# Learning from Label Proportions with Consistency Regularization

**Kuen-Han Tsai** r06922066@csie.ntu.edu.tw
**Hsuan-Tien Lin** htlin@csie.ntu.edu.tw
*Department of Computer Science and Information Engineering, National Taiwan University*

**Editors:** Sinno Jialin Pan and Masashi Sugiyama

## Abstract

The problem of learning from label proportions (LLP) involves training classifiers with weak labels on bags of instances, rather than strong labels on individual instances. The weak labels only contain the label proportion of each bag. The LLP problem is important for many practical applications that only allow label proportions to be collected because of data privacy or annotation cost, and has recently received lots of research attention. Most existing works focus on extending supervised learning models to solve the LLP problem, but the weak learning nature makes it hard to further improve LLP performance with a supervised angle. In this paper, we take a different angle from semi-supervised learning. In particular, we propose a novel model inspired by consistency regularization, a popular concept in semi-supervised learning that encourages the model to produce a decision boundary that better describes the data manifold. With the introduction of consistency regularization, we further extend our study to non-uniform bag-generation and validation-based parameter-selection procedures that better match practical needs. Experiments not only justify that LLP with consistency regularization achieves superior performance, but also demonstrate the practical usability of the proposed procedures.

**Keywords:** Learning from Label Proportions, Consistency Regularization, Semi-supervised Learning, Weakly-supervised Learning

## 1. Introduction

In traditional supervised learning, a classifier is trained on a dataset where each instance is associated with a class label. However, label annotation can be expensive or difficult to obtain for some applications. Take the embryo selection as an example (Hernández-González et al., 2018). To increase the pregnancy rate, clinicians would transfer multiple embryos to a mother at the same time. However, clinicians are unable to know the outcome of a particular embryo due to limitations of current medical techniques. The only thing we know is the proportion of embryos that implant successfully. To increase the success rate of embryo implantation, clinicians aim to select high-quality embryos through the aggregated results. In this case, only label proportions about groups of instances are provided to train the classifier, a problem setting known as learning from label proportions (LLP).

In LLP, each group of instances is called a *bag*, which is associated with a *proportion label* of different classes. A classifier is then trained on several bags and their associated proportion labels in order to predict the class of each unseen instance. Recently, LLP has attracted much attention among researchers because its problem setting occurs in many real-life scenarios. For example, the census data and medical databases are all provided in the form of label proportion data due to privacy issues (Patrini et al., 2014; Hernández-González et al., 2018). Other LLP applications include fraud detection (Rueping, 2010), object recognition (Kuck and de Freitas, 2012), video event detection (Lai et al., 2014), and ice-water classification (Li and Taylor, 2015).

The challenge in LLP is to train models using the weak supervision of proportion labels. To overcome this issue, prior work seeks to estimate either the individual label (Yu et al., 2013; Dulac-Arnold et al., 2019) or the mean of each class by proportion labels (Quadrianto et al., 2009; Patrini et al., 2014). In terms of the weak supervision, the LLP scenario is similar to the problem of semi-supervised learning, where most data examples are unlabeled and only a few labeled data are provided. Inspired by this perspective, we would like to ask the following question: Is it possible to incorporate semi-supervised techniques to tackle the LLP problem? Dulac-Arnold et al. (2019) first adapt the concept of pseudo-labeling (Lee, 2013), a straightforward semi-supervised technique, to the multi-class LLP setting by proposing the method of Relax Optimal Transport (ROT). The ROT approach seeks to estimate an individual label for each unlabeled instance within a bag and update model parameters alternatively. Nevertheless, a common drawback of pseudo-labeling methods is the propagation and amplification of estimation errors. In particular, inaccurate estimates of the pseudo-labels can misguide the model toward erroneous decision boundaries, which result in a vicious cycle of even worse pseudo-label estimates.

Another popular semi-supervised learning technique is consistency regularization, which enforces network predictions to be consistent when the input is perturbed. Consistency regularization is build on the *smoothness assumption*: if two instances are close to each other, their labels should also be similar In particular, consistency regularization introduces an auxiliary loss term to produce a decision boundary that captures the data manifold. Recently, consistency regularization methods (Verma et al., 2019; Berthelot et al., 2019) have demonstrated outstanding performance in semi-supervised learning.

Motivated by the success of consistency regularization methods, we initiate this study on solving LLP with the help of consistency regularization. We design a framework that iteratively optimizes a combined loss consisting of a bag-level loss of LLP and an instance-level loss for consistency regularization. The framework is general and allows us to study different consistency regularization techniques easily. We propose a method under the framework that uses the Virtual Adversarial Training (Miyato et al., 2018, VAT) technique for consistency regularization. To evaluate the effectiveness of the proposed method, we conduct experiments on three standard image data sets along with a standard bag generation benchmark that randomly and uniformly groups examples into same-sized bags. Experimental results on this benchmark across different data sets demonstrate that our proposed method achieves state-of-the-art performance.

However, most existing LLP works assume that bags of data are randomly generated, which is not the case for many real-world applications. For example, the data of population census are collected on region, age, or occupation with varying group sizes. Therefore, we further explore a new bag generation procedure - K-means bag generation, where data examples are grouped by the feature correlation. The new procedure of K-means bag generation not only better fits the practical LLP scenario, but is also more challenging and worth-studying. Last, since the hyperparameter selection requires a validation set with labeled data for computing the classification error, it would be more practical if the process of hyperparameter selection relies only on the proportion labels. To alleviate the need for labeled data, we propose four bag-level validation metrics, which compute the validation error on bags of instances. We empirically study the Pearson correlation coefficient between the bag-level validation error and the instance-level test error. Surprisingly, the empirical results demonstrate the feasibility of hyperparameter selection with only proportion labels.

