Learning from Label Proportions by Learning with Label Noise

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning from label proportions (LLP) is a weakly supervised classification problem where data points are grouped into bags, and the label proportions within each bag are observed instead of the instance-level labels. The task is to learn a classifier to predict the labels of future individual instances. Prior work on LLP for multi-class data has yet to develop a theoretically grounded algorithm. In this work, we propose an approach to LLP based on a reduction to learning with label noise, using the forward correction (FC) loss of Patrini et al. [30]. We establish an excess risk bound and generalization error analysis for our approach, while also extending the theory of the FC loss which may be of independent interest. Our approach demonstrates improved empirical performance in deep learning scenarios across multiple datasets and architectures, compared to the leading methods.

1 Introduction

In the weakly supervised problem of *learning from label proportions* (LLP), the learner is presented with bags of instances, where each bag is annotated with the proportions of the different classes in the bag. The learner's objective is to produce a classifier that accurately assigns labels to individual instances in the future. LLP arises in various applications including high energy physics [7], election prediction [45], computer vision [4, 20], medical image analysis [2], remote sensing [8], activity recognition [32], and reproductive medicine [12].

To date, most methods for LLP have addressed the setting of binary classification [50, 36, 39, 34, 41, 44, 24, 37, 38], although multiclass methods have also recently been investigated [9, 22, 46]. The dominant approach to LLP in the literature is "label proportion matching": train a classifier to accurately reproduce the observed label proportions on the training data, perhaps with additional regularization. In the multiclass setting, the Kullback-Leibler (KL) divergence between the observed and predicted label proportions is adopted by the leading approaches to assess proportion matching. Unfortunately, while matching the observed label proportions is intuitive and can work well in some settings, it has little theoretical basis [50, 38], especially in the multiclass setting, and there are natural settings where it fails [50, 39].

Recently, Scott and Zhang [39] demonstrated a principled approach to LLP with performance guarantees based on a reduction to learning with label noise (LLN) in the binary setting. Their basic strategy was to pair bags, and view each pair of bags as an LLN problem, where the observed label proportions are related to the "label flipping" or "noise transition" probabilities. Using an existing technique for LLN based on loss correction, which allows the learner to train directly on the noisy data, they formulated an overall objective based on a (weighted) sum of objectives for each pair of bags. They established generalization error analysis and consistency for the method, and also showed that in the context of kernel methods, their approach outperformed the leading kernel methods.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The objective of the present paper is to developed a theoretically grounded and practical approach to multiclass LLP, drawing inspiration from Scott and Zhang [39]. The primary challenge stems from the fact that Scott and Zhang [39] employed the so-called "backward correction" loss, which solves LLN by scaling the *output* of a loss function of interest according to the noise transition probabilities [28, 30, 35]. While this loss correction was demonstrated to work well for kernel methods in a binary setting, Patrini et al. [30] showed that it is poorly suited for multiclass deep learning settings . To remedy this, they proposed the "forward correction" loss, which scales the *inputs* to a loss function of interest according to the noise transition probabilities. They established Fisher consistency of the technique, and demonstrated its superiority empirically (see also [53]).

The present work is thus inspired by Scott and Zhang [39] but uses the forward correction (FC) loss in a multiclass setting. This requires a number of technical modifications to the arguments of Scott and Zhang [39]. Most notably, it now becomes necessary to demonstrate that the FC loss is *calibrated* with respect to the 0-1 loss, a critical property needed for showing consistency. Such analysis is inherently not needed when using the backward correction, due to the way in which this correction yields an unbiased estimate of the uncorrupted risk. Furthermore, Scott and Zhang [39] adopt a probabilistic setting known as a "mutual contamination model", whereas we adopt the more conventional setting based on a noise transition matrix. Finally, the multiclass setting involves new estimation challenges not present in the binary case. These factors mean that our work is not a straightforward extension of Scott and Zhang [39]. Indeed, the authors of a recent report acknowledge that it is "difficult to extend [the method of Scott and Zhang [39]] to multiclass classification" [16].

Additional related work: Much work on LLP has focused on learning specific types of models, including support vector machines [36, 50, 47, 33, 5, 19, 40], probabilistic models [18, 13, 45, 32, 12], random forests [41], neural networks [21, 1, 9, 22, 46], and clustering-based models, [3, 44]. Many of these works develop learning criteria that are specific to the model being learned.

On the theoretical front, Quadrianto et al. [34] and Patrini et al. [30] initiated the learning theoretic study of LLP, introducing Rademacher style bounds for linear methods, but they do not address consistency w.r.t. a classification performance measure. Yu et al. [51] provides support for label proportion matching but only under the assumption that the bags are very pure. Saket [37] studies learnability of linear threshold functions. Recently Saket et al. [38] introduced a condition under which label proportion matching does provably well w.r.t. a squared error loss in the binary setting, and developed an associated algorithm. This method does not scale easily to large datasets, and further requires knowledge of how bags are grouped according to different bag-generating distributions.

A handful of recent papers have studied multiclass LLP in deep learning scenarios. Dulac-Arnold et al. [9] study the KL loss for label proportion matching, and a variant based on optimal transport. Liu et al. [22, 23] examine an approach based on generative adversarial models. Tsai and Lin [46] study the use of a regularizer derived from semi-supervised learning. One challenge common to these approaches is that their implementations employ mini-batches of bags, which becomes computationally prohibitive for deep architectures when the batch size is still very small, e.g., 2 or 3. In contrast, our approach avoids this issue. Finally, a recent technical report presents a risk analysis for multiclass LLP under the assumption of fixed bag size, which we do not require [16]. Their method is not tractable for large bag sizes in which case they approximate their objective "using the bag-level loss proposed in the existing research."

Contributions and Outline: Our contributions and the paper structure are summarized as follows. In Section 2, we review the FC loss as a solution to LLN. In Section 3, we extend the theory of the FC loss for LLN. In particular, we show that the FC loss is "uniformly calibrated" with respect to the 0-1 loss using the framework of Steinwart [43], establish an excess risk bound, and determine an explicit lower bound on the calibration function in terms of the noise transition matrix. In Section 4, we extend the results of Section 3 to the setting with multiple noise transition matrices, which form the basis of our approach to LLP. In particular, we establish an excess risk bound and generalization error analysis for learning with multiple noise transition matrices, which in turn enables proofs of consistency. In Section 5, we state the probabilistic model for reducing LLP to LLN with multiple different noise transition matrices and present the LLPFC algorithms. Experiments with deep neural networks are presented in Section 6, where we observe that our approach outperforms competing methods by a substantial margin. Proofs appear in the supplemental material.

2 Learning with Label Noise and the Forward Correction Loss

This section sets notation and introduces the FC loss as a solution to learning with label noise. Let $\mathcal{X}$ be the feature space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ be the label space, $C \in \mathbb{N}$. We define the C-simplex as $\Delta^C = \{p \in \mathbb{R}^C : p_i \geq 0, \forall i = 1, 2, \ldots, C, \sum_{i=1}^C p_i = 1\}$ and denote its interior by $\mathring{\Delta}^C$. Let P be a probability measure on the space $\mathcal{X} \times \mathcal{Y}$.

