Local and Global Guidance for Multi-Complementary Label Learning

Cheng Chen[1,2,3], Ivor W Tsang[2,3,4]

The Australian Artificial Intelligence Institute (AAII), University of Technology[1], Sydney,Australia

*Centre for Frontier AI Research, A*STAR[2], Singapore*

*Institute of High Performance Computing, A*STAR[3], Singapore*

School of Computer Science and Engineering, Nanyang Technological University[4], Singapore

***Abstract*—Complementary label learning involves instances associated with a candidate set of labels, which may contain multiple labels or a label that does not belong to the true label. Even though such weak supervision is easy to obtain, the setup inevitably introduces issues of label ambiguity and potential label distribution bias. Previous works have approached the task either based on the local characteristics of the dataset (label generation-wise), improving model robustness through mechanisms like transition matrices, or leveraging the global characteristics of the dataset (instance-wise), such as data augmentation and sample selection. However, these methods fail to simultaneously exploit both local and global aspects in addressing the issues of the multi-complementary label task. To tackle this issue, our paper presents a unified approach that considers both local and global guidance. For local guidance, we design homogeneous label regularisation using optimal transportation via the Sinkhorn-Knopp algorithm, ensuring a uniform label distribution across instances. For global guidance, we implement co-occurrence class embedding regularisation to enhance the model's understanding of the underlying sample distribution. This dual approach mitigates the misleading effects of complementary labels and corrects errors introduced by less informative labels during training. Overall, our experiments have demonstrated that our method is particularly effective on large datasets like CIFAR-100.**

***Index Terms*—Entropic regularisation and Logit Adjustment, Multi-Complementary Labels.**

I. Introduction

Complementary label learning, which uses labels that are not the true label, develops significantly due to the high costs and considerable time required for high-quality annotations in large datasets. This field sees several significant contributions, which are broadly categorized into two groups: local-oriented (label generation-wise) and global-oriented (instance-wise) methods. Local-oriented approaches focus primarily on label-generation mechanisms. Yu et al. [1] and Patrini et al. [2] propose biased complementary learning, operating under the assumption that labels are generated non-uniformly and employing a transition matrix to address this issue. Ishida et al. [3] present an unbiased estimator for classification risk, adaptable to arbitrary losses and models, specifically for uniform complementary label learning. Additionally, Kim et al. [4] introduce a negative cross-entropy loss function, which treats the complementary label as a negative label in the optimization process. Extending these concepts, Feng et al. [5] and Cao et al. [6] propose a multi-complementary label learning framework. The former designs a wrapper approach by converting multi-label contexts (MLCs) into several single complementary label sub-tasks, while the latter develops an unbiased risk estimator that treats each candidate set holistically.

On the global-oriented side, which focuses on instance-wise methods, Wang et al. [7], Ge et al. [8], and Cao et al. [6] make notable contributions. Ge et al. [8] introduce dual regularisation, advocating for the use of confident instances, identified through disagreement between two neural networks, for training. This concept bears resemblance to co-teaching by Han et al. [9], where instances with high posterior probability are selected for training, and co-training methods by Blum and Mitchell [10], Nigam et al. [11], and Qiao et al. [12], which involve choosing samples with small losses for training. In summary, the evolution of complementary label learning methodologies is viewed through the lens of these two distinct but interconnected approaches: the local-oriented methods, exemplified by Yu et al. [1], Ishida et al. [13], Kim et al. [4], Feng et al. [5], and Cao et al. [6], and the global-oriented methods, represented by Wang et al. [7], Ge et al. [8], and Cao et al. [6].

These methods can be categorised based on their focus areas: some prioritise the local properties of the dataset through techniques like data augmentations and the selection of challenging samples, while others emphasise global properties, utilising approaches such as unbiased estimators or surrogate losses. In addition, methods such as the transition matrix or prior label adjustments are often employed in this context. Despite these efforts, challenges persist in simultaneously enhancing the model's understanding of both local and global properties, which is vital for effectively managing the complementary label task. Moreover, this type of supervision can lead the classifier to disproportionately emphasise parameters for these complementary labels, potentially degrading performance. Nevertheless, the dataset maintains a uniform distribution across both label and instance spaces, highlighting the importance of leveraging this inherent local and global information to optimise classifier performance. This leads us to think, *'How can we design a learning framework that enhances the robustness of the classifier by simultaneously incorporating both the local and global properties of the dataset as guidance to solve the complementary label task?'* To address this challenge, local guidance is facilitated through equal-

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00160

partition label regularisation, employing optimal transportation via the Sinkhorn-Knopp algorithm. Concurrently, global guidance is exerted by leveraging label distribution through co-occurrence class embedding and consistency regularisation. This dual approach enables the model to collaboratively correct biases introduced by the inherently less informative feature of complementary labels during the learning process.

