

I will start off with some useful properties I need to use later on in the ELBO.

## 1 Negative cross entropies

### 1.1 Two Normals

Note: All the normals are parametrized using the precision matrix.

$$q \sim \mathcal{N}(x|m, L)$$

$$p \sim \mathcal{N}(x|\mu, \Lambda)$$

$$\begin{aligned} \int q(x) \ln p(x) dx &= \int \mathcal{N}(x|m, L) \left( -\frac{K}{2} \ln 2\pi + \frac{1}{2} \ln |\Lambda| - \frac{1}{2} \left( \text{Tr} \Lambda \{ (x - \mu)(x - \mu)^T \} \right) \right) dx \\ &= -\frac{K}{2} \ln 2\pi + \frac{1}{2} \ln |\Lambda| + \int \mathcal{N}(x|m, L) \left( -\frac{1}{2} \left( \text{Tr} \Lambda \{ (x - \mu)(x - \mu)^T \} \right) \right) dx \\ &= -\frac{K}{2} \ln 2\pi + \frac{1}{2} \ln |\Lambda| + \int \mathcal{N}(x|m, L) \left( -\frac{1}{2} \left( \text{Tr} \Lambda \{ xx^T + \mu\mu^T - x\mu^T - \mu x^T \} \right) \right) dx \end{aligned}$$

$$\text{We should note that } \mathbb{E}_q [xx^T] = \text{Cov}_q + \mathbb{E}_q [x] \mathbb{E}_q [x]^T$$

$$\mathbb{E}_q [x] = m \text{ and } \text{Cov}_q = L^{-1}$$

$$\begin{aligned} \int \mathcal{N}(x|m, L) \left( -\frac{1}{2} \left( \text{Tr} \left[ \Lambda \{ xx^T + \mu\mu^T - x\mu^T - \mu x^T \} \right] \right) \right) dx &= -\frac{1}{2} \text{Tr} \left[ (\Lambda L^{-1} + \Lambda m m^T) + \Lambda (m m^T - \mu m^T - m \mu^T) \right] \\ &= -\frac{1}{2} \left( \text{Tr} [\Lambda L^{-1}] + (m - \mu)^T \Lambda (m - \mu) \right) \end{aligned}$$

Hence we have:

$$\boxed{\mathbb{E}_q [\ln p(x)] = -\frac{K}{2} \ln 2\pi + \frac{1}{2} \ln |\Lambda| - \frac{1}{2} \left( \text{Tr} [\Lambda L^{-1}] + (m - \mu)^T \Lambda (m - \mu) \right)}$$

### 1.2 Two Wisharts

$$\Lambda \sim q \sim \mathcal{W}(v, W)$$

$$\Lambda \sim p \sim \mathcal{W}(n, S)$$

$$\begin{aligned} \int q(\Lambda) \ln p(\Lambda) d\Lambda &= \mathbb{E}_q [\ln p(\Lambda)] \\ &= \mathbb{E}_q \left[ \ln \frac{|\Lambda|^{\frac{n-K-1}{2}} \exp(-\frac{1}{2} \text{Tr}(S^{-1} \Lambda))}{2^{\frac{nK}{2}} |S|^{n/2} \Gamma_p(\frac{n}{2})} \right] \\ &= \mathbb{E}_q \left[ -\frac{nk}{2} \ln 2 - \frac{n}{2} \ln |S| - \ln \Gamma_K(\frac{n}{2}) \right. \\ &\quad \left. + \frac{n-K-1}{2} \ln |\Lambda| - \frac{1}{2} \text{Tr}(S^{-1} \Lambda) \right] \\ &= -\frac{nk}{2} \ln 2 - \frac{n}{2} \ln |S| - \ln \Gamma_K(\frac{n}{2}) \\ &\quad + \frac{n-K-1}{2} \left( \psi_K(\frac{v}{2}) + K \ln 2 + \ln |W| \right) - \frac{v}{2} \text{Tr}(S^{-1} W) \end{aligned}$$

Note that:

$$\mathbb{E}_q[\Lambda] = vW$$

$$\mathbb{E}_q[\ln |\Lambda|] = \psi_K(\frac{v}{2}) + K \ln 2 + \ln |W|$$

$$\psi_K(\frac{v}{2}) = \sum_{i:1}^K \psi(\frac{v-i+1}{2})$$

$$\ln \Gamma_K(\frac{n}{2}) = \frac{K(K-1)}{4} \ln \pi + \sum_{i:1}^K \ln \Gamma(\frac{n-i+1}{2})$$

$$\begin{aligned} \mathbb{E}_q[\ln p(\Lambda)] &= -\frac{K(K+1)}{2} \ln 2 + \frac{n-K-1}{2} \psi_K(\frac{v}{2}) - \ln \Gamma_K(\frac{n}{2}) \\ &\quad - \frac{v}{2} \text{Tr}(S^{-1}W) + \frac{n-K-1}{2} \ln |W| - \frac{n}{2} \ln |S| \end{aligned}$$

so we have:

$$\mathbb{E}_q[\ln p(\Lambda)] = -\frac{K(K+1)}{2} \ln 2 + \frac{n-K-1}{2} \psi_K(\frac{v}{2}) - \ln \Gamma_K(\frac{n}{2}) - \frac{v}{2} \text{Tr}(S^{-1}W) + \frac{n-K-1}{2} \ln |W| - \frac{n}{2} \ln |S|$$

or

$$\mathbb{E}_q[\ln p(\Lambda)] = -\frac{K(K+1)}{2} \ln 2 + \frac{n-K-1}{2} \psi_K(\frac{v}{2}) - \ln \Gamma_K(\frac{n}{2}) - \frac{v}{2} \text{Tr}(S^{-1}W) - \frac{K+1}{2} \ln |W| + \frac{n}{2} \ln |S^{-1}W|$$