Viewing P as the "clean" probability measure, a noisy probability measure with label-dependent label noise can be constructed from P in terms of a $C \times C$ column-stochastic matrix T, referred to as the *noise transition matrix*. Formally, we define a measure $\bar{P}_T$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ by requiring $\forall$ events $\mathcal{A} \subset \mathcal{X}, \bar{P}_T(\mathcal{A} \times \{i\} \times \{j\}) = P(\mathcal{A} \times \{i\})t_{j,i}$ where $t_{j,i}$ is the element at the j-th row and i-th column of T. Let $(X, Y, \tilde{Y})$ have joint distribution $\bar{P}_T$ where X is the feature vector, Y is the "clean" label, and $\tilde{Y}$ is the "noisy" label. Thus the element of T at row i and column j is $t_{i,j} = \bar{P}_T(\tilde{Y} = i | Y = j)$. In addition, P is the marginal distribution of (X, Y). Define P_T to be the marginal distribution of $(X, \tilde{Y})$. Let $\mathcal{F}$ be the collection of all measurable functions from $\mathcal{X}$ to Δ^C.

The existence of a regular conditional distribution is guaranteed by the Disintegration Theorem ($e.g.$ Theorem 6.4 in Kallenberg [14]) under suitable properties ($e.g.$ when $\mathcal{X}$ is a Radon space). While the existence of regular conditional probability is beyond the scope of this paper, we assume fixed regular conditional distributions for Y and $\tilde{Y}$ given X exist, denoted by $P(\cdot \mid \cdot) : \mathcal{Y} \times \mathcal{X} \to [0, 1]$ and $P_T(\cdot \mid \cdot) : \mathcal{Y} \times \mathcal{X} \to [0, 1]$, respectively. Given $x \in \mathcal{X}$, we define the probability vectors $\eta(x) = [P(1 \mid x), \ldots, P(C \mid x)]^{tr}$ and $\eta_T(x) = [P_T(1 \mid x), \ldots, P_T(C \mid x)]^{tr}$ where we use tr to denote transposition. It directly follows that $\eta_T(x) = T\eta(x)$.

We use $\mathbb{R}_+$ to denote the positive real numbers. The goal of LLN is to learn a classifier that optimizes a performance measure defined $w.r.t.$ P, given access to corrupted training data $(X_i, \tilde{Y}_i) \overset{i.i.d.}{\sim} P_T$. In this work we assume T is known or can be estimated, as is the case when we apply LLN techniques to LLP (see Section 5). A more formal formulation of LLP is given in Section 5.

When attempting to minimize the risk associated to the 0-1 loss and the clean distribution P, it is common to employ a smooth or convex surrogate loss. For LLN problems, the idea of a *loss correction* is to modify the surrogate loss so that when optimized using the *noisy* data, it still achieves the desired goal. Below, we introduce the forward correction loss, before which we need to define inner risk and proper loss. For this purpose we focus on loss functions of the form $L : \Delta^C \times \mathcal{Y} \to \mathbb{R}$.

Definition 1. *Let $L : \Delta^C \times \mathcal{Y} \to \mathbb{R}$ be a loss function. The **inner L-risk** at x with probability measure P is $\mathcal{C}_{L,P,x} : \Delta^C \to \mathbb{R}$, $\mathcal{C}_{L,P,x}(q) := \mathbb{E}_{Y \sim P(\cdot|x)} L(q, Y)$. The **minimal inner L-risk** at x with a probability measure P is $\mathcal{C}^*_{L,P,x} := \inf_{q \in \Delta^C} \mathcal{C}_{L,P,x}(q)$.*

Definition 2. *$\ell : \Delta^C \times \mathcal{Y} \to \mathbb{R}$ is a **proper loss** if $\forall$ probability measures P on $\mathcal{X} \times \mathcal{Y}$, $\forall x \in \mathcal{X}, \mathcal{C}^*_{\ell,P,x} = \mathcal{C}_{\ell,P,x}(\eta(x))$, and a proper loss is called **strictly proper** if the minimizer of $\mathcal{C}_{\ell,P,x}$ is unique for all $x \in \mathcal{X}$.*

Commonly used proper losses include the *log loss* $\ell^{log}(q, c) = -\log q_c$, the *square loss* $\ell^{sq}(q, c) = \sum_{c'=1}^C (\mathbb{1}_{c=c'} - q_{c'})^2$, and the 0-1 *loss* $\ell^{01}(q, c) = \mathbb{1}_{c \neq \min\{\arg\max_j q_j\}}$, among which only the log loss and the square loss are strictly proper [49]. Here $\mathbb{1}$ denotes the indicator function. Note that it is common to compose proper losses with inverted link functions, leading to familiar losses like the cross-entropy loss. Such losses are discussed further in Section 4.

We are now ready to introduce the forward correction loss.

Definition 3. *Let ℓ be a strictly proper loss and let T be a noise transition matrix. Define the **forward correction loss** of ℓ as $\ell_T : \Delta^C \times \mathcal{Y} \to \mathbb{R}$, $\ell_T(q, c) := \ell(Tq, c)$.*

It follows from the definition that, if T is invertible, then the inner ℓ_T-risk under the distribution P_T has a unique minimizer $\eta(x)$. Next we introduce the definition of L-risk and L-Bayes risk associated with a loss function L.

Definition 4. *Let $L : \Delta^C \times \mathcal{Y} \to \mathbb{R}$ and P be a probability measure. Define the **L-risk** of f with distribution P to be $\mathcal{R}_{L,P} : \mathcal{F} \to \mathbb{R}$, $\mathcal{R}_{L,P}(f) := \mathbb{E}_P [L(f(X), Y)]$ and the **L-Bayes risk** to be $\mathcal{R}^*_{L,P} := \inf_{f \in \mathcal{F}} \mathcal{R}_{L,P}(f)$.*

We call $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$ the *excess L-risk* of f under distribution P. Given a proper loss ℓ, Theorem 2 of Patrini et al. [30] establishes Fisher consistency of the FC loss, meaning the minimizer of ℓ-risk under the clean distribution P is the same as the minimizer of ℓ_T-risk under noisy distribution P_T: $\arg\min_{f\in\mathcal{F}} \mathcal{R}_{L,P}(f) = \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\ell_T,P_T}(f)$. Next, we present a stronger result relating the excess ℓ_T-risk under the noisy distribution P_T to the excess 0-1 risk under the clean distribution P.

3 Calibration Analysis for the Forward Correction Loss

Our objective in this section is to show that when L is the 0-1 loss and ℓ is a continuous strictly proper surrogate loss, there exists a strictly increasing, invertible function θ with $\theta(0) = 0$ such that $\forall f \in \mathcal{F}$ and $\forall$ distributions P, $\theta\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big) \leq \mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T}$. Given such a bound, it follows that consistency $w.r.t$ the surrogate risk implies consistency $w.r.t.$ the target risk. The results in this section are standalone results for the FC loss that may be of independent interest, and will be extended in the next section in relation to LLP.

The following theorem guarantees the existence of such function θ, given that T is invertible.

