Proceedings of Machine Learning Research 129:769–774, 2020 ACML 2020

# Scalable Inference on the Soft Affiliation Graph Model for Overlapping Community Detection

**Nishma Laitonjam** NISHMA.LAITONJAM@INSIGHT-CENTRE.ORG
**Wěipéng Huáng** WEIPENG.HUANG@INSIGHT-CENTRE.ORG
**Neil J. Hurley** NEIL.HURLEY@INSIGHT-CENTRE.ORG
*Insight Centre for Data Analytics, University College Dublin, Ireland*

**Editors:** Sinno Jialin Pan and Masashi Sugiyama

## 1. ELBO for S-AGM

ELBO for S-AGM is given as

$$\begin{aligned}
\mathcal{L}(q) &= \mathrm{E}_q[\log p(\mathrm{A}, \mathrm{W}, \alpha, \pi|\eta, \beta)] - \mathrm{E}_q[\log q(\mathrm{W}, \pi, \alpha)] \\
&= \sum_{ij:i<j} \mathrm{E}_q[\log p(a_{ij}|w_i, w_j, \pi)] \\
&+ \sum_i \sum_k \mathrm{E}_q[\log p(w_{ik}|\alpha_k)] + \sum_k \mathrm{E}_q[\log p(\pi_k|\eta_{k0}, \eta_{k1})] + \sum_k \mathrm{E}_q[\log p(\alpha_k|\beta_0, \beta_1)] \\
&- \sum_i \sum_k \mathrm{E}_q[\log q(w_{ik}|\phi_{ik0}, \phi_{ik1})] - \sum_k \mathrm{E}_q[\log q(\pi_k|\lambda_{k0}, \lambda_{k1})] - \sum_k \mathrm{E}_q[\log q(\alpha_k|\tau_{k0}, \tau_{k1})]
\end{aligned}$$

Here, the term

$$\mathrm{E}_q[\log p(a_{ij}|w_i, w_j, \pi)] = a_{ij}\mathrm{E}_q[\log p_{ij}] + (1 - a_{ij})\mathrm{E}_q[\log(1 - p_{ij})]\,,$$

does not have an analytic form.

## 2. Computation of Gradients for SG-VI

In this section, the gradients i.e. $\hat{g}(\overline{\phi}_{ik0})$, $\hat{g}(\overline{\phi}_{ik1})$, $\hat{g}(\overline{\lambda}_{k0})$, $\hat{g}(\overline{\lambda}_{k1})$, $\hat{g}(\tau_{k0})$ and $\hat{g}(\tau_{k1})$ given in Algorithm 1 are computed.

### 2.1. Computation of $\hat{g}(\overline{\phi}_{ik0})$ and $\hat{g}(\overline{\phi}_{ik1})$

For $m \in \{0, 1\}$, we have $\hat{g}(\overline{\phi}_{ikm})$ representing the unbiased estimate of the gradient of the ELBO wrt $\overline{\phi}_{ikm}$,

$$\hat{g}(\overline{\phi}_{ikm}) = \nabla_{\overline{\phi}_{ikm}} \mathcal{L}_{\phi_{ik}}(q) = \hat{g}(\phi_{ikm}) \times \frac{1}{1 + \exp(-\overline{\phi}_{ikm})}$$

**Algorithm 1** VI for the S-AGM using SGD-VI at iteration $t$

Sample a mini-batch $\mathcal{E}^t$ of node pairs.
**for** Each node $i$ in $\mathcal{E}^t$ **do**
  Sample a mini-batch of nodes $\mathcal{V}_i^t$.
  **for** $k = 1 : K$ **do** ▷ utilizing the sampled $\mathcal{V}_i^t$
    $\overline{\phi}_{ik0}^{(t)} = \overline{\phi}_{ik0}^{(t-1)} + \rho_\phi^{(t)} \times \hat{g}(\overline{\phi}_{ik0})$
    $\overline{\phi}_{ik1}^{(t)} = \overline{\phi}_{ik1}^{(t-1)} + \rho_\phi^{(t)} \times \hat{g}(\overline{\phi}_{ik1})$
**for** $k = 1 : K$ **do** ▷ utilizing the sampled $\mathcal{E}^t$
  $\overline{\lambda}_{k0}^{(t)} = \overline{\lambda}_{k0}^{(t-1)} + \rho_\lambda^{(t)} \times \hat{g}(\overline{\lambda}_{k0})$
  $\overline{\lambda}_{k1}^{(t)} = \overline{\lambda}_{k1}^{(t-1)} + \rho_\lambda^{(t)} \times \hat{g}(\overline{\lambda}_{k1})$
**for** $k = 1 : K$ **do** ▷ utilizing the sampled $\mathcal{E}^t$
  $\tau_{k0}^{(t)} = \tau_{k0}^{(t-1)} + \rho_\tau^{(t)} \times \hat{g}(\tau_{k0})$
  $\tau_{k1}^{(t)} = \tau_{k1}^{(t-1)} + \rho_\tau^{(t)} \times \hat{g}(\tau_{k1})$

