Variational Inference for nCRP

1 Comparison with LDA

A generative model of LDA is

$$\begin{aligned}
\boldsymbol{\theta}_d &\sim \text{Dirichlet}(\boldsymbol{\alpha}), \boldsymbol{\theta}_d \in \mathbb{R}^{\mathrm{K}}_{>0} \\
\boldsymbol{\phi}_k &\sim \text{Dirichlet}(\boldsymbol{\beta}), \boldsymbol{\phi}_k \in \mathbb{R}^{\mathrm{H}}_{>0} \\
z_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\theta}_d) \\
w_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\phi}_{z_{d,n}})
\end{aligned}$$

A generative model of nCRP-LDA is

$$\begin{aligned}
\mathbf{c}_d &\sim \text{nCRP}(\gamma), \mathbf{c}_d \in \mathbb{R}^{\infty}_{>0} \\
\boldsymbol{\theta}_d &\sim \text{GEM}(m, \pi), \boldsymbol{\theta}_d \in \mathbb{R}^{\infty}_{>0} \\
\boldsymbol{\phi}_j &\sim \text{Dirichlet}(\boldsymbol{\beta}), \boldsymbol{\phi}_j \in \mathbb{R}^{\mathrm{H}}_{>0} \\
z_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\theta}_d) \\
w_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\phi}_{\mathbf{c}_d, z_{d,n}})
\end{aligned}$$

A tree-based stick-breaking construction version is

$$\begin{aligned}
v_j &\sim \text{Beta}(1, \gamma) \\
\mathbf{c}_d &\sim \textstyle\prod_{\mathbf{c}} \prod_{l=1}^{L} v_{c_1 c_2 \ldots c_l} \prod_{m=1}^{c_l - 1} (1 - v_{c_1 c_2 \ldots c_{l-1} m})^{I(\mathbf{c}_d = \mathbf{c})}, \mathbf{c}_d \in \mathbb{R}^{\mathrm{P}}_{>0} \\
\boldsymbol{\theta}_d &\sim \text{Dirichlet}(\boldsymbol{\alpha}), \boldsymbol{\theta}_d \in \mathbb{R}^{\mathrm{L}}_{>0} \\
\boldsymbol{\phi}_j &\sim \text{Dirichlet}(\boldsymbol{\beta}), \boldsymbol{\phi}_j \in \mathbb{R}^{\mathrm{H}}_{>0} \\
z_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\theta}_d) \\
w_{d,n} &\sim \text{Multinomial}(1, \boldsymbol{\phi}_{\mathbf{c}_d, z_{d,n}})
\end{aligned}$$

Note that the distribution of $\mathbf{c}_d$ is expressed as a categorical distribution.

$$x \sim \prod_{i=1}^{k} \pi_i^{I(x=i)}$$

The following example is helpful to understand how the truncated stick breaking process can be used for the categorical distribution. Let us consider the above categorical distribution $p(x|\boldsymbol{\pi})$ that $\boldsymbol{\pi}$ is constructed through the truncated stick-breaking process of the truncation level k, then we can derive

$$\begin{aligned}
\pi_i &= v_i \prod_{j=1}^{i-1} (1 - v_j), v_k = 1, v_i \sim \text{Beta}(1, \alpha) \text{ for } i = 1, 2, ..., k-1 \\
x &\sim \prod_{i=1}^{k} \Big(v_i \prod_{j=1}^{i-1} (1 - v_j) \Big)^{I(x=i)}
\end{aligned}$$

2 ELBO of nCRP

Define the evidence lower bound (ELBO) $\mathcal{L}(q)$ as below

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

where

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z})\text{ln}\Big\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\Big\}d\mathbf{Z} \\
\text{KL}(q||p) &= \int q(\mathbf{Z})\text{ln}\Big\{\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\Big\}d\mathbf{Z} \\
&= -\int q(\mathbf{Z})\text{ln}\Big\{\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\Big\}d\mathbf{Z} \\
\mathcal{L}(q) + \text{KL}(q||p) &= \int q(\mathbf{Z})\Big(\text{ln}p(\mathbf{X},\mathbf{Z}) - \text{ln}q(\mathbf{Z}) - \text{ln}p(\mathbf{Z}|\mathbf{X}) + \text{ln}q(\mathbf{Z})\Big)d\mathbf{Z} \\
&= \int q(\mathbf{Z})\Big(\text{ln}p(\mathbf{X},\mathbf{Z}) - \text{ln}p(\mathbf{X},\mathbf{Z}) + \text{ln}p(\mathbf{X})\Big)d\mathbf{Z}, \text{ln}p(\mathbf{Z}|\mathbf{X}) = \text{ln}p(\mathbf{X},\mathbf{Z}) - \text{ln}p(\mathbf{X}) \\
&= \text{ln}p(\mathbf{X})
\end{aligned}$$

