OT-LLP: OPTIMAL TRANSPORT FOR LEARNING FROM LABEL PROPORTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from label proportions (LLP), where the training data are arranged in form of groups with only label proportions provided instead of the exact labels, is an important weakly supervised learning paradigm in machine learning. Existing deep learning based LLP methods pursue an end-to-end learning fashion and construct the loss using Kullback-Leibler divergence, which measures the difference between the prior and posterior class distributions in each bag. However, unconstrained optimization on this objective can hardly reach a solution in accordance with the given proportions at the bag level. In addition, concerning the probabilistic classifier, it probably results in high-entropy conditional class distributions at the instance level. These issues will further degrade the performance of instance-level classification. To address these problems, we propose to impose the exact proportions on the classifier with a constrained optimization, and firstly apply the optimal transport algorithm to solve LLP. With the entropic regularization, our formulation allows to solve a convex programming efficiently and further arrive at an integer solution that meets the proportion constraint strictly. More importantly, our framework is model-agnostic, and demonstrates compelling performance improvement in extensive experiments, when it is incorporated into other deep LLP models as a post-hoc stage.

1 INTRODUCTION

Learning from label proportions (LLP) is a weakly supervised classification problem with only the label proportions in grouped data available. Still, training LLP aims to obtain an instance-level classifier for the new-come inputs. Successfully resolving LLP problems greatly contribute to many real-life applications: demographic classification (Ardehaly & Culotta, 2017), US presidential election (Sun et al., 2017), embryo implantation prediction (Hernández-González et al., 2018), spam filtering (Kuck & de Freitas, 2012), video event detection (Lai et al., 2014), visual attribute modeling (Chen et al., 2014; Yu et al., 2014a), and traffic flow prediction (Liebig et al., 2015).

On the one hand, the learnability of LLP strongly depends on the grouping of instances, as well as the proportions distribution. For example, Yu et al. (2014b) have studied the instance-level expected risk of empirical proportions risk minimization (EPRM) algorithm for LLP, with respect to the number of bags, the bag size, and the prior label distribution within the bags. They have proved that LLP is learnable with the EPRM principle and given the bound of expected learning risk.

On the other hand, EPRM strives for minimizing bag-level label proportions error. Normally, this goal is achieved by minimizing Kullback-Leibler (KL) divergence between prior and posterior class distributions in each bag. However, bag-level proportional information provides too insufficient constraints to perfectly solve LLP, because too many instance-level classifiers can satisfy proportional constraints exactly. In other words, when considering instance-level classification, LLP is ill-posed. As a consequence of this underdetermination, despite a number of achievements have been developed to resolve LLP accurately and efficiently,

it is still of great challenge to design an effective learning scheme to significantly improve the performance of high dimensional instances recognition, e.g., images, merely with bag-level proportional information.

In addition, acquiring the exact labeling congruous with label proportions somehow leads to a problematic integer programming problem (Stolpe & Morik, 2011). As a result, advanced LLP algorithms usually impose the relaxation to achieve probabilistic labeling, which is also used as the final classifier in an alternate labeling-and-classifying framework (Yu et al., 2013). However, in terms of the probabilistic classifier, the aforementioned ill-posed situation may be even worse: Infinite solutions are in accordance with the bag-level proportions. In general, suboptimal hypothesis can be probably found. In particular, in the extreme case where proportions are equal in each bag, it will lead to a trivial solution with uniform output distribution.

In this paper, we argue that the above challenge is mainly due to the inadequate KL divergence at the bag level, and propose to solve LLP through explicitly combing unsupervised clustering and proportional information. To be concrete, the unsupervised learning helps to discover the data clusters specifically corresponding to the classes. However, naively applying clustering without supervision will easily result in a trivial solution to assign all the instances to the same cluster. Fortunately, we can avoid this degeneration by imposing appropriate constraints on the label distribution (Caron et al., 2018). For example, when there is no knowledge on label distribution, we can impose discrete uniform distribution to the labels, leading to equal clustering. Besides, in a semi-supervised learning protocol, the clustering result can be restricted with the help of the labeled instance (Asano et al., 2019). Similarly, in LLP scenario, the constraint for clustering can be the label proportions. Hence, our solution is to construct a constrained optimization problem by confining the feasible solution to accurately comply the proportions in each bag. In order to tackle the weak supervision, we cast a framework by considering the classification and pseudo labeling within one objective, and leverage the label proportions as the constraints. To implement the proposed schema, we alternately update the network parameters and the pseudo labels, where optimal transport (OT) is employed to the pseudo labeling process and standard cross-entropy minimization is adopted to train the network parameters.