where $\hat{g}(\phi_{ikm})$ is gradient of ELBO corresponding to $\phi_{ikm}$ i.e. corresponding gradient wrt $\phi$ with preconditioning matrix $G(\phi) = \text{diag}(\phi)^{-1}$ where $\phi = \{\phi_{ik0}, \phi_{ik1}\}_{i=1,\dots,N;k=1,\dots,K}$, and $\mathcal{L}_{\phi_{ik}}(q)$ is an expression which is proportional to $\mathcal{L}(q)$ and dependent on $\phi_{ik0}$ and $\phi_{ik1}$. We have

$$\hat{g}(\phi_{ikm}) = (G^{-1}(\phi)\nabla_\phi \mathcal{L}(q))_{ikm} = \phi_{ikm}\nabla_{\phi_{ikm}}\mathcal{L}_{\phi_{ik}}(q)$$

where,

$$\begin{aligned}
\mathcal{L}_{\phi_{ik}}(q) &= \sum_{i\neq j} \mathrm{E}_q[\log p(a_{ij}|w_i, w_j, \pi)] + \mathrm{E}_q[\log p(w_{ik}|\alpha_k)] - \mathrm{E}_q[\log q(w_{ik}|\phi_{ik0}, \phi_{ik1})] \\
&= \mathrm{E}_{q(w_{ik}|\phi)}\left[\sum_{i\neq j} \mathrm{E}_{q^{-w_{ik}}}[\log p(a_{ij}|w_i, w_j, \pi)] + \mathrm{E}_{q(\alpha_k|\tau)}[\log p(w_{ik}|\alpha_k)]\right] \\
&\quad - \mathrm{E}_q[\log q(w_{ik}|\phi_{ik0}, \phi_{ik1})] \\
&= \mathrm{E}_{q(w_{ik}|\phi)}\left[\sum_{i\neq j} \mathrm{E}_{q^{-w_{ik}}}[\log p(a_{ij}|w_i, w_j, \pi)] + \Psi(\tau_{k0}) - \log(\tau_{k1}) + \left(\frac{\tau_{k0}}{\tau_{k1}} - 1\right)\log w_{ik}\right] \\
&\quad - \mathrm{E}_q[\log q(w_{ik}|\phi_{ik0}, \phi_{ik1})]
\end{aligned}$$

We can rewrite it as

$$\begin{aligned}
\mathcal{L}_{\phi_{ik}}(q) = \mathrm{E}_{q(w_{ik}|\phi)}[\hat{f}(w_{ik})] - \Big( &\log\Gamma(\phi_{ik0} + \phi_{ik1}) - \log\Gamma(\phi_{ik0}) - \log\Gamma(\phi_{ik1}) \\
&+ (\phi_{ik0} - 1)(\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) + (\phi_{ik1} - 1)(\Psi(\phi_{ik1}) - \Psi(\phi_{ik0} + \phi_{ik1}))\Big)
\end{aligned}$$

For a noisy gradient considering mini-batch of data i.e. considering only $j \in \mathcal{V}_i$, we have

$$\hat{f}(w_{ik}) = \frac{N}{|j \in \mathcal{V}_i^t|} \sum_{j\in\mathcal{V}_i} \mathrm{E}_{q^{-w_{ik}}}[\log p(a_{ij}|w_i, w_j, \pi)] + \Psi(\tau_{k0}) - \log(\tau_{k1}) + \left(\frac{\tau_{k0}}{\tau_{k1}} - 1\right)\log w_{ik}$$

