A1. Proof of calibrating incorrect label connections with LCC

Definition 1 *Given an undirected graph* $\mathrm{G} = \langle \mathrm{V}, \mathrm{E} \rangle$*, let* V *and* E *be its sets of nodes and edges respectively. We use* $G_c = (V_c, E_c)$ *to denote the subgraph of* G *induced by node set* $V_c \subseteq V$*, where* $E_c = (u, v) \in E, u, v \in S$*. Let* $N(u)$ *be the set of neighbors of* u *in* G*. We denote by* $G \backslash u$ *the graph obtained by removing node* u *from* V*. We use component to refer to a connected component of* G*.*

Definition 2 (Connectivity) *The connectivity (or edgeconnectivity in [1])* $\lambda(u, v)$ *between two distinct nodes* u *and* v *in* V *is the minimum number of edges whose removal disconnects* u *and* v*. The connectivity of the graph* $\lambda(G) = \min_{u,v \in V} \lambda(u, v)$ *is the minimum connectivity between any two distinct nodes in* G *(i.e., the smallest number of edges whose removal disconnects* G*).*

Definition 3 (Connected Component) *A connected component [2][3] of an undirected graph* $G = \langle V, E \rangle$*, denoted by* $G_c = (V_c, E_c)$*, is a subgraph where for every pair of nodes* $u, v \in V_c$*, there exists a path connecting* u *and* v *in* G_c*, and* $E_c = (u, v) \in E \mid u, v \in V_c$*, maximizing* V_c *under the constraint that no additional node from* $V \backslash V_c$ *can be added without violating connectivity.*

Definition 4 (Largest Connected Component) *The largest connected component [1],* $G_{\max} = (V_{\max}, E_{\max})$*, of an undirected graph* $G = \langle V, E \rangle$*, is defined by* $|V_{\max}|$ *being the maximum among all* $|V_c|$ *for any connected component* $G_c = (V_c, E_c)$ *in* G*, where* $V_{\max} \subseteq V$ *and* $E_{\max} = (u, v) \in E \mid u, v \in V_{\max}$*.*

Theorem 1 *Given an undirected graph* G*, Largest Connected Component (LCC)* $G_{\max} = \max_{|V_c|} \{G_c \mid G_c$ *is a connected component of* $G \backslash E'\}$*, where* E' *is incorrect edges affected by fine-grained classes.*

$Proof$ Process of proofing this Theorem:

1. Erroneous connections affected by fine-grained classes: Using Definitions 1 and 3, minority unknown and fine-grained classes affect the graph G by introducing incorrect edges E', where $E' \subseteq E$ and each $e \in E'$ connects unrelated nodes u and v, contradicting the connectivity condition in the definition of G_c. This leads to the formation of inaccurate connected components G'_c.

2. Criteria for Selecting the LCC: According to Definition 4, choosing $G_{\max}$ to maximize $|V_{\max}|$, automatically excluding smaller G'_c formed due to fine-grained classes. This is because $|V_{\max}| > |V'_c|$, where G'_c is a connected component formed with incorrect edges E'.

3. Direct Consequences of the Refinement Process: The direct consequences of selecting the LCC can be described using Definition 4: By excluding smaller connected components G'_c, we retain $G_{\max}$, ensuring the resulting subgraph more accurately represents the relationships between categories. This refinement process, grounded in concepts from Definitions 1 to 4, ensuring $G_{\max}$ represents true relationships among classes.

Remark The selection criterion of $G_{\max}$ inherently refines class-specific subgraphs by excluding error connected components G'_c affected fine-grained classes connections E'. This process enhances the accuracy and reliability of subgraphs in representing specific category information, offering an effective method to improve the quality of class-specific subgraphs, ensuring their effectiveness in analysis and application.

