Probabilistic task modelling for meta-learning

Cuong C. Nguyen[1] **Thanh-Toan Do**[2] **Gustavo Carneiro**[1]

[1]Australian Institute for Machine Learning, University of Adelaide, Australia
[2]Department of Data Science and AI, Monash University, Australia

A CALCULATION OF EACH TERM IN THE ELBO

As described in section 3, the variational distributions for $\mathbf{u}, \mathbf{z}$ and $\boldsymbol{\pi}$ are:

$$q(\mathbf{u}_{in}; \phi) = \mathcal{N}\left(\mathbf{u}_{in}; \mathbf{m}_{in}, (\mathbf{s}_{in})^2 \, \mathbf{I}\right) \tag{4}$$

$$q(\boldsymbol{\pi}_i; \boldsymbol{\gamma}_i) = \mathrm{Dirichlet}\left(\boldsymbol{\pi}_i; \boldsymbol{\gamma}_i\right) \tag{10}$$

$$q(\mathbf{z}_{in}; \mathbf{r}_{in}) = \mathrm{Categorical}\left(\mathbf{z}_{in}; \mathbf{r}_{in}\right). \tag{11}$$

A.1 $\mathbb{E}_{q(\mathbf{u}_i; \boldsymbol{\mu}_{u_i}, \boldsymbol{\Sigma}_{u_i})} \mathbb{E}_{q(\mathbf{z}_i, \boldsymbol{\pi}_i)} \left[\ln p(\mathbf{u}_i | \mathbf{z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right]$

$$\begin{aligned} \mathbb{E}_{q(\mathbf{z}_i, \boldsymbol{\pi}_i)} \left[\ln p(\mathbf{u}_i | \mathbf{z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right] & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \ln p(\mathbf{u}_{in} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \ln \mathcal{N}(\mathbf{u}_{in} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned} \tag{16}$$

Hence:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{u}_i; \boldsymbol{\mu}_{u_i}, \boldsymbol{\Sigma}_{u_i})} \mathbb{E}_{q(\mathbf{z}_i, \boldsymbol{\pi}_i)} \left[\ln p(\mathbf{u}_i | \mathbf{z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right] & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \underbrace{\mathbb{E}_{q(\mathbf{u}_i; \boldsymbol{\mu}_{u_i}, \boldsymbol{\Sigma}_{u_i})} \left[\ln \mathcal{N}(\mathbf{u}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]}_{\text{cross-entropy between 2 Gaussians}} \\ & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \left[-\frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\Sigma}_{u_{in}}) + \ln \mathcal{N}(\boldsymbol{\mu}_{u_{in}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]. \end{aligned} \tag{17}$$

A.2 $\mathbb{E}_{q(\mathbf{u}_i; \boldsymbol{\mu}_{u_i}, \boldsymbol{\Sigma}_{u_i})} \mathbb{E}_{q(\mathbf{z}_i, \boldsymbol{\pi}_i)} \left[\ln p(\mathbf{z}_i | \boldsymbol{\pi}_i) \right]$

$$\begin{aligned} \mathbb{E}_{q(\mathbf{u}_i; \boldsymbol{\mu}_{u_i}, \boldsymbol{\Sigma}_{u_i})} \mathbb{E}_{q(\mathbf{z}_i, \boldsymbol{\pi}_i)} \left[\ln p(\mathbf{z}_i | \boldsymbol{\pi}_i) \right] & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \int \mathrm{Dir}_K(\boldsymbol{\pi}_i; \boldsymbol{\gamma}_i) \ln \pi_{ik} \, \mathrm{d}\pi_{ik} \\ & = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{ink} \ln \tilde{\pi}_{ik}, \end{aligned} \tag{18}$$

where:

$$\boxed{\ln \tilde{\pi}_{ik} = \psi(\gamma_{ik}) - \psi\left(\sum_{j=1}^{K} \gamma_{ij}\right).} \tag{19}$$

A.3 $\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln p(\boldsymbol{\pi}_i|\boldsymbol{\alpha})\right]$

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln p(\boldsymbol{\pi}_i|\boldsymbol{\alpha})\right] &= \mathbb{E}_{\mathrm{Dir}(\boldsymbol{\pi}_i;\boldsymbol{\gamma}_i)}\left[\ln\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \left[\sum_{k=1}^{K}\ln\Gamma(\alpha_k) - (\alpha_k - 1)\ln\pi_{ik}\right]\right] \\
&= \left[\ln\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \sum_{k=1}^{K}\ln\Gamma(\alpha_k)\right] + \sum_{k=1}^{K}(\alpha_k - 1)\ln\tilde{\pi}_{ik}. \qquad (20)
\end{aligned}
$$