#### 2.1.1. Partial derivative of ELBO wrt $\phi_{ikm}$

$$\begin{aligned}\nabla_{\phi_{ikm}}\mathcal{L}_{\phi_{ik}}(q) &= \nabla_{\phi_{ikm}}\mathrm{E}_{q(w_{ik}|\phi)}[\hat{f}(w_{ik})] \\ &\quad - (\phi_{ikm}-1)\Psi'(\phi_{ikm}) + (\phi_{ik0}+\phi_{ik1}-2)\Psi'(\phi_{ik0}+\phi_{ik1}) \end{aligned} \tag{1}$$

**Computation of $\nabla_{\phi_{ikm}}\mathrm{E}_{q(w_{ik}|\phi)}[\hat{f}(w_{ik})]$ :** For a random variable $w_{ik} \sim \mathrm{Beta}(\phi_{ik0}, \phi_{ik1})$, we could rewrite $w_{ik} = \frac{w_{ik0}}{w_{ik0}+w_{ik1}}$ where $w_{ik0} \sim \mathrm{Gamma}(\phi_{ik0}, 1)$ and $w_{ik1} \sim \mathrm{Gamma}(\phi_{ik1}, 1)$. And setting $w_{ik0} = \exp(\epsilon_{ik0}\sqrt{\Psi'(\phi_{ik0})} + \Psi(\phi_{ik0}))$ and $w_{ik1} = \exp(\epsilon_{ik1}\sqrt{\Psi'(\phi_{ik1})} + \Psi(\phi_{ik1}))$, we can write

$$\nabla_{\phi_{ikm}}\mathrm{E}_{q(w_{ik}|\phi)}[\hat{f}(w_{ik})] = \nabla_{\phi_{ikm}}\mathrm{E}_{q(w_{ik0},w_{ik1}|\phi)}[\hat{f}(w_{ik0}, w_{ik1})] = g^{rep}_{\phi_{ikm}} + g^{corr}_{\phi_{ikm}}$$

where

$$\begin{aligned} g^{rep}_{\phi_{ikm}} &= \mathrm{E}_{q(w_{ik0},w_{ik1};\phi)}\left[\nabla_{w_{ikm}}\hat{f}(w_{ik0}, w_{ik1}) \times \nabla_{\phi_{ikm}} \exp\left(\epsilon_{ikm}\sqrt{\Psi'(\phi_{ikm})} + \Psi(\phi_{ikm})\right)\right] \\ &= \mathrm{E}_{q(w_{ik0},w_{ik1};\phi)}\left[\nabla_{w_{ik}}\hat{f}(w_{ik}) \times (-1)^m w_{ik}(1-w_{ik})\left((\log(w_{ikm}) - \Psi(\phi_{ikm})) \times \frac{\Psi''(\phi_{ikm})}{2\Psi'(\phi_{ikm})} + \Psi'(\phi_{ikm})\right)\right] \end{aligned}$$

and

$$\begin{aligned} g^{corr}_{\phi_{ikm}} &= \mathrm{E}_{q(w_{ik0},w_{ik1};\phi)}\left[\hat{f}(w_{ik0}, w_{ik1})\left\{\nabla_{w_{ikm}} \log q(w_{ikm};\phi_{ikm})\left(\nabla_{\phi_{ikm}} \exp\left(\epsilon_{ikm}\sqrt{\Psi'(\phi_{ikm})} + \Psi(\phi_{ikm})\right)\right)\right.\right. \\ &\quad \left.\left. + \nabla_{\phi_{ikm}} \log q(w_{ikm};\phi_{ikm}) + \nabla_{\phi_{ikm}} \log J(\epsilon_{ikm};\phi_{ikm})\right\}\right] \\ &= \mathrm{E}_{q(w_{ik0},w_{ik1};\phi)}\left[\hat{f}(w_{ik0}, w_{ik1})\left\{(\phi_{ikm} - w_{ikm})\left((\log(w_{ikm}) - \Psi(\phi_{ikm})) \times \frac{\Psi''(\phi_{ikm})}{2\Psi'(\phi_{ikm})} + \Psi'(\phi_{ikm})\right)\right.\right. \\ &\quad \left.\left. + \log(w_{ikm}) - \Psi(\phi_{ikm}) + \frac{\Psi''(\phi_{ikm})}{2\Psi'(\phi_{ikm})}\right\}\right] \end{aligned}$$

where

$$\begin{aligned} J(\epsilon_{ikm};\phi_{ikm}) &= \left|\det\nabla_{\epsilon_{ikm}} \exp\left(\epsilon_{ikm}\sqrt{\Psi'(\phi_{ikm})} + \Psi(\phi_{ikm})\right)\right| \\ &= \exp\left(\epsilon_{ikm}\sqrt{\Psi'(\phi_{ikm})} + \Psi(\phi_{ikm})\right)\sqrt{\Psi'(\phi_{ikm})} \end{aligned}$$

