Cluster-wise Hierarchical Generative Model for Deep Amortized Clustering
Supplementary Material

1. Details of ELBO

$$
\begin{aligned}
&\mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \log p_\theta(\mathcal{C}_{1:K}|\mathbf{X}) \\
&= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \log \left[\sum_{I_k=1}^{N_k} \int p_\theta(\mathbf{h}^k, \mathbf{z}^k, I^k|\mathcal{C}_{1:k-1}, \mathbf{S}^k) d\mathbf{z}^k \right] \\
&\geq \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathbf{z}^k, I^k|\mathcal{C}_{1:k}, \mathbf{S}^k)} \log \left[\frac{p_\theta(\mathbf{h}^k, \mathbf{z}^k, I^k|\mathcal{C}_{1:k-1}, \mathbf{S}^k)}{q_\phi(\mathbf{z}^k, I^k|\mathcal{C}_{1:k}, \mathbf{S}^k)} \right] \\
&= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathbf{z}^k, I^k|\mathcal{C}_{1:k}, \mathbf{S}^k)} \log \left[\frac{\prod_{i=1}^{M_k-1} p_\theta(h_i^k|\mathbf{z}^k, \mathbf{S}_{I^k}) p_\theta(\mathbf{z}^k|\mathbf{S}_{I^k}) p_\theta(I^k|\mathcal{C}_{1:k-1}, \mathbf{S}^k)}{q_\phi(\mathbf{z}^k|\mathbf{h}^k, \mathbf{S}_{I^k}) q_\phi(I^k|\mathcal{C}_{1:k}, \mathbf{S}^k)} \right] \\
&= \mathcal{L}_{\theta,\phi}
\end{aligned} \tag{1}
$$

2. Implementation Details

2.1. Multi-head attention Module (MhA)

We use Multi-head attention Module $\textsc{MhA}(\cdot)$ to exploit pair-wise or higher-order interactions between data points in both inter-and intra-cluster. Considering we want to capture the elements-wise relationship between $\mathbf{A}$ and $\mathbf{B}$, we set $\mathbf{A}$ as query, and set key and values are $\mathbf{B}$. The Multi-head attention Module is defined as follow

$$
\begin{aligned}
&\textsc{MhA}(\mathbf{A}, \mathbf{B}) = \textsc{concat}(\mathbf{O}_1, \cdots, \mathbf{O}_H)\mathbf{W}^h \\
&\text{where} \quad \mathbf{O}_i = \sigma\left(\mathbf{A}\mathbf{W}_i^Q \left(\mathbf{B}\mathbf{O}_i^K\right)^\top\right) \mathbf{B}\mathbf{W}_i^V
\end{aligned} \tag{2}
$$

where $\sigma(\cdot)$ is activation function, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ are head-specific transform matrices.

2.2. Implementation of $q_\phi(I^k|\mathcal{C}_{1:k}, \mathbf{S}^k)$

For convenience, we drawn one-hot vector $\mathbf{o}^k$ from the following categorical distribution:

$$
\begin{aligned}
\mathbf{o}^k &\sim \textsc{Categorical}(\textsc{Softmax}([s_1^k, \cdots, s_j^k, \cdots, s_{N_k}^k])) \\
s_j^k &= \frac{\textsc{Cosine}(\mathbf{e}_j^k, \mathbf{m}^k)}{\tau} - \frac{1}{k-1} \sum_{l=1}^{k-1} \frac{\textsc{Cosine}(\mathbf{m}^l, \mathbf{m}^k)}{\tau}
\end{aligned} \tag{3}
$$

here $I^k = \text{index}(\max(\mathbf{o}^k))$

3. Proof for Theorem 1

Theorem 1 *Ergodic amortized inference (EAI) objective $\mathcal{L}^*_{\theta,\phi}$ serves as a valid lower bound to the log likelihood of data and tighter than the original amortized inference objective $\mathcal{L}_{\theta,\phi}$ and SVI-based amortized inference objective $\mathcal{L}^\triangle_{\theta,\phi}$. The lower bounds satisfy*

$$
\mathcal{L}_{\theta,\phi} \leq \mathcal{L}^\triangle_{\theta,\phi} \leq \mathcal{L}^*_{\theta,\phi} \leq \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \log p_\theta(\mathcal{C}_{1:K}|\mathbf{X}) \tag{4}
$$

