VIBP derivations

Joey

2021

1 Introduction

Existing Variational Indian Buffet Process methods do not update the hyper-parameters, such as α, σ_A, and σ_n (Doshi et al., 2009). Also, it's using the mean field approximation rather than the stochastic variant. This note is for deriving the equations myself.

2 Derivations

Directly start with the infinite linear gaussian model.

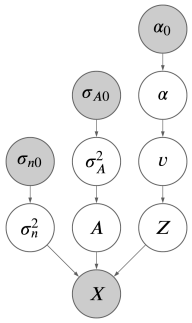

Figure 1: (a) True model $p(\boldsymbol{W})$

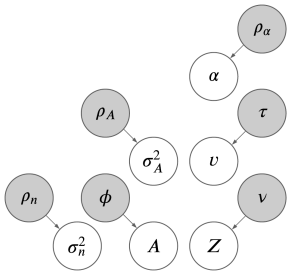

Figure 2: (b) Variational approximation $q(\boldsymbol{W})$

Denote the set of hidden variables in IBP by $\boldsymbol{W} = \{\boldsymbol{v}, \boldsymbol{Z}, \boldsymbol{A}, \alpha, \sigma_A^2, \sigma_n^2\}$ (σ_A^2 and σ_n^2 are precision, for simpler update formula later. Also, consider them as a single term own, not a squared term) and the set of hyper-parameters by $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_0, \boldsymbol{\sigma}_{A0}, \boldsymbol{\sigma}_{n0}\}$. The true log posterior

$$\log p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{1}$$

is difficult due to the intractability of computing the log marginal probability $\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \int \log p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})\boldsymbol{W}$

Mean field variational methods approximate the true posterior with a variational distribution $q_{\boldsymbol{\Phi}}(\boldsymbol{W})$, where $\boldsymbol{\Phi}$ denotes the set of parameters $\{\phi, \tau, \nu, \rho_n, \rho_A, \rho_\alpha\}$.

Since σ_A^2, σ_n^2, and α are all positive, I set them to follow a Gamma distribution. This will prevent the troublesome term if using a normal distribution. To be clear, the hyper-parameter of Gamma distribution, for example, $\boldsymbol{\alpha}_0$, is a vector includes $\{\alpha_s, \alpha_r\}$, where s stands for shape and r stands for rate. The model for p is the full stick-breaking construction for IBP:

$$\begin{aligned}
\alpha &\sim Gamma(\alpha_s, \alpha_r) \\
\sigma_A^2 &\sim Gamma(\sigma_{As}, \sigma_{Ar}) \\
\sigma_n^2 &\sim Gamma(\sigma_{ns}, \sigma_{nr}) \\
v_k &\sim Beta(\alpha, 1) \\
\pi_k &= \prod_{i=1}^{k} v_i \\
z_{nk} &\sim Bernoulli(\pi_k) \\
\boldsymbol{A}_{k\cdot} &\sim Normal(0, \sigma_A^2 \boldsymbol{I}) \ \sigma_A^2 \text{ is precision} \\
\boldsymbol{X}_{n\cdot} &\sim Normal(\boldsymbol{Z}_n \boldsymbol{A}, \sigma_n^2 \boldsymbol{I}) \ \sigma_n^2 \text{ is precision}
\end{aligned} \tag{2}$$

The joint is

$$\begin{aligned}
p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{k=1}^{\inf} \left(p(v_k|\alpha) p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I) \prod_{n=1}^{N} p(z_{nk}|v_k) \right) \\
\prod_{n=1}^{N} p(\boldsymbol{X}_n|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \sigma_n^2 I) p(\alpha|\boldsymbol{\alpha}_0) p(\sigma_A^2|\boldsymbol{\sigma}_{A0}) p(\sigma_n^2|\boldsymbol{\sigma}_{n0})
\end{aligned} \tag{3}$$

We use variational approximation. A truncated stick-breaking process is used here (Blei and Jordan, 2004).

Our mean field variational distribution is:

$$q(\boldsymbol{W}) = q_{\boldsymbol{\rho}_\alpha}(\alpha) q_{\boldsymbol{\rho}_A}(\sigma_A^2) q_{\boldsymbol{\rho}_n}(\sigma_n^2) q_{\boldsymbol{\tau}}(\boldsymbol{v}) q_{\boldsymbol{\theta}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z}) \tag{4}$$

where

$$\begin{aligned}
q_{\boldsymbol{\rho}_\alpha}(\alpha) &= Gamma(\alpha; \rho_{\alpha s}, \rho_{\alpha r}) \\
q_{\boldsymbol{\rho}_A}(\sigma_A^2) &= Gamma(\sigma_A^2; \rho_{As}, \rho_{Ar}) \\
q_{\boldsymbol{\rho}_n}(\sigma_n^2) &= Gamma(\sigma_n^2; \rho_{ns}, \rho_{nr}) \\
q_{\boldsymbol{\tau}_k}(v_k) &= Beta(v_k; \tau_{k1}, \tau_{k2}) \\
q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) &= Normal(\boldsymbol{A}_{k\cdot}; \overline{\phi}_k, \boldsymbol{\Phi}_k) \ \boldsymbol{\Phi}_k \text{ is variance} \\
q_{\boldsymbol{\nu}_{nk}}(z_{nk}) &= Bernoulli(z_{nk}; \nu_{nk})
\end{aligned} \tag{5}$$

Inference involves optimising parameters to minimise the KL divergence $D(q||p)$, or equivalently to maximise the ELBO.