To compute $g^{rep}_{\phi_{ikm}}$, the term $\nabla_{w_{ik}}\hat{f}(w_{ik})$ is derived as

$$\nabla_{w_{ik}}\hat{f}(w_{ik}) = \frac{N}{|j \in \mathcal{V}_i^t|}\sum_{j\in\mathcal{V}_i^t} \mathrm{E}_{q^{-w_{ik}}}\left[h_{ij}\frac{\pi_k w_{jk}}{1-\pi_k w_{ik} w_{jk}}\right] + \frac{\frac{\tau_{k0}}{\tau_{k1}} - 1}{w_{ik}}$$

where

$$h_{ij} = (-1)^{1-a_{ij}}\left(p_{ij}^{-1} - 1\right)^{a_{ij}} .$$

Putting the values of $\nabla_{\phi_{ikm}}\mathrm{E}_{q(w_{ik}|\phi)}[\hat{f}(w_{ik})]$ in Equation (1), we get $\nabla_{\phi_{ikm}}\mathcal{L}_{\phi_{ik}}(q)$ which is used to compute $\hat{g}(\overline{\phi}_{ikm})$.

Here, $\Gamma$, $\Psi$, $\psi'$ and $\Psi''$ are gamma function, digamma function, polygamma of order 1 and polygamma of order 2 respectively.

### 2.2. Computation of $\hat{g}(\overline{\lambda}_{k0})$ and $\hat{g}(\overline{\lambda}_{k1})$

For $m \in \{0,1\}$, we have $\hat{g}(\overline{\lambda}_{km})$ representing the unbiased estimate of the gradient of the ELBO wrt $\lambda_{km}$ which is given by

$$\hat{g}(\overline{\lambda}_{km}) = \nabla_{\overline{\lambda}_{km}} \mathcal{L}_{\lambda_k}(q) = \hat{g}(\lambda_{km}) \times \frac{1}{1+\exp(-\overline{\lambda}_{km})}$$

where $\hat{g}(\lambda_{km})$ is gradient of ELBO corresponding to $\lambda_{km}$ i.e. corresponding gradient wrt $\lambda$ with preconditioning matrix $G(\lambda) = \mathrm{diag}(\lambda)^{-1}$ where $\lambda = \{\lambda_{k0}, \lambda_{k1}\}_{k=1,\dots,K}$, and $\mathcal{L}_{\lambda_k}(q)$ is an expression which is proportional to $\mathcal{L}(q)$ and dependent on $\lambda_{k0}$ and $\lambda_{k1}$. We have

$$\hat{g}(\lambda_{km}) = (G^{-1}(\lambda)\nabla_\lambda \mathcal{L}(q))_{km} = \lambda_{km}\nabla_{\lambda_{km}}\mathcal{L}_{\lambda_k}(q)$$

where

$$\begin{aligned}
\mathcal{L}_{\lambda_k}(q) &= \sum_{ij:i<j} \mathrm{E}_q[\log p(a_{ij}|w_i,w_j,\pi)] + \mathrm{E}_q[\log p(\pi_k|\eta_{k0},\eta_{k1})] - \mathrm{E}_q[\log q(\pi_k|\lambda_{k0},\lambda_{k1})] \\
&= \mathrm{E}_{q(\pi_k|\lambda_{k0},\lambda_{k1})}\Big[\sum_{ij:i<j} \mathrm{E}_{q^{-\pi_k}}[\log p(a_{ij}|w_i,w_j,\pi)] + \log p(\pi_k|\eta_{k0},\eta_{k1})\Big] \\
&- \mathrm{E}_q[\log q(\pi_k|\lambda_{k0},\lambda_{k1})] \\
&= \mathrm{E}_{q(\pi_k|\lambda_{k0},\lambda_{k1})}\Big[\sum_{ij:i<j} \mathrm{E}_{q^{-\pi_k}}[\log p(a_{ij}|w_i,w_j,\pi)] \\
&+ \log\Gamma(\eta_{k0}+\eta_{k1}) - \log\Gamma(\eta_{k0}) - \log\Gamma(\eta_{k1}) + (\eta_{k0}-1)\log\pi_k + (\eta_{k1}-1)\log(1-\pi_k)\Big] \\
&- \mathrm{E}_q[\log q(\pi_k|\lambda_{k0},\lambda_{k1})]
\end{aligned}$$