In summary, our main contributions are four-fold: (1) We propose a novel LLP framework called OT-LLP, which leverage the exact proportional information to construct a constrained optimization, and is an essentially orthogonal LLP treatment to the existing LLP algorithms; (2) We propose an alternate optimization process to solve the optimization on both the classifier and pseudo labeling, and firstly apply the OT algorithm to obtain integer solutions for the labels; (3) Our framework is model-agnostic, and we demonstrate that it can easily fit for various deep-based LLP algorithms to further boost their performance; (4) With no additional hyper-parameter involved, our framework achieves state-of-the-art LLP performance on several benchmark datasets, based on the neural network pipeline with standard settings.

2 Related Work

To our knowledge, four end-to-end pipelines are recently proposed for LLP, using deep neural networks as the backbone architectures. Specifically, Ardehaly & Culotta (2017) first propose an end-to-end LLP algorithm, incorporating the KL divergence of prior and posterior proportions into an objective. Although their approach can learn a competent instance-level classifier, it is hardly in accordance with the proportions in training data, especially with large bag sizes (e.g., > 64). Based on guessing the individual labels of samples (Yu et al., 2013), Dulac-Arnold et al. (2019) employ a similar instance-level cross entropy loss as the objective. Then, alternating update and convex relaxation are exploited to mitigate the intractable combinatorial optimization. However, the inaccurate labeling problem is still intact, due to the KL divergence loss.

Recently, by introducing the adversarial learning mechanism, LLP-GAN (Liu et al., 2019) greatly improves the method in (Ardehaly & Culotta, 2017). In detail, the discriminator is designed as a $(K{+}1)$-way classifier, where the first K classes indicate multi-class true samples, and the $(K{+}1)^{th}$ class accounts for the generated samples. The main insight is to implement better representation learning through the adversarial mechanism,

thus boosting the downstream discriminative task performance. Despite of the substantial improvement in performance compared with previous methods, LLP-GAN greatly suffers from the training instability, which inherits the characteristics from generative adversarial networks. Furthermore, subtle network structure design and hyper-parameters selection are required, in order to obtain satisfactory results. Similarly, Tsai & Lin (2019) introduce a consistency regularization technique in semi-supervised learning to the multi-class LLP problem, where a KL divergence on the proportions is also employed in the loss.

The above four algorithms all build their losses (partly) on the KL divergence, and solve an unconstrained optimization. We advocate that their final classifiers cannot fully agree with the proportions, thus deteriorating the classification performance at the instance level. In contrast, we address this problem with a constrained optimization, which can exactly follow the proportion constraint.

Furthermore, enough diversity in unlabeled data demonstrates useful behavior for classification (Lu et al., 2018). Based on the great ability of deep models, unsupervised representation learning is achievable and appealing as the essential step towards discriminative tasks. For example, aligning the clustering result with semantic classes can be successfully applied to image classification and segmentation (Ji et al., 2019). Asano et al. (2019) have proposed to apply (entropic constrained) optimal transport (OT) (Peyré et al., 2019) to unsupervised representation learning with a self-labeling mechanism. Following their seminal framework, in this paper, we propose an OT based LLP algorithm, and develop more detailed techniques.

3 PRELIMINARIES

In order to clearly describe how to leverage OT for LLP, we first introduce several preliminaries for both OT and LLP. More details are in Appendix due to the space limitation.

3.1 OPTIMAL TRANSPORT VIA KANTOROVICH RELAXATION

In Appendix A.1, we describe the Kantorovich formulation of *coupling* as an extension for the mass transportation in the Monge problem (10) (see Appendix A.1). Different from that mass transportation should be *deterministic*, *Kantorovich relaxation* instead considers a *probabilistic* transport, which allows *mass splitting* from a source toward several targets (Villani, 2008). Note that in this way, we may get rid of the challenge in the Monge problem, because the source measure is regarded as the atom in the Monge problem while separable in the Kantorovich relaxation with respect to the source measure.

