LEVERAGED WEIGHTED LOSS FOR PARTIAL LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As an important branch of weakly supervised learning, partial label learning deals with data where each instance is assigned with a set of candidate labels, whereas only one of them is true. In this paper, we propose a family of loss functions named *Leveraged Weighted* (LW) loss function, which for the first time introduces the leverage parameter β to partial loss functions to leverage between losses on partial labels and residual labels (non-partial labels). Under mild assumptions, we achieve the relationship between the partial loss function and its corresponding ordinary loss that leads to the consistency in risk. Compared to the existing literatures, our result applies to both deterministic and stochastic scenarios, considers the loss functions of a more general form, and takes milder assumptions on the distribution of the partial label set. As special cases, with $\beta = 1$ and $\beta = 2$, the corresponding ordinary loss of our LW loss respectively match the binary classification loss and the *one-versus-all* (OVA) loss function. In this way, our theorems successfully explain the experimental results on parameter analysis, where $\beta = 1$ and especially $\beta = 2$ are considered as preferred choices for the leverage parameter β. Last but not least, real data comparisons show the high effectiveness of our LW loss over other state-of-the-art partial label learning algorithms.

1 INTRODUCTION

While labeled data is usually expensive, time consuming to collect, and sometimes requires human domain experts to annotate, partially labeled data is often relatively easier to obtain (Nguyen & Caruana, 2008b). Partially labeled problem, also called ambiguously labeled problem, refers to the task where each training example is associated a set of candidate labels, while only one is assumed to be true. Conceptually speaking, partial label learning lies between the traditional supervised learning with explicit supervision and the unsupervised learning with blind supervision, nevertheless, it is essentially different from the semi-supervised setting where both labeled and unlabeled data are available. The problem of learning from partially labeled examples naturally arises in a number of real-world scenarios such as web mining (Luo & Orabona, 2010), multimedia contents analysis (Cour et al., 2009; Zeng et al., 2013), ecoinformatics (Liu & Dietterich, 2012), etc. This field has attracted much attention of researchers and there is a rich body of literature on this problem.

Most of the existing analysis is established on the strategy of disambiguation, i.e. aiming to identify the ground-truth label from candidate label set. One intuitive strategy is to treat all the candidate labels in an equal manner and then average the outputs from all candidate labels for prediction. Following this strategy, for parametric models, Cour et al. (2011) designs the convex loss for partial label (CLPL), via which the averaged output from all candidate labels is distinguished from the outputs from non-candidate labels. In Hüllermeier & Beringer (2006), several nonparametric, instance-based algorithms for ambiguous learning based on greedy heuristics are proposed. Besides, Zhang & Yu (2015) establishes an instance-based approach, named IPAL, trying to identify the valid label of each partial label example via an iterative label propagation procedure. Although intuitive and easy to implement, the effectiveness of averaging strategy is heavily affected if the outputs of false positive labels overwhelm that of the truth label. Another way towards disambiguation treats the ground-truth label as a latent variable and identify it from the candidate label set through iterative refining procedure such as maximum likelihood criteria, or the maximum margin criteria. For instance, Jin & Ghahramani (2003) use an EM like algorithm with a discriminative log-linear model to disambiguate correct labels from incorrect ones and Grandvalet et al. (2004) adds a mini-

mum entropy term to the set of possible label distributions. Lv et al. (2020) generalize Jin's (Jin & Ghahramani, 2003) learning objectives, and seamlessly combines updating model parameters and identifying true labels.

However, many of these proposed algorithms rely on iterative non-convex learning. In order to tackle this issue, Liu & Dietterich (2012) proposes a probabilistic model, named the Logistic Stick-Breaking Conditional Multinomial Model (LSB-CMM), which maps data points to mixture components and then assigns to each mixture component a label drawn from a component-specific multinomial distribution. This optimization problem is similar to the problem of logistic regression and is also a concave maximization problem, which can be solved by any gradient-based method. In addition, within the margin-based learning framework, Nguyen & Caruana (2008a) formulates the partially labeled problem as a convex quadratic optimization through utilizing the L_2-norm regularization and the redefined hinge loss for partial label data. Unfortunately, the margin does not consider the predictive difference between the ground-truth label and other false positive labels, which may lead to suboptimal performance for the resulting maximum margin partial label approach. The fact can be easily observed that the major difficulty in making use of multi-class margin for partial label training examples lies in that the truth label information is not accessible to the learning algorithm. Therefore, an alternating optimization procedure is employed in Yu & Zhang (2016) to iteratively identify the ground-truth label and maximize the multi-class margin. Although these approaches adopting the two strategies are able to extract the relative labeling confidence of each candidate label, they fail to reflect the mutually exclusive relationships among different candidate labels and is generally conducted by focusing on manipulating the label space. Therefore, some novel partial learning methods arises taking advantage of new techniques such as feature-aware (Zhang et al., 2016), dictionaries (Chen et al., 2014), self-training (Feng & An, 2019) and label enhancement (Xu et al., 2019).