We can be rewrite it as

$$\begin{aligned}
\mathcal{L}_{\lambda_k}(q) = \mathrm{E}_{q(\pi_k|\lambda_{k0},\lambda_{k1})}[\hat{f}(\pi_k)] - \Big(&\log\Gamma(\lambda_{k0}+\lambda_{k1}) - \log\Gamma(\lambda_{k0}) - \log\Gamma(\lambda_{k1}) \\
&+ (\lambda_{k0}-1)(\Psi(\lambda_{k0}) - \Psi(\lambda_{k0}+\lambda_{k1})) + (\lambda_{k1}-1)(\Psi(\lambda_{k1}) - \Psi(\lambda_{k0}+\lambda_{k1}))\Big)
\end{aligned}$$

For a noisy gradient considering mini-batch of data i.e. considering only $(i,j) \in \mathcal{E}^t$, we have

$$\begin{aligned}
\hat{f}(\pi_k) &= s(\mathcal{E}^t) \sum_{(i,j)\in\mathcal{E}^t} \mathrm{E}_{q^{-\pi_k}}[\log p(a_{ij}|w_i,w_j,\pi)] \\
&\quad + \log\Gamma(\eta_{k0}+\eta_{k1}) - \log\Gamma(\eta_{k0}) - \log\Gamma(\eta_{k1}) + (\eta_{k0}-1)\log\pi_k + (\eta_{k1}-1)\log(1-\pi_k)
\end{aligned}$$

Here, $s(\mathcal{E}^t)$ is scale value for stratified sampling.

#### 2.2.1. Partial derivative of ELBO wrt $\lambda_{km}$

$$\begin{aligned}
\nabla_{\lambda_{km}}\mathcal{L}_{\lambda_k}(q) &= \nabla_{\lambda_{km}} \mathrm{E}_{q(\pi_k|\lambda)}[\hat{f}(\pi_k)] \\
&\quad - (\lambda_{km}-1)\Psi'(\lambda_{km}) + (\lambda_{k0}+\lambda_{k1}-2)\Psi'(\lambda_{k0}+\lambda_{k1}) \qquad (2)
\end{aligned}$$

**Computation of $\nabla_{\lambda_{km}} \mathrm{E}_{q(\pi_k|\lambda_{k0},\lambda_{k1})}[\hat{f}(\pi_k)]$ :** For a random variable $\pi_k \sim \text{Beta}(\lambda_{k0}, \lambda_{k1})$, we could rewrite $\pi_k = \frac{\pi_{k0}}{\pi_{k0}+\pi_{k1}}$ where $\pi_{k0} \sim \text{Gamma}(\lambda_{k0}, 1)$ and $\pi_{k1} \sim \text{Gamma}(\lambda_{k1}, 1)$. And setting $\pi_{k0} = \exp(\epsilon_{k0}\sqrt{\Psi'(\lambda_{k0})} + \Psi(\lambda_{k0}))$ and $\pi_{k1} = \exp(\epsilon_{k1}\sqrt{\Psi'(\lambda_{k1})} + \Psi(\lambda_{k1}))$, we can write

$$\nabla_{\lambda_{km}} \mathrm{E}_{q(\pi_k|\lambda)}[\hat{f}(\pi_k)] = \nabla_{\lambda_{km}} \mathrm{E}_{q(\pi_{k0},\pi_{k1}|\lambda)}[\hat{f}(\pi_{k0}, \pi_{k1})] = g^{rep}_{\lambda_{km}} + g^{corr}_{\lambda_{km}}$$