To achieve the mass splitting, instead of a permutation σ or a surjective map T, we need a *coupling matrix* $\mathbf{P} \in \mathbb{R}_+^{n\times m}$, where P_{ij} is the amount of mass flowing from x_i to y_j. Admissible couplings give a simpler characterization than maps in the Monge problem as: Suppose that $\boldsymbol{a} \in \Sigma^n, \boldsymbol{b} \in \Sigma^m$, we denote $U(\boldsymbol{a}, \boldsymbol{b}) = \left\{\mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbb{1}_m = \boldsymbol{a},\ \mathbf{P}^\intercal \mathbb{1}_n = \boldsymbol{b}\right\}$, where Σ^s is a probability simplex, i.e., $\Sigma^s = \left\{\boldsymbol{a} \in \mathbb{R}_+^s \middle| \sum_{i=1}^{s} a_i = 1\right\}$, and $\mathbb{1}_n$ represents the column vector of all ones with dimension n.

Remark 1 (Peyré et al. (2019)). *The Kantorovich mapping is symmetric:* $\mathbf{P} \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \mathbf{P}^\intercal \in U(\boldsymbol{b}, \boldsymbol{a})$.

Similar to the Monge problem (10), let $\langle \cdot, \cdot \rangle$ be the Frobenius dot-product, the Kantorovich's *optimal transport* (OT) problem is a linear programming problem given $\boldsymbol{a}, \boldsymbol{b} \in \Sigma^n$ and a cost matrix $\mathbf{C}$ as

$$\mathbf{L}_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\mathbf{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle = \sum_{i,j} P_{ij} C_{ij}. \tag{1}$$

An important feature of OT (1) is that it induces a distance between two probability measures in Σ^n, which are both discrete in this paper, as soon as the cost matrix satisfies the properties of a legitimate distance.

Proposition 1 (Villani (2008)). *Let* $\mathbf{D} \in \mathbb{R}_+^{n\times n}$ *be a distance on* $[\![n]\!] = \{1, 2, \cdots, n\}$ *and* $p \geqslant 1$. *We can define the* p*-Wasserstein distance on* Σ^n *as* $W_p(\boldsymbol{a}, \boldsymbol{b}) = \mathbf{L}_{\mathbf{D}^p}(\boldsymbol{a}, \boldsymbol{b})^{1/p}, \forall \boldsymbol{a}, \boldsymbol{b} \in \Sigma^n$.

Remark 2. *Particularly,* $W_1(\boldsymbol{a}, \boldsymbol{b}) = \min_{\mathbf{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbf{P}}[\|\boldsymbol{x} - \boldsymbol{y}\|]$ *is a legitimate distance between two distributions, and is used Wasserstein GAN, as a weaker distance than the Jensen-Shannon (JS) Divergence in the original GAN and Kullback-Leibler (KL) divergence in the maximizing likelihood estimation (MLE).*

3.2 Entropic regularization of optimal transport

The solution of the original transport problem (1) is non-unique and tends to be sparse, i.e., arriving at certain vertex of the polytope $U(\boldsymbol{a}, \boldsymbol{b})$. In certain scenarios, the sparsity of optimal couplings for (1) is not desirable, so Cuturi (2013) instead employs the entropic regularization term to form a more "blurred" prediction. The discrete entropy of a coupling matrix $\mathbf{P}$ is well-known as $\mathbf{H}(\mathbf{P}) = -\sum_{i,j} P_{ij} \log(P_{ij})$.

Remark 3 (Peyré et al. (2019))**.** *The entropic function* $\mathbf{H}(\cdot)$ *is 1-strongly concave, due to the negative definite Hessian matrix:* $\partial^2 \mathbf{H}(\mathbf{P}) = -\mathbf{diag}(1./\mathbf{P})$ *and* $P_{ij} \leqslant 1, \forall i, j$*, where we flatten* $\mathbf{P}$ *to a long vector.*

Note that $\boldsymbol{a}\boldsymbol{b}^\intercal$ is an admissible coupling, and $\mathbf{H}(\mathbf{P}) \leqslant \mathbf{H}(\boldsymbol{a}) + \mathbf{H}(\boldsymbol{b}) = \mathbf{H}(\boldsymbol{a}\boldsymbol{b}^\intercal)$. The idea of the *entropic regularization* of OT is to add $-\mathbf{H}(\cdot)$ to the original OT (1) to obtain an approximate solutions as:

$$\mathbf{L}^{\varepsilon}_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\mathbf{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}), \tag{2}$$

which is a constrained minimization problem of an ε-convex function, thus has a unique optimal solution.

