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.
$q \sim \mathcal{N}(x|m, L)$
$p \sim \mathcal{N}(x|\mu, \Lambda)$

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

We should note that $\mathbb{E}_q\Big[xx^T\Big] = Cov_q + \mathbb{E}_q\Big[x\Big]\mathbb{E}_q\Big[x\Big]^T$
$\mathbb{E}_q\Big[x\Big] = m$ and $Cov_q = L^{-1}$

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

Hence we have:

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

1.2 Two Wisharts

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

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

Note that:
$\mathbb{E}_q[\Lambda] = vW$
$\mathbb{E}_q[ln\,|\Lambda|] = \psi_K(\frac{v}{2}) + Kln\,2 + ln\,|W|$
$\psi_K(\frac{v}{2}) = \sum_{i:1}^{K} \psi(\frac{v-i+1}{2})$
$ln\,\Gamma_K(\frac{n}{2}) = \frac{K(K-1)}{4} ln\,\pi + \sum_{i:1}^{K} ln\,\Gamma(\frac{n-i+1}{2})$

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

so we have:
$$\boxed{\mathbb{E}_q[ln\,p(\Lambda)] = -\tfrac{K(K+1)}{2} ln\,2 + \tfrac{n-K-1}{2}\psi_K(\tfrac{v}{2}) - ln\,\Gamma_K(\tfrac{n}{2}) - \tfrac{v}{2} Tr\,(S^{-1}W) + \tfrac{n-K-1}{2} ln\,|W| - \tfrac{n}{2} ln\,|S|}$$
or
$$\boxed{\mathbb{E}_q[ln\,p(\Lambda)] = -\tfrac{K(K+1)}{2} ln\,2 + \tfrac{n-K-1}{2}\psi_K(\tfrac{v}{2}) - ln\,\Gamma_K(\tfrac{n}{2}) - \tfrac{v}{2} Tr\,(S^{-1}W) - \tfrac{K+1}{2} ln\,|W| + \tfrac{n}{2} ln\,|S^{-1}W|}$$

1.3 Two Betas

$\beta \sim q \sim Beta(b)$
$\beta \sim p \sim Beta(\eta)$

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

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

2 Entropies

2.1 Normal

$q(x) \sim \mathcal{N}(m, M)$
$$\boxed{H[q] = \tfrac{K}{2} ln\,(2\pi) + \tfrac{K}{2} - \tfrac{1}{2} ln\,|M|}$$

2.2 Wishart

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

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

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

2.3 Beta

$\beta \sim q \sim Beta(b)$

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

So,

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

2.4 Multinomial(,1) or Categorical

$z \sim q \sim Cat(\phi)$

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

so,

$$\boxed{H[q] = -\sum_k \phi_k ln\,\phi_k}$$

3 Variational ELBO

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

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

$$H_q[params] \quad = \quad H_q[\mu] + H_q[\Lambda] + H_q[\theta] + H_q[\beta] + H_q[z_\to] + H_q[z_\leftarrow]$$

Furthermore,

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

We parametrize the variational distribution as follows:

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

Using the results from above regarding the negative cross entropies:

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

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

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

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

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

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

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

Moreover, using the entropies from above:

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

Note that here I assume the following for the hyperparameters:

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

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

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

Finally, we have the following:

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

4 ELBO Gradients

4.1 Gradient with respect to m

$$\begin{aligned}
\mathcal{L}_m &= -\tfrac{1}{2}\Big(Tr\, M_0(m-m_0)(m-m_0)^T\Big]\Big) \\
&\quad -\tfrac{\ell}{2}\Big(Tr\, L\Big(\sum_a(\mu_a-m)(\mu_a-m)^T\Big)\Big) \\
&\propto Tr\, M_0(m-m_0)(m-m_0)^T \\
&\quad +\ell\Big(Tr\, L(\sum_a mm^T+\mu_a\mu_a^T-m\mu_a^T-\mu_a m^T)\Big) \\
&= \\
&\Longrightarrow \\
\nabla_m\mathcal{L}_m &\propto 2M_0(m-m_0)-2\ell L\sum_a(\mu_a-m)=0 \\
&\Longrightarrow \\
&\quad \boxed{m=(M_0+N\ell L)^{-1}(M_0m_0+\ell L\sum_a\mu_a)}
\end{aligned}$$