### 1.3 Two Betas

$$\beta \sim q \sim \text{Beta}(b)$$

$$\beta \sim p \sim \text{Beta}(\eta)$$

$$\begin{aligned} \mathbb{E}_q[\ln p(\beta)] &= \mathbb{E}_q \left[ \ln \Gamma(\eta_0 + \eta_1) - \ln \Gamma(\eta_0) - \ln \Gamma(\eta_1) + (\eta_0 - 1) \ln \beta + (\eta_1 - 1) \ln (1 - \beta) \right] \\ &= \ln \Gamma(\eta_0 + \eta_1) - \ln \Gamma(\eta_0) - \ln \Gamma(\eta_1) + (\eta_0 - 1) (\psi(b_0) - \psi(b_0 + b_1)) + (\eta_1 - 1) (\psi(b_1) - \psi(b_0 + b_1)) \\ &= \ln \Gamma(\eta_0 + \eta_1) - \ln \Gamma(\eta_0) - \ln \Gamma(\eta_1) + (\eta_0 - 1) \psi(b_0) + (\eta_1 - 1) \psi(b_1) - (\eta_0 + \eta_1 - 2) \psi(b_0 + b_1) \end{aligned}$$

$$\text{Note that } \mathbb{E}_q[\ln \beta] = \psi(b_0) - \psi(b_0 + b_1)$$

so :

$$\mathbb{E}_q[\ln p(\beta)] = \ln \Gamma(\eta_0 + \eta_1) - \ln \Gamma(\eta_0) - \ln \Gamma(\eta_1) + (\eta_0 - 1) \psi(b_0) + (\eta_1 - 1) \psi(b_1) - (\eta_0 + \eta_1 - 2) \psi(b_0 + b_1)$$

## 2 Entropies

### 2.1 Normal

$$q(x) \sim \mathcal{N}(m, M)$$

$$H[q] = \frac{K}{2} \ln(2\pi) + \frac{K}{2} - \frac{1}{2} \ln |M|$$

### 2.2 Wishart

$$\Lambda \sim q \sim \mathcal{W}(v, W)$$

$$\begin{aligned} H[q] &= -\frac{v-K-1}{2} \mathbb{E}_q \ln |\Lambda| - (-\frac{1}{2} \mathbb{E}_q \text{Tr}(W^{-1}\Lambda)) + \frac{v}{2} \ln |W| + \frac{vK}{2} \ln 2 + \ln \Gamma_K(\frac{v}{2}) \\ &= -\frac{v-K-1}{2} (\psi_K(\frac{v}{2}) + \frac{Kv}{2} + K \ln 2 + \ln |W|) + \frac{v}{2} \ln |W| + \frac{vK}{2} \ln 2 + \ln \Gamma_K(\frac{v}{2}) \\ &= \frac{K(K+1)}{2} \ln 2 + \frac{K+1}{2} \ln |W| - \frac{v-K-1}{2} \psi_K(\frac{v}{2}) + \ln \Gamma_K(\frac{v}{2}) + \frac{Kv}{2} \end{aligned}$$

so

$$H[q] = \frac{K(K+1)}{2} \ln 2 + \frac{K+1}{2} \ln |W| - \frac{v-K-1}{2} \psi_K(\frac{v}{2}) + \ln \Gamma_K(\frac{v}{2}) + \frac{Kv}{2}$$

### 2.3 Beta

$$\beta \sim q \sim \text{Beta}(b)$$

$$\begin{aligned} H[q] &= \ln \Gamma(b_0) + \ln \Gamma(b_1) - \ln \Gamma(b_0 + b_1) - (b_0 - 1) \mathbb{E}_q[\ln \beta] - (b_1 - 1) \mathbb{E}_q[\ln (1 - \beta)] \\ &= \ln \Gamma(b_0) + \ln \Gamma(b_1) - \ln \Gamma(b_0 + b_1) - (b_0 - 1) \psi(b_0) - (b_1 - 1) \psi(b_1) + (b_0 + b_1 - 2) \psi(b_0 + b_1) \end{aligned}$$

So,

$$H[q] = \ln \Gamma(b_0) + \ln \Gamma(b_1) - \ln \Gamma(b_0 + b_1) - (b_0 - 1)\psi(b_0) - (b_1 - 1)\psi(b_1) + (b_0 + b_1 - 2)\psi(b_0 + b_1)$$

## 2.4 Multinomial(,1) or Categorical

$$z \sim q \sim \text{Cat}(\phi)$$

$$H[q] = - \sum_k \mathbb{E}_q[z_k] \ln \phi_k$$

so,

$$H[q] = - \sum_k \phi_k \ln \phi_k$$

## 3 Variational ELBO

$$\mathcal{L} = \mathbb{E}_q[\ln p(\text{joint})] + H_q[\text{params}]$$

$$\begin{aligned} \ln p(\text{joint}) &= \ln p(\mu|m_0, M_0) + \ln p(\Lambda|\ell_0, L_0) + \sum_a \ln p(\theta_a|\mu, \Lambda) + \sum_a \sum_b \ln p(z_{a \rightarrow b}|\theta_a) \\ &\quad + \sum_a \sum_b \ln p(z_{a \leftarrow b}|\theta_b) + \sum_k \ln p(\beta_{kk}|\eta) + \sum_a \sum_b \ln p(y_{ab}|z_{a \rightarrow b}, z_{a \leftarrow b}, \beta) \end{aligned}$$

$$H_q[\text{params}] = H_q[\mu] + H_q[\Lambda] + H_q[\theta] + H_q[\beta] + H_q[z_{\rightarrow}] + H_q[z_{\leftarrow}]$$