More importantly, the entropic constraint for OT can guarantee a more computationally efficient process to find the solution, as a consequence of restricting the search for low cost joint probabilities within sufficient smooth tables (Cuturi, 2013). In addition, we introduce Proposition 2 in Appendix A.2 to confirm that the solution of entropic regularized OT (2) converges to that of the original OT (1) as $\varepsilon \to 0$.

3.3 Learning from label proportions

In LLP problem, because the label proportions are available, we can restrict the instance-level self-labeling procedure with these proportional information using an OT framework. Before further discussion, we fist give the formal formulation for LLP by directly considering a multi-class problem with K classes in this paper. With no prior knowledge, we further suppose that the training data consist of N *randomly generated* disjoint bags. Consequently, the training data can be expressed as $\mathcal{D} = \{(\mathcal{B}_i, \boldsymbol{p}_i)\}_{i=1}^m$, where $\mathcal{B}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{n_i}$ denotes the instances in the i^{th} bag, and $\mathcal{B}_i \cap \mathcal{B}_j = \varnothing, \forall i \neq j$. The $\boldsymbol{p}_i \in [0,1]^K$ and n_i are the known ground-truth label proportions and the bag size of the i^{th} bag, respectively.

4 Approach

4.1 Linking OT to LLP (OT-LLP)

In Appendix A.3, we introduce how to leverage equal clustering to learn discriminative representation on unsupervised data and achieve classification in a self-labeling framework (Asano et al., 2019). In LLP problem, although the class distribution is not uniform within each bag, we can easily modify the constraint in (18) (see Appendix A.3) or the admissible couplings in (19) (see Appendix A.3) to fit in the proportional information. Consequently, with p_i^y as the proportion of class y in bag i, we convert (18) into

$$\begin{aligned} &\min_{q} \;\; BCE(p, q) = -\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{y=1}^{K} \frac{q(y|\mathbf{x}_{i,j})}{n_i} \log p_\phi(y|\mathbf{x}_{i,j}), \\ &s.t. \;\; \sum_{j=1}^{n_i} q(y|\mathbf{x}_{i,j}) = p_i^y \cdot n_i, q(y|\cdot) \in [0,1], \forall y \in [\![K]\!], \forall i \in [\![m]\!]. \end{aligned} \tag{3}$$

Note that (3) is a constrained optimization and the labels should strictly comply with the proportion information of each bag. Nevertheless, (3) is combinatorial in q and thus seems to be very difficult to optimize. Fortunately, we can convert (3) to a typical OT problem, which can be solved relatively efficiently.

In order to better explain the relation between (3) and OT, we rewrite it in a matrix fashion. Formally, let $\mathbf{Q}^i=(Q^i_{jk})\in\mathbb{R}_+^{K\times n_i}$, $Q^i_{jk}=q(k|\mathbf{x}_{i,j})/n_i$, and $\mathbf{P}^i=(P^i_{jk})\in\mathbb{R}_+^{K\times n_i}$, $P^i_{jk}=p_\phi(k|\mathbf{x}_{i,j})/n_i$. We further denote $\mathbf{Q}=\text{diag}\{\mathbf{Q}^i\}_{i=1}^m$, $\mathbf{P}=\text{diag}\{\mathbf{P}^i\}_{i=1}^m$, and $\boldsymbol{p}=(\boldsymbol{p}_1^\intercal,\boldsymbol{p}_2^\intercal,\cdots,\boldsymbol{p}_m^\intercal)^\intercal$, $\boldsymbol{b}=(\mathbb{1}_{n_1}^\intercal/n_1,\mathbb{1}_{n_2}^\intercal/n_2,\cdots,\mathbb{1}_{n_m}^\intercal/n_m)^\intercal$. Accordingly, we can define a set with the form of $U(\boldsymbol{p},\boldsymbol{b})=\left\{\mathbf{Q}\in\mathbb{R}_+^{mK\times N}\mid \mathbf{Q}\mathbb{1}_N=\boldsymbol{p},\ \mathbf{Q}^\intercal\mathbb{1}_{mK}=\boldsymbol{b}\right\}$.