In this paper, we propose a family of loss functions named *Leveraged Weighted* (LW) loss function, leveraging between losses on partial labels and residual labels (non-partial labels). We examine the theoretical properties of our LW loss, especially the choice of leverage parameter β, from the perspective of risk consistency. Then we design the partial label learning algorithm by learning the weighting parameters and score functions iteratively. As follows are our contributions.

1) Under mild assumption that each untrue label appears independently in the partial label set, we achieve the relationship between the partial loss function and its corresponding ordinary loss such that their risks corresponds to each other. In this way, we theoretically guarantees the optimization of partial label learning leads to the optimization of ordinary-labeled multiclass classification. Compared to the existing literature about risk consistency, our result applies not only to the deterministic scenario but also to the stochastic scenario where the true output label is a probabilistic function of the input. Moreover, we make milder assumptions on the distribution of partial labels, and extend the previous result to the partial loss function of a more general form.
2) We propose a family of loss function for partial label learning named Leveraged Weighted (LW) loss function, which combines binary losses on each label and leverages between losses on partial labels and residual labels (non-partial labels). We highlight that it is the first time that the leverage parameter β is introduce into loss function for partial label learning. Risk consistency analysis presents the corresponding ordinary loss function of LW loss, and shows that $\beta = 1$ and especially $\beta = 2$ are preferred choices for the leverage parameter β. To be specific, with $\beta = 2$ and symmetric binary loss function, the corresponding ordinary loss of our LW loss exactly matches the *one-versus-all* (OVA) loss function proposed by Zhang (2004).
3) We conduct parameter analysis of our LW loss for partial label learning on four benchmark datasets, where the experimental results that $\beta = 1$ and especially $\beta = 2$ achieves high learning accuracy over the state-of-the-art partial label learning algorithm, which exactly verifies the theoretical analysis about the leverage parameter β. We also show the desirable performance of our algorithm under various data generation situations.

2 METHODOLOGY

We assume that $\mathcal{X} \subset \mathbb{R}^d$ is a non-empty feature space (or input space), $\mathcal{Y} := \{1, \ldots, K\} =: [K]$ is the ordinary label space and that $\boldsymbol{\mathcal{Y}} := \{\boldsymbol{Y} | \boldsymbol{Y} \subset \mathcal{Y}\} = 2^{[K]}$ is the partial label space, where $2^{[K]}$ is the collection of all subsets in $[K]$. In the partially supervised setting, for an input random variable

$X \in \mathcal{X}$, we have the corresponding ambiguity set (or partial label set) $\boldsymbol{Y} \in \boldsymbol{\mathcal{Y}}$, containing the true label, denoted by random variable $Y \in \mathcal{Y}$. The goal is to find the latent ground-truth label Y for the input X through observing the partial label set $\boldsymbol{Y}$. For the rest of this paper, y always represents the true label of input x unless otherwise specified. All proofs are shown in the supplementary material.

2.1 Fundamental Assumption

The fundamental assumptions focus on the distribution of the partial label set $\boldsymbol{Y}$.

For notational simplicity, we denote $\mathrm{P}(z \in \boldsymbol{y}|Y = y, X = x) := q_z$ for all $z \in [K]$. Assumption 1 implies that the true label y always resides in the partial label set $\boldsymbol{y}$, which exactly corresponds with the problem setting of partial label learning.

Assumption 1 *Denote y as the true label of an input x, then we assume*

$$q_y := \mathrm{P}(y \in \boldsymbol{y}|Y = y, X = x) = 1.$$

By Assumption 1, we have $\#|\boldsymbol{Y}| \geq 1$, and $\#|\boldsymbol{Y}| = 1$ holds if and only if $\boldsymbol{Y} = \{Y\}$, in which case the partial label learning problem reduces to multi-class classification with ordinary labels.

Assumption 2 *Denote y as the true label of an input x, then we assume for $z \neq y$*

$$q_z := \mathrm{P}(z \in \boldsymbol{y}|Y = y, X = x) < 1.$$

Assumption 2 corresponds with the assumptions in Cour et al. (2011); Lv et al. (2020), which guarantees the partial label learning problem to be ERM learnable.

Assumption 3 *When the true label y and the input x is given, the behavior of label $z \in \boldsymbol{y}$, where $z \neq y$ and $z \in [K]$, is independent, i.e. for $z_1, z_2 \in [K]$, $z_1 \neq z_2$ and $z_1, z_2 \neq y$, we have*

$$\mathrm{P}(\{z_1, z_2\} \in \boldsymbol{y}|Y = y, X = x) = \mathrm{P}(z_1 \in \boldsymbol{y}|Y = y, X = x) \cdot \mathrm{P}(z_2 \in \boldsymbol{y}|Y = y, X = x).$$

Assumption 3 states that when the true label y and the input x is given, the behavior of each label $z \neq y, z \in [K]$, whether belonging to the partial label set or not, is independent. According to the problem setting, the probability of each label $z \neq y$ being in the partial label set may be different. For instance, when the true label is *mule*, *dunkey* is more likely to be picked as a partial label than *cat*. However, when no additional information is given, it is perfectly natural to assume that the labelers make independent decision for each label, i.e. the situation of *cat* being a partial label doesn't affect whether *dunkey* is picked and vice versa.