$\text{ln}p(\mathbf{X})$ is constant, so maximisation of $\mathcal{L}(q)$ is equal to minimisation of $\text{KL}(q||p)$

$$\begin{aligned}
\text{ln}p(\mathbf{X}) &\geq \mathcal{L}(q) \\
&= \int q(\mathbf{Z})\text{ln}p(\mathbf{X},\mathbf{Z})d\mathbf{Z} - \int q(\mathbf{Z})\text{ln}q(\mathbf{Z})d\mathbf{Z} \\
&= \mathbb{E}_{q(\mathbf{Z})}[\text{ln}p(\mathbf{X},\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\text{ln}q(\mathbf{Z})]
\end{aligned}$$

ELBO of nCRP can be derived from the full joint distribution

$$\begin{aligned}
p(\mathbf{X}, \mathbf{Z}) &= p(\mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{c}, \boldsymbol{\phi}, \boldsymbol{\theta} | \gamma, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= p(\mathbf{v}|\gamma) p(\mathbf{c}|\mathbf{v}) p(\mathbf{w}|\mathbf{c}, \mathbf{z}, \boldsymbol{\phi}) p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\phi}|\boldsymbol{\beta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\
q(\mathbf{Z}) &= q(\mathbf{z}, \mathbf{v}, \mathbf{c}, \boldsymbol{\phi}, \boldsymbol{\theta}) \\
&= q(\mathbf{z}) q(\mathbf{v}) q(\mathbf{c}) q(\boldsymbol{\phi}) q(\boldsymbol{\theta})
\end{aligned}$$

Note the full joint distribution of LDA is

$$\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}) p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\phi}|\boldsymbol{\beta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\
q(\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) &= q(\mathbf{z}) q(\boldsymbol{\phi}) q(\boldsymbol{\theta})
\end{aligned}$$