This paper aims to resolve the previous problems. Our main contributions are listed as follows:

- We first apply a semi-supervised learning technique, consistency regularization, to the multi-class LLP problem. Consistency regularization considers an auxiliary loss term to enforce network predictions to be consistent when its input is perturbed. By exploiting the unlabeled instances, our method captures the latent structure of data and obtains the SOTA performance on three benchmark datasets.

- We explore a new bag generation procedure—the K-means bag generation, where training data are grouped by attribute similarity. Using this setup can help train models that are more applicable to actual LLP scenarios.

- We show that it is possible to select models with a validation set consisting of only bags and associated proportion labels. The experiments demonstrate correlation between bag-level validation error and instance-level test error. This potentially reduces the need of a validation set with instance-level labels.

## 2. Preliminary

### 2.1. Learning from label proportions

We consider the multi-class classification problem in the LLP setting in this paper. Let $\boldsymbol{x}_i \in \mathbb{R}^D$ be a feature vector of $i$-th example and $y_i \in \{1, \ldots, L\}$ be a class label of $i$-th example, where $L$ is the number of different classes. We define $\boldsymbol{e}^{(j)}$ to be a standard basis vector $[0, \ldots, 1, \ldots, 0]$ with 1 at $j$-th position and $\Delta_L = \{\boldsymbol{p} \in \mathbb{R}_+^L : \sum_i^L \boldsymbol{p}_i = 1\}$ to be a probability simplex. In the setting of LLP, each individual label $y_i$ is hidden from the training data. On the other hand, the training data are aggregated by a bag generation procedure. We are given $M$ bags $B_1, \ldots, B_M$, where each bag $B_m$ contains a set $\mathbb{X}_m$ of instances and a proportion label $\boldsymbol{p}_m$, defined by

$$\boldsymbol{p}_m = \frac{1}{|\mathbb{X}_m|} \sum_{i:\boldsymbol{x}_i \in \mathbb{X}_m} \boldsymbol{e}^{(y_i)}, \qquad \bigcup_{m=1}^{M} \mathbb{X}_m = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}.$$

We do not require each subset to be disjoint. Also, each bag may have different size. The task of LLP is to learn an individual-level classifier $f_\theta : \mathbb{R}^D \to \Delta_L$ to predict the correct label $y = \arg\max_i f_\theta(\boldsymbol{x})_i$ for a new instance $\boldsymbol{x}$. Figure 1 illustrates the setting of learning from label proportions in the multi-class classification (Dulac-Arnold et al., 2019).

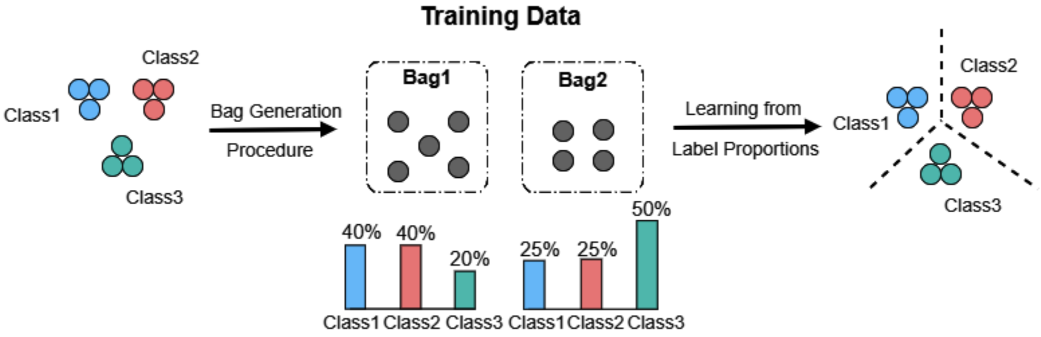


Figure 1: An illustration of multi-class learning from label proportions. Before training, the data are grouped according to a bag generation procedure. During the training stage, we are given bags of unlabeled data and their corresponding proportion labels. The goal of LLP is to learn an individual-level classifier.

### 2.2. Proportion loss

The feasibility of the binary LLP setting has been theoretically justified by Yu et al. (2014). Specifically, Yu et al. (2014) propose the framework of *Empirical Proportion Risk Minimization* (EPRM), proving that the LLP problem is PAC-learnable under the assumption that bags are i.i.d sampled from an unknown probability distribution. The EPRM framework provides a generalization bound on the expected proportion error and guarantees to learn a probably approximately correct proportion predictor when the number of bags is large enough. Furthermore, the authors prove that the instance label error can be bounded by the bag proportion error. That is, a decent bag proportion predictor guarantees a decent instance label predictor.

Based on the profound theoretical analysis, a vast number of LLP approaches learn an instance-level classifier by directly minimizing the proportion loss without acquiring the individual labels. To be more precise, given a bag $B = (\mathbb{X}, \boldsymbol{p})$, an instance-level classifier $f_\theta$ and a divergence function $d_{\text{prop}} : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, the proportion loss penalizes the difference between the real proportion label $\boldsymbol{p}_m$ and the estimated proportion label $\hat{\boldsymbol{p}} = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} f_\theta(\boldsymbol{x})$, which is an average of the instance predictions within a bag. Thus, the proportion loss $\mathcal{L}_{\text{prop}}$ can be defined as follows:

$$\mathcal{L}_{\text{prop}}(\theta) = d_{\text{prop}}(\boldsymbol{p}, \hat{\boldsymbol{p}}).$$

The commonly used divergence functions are $L^1$ and $L^2$ function in prior work (Musicant et al., 2007; Yu et al., 2013). Ardehaly and Culotta (2017) and Dulac-Arnold et al. (2019), on the other hand, consider the cross-entropy function for the multi-class LLP problem.