2.1 Lower Bound on the Marginal Likelihood

Start from the KL divergence (Kingma and Welling, 2013; Odaibo, 2019),

$$
\begin{aligned}
&D_{KL}(q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)||p(\boldsymbol{W}|\boldsymbol{X},\theta)) \geq 0 \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log\left(\frac{p(\boldsymbol{W}|\boldsymbol{X},\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right)d\boldsymbol{W} \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)p(\boldsymbol{X}|\theta)}\right)d\boldsymbol{W} \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\left[\log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right)-\log p(\boldsymbol{X}|\theta)\right]d\boldsymbol{W} \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right)d\boldsymbol{W}+\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log p(\boldsymbol{X}|\theta)d\boldsymbol{W} \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right)d\boldsymbol{W}+\log p(\boldsymbol{X}|\theta)\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)d\boldsymbol{W} \\
&= -\int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right)d\boldsymbol{W}+\log p(\boldsymbol{X}|\theta) \geq 0
\end{aligned}
\tag{6}
$$

reordering and applying rules of logarithms,

$$
\begin{aligned}
\log p(\boldsymbol{X}|\theta) &\geq \int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\Big[\log p(\boldsymbol{X}|\boldsymbol{W},\theta)+\log p(\boldsymbol{W}|\theta)-\log q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\Big]d\boldsymbol{W} \\
&= \mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\Big[\log p(\boldsymbol{X}|\boldsymbol{W},\theta)+\log p(\boldsymbol{W}|\theta)-\log q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\Big] \\
&= \mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\Big[\log p(\boldsymbol{X},\boldsymbol{W}|\theta)-\log q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)\Big] \\
&= \mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\big[\log p(\boldsymbol{X},\boldsymbol{W}|\theta)\big]+H[q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)]
\end{aligned}
\tag{7}
$$

where H is the entropy of distribution q. The term on the RHS is usually called variational lower bound on the marginal likelihood or the Evidence Lower Bound (ELBO). From Equation 6, we also have

$$
\begin{aligned}
\log p(\boldsymbol{X}|\theta) &\geq \int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta) \log\left(\frac{p(\boldsymbol{X}|\boldsymbol{W},\theta)p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right) d\boldsymbol{W} \\
&= \int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta) \log\left(\frac{p(\boldsymbol{W}|\theta)}{q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\right) d\boldsymbol{W} + \int q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta) \log p(\boldsymbol{X}|\boldsymbol{W},\theta) d\boldsymbol{W} \\
&= -D_{KL}(q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)||\log p(\boldsymbol{W}|\theta)) + \mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}[\log p(\boldsymbol{X}|\boldsymbol{W},\theta)]
\end{aligned} \tag{8}
$$

Therefore,

$$
\begin{aligned}
&\arg \min_{\tau,\theta,\nu,\rho_\alpha,\rho_A,\rho_n} D_{KL}(q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)||p(\boldsymbol{W}|\boldsymbol{X},\theta)) \\
=&\arg \max_{\tau,\theta,\nu,\rho_\alpha,\rho_A,\rho_n} -D_{KL}(q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)||\log p(\boldsymbol{W}|\theta)) + \mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}[\log p(\boldsymbol{X}|\boldsymbol{W},\theta)
\end{aligned} \tag{9}
$$

2.2 Decomposing

Continue from Equation 7, the RHS can be decomposed as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{W}\sim q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)}\big[\log p(\boldsymbol{X},\boldsymbol{W}|\theta)\big] + H[q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)] \\
=&\mathbb{E}_{\alpha\sim q}[\log p(\alpha|\boldsymbol{\alpha}_0)] + \mathbb{E}_{\sigma_A^2\sim q}[\log p(\sigma_A^2|\boldsymbol{\sigma}_{A0})] + \mathbb{E}_{\sigma_n^2\sim q}[\log p(\sigma_n^2|\boldsymbol{\sigma}_{n0})] \\
&+ \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{v},\alpha\sim q}[\log p(v_k|\alpha)] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}\sim q}[\log p(Z_{nk}|\boldsymbol{v}) \\
&+ \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A},\sigma_A^2\sim q}[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)] + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\sigma_n^2\sim q}[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)] \\
&+ H[q_{\boldsymbol{\Phi}}(\boldsymbol{W}|\boldsymbol{X},\theta)]
\end{aligned} \tag{10}
$$

To make it shorter, we will omit $\sim q$ in expectation term if possible from now on.

1-3. The first three terms are similar, since they all follow Gamma distribution. So here I derive use a random variable x as an example for α, σ_A^2, and σ_n^2.

Suppose X has $\Gamma(\alpha,\beta)$ distribution and we'd like to get the expectation of $Y = \log(X)$. Because β is a rate parameter, it will shift by $-\log(\beta)$. Therefore we start with the case $\beta = 1$. The probability element of X is

$$
f_X(x) = \frac{1}{\Gamma(\alpha)x^\alpha e^{-x}}\frac{dx}{x} \tag{11}
$$

where $\Gamma(\alpha) = \int_0^{\inf} x^\alpha e^{-x} \frac{dx}{x}$ is a constant. Substitute $x = e^y$, and $\frac{dx}{x} = dy$

$$f_Y(y) = \frac{1}{\Gamma(\alpha) e^{\alpha y - e^x}} dy \tag{12}$$

which is a differentiable function of α. Gives

$$\frac{d}{d\alpha} e^{\alpha y - e^x} dy = y\Gamma(\alpha) f_Y(y) \tag{13}$$

So,

$$\begin{aligned}
\mathbb{E}(Y) &= \int y f_Y(y) = \frac{1}{\Gamma(\alpha)} \int \frac{d}{d\alpha} e^{\alpha y - e^x} dy \\
&= \frac{1}{\Gamma(\alpha)} \frac{d}{d\alpha} \int e^{\alpha y - e^x} dy \\
&= \frac{1}{\Gamma(\alpha)} \frac{d}{d\alpha} \Gamma(\alpha) \\
&= \frac{d}{d\alpha} \log \Gamma(\alpha) \\
&= \psi(\alpha)
\end{aligned} \tag{14}$$

where $\psi(\cdot)$ is the digamma function.