Theorem 5. *Let ℓ be a continuous strictly proper loss and T be an invertible column-stochastic matrix. Let L be the 0-1 loss. Assume $\mathcal{R}^*_{\ell_T,P_T} < \infty$. Then $\exists\theta : [0,1] \to [0,\infty]$ that is strictly increasing and continuous, satisfying $\theta(0) = 0$, such that $\forall f \in \mathcal{F}, \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \leq \theta^{-1}\Big(\mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T}\Big)$.*

The function θ depends on ℓ and T. The following proposition provides a convex lower bound on θ for the commonly used log loss $\ell^{log}(q,c) = -\log q_c$. Let $M \in \mathbb{R}^{C\times C}$ be a matrix and let $\|\cdot\|$ be a norm on $\mathbb{R}^C$. The *subordinate matrix norm* induced by $\|\cdot\|$ is $\|M\| := \sup_{x\in\mathbb{R}^C : x\neq 0} \frac{\|Mx\|}{\|x\|}$. When $\|\cdot\|$ is the 1-norm on $\mathbb{R}^C$, the induced norm is denoted $\|M\|_1$, referred to as the matrix 1-norm, and can be computed as $\|M\|_1 = \max_{1\leq j\leq C} \sum_{i=1}^{C} |M(i,j)|$ [10].

Proposition 6. *Let $T \in \mathbb{R}^{C\times C}$ be an invertible, column-stochastic matrix. Define $\underline{\theta}_T : [0,\infty] \to [0,\infty]$ by $\underline{\theta}_T(\epsilon) = \frac{1}{2}\frac{\epsilon^2}{\|T^{-1}\|_1^2}$. If L is the 0/1 loss, ℓ is the log loss, then for all $f \in \mathcal{F}$ and distributions P, $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \leq \underline{\theta}_T^{-1}\Big(\mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T}\Big) = \sqrt{2}\|T^{-1}\|_1\sqrt{\mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T}}$*

The factor $\|T^{-1}\|_1$ may be viewed as a constant that captures the overall amount of label noise. The more noise, the larger the constant. For example, let I and N be the identity and the all $1/C$'s matrices, respectively. Let $\alpha \in [0,1]$ and $T = (1-\alpha)I + \alpha N$. Thus, $\alpha = 0$ represents the noise-free case and $\alpha = 1$ the noise-only case. It is easy to verify that $T^{-1} = (1-\alpha)^{-1}(I - \alpha N)$ and $\|T^{-1}\|_1 = (1-\alpha)^{-1}(1 + (1-2/C)\alpha)$.

4 Learning with Multiple Noise Transition Matrices

Our algorithm for LLP, formally stated in subsection 5.4, reduces the problem of LLP to LLN by partitioning bags into groups and modeling each group as an LLN problem. Since each group has its own noise transition matrix, this leads to a new problem that we refer to as learning with multiple noise transition matrices (LMNTM). In this section, we show how to extend the calibration analysis of section 3 to this setting. In addition, we offer a generalization error bound that justifies an empirical risk minimization learning procedure based on a weighted sum of FC losses.

4.1 Learning with Multiple Noise Transition Matrices

We first define the LMNTM problem formally. For all $n \in \mathbb{N}$, denote $\mathbb{N}_n = \{1, 2, \ldots, n\}$. Consider a clean distribution P on $\mathcal{X} \times \mathcal{Y}$ and a finite sequence of noise transition matrices $T_1, T_2, \ldots, T_N$. For each i we denote the noisy prior as the $\alpha_i \in \mathring{\Delta}^C$ where, $\forall c \in \mathcal{Y}$, $\alpha_i(c) = P_{T_i}(\tilde{Y} = c)$. We assume the α_i's are known for theoretical analysis. In practice, α_i is estimable as discussed below. In LMNTM, we observe data points $S = \big\{X_{i,c,j} : \ i \in \mathbb{N}_N, \ c \in \mathcal{Y}, \ j \in \mathbb{N}_{n_{i,c}}\big\}$ where $X_{i,c,j} \overset{i.i.d.}{\sim} P_{T_i}(\cdot \mid c)$, and $n_{i,c} \in \mathbb{N}$ is the number of data points drawn from the class conditional

distribution $P_{T_i}(\cdot \mid c)$. Assume all $X_{i,c,j}$'s are mutually independent. We make additional remarks on this setting in subsection D.1 in the appendix.

4.2 An Risk for LMNTM

The following result extends Theorem 5 to LMNTM. It establishes that the risk $\widetilde{R}_{\ell,P,\mathcal{T}}$, which can be estimated from LMNTM training data, is a valid surrogate risk. This type of result is not needed for the backward correction approach of Scott and Zhang [39].

Theorem 7. *Let L be the 0-1 loss and $N \in \mathbb{N}$. Consider a sequence of invertible column-stochastic matrices $\mathcal{T} = \{T_i\}_{i=1}^N$ and a continuous strictly proper loss function ℓ. Let $w = (w_i)_{i=1}^N \in \Delta^N$. Define $\widetilde{R}_{\ell,P,\mathcal{T}} : \mathcal{F} \to \mathbb{R}$ by $\widetilde{R}_{\ell,P,\mathcal{T}}(f) := \sum_{i=1}^N w_i \mathcal{R}_{\ell_{T_i},P_{T_i}}(f)$ and $\widetilde{R}^*_{\ell,P,\mathcal{T}} = \inf_{f\in\mathcal{F}} \widetilde{R}_{\ell,P,\mathcal{T}}(f)$. Assume $\forall i \in \{1,2,\ldots,N\}, \mathcal{R}^*_{\ell_{T_i},P_{T_i}} < \infty$. Then $\exists$ a strictly increasing continuous function θ : $[0,1] \to [0,\infty]$ with $\theta(0) = 0$ s.t. for all P, $\forall f \in \mathcal{F}$, $\theta\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big) \le \widetilde{\mathcal{R}}_{\ell,P,\mathcal{T}}(f) - \widetilde{\mathcal{R}}^*_{\ell,P,\mathcal{T}}$.*

The weights w_i allow the user flexibility, for example, to place different weights on noisier or larger subsets of data. Unlike Scott and Zhang [39], however, because the weights appear in both our excess risk bound and generalization error bound, it is not straightforward to optimize them a priori.

4.3 Generalization Error Bound

The aggregate risk $\widetilde{R}_{\ell,P,\mathcal{T}}$ is desirable because it can naturally be estimated from the given data. We propose the empirical risk

$$\hat{\mathcal{R}}_{w,S}(f) = \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \ell_{T_i}(f(X_{i,c,j}), c). \tag{1}$$

It should be noted that $\hat{\mathcal{R}}_{w,S}(f)$ is an unbiased estimate of $\tilde{\mathcal{R}}_{\ell,P,\mathcal{T}}(f)$. Here we establish a generalization error bound for this estimate, that is, we show that the empirical risk concentrates around the risk uniformly. Our results build on Rademacher complexity analysis.