- In this paper, we propose a learning framework that integrates both the local and global properties of the dataset as guidance into the learning process, aiming to effectively address the challenges of less informative supervision in multi-complementary label learning.
- For local guidance, we incorporate the Sinkhorn-Knopp algorithm for optimal transportation, using an entropic function as regularisation for equipartition. Concurrently, global guidance is achieved through the use of co-occurrence peer embedding adjustment, guiding the model to treat every class uniformly, preventing it from being misled by the frequently occurring complementary labels.

II. PRELIMINARIES

We define $\mathcal{D}$ as the distribution of a Cartesian product $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, in which X denotes the variable of instances $\boldsymbol{x}$, and Y the true labels y. The feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space $\mathcal{Y} = [K]$, where $[K] = 1, 2, \ldots, K$. In multi-complementary label learning, since true labels are not given, we define a new distribution $\bar{\mathcal{D}}$, consisting of $(X, \vec{Y}) \in \mathcal{X} \times [K]$, where $\vec{Y}$ is the complementary label $\vec{y}$. The objective is to train a classifier on i.i.d. samples from $\bar{\mathcal{D}}$, $\{(\boldsymbol{x}_1, \vec{y}_1), \ldots, (\boldsymbol{x}_n, \vec{y}_n)\}$, to achieve classification performance same to training with true labels $y \in Y$. The complementary label generation, based on [5], considers $\vec{Y}_i$ as a candidate set of $k-1$ complementary labels from a total of k classes, leading to a multi-class classification problem. $\vec{Y}_i$ may contain either a full label set or an empty set, denoted as $\vec{Y}_i \in \overline{\mathcal{Y}}$ where $\overline{\mathcal{Y}} = 2^{\mathcal{Y}} - \varnothing - \mathcal{Y}$ and $|\overline{\mathcal{Y}}| = 2^k - 2$. During the label generation process, the number of complementary labels for each instance is denoted by the random variable s with distribution $p(s)$.

A. Global Guidance through Co-occurrence Class Embedding regularisation

In this section, we present co-occurrence class embedding regularization. It is designed to enforce uniform regularization across the embedding space using knowledge of label distribution, ensuring that the embedding space is consistent with the true label distribution. Our approach diverges from previous work [14], which primarily focus on logit adjustments related to a single class label index. Our method is designed to handle scenarios involving multiple candidate labels, targeting logit adjustment across all entries corresponding to classes within the candidate label set. This makes our approach particularly adept at addressing the multi-complementary learning task. The conventional approach determines the predicted label using $\arg\max_{i \in [k]} \boldsymbol{h}_i(\boldsymbol{x})$. The traditional softmax function is typically defined as $h(x) = \frac{\sum_{y \in Y} e^{\vec{q}_y}}{\sum_{y \in Y} e^{\vec{q}_y} + \sum_{j \notin Y} e^{\vec{q}_j}}$. Nonetheless, this method may not adequately consider the problem of label ambiguity in multi-complementary labels. These weaknesses arise because the standard softmax approach assumes that the labels are true classes for an instance, overlooking the situation introduced by labels that are explicitly not associated with the true class. In order to achieve this, we modify the softmax output as follows:

$$\vec{C}(x)_y = \frac{\sum_{y \in \vec{Y}} e^{\vec{q}_y} - \epsilon_y}{\sum_{y \in \vec{Y}} e^{\vec{q}_y} - \epsilon_y + \sum_{j \notin \vec{Y}} e^{\vec{q}_j}}, \tag{1}$$

where $\vec{q}_y$ are the logits for the classes in the candidate set $\vec{Y}$. In this formulation, we apply targeted subtraction to the y-th entries of the vector $\vec{C}(x)_y$ using ϵ_y, while keeping the remaining entries, specifically those where $y \neq j$, unchanged. This strategy preserves the original values of $\vec{C}(x)_j$ throughout each training batch. The term ϵ_y, known as the identical sample margin, is computed as $\epsilon_y = \frac{L}{n_j^{1/4}}$ for each class y within the range $\{1, \ldots, k\}$, where n_j denotes the number of samples in each class and L is a hyperparameter. This modification follows the approach described in the work of Cao et al. (2019), which is designed to increase the margin for underrepresented classes and decrease it for overrepresented ones. Although our underlying dataset distribution is uniform and free from imbalance issues, the complementary labels do not represent ground truth labels. Consequently, the model may mistakenly overweight or underweight certain classes during training if adjustments to co-occurrence class embedding are not made. This potential issue arises from the inherent ambiguity of these labels. The modified softmax output, $\vec{C}(x)_y$, will be utilised in local and global guidance regularisation frameworks.