### 2.3. Consistency regularization

Since collecting labeled data is expensive and time-consuming, the semi-supervised learning approaches aim to leverage a large amount of unlabeled data to mitigate the need for labeled data. There are many semi-supervised learning methods, such as pseudo-labeling (Lee, 2013), generative approaches (Kingma et al., 2014), and consistency-based methods (Laine and Aila, 2016; Miyato et al., 2018; Tarvainen and Valpola, 2017). Consistency-based approaches encourage the network to produce consistent output probabilities between unlabeled data and the perturbed examples. These methods rely on the smoothness assumption (Chapelle et al., 2009): if two data points $x_i$ and $x_j$ are close, then so should be the corresponding output distributions $y_i$ and $y_j$. Then, the consistency-based approaches can enforce the decision boundary to traverse through the low-density region. More precisely, given a perturbed input $\hat{\boldsymbol{x}}$ taken from the input $\boldsymbol{x}$, consistency regularization penalizes the distinction of model predictions between $f_\theta(\boldsymbol{x})$ and $f_\theta(\hat{\boldsymbol{x}})$ by a distance function $d_{\text{cons}} : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$. The consistency loss can be written as follows:

$$\mathcal{L}_{\text{cons}}(\theta) = d_{\text{cons}}(f_\theta(\boldsymbol{x}), f_\theta(\hat{\boldsymbol{x}})).$$

Modern consistency-based methods (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Miyato et al., 2018; Verma et al., 2019; Berthelot et al., 2019) differ in how perturbed examples are generated for the unlabeled data. Laine and Aila (2016) introduce the Π-Model approach, which uses the additive Gaussian noise for perturbed examples and chooses the $L^2$ error as the distance function. However, a drawback to Π-Model is that the consistency target $f_\theta(\hat{\boldsymbol{x}})$ obtained from the stochastic network is unstable since the network changes rapidly during training. To address this problem, Temporal Ensembling (Laine and Aila, 2016) takes the exponential moving average of the network predictions as the consistency target. Mean Teacher (Tarvainen and Valpola, 2017), on the other hand, proposes averaging the model parametes instead of network predictions. Overall, the Mean Teacher approach significantly improves the quality of consistency targets and the empirical results on semi-supervised benchmarks.

Instead of applying stochastic perturbations to the inputs, Virtual Adversarial Training or VAT (Miyato et al., 2018) computes the perturbed examples $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{r}_{\text{adv}}$, where

$$\boldsymbol{r}_{\text{adv}} = \underset{\boldsymbol{r}:||\boldsymbol{r}||_2 \leq \epsilon}{\arg\max}\, D_{\text{KL}}(f_\theta(\boldsymbol{x}) \| f_\theta(\boldsymbol{x} + \boldsymbol{r})). \tag{1}$$

That is, the VAT approach attempts to generate a perturbation which most likely causes the model to misclassify the input in an adversarial direction. Finally, the VAT approach adopts Kullback-Leibler (KL) divergence to compute the consistency loss. In comparison to the stochastic perturbation, the VAT approach demonstrates the greater effectiveness in the semi-supervised learning problem.

## 3. LLP with consistency regularization

With regards to weak supervision, the LLP scenario is similar to the semi-supervised learning problem. In the semi-supervised learning setting, only a small portion of training examples is labeled. On the other hand, in the LLP scenario, we are given the weak supervision of label proportions instead of the strong label on individual instances. Both settings are challenging since most training examples do not have individual labels. To address this challenge, semi-supervised approaches seek to exploit the unlabeled examples to further capture the latent structure of data.

Motivated by these semi-supervised approaches, we combine the idea of leveraging the unlabeled data into the LLP problem. We make the same smoothness assumption and

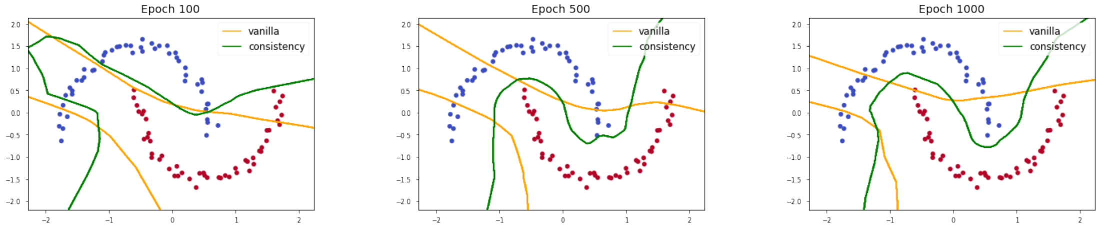


Figure 2: In this toy example, we generate 5 bags, each of which contains 20 data points uniformly sampled from the "two moons" dataset without replacement. The vanilla approach, which simply optimizes the proportion loss, suffers from poor performance as the label information is insufficient. In contrast, the "two moons" can be effectively separated into two clusters by LLP with consistency regularization. Our method enforces the network to produce consistent outputs for perturbed examples, and thus help capture the underlying structure of the data.