Proof Firstly, we show the following facts about the log-likelihood lower bound $\mathcal{L}^*_{\theta,\phi}$

$$\begin{aligned}
\mathcal{L}^*_{\theta,\phi} &= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \log \sum_{m=0}^{M} \pi_m w^{(k)}_m \\
&\leq \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \sum_{m=0}^{M} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{0:M}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} w^{(k)}_m \\
&= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \sum_{m=0}^{M} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{m}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} p_\theta(\mathcal{C}_{1:k}|\mathbf{X}) \\
&= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \log p_\theta(\mathcal{C}_{1:k}|\mathbf{X})
\end{aligned} \tag{5}$$

Secondly, we prove that $\mathcal{L}^{\triangle}_{\theta,\phi} \leq \mathcal{L}^*_{\theta,\phi}$. Let $I \subset \{1,\cdots.M\}$ with $|I| = P$ be a uniformly distributed subset of distinct indices from $\{1,\cdots,M\}$.

$$\begin{aligned}
\mathcal{L}^*_{\theta,\phi} &= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \log \sum_{m=0}^{M} \pi_m w^{(k)}_m \\
&= \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \log \mathbb{E}_{I=\{i_1,\cdots,i_P\}} \left[\frac{1}{P}\sum_{i=1}^{P} w^{(k)}_{(m,i)}\right] \\
&\overset{\text{(a)}}{\geq} \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \mathbb{E}_{I=\{i_1,\cdots,i_P\}} \left[\log \frac{1}{P}\sum_{i=1}^{P} w^{(k)}_{(m,i)}\right] \\
&\overset{\text{(b)}}{=} \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \left[\log \frac{1}{P}\sum_{i=1}^{P} w^{(k)}_{(m,i)}\right]
\end{aligned} \tag{6}$$

Inequality (a) holds due to the Jensen's inequality, and equality (b) holds since a simple observation:

$$\mathbb{E}_{I=\{i_1,\cdots,i_P\}} \left[\frac{a_{i_1}+\cdots+a_{i_P}}{P}\right] = \frac{a_1+\cdots+a_M}{M}$$

Based on above inequality, we can derive that

$$\begin{aligned}
\mathcal{L}^*_{\theta,\sigma} &\geq \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \pi_m \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \left[\log \frac{1}{P}\sum_{i=1}^{P} w^{(k)}_{(m,i)}\right] \\
&\overset{\text{(P=1)}}{=} \mathbb{E}_{p(\mathbf{X},\mathcal{C}_{1:K})} \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathbf{z}^{(k)}_{(0:M)}, I^k|\mathcal{C}_{1:k},\mathbf{S}^k)} \log w^{(0)}_k \\
&= \mathcal{L}_{\theta,\phi}
\end{aligned} \tag{7}$$

and

$$\mathcal{L}^*_{\theta,\phi} \geq \mathcal{L}^{\triangle}_{\theta,\phi} \tag{8}$$

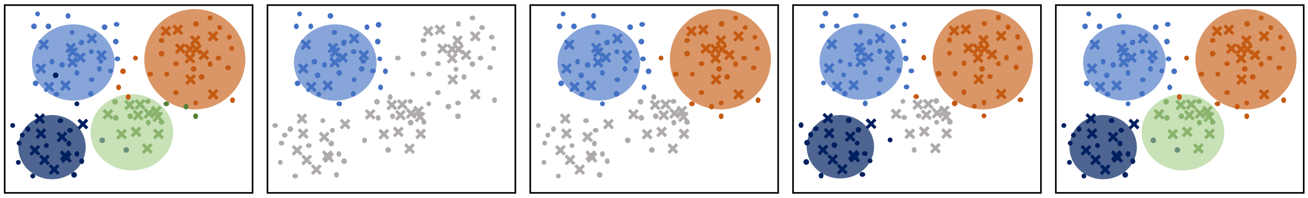
Since $\mathbf{z}^{(M)}_k$ is the optimal $\mathbf{z}_k$, we can obtain

$$\mathcal{L}_{\theta,\phi} \leq \mathcal{L}^{\triangle}_{\theta,\phi} \tag{9}$$

By combining inequalities (5), (7), (8) and (9), we established the bound as stated above. □

3.1. Infrastructure and Experimental Details

Infrastructure: We implement our model with Tensorflow, and conduct our experiments with:



(a) Ground-truth (b) Epoch 5 (c) Epoch 13 (d) Epoch 16 (e) Epoch 21

Figure 1. The cluster generative process of CHiGac on 2D MoG with 4 components.