1) Consistency Regularisation: In this section, we introduce a consistency regularisation in conjunction with Co-occurrence Class Embedding to form the global guidance, inspired by Wang et al. [7]. We start with a commonly used loss function in complementary label learning, defined as follows:

$$\overline{\mathcal{L}}_{\log}\left(f\left(\boldsymbol{x}_i\right), \bar{Y}_i\right) = -\sum_{y \in \bar{Y}_i} \log\left(1 - f_y\left(\boldsymbol{x}_i\right)\right), \tag{2}$$

The log loss function, proposed by [15] and [5], is motivated by the assumption that complementary labels do not contain the true labels. Thus, removing labels from the original multi-complementary label set provides a candidate set that likely includes the true label, offering more information than the original set. Differing from [7], we have revised the softmax output, $\vec{C}_y$, instead of using the traditional softmax output $h(x)$. Our method is designed to impose an equipartition constraint at the embedding level by subtracting a certain amount from the logits output, in line with the prior label distribution. The goal is to guide the model to treat every class uniformly, thereby preventing biased information from influencing the model due to the frequency of the occurring complementary labels, which do not accurately represent the underlying data distribution and leads to biased predictions. To

our knowledge, we are the first to integrate logit adjustment into consistency regularisation. Previous works by [15] and [5] considered only applying data augmentation to generate additional datasets from the existing dataset, arguing that minimising metric divergence between the original image and the augmented image would improve the generalisation ability of the model. However, these approaches failed to consider leveraging the label distribution as a constraint through logit adjustment. Together, these elements constitute global guidance regularisation.

$$\begin{aligned}&\overline{\mathcal{L}}_{\text{GGCR}}\left(\vec{C}_y\left(\mathcal{AUG}_j(\boldsymbol{x}_i)\right), \tilde{f}_y\right)\\&= -\sum_{j=1}^{N}\sum_{y\notin\vec{Y}_i}\tilde{f}_y\left(\boldsymbol{x}_i\right)\log\left(\vec{C}_y\left(\mathcal{AUG}_j\left(\boldsymbol{x}_i\right)\right)\right)\end{aligned}$$

Inspired by [7], our regularisation (3) is implemented by partitioning the observed candidate label set into two distinct groups: the complementary label set and the co-occurrence classes $\tilde{f}_y$. The co-occurrence classes encompass potential candidate classes, explicitly excluding those identified as complementary labels:

$$\tilde{f}_y\left(\boldsymbol{x}_i\right)=\begin{cases}\frac{\vec{y}_y(\boldsymbol{x}_i)}{\sum_{k\notin\vec{Y}_i}\vec{y}_k(\boldsymbol{x}_i)}, & \text{if } y\notin\vec{Y}_i,\\ 0, & \text{otherwise.}\end{cases}\tag{3}$$

The $\mathcal{AUG}_j\left(\boldsymbol{x}_i\right)$ denotes the data augmentation strategy applied to each instance x, as discussed by [7]. Additionally, we employ the revised softmax output $\vec{C}(x)_y$, as defined in Section A equation (1), for the prediction of the trained classifier.

B. Local Guidance Regularisation via Optimal Transportation

In this section, we propose applying optimal transport, as formulated in (4), to encourage the model to predict pseudo-labels by the marginal class prior. This approach aims to guide the model to assign a uniform weight to each class and each instance, preventing the model from being misled by multi-complementary labels to disproportionately overweight or underweight specific classes in each instance. To achieve this, we aim to minimize the following objective function $\langle Q, \log \vec{C}(x)\rangle$ with the guidance of entropic regularisation:

$$\begin{aligned}&\min_{Q\in G(r,c)}\langle Q,\log\vec{C}(x)\rangle+\gamma_{ij}\sum Q_{ij}(\log Q_{ij}-1),\\&\text{s.t. } G(r,c)=\{Q\in\mathbf{R}+^{K\times N}\mid Q\mathbf{1}=r, Q^{\top}\mathbf{1}=c\}.\end{aligned}\tag{4}$$