To state the bound, we must first introduce the notion of a *proper composite loss* [49]. This stems from the fact that in practice, a function f outputting values in Δ^C is typically obtained by composing a $\mathbb{R}^C$-valued function (such as a neural network with C output layer nodes), with another function $\mathbb{R}^C \to \Delta^C$ such as the softmax function. Thus, let $\psi : \mathcal{U} \subset \Delta^C \to \mathcal{V}$ be an invertible function where $\mathcal{V}$ is a subset of a normed space, referred to as an *invertible link function*. Consider $\mathcal{G} \subset \psi \circ \mathcal{F} := \{\psi \circ f : f \in \mathcal{F}\}$, and observe that $\forall g \in \mathcal{G}, \psi^{-1} \circ g \in \mathcal{F}$. In practice, ψ is fixed and we seek to learn $g \in \mathcal{G}$ that leads to an $f \in \mathcal{F}$ with a risk close to the Bayes risk. An example of ψ^{-1} is the softmax function so that $\psi : \mathcal{U} \to \mathcal{V}, \psi_i(p) = \log p_i - \frac{1}{C}\sum_{k=1}^C \log p_k, (\psi^{-1})_i(s) = \frac{e^{s_i}}{\sum_{k=1}^C e^{s_k}}$ where $\mathcal{U}$ is the interior of Δ^C and $\mathcal{V} = \left\{s \in \mathbb{R}^C : \sum_{i=1}^C s_i = 0\right\}$. This motivates the following definition.

Definition 8 (Proper Composite Loss)**.** *Given an invertible link function $\psi : \mathcal{U} \subset \Delta^C \to \mathcal{V}$, we define the **proper composite loss** λ_ℓ of a proper loss $\ell : \Delta^C \times \mathcal{Y} \to \mathbb{R}$ to be $\lambda_\ell : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$, $\lambda_\ell(v, c) = \ell\big(\psi^{-1}(v), c\big)$.*

For example, when ℓ is the log loss and ψ^{-1} is the softmax function, λ_ℓ is the cross-entropy (or multinomial logistic) loss. With this notation, we are now able to state our generalization error bound for LMNTM. We study two popular choices of function classes, the reproducing kernel Hilbert space (RKHS) and the multilayer perceptron (MLP).

Definition 9. *Let k be a symmetric positive definite (SPD) kernel, and let $\mathcal{H}$ be the associated reproducing kernel Hilbert space (RKHS). Assume k is bounded by K, meaning $\forall x$, $\|k(\cdot, x)\|_{\mathcal{H}} \le K$. Let $\mathcal{G}^k_{K,R}$ denote the ball of radius R in $\mathcal{H}$. Define $\mathcal{G}_1 = \mathcal{G}^k_{K,R} \times \mathcal{G}^k_{K,R} \times \cdots \times \mathcal{G}^k_{K,R}$ (C copies).*

We follow Zhang et al. [54] and define real-valued MLPs inductively:

Definition 10. *Define $\mathcal{N}_1 = \left\{x \to \langle x, v\rangle : v \in \mathbb{R}^d, \|v\|_2 \le \beta\right\}$, and for $m > 2$, inductively define $\mathcal{N}_m = \left\{x \to \sum_{j=1}^d v_j \mu(f_j(x)) : v \in \mathbb{R}^d, \|v\|_1 \le \beta, f_j \in \mathcal{N}_{m-1}\right\}$, where $\beta \in \mathbb{R}_+$ and μ is a 1-Lipschitz activation function. Define an MLP which outputs a vector in $\mathbb{R}^C$ by $\mathcal{G}_2 = \mathcal{N}_m \times \mathcal{N}_m \times \cdots \times \mathcal{N}_m$ (C copies). We additionally assume that the choice of μ satisfies $\forall m \in \mathbb{N}, 0 \in \mu \circ \mathcal{N}_m$.*

Theorem 11. *Let $T_1, T_2, \ldots, T_N$ be a finite sequence of invertible column-stochastic matrices. Let ℓ be a proper loss such that $\forall i, c$ the function $\lambda_{\ell_{T_i}}(\cdot, c)$ is Lipschitz continuous w.r.t. the 2-norm. Let S be the set of data points as defined in Section 4.1 Assume $\sup_{x\in\mathcal{X}, g\in\mathcal{G}_q}\|g(x)\|_2 \leq A_q$ for some constant A_q, $\forall q \in \{1,2\}$. Let $\hat{\mathcal{R}}_{w,S}(g) := \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \lambda_{\ell_{T_i}}(g(X_{i,c,j}), c)$, $\widetilde{\mathcal{R}}(g) := \widetilde{R}_{\ell,P,\mathcal{T}}(\psi^{-1}\circ g) = \mathbb{E}_S\left[\hat{\mathcal{R}}_{w,S}(g)\right]$. Then for each $q \in \{1,2\}$, $\forall \delta \in [0,1]$, with probability at least $1-\delta$,*

$$\sup_{g\in\mathcal{G}_q}\left|\hat{\mathcal{R}}_{w,S}(g) - \widetilde{\mathcal{R}}(g)\right| \leq (\max_i(\left|\lambda_{\ell_{T_i}}\right|A_q + \left|\lambda_{\ell_{T_i}}\right|_0)\sqrt{2\log\frac{2}{\delta}} + CB_q \max_i|\lambda_{\ell_{T_i}}|)\sqrt{\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2}{n_{i,c}}}.$$

where B_q is a constant depending on $\mathcal{G}_q$, $\left|\lambda_{\ell_{T_i}}\right|_0 = \max_c\left|\lambda_{\ell_{T_i}}(0,c)\right|$, and $\left|\lambda_{\ell_{T_i}}\right|$ is the smallest real number such that it is a Lipschitz constant of $\lambda_{\ell_{T_i}}(\cdot, c)$, $\forall i, c$.

It should be noted that Theorem 11 is a special case of of Lemma 26 which extends the notion of Rademacher complexity to the LMNTM setting and applies to arbitrary function classes. Lemma 26 is presented in the appendix.

Let HM_i denote the harmonic mean of $n_{i,1}, \ldots, n_{i,C}$, *i.e.*, $HM_i = \frac{C}{\sum_{c=1}^{C}\frac{1}{n_{i,c}}}$. The term $\sqrt{\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2}{n_{i,c}}}$ could be written as $\sqrt{C\sum_{i=1}^{N}\frac{w_i^2}{HM_i}}$ and is optimized by $w_i = \frac{HM_i}{\sum_{m=1}^{N} HM_m}$, leading to $\sqrt{\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2}{n_{i,c}}} = \sqrt{\frac{C}{\sum_{i=1}^{N} HM_i}}$. The term $\sqrt{\frac{C}{\sum_{i=1}^{N} HM_i}}$ vanishes (needed to establish consistency) when N goes to infinity, or when $\exists i\ s.t.\ \forall c$, $n_{i,c}$ goes to infinity. For the special case where all bags have the same size n and all weights w_i are $1/N$, $\sqrt{\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2}{n_{i,c}}} = \sqrt{\frac{C}{Nn}}$. Thus, consistency is possible even if bag size remains bounded.

Assuming ℓ is the log loss and ψ^{-1} is the softmax function, we next study the constants $\left|\lambda_{\ell_{T_i}}\right|$ and $\left|\lambda_{\ell_{T_i}}\right|_0$.