$$
\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\mathbf{v})}[\ln p(\mathbf{v}|\gamma)] + \mathbb{E}_{q(\mathbf{v},\mathbf{c})}[\ln p(\mathbf{c}|\mathbf{v})] + \mathbb{E}_{q(\mathbf{c},\mathbf{z},\boldsymbol{\phi})}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] + \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[\ln p(\mathbf{z}|\boldsymbol{\theta})] \\
&+ \mathbb{E}_{q(\boldsymbol{\phi})}[\ln p(\boldsymbol{\phi}|\boldsymbol{\beta})] + \mathbb{E}_{q(\boldsymbol{\theta})}[\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})] - \mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{v})}[\ln q(\mathbf{v})] - \mathbb{E}_{q(\mathbf{c})}[\ln q(\mathbf{c})] \\
&- \mathbb{E}_{q(\boldsymbol{\phi})}[\ln q(\boldsymbol{\phi})] - \mathbb{E}_{q(\boldsymbol{\theta})}[\ln q(\boldsymbol{\theta})] \\
&= \sum_{j}(\gamma-1)\mathbb{E}_{v}[\ln(1-v_j)] \\
&+ \sum_{d=1}^{D}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\sum_{l=1}^{L}\Big(\mathbb{E}_{v}[\ln v_{c_1c_2\ldots c_l}] + \sum_{m=1}^{c_l-1}\mathbb{E}_{v}[\ln(1-v_{c_1c_2\ldots c_{l-1}m})]\Big) \\
&+ \sum_{d=1}^{D}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c_d}=\mathbf{c})]\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_{z}[I(z_{d,n}=l)]\Big(\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c},h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big)\Big) \\
&+ \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_{z}[I(z_{d,n}=l)]\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}] \\
&+ \sum_{j}\Big(\sum_{h=1}^{H}(\beta_h-1)\mathbb{E}_{\phi}[\ln\phi_{j,h}]\Big) \\
&+ \sum_{d=1}^{D}\Big(\sum_{l=1}^{L}(\alpha_l-1)\mathbb{E}_{\theta}[\ln\theta_{d,l}]\Big) \\
&- \sum_{j}\Big((\xi_a^{v_j}-1)\mathbb{E}_{v}[\ln v_j] + (\xi_b^{v_j}-1)\mathbb{E}_{v}[\ln(1-v_j)] + \ln\Gamma(\xi_a^{v_j}+\xi_b^{v_j}) - \ln\Gamma(\xi_a^{v_j}) - \ln\Gamma(\xi_b^{v_j})\Big) \\
&- \sum_{d=1}^{D}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c_d}=\mathbf{c})]\ln\xi_{\mathbf{c}}^{\mathbf{c}_d} \\
&- \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_{z}[I(z_{d,n}=l)]\ln\xi_{n,l}^{\mathbf{z}_d} \\
&- \sum_{j}\Big(\sum_{h=1}^{H}(\xi_h^{\boldsymbol{\phi}_j}-1)\mathbb{E}_{\phi}[\ln\phi_{j,h}] + \ln\Gamma(\sum_{h}\xi_h^{\boldsymbol{\phi}_j}) - \sum_{h}\ln\Gamma(\xi_h^{\boldsymbol{\phi}_j})\Big) \\
&- \sum_{d=1}^{D}\Big(\sum_{l=1}^{L}(\xi_l^{\boldsymbol{\theta}_d}-1)\mathbb{E}_{\theta}[\ln\theta_{d,l}] + \ln\Gamma(\sum_{l}\xi_l^{\boldsymbol{\theta}_d}) - \sum_{l}\ln\Gamma(\xi_l^{\boldsymbol{\theta}_d})\Big) + \text{const.}
\end{aligned}
$$

3 Log-likelihood of nCRP

Here is the equation of test-set log-likelihood introduced in ”Collapsed Variational Inference for HDP” (NIPS, 2008)

$$\begin{aligned} p(\mathbf{w}^{\text{test}}) &= \prod_{ij} \sum_{k} \bar{\theta}_{jk} \bar{\phi}_{k w_{ij}^{\text{test}}} \\ \bar{\theta}_{jk} &= \frac{\alpha \pi_k + \mathbb{E}_q[n_{jk.}]}{\alpha + \mathbb{E}_q[n_{j..}]} \\ \bar{\phi}_{kv} &= \frac{\beta \tau_v + \mathbb{E}_q[n_{.kv}]}{\beta + \mathbb{E}_q[n_{.k.}]} \end{aligned}$$

Here is the LDA version.

$$\begin{aligned} p(w_{ij}^{\text{test}} = v | \mathbf{w}^{\text{train}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\approx \sum_{k} \int \phi_{kv} q(\boldsymbol{\phi}_k) d\boldsymbol{\phi}_k \int \theta_{jk} q(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \\ &= \sum_{k} \frac{\mathbb{E}_{q(\mathbf{z}^{\text{train}})}[n_{kv}] + \beta_v}{\sum_{v'} (\mathbb{E}_{q(\mathbf{z}^{\text{train}})}[n_{kv'}] + \beta_{v'})} \frac{\mathbb{E}_{q(\mathbf{z}^{\text{train}})}[n_{jk}] + \alpha_k}{\sum_{k'} (\mathbb{E}_{q(\mathbf{z}^{\text{train}})}[n_{jk'}] + \alpha_{k'})} \end{aligned}$$

The equations above are based on the expectation of the Dirichlet distribution

$$\mathbb{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_i}$$

Note that the joint distribution of data and latent variables of LDA is

$$p(\mathbf{w}, \mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{ij} \prod_{k} (\theta_{jk} \phi_{k w_{ij}})^{I(z_{ij}=k)}$$

and the marginal distribution is

$$\begin{aligned} p(\mathbf{w} | \boldsymbol{\theta}, \boldsymbol{\phi}) &= \sum_{z} p(\mathbf{w}, \mathbf{z} | \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \prod_{ij} \Big\{ \sum_{z} \prod_{k} (\theta_{jk} \phi_{k w_{ij}})^{I(z_{ij}=k)} \Big\} \\ &= \prod_{ij} (\sum_{k} \theta_{jk} \phi_{k w_{ij}}) \end{aligned}$$