In the objective function (4), $\mathbf{1}$ denotes a vector containing all ones, with dimensions corresponding to the row and column sums of the matrix Q. The matrix Q represents the label assignment matrix, and γ, the Lagrangian multiplier, controls the degree of homogeneity of the matrix Q. The vectors α and β are obtained through matrix scaling iterations. The matrix Q is a conditional probability distribution, and we employ the Sinkhorn-Knopp algorithm to expedite label transportation for large-scale instances. The term $\vec{C}(x)y$ denotes the posterior probability after applying a softmax layer in a convolutional neural network. The constraint $G(r, c)$ ensures that each instance x_i is allocated exactly one label and the N instances are partitioned uniformly among the K classes. The vectors r and c represent the summation of the rows and columns of the matrix Q, and the marginal probabilities of Q, respectively. The Sinkhorn-Knopp algorithm iteratively modifies the matrix Q to meet these marginal constraints while simultaneously minimising $\min_{Q\in G(r,c)}\langle Q, \log \vec{C}(x)\rangle$. The vectors α and β have lengths K and N, respectively, corresponding to the number of columns and rows in the matrix Q. Our work follows the same instructions and derivation as [16]. We can formulate the equation (12) into the following:

Dataset	CIFAR100
SCL-EXP	45.36 ± 1.12
SCL-LOG	46.82 ± 0.64
UB-EXP	28.03 ± 1.42
UB-LOG	47.92 ± 2.62
UB-LOG with A&C	21.94 ± 0.93
POCR w/o Re-norm	50.49 ± 0.18
POCR	$\mathbf{54.17 \pm 0.89}$
Our	$\mathbf{69.69 \pm 0.55}$

TABLE I

MAIN EXPERIMENT: THE TABLE SHOWS AN ACCURACY COMPARISON BETWEEN THE MOST RECENT WORKS AND POCR (BASELINE) AND $\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log} + \mathcal{L}(Q, \alpha, \beta)$ FOR CIFAR100 DATASET.

$$\begin{aligned}&\mathcal{L}(Q,\alpha,\beta)=\textstyle\sum_i\sum_j\left(Q_{ij}\vec{C}_{ij}(x)\right)+\gamma Qij\left(\log Q_{ij}-1\right)\\&+\alpha^{\top}\left(Q\mathbf{1}-r\right)+\beta^{\top}\left(Q^{\top}\mathbf{1}-c\right).\end{aligned}\tag{5}$$

Deriving the objective function (12), we can obtain the following (The Full Derivation is shown in the appendix):

$$Q=\operatorname{diag}(u)\cdot K\cdot\operatorname{diag}(v).\tag{6}$$

We have the element of the matrix K_{ij}, which is $\left(-\frac{C(\vec{x})ij}{\gamma}\right)$ and $u_i = \exp(-\frac{\alpha_i}{\gamma})$, $v_j = \exp(-\frac{\beta_i}{\gamma})$. Given that we want Q to be constrained by r and c as follows:

$$Q\mathbf{1_m}=\operatorname{diag}(u)Kv=r, Q^{\top}\mathbf{1_n}\quad=\operatorname{diag}(v)K^{\top}u=c.\tag{7}$$

We have the element-wise product of each element as follows:

$$u\odot(Kv)=r, v\odot(K^{\top}u)\quad=c.\tag{8}$$

Here, r is obtained by the element-wise product between u (vector) and Kv (matrix-vector product), and c is obtained by the element-wise product between v (vector) and $K^{\top}u$ (matrix-vector product). The vectors u and v are updated iteratively using the following equations:

$$u=r\oslash(Kv), v=c\oslash(K^{\top}u),\tag{9}$$

where $\oslash$ denotes element-wise division. In a nutshell, we initiate α and β according to [17] to estimate scalar coefficient vectors v and u. The variable K is given (given the estimated $\vec{C}(x)_y$ matrix over γ). Ultimately, the objective function is to minimise Q and $\vec{C}(x)_y$ while maintaining equal partition constraints. Our goal is to estimate u and v to update Q until the convergence of the objective function. The final loss function can be formalised as follows:

$$\overline{\mathcal{L}}_{LG}=\overline{\mathcal{L}}_{\text{GGCR}}+\overline{\mathcal{L}}_{\log}+\mathcal{L}(Q,\alpha,\beta)\tag{10}$$

Dataset	POCR	$\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log}$	$\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log} + \mathcal{L}(Q, \alpha, \beta)$
CIFAR100	54.17 ± 0.89	67.60 ± 1.01	69.69 ± 0.55

TABLE II

ABLATION STUDY:THE TABLE SHOWS A COMPARISON BETWEEN POCR (BASELINE) AND $\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log}$ AND $\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log} + \mathcal{L}(Q, \alpha, \beta)$

III. EXPERIMENT

Datasets: We have conducted experiments with our method on the CIFAR100 dataset [18]. It contains one hundred classes for all images, totalling 60,000 images, with 50,000 for training and 10,000 for testing.