Reintroducing β

$$\mathbb{E}(\log(X)) = -\log(\beta) + \psi(\alpha) \tag{15}$$

Therefore, the expectation we want is

$$\begin{aligned}
\mathbb{E}_{x \sim q(\alpha_q, \beta_q)} \Big[\log p(x|\alpha_p, \beta_p) \Big] &= \mathbb{E}_{\sim q} \Big[\log(\frac{\beta_p^{\alpha_p}}{\Gamma(\alpha_p)} x^{\alpha_p - 1} e^{-\beta_p x}) \Big] \\
&= \mathbb{E}[\frac{\beta_p^{\alpha_p}}{\Gamma(\alpha_p)}] + (\alpha_p - 1)\mathbb{E}[\log x] - \beta_p \mathbb{E}[x] \\
&= \frac{\beta_p^{\alpha_p}}{\Gamma(\alpha_p)} + (\alpha_p - 1)(-\log(\beta_q) + \psi(\alpha_q)) - \beta_p \frac{\alpha_q}{\beta_q}
\end{aligned} \tag{16}$$

As to α, σ_A^2, and σ_n^2,

$$\begin{aligned}
\mathbb{E}_\alpha[\log p(\alpha|\boldsymbol{\alpha}_0)] &= \frac{\alpha_r^{\alpha_s}}{\Gamma(\alpha_s)} + (\alpha_s - 1)(-\log(\rho_{\alpha r}) + \psi(\rho_{\alpha s})) - \alpha_s \frac{\rho_{\alpha s}}{\rho_{\alpha r}} \\
\mathbb{E}_{\sigma_A^2}[\log p(\sigma_A^2|\boldsymbol{\sigma}_{A0})] &= \frac{\sigma_{Ar}^{\sigma_{As}}}{\Gamma(\sigma_{As})} + (\sigma_{As} - 1)(-\log(\rho_{Ar}) + \psi(\rho_{As})) - \sigma_{As} \frac{\rho_{As}}{\rho_{Ar}} \\
\mathbb{E}_{\sigma_n^2}[\log p(\sigma_n^2|\boldsymbol{\sigma}_{n0})] &= \frac{\sigma_{nr}^{\sigma_{ns}}}{\Gamma(\sigma_{ns})} + (\sigma_{ns} - 1)(-\log(\rho_{nr}) + \psi(\rho_{ns})) - \sigma_{ns} \frac{\rho_{ns}}{\rho_{nr}}
\end{aligned} \tag{17}$$

4. According to the derivation in the previous part, the forth term becomes

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\alpha}[\log p(v_k|\alpha)] &= \mathbb{E}_{\boldsymbol{v},\alpha}[\log(\alpha v_k^{\alpha-1})] \\ &= \mathbb{E}_{\alpha}[\log\alpha] + \mathbb{E}_{\alpha}[(\alpha-1)]\mathbb{E}_{\boldsymbol{v}}[\log(v_k)] \\ &= \big(-\log(\rho_{\alpha r}) + \psi(\rho_{\alpha s})\big) + \big(\frac{\rho_{\alpha s}}{\rho_{\alpha r}} - 1\big)\big(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\big)\end{aligned} \tag{18}$$

5. The fifth term becomes

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}[\log p(Z_{nk}|\boldsymbol{v}) &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}[\log p(z_{nk}=1|\boldsymbol{v})^{z_{nk}} p(z_{nk}=0|\boldsymbol{v})^{1-z_{nk}}] \\ &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}[z_{nk}\log p(z_{nk}=1|\boldsymbol{v}) + (1-z_{nk})\log p(z_{nk}=0|\boldsymbol{v})] \\ &= \mathbb{E}_{\boldsymbol{Z}} z_{nk}\mathbb{E}_{\boldsymbol{v}}\log p(z_{nk}=1|\boldsymbol{v}) + \mathbb{E}_{\boldsymbol{Z}}(1-z_{nk})\mathbb{E}_{\boldsymbol{v}}\log p(z_{nk}=0|\boldsymbol{v}) \\ &= \nu_{nk}\mathbb{E}_{\boldsymbol{v}}\left[\log\prod_{m=1}^{k} v_m\right] + (1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right] \\ &= \nu_{nk}\left(\sum_{m=1}^{k}\big(\psi(\tau_{m1}) - \psi(\tau_{m1}+\tau_{m2})\big)\right) + (1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]\end{aligned} \tag{19}$$

The last expectation can be bounded using an auxiliary multinomial approximation $q_k = (q_{k1}, q_{k2}, \ldots, q_{kk})$. The detailed derivation is in Doshi et al., 2009 Section 5.1. Only list the final expression here

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{k=1}^{k} v_m\right)\right] \geq &\left(\sum_{m=1}^{k} q_{km}\psi(\tau_{m2})\right) + \left(\sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k} q_{kn}\right)\psi(\tau_{m1})\right) \\ &-\left(\sum_{m=1}^{k}\left(\sum_{n=m}^{k} q_{kn}\right)\psi(\tau_{m1}+\tau_{m2})\right) - \sum_{m=1}^{k} q_{km}\log q_{km}\end{aligned} \tag{20}$$

where

$$q_{ki} \propto \exp\left(\psi(\tau_{i2}) + \sum_{m=1}^{i-1}\psi(\tau_{m1}) - \sum_{m=1}^{i}\psi(\tau_{m1}+\tau_{m2})\right) \tag{21}$$

6. First, the expectation of the inverse of a $Gamma(\alpha,\beta)$ distributed random variable is required.

$$\begin{aligned}\mathbb{E}_{x\sim p(\alpha,\beta)}[\frac{1}{X}] =& \int \frac{1}{x}\beta\frac{(\beta x)^{\alpha-1}}{\Gamma(\alpha)}e^{-\beta x}dx \\ =&\beta\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}\int\frac{(\beta x)^{\alpha-2}}{\Gamma(\alpha-1)}e^{-\beta x}dx \\ =&\beta\frac{(\alpha-1)!}{(\alpha-2)!}\cdot 1 \\ =&\frac{\beta}{\alpha-1}\end{aligned} \tag{22}$$

The integral is 1 because it's the density of $Gamma(\alpha-1,\beta)$.