Lemma 1 follows directly from Assumption 1 and 3. It shows the conditional distribution of partial label set $\boldsymbol{Y}$, which is essential in achieving the risk of our partial label learning algorithm.

Lemma 1 *When Assumption 1 and Assumption 3 hold, for all $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, we have*

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}|Y = y, X = x) = \prod_{s \in \boldsymbol{y}, s \neq y} q_s \cdot \prod_{t \notin \boldsymbol{y}} (1 - q_t).$$

2.2 Relationship between partial loss and ordinary loss

We first of all introduce some notations and key concepts. We denote $g(X) = (g_1(X), \ldots, g_K(X))$ as the decision function learned by an algorithm, where $g_z(X)$ is the score function for label $z \in [K]$. Larger $g_z(x)$ implies that x is more likely to come from class $z \in [K]$. Then the resulting classifier is $f(X) = \arg\max_{z \in [K]} g_z(X)$. We denote $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X))$ as the risk, also called generalization error, for the decision function $g(X)$ w.r.t. the partial loss function $\bar{\mathcal{L}}$. By definition, $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X)) := \mathbb{E}_{(X,\boldsymbol{Y})}[\bar{\mathcal{L}}(\boldsymbol{Y}, g(X))]$, measuring the average loss of a $g(X)$ learned through partial labels w.r.t. the joint distribution of $(X, \boldsymbol{Y})$. Similarly, we denote $\mathcal{R}(\mathcal{L}, g(X))$ as the ordinary risk for $g(X)$, where $\mathcal{L}$ is the loss function for learning with ordinary labels (X, Y), and by definition, $\mathcal{R}(\mathcal{L}, g(X)) := \mathbb{E}_{(X,Y)}[\mathcal{L}(Y, g(X))]$, w.r.t. the joint distribution of ordinary-labeled data (X, Y).

In this part, we are interested in the partial risk $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X))$. We wonder under what circumstances the partial risk $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X))$ corresponds to the ordinary risk $\mathcal{R}(\mathcal{L}, g(X))$, i.e. what algorithm design

for partial label learning will achieve the same theoretical effectiveness as when using ordinary labels. Theorem 1 answers this problem from the perspective of loss function.

Theorem 1 *Denote $\mathcal{L}: \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}^+$ as a loss function for the ordinary-label classification, and $\bar{\mathcal{L}}: \boldsymbol{\mathcal{Y}} \times \mathbb{R}^K \to \mathbb{R}^+$ as the loss function for the partial-label classification. If the loss function for ordinary classification problem $\mathcal{L}(y, g(x))$ is of the form*

$$\mathcal{L}(y, g(x)) = \sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}} \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} | Y = y, X = x) \bar{\mathcal{L}}(\boldsymbol{y}, g(x)),$$

we have $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X)) = \mathcal{R}(\mathcal{L}, g(X))$. Moreover, under Assumption 1 and Assumption 3, we have

$$\mathcal{L}(y, g(x)) = \sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}^y} \prod_{s \in \boldsymbol{y}, s \neq y} q_s \prod_{t \notin \boldsymbol{y}} (1 - q_t) \bar{\mathcal{L}}(\boldsymbol{y}, g(x)),$$

where $\boldsymbol{\mathcal{Y}}^y := \{\boldsymbol{y} \in \boldsymbol{Y} | y \in \boldsymbol{y}\}$ with y denoting the true label of x.

2.3 Leveraged Weighted (LW) Loss Function

In this paper, we propose a family of loss function for partial label learning named *Leveraged Weighted* (LW) loss function. We adopt a multiclass scheme frequently used for the fully supervised setting (Crammer & Singer, 2001; Rifkin & Klautau, 2004; Zhang, 2004; Tewari & Bartlett, 2005), that combines binary losses $\psi(\cdot): \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ on the score functions g_z, $z \in [K]$, to create a multiclass loss. We highlight that it is the first time that the leverage parameter β is introduce into loss function for partial label learning, which leverages between losses on partial labels and residual labels (non-partial labels). To be specific, the partial loss function of concern is of the form

$$\bar{\mathcal{L}}_\psi(\boldsymbol{y}, g(x)) = \sum_{z \in \boldsymbol{y}} w_z \psi(g_z(x)) + \beta \cdot \sum_{z \notin \boldsymbol{y}} w_z \psi(-g_z(x)). \tag{1}$$

It consists of three components.

