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 $Cov_q = L^{-1}$

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

Hence we have:

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

## 1.2 Two Wisharts

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

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

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

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

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

### 1.3 Two Betas

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

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

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

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

### 2.2 Wishart

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

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

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

### 2.3 Beta

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

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

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

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

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

so,

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

## 3 Variational ELBO

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

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

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

Furthermore,

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

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:

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

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

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

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

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

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:

$$
\begin{aligned}
m_0 &= 0 \\
M_0 &= I \\
\ell_0 &= K \\
L_0 &= \tfrac{1}{K} I \\
\eta_0 &> 1 \\
\eta_1 &= 1
\end{aligned}
$$

Finally, we have the following:

$$
\begin{aligned}
\mathcal{L} \;=\;& -\tfrac{1}{2}\Big(Kln\,2\pi + tr\,(mm^T) + tr\,M^{-1}\Big) \\
& -\tfrac{1}{2}\Big( -K^2 ln\,K + ln\,|L| + \ell K + tr\,L + \tfrac{K(K-1)}{2} ln\,\pi + \\
& 2\sum_i ln\,\Gamma(\tfrac{K-i+1}{2}) + \sum_i \Psi(\tfrac{\ell-i+1}{2}) + K(K+1) ln\,2\Big) \\
& -\tfrac{1}{2}\Big(Kln\,2\pi - \sum_i \Psi(\tfrac{\ell-i+1}{2}) - Kln\,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 (\eta_0 - 1)\Psi(b_{k0}) - \sum_k (\eta_0 - 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(Kln\,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}) - \tfrac{\ell-K-1}{2}\sum_i \Psi(\tfrac{\ell-i+1}{2})\Big) \\
& +\tfrac{1}{2}\sum_a \Big(Kln\,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 $m$

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

In minibatch node sampling this would be

$$
\boxed{m = \Big(\#mbnodes \times (\tfrac{1}{\#mbnodes}I + \ell L)\Big)^{-1}(\ell L\sum_{a\in mbnodes}\mu_a)}
$$

### 4.2 Gradient with respect to $M$

$$
\begin{aligned}
\mathcal{L}_M &= -\tfrac{1}{2}\Big(Tr\, M^{-1}\Big) \\
&\quad -\tfrac{\ell}{2}Tr\, NLM^{-1} \\
&\quad -\tfrac{1}{2}ln\,|M| \\
&\propto Tr\, M^{-1} + \ell Tr\, NLM^{-1} + ln\,|M| \\
&\implies \\
\nabla_M \mathcal{L}_M &= -(M^{-1}M^{-1})^T - \ell N(M^{-1}LM^{-1})^T + (M^{-1})^T = 0 \\
&\underset{\overbrace{\implies}^{M\times()\times M}}{transpose} \\
&\quad -I - N\ell L + M = 0 \\
&\quad \boxed{M = I + N\ell L}
\end{aligned}
$$

In minibatch node sampling this would be

$$
\boxed{M = \#mbnodes \times (\tfrac{1}{\#mbnodes}I + \ell L)}
$$

### 4.3 Gradient with respect to $L$

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

optimizing simultaeneously with $\ell$:

$$
\boxed{L = \left((KI + \sum_{a\in mbnodes} \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big)) + \#mbnodes \times M^{-1})\right)^{-1}}
$$