The basic formula of the predictive distribution of LDA is

$$p(w^* | \mathbf{w}^{\text{train}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(w^* | \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{w}^{\text{train}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} d\boldsymbol{\phi}$$

So the nCRP version of the joint distribution is

$$p(\mathbf{w}, \mathbf{c}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{d,n} \prod_{\mathbf{c}} \pi_{d,\mathbf{c}} \{\prod_{l} (\theta_{d,l} \phi_{\mathbf{c},l,w_{d,n}})^{I(z_{d,n}=l)}\}^{I(\mathbf{c}_d=\mathbf{c})}$$

and the marginal distribution is

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\theta}, \boldsymbol{\phi}) &= \prod_{d,n} \sum_{\mathbf{c}} \sum_{z} \prod_{\mathbf{c}} \pi_{d,\mathbf{c}} \{\prod_{l} (\theta_{d,l} \phi_{\mathbf{c},l,w_{d,n}})^{I(z_{d,n}=l)}\}^{I(\mathbf{c}_d=\mathbf{c})} \\ &= \prod_{d,n} (\sum_{\mathbf{c}} \pi_{d,\mathbf{c}} \sum_{l} \theta_{d,l} \phi_{\mathbf{c},l,w_{d,n}}) \end{aligned}$$

Log-likelihood can be calculated by replacing parameters to the expectations of those and training data to test data. For example, "Variational Inference for the Nested Chinese Restaurant Process" (NIPS, 2009) uses the equation below to calculate likelihood.

$$p(\mathbf{w}_{1:D}^{\text{test}}) = \prod_{d=1}^{D} \sum_{\mathbf{c}} q(\mathbf{c}_d = \mathbf{c}) \prod_{n} \sum_{l} \bar{\theta}_{d,l} \bar{\phi}_{\mathbf{c},l,w_{d,n}}$$

For the reference, basic mixture models can be formalised as below.

$$\begin{aligned} p(\mathbf{x}) &= \sum_{k=1}^{K} \theta_k p(\mathbf{x}|\boldsymbol{\phi}_k) \\ p(\mathbf{z}) &= \prod_{k=1}^{K} \theta_k^{z_k} \\ p(\mathbf{x}|\mathbf{z}) &= \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\phi}_k)^{z_k} \\ p(\mathbf{x}, \mathbf{z}) &= p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \\ &= \prod_{k=1}^{K} \Big(\theta_k p(\mathbf{x}|\boldsymbol{\phi}_k)\Big)^{z_k} \\ p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \end{aligned}$$

4 Variational inference of $q(v_j)$

$\prod\limits_{j \in \mathcal{M}_T} q(v_j)$, it has the same distribution formula with $p(v_j|\gamma)$

$$\begin{aligned}
\ln q^*(\mathbf{v}) &= \mathbb{E}_{\mathbf{z},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}}[\ln p(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}|\gamma,\boldsymbol{\alpha},\boldsymbol{\beta})] + \text{const.} \\
&= \ln p(\mathbf{v}|\gamma) + \mathbb{E}_{\mathbf{c}}[\ln p(\mathbf{c}|\mathbf{v})] + \text{const.}
\end{aligned}$$

According to $v_j \sim \text{Beta}(1,\gamma)$, the first term is

$$\ln p(\mathbf{v}|\gamma) = \sum_j \Big(\ln(1-v_j)^{\gamma-1} - \ln \mathrm{B}(1,\gamma)\Big)$$

where $\mathrm{B}(1,\gamma)$ is beta function. The second term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{c}}[\ln p(\mathbf{c}|\mathbf{v})] &= \sum_{d=1}^{D}\sum_{\mathbf{c}} \mathbb{E}_{\mathbf{c}}\Big[\ln \prod_{l=1}^{L} v_{c_1 c_2 \ldots c_l} \prod_{m=1}^{c_l-1}(1-v_{c_1 c_2 \ldots c_{l-1} m})^{I(\mathbf{c}_d=\mathbf{c})}\Big] \\
&= \sum_{d=1}^{D}\sum_{\mathbf{c}} \mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})] \sum_{l=1}^{L}\Big(\ln v_{c_1 c_2 \ldots c_l} + \sum_{m=1}^{c_l-1}\ln(1-v_{c_1 c_2 \ldots c_{l-1} m})\Big)
\end{aligned}$$