Proposition 12. *Let ℓ be the log loss, ψ^{-1} be the softmax function, and T be a column-stochastic matrix. Then $|\lambda_{\ell_T}| \leq \sqrt{2}$.*

The constant $|\lambda_{\ell_T}|_0 = \max_c|\lambda_{\ell_T}(0,c)| = \max_c -\log\left(\frac{1}{C}\sum_{j=1}^{C} t_{c,j}\right)$. The invertibility of T guarantees $\sum_{j=1}^{C} t_{c,j}$ is positive and hence the finiteness of $|\lambda_{\ell_T}|_0$. However, if we have a "bad" T, $\sum_{j=1}^{C} t_{c,j}$ could be arbitrarily close to 0 leading to a large $|\lambda_{\ell_T}|_0$.

Following Theorem 11, if the function class $\mathcal{G}$ has a universal approximation property, such as an RKHS associated to a universal kernel, or an MLP with increasing number of nodes, consistency for LMNTM via (regularized) minimization of $\hat{\mathcal{R}}_{w,S}(g)$ can be shown by leveraging standard techniques, provided $N \to \infty$ (bag size may remain bounded). Then the excess risk bound in Theorem 7 would automatically imply consistency with respect to 0-1 loss.

5 The LLPFC algorithms

In this section, we define a probabilistic model for LLP, show how LLP reduces to LMNTM, and introduce algorithms that we refer to as the LLPFC algorithms.

5.1 Probabilistic Model for LLP

Given a measure P on the space $\mathcal{X}\times\mathcal{Y}$, let $\{P_c : c \in \mathcal{Y}\}$ denote the class-conditional distributions of $\mathcal{X}$, *i.e.*, $\forall$ events $\mathcal{A} \subset \mathcal{X}$, $P_c(\mathcal{A}) = P(\mathcal{A} \mid Y = c)$. Let $\sigma(c) = P(Y = c), \forall c \in \mathcal{Y}$ and call $\sigma = (\sigma(1), \ldots, \sigma(C))$ the clean prior. Assume $\forall c \in \mathcal{Y}, \sigma(c) \neq 0$. Given $z = (z(1), \ldots, z(C)) \in \Delta^C$, let P_z be the probability measure on $\mathcal{X}\times\mathcal{Y}$ s.t. $\forall$ events $\mathcal{A} \subset \mathcal{X}$, $\forall i \in \mathcal{Y}$, $P_z(\mathcal{A}\times\{i\}) = z(i)P_i(\mathcal{A})$. Thus P_z has the same class-conditional distributions as P but a variable prior z.

We first define a model for a single bag. Given $z \in \Delta^C$, we say that bag b is *governed* by $z \in \Delta^C$ if b is a collection of feature vectors $\left\{X_j : j \in \mathbb{N}_{|b|}\right\}$ annotated by label proportion

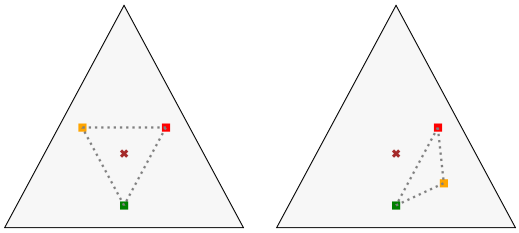

Figure 1: The gray triangle represents the probability simplex Δ^3. The squares represent γ_1, γ_2, and γ_3. The cross is σ. The ternary graph on the left visualizes an example where Assumption 13 holds. The one on the right visualizes an example where Assumption 13 fails.

$\hat{z} = (\hat{z}(1), \hat{z}(2), \ldots, \hat{z}(C))$, where $|b|$ denotes the cardinality of the bag, each X_j is paired with an unobserved label Y_j *s.t.* $(X_j, Y_j) \overset{iid}{\sim} P_z$, and $\hat{z}(c) = \frac{1}{|b|}\sum_{j=1}^{|b|} \mathbb{1}_{Y_j=c}$. Note $\mathbb{E}_{P_z}[\hat{z}] = z$ and $P_z(Y_j = c) = z(c)$. We think of z as the true label proportion and $\hat{z}$ as the empirical label proportion. Using this model for individual bags, we now formally state a model for LLP. Given bags $\{b_k\}$, let each b_k be governed by γ_k. Each b_k is a collection of feature vectors $\left\{X_j^k : j \in \mathbb{N}_{|b_k|}\right\}$ where $(X_j^k, Y_j^k) \overset{i.i.d.}{\sim} P_{\gamma_k}$ and Y_j^k is unknown. Further assume the X_j^k's are independent for all k and j. In practice, γ_k is unknown and we observe $\hat{\gamma}_k$ with $\hat{\gamma}_k(c) = \frac{1}{|b_k|}\sum_{j=1}^{|b_k|} \mathbb{1}_{Y_j^k=c}$ instead. The goal is learn an f that minimizes the risk $\mathcal{R}_{L,P} = \mathbb{E}_{(X,Y)\sim P}[L(f(X), Y)]$ where L is the 0-1 loss, given access to the training data $\{(b_k, \hat{\gamma}_k)\}$.

5.2 The Case of C Bags: Reduction to LLN

To explain our reduction of LLP to LLN, we first consider the case of exactly C bags $b_1, b_2, \ldots, b_C$, governed by respective (unobserved) $\gamma_1, \ldots, \gamma_C \in \Delta^C$, and annotated with label proportions $\hat{\gamma}_1, \ldots \hat{\gamma}_C$. Define $\Gamma \in \mathbb{R}^{C\times C}$ by $\Gamma(i,j) = \gamma_i(j)$, and let Γ^{tr} denote the transpose of Γ. Recall that σ is the class prior associated to P. To model LLP with C bags as an LLN problem, we make the following assumption on Γ and σ:

Assumption 13. $\exists$ *unique* $\alpha \in \mathring{\Delta}^C$ *s.t.* $\Gamma^{tr}\alpha = \sigma$.

We write $\alpha = (\alpha(1), \ldots, \alpha(C))$. Assumption 13 is equivalent to: $\{\gamma_1, \ldots, \gamma_C\}$ is a linearly independent set and σ is in the interior of the convex hull of $\{\gamma_1, \ldots, \gamma_C\}$. Ternary plots in Figure 5.2 visualize examples where assumption 13 holds and fails when $C = 3$. Intuitively, assumption 13 is more likely to hold when $\{\gamma_i : i \in \mathbb{N}_C\}$ are more "spread out" in Δ^C, in which case it is more likely for σ to reside in the convex hull of $\{\gamma_i : i \in \mathbb{N}_C\}$.

To reduce LLP with C bags to LLN, we simply propose to assign the "noisy label" $\tilde{Y} = i$ to all elements of bag b_i and to construct a noise transition matrix T with $T(i,j) = \frac{\gamma_i(j)\alpha(i)}{\sigma(j)}$. Assumption 13 ensures T is indeed a column-stochastic matrix. Thus, the probability measure $\bar{P}_T$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ satisfies $\alpha(i) = \bar{P}_T(\tilde{Y} = i)$ and $P_{\gamma_i}(\cdot) = \bar{P}_T(\cdot \mid \tilde{Y} = i)$, which further implies $\gamma_i(c) = \bar{P}_T(Y = c \mid \tilde{Y} = i)$. We confirm these facts in Section E in the appendix. Such construction transforms LLP with C bags into LLN with an estimable noise transition matrix T. Each element of a bag can then be viewed as a triplet $(X, Y, \tilde{Y})$, with Y unobserved, such that (X, Y) is drawn from $P_{\gamma_{\tilde{Y}}}$.