Also, according to the multivariate quadratic form

$$\mathbb{E}[\boldsymbol{A}^T\boldsymbol{\Lambda}\boldsymbol{A}] = tr[\boldsymbol{\Lambda}\boldsymbol{\Sigma}] + \mu^T\boldsymbol{\Lambda}\mu \tag{23}$$

where μ and $\boldsymbol{\Sigma}$ are the expected value and covariance matrix of vector $\boldsymbol{A}$. Therefore, the sixth term becomes

$$\begin{aligned}\mathbb{E}_{\boldsymbol{A},\sigma_A^2}[\log p(\boldsymbol{A}_{k\cdot}|0,\sigma_A^2 I)] &= \mathbb{E}_{\boldsymbol{A},\sigma_A^2}\left[\log\left(\frac{\sigma_A^2}{(2\pi)^{D/2}}^{D/2}e^{-\frac{1}{2}\boldsymbol{A}_{k\cdot}^T\sigma_A^2\boldsymbol{A}_{k\cdot}}\right)\right] \\ &= \frac{D}{2}\mathbb{E}_{\sigma_A^2}[\log(\frac{\sigma_A^2}{2\pi})] - \frac{1}{2}\mathbb{E}_{\sigma_A^2}[\sigma_A^2]\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}] \\ &= \frac{D}{2}\left(\mathbb{E}_{\sigma_A^2}\log(\sigma_A^2) - \log(2\pi)\right) - \frac{1}{2}\mathbb{E}_{\sigma_A^2}[\sigma_A^2]\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}] \\ &= \frac{D}{2}\left((-\log(\rho_{Ar}) + \psi(\rho_{As}) - \log(2\pi)\right) - \frac{\rho_{As}}{2\rho_{Ar}}\left(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T\right)\end{aligned} \tag{24}$$

7. The seventh term, which is the likelihood, becomes

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\sigma_n^2}[\log(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)]\\
=&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\sigma_n^2}\left[\log\left((\frac{\sigma_n^2}{2\pi})^{D/2}e^{-\frac{1}{2}(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}A)\sigma_n^2(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}A)^T}\right)\right]\\
=&\mathbb{E}_{\sigma_n^2}[\log(\frac{\sigma_n^2}{2\pi})^{D/2}]-\frac{1}{2}\mathbb{E}_{\sigma_n^2}[\sigma_n^2]\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}A)(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}A)^T]\\
=&\frac{D}{2}\Big(\big(-\log(\rho_{nr})+\psi(\rho_{ns})\big)-\log(2\pi)\Big)\\
&-\frac{1}{2}\mathbb{E}_{\sigma_n^2}[\sigma_n^2]\Big(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T-2\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_{n\cdot}]\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}]\boldsymbol{X}_{n\cdot}^T+\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T]\Big)\\
=&\frac{D}{2}\Big(\big(-\log(\rho_{nr})+\psi(\rho_{ns})\big)-\log(2\pi)\Big)\\
&-(\frac{\rho_{ns}}{2\rho_{nr}})\Big(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T-2\sum_{k=1}^{K}\nu_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T+2\sum_{k<k'}\nu_{nk}\nu_{nk'}\overline{\phi}_k\overline{\phi}_{k'}^T+\sum_{k=1}^{K}\nu_{nk}(tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T)\Big)
\end{aligned}
\tag{25}
$$

where the last expectation is derived by

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T]=&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\left(\sum_{k=1}^{K}z_{nk}\boldsymbol{A}_{k\cdot}\right)\left(\sum_{k=1}^{K}z_{nk}\boldsymbol{A}_{k\cdot}\right)^T\right]\\
=&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\sum_{d=1}^{D}\left(\sum_{k=1}^{K}z_{nk}^2A_{kd}^2+2\sum_{k,k':k'\neq k}z_{nk}z_{nk'}\boldsymbol{A}_{kd}\boldsymbol{A}_{k'd}\right)^T\right]\\
=&\sum_{k=1}^{K}\nu_{nk}(tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T)+2\sum_{k<k'}\nu_{nk}\nu_{nk'}\overline{\phi}_k\overline{\phi}_{k'}^T
\end{aligned}
\tag{26}
$$

8. The last term entropy

$$
\begin{aligned}
H[q]=&-\mathbb{E}_q\log\left[q_{\boldsymbol{\rho}_\alpha}(\alpha)q_{\boldsymbol{\rho}_A}(\sigma_A^2)q_{\boldsymbol{\rho}_n}(\sigma_n^2)\prod_{k=1}^{K}q_{\boldsymbol{\tau}_k}(\boldsymbol{v}_k)\prod_{k=1}^{K}q_{\boldsymbol{\theta}_k}(\boldsymbol{A}_{k\cdot})\prod_{k=1}^{K}\prod_{n=1}^{N}q_{\boldsymbol{\nu}_{nk}}(z_{nk})\right]\\
=&\mathbb{E}_\alpha(-\log q_{\boldsymbol{\rho}_\alpha}(\alpha))+\mathbb{E}_{\sigma_A^2}(-\log q_{\boldsymbol{\rho}_A}(\sigma_A^2))+\mathbb{E}_{\sigma_n^2}(-\log q_{\boldsymbol{\rho}_n}(\sigma_n^2))\\
&+\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{v}}(-\log q_{\boldsymbol{\tau}_k}(\boldsymbol{v}_k))+\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\theta}_k}(\boldsymbol{A}_{k\cdot}))+\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\boldsymbol{\nu}_{nk}}(z_{nk}))
\end{aligned}
\tag{27}
$$

The first three term are still in the same form. So again, as an example,

$$\begin{aligned}\mathbb{E}_{x\sim\Gamma(\alpha,\beta)}(-\log(q(x))) =& \mathbb{E}_x\left[-\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\right)\right] \\ =& \log(\frac{\Gamma(\alpha)}{\beta^\alpha}) - (\alpha-1)\mathbb{E}[\log x] + \beta\mathbb{E}[x] \\ =& \log(\frac{\Gamma(\alpha)}{\beta^\alpha}) - (\alpha-1)(-\log(\beta)+\psi(\alpha)) + \alpha\end{aligned} \tag{28}$$