A.4 $\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln q(\mathbf{z}_i|\mathbf{r}_i)\right]$

$$
\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln q(\mathbf{z}_i|\mathbf{r}_i)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ink}\ln r_{ink}. \qquad (21)
$$

A.5 $\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln q(\boldsymbol{\pi}_i|\boldsymbol{\gamma}_i)\right]$

$$
\mathbb{E}_{q(\mathbf{u}_i;\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})}\mathbb{E}_{q(\mathbf{z}_i,\boldsymbol{\pi}_i)}\left[\ln q(\boldsymbol{\pi}_i|\boldsymbol{\gamma}_i)\right] = \ln\Gamma\left(\sum_{j=1}^{K}\gamma_{ij}\right) - \sum_{k=1}^{K}\left[\ln\Gamma(\gamma_{ik}) - (\gamma_{ik} - 1)\ln\tilde{\pi}_{ik}\right]. \qquad (22)
$$

B MAXIMISATION OF THE ELBO

Since the ELBO can be evaluated as shown in Appendix A, we can maximise the ELBO w.r.t. "task-specific" variational parameters by taking derivative, setting it to zero and solving for the parameters of interest.

B.1 VARIATIONAL CATEGORICAL DISTRIBUTION

Note that:

$$
\sum_{k=1}^{K} r_{ink} = 1. \qquad (23)
$$

The derivative of L_i with respect to r_{ink} can be expressed as:

$$
\frac{\partial\mathsf{L}}{\partial r_{ink}} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}}) + \ln\mathcal{N}(\boldsymbol{\mu}_{u_{in}};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \ln\tilde{\pi}_{ik} - \ln r_{ink} - 1 + \lambda, \qquad (24)
$$

where: λ is the Lagrange multiplier and $\ln\tilde{\pi}_{ik}$ is defined in Eq. (19). Setting the derivative to zero and solving for r_{ink} give:

$$
\boxed{r_{ink} \propto \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}}) + \ln\mathcal{N}(\boldsymbol{\mu}_{u_{in}};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \ln\tilde{\pi}_{ik}\right].} \qquad (25)
$$

B.2 VARIATIONAL DIRICHLET DISTRIBUTION

The lower-bound related to γ_{ik} can be written as:

$$
\begin{aligned}
\mathsf{L} &= \sum_{k=1}^{K}\sum_{n=1}^{N} r_{ink}\ln\tilde{\pi}_{ik} + \sum_{k=1}^{K}(\alpha_k - 1)\ln\tilde{\pi}_{ik} - \ln\Gamma\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \sum_{k=1}^{K}\left[\ln\Gamma(\gamma_{ik}) - (\gamma_{ik} - 1)\ln\tilde{\pi}_{ik}\right] \\
&= -\ln\Gamma\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \sum_{k=1}^{K}\ln\tilde{\pi}_{ik}\left(\alpha_k - \gamma_{ik} + \sum_{n=1}^{N} r_{ink}\right) + \ln\Gamma(\gamma_{ik}) \\
&= -\ln\Gamma\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \sum_{k=1}^{K}\left[\psi(\gamma_{ik}) - \psi\left(\sum_{j=1}^{K}\gamma_{ij}\right)\right]\left(\alpha_k - \gamma_{ik} + \sum_{n=1}^{N} r_{ink}\right) + \ln\Gamma(\gamma_{ik}). \qquad (26)
\end{aligned}
$$

Hence, the lower-bound related to γ_{ik} is:

$$\begin{aligned}\mathsf{L}[\gamma_{ik}] = &-\ln\Gamma\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \psi(\gamma_{ik})\left(\alpha_k - \gamma_{ik} + \sum_{n=1}^{N} r_{ink}\right) \\ &- \psi\left(\sum_{j=1}^{K}\gamma_{ij}\right)\left(\sum_{j=1}^{K}\alpha_j - \gamma_{ij} + \sum_{n=1}^{N} r_{inj}\right) + \ln\Gamma(\gamma_{ik})\end{aligned} \tag{27}$$

Taking derivative w.r.t. γ_{ik} gives:

$$\begin{aligned}\frac{\partial \mathsf{L}}{\partial \gamma_{ik}} &= -\psi\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \Psi(\gamma_{ik})\left(\alpha_k - \gamma_{ik} + \sum_{n=1}^{N} r_{ink}\right) - \psi(\gamma_{ik}) \\ &\quad - \Psi\left(\sum_{j=1}^{K}\right)\left(\sum_{j=1}^{K}\alpha_j - \gamma_{ij} + \sum_{n=1}^{N} r_{inj}\right) + \psi\left(\sum_{j=1}^{K}\gamma_{ij}\right) + \psi(\gamma_{ik}) \\ &= \Psi(\gamma_{ik})\left(\alpha_k - \gamma_{ik} + \sum_{n=1}^{N} r_{ink}\right) - \Psi\left(\sum_{j=1}^{K}\right)\sum_{j=1}^{K}\alpha_j - \gamma_{ij} + \sum_{n=1}^{N} r_{inj},\end{aligned} \tag{28}$$

where $\Psi(.)$ is the trigamma function.