Then, we can give an equivalent problem for (3) with the following OT problem:

$$\min_{\mathbf{Q}\in U(\boldsymbol{p},\boldsymbol{b})}\langle\mathbf{Q},-\log\mathbf{P}\rangle=BCE(p,q)+\log\prod_{i=1}^m n_i. \tag{4}$$

On the other hand, with $\lambda\to+\infty$, we can instead solve the entropic regularized OT problem to accelerate the process of convergence, as well as avoiding the non-unique sparse solution.

$$\mathbf{L}^{1/\lambda}_{-\log\mathbf{P}}(\boldsymbol{p},\boldsymbol{b})=\min_{\mathbf{Q}\in U(\boldsymbol{p},\boldsymbol{b})}\langle\mathbf{Q},-\log\mathbf{P}\rangle-\frac{1}{\lambda}\mathbf{H}(\mathbf{Q}). \tag{5}$$

4.2 Alternating optimization

In the proposed learning framework, the network parameters $\phi=(\varphi,\theta)$ and self-labels $\mathbf{Q}$ are alternately updated. Now, we further describe the details as follows.

Training the network with fixed q: Because the cross-entropy is differentiable with respect to the network parameters $\phi=(\varphi,\theta)$, we can directly conduct common optimizer, e.g., ADAM, on the objective in (3) by fixing q.

Updating the labels $\mathbf{Q}$ with fixed ϕ: When the model is fixed, the label assignment matrix $\mathbf{Q}$ are obtained by OT or entropic regularized OT. When performing the original OT, the solution $\mathbf{Q}^*$ is with binary elements of 0 or 1. However, when performing entropic regularized OT, the elements of $\mathbf{Q}^*$ are in [0,1]. In practice, we employ two strategies for label update: hard labeling and soft labeling. In hard labeling, we update $\mathbf{Q}$ as:

$$Q^i_{js}=\begin{cases}1, & \text{if } s=\arg\max_k\ q(k|\mathbf{x}_{i,j})\\ 0, & \text{otherwise}\end{cases},\quad i=1,2,\cdots,m. \tag{6}$$

In soft labeling, we directly use the labels obtained by entropic regularized OT. In the experimental part, we provide the performance comparison on hard and soft labeling: Hard labeling strategy outperforms the soft labeling in convolution network while the soft labeling is superior to hard labeling in fully-connected network. As a results, we employ the hard labeling result to report the final performance for CIFAR-10 and CIFAR-100, while soft labeling for MNIST, F-MNIST, and K-MNIST.

4.3 The entropic regularized OT based LLP algorithm (EROT-LLP)

In practice, we consider the LLP problem in every single bag, and perform the clustering in one bag, with the proportions as the constraint for instances number in each cluster. In detail, we conduct the constrained OT problem (4) with respect to $\mathbf{Q}^i$ and $\mathbf{P}^i$, with minor revision on $U(\boldsymbol{p},\boldsymbol{b})$, i.e., $U(\boldsymbol{p}_i,\boldsymbol{b}_i)$, where $\boldsymbol{b}_i=\mathbb{1}_{n_i}/n_i$. On the other hand, we can instead solve the entropic regularized OT problem (5) with the same revision as (7) to accelerate the training, as well as obtain non-sparse solution to perform soft labeling:

$$\mathbf{L}^{1/\lambda}_{-\log\mathbf{P}}(\boldsymbol{p}_i,\boldsymbol{b}_i)=\min_{\mathbf{Q}^i\in U(\boldsymbol{p}_i,\boldsymbol{b}_i)}\langle\mathbf{Q}^i,-\log\mathbf{P}\rangle-\frac{1}{\lambda}\mathbf{H}(\mathbf{Q}^i). \tag{7}$$

Based on the Sinkhorn's algorithm for the entropic regularized OT (Cuturi, 2013), we describe the entropic regularized OT based LLP algorithm, LLP-EROT, in Algorithm 1, as a realization of OT-LLP framework.