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

$$
\begin{aligned}
\mathcal{L}_\ell \quad &= \quad \tfrac{-1}{2}\psi_K(\tfrac{\ell}{2}) + \tfrac{1}{2}\sum_a \psi_K(\tfrac{\ell}{2}) - \tfrac{\ell}{2}Tr\,(KIL) \\
&\quad -\tfrac{\ell}{2}\Big(Tr\,\Big[L\Big(\sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\}\Big) \\
&\quad -\tfrac{\ell-K-1}{2}\psi_K(\tfrac{\ell}{2}) + ln\,\Gamma_K(\tfrac{\ell}{2}) + \tfrac{K\ell}{2} \\
&\propto \quad (-1)\psi_K(\tfrac{\ell}{2}) + \sum_a \psi_K(\tfrac{\ell}{2}) - \ell Tr\,(KIL) \\
&\quad -\ell\Big(Tr\,\Big[L\Big(\sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\}\Big) \\
&\quad -(\ell-K-1)\psi_K(\tfrac{\ell}{2}) + 2ln\,\Gamma_K(\tfrac{\ell}{2}) + K\ell \\
&\Longrightarrow \quad (\sum_{i:1}^{K}\psi(\tfrac{\ell-i+1}{2}))\Big(K - \cancel{K} - \cancel{1} + N - l + \cancel{K} + \cancel{1}\Big) + 2(\tfrac{K(K-1)}{4}ln\,\pi + \sum_{i:1}^{K} ln\,\Gamma(\tfrac{\ell-i+1}{2})) \\
&\quad +\ell(K - Tr\,\Big[L\Big(KI + \sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\}) \\
&\propto \quad (\sum_{i:1}^{K}\psi(\tfrac{\ell-i+1}{2}))\Big(K + N - \ell\Big) + 2(\sum_{i:1}^{K} ln\,\Gamma(\tfrac{\ell-i+1}{2})) \\
&\quad +\ell(K - Tr\,\Big[L\Big(KI + \sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\}) \\
&\Longrightarrow \\
\nabla_\ell \mathcal{L}_\ell \quad &= \quad \tfrac{1}{2}(\sum_{i:1}^{K}\psi'(\tfrac{\ell-i+1}{2}))\Big(K + N - \ell\Big) \\
&\quad +K - Tr\,\Big[L\Big(KI + \sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\} = 0 \\
&\quad \text{simultaenously with } L \\
&\Longrightarrow \quad L\Big(KI + \sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big) = I_K \\
&hence, \quad K - Tr\,I_K = 0 \\
&\Longrightarrow \\
&\quad \boxed{\ell = K + N}
\end{aligned}
$$