- A binary loss function $\psi(\cdot): \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$, which forces g_z to be larger when z resides in the partial label set $\boldsymbol{y}$, while $\psi(-g_z)$ punishes large g_z when $z \notin \boldsymbol{y}$.
- Weighting parameters $w_z \geq 0$ on $\psi(g_z)$ for $z \in [K]$. Generally speaking, we would like to assign more weights to the loss of labels that are more likely to be the true label.
- The leverage parameter $\beta \geq 0$ that distinguishes between partial labels and non-partial ones. Larger β quickly rules out non-partial labels during training. However, it also lessens the weights assigned to partial labels.

We mention that the partial loss proposed in (1) is a general form. Some special cases include

1) Taking $\beta = 1$, $w_z = 1/\#|\boldsymbol{y}|$ for $z \in \boldsymbol{y}$ and $w_z = 1$ for $z \notin \boldsymbol{y}$, we achieve the "naïve" partial loss proposed by Jin & Ghahramani (2002), the form of which is

$$\bar{\mathcal{L}}_\psi^{\text{naïve}}(\boldsymbol{y}, g(x)) = \frac{1}{\#|\boldsymbol{y}|} \sum_{y \in \boldsymbol{y}} \psi(g_y(x)) + \sum_{y \notin \boldsymbol{y}} \psi(-g_y(x)). \tag{2}$$

2) By taking $w_{z^*} = 1$ where $z^* = \arg\max_{z \in \boldsymbol{y}} g_z$, $w_z = 0$ for $z \in \boldsymbol{y} \setminus \{z^*\}$, $w_z = 1$ for $z \notin \boldsymbol{y}$, and $\beta = 1$, we achieve the partial loss function adopting the "hardmax" scheme proposed by Cour et al. (2011), with the form of

$$\bar{\mathcal{L}}_\psi^{\text{hardmax}}(\boldsymbol{y}, g(x)) = \psi(\max_{y \in \boldsymbol{y}} g_y(x)) + \sum_{y \notin \boldsymbol{y}} \psi(-g_y(x)). \tag{3}$$

3) By taking $w_{z^*} = 1$ where $z^* = \arg\max_{z \in \boldsymbol{y}} g_z$, $w_z = 0$ for $z \in \boldsymbol{y} \setminus \{z^*\}$, $w_z = 0$ for $z \notin \boldsymbol{y}$, and $\beta = 0$, we achieve the partial loss function adopting the "softmax" scheme proposed by Lv et al. (2020), with the form of

$$\bar{\mathcal{L}}_\psi^{\text{softmax}}(\boldsymbol{y}, g(x)) = \psi(\max_{y \in \boldsymbol{y}} g_y(x)) = \min_{y \in \boldsymbol{y}} \psi(g_y(x)). \tag{4}$$

2.4 RISK CONSISTENCY OF LW LOSS

Theorem 2 *If the partial loss function is of the form in (1), then its corresponding ordinary loss function has the form*

$$\mathcal{L}_\psi(y, g(x)) = w_y\psi(g_y(x)) + \sum_{z\neq y} w_z q_z \big[\psi(g_z(x)) + \beta\psi(-g_z(x))\big], \tag{5}$$

such that $\bar{\mathcal{R}}(\bar{\mathcal{L}}, g(X)) = \mathcal{R}(\mathcal{L}, g(X))$.

Theorem 2 shows the consistency in risk between LW loss for partial label learning and its corresponding loss for ordinary label learning. Compared to the previous result in Lv et al. (2020), where partial loss function (5) is considered under the deterministic scenario, i.e. the true label of a point can be uniquely determined by some measurable function $f : \mathcal{X} \to \mathcal{Y}$, we highlight that our result applies to both deterministic and stochastic scenarios, i.e. Theorem 2 also holds when the output label is a probabilistic function of the input. We extend the previous result to the partial loss function of a more general form, with (5) being one of the special cases. This enables us to take a step further to compare between various partial loss functions w.r.t. their corresponding ordinary loss. Moreover, compared to Feng et al. (2020), where the partial label set is assumed to be uniformly sampled, our assumptions are much weaker and closer to the reality.

In the following analysis, we focus on symmetric binary loss $\psi(\cdot)$ for their fine theoretical properties. We remark that commonly adopted loss functions such as zero-one loss, Sigmoid loss, Ramp loss, etc. satisfies the symmetric condition. (See Ishida et al. (2017))

Corollary 1 *If* $\psi(\cdot)$ *is symmetric, i.e.* $\psi(g_z(x)) + \psi(-g_z(x)) = 1$, *and*

1) $\beta = 0$, *then we have* $\mathcal{L}_\psi(y, g(x)) = w_y\psi(g_y(x)) + \sum_{z\in\boldsymbol{y}, z\neq y} w_z q_z \psi(g_z(x))$.

