Evidence Lower Bound (ELBO)

Minimizing the Kullback-Leibler divergence between the true posterior $p(\theta|\{w_{nd}\})$ and the variational distribution $q(\theta)$ corresponds to maximizing the lower bound on the evidence $E_q[\log p(\theta|\{w_{nd}\})] - H(q)$ where $H(q)$ is the entropy. We expand the terms in $E_q[\log p(\theta|\{w_{nd}\})] - H(q)$ into

$$L(t, y, \beta, \pi, z) = E_q[\log p(\pi|\alpha)]$$

$$+ \sum N E_q[\log p(z_n|\pi)]$$

$$+ \sum N \sum D E_q[\log p(w_{nd}|i_{ng}, f, l_d)]$$

$$+ \sum N \sum G E_q[\log p(i_{ng}|\beta, z_n)]$$

$$+ \sum G \sum K E_q[\log p(\beta_{gk}|t_{gk}, y_g)]$$

$$+ \sum G \sum K E_q[\log p(t_{gk}|\gamma_g)]$$

$$+ \sum G E_q[\log p(\gamma_g|\sigma)]$$

$$+ \sum G E_q[\log p(y_g|\rho)]$$

$$+ \sum G E_q[\log p(f_g|\iota)]$$

$$+ \sum D E_q[\log p(l_d|\kappa)]$$

$$- H(q)$$

and derive the value of each of the terms below.

**1st term** $\pi$ are the cluster popularities. $\Psi$ is the digamma function.

$$E_q[\log p(\pi|\alpha)] = E_q[\log(\pi^\alpha) - \alpha + \sum K \log \Gamma(\alpha_i) + \log \Gamma(\sum \alpha_i)]$$

$$= \sum (\alpha_i - 1) E_q[\log(\pi_k)] - \sum K \log \Gamma(\alpha_i) + \log \Gamma(\sum \alpha_i)$$

$$= \sum (\alpha_i - 1)(\Psi(\pi_k) - \Psi(\sum \tau_j)) - \sum K \log \Gamma(\alpha_i) + \log \Gamma(\sum \alpha_i)$$
2nd term \( z_n \) encodes the assignments of observations to clusters.

\[
E_q[\log p(z_n|\pi)] = E_q[\sum_k^K \log(\pi_k z_n=k)] \\
= \sum_k^K E[1(z_n=k)] E[\log \pi_k] \\
= \sum_k^K \nu_{nk} \log(\tau_k) - \nu(\sum_j \tau_j)
\]

3rd term Likelihood.

\[
E_q[\log p(w_{nd}|i_{ng}, f, l_d)] = E_q[\log(\prod_g 1(l_d=g)(1(f_g=1)1(i_{ng}=1)1(w_{nd}=1)(0)1(f_g=1)1(i_{ng}=1)1(w_{nd}=0)) \\
\frac{1}{2}1(f_g=0)1(i_{ng}=1)1(w_{nd}=1)(0)1(f_g=1)1(i_{ng}=0)1(w_{nd}=0)] \\
= \sum_g E_q[1(l_d=g)1(f_g=1)1(i_{ng}=1)1(w_{nd}=1) \log(1) \\
+1(l_d=g)1(f_g=1)1(i_{ng}=1)1(w_{nd}=0) \log(0) \\
+1(l_d=g)1(f_g=0)1(i_{ng}=1) \log(\frac{1}{2}) \\
+1(l_d=g)1(i_{ng}=0)1(w_{nd}=1) \log(0) \\
+1(l_d=g)1(i_{ng}=0)1(w_{nd}=0) \log(1)]
\]

4th term \( i_{ng} \) indicates the presence of group \( g \) in observation \( n \).

\[
E_q[\log p(i_{ng}|\beta, z_n)] = E_q[\log(\beta_{i_{ng}}(1-\beta_{i_{ng}}))^{1-i_{ng}}] \\
= i_{ng} E_q[\log \beta_{i_{ng}}] + (1-i_{ng})E_q[\log(1-\beta_{i_{ng}})] \\
= i_{ng} \sum_k^K 1(z_n=k) \log \beta_{gk} + (1-i_{ng})\sum_k^K 1(z_n=k) \log(1-\beta_{gk}) \\
= i_{ng} \sum_g E_q[\nu_{nk} \log \beta_{gk}] + (1-i_{ng})\sum_g E_q[\nu_{nk} \log(1-\beta_{gk})] \\
= o_{ng} \sum_g \nu_{nk}(\Psi(\phi_{gk1}) - \Psi(\phi_{gk1} + \phi_{gk2})) + (1-o_{ng})\sum_g \nu_{nk}(\Psi(\phi_{gk2}) - \Psi(\phi_{gk1} + \phi_{gk2}))
\]
5th term: \( \beta_{gk} \) is the probability of each group for each cluster.