A2. Proof of GMM can facilitate the accurate estimation of classes

Theorem 2. Assume we have a dataset $X = \{x_1, x_2, \ldots, x_n\}$, where each x_i represents an observed sample. Gaussian Mixture Models (GMMs) model this data by combining K distinct Gaussian distributions, each corresponding to a different class, defined by parameters (mean μ_k and covariance Σ_k) and mixture coefficients π_k, where $k = 1, \ldots, K$. Therefore, for any sample x_i, the probability of being generated by the k^{th} Gaussian component is given by:

$$p\left(x_i \mid \mu_k, \Sigma_k\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_k\right|}} \exp\left(-\frac{1}{2}\left(x_i - \mu_k\right)^T \Sigma_k^{-1}\left(x_i - \mu_k\right)\right)$$

Here, d represents the dimensionality of the sample space, and $|\Sigma_k|$ is the determinant of the covariance matrix Σ_k. The posterior probability that sample x_i belongs to class k, meaning the probability of assigning sample x_i to class k, is calculated using Bayes' theorem as:

$$P\left(y_i = k \mid x_i\right) = \frac{\pi_k p\left(x_i \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j p\left(x_i \mid \mu_j, \Sigma_j\right)}$$

Where $P\left(y_i = k \mid x_i\right)$ denotes the posterior probability that sample x_i belongs to class k given the sample x_i, and y_i is the class label of sample x_i.

$Proof$ The EM algorithm for GMM consists of two main steps: the Expectation step (E-step) and the Maximization step (M-step). We prove the theorem by showing that each step increases or maintains the log-likelihood of the observed data.

1. E-step: In the E-step, we compute the expected log-likelihood of the complete data (observed + latent variables), given the observed data and current estimates of the parameters. Let's denote the log-likelihood function as $\mathcal{L}$ and the expected loglikelihood calculated in the E-step as $\mathcal{Q}\left(\theta \mid \theta^{(t)}\right)$, where θ represents the parameters of the model and $\theta^{(t)}$ the parameters at iteration t. According to Jensen's inequality [4], we have:

$$\mathcal{L}\left(\theta^{(t)}\right) \leq \mathcal{Q}\left(\theta \mid \theta^{(t)}\right)$$

2. M-step: In the M-step, we find the parameters $\theta^{(t+1)}$ that maximize the expected log-likelihood $\mathcal{Q}\left(\theta \mid \theta^{(t)}\right)$ calculated in the E-step. This implies:

$$\mathcal{Q}\left(\theta \mid \theta^{(t)}\right) \leq \mathcal{Q}\left(\theta^{(t+1)} \mid \theta^{(t)}\right)$$

Thus, by the end of the M-step, we have:

$$\mathcal{L}\left(\theta^{(t)}\right) \leq \mathcal{Q}\left(\theta \mid \theta^{(t)}\right) \leq \mathcal{L}\left(\theta^{(t+1)}\right)$$

This sequence ensures that the log-likelihood $\mathcal{L}$ of the observed data under the model parameters does not decrease across iterations of the EM algorithm.

Corollary: Given sufficient iterations and under suitable conditions, the EM algorithm for GMM will converge to a set of parameters that represent a local maximum of the log-likelihood $\mathcal{L}$ of the observed data.

$Proof$ Given the non-decreasing nature of the log-likelihood across EM iterations, and assuming the log-likelihood $\mathcal{L}$ is bounded above, the sequence of parameter estimates will converge to a local optimum of the log-likelihood $\mathcal{L}$. This is due to the property of compactness in the parameter space and the continuity of the loglikelihood $\mathcal{L}$, which together ensure the existence of convergent subsequences.

These proofs collectively underscore the effectiveness of the EM algorithm in iteratively refining parameter estimates for GMM, thereby facilitating the accurate estimation of classes.

Reference:

[1] Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5(Oct), 1225-1251.

[2] Bartlett, P. L., Jordan, M. I., McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473), 138-156.

[3] Cheng, Weiwei, Eyke Hüllermeier, and Krzysztof J. Dembczynski. "Bayes optimal multilabel classification via probabilistic classifier chains." In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 279-286. 2010.

[4] McShane, Edward James. "Jensen's inequality." (1937): 521-527.