Algorithm 1 LLP based on the entropic regularized OT (LLP-EROT)

Require: $\mathcal{D} = \{(\mathbf{B}_i, \boldsymbol{p}_i)\}_{i=1}^m$, $\lambda \in (0, +\infty)$, the threshold $\varepsilon > 0$, and $\delta > \varepsilon$, the initialization $\mathbf{P}^{\Delta} = \text{diag}\{\mathbf{P}^i\}_{i=1}^m = \text{diag}\{\mathbb{1}_{K\times n_i}/(n_i K)\}_{i=1}^m$, $\boldsymbol{v}^{(0)} = \mathbb{1}_{n_i}$, and $\boldsymbol{b}_i = \frac{1}{n_i}\mathbb{1}_{n_i}$.

1: **while** $\delta > \varepsilon$ **do**

2: **for** each $i \in [\![m]\!]$ **do**

3: Solve the entropic regularized OT problem (7) by iteratively update to obtain $\mathbf{Q}^i$ for the assignment of bag i, using $\boldsymbol{Q}_l^i = \text{diag}\{\boldsymbol{u}^{(l)}\}\mathbf{K}^{\lambda}\text{diag}\{\boldsymbol{v}^{(l)}\}$, with $\mathbf{K}^{\lambda} = \exp\{\lambda \log \mathbf{P}^i\}$ and

$$\boldsymbol{u}^{(l)} = \boldsymbol{p}_i./\mathbf{K}^{\lambda}\boldsymbol{v}^{(l)} \;\; and \;\; \boldsymbol{v}^{(l+1)} = \boldsymbol{b}_i./(\mathbf{K}^{\lambda})^{\intercal}\boldsymbol{u}^{(l)} \tag{8}$$

4: $\boldsymbol{Q}^i = \lim_{l\to+\infty} \boldsymbol{Q}_l^i$. (The convergence is element-wise and proved in (Peyré et al., 2019), Remark 4.8.)

5: **end for**

6: Combine $\{\mathbf{Q}^i\}_{i=1}^m$ as the diagonal element to obtain the block diagonal matrix $\mathbf{Q} = \text{diag}\{\mathbf{Q}^i\}_{i=1}^m$.

7: Fixing $\mathbf{Q}$, solve the following unconstrained programming (9) with respect to the network parameters $\phi = (\varphi, \theta)$ on the whole training data:

$$\min_{\phi=(\varphi,\theta)} CE(p_\phi, q) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q(y|\mathbf{x}_i)\log p_\phi(y|\mathbf{x}_i). \tag{9}$$

8: $\delta = \|\mathbf{P} - \mathbf{P}^{\Delta}\|_F$.

9: $\mathbf{P}^{\Delta} = \mathbf{P}$.

10: **end while**

Ensure: The final network parameters $\phi = (\varphi, \theta)$.

5 Numerical Experiments

In order to demonstrate that our proposed OL-LLP framework is model-agnostic, in this section, we conduct extensive numerical experiments to study the improvement of former LLP methods, when combined with OT-LLP. The evaluation strategy is to conduct a two-phase training, where the first phase is to train with the former models, and OT-LLP is intergrated in the second phase. Four benchmark datasets: MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100, are used in our experiments. The comparisons are performed on two recently proposed algorithms DLLP (Ardehaly & Culotta, 2017) and LLP-GAN (Liu et al., 2019).

5.1 Experimental setting

5.1.1 Label proportions generation

As there is no off-the-shelf LLP datasets, we first generate the bag-based LLP datasets, and obtain the label proportions with four supervised benchmark datasets. Following the setting from Liu et al. (2019), we construct four kinds of bags, with bag sizes of 16, 32, 64, and 128, respectively. In order to avoid the influence of different label distributions, the bag setting is fixed across different algorithms.

5.1.2 TRAINING SETTING

We choose a 5-hidden-layer fully connected network for MNIST, K-MNIST, and F-MNIST, and a conv-based 13-layer max-pooling network for CIFAR-10. The details of network architectures are given in Appendix A.4. Meanwhile, ADAM optimizer is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is $1e-4$ consistently for all datasets, divided by 2 for every 100 epochs. Data augmentation is employed for CIFAR-10 and CIFAR-100 by random horizontal flip and random crop with padding the original images.