Experimental Results: The classification performance for the CIFAR100 dataset is shown in Table I. We have shown how our method performs in comparison to SCL-EXP and SCL-LOG [15], UB-EXP, UB-LOG and UB-LOG[5] with A and C, POCR w/o Re-norm, and POCR [7]. Our method has shown significant improvement compared to POCR, which is the state-of-the-art method in multi-complementary label learning. The result is calculated as the average over 5 random seeds.

1) Ablation Study: In this Table II, we have shown a comparison between the POCR (BaseLine) and $\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log}$ as well as $\overline{\mathcal{L}}_{\text{GGCR}} + \overline{\mathcal{L}}_{log} + \mathcal{L}(Q, \alpha, \beta)$. It has shown that both our proposed methods have improved the robustness of the model against the multi-complementary labels.

IV. CONCLUSION

In this paper, we address the learning of multi-complementary labels by simultaneously incorporating the local and global knowledge of the dataset into the model to make it aware of potential biases inherent in complementary labels. More specifically, our work provides a new perspective on dealing with multi-complementary labels. For local guidance, we apply equal-partition label regularisation using optimal transportation via the Sinkhorn-Knopp algorithm. Meanwhile, we use label distribution to conduct global guidance using co-occurrence class embedding regularisation. These unified methods aid in enhancing the robustness and accuracy of the classifier in tackling the less informative multi-complementary labels problem.

REFERENCES

[1] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 68–83.

[2] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1944–1952.

[3] T. Ishida, G. Niu, A. Menon, and M. Sugiyama, "Complementary-label learning for arbitrary losses and models," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2971–2980.

[4] Y. Kim, J. Yim, J. Yun, and J. Kim, "Nlnl: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 101–110.

[5] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3072–3081.

[6] Y. Cao, S. Liu, and Y. Xu, "Multi-complementary and unlabeled learning for arbitrary losses and models," *arXiv preprint arXiv:2001.04243*, 2020.

[7] D.-B. Wang, L. Feng, and M.-L. Zhang, "Learning from complementary labels via partial-output consistency regularization." in *IJCAI*, 2021, pp. 3075–3081.

[8] L. Ge, M. Gong, Y. Lin, and B. Du, "Dual-regularization complementary learning for image classification," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[9] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *arXiv preprint arXiv:1804.06872*, 2018.

[10] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[11] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 86–93.

[12] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 135–152.

[13] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," *arXiv preprint arXiv:1705.07541*, 2017.

[14] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[15] Y.-T. Chou, G. Niu, H.-T. Lin, and M. Sugiyama, "Unbiased risk estimators can mislead: A case study of learning with complementary labels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1929–1938.

[16] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, pp. 2292–2300, 2013.

[17] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," *arXiv preprint arXiv:1911.05371*, 2019.

[18] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

V. Appendix

A. Sinkhorn Knop algorithm

Basically, we are trying to minimise the following objective function $\langle Q, \vec{C}(x)\rangle$ with the guidance of entropic regularisation.

$$\begin{aligned} &\min_{Q\in G(r,c)} \langle Q, log\vec{C}(x)\rangle + \gamma_{ij}\sum Q_{ij}(\log Q_{ij}-1), \\ &\text{s.t. } G(r,c) = \{Q \in \mathbf{R}_{+}^{K\times N} \mid Q\mathbf{1_m} = r, Q^{\top}\mathbf{1_n} = c.\} \end{aligned} \tag{11}$$

The objective function (11) contains the following variables, Q label assignment matrix, γ Lagrangian multiplier, α and β are two vectors which need to be acquired through a matrix scaling iteration;

- The Q is a conditional probability distribution, using the Sinkhorn-Knop algorithm to fasten the large-scale instance label transportation, whereas the $\vec{C}(x)_y$ denoted as the posterior probability after the layer of softmax using convolutional neural networks.
- The constraint $G(r,c)$ enforces model to allocate each instance x_i with exact one label, subsequently, the N instances are partitioned uniformly among the K classes.
- The vectors r and c represent the summation of the rows and columns of matrix Q, the marginal probabilities of Q, respectively. The Sinkhorn-Knopp algorithm iteratively modifies the matrix Q to meet these marginal constraints while simultaneously minimizing $\min_{Q\in G(r,c)} \langle Q, \log \vec{C}(x)\rangle$.
- The γ, which is the Lagrangian multiplier, controls the degree of homogeneity for the matrix Q.
- The α is a vector with length K, in which K is the number of columns in the matrix Q. β is a vector of length N, where N is the number of rows in the matrix Q. In addition, $\alpha \in \mathbb{R}^N$ and $\beta \in \mathbb{R}^K$.