After assigning the noisy labels, we have a dataset $\bigcup_{c=1}^{C}\left\{\left(X_j^c, c\right) : j \in \mathbb{N}_{|b_c|}\right\}$ along with the noise transition matrix T. This allows us to leverage the forward correction loss ℓ_T to minimize the objective $\mathcal{R}_{\ell_T,P_T}(f) = \mathbb{E}_{P_T}\left[\ell_T(f(X), \tilde{Y})\right]$ which can be estimated by the empirical risk $\sum_{c=1}^{C} \frac{\alpha(c)}{|b_c|}\sum_{j=1}^{|b_c|} \ell_T\left(f(X_j^c), c\right)$.

5.3 The General Case: Reduction to LMNTM

More generally, consider LLP with NC bags, $N \in \mathbb{N}$. We propose to randomly partition the bags into N groups, each with C bags indexed from 1 to C. Let $k_{i,c}$ denote the index of the c-th bag in the

Algorithm 1 LLPFC-ideal

Input: $\{(b_k, \gamma_k)\}_{k=1}^{NC}$ and $w \in \Delta^N$ where $b_k = \{X_j^k : j \in \mathbb{N}_{|b_k|}\}$.
Randomly partition the bags into N groups $\{G_i\}_{i=1}^N$ *s.t.* $G_i = \{(b_{k_{i,c}}, \gamma_{k_{i,c}}) : c \in \mathcal{Y}\}$ and $\{k_{i,c} : i \in \mathbb{N}_N, c \in \mathcal{Y}\} = \mathbb{N}_{NC}$.
for $i = 1 : N$ **do**
  $\Gamma_i \leftarrow [\gamma_{k_{i,1}}, \gamma_{k_{i,2}}, \ldots, \gamma_{k_{i,C}}]^{tr}$
  $\alpha_i \leftarrow \Gamma_i^{-tr}\sigma$
  for $c_1 = 1 : C, c_2 = 1 : C$ **do**
    $T_i(c_1, c_2) \leftarrow \frac{\gamma_{k_{i,c_1}}(c_2)\alpha_i(c_1)}{\sigma(c_2)}$
  end for
end for
Train f with the empirical objective (1)
Return: the trained model f.

i-th group. Thus, $b_{k_{i,c}}$ is the c-th bag in the i-th group and it is governed by $\gamma_{k_{i,c}}$. For $i \in \mathbb{N}_N$, define the matrix $\Gamma_i \in \mathbb{R}^{C\times C}$ by $\Gamma_i(c_1, c_2) = \gamma_{k_{i,c_1}}(c_2), \forall c_1, c_2 \in \mathcal{Y}$. We make the following assumption on the Γ_i's and σ:

Assumption 14. *For each* $i \in \mathbb{N}_N$, $\exists$ *unique* $\alpha_i \in \mathring{\Delta}^C$ *s.t.* $\Gamma_i^{tr}\alpha_i = \sigma$.

Thus, every group i can be modeled as above as an LLN problem with noise transition matrix T_i where $T_i(c_1, c_2) = \frac{\gamma_{k_{i,c_1}}(c_2)\alpha_i(c_1)}{\sigma(c_2)}$. Data points in the bag assigned with noisy label c in the i-th group can be viewed as drawn $i.i.d.$ from the class conditional distribution $P_{T_i}(\cdot \mid c)$. This problem now maps directly to LMNTM as described in Section 4.1, and satisfies the associated performance guarantees. In the next subsection, we spell out the associated algorithm.

5.4 Algorithm

As above, assume we have NC bags where $N \in \mathbb{N}$. Let each bag b_k be governed by $\gamma_k \in \Delta^C$ and be annotated by label proportion $\hat{\gamma}_k$. Denote the size of b_k by $|b_k|$. We first present the LLPFC-ideal algorithm in an ideal setting where σ, the γ_k's and the α_i's are known precisely and Assumption 14 holds. We then present the real-world adaptations LLPFC-uniform and LLPFC-approx in practical settings where these assumptions might not hold.

The LLPFC-ideal algorithm is presented in Algorithm 1. We follow the idea in section 5.3 to partition the bags into N groups of C bags, and model each group as an LLN problem. In Algorithm 1, we assume γ_k and σ are known and Assumption 14 holds so that $\alpha_i \in \mathring{\Delta}^C$ for every i. The theoretical analysis in Section 4 is immediately applicable to the LLPFC-ideal algorithm. We partition the bags by uniformly randomly partitioning the set of indices $\mathbb{N}_{NC}$ into disjoint subsets $\{k_{i,c} : c \in \mathcal{Y}\}$, $i \in \mathbb{N}_N$, where $k_{i,c}$ denotes the index of the c-th bag in the i-th group. Hence, the c-th bag in the i-th group is denoted as $b_{k_{i,c}}$ and it is governed by $\gamma_{k_{i,c}}$. Following our convention, in Algorithms 1, 2, and 3, we denote the element at c_1-th row and c_2-th column in a matrix T by $T(c_1, c_2)$ and the c_1-th element in vector v by $v(c_1)$. We denote the inverse transpose of Γ_i by ${\Gamma_i}^{-tr}$.

In practice, when γ_k is unknown, we replace γ_k with $\hat{\gamma}_k$ as a plug-in method. Hence, we work with $\hat{\sigma} = \frac{\sum_{k=1}^{NC}|b_k|\hat{\gamma}_k}{\sum_{k=1}^{NC}|b_k|}$ and $\hat{\Gamma}_i = [\hat{\gamma}_{k_{i,1}}, \hat{\gamma}_{k_{i,2}}, \ldots, \hat{\gamma}_{k_{i,C}}]^{tr}$ instead of σ and Γ_i in Algorithm 1, respectively. Here $\hat{\sigma}$ is the label proportion of all training data points and we use it as an estimate of the clean prior σ. Likewise, $\alpha_i = \Gamma_i^{-tr}\sigma$ in Algorithm 1 should be replaced with $\hat{\alpha}_i = \hat{\Gamma}_i^{-tr}\hat{\sigma}$ and we would like to use $\hat{\Gamma}_i$, $\hat{\sigma}$, and $\hat{\alpha}_i$ to calculate $\hat{T}_i$ as an estimate of T_i. For this to make sense, we need $\hat{\alpha}_i = \hat{\Gamma}_i^{-tr}\hat{\sigma} \in \mathring{\Delta}^C$, which is equivalent to $\hat{\sigma}$ being in the interior of the convex hull of $\{\hat{\gamma}_{k_{i,c}} : c \in \mathcal{Y}\}$ for all i. However, this may not be the case in practice. Thus, we consider two heuristics to estimate $\hat{T}_i$ as real-world adaptations of the LLPFC-ideal algorithm. The first, called LLPFC-uniform, is presented in Algorithm 2 in the Appendix which sets $\hat{\alpha}_i$ by counting the occurrences of the noisy labels. This is motivated by our model wherein α_i is the noisy class prior for the i-th group. The second, called LLPFC-approx, is presented in Algorithm 3 in the Appendix and sets $\hat{\alpha}_i$ to be the solution of $\arg\min_{\alpha\in\Delta^C} \left\|\hat{\sigma} - \hat{\Gamma}_i\alpha\right\|_2^2$. Both real-world adaptations perform reasonably well in experiments.

