A Review from Neurips 2020

Insufficient evaluation for the task selection, e.g. adding a baseline, varying the number of ways and the number of training examples. Originally, we only ran the experiment shown in Figure 3 without the baseline Task2Vec and fixing the setting for 5-way 5-shot. In the current version, we add more experiments as shown in Figure 4 with several variations in term of number of classes within a task, and the number of training tasks. The baseline Task2Vec is also included in the comparison.

The proposed approach adds an inefficient step. The proposed approach is not about meta-learning, but to learn a generative model to represent tasks in an embedding space. The obtained representation can be used either for task similarity (as presented in the paper) or further downstream transfer-learning works, such as task augmentation. Here, we use it as a pre-processing step before performing meta-learning for demonstration purpose without claiming its efficiency. If considering such applications, other baselines, such as Task2Vec, also share the same computational burden when calculating task-to-task distances.

Is the proposed approach restricted to gradient based meta-learning algorithms? Meta-learning is used to demonstrate the concept of task similarity quantification proposed in our paper, so our method is not restricted to any particular meta-learning algorithm. To demonstrate that, the submitted paper shows the results using the metric-based Prototypical Networks (we replicate the paper results in Fig. 4b and add the 10-way result in Fig. 4d). The extension to other meta-learning algorithms is trivial.

Experimenting on only 2 data sets. The reason for using those two data sets is that the our proposed approach works with low-dimensional data (Omniglot) or extracted features (mini-ImageNet). We will demonstrate our method on other data sets by embedding images into a lower-dimensional embedding space before performing LDCC. The training at that stage will be to learn both the image embedding and LDCC simultaneously.

B Detail derivation of each term in ELBO

B.1 $\mathbb{E}_q \left[\ln p(\mathbf{z}|\boldsymbol{\theta}) \right]$

$$
\begin{aligned}
\mathbb{E}_q \left[\ln p(\mathbf{z}|\boldsymbol{\theta}) \right] &= \sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_{dcnk} = 1; \mathbf{r}_{dcn}) \int q(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) \ln p(z_{dcnk} = 1 | \boldsymbol{\theta}_{dc}) \, d\boldsymbol{\theta}_{dc} \\
&= \sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{dcnk} \int \mathrm{Dir}_K \left(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc} \right) \ln \theta_{dck} \, d\boldsymbol{\theta}_{dck} \\
&= \sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{dcnk} \ln \tilde{\theta}_{dck},
\end{aligned} \tag{17}
$$

where:

$$
\ln \tilde{\theta}_{dck} = \psi \left(\gamma_{dck} \right) - \psi \left(\sum_{j=1}^{K} \gamma_{dcj} \right). \tag{18}
$$

B.2 $\mathbb{E}_q \left[\ln p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha}) \right]$

$$
\begin{aligned}
\mathbb{E}_q \left[\ln p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha}) \right] &= \sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{l=1}^{L} q(y_{dcl} = 1; \boldsymbol{\eta}_{dc}) \int q(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) \ln p(\boldsymbol{\theta}_{dc} | y_{dcl} = 1, \boldsymbol{\alpha}) \, d\boldsymbol{\theta}_{dc} \\
&= \sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{l=1}^{L} \eta_{dcl} \int \mathrm{Dir}_K (\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) \ln \mathrm{Dir}_K (\boldsymbol{\theta}_{dc}; \boldsymbol{\alpha}_l) \, d\boldsymbol{\theta}_{dc}.
\end{aligned} \tag{19}
$$

Note that the cross-entropy between 2 Dirichlet distributions can be expressed as:

$$\begin{aligned}
\mathcal{H}\left[\operatorname{Dir}\left(\mathbf{x};\boldsymbol{\alpha}_0\right),\operatorname{Dir}\left(\mathbf{x};\boldsymbol{\alpha}_1\right)\right] &= -\mathbb{E}_{\operatorname{Dir}(\mathbf{x};\boldsymbol{\alpha}_0)}\left[\ln \operatorname{Dir}\left(\mathbf{x};\boldsymbol{\alpha}_1\right)\right] \\
&= -\mathbb{E}_{\operatorname{Dir}(\mathbf{x};\boldsymbol{\alpha}_0)}\left[-\ln B(\boldsymbol{\alpha_1}) + \sum_{k=1}^{K}(\alpha_{1k}-1)\ln x_k\right] \\
&= \ln B(\boldsymbol{\alpha}_1) - \sum_{k=1}^{K}(\alpha_{1k}-1)\left[\psi(\alpha_{0k}) - \psi\left(\sum_{k'=1}^{K}\alpha_{0k'}\right)\right], \qquad (20)
\end{aligned}$$

where:

$$\ln B(\boldsymbol{\alpha}_1) = \sum_{k=1}^{K}\ln\Gamma\left(\alpha_{1k}\right) - \ln\Gamma\left(\sum_{j=1}^{K}\alpha_{1j}\right). \qquad (21)$$