introduce a new concept incorporating consistency regularization with LLP. In particular, we consider the typical cross-entropy function between real label proportions and estimated label proportions. Given a bag $B = (\mathbb{X}, \boldsymbol{p})$, we define the proportion loss $L_{\text{prop}}$ as follows:

$$\mathcal{L}_{\text{prop}}(\theta) = -\sum_{i=1}^{L} \boldsymbol{p}_i \log \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} f_\theta(\boldsymbol{x})_i.$$

Interestingly, the proportion loss $\mathcal{L}_{\text{prop}}$ boils down to standard cross-entropy loss for fully-supervised learning when the bag size is one. To learn a decision boundary that better reflects the data manifold, we add an auxiliary consistency loss that leverages the unlabeled data. More formally, we compute the average consistency loss across all instances within the bag. Given a bag $B = (\mathbb{X}, \boldsymbol{p})$, the consistency loss $\mathcal{L}_{\text{cons}}$ can be written as follows:

$$\mathcal{L}_{\text{cons}}(\theta) = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} d_{\text{cons}}(f_\theta(\boldsymbol{x}), f_\theta(\hat{\boldsymbol{x}})),$$

where $d_{\text{cons}}$ is a distance function, and $\hat{\boldsymbol{x}}$ is a perturbed input of $\boldsymbol{x}$. We can use any consistency-based approach to generate the perturbed examples and compute the consistency loss. Finally, we mix the two loss functions $\mathcal{L}_{\text{prop}}$ and $\mathcal{L}_{\text{cons}}$ with a hyperparameter $\alpha > 0$, yielding the combined loss $\mathcal{L}$ for LLP:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{prop}}(\theta) + \alpha \mathcal{L}_{\text{cons}}(\theta),$$

where $\alpha$ controls the balance between the bag-level estimation of proportion labels and instance-level consistency regularization.

To understand the intuition behind combining consistency regularization into LLP, we follow the Π-Model approach (Laine and Aila, 2016) to adopt the stochastic Gaussian noise as the perturbation and to use $L^2$ as the distance function $d_{\text{cons}}$ in a toy example. Figure 2 illustrates how our method is able to produce a decision boundary that passes through the low-density region and captures the data manifold. On the other hand, the vanilla approach, which simply optimizes the proportion loss, gets easily stuck at a poor solution due to the lack of label information. This toy example shows the advantage of applying consistency regularization into LLP.

**Algorithm 1** LLP-VAT algorithm

```
Require: $\mathcal{D} = \{(\mathbb{X}_m, \boldsymbol{p}_m)\}_{m=1}^{M}$: collection of bags
Require: $f_\theta(\boldsymbol{x})$: instance-level classifier with trainable parameters $\theta$
Require: $g(\boldsymbol{x}; \theta) = \boldsymbol{x} + \boldsymbol{r}_{\text{adv}}$: VAT augmentation function according to Equation 1
Require: $w(t)$: ramp-up function for increasing the weight of consistency regularization
Require: $T$: total number of iterations
  for $t = 1, \ldots, T$ do
    for each bag $(\mathbb{X}, \boldsymbol{p}) \in \mathcal{D}$ do
      $\hat{\boldsymbol{p}} \leftarrow \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} f_\theta(\boldsymbol{x})$                                   ▷ Estimated proportion label
      $\mathcal{L}_{\text{prop}} = -\sum_{i=1}^{L} \boldsymbol{p}_i \log \hat{\boldsymbol{p}}_i$                                   ▷ Proportion loss
      $\mathcal{L}_{\text{cons}} = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} D_{\text{KL}}(f_\theta(\boldsymbol{x}) \| f_\theta(g(\boldsymbol{x}; \theta)))$     ▷ Consistency loss
      $\mathcal{L} = \mathcal{L}_{\text{prop}} + w(t) \cdot \mathcal{L}_{\text{cons}}$                                   ▷ Total loss
      update $\theta$ by gradient $\nabla_\theta \mathcal{L}$                                   ▷ e.g. SGD, Adam
    end for
  end for
  return $\theta$
```

According to Miyato et al. (2018), VAT is more effective and stable than Π-Model due to the way it generates the perturbed examples. For each data example, the Π-Model approach stochastically perturbs inputs and trains the model to assign the same class distributions to all neighbors. In contrast, the VAT approach focuses on neighbors that are *sensitive* to the model. That is, VAT aims to generate a perturbed input whose prediction is the most different from the model prediction of its original input. The learning of VAT approach tends to be more effective in improving model generalization. Therefore, we adopt the VAT approach to compute the consistency loss for each instance in the bag. Additionally, to prevent the model from getting stuck at a local optimum in the early stage, we use the exponential ramp-up scheduling function (Laine and Aila, 2016) to increase the consistency weight gradually to the maximum value $\alpha$. The full algorithm of LLP with VAT (LLP-VAT) is described in Algorithm 1.

## 4. Experiments

We evaluate our LLP-VAT on three benchmarks, including SVHN, CIFAR10, and CIFAR100. We choose hyperparameters using a validation set without individual labels for model selection. For each run of experiment, we report the test instance accuracy averaged over the last 10 epochs. Experiments on each benchmark are repeated 3 times. The full experiment details are provided in the appendix A.

### 4.1. Uniform bag generation

For convenience, most LLP works validate their proposed methods with the uniform bag generation where the training data are randomly partitioned into bags of the same size. We evaluate our method using this bag generation procedure with the bag size $n \in \{16, 32, 64, 128, 256\}$.