6 Experiments

We compare against three previous works that have studied LLP applying deep learning to image data: Dulac-Arnold et al. [9] study the KL loss described in the introduction, and a novel loss based on optimal transport. They find that KL performs just as well as the novel loss. Liu et al. [22] employ the KL loss within a generative adversarial framework (LLPGAN). Tsai and Lin [46] propose augmenting the KL loss with a regularizer from semi-supervised learning and show improved performance (LLPVAT). We compare both LLPFC-uniform and LLPFC-approx against the KL loss, LLPGAN, and LLPVAT to clearly establish which empirical objective is better. Recent papers on multiclass LLP for which code is not available were not included [23, 16].

We generate bags with fixed, equal sizes in $\{32, 64, 128, 256, 512, 1024, 2048\}$. To generate each bag, we first sample a label proportion γ from the uniform distribution on Δ^C. Then we sample data points from a benchmark dataset without replacement using a multinomial distribution with parameter γ. It should be noted that Tsai and Lin [46], Dulac-Arnold et al. [9], and Liu et al. [22] generate bags by shuffling all data points and making every B data points a bag where B is a fixed bag size. Their method is equivalent to sampling data points without replacement using a multinomial distribution with a fixed parameter $\gamma = \left\{\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}\right\}$. As noted by Scott and Zhang [39], this leads to bags with very similar label proportions which makes the learning task much more challenging. All models are trained on a single Nvidia Tesla v100 GPU with 16GB RAM. We repeat each experiment 5 times and report the mean test accuracy and standard deviation.

For the comparison against KL and LLPVAT, we perform experiments on three benchmark image datasets: the "letter" split of EMNIST [6], SVHN [29], and CIFAR10 [17]. To show that our approach is robust to the choice of architecture, we experiment with three different networks: Wide ResNet-16-4 [52], ResNet18 [11], and VGG16 [42]. We train these networks with the parameters suggested in the original papers. The test accuracies are reported in Tables 1, 2, and 3 in the appendix.

Since convergence in the GAN framework is sensitive to the choice of architecture and hyperparameters, we compare LLPFC against LLPGAN using the architecture proposed in the original paper along with the hyperparameters suggested in their code[1]. It should be noted that for LLPFC we only use the discriminator for classification and did not use the generator to augment data. Since Liu et al. [22] only provide hyperparameters for colored images, we perform experiments on SVHN and CIFAR10 only. The test accuracies are reported in Table 4 in the appendix.

LLPFC-uniform and LLPFC-approx substantially outperform the competitors in a clear majority of settings. In a few settings, the competitors perform better, but LLPFC is not far behind. On the other hand, in many settings, the competitors perform very poorly. Quantitatively, comparing mean accuracies across all models and bag sizes, min(llpfc-uniform, llpfc-approx) - max(kl, llpgan, llpvat) $> 22\%$ on CIFAR10, > 27 % on SVHN, and > 28 % on EMNIST. In addition, all three competitors perform gradient descent with minibatches of bags and the GPU could potentially run out of memory when the bag size is large. Our implementation, which also uses stochastic optimization (with random regrouping of bags at every epoch), does not suffer from this phenomenon. Full experimental details are in an appendix.

7 Conclusions and Future Work

We propose a theoretically supported approach to LLP by reducing it to learning with label noise and using the forward correction (FC) loss. An excess risk bound and generalization error analysis are established. Our approach outperforms leading existing methods in deep learning scenarios across multiple datasets and architectures. A limitation of our approach is that the theory makes an assumption that may not be valid in practice. Future research directions include optimizing the grouping of bags and adapting LLPFC to other objectives beyond accuracy.

[1] https://github.com/liujiabin008/LLP-GAN

References

[1] Ehsan M. 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)*. 2017, pp. 1017–1024.

[2] 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, 2018, pp. 768–776. ISBN: 978-3-030-00934-2.

[3] S. Chen, B. Liu, M. Qian, and C. Zhang. "Kernel K-means Based Framework for Aggregate Outputs Classification". In: *2009 IEEE International Conference on Data Mining Workshops*. 2009, pp. 356–361.

[4] 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*. MM '14. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 367–376. ISBN: 9781450330633. DOI: 10.1145/2647868.2654935. URL: https://doi.org/10.1145/2647868.2654935.

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

[6] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. "EMNIST: an extension of MNIST to handwritten letters". In: *ArXiv* abs/1702.05373 (2017).

[7] Lucio Mwinmaarong Dery, Benjamin Nachman, Francesco Rubbo, and Ariel Schwartzman. "Weakly Supervised Classification For High Energy Physics". In: *Journal of Physics: Conference Series* 1085 (Sept. 2018), p. 042006. DOI: 10.1088/1742-6596/1085/4/042006. URL: https://doi.org/10.1088/1742-6596/1085/4/042006.

[8] Yongke Ding, Yuanxiang Li, and W. Yu. "Learning from label proportions for SAR image classification". In: *EURASIP Journal on Advances in Signal Processing* 2017 (2017), pp. 1–12.

[9] Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. "Deep multi-class learning from label proportions". In: *ArXiv* abs/1905.12909 (2019).

[10] Jean Gallier and Jocelyn Quaintance. *Linear Algebra and Optimization with Applications to Machine Learning*. World Scientific, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[12] J. Hernández-González, I. Inza, Lorena Crisol-Ortíz, M. A. Guembe, M. J. Iñarra, and J. A. Lozano. "Fitting the data from embryo implantation prediction: Learning from label proportions". In: *Statistical Methods in Medical Research* 27.4 (2018), pp. 1056–1066.

[13] J. Hernández-González, I. Inza, and J. A. Lozano. "Learning Bayesian network classifiers from label proportions". In: *Pattern Recognition* 46.12 (2013), pp. 3425–3440.

[14] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

[15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[16] Ryoma Kobayashi, Yusuke Mukuta, and Tatsuya Harada. *Risk Consistent Multi-Class Learning from Label Proportions*. 2022. DOI: 10.48550/ARXIV.2203.12836. URL: https://arxiv.org/abs/2203.12836.

[17] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: *University of Toronto* (May 2012).

[18] Hendrik Kück and Nando de Freitas. "Learning about Individuals from Group Statistics". In: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. 2005, pp. 332–339.

[19] K. Lai, F. X. Yu, M. Chen, and S. Chang. "Video Event Detection by Inferring Temporal Instance Labels". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2251–2258.