Therefore, we can write down each term explicitly as below

$$\begin{aligned}\mathbb{E}_{\alpha}(-\log q_{\boldsymbol{\rho}_\alpha}(\alpha)) =& \log(\frac{\Gamma(\rho_{\alpha s})}{\rho_{\alpha r}^{\rho_{\alpha s}}}) - (\rho_{\alpha s}-1)(-\log(\rho_{\alpha r})+\psi(\rho_{\alpha s})) + \rho_{\alpha s} \\ \mathbb{E}_{\sigma_A^2}(-\log q_{\boldsymbol{\rho}_A}(\sigma_A^2)) =& \log(\frac{\Gamma(\rho_{As})}{\rho_{Ar}^{\rho_{As}}}) - (\rho_{As}-1)(-\log(\rho_{Ar})+\psi(\rho_{As})) + \rho_{As} \\ \mathbb{E}_{\sigma_n^2}(-\log q_{\boldsymbol{\rho}_n}(\sigma_n^2)) =& \log(\frac{\Gamma(\rho_{ns})}{\rho_{nr}^{\rho_{ns}}}) - (\rho_{ns}-1)(-\log(\rho_{nr})+\psi(\rho_{ns})) + \rho_{ns} \\ \mathbb{E}_{\boldsymbol{v}}(-\log q_{\boldsymbol{\tau}_k}(v_k)) =& \log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) \\ & -(\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2}) \\ \mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) =& \frac{1}{2}\log\left((2\pi e)^D|\boldsymbol{\Phi}_k|\right) \\ \mathbb{E}_{\boldsymbol{Z}}(-\log q_{\boldsymbol{\nu}_{nk}}(z_{nk})) =& -\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk})\end{aligned} \tag{29}$$

Put them together gives the variational lower bound.

2.3 Parameter Updates

Since they are in the exponential family, we use an alternate formula to update the parameters (Beal, 2003)

For any variational parameters ξ_i that correspond to W_i, the optimal ξ_i are the solution to

$$\log q_{\xi_i}(W_i) = \mathbb{E}_{\boldsymbol{W}_{-i}}[\log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})] + c \tag{30}$$

1. Updates the feature weight matrix $\boldsymbol{A}$

$$
\begin{aligned}
\log q_{\phi_k}(\boldsymbol{A}_k) =& \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\sigma_A^2,\sigma_n^2}[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})] + c \\
=& \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\sigma_A^2,\sigma_n^2}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2) + \sum_{n=1}^{N}\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right] + c \\
=& \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\sigma_A^2,\sigma_n^2}\Bigg[\log\Big((\frac{\sigma_A^2}{2\pi})^{D/2}e^{-\frac{\boldsymbol{A}_{k\cdot}\sigma_A^2\boldsymbol{A}_{k\cdot}^T}{2}}\Big) \\
&+ \sum_{n=1}^{N}\log\Big((\frac{\sigma_n^2}{2\pi})^{D/2}e^{-\frac{1}{2}(\boldsymbol{X}_n-\boldsymbol{Z}_{n\cdot}\boldsymbol{A})\sigma_n^2(\boldsymbol{X}_n-\boldsymbol{Z}_{n\cdot}\boldsymbol{A})^T}\Big)\Bigg] + c \\
=& \frac{D}{2}\mathbb{E}_{\sigma_A^2}[\sigma_A^2]\mathbb{E}_{\boldsymbol{A}_{-k}}(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T) \\
&- \frac{1}{2}\mathbb{E}_{\sigma_n^2}[\sigma_n^2]\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}[(\boldsymbol{X}_n-\boldsymbol{Z}_{n\cdot}\boldsymbol{A})(\boldsymbol{X}_n-\boldsymbol{Z}_{n\cdot}\boldsymbol{A})^T] + c \\
=& \frac{D}{2}\mathbb{E}_{\sigma_A^2}[\sigma_A^2](\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T) \\
&- \frac{1}{2}\mathbb{E}_{\sigma_n^2}[\sigma_n^2]\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}[(\boldsymbol{X}_n-\boldsymbol{Z}_{n(-k)}\boldsymbol{A}_{-k}-\boldsymbol{Z}_{nk}\boldsymbol{A}_k)(\boldsymbol{X}_n-\boldsymbol{Z}_{n(-k)}\boldsymbol{A}_{-k}-\boldsymbol{Z}_{nk}\boldsymbol{A}_k)^T] + c \\
=& \frac{D\rho_{As}}{2\rho_{Ar}}(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T) \\
&- \frac{\rho_{ns}}{2\rho_{nr}}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}\Bigg[-2\boldsymbol{Z}_{nk}\boldsymbol{A}_{k\cdot}(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n(-k)}\boldsymbol{A}_{-k})^T + \boldsymbol{Z}_{nk}\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T\boldsymbol{Z}_{nk}^T\Bigg] + c \\
=& \frac{D\rho_{As}}{2\rho_{Ar}}(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T) \\
&- \frac{\rho_{ns}}{2\rho_{nr}}\sum_{n=1}^{N}\left(-2\nu_{nk}\boldsymbol{A}_{k\cdot}\left(\boldsymbol{X}_{n\cdot}-\sum_{l:l\neq k}\nu_{nl}\cdot\overline{\phi}_l\right) + \mathbb{E}[\boldsymbol{Z}_{nk}^2](\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T)\right) + c \\
=& \frac{D\rho_{As}}{2\rho_{Ar}}(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T) \\
&- \frac{\rho_{ns}}{2\rho_{nr}}\sum_{n=1}^{N}\left(-2\nu_{nk}\boldsymbol{A}_{k\cdot}\left(\boldsymbol{X}_{n\cdot}-\sum_{l:l\neq k}\nu_{nl}\cdot\overline{\phi}_l\right)^T + \nu_{nk}(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T)\right) + c \\
=& \frac{1}{2}\left(\boldsymbol{A}_{k\cdot}\left(\frac{D\rho_{As}}{\rho_{Ar}} - \frac{\rho_{ns}\sum_{n=1}^{N}\nu_{nk}}{\rho_{nr}}\right)\boldsymbol{A}_{k\cdot}^T - 2\boldsymbol{A}_{k\cdot}\frac{\rho_{ns}\sum_{n=1}^{N}\nu_{nk}}{\rho_{nr}}\left(\boldsymbol{X}_{n\cdot}-\sum_{l:l\neq k}\nu_{nl}\cdot\overline{\phi}_l\right)^T\right) + c
\end{aligned}
\tag{31}
$$