Our work has followed the same instructions and derivation as [16]. We can formulate the equation (11) into the following,

$$\begin{aligned} &\mathcal{L}(Q,\alpha,\beta) = \textstyle\sum_i \sum_j \left(Q_{ij}\vec{C}_{ij}(x)\right) + \gamma Q_{ij}(\log Q_{ij}-1) \\ &+\alpha^{\top}(Q1-r) + \beta^{\top}\left(Q^{\top}1-c\right). \end{aligned} \tag{12}$$

Deriving the objective function (12) as following,

$$\frac{\partial \mathcal{L}}{\partial Q_{ij}} = \vec{C}_{ij}(x) + \gamma(\log Q_{ij}-1) + \gamma Q_{ij} + \frac{1}{Q_{ij}} + \alpha_i + \beta_j = 0. \tag{13}$$

$$\log Q_{ij} = -\frac{1}{\gamma}\cdot\alpha_i - \frac{1}{\gamma}\vec{C}_{ij}(x) - \frac{1}{\gamma}\beta_j \tag{14}$$

$$Q_{ij} = \underbrace{\exp\left(-\frac{\alpha_i}{\gamma}\right)}_{U_i} \underbrace{\exp\left(-\frac{\vec{C}_{ij}(x)}{\gamma}\right)}_{K_{ij}} \underbrace{\exp\left(-\frac{\beta_j}{\gamma}\right)}_{V_j}. \tag{15}$$

$$Q = \mathrm{diag}(u)\cdot K \cdot \mathrm{diag}(v). \tag{16}$$

- We have the element of the matrix K_{ij}, which is $\left(-\frac{\vec{C(x)}ij}{\gamma}\right)$ and $U_i = \exp(-\frac{\alpha_i}{\gamma})$, $V_j = \exp(-\frac{\beta_i}{\gamma})$.

Our goal is to iteratively refine Q by calculating v, u and K; The Q calculation is shown as following:

- We first initiate α and β according to [17] to obtain unknown hyper-parameters v and u.
- After obtained v and u and since K is given (given the estimated $\vec{C}(x)_y$ matrix over γ).
- Ultimately, our objective function can reach the convergence where Q and $\vec{C}(x)_y$ is minimised while maintaining equal partition constraints.

Given that we want Q to be constrained by r and c as follows:

$$\begin{aligned} Q\mathbf{1} &= \mathrm{diag}(u)K\,\mathrm{diag}(v)\mathbf{1}, \\ Q\mathbf{1_m} &= \mathrm{diag}(u)Kv = r, \\ Q^{\top}\mathbf{1_n} &= \mathrm{diag}(v)K^{\top}u = c. \end{aligned} \tag{17}$$

We have the element-wise product of each element as follows:

$$\begin{aligned} u \odot (Kv) &= r, \\ v \odot (K^{\top}u) &= c. \end{aligned} \tag{18}$$

Here, r is obtained by the element-wise product between u (vector) and Kv (matrix-vector product), and c is obtained by the element-wise product between v (vector) and $K^{\top}u$ (matrix-vector product). The vectors u and v are updated iteratively using the following equations:

$$u = r \oslash (Kv), \tag{19}$$

$$v = c \oslash (K^{\top}u), \tag{20}$$

where $\oslash$ denotes element-wise division. Our goal is to estimate u and v to update Q until the convergence of the objective function.

B. Experimental settings

We have used the multi-complementary label according to [6, 7] as our observed candidate label set. The data augmentation strategy used is based on [7]. For the CIFAR100 dataset, we have used the PreAct-ResNet-18 neural network [19]. The total number of epochs for training is 200. We train the PreAct-ResNet-18 neural network with an initial learning rate of 0.1 and weight decay of 9e-4. Subsequently, we will divide the learning rate by 0.1 for every 50 epochs onward until the end of the training. This learning schedule also applies to weight decay. The optimiser used is SGD.