2) $\beta = 1$, *then we have* $\mathcal{L}_\psi(y, g(x)) = w_y\psi(g_y(x)) + \sum_{z\neq y} w_z q_z$.

3) $\beta = 2$, *then we have* $\mathcal{L}_\psi(y, g(x)) = w_y\psi(g_y(x)) + \sum_{z\neq y} w_z q_z \psi(-g_z(x)) + \sum_{z\neq y} w_z q_z$.

When $\psi(\cdot)$ is symmetric $\beta = 0$, the average loss $\psi(-g_z)$ on the residual labels $z \notin \boldsymbol{y}$ failed to offset the average loss $\psi(g_z)$ on partial labels $z \in \boldsymbol{y} \setminus \{y\}$. Therefore, in addition to focus on the true label y, the corresponding ordinary loss also give positive weights to those partial, but unfortunately untrue, labels, which may harm the effectiveness of the partial label learning. This problem can only be avoided when $w_z = 0$ for $z \in \boldsymbol{y} \setminus \{y\}$, which corresponds to the "softmax" loss function (4).

When $\psi(\cdot)$ is symmetric and $\beta = 1$, the corresponding ordinary loss function $\mathcal{L}_\psi(y, g(x))$ is a linear transformation of $\psi(g_y(x))$. In this case, when optimizing the partial loss function $\bar{\mathcal{L}}_\psi$, we are at the same time optimizing the corresponding ordinary loss $\mathcal{L}_\psi := \psi(g_z(x))$.

When $\psi(\cdot)$ is symmetric and $\beta = 2$, the corresponding ordinary loss function $\mathcal{L}_\psi(y, g(x))$ is a linear combination of $\psi(g_y(x))$ and $\psi(-g_z(x))$ for $z \neq y$. When taking $w_z = 1/q_z$ for $z \in [K]$, we have

$$\mathcal{L}_\psi(y, g(x)) = \psi(g_y(x)) + \sum_{z\neq y} \psi(-g_z(x)) + K - 1,$$

which corresponds to the *one-versus-all* (OVA) loss function proposed by Zhang (2004).

As a matter of fact, the leverage parameter β decides to what extent the average extra loss $\psi(g_z(x))$ on $z \in \boldsymbol{y} \setminus \{y\}$ is compensated by the average $\psi(-g_z(x))$ on $z \notin \boldsymbol{y}$, and Corollary 1 indicates that $\beta = 1$ and especially $\beta = 2$ can be good choices.

3 MAIN ALGORITHM

In the theoretical analysis of the previous section, we focus on partial and ordinary loss functions that consistent in risk. However, in experiment, the risk for partial label loss is not directly accessible

Algorithm 1 Leveraged Weighted Loss for Partial Label Learning

Input: Training data $D_n := \{(x_1, \boldsymbol{y}_1), \ldots, (x_n, \boldsymbol{y}_n)\}$;
Number of Training Epochs T;
Learning rate $\rho > 0$;
For $i = 1, \ldots, n$ initialize $w_{z,i}^{(0)} = \frac{1}{\#|\boldsymbol{y}_i|}$ for $z \in \boldsymbol{y}_i$ and $w_{z,i}^{(0)} = \frac{1}{K - \#|\boldsymbol{y}_i|}$ for $z \notin \boldsymbol{y}_i$.
for $t = 1$ **to** T **do**
Calculate empirical risk $\bar{\mathcal{R}}_{D_n}^{(t)}(\bar{\mathcal{L}}^{(t-1)}, g(x; \theta^{(t-1)}))$ by (7);
Update parameter $\theta^{(t)}$ for score functions by (8) and achieve $g(x; \theta^{(t)})$.
Update weighting parameters $w_{z,i}^{(t)}$ by (6);
end for
Output: Decision function achieved by $\hat{y} = \arg\min_{z \in [K]} g_z(x; \theta^{(T)})$.

since the underlying distribution of $\mathrm{P}(X, \boldsymbol{Y})$ is unknown. Instead, the we can measure the empirical risk of a learning algorithm on the partially labeled sample $D_n := \{(x_1, \boldsymbol{y}_1), \ldots, (x_n, \boldsymbol{y}_n)\}$, which is $\bar{\mathcal{R}}_{D_n}(\bar{\mathcal{L}}, g(X)) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathcal{L}}(\boldsymbol{y}_i, g(x_i))$. In the following experiments, we select sigmoid loss as the binary loss in our LW loss function for our method due to its symmetric property. To be specific, we let $\psi(g_z(x)) := 1/(1 + \exp(g_z(x)))$ for $z \in [K]$.