where

$$\begin{aligned} g^{rep}_{\lambda_{km}} &= \mathrm{E}_{q_{\pi_{km}}(\pi_{k0},\pi_{k1};\lambda)}\left[\nabla_{\pi_{km}} \hat{f}(\pi_{k0}, \pi_{k1}) \times \nabla_{\lambda_{km}} \exp(\epsilon_{km}\sqrt{\Psi'(\lambda_{km})} + \Psi(\lambda_{km}))\right] \\ &= \mathrm{E}_{q_{\pi_{km}}(\pi_{k0},\pi_{k1};\lambda)}\left[\nabla_{\pi_k} \hat{f}(\pi_k) \times (-1)^m \pi_k(1-\pi_k)\left((\log(\pi_{km}) - \Psi(\lambda_{km})) \times \frac{\Psi''(\lambda_{km})}{2\Psi'(\lambda_{km})} + \Psi'(\lambda_{km})\right)\right] \end{aligned}$$

and

$$\begin{aligned} g^{corr}_{\lambda_{km}} &= \mathrm{E}_{q_{\pi_{km}}(\pi_{k0},\pi_{k1};\lambda)}\Bigg[\hat{f}(\pi_{k0}, \pi_{k1})\Bigg\{\nabla_{\pi_{km}} \log q(\pi_{km}; \lambda_{km})\left(\nabla_{\lambda_{km}} \exp(\epsilon_{km}\sqrt{\Psi'(\lambda_{km})} + \Psi(\lambda_{km}))\right) \\ &\quad + \nabla_{\lambda_{km}} \log q(\pi_{km}; \lambda_{km}) + \nabla_{\lambda_{km}} \log J(\epsilon_{km}; \lambda_{km}) \\ &= \mathrm{E}_{q_{\pi_{km}}(\pi_{k0},\pi_{k1};\lambda)}\Bigg[\hat{f}(\pi_{k0}, \pi_{k1})\Bigg\{(\lambda_{km} - \pi_{km})\left((\log(\pi_{km}) - \Psi(\lambda_{km})) \times \frac{\Psi''(\lambda_{km})}{2\Psi'(\lambda_{km})} + \Psi'(\lambda_{km})\right) \\ &\quad + \log(\pi_{km}) - \Psi(\lambda_{km}) + \frac{\Psi''(\lambda_{km})}{2\Psi'(\lambda_{km})}\Bigg\}\Bigg] \end{aligned}$$

where

$$\begin{aligned} J(\epsilon_{km}; \lambda_{km}) &= |\det \nabla_{\epsilon_{km}} \exp(\epsilon_{km}\sqrt{\Psi'(\lambda_{km})} + \Psi(\lambda_{km}))| \\ &= \exp(\epsilon_{km}\sqrt{\Psi'(\lambda_{km})} + \Psi(\lambda_{km}))\sqrt{\Psi'(\lambda_{km})} \end{aligned}$$

To compute $g^{rep}_{\lambda_{km}}$, the term $\nabla_{\pi_k} \hat{f}(\pi_k)$ is derived as

$$\nabla_{\pi_k} \hat{f}(\pi_k) = s(\mathcal{E}^t) \sum_{(i,j)\in\mathcal{E}^t} \mathrm{E}_{q^{-\pi_k}}\left[h_{ij} \frac{w_{ik}w_{jk}}{1 - \pi_k w_{ik} w_{jk}}\right] + \frac{\eta_{k0} - 1}{\pi_k} - \frac{\eta_{k1} - 1}{1 - \pi_k}$$

Putting the values of $\nabla_{\lambda_{km}} \mathrm{E}_{q(\pi_k|\lambda_{k0},\lambda_{k1})}[\hat{f}(\pi_k)]$ in Equation (2), we get $\nabla_{\lambda_{km}} \mathcal{L}_{\lambda_k}(q)$ which is used to compute $\hat{g}(\overline{\lambda}_{km})$.

### 2.3. Computation of $\hat{g}(\tau_{k0})$ and $\hat{g}(\tau_{k1})$

For $m \in \{0, 1\}$, we have $\hat{g}(\tau_{km})$ representing the unbiased estimate of natural gradient of the ELBO wrt $\tau_{km}$ which is given by

$$\hat{g}(\tau_{km}) = (G^{-1}(\tau_{k0}, \tau_{k1})\nabla_{\tau_k} \mathcal{L}_{\tau_k}(q))_m$$

which is natural gradient corresponding to $\tau_{km}$ i.e. corresponding gradient of the ELBO wrt $\tau_k$ with Fisher Information matrix as preconditioning matrix $G(\tau_{k0}, \tau_{k1})$, and $\mathcal{L}_{\tau_k}(q)$ is proportional to $\mathcal{L}(q)$ and dependent on $\tau_{k0}$ and $\tau_{k1}$.