Table 1: Test accuracy with the uniform bag generation. The performance of the vanilla approach with a bag size of one corresponds to the fully-supervised setting.

| | | Bag Size | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset | Method | 1 | 16 | 32 | 64 | 128 | 256 |
| SVHN | vanilla | 95.52 | 95.49 ± 0.20 | 95.30 ± 0.10 | 93.72 ± 0.61 | 89.96 ± 1.02 | 14.44 ± 1.78 |
| | ROT | - | 95.29 ± 0.15 | 94.82 ± 0.11 | 94.04 ± 0.32 | 91.73 ± 0.49 | **17.2** ± **3.64** |
| | LLP-VAT | - | **95.83** ± **0.20** | **95.57** ± **0.22** | **94.41** ± **0.18** | **92.19** ± **0.83** | 15.24 ± 3.58 |
| CIFAR10 | vanilla | 90.64 | 88.81 ± 0.13 | **85.41** ± **0.42** | 70.96 ± 0.65 | 42.89 ± 4.95 | 38.69 ± 3.59 |
| | ROT | - | 86.86 ± 0.13 | 78.04 ± 0.92 | 61.61 ± 2.58 | **51.42** ± **2.52** | 37.89 ± 4.49 |
| | LLP-VAT | - | **89.30** ± **0.10** | 84.99 ± 0.69 | **71.52** ± **2.92** | 51.11 ± 1.01 | **39.29** ± **3.76** |
| CIFAR100 | vanilla | 59.79 | 58.15 ± 1.29 | 47.28 ± 0.88 | 20.39 ± 0.59 | 6.05 ± 1.01 | 2.98 ± 0.23 |
| | ROT | - | 55.35 ± 1.54 | 47.35 ± 1.27 | **29.08** ± **0.58** | 8.25 ± 0.37 | 2.86 ± 0.29 |
| | LLP-VAT | - | **59.53** ± **0.12** | **48.91** ± **0.31** | 24.20 ± 3.29 | **8.51** ± **0.86** | **3.85** ± **0.52** |

We drop the last incomplete bag if the number of training data is indivisible by the bag size. Table 1 shows the experimental results for the LLP scenario with a uniform bag generation.

In comparison to the vanilla approach, our LLP-VAT significantly improves the performance on CIFAR10 and CIFAR100. This indicates that applying consistency regularization into LLP does help learn a better classifier. As for SVHN, since the test accuracy is close to the fully-supervised performance when the bag size is small, there is no clear difference among three methods. In addition, the results also show that the performance of ROT is unstable and lead us to conclude that the unhelpful pseudo-labels would easily result in a worse classifier. Conversely, our LLP-VAT is more stable and obtains better test accuracy in most cases.

### 4.2. K-means bag generation

In this section, we further investigate our LLP-VAT in a more practical scenario. We observe that the uniform bag generation barely fits the real-world LLP situation because of following two reasons. First, the real-life data are usually grouped by attribute similarity instead of uniformly sampled. Second, each bag may have different bag sizes, i.e., the distribution of bag sizes is diverse. Consider the US presidential election results (Sun et al., 2017), where the statistics of voting results are collected by geological regions (e.g., states). Also, each state have varying number of voters. Therefore, we introduce a new bag generation procedure—the K-means bag generation, where we cluster examples into bags by the K-means algorithm. Although those bags generated from the K-means bag generation are dependent on each other, violating the i.i.d. assumption, this setting is both challenging and worth-studying.

Since we perform experiments on image datasets, it is meaningless to cluster data examples based on RGB pixels. We first adopt the principle component analysis algorithm, which is an unsupervised dimension reduction technique, to project the data into a low-dimensional representation space. This space may capture more important patterns in an images. Then we group the low-dimensional representations of the images following the K-means bag generation procedure. We conduct experiments with the number of clusters $K \in \{3120, 1560, 780, 390, 195\}$ on CIFAR10 and CIFAR100, and

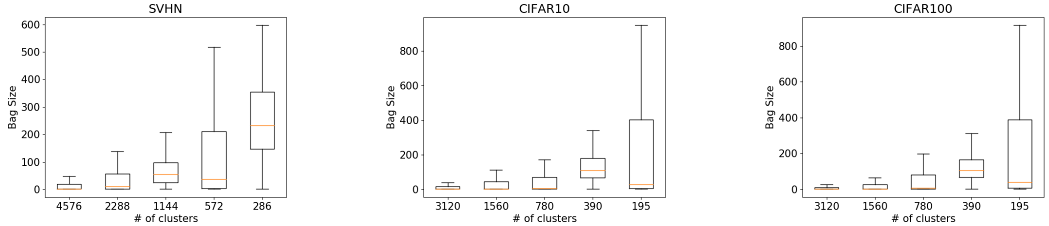


Figure 3: The distribution of bag sizes from the K-means procedure on three benchmarks. When the number of clusters increases, the distribution of bag sizes becomes various.

Table 2: Test accuracy with the K-means bag generation on SVHN.

| | | K | | | | |
|---|---|---|---|---|---|---|
| Dataset | Method | 4576 | 2288 | 1144 | 572 | 286 |
| SVHN | vanilla | 66.3 ± 16.58 | 79.25 ± 2.07 | 91.49 ± 0.90 | 90.79 ± 0.36 | 92.68 ± 0.57 |
| | LLP-VAT | **68.37 ± 19.04** | **80.36 ± 1.53** | **92.95 ± 0.56** | **90.94 ± 0.66** | **93.27 ± 0.14** |

$K \in \{4576, 2288, 1144, 572, 286\}$ on SVHN. These numbers are selected to match the number of proportion labels in the uniform bag generation procedure. The distribution of bag sizes generated from the K-means procedure are shown in Figure 3.