Furthermore,

$$\begin{aligned} \mathbb{E}_q[\ln p(\text{joint})] &= \mathbb{E}_q[\ln p(\mu|m_0, M_0)] + \mathbb{E}_q[\ln p(\Lambda|\ell_0, L_0)] + \sum_a \mathbb{E}_q[\ln p(\theta_a|\mu, \Lambda)] + \sum_a \sum_b \mathbb{E}_q[\ln p(z_{a \rightarrow b}|\theta_a)] \\ &\quad + \sum_a \sum_b \mathbb{E}_q[\ln p(z_{a \leftarrow b}|\theta_b)] + \sum_k \mathbb{E}_q[\ln p(\beta_{kk}|\eta)] + \sum_a \sum_b \mathbb{E}_q[\ln p(y_{ab}|z_{a \rightarrow b}, z_{a \leftarrow b}, \beta)] \end{aligned}$$

We parametrize the variational distribution as follows:

$$\begin{aligned} \mu &\sim q(\mu|m, M) \sim \mathcal{N}(\mu|m, M) \\ \Lambda &\sim q(\Lambda|\ell, L) \sim \mathcal{W}(\Lambda|\ell, L) \\ \theta_a &\sim q(\theta_a|\mu_a, \Lambda_a) \sim \mathcal{N}(\theta_a|\mu_a, \Lambda_a) \\ \beta_{kk} &\sim q(\beta_{kk}|b_k) \sim \mathcal{B}(b_{k0}, b_{k1}) \\ z_{a \rightarrow b} &\sim q(z_{a \rightarrow b}|\phi_{a \rightarrow b}) \sim \text{Cat}(z_{a \rightarrow b}|\phi_{a \rightarrow b}) \\ z_{a \leftarrow b} &\sim q(z_{a \leftarrow b}|\phi_{a \leftarrow b}) \sim \text{Cat}(z_{a \leftarrow b}|\phi_{a \leftarrow b}) \end{aligned}$$

Using the results from above regarding the negative cross entropies:

$$\begin{aligned}
\mathbb{E}_q[\ln p(\text{joint})] &= -\frac{K}{2} \ln 2\pi + \frac{1}{2} \ln |M_0| - \frac{1}{2} \left( \text{Tr} M_0 \left[ M^{-1} + (m - m_0)(m - m_0)^T \right] \right) \\
&\quad - \frac{K(K+1)}{2} \ln 2 + \frac{\ell_0 - K - 1}{2} \psi_K\left(\frac{\ell}{2}\right) - \ln \Gamma_K\left(\frac{\ell_0}{2}\right) - \frac{\ell}{2} \text{Tr} (L_0^{-1} L) - \frac{K+1}{2} \ln |L| + \frac{\ell_0}{2} \ln |L_0^{-1} L| \\
&\quad - \sum_a \frac{K}{2} \ln 2\pi + \frac{1}{2} \sum_a \psi_K\left(\frac{\ell}{2}\right) + \frac{1}{2} \sum_a K \ln 2 + \frac{1}{2} \sum_a \ln |L| \\
&\quad - \frac{\ell}{2} \left( \text{Tr} \left[ L \left( \sum_a (\Lambda_a^{-1} + (m - \mu_a)(m - \mu_a)^T) + \sum_a M^{-1} \right) \right] \right) \\
&\quad + \sum_a \sum_b \sum_k \phi_{a \rightarrow b, k} \mu_{a, k} - \sum_a \sum_b \mathbb{E}_q[\ln (\sum_l \exp(\theta_{a, l}))] \\
&\quad + \sum_a \sum_b \sum_k \phi_{a \leftarrow b, k} \mu_{b, k} - \sum_a \sum_b \mathbb{E}_q[\ln (\sum_l \exp(\theta_{b, l}))] \\
&\quad + \sum_k \ln \Gamma(\eta_0 + \eta_1) - \sum_k \ln \Gamma(\eta_0) - \sum_k \ln \Gamma(\eta_1) + \sum_k (\eta_0 - 1) \psi(b_{k0}) \\
&\quad + \sum_k (\eta_1 - 1) \psi(b_{k1}) - \sum_k (\eta_0 + \eta_1 - 2) \psi(b_{k0} + b_{k1}) \\
&\quad + \sum_{a, b \in \text{link}} \sum_k \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) + \ln \epsilon \\
&\quad + \sum_{a, b \notin \text{link}} \sum_k \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) + \ln(1 - \epsilon)
\end{aligned}$$

$$\mathbb{E}_q[\Lambda] = \ell L$$

$$\mathbb{E}_q[\ln |\Lambda|] = \psi_K\left(\frac{\ell}{2}\right) + K \ln 2 + \ln |L|$$

$$\begin{aligned}
& - \sum_a \frac{K}{2} \ln 2\pi + \sum_a \frac{1}{2} \mathbb{E}_q \left\{ \ln |\Lambda| \right\} \\
& - \sum_a \frac{1}{2} \left( \text{Tr} \left[ \mathbb{E}_q \left\{ \Lambda \right\} \Lambda_a^{-1} \right] + \mathbb{E}_q \left\{ (\mu_a - \mu)^T \Lambda (\mu_a - \mu) \right\} \right) = \\
& - \sum_a \frac{K}{2} \ln 2\pi + \sum_a \psi_K\left(\frac{\ell}{2}\right) + \sum_a K \ln 2 + \sum_a \ln |L| \\
& - \frac{\ell}{2} \left( \text{Tr} \left[ L \left( \sum_a (\Lambda_a^{-1} + (m - \mu_a)(m - \mu_a)^T) \right) \right] \right)
\end{aligned}$$