In minibatch node sampling this would be

$$\boxed{\ell = K + \#mbnodes}$$

### 4.5 Gradient with respect to $b_k$

$$\begin{aligned}
\mathcal{L}_{b_k} &= (\eta_0-1)\psi(b_{k0}) + -(\eta_0-2)\psi(b_{k0}+b_{k1}) \\
&\quad + \sum_{a,b\in link} \phi_{a\to b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k0})-\psi(b_{k0}+b_{k1})\big) \\
&\quad + \sum_{a,b\notin link} \phi_{a\to b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k1})-\psi(b_{k0}+b_{k1})\big) \\
&\quad + ln\,\Gamma(b_{k0}) + ln\,\Gamma(b_{k1}) - ln\,\Gamma(b_{k0}+b_{k1}) - (b_{k0}-1)\psi(b_{k0}) \\
&\quad - (b_{k1}-1)\psi(b_{k1}) + (b_{k0}+b_{k1}-2)\psi(b_{k0}+b_{k1}) \\
&\quad \text{simultaenously optimizing } b_{k0}, b_{k1} \\
&\Longrightarrow \text{Similar to our previous results} \\
\nabla_{b_{k0}}\mathcal{L}_{b_k} &= 0 \\
&\Longrightarrow \boxed{b_{k0} = \eta_0 + \sum_{a,b\in mblinks} \phi_{a\to b,k}\phi_{a\leftarrow b,k}} \\
\nabla_{b_{k1}}\mathcal{L}_{b_k} &= 0 \\
& \boxed{b_{k1} = 1 + \sum_{a,b\notin mblinks} \phi_{a\to b,k}\phi_{a\leftarrow b,k}}
\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\mathcal{L}_{\mu_{a,k}} &= -\tfrac{\ell}{2}\big[(\mu_{a,k}-m_k)^T L_{kk}(\mu_{a,k}-m_k) + \\
&\quad \sum_{b\in sink(a)} \phi_{a\to b,k}\mu_{a,k} + \\
&\quad \sum_{b\notin sink(a)} \phi_{a\to b,k}\mu_{a,k} + \\
&\quad \sum_{b\in source(a)} \phi_{b\leftarrow a,k}\mu_{a,k} + \\
&\quad \sum_{b\notin source(a)} \phi_{b\leftarrow a,k}\mu_{a,k} - \\
&\quad \sum_b \zeta_a^{-1}\sum_l exp(\mu_{a,l}+\tfrac{1}{2}diag(\Lambda_a^{-1})_{ll}) \\
&=
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\mu_{a,k}}\mathcal{L}_{\mu_{a,k}} &= -\ell L_{kk}(\mu_{a,k}-m_k) + \sum_b(\phi_{a\to b,k}+\phi_{b\leftarrow a,k}) - \sum_b \zeta_a^{-1} exp(\mu_{a,k}+\tfrac{1}{2}diag(\Lambda_{a,k}^{-1})) \\
& Also \\
\mathcal{L}_{\zeta_a} &= -\zeta_a^{-1}\sum_l exp(\mu_{a,l}+\tfrac{1}{2}diag(\Lambda_a^{-1})_{ll}) - ln\,\zeta_a \\
\nabla_{\zeta_a}\mathcal{L}_{\zeta_a} &= -\zeta_a^{-2}\sum_l exp(\mu_{a,l}+\tfrac{1}{2}diag(\Lambda_a^{-1})_{ll}) - \zeta_a^{-1} = 0 \\
&= \boxed{\sum_l exp(\mu_{a,l}+\tfrac{1}{2}diag(\Lambda_a^{-1})_{ll}) = \zeta_a} \\
So \quad & \zeta_a^{-1}exp(\mu_{a,k}+\tfrac{1}{2}diag(\Lambda_{a,k}^{-1})) = softmax(\mu_{a,k}+\tfrac{1}{2}diag(\Lambda_{a,k}^{-1})) \\
\nabla_{\mu_{a,k}}\mathcal{L}_{\mu_{a,k}} &\propto \boxed{-\ell L_{kk}(\mu_{a,k}-m_k) + \sum_b(\phi_{a\to b,k}+\phi_{b\leftarrow a,k}) - \sum_b softmax(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1})} \\
\nabla^2_{\mu_{a,k}}\mathcal{L}_{\mu_{a,k}} &\propto \boxed{-\ell L_{kk} - \sum_b \Big(softmax(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1}) - (softmax(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1}))^2\Big)}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{L}_{\Lambda_{a,k}} = & -\tfrac{\ell}{2}tr\,(L\Lambda_a^{-1}) - \tfrac{1}{2}ln\,|\Lambda_a| - \sum_b \zeta_a^{-1}exp(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1}) \\
= & -\tfrac{\ell}{2}(L_{kk}\Lambda_{a,k}^{-1}) - \tfrac{1}{2}\sum_{j!=k} ln\,\Lambda_{a,j} - \tfrac{1}{2}ln\,\Lambda_{a,k} - \sum_b \zeta_a^{-1}exp(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1}) \\
\nabla_{\Lambda_{a,k}}\mathcal{L}_{\Lambda_{a,k}} = & \boxed{+\tfrac{\ell}{2}L_{kk}\Lambda_{a,k}^{-2} - \tfrac{1}{2}\Lambda_{a,k}^{-1} + \tfrac{1}{2}\sum_b \zeta_a^{-1}\Lambda_{a,k}^{-2}exp(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1})} \\
or \propto & \boxed{\tfrac{\ell}{2}L_{kk} - \tfrac{1}{2}\Lambda_{a,k} + \tfrac{1}{2}\sum_b softmax(\mu_{a,k}+\tfrac{1}{2}\Lambda_{a,k}^{-1})} \\
\nabla^2_{\Lambda_{a,k}}\mathcal{L}_{\Lambda_{a,k}} \propto & \boxed{-\tfrac{1}{2} - \tfrac{1}{4}\sum_b \Lambda_{a,k}^{-2}\Big(softmax(.) - softmax^2(.)\Big)}
\end{aligned}
$$