Let the path reaching to the node of v_j be $\bar{\mathbf{c}} = [\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_{l_0}]$ where $l_0 \leq L$. Index of v and elements of $\mathbf{c}$ has one-to-one relation so $j \iff \bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}$. Based on factorization property of mean-field theory, particular $q(v_j)$ is

$$\begin{aligned}
q(v_j) &= q(v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}}) \\
&\propto \exp\Big\{\ln(1-v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}})^{\gamma-1} + \sum_{d=1}^{D}\sum_{\mathbf{c}} \mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})] \sum_{l=1}^{L}\Big(\ln v_{c_1 c_2 \ldots c_l} + \sum_{m=1}^{c_l-1}\ln(1-v_{c_1 c_2 \ldots c_{l-1} m})\Big)\Big\} \\
&= \exp\Big\{\sum_{d=1}^{D} \mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\bar{\mathbf{c}})]\ln v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}} \\
&+ \Big(\gamma - 1 + \sum_{d=1}^{D}\sum_{\mathbf{c}\in[\bar{c}_1,\bar{c}_2,\ldots,\bar{c}_{l_0-1},m:m>\bar{c}_{l_0}]} \mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\ln(1-v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}})\Big)\Big\} \\
&= v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}}^{\sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\bar{\mathbf{c}})]}(1-v_{\bar{c}_1\bar{c}_2\ldots\bar{c}_{l_0}})^{\gamma+\sum_{d=1}^{D}\sum_{\mathbf{c}\in[\bar{c}_1,\bar{c}_2,\ldots,\bar{c}_{l_0-1},m:m>\bar{c}_{l_0}]}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]-1} \\
&\Rightarrow \text{Beta}\Big(1+\textstyle\sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\bar{\mathbf{c}})], \gamma + \sum_{d=1}^{D}\sum_{\mathbf{c}\in[\bar{c}_1,\bar{c}_2,\ldots,\bar{c}_{l_0-1},m:m>\bar{c}_{l_0}]}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\Big)
\end{aligned}$$

5 Variational inference of $q(\mathbf{c}_d)$

$\prod_{d=1}^{D} q(\mathbf{c}_d)$, it has the same distribution formula with $p(\mathbf{c}_d|\mathbf{v})$

$$\begin{aligned}
\ln q^*(\mathbf{c}) &= \mathbb{E}_{\mathbf{z},\mathbf{v},\boldsymbol{\phi},\boldsymbol{\theta}}[\ln p(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}|\gamma,\boldsymbol{\alpha},\boldsymbol{\beta})] + \text{const.} \\
&= \mathbb{E}_{\mathbf{v}}[\ln p(\mathbf{c}|\mathbf{v})] + \mathbb{E}_{\mathbf{z},\boldsymbol{\phi}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] + \text{const.}
\end{aligned}$$

The first term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{v}}[\ln p(\mathbf{c}|\mathbf{v})] &= \sum_{d=1}^{D}\sum_{\mathbf{c}} \mathbb{E}_{\mathbf{v}}\Big[\ln \prod_{l=1}^{L} v_{c_1c_2\dots c_l} \prod_{m=1}^{c_l-1}(1-v_{c_1c_2\dots c_{l-1}m})^{I(\mathbf{c}_d=\mathbf{c})}\Big] \\
&= \sum_{d=1}^{D}\sum_{\mathbf{c}} I(\mathbf{c}_d=\mathbf{c}) \sum_{l=1}^{L}\Big(\mathbb{E}_v[\ln v_{c_1c_2\dots c_l}] + \sum_{m=1}^{c_l-1}\mathbb{E}_v[\ln(1-v_{c_1c_2\dots c_{l-1}m})]\Big)
\end{aligned}$$