5.2 OVERALL ACCURACY

We first provide the overall accuracy of DLLP and OT-LLP on five datasets in Table 1. As introduced above, OT-LLP means a two-stage training, where the first stage is to train the KL divergence based DLLP as the *teacher model*, and the second one is to update the *student model* based on the proposed OT-LLP framework (*c.f.* Algorithm 1). In practice, the first stage can be any other previous deep LLP algorithms (Dulac-Arnold et al., 2019; Liu et al., 2019; Tsai & Lin, 2019), then using our model to further boost their performance.

Table 1: Test accuracy rates and standard deviations (%) on benchmark datasets with different bag sizes.

Dataset	Algorithm	Bag Size			
		16	32	64	128
MNIST	DLLP	98.47 (0.09)	98.40 (0.10)	98.01 (0.16)	97.14 (0.15)
	OT-LLP	**98.63** (0.03)	**98.59** (0.04)	**98.35** (0.06)	**97.82** (0.09)
F-MNIST	DLLP	88.36 (0.29)	87.01 (0.23)	85.53 (0.28)	82.93 (0.21)
	OT-LLP	**89.31** (0.12)	**87.89** (0.11)	**86.75** (0.21)	**83.98** (0.33)
K-MNIST	DLLP	92.58 (0.22)	92.03 (0.23)	89.01 (0.29)	82.14 (0.28)
	OT-LLP	**92.95** (0.15)	**92.44** (0.23)	**90.31** (0.21)	**82.54** (0.35)
CIFAR-10	DLLP	88.78 (0.37)	84.29 (0.66)	66.65 (1.19)	39.14 (0.78)
	OT-LLP	**90.55** (0.31)	**88.24** (0.31)	**76.26** (0.48)	**44.88** (0.51)
CIFAR-100	DLLP	63.47 (0.61)	48.50 (0.66)	1.65 (0.19)	1.14 (0.18)
	OT-LLP	**66.21** (0.46)	**59.66** (0.34)	**4.86** (0.18)	**2.27** (0.19)

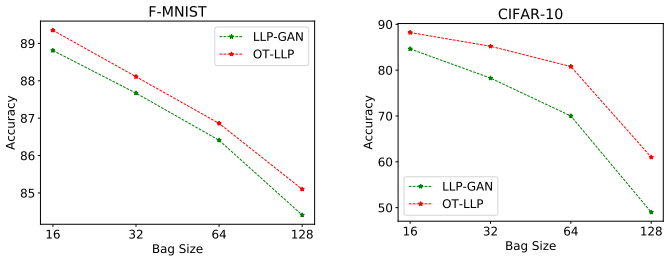
In Table 1, we observe significant improvement on DLLP by adding OT-LLP as second stage. In particular, the advantage is more apparent for CIFAR-10 and CIFAR-100, which are two harder datasets compared with the other three.

5.3 COMBINING WITH OTHER METHOD

As shown above, our model is orthogonal to previous KL-based LLP algorithms, and can be combined with these methods by adding OT-LLP as the second phase. In this section, we further investigate the improvement with our framework on datasets F-MNIST and CIFAR-10 when the first stage is LLP-GAN (Liu et al., 2019), which is the currently SoTA LLP solver. The final performance is shown in Figure 1, where the performance of LLP-GAN is obtained within 300 epochs. Similar to Table 1, our model can also boost LLP-GAN by a large margin.

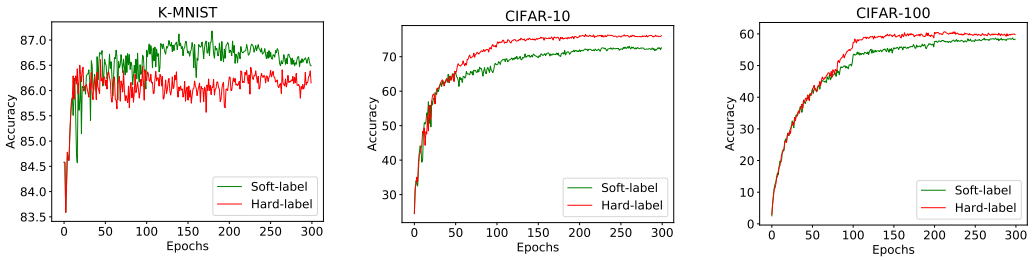
5.4 HARD-LABEL VS. SOFT-LABEL

In our framework, we can employ two pseudo labeling strategies: hard-label and soft-label. The update details are shown in (6). To further study the difference of these two labelings, we compare their performance under different bag sizes on K-MNIST, CIFAR-10, and CIFAR-100. Furthermore, the performance of first