Hence:

$$\mathbb{E}_q\left[\ln p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\alpha})\right] = \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{L}\eta_{dcl}\left[-\ln B(\boldsymbol{\alpha}_l) + \sum_{k=1}^{K}(\alpha_{lk}-1)\ln\tilde{\theta}_{dck}\right]. \qquad (22)$$

B.3 $\mathbb{E}_q\left[\ln p(\mathbf{y}|\boldsymbol{\phi})\right]$

$$\begin{aligned}
\mathbb{E}_q\left[\ln p(\mathbf{y}|\boldsymbol{\phi})\right] &= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{L} q(y_{dcl}=1;\boldsymbol{\eta}_{dc})\int q(\boldsymbol{\phi}_d;\boldsymbol{\lambda}_d)\ln p(y_{dcl}=1|\boldsymbol{\phi}_d)\,d\phi_d \\
&= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{L}\eta_{dcl}\int \operatorname{Dir}(\boldsymbol{\phi}_d;\boldsymbol{\lambda}_d)\ln\phi_{dl}\,d\phi_d \\
&= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{L}\eta_{dcl}\ln\tilde{\phi}_{dl}, \qquad (23)
\end{aligned}$$

where:

$$\ln\tilde{\phi}_{dl} = \psi\left(\lambda_{dl}\right) - \psi\left(\sum_{j=1}^{K}\lambda_{dl}\right) \qquad (24)$$

B.4 $\mathbb{E}_q\left[\ln p(\boldsymbol{\phi}|\boldsymbol{\delta})\right]$

$$\begin{aligned}
\mathbb{E}_q\left[\ln p(\boldsymbol{\phi}|\boldsymbol{\delta})\right] &= \sum_{d=1}^{M}\int q(\boldsymbol{\phi}_d;\boldsymbol{\lambda}_d)\ln p(\boldsymbol{\phi}_d|\boldsymbol{\delta})\,d\boldsymbol{\phi}_d \\
&= \sum_{d=1}^{M}\int \operatorname{Dir}_L(\boldsymbol{\phi}_d;\boldsymbol{\lambda}_d)\ln\operatorname{Dir}_L(\boldsymbol{\phi}_d|\boldsymbol{\delta})\,d\boldsymbol{\phi}_d \\
&= \sum_{d=1}^{M}\ln\frac{\Gamma\left(\sum_{l=1}^{L}\delta_l\right)}{\prod_{l=1}^{L}\Gamma\left(\delta_l\right)} + \sum_{l=1}^{L}(\delta_l-1)\ln\tilde{\phi}_{dl}. \qquad (25)
\end{aligned}$$

B.5 $\mathbb{E}_q\left[\ln p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]$

$$
\begin{aligned}
&\mathbb{E}_q\left[\ln p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] \\
&= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{dcnk}=1;\mathbf{r}_{dcn}) \\
&\qquad \times \int q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k;\mathbf{m}_k,\kappa_k,\mathbf{W}_k,\nu_k)\ln p(\mathbf{x}_{dcn}|z_{dcnk}=1,\boldsymbol{\mu},\boldsymbol{\Lambda})\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda} \\
&= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{dcnk} \\
&\qquad \times \int \mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right)\,\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\ln\mathcal{N}(\mathbf{x}_{dcn}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\,d\boldsymbol{\mu}_k\,d\boldsymbol{\Lambda}_k \\
&= \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{r_{dcnk}}{2}\int \mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right)\,\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\times \\
&\qquad\qquad \times\left[\ln|\boldsymbol{\Lambda}_k| - (\mathbf{x}_{dcn}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\boldsymbol{\mu}_k) - D\ln(2\pi)\right]\,d\boldsymbol{\mu}_k\,d\boldsymbol{\Lambda}_k
\end{aligned}
\tag{26}
$$