For experiments, we do not compare our proposed method to the ROT loss, which needs to estimate individual labels iteratively for each bag. The procedure of the ROT algorithm is time-consuming and cannot be accelerated if bags are of varying sizes. Besides, for the K-means bag generation, there may be some large bags when the value of $K$ is small. Because of the limited computational resource, we take a subsample in each bag if the bag size is larger than the threshold of 256. Particularly, when a large bag is sampled, we randomly sample 256 instances and assign the original label proportions to the reduced bag.

The experimental results of the K-means bag generation are shown in Table 2 and Table 3. Although this scenario violates the i.i.d. assumption, the results demonstrate that it is feasible to learn an instance-level classifier by simply minimizing the proportion loss. Also, our LLP-VAT significantly brings benefits for the k-means bag generation scenario on SVHN and CIFAR10, while showing comparable performance on CIFAR100. Interestingly, the performance of a model is not well-correlated with the value of $K$. One possible reason is that we might drop informative bags as we randomly split bags into validation and training.

### 4.3. Validation metrics

Many modern machine learning models require a wide range of hyperparameter selections about the architecture, optimizer and regularization. However, for the realistic LLP scenario, we have no access to labeled instances during training. It is crucial to choose appropriate hyperparameters based on the bag-level validation error that is computed with only proportion labels. To evaluate the performance at the bag level, we consider four validation metrics: soft $L^1$ error, hard $L^1$ error, soft KL divergence, and hard KL divergence. Their definitions are given as follows. First, we define the output probabilities of an instance as the soft prediction and its one-hot encoding as the hard prediction. For each bag, we

Table 3: Test accuracy with the K-means bag generation on CIFAR10 and CIFAR100.

| | | K | | | | |
|---|---|---|---|---|---|---|
| Dataset | Method | 3120 | 1560 | 780 | 390 | 195 |
| CIFAR10 | vanilla | 36.03 ± 5.10 | 49.08 ± 0.71 | 53.29 ± 7.94 | 66.09 ± 0.39 | 73.91 ± 1.39 |
| | LLP-VAT | **37.76 ± 0.65** | **50.77 ± 0.59** | **54.46 ± 2.97** | **68.01 ± 0.83** | **76.85 ± 0.56** |
| CIFAR100 | vanilla | 9.26 ± 1.27 | 13.85 ± 1.42 | **15.33 ± 0.71** | **22.85 ± 0.70** | 37.88 ± 0.71 |
| | LLP-VAT | **9.32 ± 1.05** | **15.11 ± 0.45** | 15.31 ± 0.30 | 22.35 ± 0.82 | **38.74 ± 0.67** |

Table 4: The Pearson correlation coefficient between the test error rate and the following validation metrics on benchmarks.

| | Uniform | | | K-means | | |
|---|---|---|---|---|---|---|
| | SVHN | CIFAR10 | CIFAR100 | SVHN | CIFAR10 | CIFAR100 |
| Hard $L^1$ | **0.96** | 0.81 | **0.85** | **0.98** | **0.76** | **0.47** |
| Soft $L^1$ | 0.83 | 0.36 | -0.47 | 0.85 | 0.68 | -0.17 |
| Hard KL | 0.83 | 0.03 | 0.71 | 0.76 | 0.40 | 0.13 |
| Soft KL | 0.65 | **0.85** | -0.09 | 0.60 | **0.76** | 0.44 |

then compute the estimated label proportions by averaging these soft or hard predictions. Finally, we use the $L^1$ error or KL divergence to measure the bag-level prediction error.

To investigate the relationship between the instance-level test error and the bag-level validation error, we compute the Pearson correlation coefficient between them on models trained for 400 epochs. The results are shown in Table 4. Surprisingly, we find that the hard $L^1$ error has a strong positive correlation to test error rate on all benchmarks. This implies that it is feasible to select hyperparameters with only label proportions in realistic LLP scenarios. Interestingly, our finding is coherent to Yu et al. (2013). Although their and our works both adopt the hard $L^1$ error for model selection, we focus on the multi-class LLP scenario instead of the binary classification problem they considered. Therefore, we suggest future multi-class LLP works could adopt the hard $L^1$ validation metric for model selection.[1]

## 5. Related work

Kuck and de Freitas (2012) first introduce the LLP scenario and formulate the probabilistic model with the MCMC algorithm to generate consistent label proportions. Several following works (Chen et al., 2006; Musicant et al., 2007) extend the LLP setting to a variety of standard supervised learning algorithms. Without directly inferring instance labels, Quadrianto et al. (2009) propose a Mean Map algorithm with exponential-family parametric models. The algorithm uses empirical mean operators of each bag to solve a convex optimization problem. However, the success of the Mean Map algorithm is based on a strong assump-

[1] Nevertheless, we do not suggest using our validation metric for early stopping since the correlation is computed after the model converges.

tion that the class-conditional distribution of data is independent of bags. To loosen the restriction, Patrini et al. (2014) propose a Laplacian Mean Map algorithm imposing an additional Laplacian regularization. Nevertheless, these Mean Map algorithms suffer from a fundamental drawback: they require the classifier to be a linear model.