For the expression  $\mathbb{E}_q[\ln (\sum_l \exp(\theta_{a, l}))]$ , we use the Jensen's inequality to acquire:

$$\begin{aligned}
\mathbb{E}_q[\ln (\sum_l \exp(\theta_{a, l}))] &\leq \ln (\sum_l \mathbb{E}_q[\exp(\theta_{a, l})]) \\
&= \ln (\sum_l \exp(\mu_{a, l} + \frac{1}{2} \text{diag}(\Lambda_a^{-1})_{ll}))
\end{aligned}$$

We can introduce another bound that introduces a new variational parameter per individual:

$$\mathbb{E}_q[\ln (\sum_l \exp(\theta_{a, l}))] \leq \zeta_a^{-1} \sum_l \exp(\mu_{a, l} + \frac{1}{2} \text{diag}(\Lambda_a^{-1})_{ll}) + \ln \zeta_a - 1$$

Moreover, using the entropies from above:

$$\begin{aligned}
H_q[params] = & \frac{K}{2} \ln(2\pi) + \frac{K}{2} - \frac{1}{2} \ln|M| \\
& + \frac{K(K+1)}{2} \ln 2 + \frac{K+1}{2} \ln|L| - \frac{\ell-K-1}{2} \psi_K\left(\frac{\ell}{2}\right) + \ln \Gamma_K\left(\frac{\ell}{2}\right) + \frac{K\ell}{2} \\
& + \sum_a \frac{K}{2} \ln(2\pi) + \sum_a \frac{K}{2} - \sum_a \frac{1}{2} \ln|\Lambda_a| \\
& + \sum_k \ln \Gamma(b_{k0}) + \sum_k \ln \Gamma(b_{k1}) - \sum_k \ln \Gamma(b_{k0} + b_{k1}) - \sum_k (b_{k0} - 1) \psi(b_{k0}) \\
& - \sum_k (b_{k1} - 1) \psi(b_{k1}) + \sum_k (b_{k0} + b_{k1} - 2) \psi(b_{k0} + b_{k1}) \\
& - \sum_a \sum_b \sum_k \phi_{a \rightarrow b, k} \ln \phi_{a \rightarrow b, k} \\
& - \sum_a \sum_b \sum_k \phi_{a \leftarrow b, k} \ln \phi_{a \leftarrow b, k}
\end{aligned}$$

Note that here I assume the following for the hyperparameters:

$$\begin{aligned}
m_0 &= \mathbf{0} \\
M_0 &= 10 \times \mathbf{I} \\
\ell_0 &= K + 2 \\
L_0 &= \frac{.1}{\ell_0} \mathbf{I} \\
\eta_0 &> 1 = 9 \\
\eta_1 &= 1
\end{aligned}$$

Note that here I assume the following for the variational parameters:

$$\begin{aligned}
m &= \mathbf{0} \\
M &= 10 \times \mathbf{I} \\
\ell &= K + 2 \\
L &= \frac{.1}{\ell_0} \mathbf{I} \\
b_0 &> 1 = 9 \\
b_1 &= 1
\end{aligned}$$

Finally, we have the following:

$$\begin{aligned}
\mathcal{L} = & -\frac{1}{2} \left( K \ln 2\pi - \ln |M_0| + \text{tr } M_0(m - m_0)(m - m_0)^T + \text{tr } M_0 M^{-1} \right) \\
& + \frac{1}{2} \left( -K(K+1) \ln 2 + (\ell_0 - K - 1) \sum_i \Psi\left(\frac{\ell-i+1}{2}\right) - \frac{K(K-1)}{2} \ln \pi - 2 \sum_i \ln \Gamma\left(\frac{\ell_0-i+1}{2}\right) \right. \\
& \left. - \ell \text{tr}(L_0^{-1} L) - (K+1) \ln |L| + \ell_0 \ln |L_0^{-1} L| \right) \\
& - \frac{1}{2} \sum_a \left( K \ln 2\pi - \sum_i \Psi\left(\frac{\ell-i+1}{2}\right) - K \ln 2 - \ln |L| + \right. \\
& \left. \ell \text{tr} \left\{ L[(\mu_a - m)(\mu_a - m)^T + M^{-1} + \Lambda_a^{-1}] \right\} \right) \\
& + \sum_a \sum_{b \in \text{sink}(a)} \left( \sum_k \phi_{a \rightarrow b, k} \mu_{a, k} - \ln \sum_l \exp(\mu_{a, l} + \frac{1}{2} \Lambda_{a, l}^{-1}) \right) \\
& + \sum_a \sum_{b \notin \text{sink}(a)} \left( \sum_k \phi_{a \rightarrow b, k} \mu_{a, k} - \ln \sum_l \exp(\mu_{a, l} + \frac{1}{2} \Lambda_{a, l}^{-1}) \right) \\
& + \sum_a \sum_{b \in \text{source}(a)} \left( \sum_k \phi_{b \leftarrow a, k} \mu_{a, k} - \ln \sum_l \exp(\mu_{a, l} + \frac{1}{2} \Lambda_{a, l}^{-1}) \right) \\
& + \sum_a \sum_{b \notin \text{source}(a)} \left( \sum_k \phi_{b \leftarrow a, k} \mu_{a, k} - \ln \sum_l \exp(\mu_{a, l} + \frac{1}{2} \Lambda_{a, l}^{-1}) \right) \\
& + \sum_k \ln \Gamma(\eta_0 + \eta_1) - \sum_k \ln \Gamma(\eta_0) - \sum_k \ln \Gamma(\eta_1) + \sum_k (\eta_0 - 1) \Psi(b_{k0}) + \sum_k (\eta_1 - 1) \Psi(b_{k1}) \\
& - \sum_k (\eta_0 + \eta_1 - 2) \Psi(b_{k0} + b_{k1}) \\
& + \sum_a \sum_{b \in \text{sink}(a)} \sum_k \left( \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\Psi(b_{k0}) - \Psi(b_{k0} + b_{k1}) - \ln \epsilon) + \ln \epsilon \right) \\
& + \sum_a \sum_{b \notin \text{sink}(a)} \sum_k \left( \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\Psi(b_{k1}) - \Psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) + \ln(1 - \epsilon) \right) \\
& + \frac{1}{2} \left( K \ln 2\pi + K - \ln |M| \right) \\
& + \frac{1}{2} \left( (K+1) \ln |L| + K(K+1) \ln 2 + \ell K + \frac{1}{2} K(K-1) \ln \pi \right. \\
& \left. + 2 \sum_i \ln \Gamma\left(\frac{\ell-i+1}{2}\right) - (\ell - K - 1) \sum_i \Psi\left(\frac{\ell-i+1}{2}\right) \right) \\
& + \frac{1}{2} \sum_a \left( K \ln 2\pi - \ln |\Lambda_a| + K \right) \\
& + \sum_k \left( \ln \Gamma(b_{k0}) + \ln \Gamma(b_{k1}) - \ln \Gamma(b_{k0} + b_{k1}) - (b_{k0} - 1) \Psi(b_{k0}) - \right. \\
& \left. (b_{k1} - 1) \Psi(b_{k1}) + (b_{k0} + b_{k1} - 2) \Psi(b_{k0} + b_{k1}) \right) \\
& - \sum_a \sum_{b \in \text{sink}(a)} \sum_k \left( \phi_{a \rightarrow b, k} \ln \phi_{a \rightarrow b, k} \right) \\
& - \sum_a \sum_{b \notin \text{sink}(a)} \sum_k \left( \phi_{a \rightarrow b, k} \ln \phi_{a \rightarrow b, k} \right) \\
& - \sum_a \sum_{b \in \text{sink}(a)} \sum_k \left( \phi_{a \leftarrow b, k} \ln \phi_{a \leftarrow b, k} \right) \\
& - \sum_a \sum_{b \notin \text{sink}(a)} \sum_k \left( \phi_{a \leftarrow b, k} \ln \phi_{a \leftarrow b, k} \right)
\end{aligned}$$