Note that:

$$
\ln\tilde{\Lambda}_k = \int \mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\ln|\boldsymbol{\Lambda}_k|\,d\boldsymbol{\Lambda}_k = \sum_{i=1}^{D}\psi\left(\frac{\nu_k-i+1}{2}\right) + D\ln(2) + \ln|\mathbf{W}_k|, \tag{27}
$$

and

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{x}_{dcn}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\boldsymbol{\mu}_k)\right] \\
&= \mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + (\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{m}_k-\boldsymbol{\mu}_k)\right. \\
&\qquad \left.+(\mathbf{m}_k-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}(\mathbf{x}_{dcn}-\mathbf{m}_k) + (\mathbf{m}_k-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}(\mathbf{m}_k-\boldsymbol{\mu}_k)\right] \\
&= (\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + \mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{m}_k-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}(\mathbf{m}_k-\boldsymbol{\mu}_k)\right] \\
&= (\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + \mathrm{tr}\left[\boldsymbol{\Lambda}\mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{m}_k-\boldsymbol{\mu}_k)^\top(\mathbf{m}_k-\boldsymbol{\mu}_k)\right]\right] \\
&= (\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + \mathrm{tr}\left[\boldsymbol{\Lambda}\left(\kappa_k\boldsymbol{\Lambda}_k\right)^{-1}\right] \\
&= (\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + D\kappa_k^{-1}
\end{aligned}
\tag{28}
$$

$$
\begin{aligned}
\Rightarrow \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_{dcn}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\boldsymbol{\mu}_k)\right] &= \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\boldsymbol{\Lambda}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + D\kappa_k^{-1}\right] \\
&= \nu_k(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\mathbf{W}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) + D\kappa_k^{-1}.
\end{aligned}
\tag{29}
$$

Hence:

$$
\begin{aligned}
\mathbb{E}_q\left[\ln p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] = \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{r_{dcnk}}{2}&\left[\ln\tilde{\Lambda}_k - D\kappa_k^{-1} - D\ln(2\pi) - \right. \\
&\left. -\nu_k(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\mathbf{W}_k(\mathbf{x}_{dcn}-\mathbf{m}_k)\right]
\end{aligned}
\tag{30}
$$

B.6 $\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \mathbf{m}_0, \kappa_0, \mathbf{W}_0, \nu_0)\right]$

$$
\begin{aligned}
&\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \mathbf{m}_0, \kappa_0, \mathbf{W}_0, \nu_0)\right] \\
&= \sum_{k=1}^{K} \int \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \\
&\qquad \times \ln \left[\mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_0, \nu_0\right)\right] d\boldsymbol{\mu}_k \, d\boldsymbol{\Lambda}_k \\
&= \sum_{k=1}^{K} \int \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \left[\int \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \ln \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right) d\boldsymbol{\mu}_k\right] d\boldsymbol{\Lambda}_k \\
&\qquad + \int \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \ln \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_0, \nu_0\right) d\boldsymbol{\Lambda}_k.
\end{aligned}
\tag{31}
$$

Note that:

$$
\begin{aligned}
&\int \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \ln \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right) d\boldsymbol{\mu}_k \\
&= -\frac{1}{2} \int \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \\
&\qquad \times \left[D \ln(2\pi) - D \ln \kappa_0 - \ln |\boldsymbol{\Lambda}_k| + \kappa_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^\top \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0)\right] d\boldsymbol{\mu}_k \\
&= \frac{1}{2} \left[D \ln\left(\frac{\kappa_0}{2\pi}\right) + \ln |\boldsymbol{\Lambda}_k| - \kappa_0 (\mathbf{m}_k - \mathbf{m}_0)^\top \boldsymbol{\Lambda}_k (\mathbf{m}_k - \mathbf{m}_0) - \frac{D\kappa_0}{\kappa_k}\right]
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
&\Rightarrow \int \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \left[\int \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \ln \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right) d\boldsymbol{\mu}_k\right] d\boldsymbol{\Lambda}_k \\
&= \frac{1}{2} \left[D \ln\left(\frac{\kappa_0}{2\pi}\right) + \ln \tilde{\Lambda}_k - \nu_k \kappa_0 (\mathbf{m}_k - \mathbf{m}_0)^\top \mathbf{W}_k (\mathbf{m}_k - \mathbf{m}_0) - \frac{D\kappa_0}{\kappa_k}\right]
\end{aligned}
\tag{33}
$$

and the cross-entropy between 2 Wishart distributions can be written as:

$$
\begin{aligned}
&\int \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \ln \mathcal{W}\left(\boldsymbol{\Lambda}_k; \mathbf{W}_0, \nu_0\right) d\boldsymbol{\Lambda}_k = -\mathcal{H}\left[(\mathbf{W}_k, \nu_k), (\mathbf{W}_0, \nu_0)\right] \\
&= \frac{\nu_0}{2} \ln \left|\mathbf{W}_0^{-1} \mathbf{W}_k\right| - \frac{D+1}{2} \ln |\mathbf{W}_k| - \frac{\nu_k}{2} \operatorname{tr}\left(\mathbf{W}_0^{-1} \mathbf{W}_k\right) - \log \Gamma_D\left(\frac{\nu_0}{2}\right) + \\
&\qquad + \frac{\nu_0 - D - 1}{2} \psi_D\left(\frac{\nu_k}{2}\right) - \frac{D(D+1)}{2} \ln(2) \\
&= -\frac{\nu_0}{2} \ln |\mathbf{W}_0| + \frac{\nu_0 - D - 1}{2} \left[\psi_D\left(\frac{\nu_k}{2}\right) + \ln |\mathbf{W}_k| + D \ln(2)\right] - \frac{\nu_k}{2} \operatorname{tr}\left(\mathbf{W}_0^{-1} \mathbf{W}_k\right) - \\
&\qquad - \log \Gamma_D\left(\frac{\nu_0}{2}\right) - \frac{\nu_0 D}{2} \ln(2) \\
&= -\frac{\nu_0}{2} \ln |\mathbf{W}_0| + \frac{\nu_0 - D - 1}{2} \ln \tilde{\Lambda}_k - \frac{\nu_k}{2} \operatorname{tr}\left(\mathbf{W}_0^{-1} \mathbf{W}_k\right) - \log \Gamma_D\left(\frac{\nu_0}{2}\right) - \frac{\nu_0 D}{2} \ln(2),
\end{aligned}
\tag{34}
$$

where $\Gamma_D(.)$ and $\psi_D(.)$ are the multivariate gamma and digamma functions, respectively.

Hence, the expectation of interest can be expressed as:

$$
\begin{aligned}
&\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \mathbf{m}_0, \kappa_0, \mathbf{W}_0, \nu_0)\right] \\
&= \sum_{k=1}^{K} \frac{1}{2} \left[D \ln\left(\frac{\kappa_0}{2\pi}\right) + \ln \tilde{\Lambda}_k - \nu_k \kappa_0 (\mathbf{m}_k - \mathbf{m}_0)^\top \mathbf{W}_k (\mathbf{m}_k - \mathbf{m}_0) - \frac{D\kappa_0}{\kappa_k}\right] - \\
&\qquad - \frac{\nu_0}{2} \ln |\mathbf{W}_0| + \frac{\nu_0 - D - 1}{2} \ln \tilde{\Lambda}_k - \frac{\nu_k}{2} \operatorname{tr}\left(\mathbf{W}_0^{-1} \mathbf{W}_k\right) - \log \Gamma_D\left(\frac{\nu_0}{2}\right) - \frac{\nu_0 D}{2} \ln(2).
\end{aligned}
\tag{35}
$$

The result is identical to the previous derivation in the literature Bishop, 2006, Eq. (10.74).

B.7 $\mathbb{E}_q\left[\ln q(\mathbf{z})\right]$

$$\mathbb{E}_q\left[\ln q(\mathbf{z})\right] = \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{dcnk}\ln r_{dcnk}. \tag{36}$$

B.8 $\mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right]$

$$\mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right] = -\sum_{d=1}^{M}\sum_{c=1}^{C}\ln B(\gamma_{dc}) + \sum_{j=1}^{K}(\gamma_{dck}-1)\ln\tilde{\theta}_{dck}. \tag{37}$$

B.9 $\mathbb{E}_q\left[\ln q(\mathbf{y})\right]$

$$\mathbb{E}_q\left[\ln q(\mathbf{y})\right] = \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{L}\eta_{dcl}\ln\eta_{dcl}. \tag{38}$$

B.10 $\mathbb{E}_q\left[\ln q(\boldsymbol{\phi})\right]$

$$\mathbb{E}_q\left[\ln q(\boldsymbol{\phi})\right] = -\sum_{d=1}^{M}\ln B(\boldsymbol{\lambda}_d) - \sum_{l=1}^{L}(\lambda_{dl}-1)\ln\tilde{\phi}_{dl}. \tag{39}$$

B.11 $\mathbb{E}_q\left[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right]$