Setting the derivative to zero yields a maximum at:

$$\boxed{\gamma_{ik} = \alpha_k + N_{ik},} \tag{29}$$

where:

$$N_{ik} = \sum_{n=1}^{N} r_{ink}. \tag{30}$$

B.3 MAXIMUM LIKELIHOOD FOR THE TASK-THEME $\boldsymbol{\mu}_k$ AND $\boldsymbol{\Sigma}_k$

The terms in the objective function relating to μ_k can be written as:

$$\begin{aligned}\mathsf{L}[\mu_k] &= \sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\ln\mathcal{N}\left(\boldsymbol{\mu}_{u_{in}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\ &= -\frac{1}{2}\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)\end{aligned} \tag{31}$$

Taking derivative w.r.t. $\boldsymbol{\mu}_k$ gives:

$$\frac{\partial \mathsf{L}}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k). \tag{32}$$

Setting the derivative to zero yields a maximum at:

$$\boxed{\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\boldsymbol{\mu}_{u_{in}}}{\sum_{i=1}^{T} N_{ik}}.} \tag{33}$$

The terms in the objective function relating to $\boldsymbol{\Sigma}_k$ is given as:

$$\begin{aligned}\mathsf{L} &= \sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}}) + \ln\mathcal{N}(\boldsymbol{\mu}_{u_{in}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] \\ &= -\frac{1}{2}\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left[\mathrm{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}}) + d\ln(2\pi) + \ln|\boldsymbol{\Sigma}_k| + \left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)\right].\end{aligned} \tag{34}$$

Taking derivative w.r.t. $\boldsymbol{\Sigma}_k$ gives:

$$\begin{aligned}\frac{\partial \mathsf{L}}{\partial \boldsymbol{\Sigma}_k} &= -\frac{1}{2}\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left[-\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}}\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\right] \\ &= \frac{1}{2}\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{u_{in}} - \mathbf{I} + \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)^{\top}\right]\boldsymbol{\Sigma}_k^{-1}.\end{aligned} \tag{35}$$

Setting the derivative to zero gives:

$$\boxed{\boldsymbol{\Sigma}_k = \frac{1}{\sum_{i=1}^{T} N_{ik}}\sum_{i=1}^{T}\sum_{n=1}^{N} r_{ink}\left[\boldsymbol{\Sigma}_{u_{in}} + \left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{\mu}_{u_{in}} - \boldsymbol{\mu}_k\right)^{\top}\right].} \tag{36}$$

B.4 MAXIMUM LIKELIHOOD FOR $\boldsymbol{\alpha}$

The lower-bound with terms relating to α_k can be expressed as:

$$\mathsf{L} = T\left[\ln\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \sum_{k=1}^{K}\ln\Gamma(\alpha_k)\right] + \sum_{i=1}^{T}\sum_{k=1}^{K}(\alpha_k - 1)\left[\psi(\gamma_{ik}) - \psi\left(\sum_{j=1}^{K}\gamma_{ij}\right)\right]. \tag{37}$$

Taking derivative w.r.t. α_k gives:

$$g_k = \frac{\partial \mathsf{L}}{\partial \alpha_k} = T\left[\psi\left(\sum_{j=1}^{K}\alpha_j\right) - \psi(\alpha_k)\right] + \sum_{i=1}^{T}\left[\psi(\gamma_{ik}) - \psi\left(\sum_{j=1}^{K}\gamma_{ij}\right)\right]. \tag{38}$$

The second derivative is, therefore, obtained as:

$$\frac{\partial^2 \mathsf{L}}{\partial \alpha_k \partial \alpha_{k'}} = T\left[\Psi\left(\sum_{j=1}^{K}\alpha_j\right) - \delta(k - k')\Psi(\alpha_k)\right]. \tag{39}$$

The Hessian can be written in matrix form [Minka, 2000] as:

$$\mathbf{H} = \mathbf{Q} + \mathbf{1}\mathbf{1}^T a \tag{40}$$

$$q_{kk'} = -T\delta(k - k')\Psi(\alpha_k) \tag{41}$$

$$a = T\Psi\left(\sum_{j=1}^{K}\alpha_j\right). \tag{42}$$

One Newton step is therefore:

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \mathbf{H}^{-1}\mathbf{g} \tag{43}$$

$$(\mathbf{H}^{-1}\mathbf{g})_k = \frac{g_k - b}{q_{kk}}, \tag{44}$$

where:

$$b = \frac{\sum_{j=1}^{K} g_j/q_{jj}}{1/a + \sum_{j=1}^{K} 1/q_{jj}}. \tag{45}$$