Iterative Learning Process of Weighting Parameters. We take the network parameters θ for score functions $g(x) := \big(g_1(x), \ldots, g_K(x)\big)$ into consideration, and write $g(x; \theta)$ and $g_z(x; \theta)$ instead. Now we turn to determine the weighting parameters, another important component in our LW loss. Our goal is to assign larger weight to the binary loss of true label y, and assign weight as small as possible for the binary losses for $z \neq y$. However, since we cannot directly observe the true label y for input x from the partial labeled data, the weighting parameters cannot be directly assigned. Therefore, inspired by the EM algorithm and PRODEN proposed by Lv et al. (2020), we learn the weighting parameters through an iterative process instead of assigning fixed values. The overall algorithm is shown in Algorithm 1.

At the first glance, it seems natural to perform the softmax operation on all score functions $g_z(x; \theta)$ for $z \in [K]$, i.e. to update the weighting parameters at t-th step by $w_z^{(t)} = \frac{\exp(g_z(x; \theta^{(t)}))}{\sum_{z \in [K]} \exp(g_z(x; \theta^{(t)}))}$. However, since larger score implies higher probability of a label to be the true label, the weights for partial labels tend to grow rapidly through training, resulting in much larger weights for the partial losses than the residual ones. Finally, as the training epochs grow, the losses on residual labels as well as the leverage parameter β gradually lose their functions, which we are not pleased to see.

To conquer this problem, we perform the softmax operation on the score functions $g_z(x; \theta)$ for $z \in \boldsymbol{y}$ and those for $z \notin \boldsymbol{y}$ respectively, i.e.

$$w_z^{(t)} = \frac{\exp(g_z(x; \theta^{(t)}))}{\sum_{z \in \boldsymbol{y}} \exp(g_z(x; \theta^{(t)}))} \text{ for } z \in \boldsymbol{y}, \text{ and } w_z^{(t)} = \frac{\exp(g_z(x; \theta^{(t)}))}{\sum_{z \notin \boldsymbol{y}} \exp(g_z(x; \theta^{(t)}))} \text{ for } z \notin \boldsymbol{y}. \tag{6}$$

By this means we have $\sum_{z \in \boldsymbol{y}} w_z^{(t)} = \sum_{z \notin \boldsymbol{y}} w_z^{(t)} = 1$. Thus the leverage parameter β can gain the full control of the relative scale of losses on partial labels and residual labels. Note that $w_z^{(t)}$ varies with sample instances. Thus for each instance $(x_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, we denote the weighting parameter as $w_{z,i}^{(t)}$.

Then we achieve the empirical risk function for the t-th step as

$$\begin{aligned}
&\bar{\mathcal{R}}_{D_n}^{(t)}(\bar{\mathcal{L}}^{(t-1)}, g(x; \theta^{(t-1)})) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathcal{L}}^{(t-1)}(\boldsymbol{y}_i, g(x_i; \theta^{(t-1)})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Bigg(\sum_{z \in \boldsymbol{y}_i} w_{z,i}^{(t-1)} / \big(1 + \exp(g_z(x; \theta^{(t-1)}))\big) + \beta \cdot \sum_{z \notin \boldsymbol{y}_i} w_{z,i}^{(t-1)} / \big(1 + \exp(-g_z(x; \theta^{(t-1)}))\big) \Bigg),
\end{aligned} \tag{7}$$

and update the parameters in score functions, i.e. let $\rho > 0$ be the learning rate, we have for $z \in [K]$,

$$\theta^{(t)} = \theta^{(t-1)} - \rho \cdot \partial \bar{\mathcal{R}}_{D_n}^{(t)}(\bar{\mathcal{L}}^{(t-1)}, g(x; \theta^{(t-1)}))/\partial\theta. \tag{8}$$

4 EXPERIMENTS

We base our experiments on four benchmark datasets: MNIST (LeCun et al., 1998), Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2018), Fashion-MNIST (FMNIST) (Xiao et al., 2017), and CIFAR-10 (Krizhevsky et al., 2009). According to Assumptions 1, 2 and 3, we generate partially labeled data by making $K-1$ independent decisions for labels $z \neq y$, where each label z has probability q_z to enter the partial label set. Note that the true label y always resides in the partial label set $\boldsymbol{y}$, and we accept the occasion that $\boldsymbol{y} = [K]$.

4.1 PARAMETER ANALYSIS

In this part, we let the partially labeled data be generated with equal probability, i.e. $q_z = q \in [0, 1)$ for $z \in [K]$. We take both linear model and 5-layer perceptron (MLP) to verify the effectiveness of our LW loss and algorithm. More implementation details are shown in the supplementary materials.