$$\begin{aligned}
&\mathbb{E}_q\left[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] \\
&= \sum_{k=1}^{K}\mathbb{E}_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}\left[\ln q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)\right] \\
&= \sum_{k=1}^{K}\int\mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\Big[\ln\mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right) \\
&\qquad + \ln\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\Big]\,d\boldsymbol{\mu}_k\,d\boldsymbol{\Lambda}_k \\
&= \sum_{k=1}^{K}\int\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\left[\int\mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right)\ln\mathcal{N}\left(\boldsymbol{\mu}_k;\mathbf{m}_k,(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right)\,d\boldsymbol{\mu}_k\right]d\boldsymbol{\Lambda}_k \\
&\qquad + \int\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\ln\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\,d\boldsymbol{\Lambda}_k \\
&= -\sum_{k=1}^{K}\int\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\frac{1}{2}\ln\left|2\pi e(\kappa_k\boldsymbol{\Lambda}_k)^{-1}\right|d\boldsymbol{\Lambda}_k + \mathcal{H}\left[\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right),\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\right] \\
&= \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\tilde{\Lambda}_k + \frac{D}{2}\ln\left(\frac{\kappa_k}{2\pi}\right) - \frac{D}{2} - \mathcal{H}\left[\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right),\mathcal{W}\left(\boldsymbol{\Lambda}_k;\mathbf{W}_k,\nu_k\right)\right]\right\}.
\end{aligned} \tag{40}$$

C Optimise ELBO

C.1 Variational parameter for categorical distribution of z

Note that $\mathbf{r}_{dcn}$ is a K-dimensional vector parameterized for a categorical distribution. Hence, one constrain for $\mathbf{r}_{dcn}$ is:

$$\sum_{k=1}^{K} r_{dcnk} = 1. \tag{41}$$

We form the Lagrangian which consists of the lower-bound with isolating terms relating r_{dcnk} and add the appropriate Lagrange multiplier:

$$\begin{aligned}\mathsf{L}[r_{dcnk}] &= \mathbb{E}_q\left[\ln p(\mathbf{z}|\boldsymbol{\theta}) + \ln p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda}) - \ln q(\mathbf{z})\right] + \zeta\left(\sum_{k=1}^{K} r_{dcnk} - 1\right)\\ &= r_{dcnk}\ln\tilde{\theta}_{dck} + \frac{r_{dcnk}}{2}\left[\ln\tilde{\Lambda}_k - \nu_k(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\mathbf{W}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) - D\kappa_k^{-1} - D\ln(2\pi)\right]\\ &\qquad - r_{dcnk}\ln r_{dcnk} + \zeta\left(\sum_{k=1}^{K} r_{dcnk} - 1\right),\end{aligned} \tag{42}$$

where: $\ln\tilde{\theta}_{dck}$ and $\ln\tilde{\Lambda}_k$ are defined in Eqs. (18) and (27).

Taking the derivative w.r.t. r_{dcnk} gives:

$$\frac{\partial\mathsf{L}}{\partial r_{dcnk}} = \ln\tilde{\theta}_{dck} + \frac{1}{2}\left[\ln\tilde{\Lambda}_k - \nu_k(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\mathbf{W}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) - D\kappa_k^{-1} - D\ln(2\pi)\right] - \ln r_{dcnk} - 1 + \zeta. \tag{43}$$

Setting this derivative to zero and solving for r_{dcnk} gives:

$$r_{dcnk} \propto \exp\left\{\ln\tilde{\theta}_{dck} + \frac{1}{2}\left[\ln\tilde{\Lambda}_k - \nu_k(\mathbf{x}_{dcn}-\mathbf{m}_k)^\top\mathbf{W}_k(\mathbf{x}_{dcn}-\mathbf{m}_k) - D\kappa_k^{-1}\right]\right\}. \tag{44}$$

C.2 Variational parameter for Dirichlet distribution of θ

The lower-bound with isolating terms relating to only γ_{dck} is:

$$\begin{aligned}\mathsf{L}[\gamma_{dck}] &= \mathbb{E}_q\left[\ln p(\mathbf{z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\alpha}) - \ln q(\boldsymbol{\theta})\right]\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{dcnk}\ln\tilde{\theta}_{dck} + \sum_{l=1}^{L}\eta_{dcl}\sum_{k=1}^{K}(\alpha_{lk}-1)\ln\tilde{\theta}_{dck} + \ln B(\boldsymbol{\gamma}_{dc}) - \sum_{k=1}^{K}(\gamma_{dck}-1)\ln\tilde{\theta}_{dck}\\ &= \sum_{k=1}^{K}\ln\tilde{\theta}_{dck}\left[\sum_{n=1}^{N} r_{dcnk} + \sum_{l=1}^{L}\eta_{dcl}(\alpha_{lk}-1) - \gamma_{dck} + 1\right] + \ln B(\boldsymbol{\gamma}_{dc}),\end{aligned} \tag{45}$$

where $B(\mathbf{u})$ is defined in Eq. (21) – the normalizing constant of the Dirichlet distribution parameterized by $\mathbf{u}$. Note that: the sum notations are applicable only on their own single lines.