(a) F-MNIST (b) CIFAR-10

Figure 1: The accuracy rate (%) comparison with LLP-GAN on F-MNIST and CIFAR-10.



(a) K-MNIST (b) CIFAR-10 (c) CIFAR-100

Figure 2: The accuracy rate (%) convergence curves with hard and soft labelings.

stage is fixed with soft-label and hard-label for fair comparison. The results are shown in Figure 2, where we provide the convergence curve of accuracy in the second stage for both hard and soft labelings.

From Figure 2, we can find that the performance with hard labeling is superior to that with soft labeling on CIFAR-10 and CIFAR-100, while soft labeling outperforms hard labeling on K-MNIST. Indeed, soft labels are more informative than hard labels. However, it may lead to unstable training process by directly using the outputs of entropic regularized OT, thus degrading the performance with convolution networks.

6 Conclusion

In this paper, we analyze the common challenge in existing LLP approaches, and point out that the minimization on the KL divergence between the prior and posterior class distributions is inadequate to comply with the label proportional information exactly. From this perspective, we propose to solve the LLP problem with a framework to combine instance-level classification and pseudo labeling, and alternately fulfill the optimization on these two objectives. Compared with the former LLP solvers, the main improvement in our method is the introduction of pseudo labeling, which converts the KL divergence based unconstrained optimization into a constrained one, so that the resulting labeling can strictly meet the proportional information and avoid suboptimal solutions. Thanks to OT and its entropic regularized variant, the above two processes can be efficiently conducted with a line search optimizer (e.g., ADAM) on differentiable objective and Sinkhorn's algorithm for entropic regularized OT, respectively. In the experimental part, by integrating OT-LLP as the second phase, we elaborately demonstrate that our framework can further improve the performance of DLLP and LLP-GAN, thus is model-agnostic. Besides that, we empirically study the difference of hard and soft labeling strategies in our framework, and provide suggestions for the practical usage.

REFERENCES

Ehsan Mohammady Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *IEEE International Conference on Data Mining Workshops*, pp. 1017–1024, 2017.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 367–376, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*, 2019.

Jerónimo Hernández-González, Inaki Inza, Lorena Crisol-Ortíz, María A Guembe, María J Iñarra, and Jose A Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical Methods in Medical Research*, pp. 1056–1066, 2018.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874, 2019.

Hendrik Kuck and Nando de Freitas. Learning about individuals from group statistics. *arXiv preprint arXiv:1207.1393*, 2012.

Kuan-Ting Lai, Felix X Yu, Ming-Syan Chen, and Shih-Fu Chang. Video event detection by inferring temporal instance labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2243–2250, 2014.

Thomas Liebig, Marco Stolpe, and Katharina Morik. Distributed traffic flow prediction with label proportions: from in-network towards high performance computation with MPI. In *International Conference on Mining Urban Data-Volume 1392*, pp. 36–43. CEUR-WS. org, 2015.

Jiabin Liu, Bo Wang, Zhiquan Qi, Yingjie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2019.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*, 2018.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 349–364. Springer, 2011.

Tao Sun, Dan Sheldon, and Brendan O'Connor. A probabilistic approach for learning with label proportions applied to the US presidential election. In *IEEE International Conference on Data Mining (ICDM)*, pp. 445–454. IEEE, 2017.

Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization. *arXiv preprint arXiv:1910.13188*, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Felix X Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$SVM for learning with label proportions. In *International Conference on Machine Learning*, 2013.

Felix X Yu, Liangliang Cao, Michele Merler, Noel Codella, Tao Chen, John R Smith, and Shih-Fu Chang. Modeling attributes from category-attribute proportions. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 977–980, 2014a.

Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014b.