q_ϕ is supposed to be a matrix normal. Therefore, it gives the updates

$$
\begin{aligned}
\overline{\phi}_k &= \left[\frac{\rho_{ns}\sum_{n=1}^{N}\nu_{nk}}{\rho_{nr}}\left(\boldsymbol{X}_{n\cdot} - \sum_{l:l\neq k}\nu_{nl}\cdot\overline{\phi}_l\right)\right]\left(\frac{D\rho_{As}}{\rho_{Ar}} - \frac{\rho_{ns}\sum_{n=1}^{N}\nu_{nk}}{\rho_{nr}}\right)^{-1} \\
\boldsymbol{\Phi}_k &= \left(\frac{D\rho_{As}}{\rho_{Ar}} - \frac{\rho_{ns}\sum_{n=1}^{N}\nu_{nk}}{\rho_{nr}}\right)^{-1} I
\end{aligned}
\tag{32}
$$

2. Updates the feature matrix $\boldsymbol{Z}$ at position nk

$$
\begin{aligned}
\log q_{\nu_{nk}}(z_{nk}) =& \mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk},\sigma_n^2}[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})] + c \\
=& \mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk},\sigma_n^2}[\log p(z_{nk}|v_k) + \log p(\boldsymbol{X}_n|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)] + c
\end{aligned}
\tag{33}
$$

where the first term

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}_{-nk}}[\log p(z_{nk}|v_k)] =& \mathbb{E}_{\boldsymbol{Z}_{-nk}} z_{nk}\mathbb{E}_{\boldsymbol{v}}\log p(z_{nk}=1|\boldsymbol{v}) + \mathbb{E}_{\boldsymbol{Z}_{-nk}}(1-z_{nk})\mathbb{E}_{\boldsymbol{v}}\log p(z_{nk}=0|\boldsymbol{v}) \\
=& z_{nk}\left(\sum_{m=1}^{k}\left(\psi(\tau_{m1}) - \psi(\tau_{m1}+\tau_{m2})\right)\right) + (1-z_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1 - \prod_{m=1}^{k} v_m\right)\right]
\end{aligned}
\tag{34}
$$

The remaining expectation is approximated following equation 19. The second term

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk},\sigma_n^2}[\log p(\boldsymbol{X}_n|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)] =& -\frac{1}{2}\mathbb{E}_{\sigma_n^2}[\sigma_n^2]\Big(-2\mathbb{E}_{\boldsymbol{Z}_{-nk}}[\boldsymbol{Z}_{n\cdot}]\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}]\boldsymbol{X}_{n\cdot}^T + \mathbb{E}_{\boldsymbol{Z}_{-nk},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T]\Big) + c \\
=& -\left(\frac{\rho_{ns}}{2\rho_{nr}}\right)\Big(-2z_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + \mathbb{E}_{\boldsymbol{Z}_{-nk},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T]\Big) + c
\end{aligned}
\tag{35}
$$

The last term is decomposed as follow ($z_{nk}^2 = z_{nk}$, since z is binary)

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}_{-nk},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T] =& \mathbb{E}_{\boldsymbol{Z}_{-nk},\boldsymbol{A}}\left[\left(\sum_{k=1}^{K} z_{nk}\boldsymbol{A}_{k\cdot}\right)\left(\sum_{k=1}^{K} z_{nk}\boldsymbol{A}_{k\cdot}\right)^T\right] \\
=& \mathbb{E}_{\boldsymbol{Z}_{-nk},\boldsymbol{A}}\left[\sum_{d=1}^{D}\left(\sum_{k=1}^{K} z_{nk}^2 A_{kd}^2 + 2\sum_{k,k':k'\neq k} z_{nk}z_{nk'}\boldsymbol{A}_{kd}\boldsymbol{A}_{k'd}\right)^T\right] \\
=& z_{nk}^2(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T) + 2z_{nk}\overline{\phi}_k\sum_{k'\neq k}\nu_{nk'}\overline{\phi}_{k'}^T \\
=& z_{nk}(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T) + 2z_{nk}\overline{\phi}_k\sum_{k'\neq k}\nu_{nk'}\overline{\phi}_{k'}^T
\end{aligned}
\tag{36}
$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}_{-nk}}[\log p(z_{nk}|v_k)] =& z_{nk}\Bigg(\sum_{m=1}^{k}\big(\psi(\tau_{m1})-\psi(\tau_{m1}+\tau_{m2})\big)\Bigg) - z_{nk}\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right] \\ &-(\frac{\rho_{ns}}{2\rho_{nr}})\Bigg(-2z_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + z_{nk}(tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T) + 2z_{nk}\overline{\phi}_k\sum_{k'\neq k}\nu_{nk'}\overline{\phi}_{k'}^T\Bigg)\end{aligned} \tag{37}$$

From the canonical parameterisation of Bernoulli distribution, we can get
(My understanding is that, assume a Bernoulli distribution $p(z;\nu)=\nu^z(1-\nu)^{1-z}$, $\log(p)=z(\log\frac{\nu}{1-\nu})$

$$\begin{aligned}\log\frac{\nu_{nk}}{1-\nu_{nk}} =& \sum_{m=1}^{k}\big(\psi(\tau_{m1})-\psi(\tau_{m1}+\tau_{m2})\big) - \mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right] \\ &-(\frac{\rho_{ns}}{2\rho_{nr}})\Bigg(-2\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + (tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T) + 2\overline{\phi}_k\sum_{k'\neq k}\nu_{nk'}\overline{\phi}_{k'}^T\Bigg) \\ \equiv& \vartheta\end{aligned} \tag{38}$$