[20] Kuan-Ting Lai, Felix Yu, Ming-Syan Chen, and S. Chang. "Video Event Detection by Inferring Temporal Instance Labels". In: June 2014. DOI: 10.1109/CVPR.2014.288.

[21] 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.

[22] 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*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/4fc848051e4459b8a6afeb210c3664ec-Paper.pdf.

[23] Jiabin Liu, Bo Wang, Xin Shen, Zhiquan Qi, and Yingjie Tian. "Two-stage Training for Learning from Label Proportions". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 2737–2743. DOI: 10.24963/ijcai.2021/377. URL: https://doi.org/10.24963/ijcai.2021/377.

[24] Nan Lu, Shida Lei, Gang Niu, Issei Sato, and Masashi Sugiyama. "Binary Classification from Multiple Unlabeled Datasets via Surrogate Set Classification". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 7134–7144. URL: https://proceedings.mlr.press/v139/lu21c.html.

[25] Andreas Maurer. "A vector-contraction inequality for Rademacher complexities". In: *International Conference on Algorithmic Learning Theory*. Springer. 2016, pp. 3–17.

[26] Colin McDiarmid. "On the method of bounded differences". In: *Surveys in combinatorics* 141.1 (1989), pp. 148–188.

[27] Ron Meir and Tong Zhang. "Generalization Error Bounds for Bayesian Mixture Algorithms". In: *J. Mach. Learn. Res.* 4.null (Dec. 2003), pp. 839–860. ISSN: 1532-4435.

[28] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. "Learning with Noisy Labels". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[30] Giorgio Patrini, A. Rozza, A. Menon, R. Nock, and Lizhen Qu. "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2233–2241.

[31] Mark S Pinsker. *Information and information stability of random variables and processes*. Holden-Day, 1964.

[32] R. Poyiadzi, R. Santos-Rodriguez, and N. Twomey. "LABEL PROPAGATION FOR LEARNING WITH LABEL PROPORTIONS". In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2018, pp. 1–6.

[33] Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. "Learning With Label Proportions via NPSVM". In: *IEEE Transactions on Cybernetics* 47 (2017), pp. 3293–3305.

[34] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, and Quoc V. Le. "Estimating Labels from Label Proportions". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 776–783. ISBN: 9781605582054. DOI: 10.1145/1390156.1390254. URL: https://doi.org/10.1145/1390156.1390254.

[35] Brendan van Rooyen and Robert C. Williamson. "A Theory of Learning with Corrupted Labels". In: *Journal of Machine Learning Research* 18.228 (2018), pp. 1–50.

[36] Stefan Rueping. "SVM Classifier Estimation from Group Probabilities". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 911–918. ISBN: 9781605589077.

[37] Rishi Saket. "Learnability of Linear Thresholds from Label Proportions". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. URL: https://openreview.net/forum?id=5BnaKeEwuYk.

[38] Rishi Saket, Aravindan Raghuveer, and Balaraman Ravindran. "On Combining Bags to Better Learn from Label Proportions". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 5913–5927. URL: https://proceedings.mlr.press/v151/saket22a.html.

[39] Clayton Scott and Jianxin Zhang. "Learning from Label Proportions: A Mutual Contamination Framework". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 22256–22267. URL: https://proceedings.neurips.cc/paper/2020/file/fcde14913c766cf307c75059e0e89af5-Paper.pdf.

[40] Yong Shi, Limeng Cui, Zhensong Chen, and Zhiquan Qi. "Learning from label proportions with pinball loss". In: *International Journal of Machine Learning and Cybernetics* 10 (2017), pp. 187–205.

[41] Yong Shi, Jiabin Liu, Zhiquan Qi, and Bo Wang. "Learning from label proportions on high-dimensional data". In: *Neural Networks* 103 (2018), pp. 9–18. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2018.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0893608018300893.

[42] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[43] Ingo Steinwart. "How to Compare Different Loss Functions and Their Risks". In: *Constructive Approximation* 26 (2007), pp. 225–287.

[44] Marco Stolpe and Katharina Morik. "Learning from Label Proportions by Optimizing Cluster Model Selection". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 349–364. ISBN: 978-3-642-23808-6.

[45] 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)*. 2017, pp. 445–454. DOI: 10.1109/ICDM.2017.54.

[46] Kuen-Han Tsai and Hsuan-Tien Lin. "Learning from Label Proportions with Consistency Regularization". In: *Proceedings of The 12th Asian Conference on Machine Learning, ACML 2020, 18-20 November 2020, Bangkok, Thailand*. Ed. by Sinno Jialin Pan and Masashi Sugiyama. Vol. 129. Proceedings of Machine Learning Research. PMLR, 2020, pp. 513–528. URL: http://proceedings.mlr.press/v129/tsai20a.html.

[47] B. Wang, Z. Chen, and Z. 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)*. Vol. 3. 2015, pp. 56–59.

[48] Nik Weaver. *Lipschitz Algebras*. WORLD SCIENTIFIC, 1999. DOI: 10.1142/4100, eprint: https://www.worldscientific.com/doi/pdf/10.1142/4100. URL: https://www.worldscientific.com/doi/abs/10.1142/4100.

[49] Robert C. Williamson, Elodie Vernet, and Mark D. Reid. "Composite Multiclass Losses". In: *Journal of Machine Learning Research* 17.222 (2016), pp. 1–52. URL: http://jmlr.org/papers/v17/14-294.html.

[50] Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. "$\propto$SVM for Learning with Label Proportions". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, June 2013, pp. 504–512. URL: http://proceedings.mlr.press/v28/yu13a.html.

[51] Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. *On Learning from Label Proportions*. Tech. rep. arXiv:1402.5902. 2015.

[52] Sergey Zagoruyko and Nikos Komodakis. "Wide Residual Networks". In: *Proceedings of the British Machine Vision Conference (BMVC)*. Ed. by Edwin R. Hancock Richard C. Wilson and William A. P. Smith. BMVA Press, Sept. 2016, pp. 87.1–87.12. ISBN: 1-901725-59-6. DOI: 10.5244/C.30.87. URL: https://dx.doi.org/10.5244/C.30.87.

[53] Mingyuan Zhang, Jane Lee, and Shivani Agarwal. "Learning from Noisy Labels with No Change to the Training Process". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 12468–12478. URL: https://proceedings.mlr.press/v139/zhang21k.html.

[54] Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael I. Jordan. "On the learnability of fully-connected neural networks". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Apr. 2017, pp. 83–91. URL: http://proceedings.mlr.press/v54/zhang17a.html.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are described in Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are included in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is included in the supplemental material. Datasets are public and we provide code to download them. We include a README file with instructions on how to reproduce experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] A concise description of experiments is in Section 6 in the main paper with full details in section B in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In table 1, 2, 3 and 4, we report the mean and standard deviation for 5 trials with different random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We provide the hardware information in Section 6 but not the amount of compute. The computational time varies for different experimental settings. It would be too exhaustive to present them given that we have nearly 300 different settings.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes. Creators of code, data, and models are all cited.
   (b) Did you mention the license of the assets? [N/A] We did not directly run experiments using others' code. However, we implement their algorithms with their code as a reference.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our experiment code is included in the supplemental materials.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]