## 4 ELBO Gradients

### 4.1 Gradient with respect to $m$

$$\begin{aligned}
\mathcal{L}_m &= -\frac{1}{2} \left( \text{Tr } M_0 (m - m_0)(m - m_0)^T \right) \\
&\quad - \frac{\ell}{2} \left( \text{Tr } L \left( \sum_a (\mu_a - m)(\mu_a - m)^T \right) \right) \\
&\propto \text{Tr } M_0 (m - m_0)(m - m_0)^T \\
&\quad + \ell \left( \text{Tr } L \left( \sum_a m m^T + \mu_a \mu_a^T - m \mu_a^T - \mu_a m^T \right) \right) \\
&= \\
&\Rightarrow \\
\nabla_m \mathcal{L}_m &\propto 2M_0(m - m_0) - 2\ell L \sum_a (\mu_a - m) = 0 \\
&\Rightarrow \\
&\boxed{m = (M_0 + N\ell L)^{-1} (M_0 m_0 + \ell L \sum_a \mu_a)}
\end{aligned}$$

In minibatch node sampling this would be

$$\boxed{m = M^{-1} (M_0 m_0 + \ell L \frac{N}{\#mbnodes} \sum_{a \in mbnodes} \mu_a)}$$

### 4.2 Gradient with respect to $M$

$$\begin{aligned}
\mathcal{L}_M &= -\frac{1}{2} \left( \text{Tr } M_0 M^{-1} \right) \\
&\quad - \frac{\ell}{2} \text{Tr } N L M^{-1} \\
&\quad - \frac{1}{2} \ln |M| \\
&\propto \text{Tr } M_0 M^{-1} + \ell \text{Tr } N L M^{-1} + \ln |M| \\
&\Rightarrow \\
\nabla_{M^{-1}} \mathcal{L}_M &= 0 \\
&= -M_0 - N\ell L + M = 0 \\
&\boxed{M = M_0 + N\ell L}
\end{aligned}$$

### 4.3 Gradient with respect to $L$

$$\begin{aligned}
\mathcal{L}_L &= -\frac{\ell}{2} \text{Tr}(L_0^{-1}L) - \frac{K+1}{2} \ln |L| + \frac{\ell_0}{2} \ln |L_0^{-1}L| \\
&\quad + \frac{1}{2} \sum_a \ln |L| - \frac{\ell}{2} \left( \text{Tr} \left[ L \left( \sum_a (\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T) + \sum_a M^{-1} \right) \right] \right) \\
&\quad + \frac{K+1}{2} \ln |L| \\
&\propto -\ell \text{Tr}(L_0^{-1}L) - (K+1) \ln |L| + \ell_0 \ln |L_0^{-1}L| \\
&\quad + \sum_a \ln |L| - \ell \left( \text{Tr} \left[ L \left( \sum_a (\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T) + \sum_a M^{-1} \right) \right] \right) \\
&\quad + (K+1) \ln |L| \\
&\Rightarrow \\
\nabla_L \mathcal{L}_L &= -\ell L_0^{-1} + \frac{1}{2}(\ell_0 + N)L^{-1} - \ell \left( \sum_a (\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T) + \sum_a M^{-1} \right)^T = 0 \\
&\quad \ell(L_0^{-1} + \sum_a \Lambda_a^{-1} + \sum_a (\mu_a - m)(\mu_a - m)^T + NM^{-1}) = (N + \ell_0)L^{-1} \\
&\Rightarrow \boxed{L = \frac{(N + \ell_0)}{\ell} \left( (L_0^{-1} + \sum_a (\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T) + \sum_a M^{-1}) \right)^{-1}}
\end{aligned}$$

optimizing simultaeneously with  $\ell$  in the minibatch setting:

$$\boxed{L = \left( (L_0^{-1} + \frac{N}{\#mbnodes} \{ \sum_a \Lambda_a^{-1} + \sum_a (\mu_a - m)(\mu_a - m)^T \} + NM^{-1}) \right)^{-1}}$$