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

$$\begin{aligned}\frac{\partial L[\gamma_{dck}]}{\partial\gamma_{dck}} &= \Psi(\gamma_{dck})\left[\sum_{n=1}^{N} r_{dcnk} + \sum_{l=1}^{L}\eta_{dcl}(\alpha_{lk}-1) - \gamma_{dck} + 1\right]\\ &\qquad - \Psi\left(\sum_{j=1}^{K}\gamma_{dcj}\right)\sum_{j=1}^{K}\left[\sum_{n=1}^{N} r_{dcnj} + \sum_{l=1}^{L}\eta_{dcl}(\alpha_{lj}-1) - \gamma_{dcj} + 1\right].\end{aligned} \tag{46}$$

Setting the derivative to zero and solve for γ_{dck} yields:

$$\gamma_{dck} = 1 + \sum_{n=1}^{N} r_{dcnk} + \sum_{l=1}^{L}\eta_{dcl}(\alpha_{lk}-1). \tag{47}$$

C.3 Variational parameter for $\mathbf{y}$

Note that the L-dimensional vector $\boldsymbol{\eta}_{dc}$ is the parameter of a categorical distribution for $\mathbf{y}_{dc}$, it satisfies the following constrain:

$$\sum_{l=1}^{L}\eta_{dcl} = 1. \tag{48}$$

The Lagrangian can be expressed as:

$$\mathsf{L}[\mathbf{y}_{dc}] = \sum_{l=1}^{L} \eta_{dcl} \left[-\ln B(\boldsymbol{\alpha}_l) + \sum_{k=1}^{K} (\alpha_{lk} - 1) \ln \tilde{\theta}_{dck} \right] + \sum_{l=1}^{L} \eta_{dcl} \ln \tilde{\phi}_{dl} - \sum_{l=1}^{L} \eta_{dcl} \ln \eta_{dcl} + \xi \left(\sum_{l=1}^{L} \eta_{dcl} - 1 \right). \tag{49}$$

Taking the derivative w.r.t. η_{dcl} gives:

$$\frac{\partial \mathsf{L}}{\partial \eta_{dcl}} = -\ln B(\boldsymbol{\alpha}_l) + \sum_{k=1}^{K} (\alpha_{lk} - 1) \ln \tilde{\theta}_{dck} + \psi\left(\lambda_{dl}\right) - \psi \left(\sum_{j=1}^{K} \lambda_{dl} \right) - \ln \eta_{dcl} - 1 + \xi. \tag{50}$$

Setting the derivative to zero and solve for η_{dcl} yields:

$$\eta_{dcl} \propto \exp \left[\ln \tilde{\phi}_{dl} - \ln B(\boldsymbol{\alpha}_l) + \sum_{k=1}^{K} (\alpha_{lk} - 1) \ln \tilde{\theta}_{dck} \right]. \tag{51}$$

C.4 Variational parameter for ϕ

$$\begin{aligned}
\mathsf{L}[\boldsymbol{\lambda}_d] &= \sum_{c=1}^{C} \sum_{l=1}^{L} \eta_{dcl} \ln \tilde{\phi}_{dl} + \sum_{l=1}^{L} (\delta_l - 1) \ln \tilde{\phi}_{dl} + \ln B(\boldsymbol{\lambda}_d) - \sum_{l=1}^{L} (\lambda_{dl} - 1) \ln \tilde{\phi}_{dl} \\
&= \ln B(\boldsymbol{\lambda}_d) + \sum_{l=1}^{L} \ln \tilde{\phi}_{dl} \left(\delta_l - \lambda_{dl} + \sum_{c=1}^{C} \eta_{dcl} \right).
\end{aligned} \tag{52}$$

Taking derivative gives:

$$\frac{\partial \mathsf{L}}{\partial \lambda_{dl}} = \Psi(\lambda_{dl}) \left(\delta_l - \lambda_{dl} + \sum_{c=1}^{C} \eta_{dcl} \right) - \Psi \left(\sum_{j=1}^{L} \lambda_{dj} \right) \sum_{l=1}^{L} \left(\delta_l - \lambda_{dl} + \sum_{c=1}^{C} \eta_{dcl} \right). \tag{53}$$