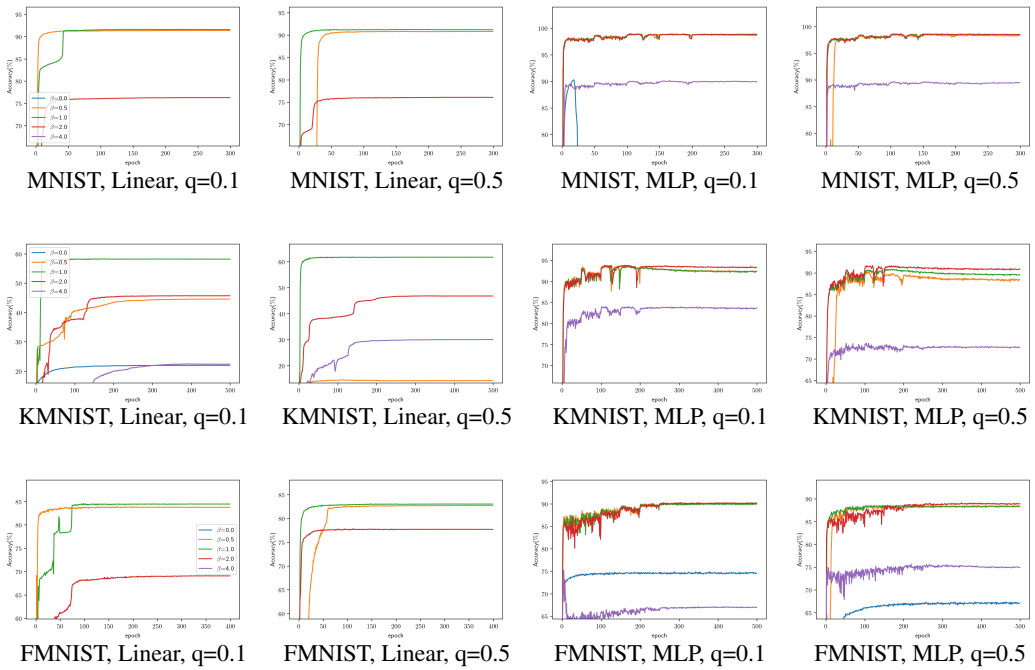

Figure 1: The study of leverage parameter β on different datasets, models and q.

Figure 1 presents the test accuracy during the training process. Since models converges fast in some cases, we just show the results of first 300 or 400 epochs. We mention that the result of $\beta = 0$ is sometimes too bad to be in the scope, thus not shown in some figures. From Figure 1, we observe that under linear models, $\beta = 1$ brings satisfactory results compared with β with other values. With MLP model, $\beta = 2$ is always the optimal parameter and $\beta = 1$ is the second best. These results exactly correspond to the the theoretical results presented in Corollary 1, which explains the experimental results from the perspective of the relationship between partial loss and ordinary multi-classification loss. It is also worth mentioning that the case of $\beta = 0$ is equivalent to using the PRODEN algorithm with Sigmoid loss function instead of with cross entropy loss in the original paper. Figure 1 also shows that our algorithm (with $\beta = 1$ or $\beta = 2$) performs far better than PRODEN (with $\beta = 0$). In addition, for $\beta = 4$ and $\beta = 0$, the performance deteriorates drastically since the loss concentrates too much on either partial set or residual set.

4.2 OVERALL ACCURACY

In this section, we compare our algorithm with the state-of-the-art partial label learning algorithm PRODEN Lv et al. (2020) (with cross entropy loss), which only focuses on the loss induced by the

partial label set, i.e. the case $\beta = 0$. However, they use cross entropy loss function instead of sigmoid function and obtain the superior results. First, we provide the accuracy on the four benchmark datasets. The parital label sets are generated by five random samplings and the parameters of PRODEN are selected according to the original paper. We follow the experimental settings in Section 4.1, and in addition, we train a 12-layer ConvNet (Laine & Aila, 2017) for CIFAR-10. Moreover, we take average of test accuracy during the last ten epochs as the final results to obtain stable results.

Table 1: Accuracy comparisons on benchmark datasets.

Dataset	Model	$q = 0.1$		$q = 0.3$		$q = 0.5$		$q = 0.7$	
		β^*	Accuracy	β^*	Accuracy	β^*	Accuracy	β^*	Accuracy
MNIST	OURS	2	**98.78 (0.08)***	2	**98.65 (0.03)***	2	**98.52 (0.07)***	2	**98.12 (0.10)***
	PRODEN	0	98.55 (0.13)	0	98.47 (0.12)	0	98.36 (0.08)	0	98.02 (0.05)
KMNIST	OURS	2	**93.38 (0.18)***	2	**92.24 (0.14)***	2	**90.70 (0.48)***	2	**88.84 (0.41)***
	PRODEN	0	91.07 (0.20)	0	90.45 (0.12)	0	88.67 (0.22)	0	85.53 (0.61)
FMNIST	OURS	1	**89.98 (0.10)***	1	**89.55 (0.14)***	2	**88.86 (0.17)**	4	**87.66 (0.09)***
	PRODEN	0	89.48 (0.11)	0	89.18 (0.18)	0	88.85 (0.16)	0	87.40 (0.14)
CIFAR-10	OURS	1	**90.62 (0.08)***	1	**89.53 (0.11)***	1	**86.10 (0.11)***	-	-
	PRODEN	0	89.00 (0.18)	0	87.76 (0.23)	0	84.94 (0.31)	-	-

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis beside each value. we apply the Wilcoxon signed-rank test, and mark with $*$ the results that are significantly better than others with significance level $\alpha = 0.05$.

