distribution parameter $\boldsymbol{\gamma}_d$, luckily, for Dirichlet distribution, the natural and "normal" parameter are the same.

$$\begin{aligned}\boldsymbol{\gamma}_d &= \begin{bmatrix}\left(\alpha_1 + \sum_{n=1}^{N}\phi_{d,n}^1\right) & \dots & \left(\alpha_K + \sum_{n=1}^{N}\phi_{d,n}^K\right)\end{bmatrix} \\ &= \boldsymbol{\alpha} + \sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}\end{aligned} \tag{112}$$

7.4 Updating $q(\boldsymbol{\beta}_k|\boldsymbol{\lambda}_k)$ optimize $\boldsymbol{\lambda}_k$

7.4.1 find natural parameter of posterior $p(\boldsymbol{\beta}_k|\mathbf{z},\mathbf{w})$

$$\begin{aligned}p(\boldsymbol{\beta}_k|\mathbf{z},\mathbf{w}) &= p(\boldsymbol{\beta}_k|\xi)\prod_{d=1}^{D}\prod_{n=1}^{N} p\left(w_{d,n}|\boldsymbol{\beta}_k\right)^{\mathbb{1}(z_{d,n}=k)} = \text{Dir}(\eta)\times\prod_{d=1}^{D}\prod_{n=1}^{N}\beta_{k,w_{d,n}}^{\mathbb{1}(z_{d,n}=k)} \\ &\propto \exp\left(\underbrace{\left(\xi - 1 + \sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\mathbb{1}(z_{d,n}=k)\right)}_{\eta_l(\eta,Z,W)}\times\underbrace{\log(\boldsymbol{\beta}_k)}_{T(\boldsymbol{\beta}_k)}\right) \\ &= \text{Dir}(\eta_l\left(\xi,Z,W\right))\end{aligned} \tag{113}$$

7.4.2 optimize $\boldsymbol{\lambda}_k$

let's use $q(\eta(\boldsymbol{\lambda}_k)) = \text{Dir}((\eta(\boldsymbol{\lambda}_k))$ to approximate $p(\boldsymbol{\beta}_k|\mathbf{z},\mathbf{w})$ (another Dirichlet distribution), we apply for the exponential family update formula Eq.(53):

$$
\begin{aligned}
\eta(\boldsymbol{\lambda}_k) &= \mathbb{E}_{\prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}^{k})}\left[\eta_l(\xi, \mathbf{z}, \mathbf{w})\right] \\
&= \mathbb{E}_{\prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}^{k})}\left[\xi - 1 + \sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\mathbb{1}(z_{d,n}=k)\right] \\
&= \eta - 1 + \sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\phi_{d,n}^{k}
\end{aligned}
\tag{114}
$$

$$
\boldsymbol{\lambda}_k = \xi + \sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\phi_{d,n}^{k} \tag{115}
$$

8 Collapsed Variational Inference Optional for Exam

$$q(z_{d,n}) = \text{Mult}(\boldsymbol{\phi}_{d,n}) \text{ or } q(z_{d,n} = k) = \phi_{d,n}^k \quad q(\boldsymbol{\beta}_k) = \text{Dir}(\boldsymbol{\lambda}_k) \quad q(\boldsymbol{\theta}_d) = \text{Dir}(\boldsymbol{\gamma}_d) \tag{116}$$

$$\begin{aligned} \implies q(Z, \theta_1 \dots \boldsymbol{\theta}_d, \beta_1 \dots \boldsymbol{\beta}_k) &= \left(\prod_{d=1}^{d=D} \prod_{n=1}^{N} q(z_{d,n}|\boldsymbol{\phi}_{d,n}) \right) \prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{k=1}^{K} q(\theta_k|\boldsymbol{\lambda}_k) \\ \text{now change to: } &= \underbrace{\left(\prod_{d=1}^{d=D} \prod_{n=1}^{N} q(z_{d,n}|\boldsymbol{\phi}_{d,n}) \right)}_{q(Z)} q(\Theta, \beta|Z) \end{aligned} \tag{117}$$

Maximize ELOB, it becomes: (remove X for clarity)
Let $U = \{\Theta, \beta\}$:

$$\begin{aligned} \text{ELBO}(q) &\triangleq \mathbb{E}_{q(U,Z)}[\log p(Z,U)] - \mathbb{E}_{q(U,Z)}[\log q(Z,U)] \\ &= \mathbb{E}_{q(U,Z)}[\log p(Z,U)] - \mathbb{E}_{q(U,Z)}[\log q(U|Z) - \log q(Z)] \\ &= \mathbb{E}_{q(Z)}\left(\mathbb{E}_{q(U|Z)}[\log p(Z,U)]\right) - \mathbb{E}_{q(Z)}\left(\mathbb{E}_{q(U|Z)}[\log q(U|Z)]\right) - \mathbb{E}_{q(Z,U)}[\log q(Z)] \\ &= \mathbb{E}_{q(Z)}\left(\underbrace{\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)}_{\mathcal{L}(q(U|Z))} \right) - \mathbb{E}_{q(Z)}[\log q(Z)] \end{aligned} \tag{118}$$

Think this as treating Z as X.
(removed X for clarity)

$$\begin{aligned} \arg\max_{q(U|Z)}(\text{ELBO}(q)) &= \arg\max_{q(U|Z)} \left[\mathbb{E}_{q(Z)}\left(\underbrace{\mathbb{E}_{q(U|Z)}\left([\log p_X(Z,U)] - [\log q(U|Z)]\right)}_{\mathcal{L}(q(U|Z))} \right) - \mathbb{E}_{q(Z)}[\log q(Z)] \right] \\ &= \mathbb{E}_{q(Z)}\left(\underbrace{\arg\max_{q(U|Z)} \left[\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)\right]}_{} \right) - \mathbb{E}_{q(Z)}[\log q(Z)] \\ &= \mathbb{E}_{q(Z)}[\underbrace{p(Z)}_{}] - \mathbb{E}_{q(Z)}[\log q(Z)] \end{aligned} \tag{119}$$

$$\arg\max_{q(U|Z)} \left[\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)\right] = p(Z) \tag{120}$$

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