$$
\begin{aligned}
\mathcal{L}_{\tau_k}(q) &= \sum_i \mathrm{E}_q[\log p(w_{ik}|\alpha_k)] + \mathrm{E}_q[\log p(\alpha_k|\beta_0, \beta_1)] - \mathrm{E}_q[\log q(\alpha_k|\tau_{k0}, \tau_{k1})] \\
&\propto (N + \beta_0 - \tau_{k0})(\Psi(\tau_{k0}) - \log(\tau_{k1})) \\
&- (\beta_1 - \sum_i (\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) - \tau_{k1})\frac{\tau_{k0}}{\tau_{k1}} + \Big( \log \Gamma(\tau_{k0}) - \tau_{k0} \log(\tau_{k1}) \Big)
\end{aligned}
$$

### 2.3.1. Partial derivative of ELBO wrt $\tau_{k0}$ and $\tau_{km}$

For a noisy gradient considering mini-batch of data i.e. considering only $i \in \mathcal{E}^t$, we have

$$
\nabla_{\tau_{k0}} \mathcal{L}_{\tau_k}(q) = \begin{bmatrix} \Psi'(\tau_{k0}) & \frac{-1}{\tau_{k1}} \end{bmatrix} \begin{bmatrix} N + \beta_0 - \tau_{k0} \\ \beta_1 - \frac{N}{|i \in \mathcal{E}^t|} \sum_{i \in \mathcal{E}^t} (\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) - \tau_{k1} \end{bmatrix}
$$

$$
\nabla_{\tau_{k1}} \mathcal{L}_{\tau_k}(q) = \begin{bmatrix} \frac{-1}{\tau_{k1}} & \frac{\tau_{k0}}{\tau_{k1}^2} \end{bmatrix} \begin{bmatrix} N + \beta_0 - \tau_{k0} \\ \beta_1 - \frac{N}{|i \in \mathcal{E}^t|} \sum_{i \in \mathcal{E}^t} (\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) - \tau_{k1} \end{bmatrix}
$$

Now,

$$
\begin{aligned}
\nabla_{\tau_k} \mathcal{L}_{\tau_k}(q) &= \begin{bmatrix} \nabla_{\tau_{k0}} \mathcal{L}_{\tau_k}(q) \\ \nabla_{\tau_{k1}} \mathcal{L}_{\tau_k}(q) \end{bmatrix} \\
&= \begin{bmatrix} \Psi'(\tau_{k0}) & \frac{-1}{\tau_{k1}} \\ \frac{-1}{\tau_{k1}} & \frac{\tau_{k0}}{\tau_{k1}^2} \end{bmatrix} \begin{bmatrix} N + \beta_0 - \tau_{k0} \\ \beta_1 - \frac{N}{|i \in \mathcal{E}^t|} \sum_{i \in \mathcal{E}^t} (\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) - \tau_{k1} \end{bmatrix}
\end{aligned}
$$

### 2.3.2. Noisy natural gradients, $\hat{g}(\tau_{k0})$ and $\hat{g}(\tau_{k1})$

Using the Fisher information matrix for Gamma$(\tau_{k0}, \tau_{k1})$, i.e. $G(\tau_{k0}, \tau_{k1}) = \begin{bmatrix} \Psi'(\tau_{k0}) & \frac{-1}{\tau_{k1}} \\ \frac{-1}{\tau_{k1}} & \frac{\tau_{k0}}{\tau_{k1}^2} \end{bmatrix}$, the natural gradients are given as

$$
\begin{aligned}
\begin{bmatrix} \hat{g}(\tau_{k0}) \\ \hat{g}(\tau_{k1}) \end{bmatrix} &= G^{-1}(\tau_{k0}, \tau_{k1}) \nabla_{\tau_k} \mathcal{L}_{\tau_k}(q) \\
&= \begin{bmatrix} N + \beta_0 - \tau_{k0} \\ \beta_1 - \frac{N}{|i \in \mathcal{E}^t|} \sum_{i \in \mathcal{E}^t} (\Psi(\phi_{ik0}) - \Psi(\phi_{ik0} + \phi_{ik1})) - \tau_{k1} \end{bmatrix}
\end{aligned}
$$