The second term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{z},\boldsymbol{\phi}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] &= \mathbb{E}_{\mathbf{z},\boldsymbol{\phi}}\Big[\ln \prod_{d=1}^{D}\prod_{\mathbf{c}}\Big(\prod_{n=1}^{N_d}\prod_{l=1}^{L}\Big(\prod_{h=1}^{H}\phi_{c_1c_2\dots c_l,h}^{w_{d,n,h}}\Big)^{z_{d,n,l}}\Big)^{I(\mathbf{c}_d=\mathbf{c})}\Big] \\
&= \sum_{d=1}^{D}\sum_{\mathbf{c}} I(\mathbf{c}_{\mathbf{d}}=\mathbf{c})\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_z[I(z_{d,n}=l)]\Big(\sum_{h=1}^{H} w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c},h}}[\ln\phi_{c_1c_2\dots c_l,h}]\Big)\Big)
\end{aligned}$$

Note that the formula below is full joint word distribution used in LDA,

$$p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi}) = \prod_{d=1}^{D}\Big(\prod_{n=1}^{N_d}\prod_{k=1}^{K}\Big(\prod_{h=1}^{H}\phi_{k,h}^{w_{d,n,h}}\Big)^{z_{d,n,k}}\Big)$$

Based on factorization property of mean-field theory, a particular $q(\mathbf{c}_d)$ is

$$\begin{aligned}
q(\mathbf{c}_d) \quad & \propto \quad \exp\Big(\sum_{\mathbf{c}} I(\mathbf{c_d}=\mathbf{c}) \sum_{l=1}^{L} \Big(\mathbb{E}_v[\ln v_{c_1 c_2 \ldots c_l}] + \sum_{m=1}^{c_l-1} \mathbb{E}_v[\ln(1-v_{c_1 c_2 \ldots c_{(l-1)} m})]\Big) \\
& + \quad \sum_{\mathbf{c}} I(\mathbf{c_d}=\mathbf{c}) \sum_{n=1}^{N_d} \sum_{l=1}^{L} \mathbb{E}_z[I(z_{d,n}=l)]\Big(\sum_{h=1}^{H} w_{d,n,h} \mathbb{E}_{\phi_{\mathbf{c},h}}[\ln\phi_{c_1 c_2 \ldots c_l, h}]\Big)\Big) \\
& = \quad \prod_{\mathbf{c}} \exp\Big(\sum_{l=1}^{L} \Big(\mathbb{E}_v[\ln v_{c_1 c_2 \ldots c_l}] + \sum_{m=1}^{c_l-1} \mathbb{E}_v[\ln(1-v_{c_1 c_2 \ldots c_{(l-1)} m})]\Big) \\
& + \quad \sum_{n=1}^{N_d} \sum_{l=1}^{L} \mathbb{E}_z[I(z_{d,n}=l)]\Big(\sum_{h=1}^{H} w_{d,n,h} \mathbb{E}_{\phi_{\mathbf{c},h}}[\ln\phi_{c_1 c_2 \ldots c_l, h}]\Big)\Big)^{I(\mathbf{c_d}=\mathbf{c})} \\
& = \quad \prod_{\mathbf{c}} \pi_{d\mathbf{c}}^{I(\mathbf{c_d}=\mathbf{c})} \\
& \Rightarrow \quad \text{Multinomial}(1, \boldsymbol{\pi}_d)
\end{aligned}$$

6 Variational inference of $q(\boldsymbol{\phi}_j)$

$\prod_{j \in \mathcal{M}_T} q(\boldsymbol{\phi}_j)$, it has the same distribution formula with $p(\boldsymbol{\phi}_j|\boldsymbol{\beta})$

$$\begin{aligned}
\ln q^*(\boldsymbol{\phi}) &= \mathbb{E}_{\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\theta}}[\ln p(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}|\gamma,\boldsymbol{\alpha},\boldsymbol{\beta})] + \text{const.} \\
&= \ln p(\boldsymbol{\phi}|\boldsymbol{\beta}) + \mathbb{E}_{\mathbf{z},\mathbf{c}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] + \text{const.}
\end{aligned}$$