Several works tackle the LLP problem from Bayesian perspectives. For example, Fan et al. (2014) propose an RBM-based generative model to estimate the group-conditional likelihood of data. Hernández-González et al. (2013), on the other hand, develop a Bayesian classifier with an EM algorithm. Recently, Sun et al. (2017) propose a graphical model using counting potential to predict instance labels for the US presidential election. Furthermore, other works (Chen et al., 2009; Stolpe and Morik, 2011) adopt a k-means approach to cluster training data by label proportions. While some works (Fan et al., 2014; Sun et al., 2017) claim that they are suitable for large-scale settings, both Bayesian methods and clustering-based algorithms are rather inefficient and computationally expensive when applied to large image datasets.

Another line of work adopts a large-margin framework for the problem of LLP. Stolpe and Morik (2011) propose a variant of support vector regression using the inverse calibration method to estimate the class-conditional probability for bags. On the other hand, Yu et al. (2013) propose a procedure that alternates between assigning a label to each instance, also known as *pseudo-labeling* in the literature, and fitting an SVM classifier. Motivated by this idea, a number of works (Wang et al., 2015; Qi et al., 2016; Chen et al., 2017) infer individual labels and updated model parameters alternately. One major drawback of SVM-based approaches is that they are tailored for binary classification; they cannot extend to the multi-class classification setting efficiently.

As deep learning has garnered huge success in a number of areas, such as natural language processing, speech recognition, and computer vision, many works leverage the power of neural networks for the LLP problem. Ardehaly and Culotta (2017) are the first to apply deep models to the multi-class LLP setting. Also, Bortsova et al. (2018) propose a deep LLP method learning the extent of emphysema from the proportions of disease tissues. Concurrent to our work, Dulac-Arnold et al. (2019) also considers the multi-class LLP setting with bag-level cross-entropy loss. They introduce a ROT loss that combines two goals: jointly maximizing the probability of instance predictions and minimizing the bag proportion loss.

## 6. Conclusion

In this paper, we first apply a novel semi-supervised learning technique, consistency regularization, to the multi-class LLP problem. Our proposed approach leverages the unlabeled data to learn a decision boundary that better depicts the data manifold. The empirical results validate that our approach obtains better performance than that achieved by existing LLP works. Furthermore, we introduce a non-uniform bag scenario - the K-means bag generation, where training instances are clustered by attribute relationships. This setting simulates more practical LLP situations than the uniform bag generation setting, which is often used in previous works. Lastly, we introduce a bag-level validation metrics, hard $L^1$ error, for model selection with only label proportions. We empirically show that the bag-level hard $L^1$ error has a strong correlation to the test classification error. For real-world applicability, we suggest that multi-class LLP methods relying on hyper-parameter tuning could evaluate their methodology based on the bag-level hard $L^1$ error. In a nutshell, we hope that future LLP work can further explore the ideas presented in this paper.

## Acknowledgments

We would like to acknowledge Chien-Min Yu, Michelle Yuan and Si-An Chen for fruitful discussion and valuable suggestions that helped to improve this article. The work was partially supported by MOST 107-2628-E-002-008-MY3 and 108-2119-M-007-010. We are grateful to the National Center for High-performance Computing for computer time and facilities.

## References

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

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Gerda Bortsova, Florian Dubost, Silas Ørting, Ioannis Katramados, Laurens Hogeweg, Laura Thomsen, Mathilde Wille, and Marleen de Bruijne. Deep learning from label proportions for emphysema quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 768–776. Springer, 2018.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20 (3):542–542, 2009.

Bee-Chung Chen, Lei Chen, Raghu Ramakrishnan, and David R Musicant. Learning from aggregate views. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 3–3. IEEE, 2006.

Shuo Chen, Bin Liu, Mingjie Qian, and Changshui Zhang. Kernel k-means based framework for aggregate outputs classification. In *2009 IEEE International Conference on Data Mining Workshops*, pages 356–361. IEEE, 2009.

Zhensong Chen, Zhiquan Qi, Bo Wang, Limeng Cui, Fan Meng, and Yong Shi. Learning with label proportions based on nonparallel support vector machines. *Knowledge-Based Systems*, 119:126–141, 2017.

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.

Kai Fan, Hongyi Zhang, Songbai Yan, Liwei Wang, Wensheng Zhang, and Jufu Feng. Learning a generative classifier from label proportions. *Neurocomputing*, 139:47–55, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

Jerónimo Hernández-González, Iñaki Inza, and Jose A Lozano. Learning bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12):3425–3440, 2013.

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*, 27(4):1056–1066, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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 *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 2243–2250, 2014.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

Fan Li and Graham Taylor. Alter-cnn: An approach to learning from label proportions with application to ice-water classification. In *Neural Information Processing Systems Workshops (NIPSW) on Learning and privacy with incomplete data and weak supervision*, 2015.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 252–261. IEEE, 2007.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *Advances in neural information processing systems*, 01 2011.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. Learning with label proportions via npsvm. *IEEE transactions on cybernetics*, 47(10):3293–3305, 2016.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(Oct):2349–2374, 2009.

Stefan Rueping. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 911–918, 2010.

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*, pages 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 *2017 IEEE International Conference on Data Mining (ICDM)*, pages 445–454. IEEE, 2017.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*, 2019.

Bo Wang, Zhensong Chen, and Zhiquan Qi. Linear twin svm for learning from label proportions. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 56–59. IEEE, 2015.

Felix X Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$svm for learning with label proportions. *arXiv preprint arXiv:1306.0886*, 2013.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Procedings of the British Machine Vision Conference 2016*, 2016.