Gives the update

$$\nu_{nk} = \frac{1}{1+e^{-\vartheta}} \tag{39}$$

3. Updates the feature probability τ
Identify terms that contain τ from the lower bound in equation 10

$$\begin{aligned}\mathcal{L}_\tau =& \sum_{k=1}^{K}[\log\alpha+(\alpha-1)(\psi(\tau_{k1})-\psi(\tau_{k1}+\tau_{k2}))] \\ &+\sum_{k=1}^{K}\sum_{n=1}^{N}\left[\nu_{nk}\Bigg(\sum_{m=1}^{k}(\psi(\tau_{m1}-\psi(\tau_{m1}+\tau_{m2})\Bigg)+(1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{m=1}^{k}v_m)]\right] \\ &+\sum_{k=1}^{K}\left[\log(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})})-(\tau_{k1}-1)\psi(\tau_{k1})-(\tau_{k2}-1)\psi(\tau_{k2})+(\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right]\end{aligned} \tag{40}$$

The expectation is bounded using the multinomial approximation

$$\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{m=1}^{k} v_m)] \geq \Big(\sum_{m=1}^{k} q_{km}\psi(\tau_{m2})\Big) + \Big(\sum_{m=1}^{k-1}(\sum_{j=m+1}^{k} q_{kj}\psi(\tau_{m1}))\Big) - \Big(\sum_{m=1}^{k}\sum_{j=m}^{k} q_{kj}\psi(\tau_{m1}+\tau_{m2})\Big) - \sum_{m=1}^{k} q_{km}\log(q_{km}) \tag{41}$$

Substitute into equation 40,

$$\begin{aligned}
\mathcal{L}_\tau =& \sum_{k=1}^{K}[\log\alpha + (\alpha-1)(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2}))] \\
&+ \sum_{k=1}^{K}\sum_{n=1}^{N}\Bigg[\nu_{nk}\Bigg(\sum_{m=1}^{k}(\psi(\tau_{m1} - \psi(\tau_{m1}+\tau_{m2})\Bigg) \\
&+ (1-\nu_{nk})\Bigg(\Big(\sum_{m=1}^{k} q_{km}\psi(\tau_{m2})\Big) + \Big(\sum_{m=1}^{k-1}(\sum_{j=m+1}^{k} q_{kj}\psi(\tau_{m1}))\Big) \\
&- \Big(\sum_{m=1}^{k}\sum_{j=m}^{k} q_{kj}\psi(\tau_{m1}+\tau_{m2})\Big) - \sum_{m=1}^{k} q_{km}\log(q_{km})\Bigg)\Bigg] \\
&+ \sum_{k=1}^{K}\Bigg[\log(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\Bigg]
\end{aligned} \tag{42}$$

Remove terms that don't contain τ_k. We need to break down the summations. It's tricky especially for the second term. Since k is a constant now, it cannot be used as an iterative index. We change the notation first, replace k with m, replace m with i, then do the cleaning

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{n=1}^{N}\Bigg[\nu_{nk}\Bigg(\sum_{m=1}^{k}(\psi(\tau_{m1}) - \psi(\tau_{m1}+\tau_{m2})\Bigg) + (1-\nu_{nk})\Bigg(\Big(\sum_{m=1}^{k} q_{km}\psi(\tau_{m2})\Big) + \Big(\sum_{m=1}^{k-1}(\sum_{j=m+1}^{k} q_{kj}\psi(\tau_{m1}))\Big) \\
&\qquad - \Big(\sum_{m=1}^{k}\sum_{j=m}^{k} q_{kj}\psi(\tau_{m1}+\tau_{m2})\Big) - \sum_{m=1}^{k} q_{km}\log(q_{km})\Bigg)\Bigg] \\
\xrightarrow{renotate}& \sum_{m=1}^{K}\sum_{n=1}^{N}\Bigg[\nu_{nm}\Bigg(\sum_{i=1}^{m}(\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\Bigg) + (1-\nu_{nm})\Bigg(\Big(\sum_{i=1}^{m} q_{mi}\psi(\tau_{i2})\Big) + \Big(\sum_{i=1}^{m-1}(\sum_{j=i+1}^{m} q_{mj}\psi(\tau_{i1}))\Big) \\
&\qquad - \Big(\sum_{i=1}^{m}\sum_{j=i}^{m} q_{mj}\psi(\tau_{i1}+\tau_{i2})\Big) - \sum_{i=1}^{m} q_{mi}\log(q_{mi})\Bigg)\Bigg]
\end{aligned} \tag{43}$$

To get all terms, the bottleneck is at the $\sum_{i=1}^{m-1}$, which means m should be $\geq k+1$ for it, and $\geq k$ for other terms

$$\xrightarrow{m\geq k \text{ or } k+1} \sum_{m=k}^{K}\sum_{n=1}^{N}\left[\nu_{nm}\left(\left(\psi(\tau_{k1})-\psi(\tau_{k1}+\tau_{k2})\right)\right]+(1-\nu_{nm})\left(\left(q_{mk}\psi(\tau_{k2})\right)\right.\right.$$
$$\left.\left.-\left(\sum_{j=k}^{m}q_{mj}\psi(\tau_{k1}+\tau_{k2})\right)-q_{mk}\log(q_{mk})\right)\right]+\sum_{m=k+1}^{K}\sum_{n=1}^{N}(1-\nu_{nm})\left(\sum_{j=k+1}^{m}q_{mj}\psi(\tau_{k1})\right) \tag{44}$$