### 4.4 Gradient with respect to $\ell$

$$\mathcal{L}_\ell = \text{revise}$$

$$\propto$$

$$\Rightarrow$$

$$\propto$$

$$\Rightarrow$$

$$\nabla_\ell \mathcal{L}_\ell =$$

$$\Rightarrow$$

hence,

$$\Rightarrow$$

$$\boxed{\ell = \ell_0 + N}$$



#### 4.5 Gradient with respect to $b_k$

$$\mathcal{L}_{b_k} = \text{revise}$$

$$\begin{aligned}
& \text{simultaneously optimizing } b_{k0}, b_{k1} \\
\Rightarrow & \text{Similar to our previous results} \\
\nabla_{b_{k0}} \mathcal{L}_{b_k} &= 0 \\
\Rightarrow & \boxed{b_{k0} = \eta_0 + \frac{\#trainlinks}{\#mblinks} \sum_{a,b \in mblinks} \phi_{a \rightarrow b,k} \phi_{a \leftarrow b,k}} \\
\nabla_{b_{k1}} \mathcal{L}_{b_k} &= 0 \\
& \boxed{b_{k1} = \eta_1 + \frac{\#trainnonlinks}{\#mbnonlinks} \sum_{a,b \notin mblinks} \phi_{a \rightarrow b,k} \phi_{a \leftarrow b,k}}
\end{aligned}$$

#### 4.6 Gradient with respect to $\phi_{a \rightarrow b,k}$ for links

$$\begin{aligned}
\mathcal{L}_{\phi_{a \rightarrow b,k}} &= \phi_{a \rightarrow b,k} \mu_{a,k} \\
&+ \phi_{a \rightarrow b,k} \phi_{a \leftarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) \\
&- \phi_{a \rightarrow b,k} \ln \phi_{a \rightarrow b,k} \\
&= \phi_{a \rightarrow b,k} \left( \mu_{a,k} + \phi_{a \leftarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) - \ln \phi_{a \rightarrow b,k} \right) \\
\nabla_{\phi_{a \rightarrow b,k}} \mathcal{L}_{\phi_{a \rightarrow b,k}} &= \mu_{a,k} + \phi_{a \leftarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) - \ln \phi_{a \rightarrow b,k} = 0 \\
&\boxed{\phi_{a \rightarrow b,k} \propto \exp \left\{ \mu_{a,k} + \phi_{a \leftarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) \right\}}
\end{aligned}$$

#### 4.7 Gradient with respect to $\phi_{a \leftarrow b,k}$ for links

$$\begin{aligned}
\mathcal{L}_{\phi_{a \leftarrow b,k}} &= \phi_{a \leftarrow b,k} \mu_{b,k} \\
&+ \phi_{a \rightarrow b,k} \phi_{a \leftarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) \\
&- \phi_{a \leftarrow b,k} \ln \phi_{a \leftarrow b,k} \\
&= \phi_{a \leftarrow b,k} \left( \mu_{b,k} + \phi_{a \rightarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) - \ln \phi_{a \leftarrow b,k} \right) \\
\nabla_{\phi_{a \leftarrow b,k}} \mathcal{L}_{\phi_{a \leftarrow b,k}} &= \mu_{b,k} + \phi_{a \rightarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) - \ln \phi_{a \leftarrow b,k} = 0 \\
&\boxed{\phi_{a \leftarrow b,k} \propto \exp \left\{ \mu_{b,k} + \phi_{a \rightarrow b,k} (\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - \ln \epsilon) \right\}}
\end{aligned}$$

#### 4.8 Gradient with respect to $\phi_{a \rightarrow b, k}$ for nonlinks

$$\begin{aligned}
\mathcal{L}_{\phi_{a \rightarrow b, k}} &= \phi_{a \rightarrow b, k} \mu_{a, k} \\
&\quad + \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) \\
&\quad - \phi_{a \rightarrow b, k} \ln \phi_{a \rightarrow b, k} \\
&= \phi_{a \rightarrow b, k} \left( \mu_{a, k} + \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) - \ln \phi_{a \rightarrow b, k} \right) \\
\nabla_{\phi_{a \rightarrow b, k}} \mathcal{L}_{\phi_{a \rightarrow b, k}} &= \mu_{a, k} + \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) - \ln \phi_{a \rightarrow b, k} = 0 \\
&\quad \boxed{\phi_{a \rightarrow b, k} \propto \exp \left\{ \mu_{a, k} + \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) \right\}}
\end{aligned}$$

#### 4.9 Gradient with respect to $\phi_{a \leftarrow b, k}$ for nonlinks

$$\begin{aligned}
\mathcal{L}_{\phi_{a \leftarrow b, k}} &= \phi_{a \leftarrow b, k} \mu_{b, k} \\
&\quad + \phi_{a \rightarrow b, k} \phi_{a \leftarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) \\
&\quad - \phi_{a \leftarrow b, k} \ln \phi_{a \leftarrow b, k} \\
&= \phi_{a \leftarrow b, k} \left( \mu_{b, k} + \phi_{a \rightarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) - \ln \phi_{a \leftarrow b, k} \right) \\
\nabla_{\phi_{a \leftarrow b, k}} \mathcal{L}_{\phi_{a \leftarrow b, k}} &= \mu_{b, k} + \phi_{a \rightarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) - \ln \phi_{a \leftarrow b, k} = 0 \\
&\quad \boxed{\phi_{a \leftarrow b, k} \propto \exp \left\{ \mu_{b, k} + \phi_{a \rightarrow b, k} (\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - \ln(1 - \epsilon)) \right\}}
\end{aligned}$$