From Table 1, we find our method with optimal $\beta = 2$ or $\beta = 1$ outperforms PRODEN in most cases, which corresponds exactly to Corollary 1.

4.3 The Influence of Data Generation

In the data generation of previous subsections, the untrue partial labels are selected with equal probabilities, i.e. $q_z = q$ for $z \neq y$. In reality, however, some labels may be more analogous to the true label than others, and thus the probabilities q_z for these labels may naturally be higher than others. In this subsection we simulate the situation where each true label has two similar labels (adjacent labels in experiment) with higher probability $q_{adj} > q$ to be partial labels, while all other labels enjoy equal probability $q = 0.1$. Note that when $q_{adj} = q$, the data generation reduced to the equal probability case.

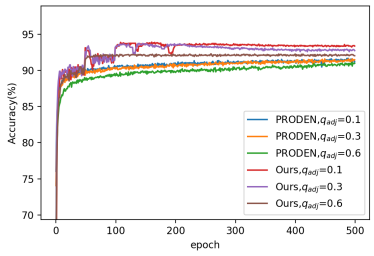

Figure 2: The influence of q_{adj}.

Figure 2 shows the comparison between our algorithm and PRODEN with various q_{adj}. Since higher q_{adj} indicates that true labels are more likely to be confused by similar labels, both our algorithm and PRODEN performs worse as q_{adj} increases. Nonetheless, partial label learning with our LW loss always enjoys higher accuracy than PRODEN.

5 Conclusion

In this paper, we propose a family of loss functions named *Leveraged Weighted* (LW) loss function, where we for the first time introduce the leverage parameter β to the partial loss function. From the theoretical perspective, we prove the relationship between our proposed LW loss for partial label learning and the ordinary multiclass classification losses that achieve the same risk as the LW risk. We mention that our theoretical analysis considers more general situations than the existing literatures. Then we examine the weighting parameter β from both the theoretical and empirical perspectives, both of which reaches the concensus that $\beta = 1$ and especially $\beta = 2$ are preferred choices for our LW loss. Specifically, with $\beta = 2$, the corresponding ordinary loss of LW matches exactly to the *one-versus-all* (OVA) loss function, which theoretically guarantees the performance of our LW loss. Last but not least, comparisons with the state-of-the-art PRODEN shows the advantage of our methods under various data generation situations.

REFERENCES

Yi-Chen Chen, Vishal M Patel, Rama Chellappa, and P Jonathon Phillips. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security*, 9(12):2076–2088, 2014.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018. URL http://arxiv.org/abs/1812.01718.

Timothee Cour, Benjamin Sapp, Chris Jordan, and Ben Taskar. Learning from ambiguously labeled images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 919–926. IEEE, 2009.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2001. URL http://jmlr.org/papers/v2/crammer01a.html.

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3542–3549, 2019.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. *CoRR*, abs/2007.08929, 2020. URL https://arxiv.org/abs/2007.08929.

Yves Grandvalet, Yoshua Bengio, et al. Learning from partial labels with minimum entropy. Technical report, CIRANO, 2004.

Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *neural information processing systems*, pp. 5639–5649, 2017.

R. Jin and Z. Ghahramani. Learning with multiple labels. *In Advances in Neural Information Processing Systems*, pp. 897–904, 2002.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in neural information processing systems*, pp. 921–928, 2003.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJ6oOfqge.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liping Liu and Thomas G Dietterich. A conditional multinomial mixture model for superset label learning. In *Advances in neural information processing systems*, pp. 548–556, 2012.

Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Advances in neural information processing systems*, pp. 1504–1512, 2010.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. *CoRR*, abs/2002.08053, 2020. URL https://arxiv.org/abs/2002.08053.

Nam Nguyen and Rich Caruana. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–559, 2008a.

Nam Nguyen and Rich Caruana. Improving classification with pairwise constraints: a margin-based approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 113–124. Springer, 2008b.

Ryan M. Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, 2004. URL http://jmlr.org/papers/v5/rifkin04a.html.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. In Peter Auer and Ron Meir (eds.), *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, volume 3559 of *Lecture Notes in Computer Science*, pp. 143–157. Springer, 2005. doi: 10.1007/11503415_10. URL https://doi.org/10.1007/11503415_10.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5557–5564, 2019.

Fei Yu and Min-Ling Zhang. Maximum margin partial label learning. In *Asian Conference on Machine Learning*, pp. 96–111, 2016.

Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 708–715, 2013.

Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, pp. 4048–4054, 2015.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, 2016.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5:1225–1251, 2004. URL http://jmlr.org/papers/volume5/zhang04b/zhang04b.pdf.