Setting to zero and solving for λ_{dl} gives:

$$\lambda_{dl} = \delta_l + \sum_{c=1}^{C} \eta_{dcl}. \tag{54}$$

C.5 Variational parameter for word-topics

The lower-bound with the terms relating only to $\boldsymbol{\mu}_k, \mathbf{\Lambda}_k$ is:

$$\begin{aligned}
\mathsf{L}[\boldsymbol{\mu}, \mathbf{\Lambda}] = &\sum_{d=1}^{M} \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{r_{dcnk}}{2} \left[\ln \tilde{\Lambda}_k - \nu_k (\mathbf{x}_{dcn} - \mathbf{m}_k)^\top \mathbf{W}_k (\mathbf{x}_{dcn} - \mathbf{m}_k) - \frac{D}{\kappa_k} \right] \\
&- \sum_{k=1}^{K} \frac{1}{2} \left[\nu_k \kappa_0 (\mathbf{m}_k - \mathbf{m}_0)^\top \mathbf{W}_k (\mathbf{m}_k - \mathbf{m}_0) + \frac{D \kappa_0}{\kappa_k} + D \ln \kappa_k \right] \\
&- D_{\mathrm{KL}} \left[\mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \| \mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_0, \nu_0\right) \right].
\end{aligned} \tag{55}$$

Note that:

$$\frac{\partial}{\partial \mathbf{W}_k} \ln \tilde{\Lambda}_k = \left(\mathbf{W}_k^{-1} \right)^\top = \mathbf{W}_k^{-1} \tag{56}$$

$$\frac{\partial}{\partial \nu_k} \ln \tilde{\Lambda}_k = \frac{1}{2} \sum_{i=1}^{D} \Psi \left(\frac{\nu_k - i + 1}{2} \right), \tag{57}$$

and:

$$\begin{aligned}
&D_{\mathrm{KL}} \left[\mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \| \mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_0, \nu_0\right) \right] \\
&= \mathcal{H} \left[\mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_k, \nu_k\right), \mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_0, \nu_0\right) \right] - \mathcal{H} \left[\mathcal{W}\left(\mathbf{\Lambda}_k; \mathbf{W}_k, \nu_k\right) \right] \\
&= -\frac{\nu_0}{2} \ln \left| \mathbf{W}_0^{-1} \mathbf{W}_k \right| + \frac{\nu_k}{2} \left[\mathrm{tr} \left(\mathbf{W}_0^{-1} \mathbf{W}_k \right) - D \right] + \ln \Gamma_D \left(\frac{\nu_0}{2} \right) - \ln \Gamma_D \left(\frac{\nu_k}{2} \right) \\
&\quad + \frac{\nu_k - \nu_0}{2} \psi_D \left(\frac{\nu_k}{2} \right),
\end{aligned} \tag{58}$$

where: Γ_D and ψ_D are the multivariate gamma and digamma function.

Therefore:

$$\begin{cases} \dfrac{\partial}{\partial \mathbf{W}_k} D_{\mathrm{KL}} & = -\dfrac{\nu_0}{2}\mathbf{W}_k^{-1} + \dfrac{\nu_k}{2}\mathbf{W}_0^{-1} \\ \dfrac{\partial}{\partial \nu_k} D_{\mathrm{KL}} & = \operatorname{tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right) - D + \dfrac{\nu_k - \nu_0}{4}\Psi_D\left(\dfrac{\nu_k}{2}\right). \end{cases} \tag{59}$$

Taking derivative of the lower-bound w.r.t. the variational parameters of $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$ gives:

$$\frac{\partial \mathsf{L}}{\partial \mathbf{m}_k} = \nu_k \kappa_0 \mathbf{W}_k(\mathbf{m}_0 - \mathbf{m}_k) + \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} r_{dcnk}\nu_k \mathbf{W}_k(\mathbf{x}_{dcn} - \mathbf{m}_k). \tag{60}$$

$$\frac{\partial \mathsf{L}}{\partial \kappa_k} = \frac{D\kappa_0}{2\kappa_k^2} - \frac{D}{2\kappa_k} + \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} \frac{D r_{dcnk}}{2\kappa_k^2} = \frac{D}{2\kappa_k^2}\left[\kappa_0 - \kappa_k + \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} r_{dcnk}\right]. \tag{61}$$

$$\begin{aligned} \frac{\partial \mathsf{L}}{\partial \mathbf{W}_k} = & \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} \frac{r_{dcnk}}{2}\left[\mathbf{W}_k^{-1} - \nu_k(\mathbf{x}_{dcn} - \mathbf{m}_k)^\top(\mathbf{x}_{dcn} - \mathbf{m}_k)\right] \\ & - \frac{1}{2}\nu_k\kappa_0(\mathbf{m}_k - \mathbf{m}_0)^\top(\mathbf{m}_k - \mathbf{m}_0) + \frac{\nu_0}{2}\mathbf{W}_k^{-1} - \frac{\nu_k}{2}\mathbf{W}_0^{-1}. \end{aligned} \tag{62}$$