In minibatch node sampling this would be

$$\boxed{m=M^{-1}(M_0m_0+\ell L\frac{N}{\#mbnodes}\sum_{a\in mbnodes}\mu_a)}$$

4.2 Gradient with respect to M

$$\begin{aligned}
\mathcal{L}_M &= -\tfrac{1}{2}\Big(Tr\, M_0M^{-1}\Big) \\
&\quad -\tfrac{\ell}{2}Tr\, NLM^{-1} \\
&\quad -\tfrac{1}{2}ln\,|M| \\
&\propto Tr\, M_0M^{-1}+\ell Tr\, NLM^{-1}+ln\,|M| \\
&\Longrightarrow \\
\nabla_{M^{-1}}\mathcal{L}_M &= 0 \\
&= -M_0-N\ell L+M=0 \\
&\quad \boxed{M=M_0+N\ell L}
\end{aligned}$$

4.3 Gradient with respect to L

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

optimizing simultaeneously with ℓ in the minibatch setting:

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

4.4 Gradient with respect to ℓ

$$
\begin{aligned}
\mathcal{L}_\ell &= revise \\
&\propto \\
&\implies \\
&\propto \\
&\implies \\
\nabla_\ell \mathcal{L}_\ell &= \\
&\implies \\
&hence, \\
&\implies \\
& \boxed{\ell = \ell_0 + N}
\end{aligned}
$$

4.5 Gradient with respect to b_k

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

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

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

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

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

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

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

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

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

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

4.10 Gradient with respect to μ_a

μ_a and Λ_a are two of the scarier ones.

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

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

Hence the geradient is

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

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

so all in all the gradient is :

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

Similarly the Hessian will be as follows:

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

The newton step would look like:
$\mu_{a,k} = \mu_{a,k} - H^{-1}_{\mu_{a,k}} G_{\mu_{a,k}}$

4.11 Gradient with respect to Λ_a

similarly assuming that Λ_a is a diagonal matrix(or a column vector).

$$\begin{aligned}
\mathcal{L}_{\Lambda_a^{-1}} &= -\tfrac{\ell}{2} tr\left(L\Lambda_a^{-1}\right) + \tfrac{1}{2} ln\, |\Lambda_a^{-1}| - \sum_b log\left(\mathbf{1}^T \underline{\text{f}}(\mu_a, \Lambda_a)\right) \\
&= \\
\nabla_{\Lambda_a^{-1}}\mathcal{L}_{\Lambda_a^{-1}} = G_{\Lambda_a^{-1}} &= \boxed{-\tfrac{\ell}{2} diag(L) + \tfrac{1}{2} diagm(\Lambda_a) - \tfrac{1}{2}\sum_b diagm(\underline{\text{sfx}}(a))} \#\#CHECK
\end{aligned}$$

□

$$\nabla^2_{\Lambda_a^{-1}}\mathcal{L}_{\Lambda_a^{-1}} = H_{\Lambda_a^{-1}} \propto \boxed{-\tfrac{1}{2} diagm(\Lambda_a \odot \Lambda_a) - \tfrac{1}{4}\sum_b \left(diagm(\underline{\text{sfx}}_a) - \underline{\text{sfx}}_a \underline{\text{sfx}}_a^T \right) \#\#\#CHECK}$$

The newton step would look like:
$\Lambda_a^{-1} = \Lambda_a^{-1} - H^{-1}_{\Lambda_a^{-1}} G_{\Lambda_a^{-1}}$

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 inifinite streams of data) is defined as:

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

where the cross entropy

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

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:

$$perplexity(p_{emp},q) = \qquad q(y^*_{1:N})^{1/N} =^N \sqrt{\prod_{i=1}^{N}\frac{1}{q(y^*_i|y^*_{1:i-1})}}$$

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:

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

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.