According to $\boldsymbol{\phi}_j \sim \text{Dirichlet}(\boldsymbol{\beta})$, the first term is

$$\begin{aligned}
\ln p(\boldsymbol{\phi}|\boldsymbol{\beta}) &= \ln\Big(\prod_j \frac{\Gamma(\sum \beta_h)}{\prod \Gamma(\beta_h)} \prod_{h=1}^{H} \phi_{j,h}^{\beta_h - 1}\Big) \\
&= \sum_j \Big(\sum_{h=1}^{H} (\beta_h - 1)\ln\phi_{j,h}\Big) + \text{const.}
\end{aligned}$$

The second term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{z},\mathbf{c}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] &= \mathbb{E}_{\mathbf{z},\mathbf{c}}\Big[\ln \prod_{d=1}^{D}\prod_{\mathbf{c}}\Big(\prod_{n=1}^{N_d}\prod_{l=1}^{L}\Big(\prod_{h=1}^{H}\phi_{c_1c_2\ldots c_l,h}^{w_{d,n,h}}\Big)^{z_{d,n,l}}\Big)^{I(\mathbf{c}_d=\mathbf{c})}\Big] \\
&= \sum_{d=1}^{D}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_{\mathbf{d}} = \mathbf{c})]\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_z[I(z_{d,n} = l)]\Big(\sum_{h=1}^{H} w_{d,n,h}\ln\phi_{c_1c_2\ldots c_l,h}\Big)\Big)
\end{aligned}$$

Based on factorization property of mean-field theory, a particular $q(\boldsymbol{\phi}_j)$ is

$$\begin{aligned}
q(\boldsymbol{\phi}_j) &\propto \exp\Big(\sum_{h=1}^{H}(\beta_h - 1)\ln\phi_{j,h} + \sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d = \mathbf{c}_j)]\sum_{n=1}^{N_d}\mathbb{E}_z[I(z_{d,n} = l_0)]\Big(\sum_{h=1}^{H} w_{d,n,h}\ln\phi_{j,h}\Big)\Big) \\
&= \exp\Big(\sum_{h=1}^{H}\ln\phi_{j,h}^{\beta_h - 1} + \sum_{h=1}^{H}\ln\phi_{j,h}^{\sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c}_j)]\sum_{n=1}^{N_d}\mathbb{E}_z[I(z_{d,n}=l_0)]w_{d,n,h}}\Big) \\
&= \prod_{h=1}^{H}\phi_{j,h}^{\beta_h + \sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c}_j)]\sum_{n=1}^{N_d}\mathbb{E}_z[I(z_{d,n}=l_0)]w_{d,n,h} - 1} \\
&\Rightarrow \text{Dirichlet}(\boldsymbol{\beta} + \sum_{d=1}^{D}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d = \mathbf{c}_j)]\sum_{n=1}^{N_d}\mathbb{E}_z[I(z_{d,n} = l_0)]w_{d,n,h})
\end{aligned}$$

7 Variational inference of $q(\mathbf{z}_d)$

$\prod_{d=1}^{D} q(\mathbf{z}_d)$, it has the same distribution formula with $p(\mathbf{z}_d|\boldsymbol{\theta})$

$$\begin{aligned}
\ln q^*(\mathbf{z}) &= \mathbb{E}_{\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}}[\ln p(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}|\gamma,\boldsymbol{\alpha},\boldsymbol{\beta})] + \text{const.} \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{z}|\boldsymbol{\theta})] + \mathbb{E}_{\mathbf{c},\boldsymbol{\phi}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] + \text{const.}
\end{aligned}$$

The first term is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{z}|\boldsymbol{\theta})] &= \mathbb{E}_{\boldsymbol{\theta}}\Big[\ln \prod_{d=1}^{D}\prod_{n=1}^{N_d}\Big(\prod_{l=1}^{L}\theta_{d,l}^{z_{d,n,l}}\Big)\Big] \\
&= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{l=1}^{L} I(z_{d,n}=l)\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}]
\end{aligned}$$