## Appendix A. Experiment details

### A.1. Datasets

To evaluate the effectiveness of our proposed method, we conduct experiments on three benchmark datasets, including SVHN (Netzer et al., 2011), CIFAR10, and CIFAR100 (Krizhevsky and Hinton, 2009). The SVHN dataset consists of 32x32 RGB digit images with 73,257 examples for training, 26,032 examples for testing, and 531,131 extra training examples that are not used in our experiments. The CIFAR10 and CIFAR100 datasets both consist of 50,000 training examples and 10,000 test examples. Each example is a 32x32 colored natural image, drawn from 10 classes and 100 classes respectively.

### A.2. Experiment Setup

**Implementation details.** For all experiments in this section, we adopt the Wide Residual Network with depth 28 and width 2 (WRN-28-2) following the standard specification in the paper (Zagoruyko and Komodakis, 2016).We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0003. Additionally, we train models for a maximum of 400 epochs with a scheduler that scales the learning rate by 0.2 once the model finishes 320 epochs. To simulate the LLP setting, we split the training data by two bag generation algorithms described in Section 4.1 and 4.2. Once completing the bag generation, we then compute the proportion labels by averaging the class labels over each bag. To avoid overfitting, we follow the common practice of data augmentation (He et al., 2016; Lin et al., 2013) padding an image by 4 pixels on each side, taking a random 32x32 crop and randomly flipping the image horizontally with the probability of 0.5 for all benchmarks.

**Hyperparameters.** We compare our method, LLP-VAT, to ROT (Dulac-Arnold et al., 2019) and the vanilla approach, which simply minimizes the proportion loss. For ROT, we conduct experiments with a hyperparameter of $\alpha \in \{0.1, 0.4, 0.7, 0.9\}$ to compute the ROT loss. Following Oliver et al. (2018), we adopt the VAT approach to generate perturbed examples with a perturbation weight $\epsilon$ of 1 and 6 for SVHN and CIFAR10 (or CIFAR100) respectively. We measure the consistency loss with the KL divergence and a consistency weight of $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$.

**Model selection.** For a fair comparison, we randomly sample 90% of bags for training and reserve the rest for validation. In the LLP setting, since there are no individual labels available in the validation set, we select hyperparameters based on the *hard $L^1$ error* which is computed with only proportion labels. To be more specific, the hard $L^1$ error for a bag $B = (\mathbb{X}, \boldsymbol{p})$ is defined by

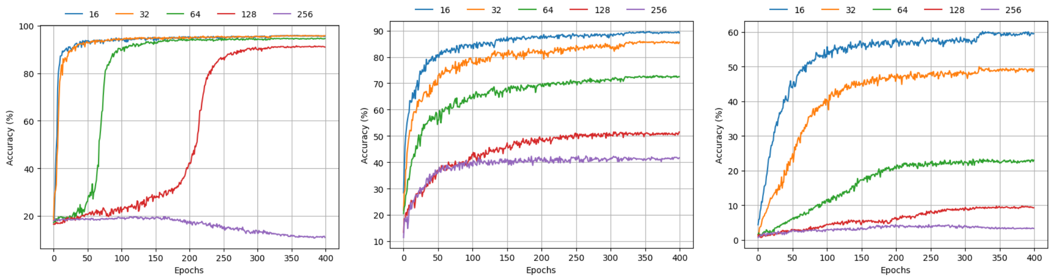
$$Err = ||\boldsymbol{p} - \hat{\boldsymbol{p}}||_1, \quad \hat{\boldsymbol{p}} = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} \boldsymbol{e}^{(i^*)},$$

where $i^* = \arg\max_i f_\theta(\boldsymbol{x})_i$ and $\boldsymbol{e}^{(i^*)}$ is the one-hot encoding of the prediction. Lastly, we report the test instance accuracy averaged over the last 10 epochs.

## Appendix B. Convergence analysis of LLP-VAT

To analyze the convergence performance of LLP-VAT, we plot the instance accuracy on the test set over training epochs. Figure 4 and 5 show the accuracy curve on the test set with the uniform bag generation and the K-means bag generation respectively. As shown in Figure 4 and 5, the experimental results demonstrate the stability of our LLP-VAT. When the training epoch gradually increases, the test instance accuracy goes up quickly and converges in the end.

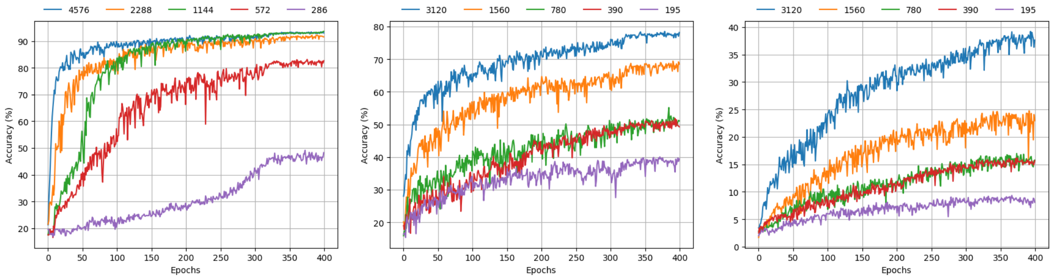
## B.1. Uniform Bag Generation



($a$) SVHN ($b$) CIFAR10 ($c$) CIFAR100

Figure 4: Evolution of the test accuracy on benchmarks with the uniform bag generation of varying bag sizes.

## B.2. K-means Bag Generation



($a$) SVHN ($b$) CIFAR10 ($c$) CIFAR100

Figure 5: Evolution of the test accuracy on benchmarks with the K-means bag generation of varying number of clusters.