Therefore,

$$\begin{aligned}\mathcal{L}_{\tau_k}=&\left[(\alpha-1)+\sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}+\sum_{m=k+1}^{K}\sum_{n=1}^{N}(1-\nu_{nm})\left(\sum_{j=k+1}^{m}q_{mj}\right)-(\tau_{k1}-1)\right](\psi(\tau_{k1})-\psi(\tau_{k1}+\tau_{k2}))\\&+\left[\sum_{m=k}^{K}\sum_{n=1}^{N}(1-\nu_{nm})q_{mk}-(\tau_{k2}-1)\right](\psi(\tau_{k2})-\psi(\tau_{k1}+\tau_{k2}))+c\end{aligned} \tag{45}$$

gives the update formula

$$\begin{aligned}\tau_{k1}=&\alpha+\sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}+\sum_{m=k+1}^{K}\sum_{n=1}^{N}(1-\nu_{nm})\left(\sum_{j=k+1}^{m}q_{mj}\right)\\\tau_{k2}=&\sum_{m=k}^{K}\sum_{n=1}^{N}(1-\nu_{nm})q_{mk}+1\end{aligned} \tag{46}$$

4. Updates the feature variance σ_A^2

$$\begin{aligned}\log q_{\boldsymbol{\rho}_A}(\sigma_A^2)=&\mathbb{E}_{\boldsymbol{A}}[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})]+c\\=&\mathbb{E}_{\boldsymbol{A}}[\log p(\sigma_A^2|\boldsymbol{\sigma}_{A0})+\sum_{k=1}^{K}\log p(\boldsymbol{A}_{k\cdot}|0,\sigma_A^2 I)]+c\\=&\log(\frac{\sigma_{Ar}^{\sigma_{As}}}{\Gamma(\sigma_{As})}{\sigma_A^2}^{\sigma_{As}-1}e^{-\sigma_{Ar}\sigma_A^2})+\sum_{k=1}^{K}(\frac{D}{2}\log(\frac{\sigma_A^2}{2\pi})-\frac{\sigma_A^2}{2}\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}])+c\\=&(\sigma_{As}-1)\log(\sigma_A^2)-\sigma_{Ar}\sigma_A^2+\frac{DK}{2}\log(\sigma_A^2)-\frac{\sigma_A^2}{2}\sum_{k=1}^{K}\left(tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T\right)+c\\=&(\sigma_{As}+\frac{DK}{2}-1)\log(\sigma_A^2)-\left(\sigma_{Ar}+\frac{1}{2}\sum_{k=1}^{K}\left(tr(\boldsymbol{\Phi}_k)+\overline{\phi}_k\overline{\phi}_k^T\right)\right)\sigma_A^2+c\end{aligned} \tag{47}$$

Therefore, equations for updating σ_A^2 is

$$\begin{aligned}\rho_{As} =&\sigma_{As} + \frac{DK}{2}\\ \rho_{Ar} =&\sigma_{Ar} + \frac{1}{2}\sum_{k=1}^{K}\left(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T\right)\end{aligned} \tag{48}$$

5-6. Similarly, σ_n^2 and α can be updated.

$$\begin{aligned}\log q_{\boldsymbol{\rho}_n}(\sigma_n^2) =&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})] + c\\ =&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\log p(\sigma_n^2|\boldsymbol{\sigma}_{n0}) + \sum_{n=1}^{N}\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)\right] + c\\ =&\log(\frac{\sigma_{nr}^{\sigma_{ns}}}{\Gamma(\sigma_{ns})}{\sigma_n^2}^{\sigma_{ns}-1}e^{-\sigma_{nr}\sigma_n^2}) + \sum_{n=1}^{N}\left(\frac{D}{2}\log(\frac{\sigma_n^2}{2\pi})\right.\\ &\left.-\frac{1}{2}\sigma_n^2\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\overline{\phi}_k\overline{\phi}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T)\right)\right)\\ =&(\sigma_{ns}-1)\log(\sigma_n^2) - \sigma_{nr}\sigma_n^2 + \frac{DN}{2}\log(\sigma_n^2)\\ &-\frac{1}{2}\sigma_n^2\sum_{n=1}^{N}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\overline{\phi}_k\overline{\phi}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T)\right)\end{aligned} \tag{49}$$

Therefore, equations for updating σ_n^2 is

$$\begin{aligned}\rho_{ns} =&\sigma_{ns} + \frac{DN}{2}\\ \rho_{nr} =&\sigma_{nr} + \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\overline{\phi}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\overline{\phi}_k\overline{\phi}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}(tr(\boldsymbol{\Phi}_k) + \overline{\phi}_k\overline{\phi}_k^T)\right)\end{aligned} \tag{50}$$

As to α

$$
\begin{aligned}
\log q_{\boldsymbol{\rho}_\alpha}(\alpha) =& \mathbb{E}_{\boldsymbol{v}}[\log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})] + c \\
=& \mathbb{E}_{\boldsymbol{v}}[\log p(\alpha|\boldsymbol{\alpha}_0) + \sum_{k=1}^{K} \log p(v_k|\alpha)]] + c \\
=& \log(\frac{\alpha_r^{\alpha_s}}{\Gamma(\alpha_s)} \alpha^{\alpha_s - 1} e^{-\alpha_r \alpha}) + \sum_{k=1}^{K} \Bigg(\log(\alpha) + (\alpha - 1)\mathbb{E}_{\boldsymbol{v}} \log(v_k) \Bigg) + c \\
=& (\alpha_s - 1)\log(\alpha) - \alpha_r \alpha + K \log(\alpha) + (\alpha - 1) \sum_{k=1}^{K} \big(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2}) \big) + c
\end{aligned}
\tag{51}
$$

Equations for updating α is

$$
\begin{aligned}
\rho_{\alpha s} =& \alpha_s + K \\
\rho_{\alpha r} =& \alpha_r - (\alpha - 1) \sum_{k=1}^{K} \big(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2}) \big)
\end{aligned}
\tag{52}
$$

Only concern is that $\rho_{\alpha r}$ may go below 0

References

[1] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[2] David M Blei and Michael I Jordan. "Variational methods for the Dirichlet process". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 12.

[3] Finale Doshi et al. "Variational inference for the Indian buffet process". In: *Artificial Intelligence and Statistics*. PMLR. 2009, pp. 137–144.

[4] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[5] Stephen Odaibo. "Tutorial: Deriving the standard variational autoencoder (VAE) loss function". In: *arXiv preprint arXiv:1907.08956* (2019).