The second term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{c},\boldsymbol{\phi}}[\ln p(\mathbf{w}|\mathbf{c},\mathbf{z},\boldsymbol{\phi})] &= \mathbb{E}_{\mathbf{c},\boldsymbol{\phi}}\Big[\ln \prod_{d=1}^{D}\prod_{\mathbf{c}}\Big(\prod_{n=1}^{N_d}\prod_{l=1}^{L}\Big(\prod_{h=1}^{H}\phi_{c_1c_2\ldots c_l,h}^{w_{d,n,h}}\Big)^{z_{d,n,l}}\Big)^{I(\mathbf{c}_d=\mathbf{c})}\Big] \\
&= \sum_{d=1}^{D}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c_d}=\mathbf{c})]\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}I(z_{d,n}=l)\Big(\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c}h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big)\Big)
\end{aligned}$$

Based on factorization property of mean-field theory, a particular $q(\mathbf{z}_d)$ is

$$\begin{aligned}
q(\mathbf{z}_d) &\propto \exp\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}z_{d,n,l}\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}] + \sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\sum_{n=1}^{N_d}\sum_{l=1}^{L}z_{d,n,l}\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c}h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big) \\
&= \exp\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}z_{d,n,l}\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}] + \sum_{n=1}^{N_d}\sum_{l=1}^{L}z_{d,n,l}\sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c}h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big) \\
&= \exp\Big(\sum_{n=1}^{N_d}\sum_{l=1}^{L}z_{d,n,l}\Big(\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}] + \sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c}h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big)\Big) \\
&= \prod_{n=1}^{N_d}\prod_{l=1}^{L}\exp\Big(\mathbb{E}_{\theta_{d,l}}[\ln\theta_{d,l}] + \sum_{\mathbf{c}}\mathbb{E}_{\mathbf{c}}[I(\mathbf{c}_d=\mathbf{c})]\sum_{h=1}^{H}w_{d,n,h}\mathbb{E}_{\phi_{\mathbf{c}h}}[\ln\phi_{c_1c_2\ldots c_l,h}]\Big)^{z_{d,n,l}}
\end{aligned}$$

8 Variational inference of $q(\boldsymbol{\theta}_d)$

$\prod_{d=1}^{D} q(\boldsymbol{\theta}_d)$, it has the same distribution formula with $p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})$

$$\begin{aligned}
\ln q^*(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi}}[\ln p(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{c},\boldsymbol{\phi},\boldsymbol{\theta}|\gamma,\boldsymbol{\alpha},\boldsymbol{\beta})] + \text{const.} \\
&= \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{z}|\boldsymbol{\theta})] + \text{const.}
\end{aligned}$$

According to $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$, the first term is

$$\begin{aligned}
\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha}) &= \ln\Big(\prod_{d=1}^{D} \frac{\Gamma(\sum \alpha_l)}{\prod \Gamma(\alpha_l)} \prod_{l=1}^{L} \theta_{d,l}^{\alpha_l - 1}\Big) \\
&= \sum_{d=1}^{D}\Big(\sum_{l=1}^{L}(\alpha_l - 1)\ln\theta_{d,l}\Big) + \text{const.}
\end{aligned}$$

The second term is

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{z}|\boldsymbol{\theta})] &= \mathbb{E}_{\mathbf{z}}\Big[\ln \prod_{d=1}^{D}\prod_{n=1}^{N_d}\Big(\prod_{l=1}^{L}\theta_{d,l}^{z_{d,n,l}}\Big)\Big] \\
&= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_{z_{d,n}}[I(z_{d,n} = l)]\ln\theta_{d,l}
\end{aligned}$$

Based on factorization property of mean-field theory, a particular $q(\boldsymbol{\theta}_d)$ is

$$\begin{aligned}
q(\boldsymbol{\theta}_d) &\propto \exp\Big(\sum_{l=1}^{L}(\alpha_l - 1)\ln\theta_{d,l} + \sum_{n=1}^{N_d}\sum_{l=1}^{L}\mathbb{E}_{z_{d,n}}[I(z_{d,n} = l)]\ln\theta_{d,l}\Big) \\
&= \exp\Big(\sum_{l=1}^{L}\Big(\alpha_l - 1 + \sum_{n=1}^{N_d}\mathbb{E}_{z_{d,n}}[I(z_{d,n} = l)]\Big)\ln\theta_{d,l}\Big) \\
&= \prod_{l=1}^{L}\theta_{d,l}^{\alpha_l - 1 + \sum_{n=1}^{N_d}\mathbb{E}_{z_{d,n}}[I(z_{d,n}=l)]}
\end{aligned}$$