#### 4.10 Gradient with respect to $\mu_a$

$\mu_a$  and  $\Lambda_a$  are two of the scarier ones.

$$\begin{aligned}
\mathcal{L}_{\mu_a} &= -\frac{\ell}{2} [(\mu_a - m)^T L (\mu_a - m)] + \\
&\quad \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b}^T \mu_a + \\
&\quad \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b}^T \mu_a + \\
&\quad \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a}^T \mu_a + \\
&\quad \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a}^T \mu_a - \\
&\quad \sum_b \log \left( \mathbf{1}^T \underline{\mathbf{f}}(\mu_a, \Lambda_a) \right)
\end{aligned}$$

$$\text{where } \underline{\mathbf{f}}(\mu_a, \Lambda_a) = \begin{pmatrix} \exp(\mu_{a,1} + \frac{1}{2} \Lambda_{a,1}^{-1}) \\ \vdots \\ \exp(\mu_{a,k} + \frac{1}{2} \Lambda_{a,k}^{-1}) \\ \vdots \\ \exp(\mu_{a,K} + \frac{1}{2} \Lambda_{a,K}^{-1}) \end{pmatrix}, \text{ and we may for convenience interchangeably use } \underline{\mathbf{f}}_a \text{ to refer to } \underline{\mathbf{f}}(\mu_a, \Lambda_a) :$$

Hence the gradient is

$$\begin{aligned}
\nabla_{\mu_a} \mathcal{L}_{\mu_a} &= -\ell L(\mu_a - m) + \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a} + \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a} - \sum_b \frac{\partial \mathbf{f}(\mu_a, \Lambda_a)}{\partial \mu_a} (1) \\
&= -\ell L(\mu_a - m) + \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a} + \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a} - \sum_b \frac{\mathbf{J}_{\mathbf{f}} \times \mathbf{1}}{\mathbf{1}^T \mathbf{f}(\mu_a, \Lambda_a)} \\
&= -\ell L(\mu_a - m) + \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a} + \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a} - \sum_b \frac{\begin{pmatrix} \frac{\partial \mathbf{f}_{a1}}{\partial \mu_{a1}} & \cdots & \frac{\partial \mathbf{f}_{a1}}{\partial \mu_{ak}} & \cdots & \frac{\partial \mathbf{f}_{a1}}{\partial \mu_{aK}} \\ \vdots & \ddots & \vdots & & \vdots \\ \frac{\partial \mathbf{f}_{ak}}{\partial \mu_{a1}} & \cdots & \frac{\partial \mathbf{f}_{ak}}{\partial \mu_{ak}} & \cdots & \frac{\partial \mathbf{f}_{ak}}{\partial \mu_{aK}} \\ \vdots & & \ddots & & \vdots \\ \frac{\partial \mathbf{f}_{aK}}{\partial \mu_{a1}} & \cdots & \cdots & \ddots & \frac{\partial \mathbf{f}_{aK}}{\partial \mu_{aK}} \end{pmatrix}}{\mathbf{1}^T \mathbf{f}(\mu_a, \Lambda_a)} \\
&= -\ell L(\mu_a - m) + \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a} + \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a} - \sum_b \frac{\text{sfx}(a)}{b}
\end{aligned}$$

$$\text{where } \text{sfx}(a) = \begin{pmatrix} \frac{\exp(\mu_{a,1} + \frac{1}{2} \Lambda_{a,1}^{-1})}{\sum_l \exp(\mu_{a,l} + \frac{1}{2} \Lambda_{a,l}^{-1})} \\ \vdots \\ \frac{\exp(\mu_{a,k} + \frac{1}{2} \Lambda_{a,k}^{-1})}{\sum_l \exp(\mu_{a,l} + \frac{1}{2} \Lambda_{a,l}^{-1})} \\ \vdots \\ \frac{\exp(\mu_{a,1} + \frac{1}{2} \Lambda_{a,1}^{-1})}{\sum_l \exp(\mu_{a,l} + \frac{1}{2} \Lambda_{a,l}^{-1})} \end{pmatrix}$$

so all in all the gradient is :

$$\nabla_{\mu_a} \mathcal{L}_{\mu_a} = \boxed{-\ell L(\mu_a - m) + \sum_{b \in \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \notin \text{sink}(a)} \phi_{a \rightarrow b} + \sum_{b \in \text{source}(a)} \phi_{b \leftarrow a} + \sum_{b \notin \text{source}(a)} \phi_{b \leftarrow a} - \sum_b \underline{\text{sfx}}(a)}$$

Similarly the Hessian will be as follows:

$$\begin{aligned} \nabla_{\mu_a}^2 \mathcal{L}_{\mu_a} &= \begin{aligned} &-\ell L- \\ &\sum_b \frac{\partial \underline{\text{sfx}}(a)}{\partial \mu_a^T} \\ &\ell L- \\ &\sum_b \underline{J}_{\underline{\text{sfx}}(a)} \\ &-\ell L- \end{aligned} \\ &= \\ &= \sum_b \begin{pmatrix} \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{a1}} & \cdots & \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{ak}} & \cdots & \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{aK}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{a1}} & \cdots & \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{ak}} & \cdots & \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{aK}} \\ \vdots & & \vdots & \ddots & \vdots \\ \frac{\partial \underline{\text{sfx}}_{aK}}{\partial \mu_{a1}} & \cdots & \cdots & \cdots & \frac{\partial \underline{\text{sfx}}_{aK}}{\partial \mu_{aK}} \end{pmatrix} \\ &= \sum_b \begin{pmatrix} \underline{\text{sfx}}_{a1} - \underline{\text{sfx}}_{a1}^2 & \cdots & -\underline{\text{sfx}}_{a1} \underline{\text{sfx}}_{ak} & \cdots & -\underline{\text{sfx}}_{a1} \underline{\text{sfx}}_{aK} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ -\underline{\text{sfx}}_{a1} \underline{\text{sfx}}_{ak} & \cdots & \underline{\text{sfx}}_{ak} - \underline{\text{sfx}}_{ak}^2 & \cdots & -\underline{\text{sfx}}_{ak} \underline{\text{sfx}}_{aK} \\ \vdots & & \vdots & \ddots & \vdots \\ -\underline{\text{sfx}}_{a1} \underline{\text{sfx}}_{aK} & \cdots & \cdots & \cdots & -\underline{\text{sfx}}_{aK} - \underline{\text{sfx}}_{aK}^2 \end{pmatrix} \\ &= \boxed{-\ell L - \sum_b \left( \text{diagm}(\underline{\text{sfx}}_a) - \underline{\text{sfx}}_a \underline{\text{sfx}}_a^T \right)} \end{aligned}$$

The newton step would look like:

$$\mu_{a,k} = \mu_{a,k} - H_{\mu_{a,k}}^{-1} G_{\mu_{a,k}}$$

#### 4.11 Gradient with respect to $\Lambda_a$

similarly assuming that  $\Lambda_a$  is a diagonal matrix(or a column vector).

$$\begin{aligned} \mathcal{L}_{\Lambda_a^{-1}} &= -\frac{\ell}{2} \text{diag}(L)' \Lambda_a^{-1} + \frac{1}{2} \ln |\text{diagm}(\Lambda_a^{-1})| - \sum_b \log \left( \mathbf{1}^T \underline{\text{f}}(\mu_a, \Lambda_a) \right) \\ &= \\ \nabla_{\Lambda_a^{-1}} \mathcal{L}_{\Lambda_a^{-1}} &= G_{\Lambda_a^{-1}} = \boxed{-\frac{\ell}{2} \text{diag}(L) + \frac{1}{2} (\Lambda_a) - \frac{1}{2} \sum_b (\underline{\text{sfx}}(a))} \\ &\quad \square \\ \nabla_{\Lambda_a^{-1}}^2 \mathcal{L}_{\Lambda_a^{-1}} &= H_{\Lambda_a^{-1}} \propto \boxed{-\frac{1}{2} \text{diagm}(\Lambda_a \odot \Lambda_a) - \frac{1}{4} \sum_b \left( \text{diagm}(\underline{\text{sfx}}_a) - \underline{\text{sfx}}_a \underline{\text{sfx}}_a^T \right)} \end{aligned}$$

The newton step would look like:

$$\Lambda_a^{-1} = \Lambda_a^{-1} - H_{\Lambda_a^{-1}}^{-1} G_{\Lambda_a^{-1}}$$

## 5 Other notes

### 5.1 Checknig ELBO

I need to use the training data, perhaps all the training links and same number sampled from nonlinks. Depending on the size of the network and the sparsity, this could be costly, so this only is needed every once in a while(not frequently) and only to check whether the ELBO is improving with our optimization algorithm or not.

### 5.2 Other forms of evluations of the model

Perplexity of a model  $q$ , given a stochastic process  $p$  (here a joint generative model that can produce infinite streams of data) is defined as:

$$\text{perplexity}(p, q) \triangleq 2^{H(p, q)}$$

where the cross entropy

$$H(p, q) \triangleq \lim_{N \rightarrow \infty} -\frac{1}{N} \sum_{y_{1:N}} p(y_{1:N}) \log q(y_{1:N})$$

and perplexity the lower the better. But we use the empirical distribution for  $p$ .  $p_{emp}(y_{1:N}) = \delta_{y_{1:N}^*}(y_{1:N})$ , where  $y^*$  is a single long test sequence. Then  $H(p_{emp}, q) = -\frac{1}{N} \log q(y_{1:N}^*)$ , so the perplexity becomes:

$$\text{perplexity}(p_{emp}, q) = q(y_{1:N}^*)^{1/N} = \sqrt[N]{\prod_{i=1}^N \frac{1}{q(y_i^* | y_{1:i-1}^*)}}$$

which is the geometric mean of the inverse of the predictive probability.

We could also think of it as exponential of the negative average log likelihood of the data. average perplexity on a test set is:

$$\text{perp}_{avg}(\text{Test} | \text{params}) = \exp \left( -\frac{\sum_{(a,b) \in \text{Test}} \log \left( \frac{1}{T} \right) \sum_{t=1}^T p(y_{ab} | \text{params})}{|\text{Test}|} \right)$$

We can set for example a 1% of the links and the nonlinks as the test set.

For the data with ground truth we can also use normalized mutual information(however, there have been ups and downs for this approach).

Further write more about some post predictive checks(PPC).

### 5.3 some fixed ones

I should ensure that  $M$  remains  $I + N\ell L$  with  $N$  the actual size of the nodes in the training set. This also affects the update of  $m$  which is  $M^{-1}(\ell L \frac{N}{\#mbnodes} \sum_{a \in mbnodes} \mu_a)$ . Moreover  $\ell = K + N$  can be fixed in advanced, and there is no need for computation in the variational loop.

### 5.4 Newton step behavior

I should ensure if the ELBO does not improve, I should half the step in the learning rate, other wise keep it the same. In general ELBO should be in sample and not out-of-sample, so I should think whether I want this ELBO to be on a sample training or just the minibatch.

### 5.5 Initialization

I should check whether the initialization is very off or should I adopt the Gopalan's initialization algorithm.