- CPU: Intel Xeon Silver 4116 @2.1GHz.
- GPU: 8x GeForce RTX 2080Ti.
- RAM: DDR4 256GB.
- ROM: 8x 1TB 7.2K 6Gb SATA and 1x 960G SATA 6Gb R SSD
- Operating system: Ubuntu 18.04 LTS.
- Environments: Python 3.7; NumPy 1.18.1; SciPy 1.2.1; scikit-learn 0.23.2; seabornn 0.1; torch_geometric 1.6.1; matplotlib 3.1.3; dgl 0.4.2; pytorch 1.6

Hyper-parameter search: We trained with the following hyperparameters: The neural network (e.g., $f_\theta(\cdot), f_\phi(\cdot), g_{\phi,i}(\cdot)$) in our model is a multilayer perceptron (MLP). We use the *tanh* activation function. We apply dropout before every layers, except the last layer. The model is trained using Adam. We then tune the other hyper-parameters of both our approachs and our baselines automatically using the TPE method implemented by Hyepropt. We let Hyperopt conduct 200 trials to search for the optimal hyper-parameter configuration for each method on the validation of each dataset. The hyper-parameter search space is specified as follows:

- The number of hidden layers in a neural network: {0, 1, 2, 3 }.
- The number of neurons in a hidden layer: {100, 200, $\cdots$, 1000 }.
- Learning rate: $[10^{-8}, \cdots, 1]$.
- L2 regularization: $[10^{-12}, \cdots, 1]$.
- Dropout rate:$[0.05, \cdots, 1]$.
- Regularization coefficient λ: $[1, 10]$.
- The standard deviation of the prior for { θ, ϕ }: $[0.01, 0.5]$.

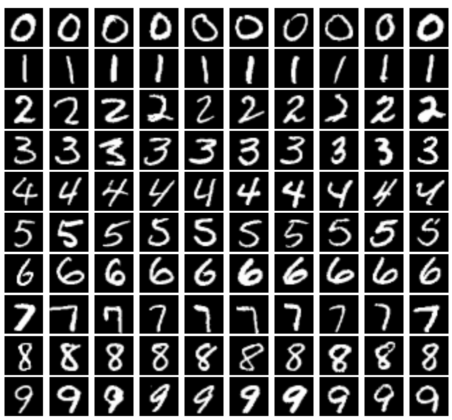
2D MoG data generation: We generate synthetic data by the following process.

$$\alpha \sim Exp(1) \qquad c_{1:N} \sim CRP(\alpha)$$
$$\mu_k \sim N(0, \sigma_\mu^2 \mathbf{1}) \qquad \mathbf{x}_i \sim N(\mu_{c_i}, \sigma^2 \mathbf{1})$$

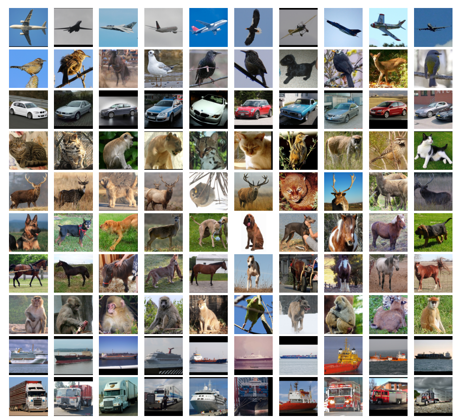
Here we give a more large-size illustration for the generative process, as shown in Figure 1

The learned clusters: In Figure 2 we show top-10 scoring images from each cluster in *MNIST* and *STL-10*. Each row corresponds to a cluster and images are sorted from left to right based on the learned h_i^k. We observe that for MNIST, the cluster assignment corresponds to natural clusters very well, while for *STL-10*, the results are mostly correct with airplanes, trucks and cars, but spends part of its attention on poses instead of categories when it comes to animal classes.

References



(a) *MNIST*



(b) *STL-10*

Figure 2. Illustration of the learned top-10 images in each cluster of *MNIST* and *STL-10*.