$$\begin{aligned} \frac{\partial \mathsf{L}}{\partial \nu_k} = & \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} \frac{r_{dcnk}}{2}\left[\frac{1}{2}\sum_{i=1}^{D}\Psi\left(\frac{\nu_k - i + 1}{2}\right) - (\mathbf{x}_{dcn} - \mathbf{m}_k)^\top \mathbf{W}_k(\mathbf{x}_{dcn} - \mathbf{m}_k)\right] \\ & - \frac{1}{2}\kappa_0(\mathbf{m}_k - \mathbf{m}_0)^\top\mathbf{W}_k(\mathbf{m}_k - \mathbf{m}_0) \\ & - \operatorname{tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right) + D - \frac{\nu_k - \nu_0}{4}\Psi_D\left(\frac{\nu_k}{2}\right). \end{aligned} \tag{63}$$

Setting these partial derivatives to zero and solving for the 4 parameters of interest give:

$$\begin{aligned} \mathbf{m}_k &= \frac{\kappa_0}{\kappa_0 + N_k}\mathbf{m}_0 + \frac{N_k}{\kappa_0 + N_k}\bar{\mathbf{x}}_k \\ \kappa_k &= \kappa_0 + N_k \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\kappa_0 N_k}{\kappa_0 + N_k}(\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^\top \\ \nu_k &= \nu_0 + N_k, \end{aligned} \tag{64}$$

where:

$$N_k = \sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} r_{dcnk} \tag{65}$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k}\sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} r_{dcnk}\mathbf{x}_{dcn} \tag{66}$$

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{d=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} r_{dcnk}(\mathbf{x}_{dcn} - \bar{\mathbf{x}}_k)(\mathbf{x}_{dcn} - \bar{\mathbf{x}}_k)^\top. \tag{67}$$

D Experiments on Omniglot and mini-ImageNet

For Omniglot, we follow the pre-processing steps as in few-shot image classification without any data augmentation, and use the standard train-test split in the original paper to prevent information leakage. For mini-ImageNet, we follow the common train-test split with 80 classes for training and 20 classes for testing Ravi and Larochelle, 2017. Since the dimension of raw images in mini-ImageNet is large, we employ the 640-dimensional features extracted from a wide-residual-network Rusu et al., 2019 to ease the calculation.

We follow Algorithm 1 to obtain the parameters of the image-topics posterior using tasks in training set. We use $L = 8$ task-topics and $K = 16$ image-topics for Omniglot, and $L = 3$ and $K = 10$ for mini-ImageNet. The Dirichlet distribution for task-topic proportion is chosen to be symmetric with $\delta = 0.01$. The hyper-parameter $\boldsymbol{\alpha}$ is uniformly chosen in $[0.01, 0.1]$ and held fixed in each experiment. The training for LDCC is carried out with 20 images per class to fit into the memory of a Nvidia 1080 Ti GPU, while the inference of the variational parameter $\boldsymbol{\lambda}$ is done on all available labelled data in each class (20 for Omniglot and 600 for mini-ImageNet). Note that this is for the inference of LDCC used in the correlation diagram. For the task selection, this number matches the number of shots in the few-shot learning setting.

D.1 Pre-processing

For Omniglot, the grey-scale images of each character are resized to 28-by-28 pixels, resulting in a 728-dimensional vector. We use the original train-test split, where 30 alphabets are used for training, and the other 20 alphabets are used for testing. No rotation is applied to augment classes.

For mini-ImageNet, we use the extracted features that are presented by 640-dimensional vectors. In addition, we normalise those features to be within [0, 1] by multiplying with a factor of 3.

D.2 Hyper-parameters in LDCC

The prior parameters of image-topics, $\mathbf{m}_0, \kappa_0, \mathbf{W}_0, \nu_0$, are selected as follows:

- The mean $\mathbf{m}_0$ is selected as the mean of all the images in the training of each data set,
- κ_0 is set to 0.01,
- The scale matrix $\mathbf{W}_0$ is selected as the covariance matrix of all the images in the training set,
- ν_0 is set to the dimension of images (728 for Omniglot and 640 for mini-ImageNet) adding 2.

The hyper-parameters for the learning rate used in online learning are $\tau_0 = 100$ and $\tau_1 = 0.5$. The mini-batch used in both data sets for LDCC is 500.

D.3 Network architectures used in meta-learning

We use two different network architectures on the two data sets. For Omniglot, we follow the "standard" 4 module CNN network, where each module consists of 64 3-by-3 filter convolutional layer, followed by batch normalisation, activated by ReLU and 2-by-2 max-pooling. For mini-ImageNet, we use a fully connected network with 1 hidden layer consisting of 128 hidden units, activated by ReLU.