\[
E_q[\log p(\beta_{gk}|t_{gk}, y_g)] = E_q[\log \left( \text{Beta}(\alpha_t, \beta_t)^{t_{gk}} \text{Beta}(\alpha_b, \beta_b)^{(1-t_{gk})} \right)^{y_g} \text{Beta}(\alpha_u, \beta_u)^{1-y_g}]
\]

\[
= E_q[y_{g|t_{gk}} \log \text{Beta}(\alpha_t, \beta_t) + y_{g}(1-t_{gk}) \log \text{Beta}(\alpha_u, \beta_u) + (1-y_g) \log \text{Beta}(\alpha_b, \beta_b)]
\]

\[
= E_q[y_{g|t_{gk}} \{ (\alpha_t - 1) \log(\beta_{gk}) + (\beta_t - 1) \log(1 - \beta_{gk}) - \log \text{BetaFun}(\alpha_t, \beta_t) \} + y_{g}(1 - t_{gk}) \{ (\alpha_b - 1) \log(\beta_{gk}) + (\beta_t - 1) \log(1 - \beta_{gk}) - \log \text{BetaFun}(\alpha_b, \beta_b) \} + (1 - y_g) \{ (\alpha_u - 1) \log(\beta_{gk}) + (\beta_t - 1) \log(1 - \beta_{gk}) - \log \text{BetaFun}(\alpha_u, \beta_u) \}]
\]

6th term: \( t_g \) is the auxiliary variable indicating from which mode an important \( \beta_{gk} \) was drawn.

\[
E_q[\log p(t_{g|\gamma_g})] = E_q[\log(\gamma_g^{t_{g|\gamma_g}}(1 - \gamma_g)^{(1-t_{g|\gamma_g}))}]
\]

\[
= E_q[t_{g|\gamma_g}] E[\log \gamma_g] + (1 - E[t_{g|\gamma_g}]) E[\log(1 - \gamma_g)]
\]

\[
= \lambda_{gk}(\Psi(\ell_{g1}) - \Psi(\ell_{g1} + \ell_{g2})) + (1 - \lambda_{gk})(\Psi(\ell_{g2}) - \Psi(\ell_{g1} + \ell_{g2}))
\]

7th term: \( \gamma_g \) is the proportion of each mode in multimodal distribution.

\[
E_q[\log p(\gamma_g|\sigma)] = E_q[\log \left( \gamma_g^{\sigma_g-1}(1 - \gamma_g)^{\sigma_g-2} \right)^{\text{BetaFun}(\sigma_1, \sigma_2)}]
\]

\[
= E_q[\log \left( \gamma_g^{\sigma_g-1}(1 - \gamma_g)^{\sigma_g-2} \right)^{\text{BetaFun}(\sigma_1, \sigma_2)}] + \text{Const}
\]

\[
= \{(\sigma_1 - 1) E[\log \gamma_g] + (\sigma_2 - 1) E[\log(1 - \gamma_g)] - \log \text{BetaFun}(\sigma_1, \sigma_2)\}
\]

\[
= \{(\sigma_1 - 1)\Psi(\ell_{g1}) - \Psi(\ell_{g1} + \ell_{g2}) + (\sigma_2 - 1)\Psi(\ell_{g2}) - \Psi(\ell_{g1} + \ell_{g2})\}
\]

\[
- \log \text{BetaFun}(\sigma_1, \sigma_2)
\]

8th term: \( y_g \) indicates whether a dimension is important.

\[
E_q[\log p(y_g)] = E_q[\log \rho^{y_g}(1 - \rho)^{1-y_g}]
\]

\[
= E_q[y_g] \log \rho + (1 - E[y_g]) \log(1 - \rho)
\]

\[
= \eta_g \log \rho + (1 - \eta_g) \log(1 - \rho)
\]

9th term: \( f_g \) indicates which formula is associated with group \( g \).

\[
E_q[\log p(f_{g|\iota})] = E_q[\log \left( \lambda_{g}^{1-f_{g}} \right)^{1-f_{g}}]
\]

\[
= E_q[f_g] \log \iota + (1 - E[f_g]) \log(1 - \iota)
\]

\[
= \eta_g \log \iota + (1 - \eta_g) \log(1 - \iota)
\]
10th term: \( l_d \) indicates to which group dimension \( d \) belongs.

\[
E_q[\log p(l_d|\kappa)] = E_q \left[ \sum_g \log (\kappa^1_{l_d=g}) \right] \\
= \sum G \left[ E[\mathbf{1}(l_d = g)] E[\log \kappa_g] \right] \\
= \sum G \left[ c_{dg} \ast (\Psi(h_g) - \Psi(\sum_g h_g)) \right]
\]
11th term: Entropy term.

\[ H(q) = E_q \left[ \log(q(t)q(y)q(\pi)q(\beta)q(z)q(\gamma)q(i)q(l)q(\theta)) \right] \]

\[ = E_q \left[ \sum_g^{G} \sum_k^K \log q(t_{gk}) + \sum_g^K \log q(y_g) + \log q(\pi) \right. \]

\[ + \sum_g^K \sum_k^K \log q(\beta_{gk}) + \sum_n^{N} \sum_k^K \log q(z_{nk}) + \sum_g^K \log q(\gamma_g) + \sum_n^{N} \sum_g^K \log q(i_{ng}) + \sum_g^K \log q(\theta_g) + \log p(\kappa) \right] \]

\[ = \sum_g^K \sum_k^K E_q[t_{gk}] \log \lambda_{gk} + (1 - E_q[t_{gk}]) \log(1 - \lambda_{gk}) \]

\[ + \sum_g^K E_q[y_g] \log(\eta_g) + (1 - E_q[y_g]) \log(1 - \eta_g) \]

\[ + \sum_k^K (\tau_k - 1) E_q[\log(\pi_k)] - \sum_k^K \log \Gamma(\tau_k) + \log \Gamma(\sum_k^K \tau_k) \]

\[ + \sum_g^K \sum_k^K ((\phi_{gk1} - 1) E[\log(\beta_{gk})] + (\phi_{gk2} - 1) E[\log(1 - \beta_{gk})]) - \log \text{BetaFun}(\phi_{gk1}, \phi_{gk2}) \]

\[ + \sum_n^K \sum_k^K \ell_{nk} \log v_{nk} \]

\[ + \sum_g^K (\ell_{g1} - 1) E[\log(\gamma_g)] + (\ell_{g2} - 1) E[\log(1 - \gamma_g)] - \log \text{BetaFun}(\ell_{g1}, \ell_{g2}) \]

\[ + \sum_n^K \sum_g^{G} E_q[i_{ng}] \log(o_{ng}) + (1 - E_q[i_{ng}]) \log(1 - o_{ng}) \]

\[ + \sum_g^K \sum_g^{D} c_{dg} \log c_{dg} \]

\[ = \sum_g^K \sum_k^K \lambda_{gk} \log \lambda_{gk} + (1 - \lambda_{gk}) \log(1 - \lambda_{gk}) \]

\[ + \sum_g^K \eta_g \log(\eta_g) + (1 - \eta_g) \log(1 - \eta_g) \]

\[ + \sum_k^K (\tau_k - 1)(\Psi(\tau_k) - \Psi(\sum_j^K \tau_j)) - \sum_k^K \log \Gamma(\tau_k) + \log \Gamma(\sum_k^K \tau_k) \]

\[ + \sum_g^K \sum_k^K ((\phi_{gk1} - 1)(\Psi(\phi_{gk1}) - \Psi(\phi_{gk1} + \phi_{gk2})) + (\phi_{gk2} - 1)(\Psi(\phi_{gk2}) - \Psi(\phi_{gk1} + \phi_{gk2}))) - \log \text{BetaFun}(\phi_{gk1}, \phi_{gk2}) \]

\[ + \sum_n^K \sum_k^K \ell_{nk} \log v_{nk} \]

\[ + \sum_g^K (\ell_{g1} - 1)(\Psi(\ell_{g1}) - \Psi(\ell_{g1} + \ell_{g2})) + (\ell_{g2} - 1)(\Psi(\ell_{g2}) - \Psi(\ell_{g1} + \ell_{2})) - \log \text{BetaFun}(\ell_{g1}, \ell_{g2}) \]

\[ + \sum_n^K \sum_o^{G} o_{ng} \log(o_{ng}) + (1 - o_{ng}) \log(1 - o_{ng}) \]

\[ + \sum_g^K \sum_g^{D} c_{dg} \log c_{dg} \]

\[ + \sum_g^K e \log(e_g) + (1 - e_g) \log(1 - e_g) \]