Multi-Complementary and Unlabeled Learning for Arbitrary Losses and Models

Yuzhou Cao¹ and Yitian Xu¹

¹College of Science, China Agriculture University

Abstract

A weakly-supervised learning framework named as complementary-label learning has been proposed recently, where each sample is equipped with a complementary label that denotes one of the classes the sample does not belong to. However, the previous complementary-label learning methods can not learn from samples with arbitrary number of complementary labels, which are more informative. The unlabeled samples are also neglected in the complementary-label learning setting. This paper gives the multi-complementary and unlabeled learning framework with two unbiased estimators of the classification risk for arbitrary losses and models. We first give an unbiased estimator of classification risk with multi-complementarily labeled samples and further incorporate unlabeled samples into the risk formulation. The estimation error bounds show that the proposed methods are in the optimal parametric convergence rate. Finally, the experiments on both linear and deep models show the effectiveness of our methods.

1 Introduction

In ordinary supervised classification problem, each training sample is equipped with an exact label that denotes the class the sample belongs to. However, the preparation of a massive of exactly labeled data is usually laborious and unrealistic in practical. Therefore, a lot of studies on learning from weak supervision have been made to tackle this problem in different scenarios, e.g. semi-supervised learning [1, 19, 27], partial label learning [4, 26], and positive-unlabeled learning [6, 5, 20, 9]. Recently, another weakly-supervised learning scenario called complementary-label learning (CLL) has been proposed. In the CLL setting, ordinary label is substituted with a complementary label, which denotes one of the classes that a training sample does not belong to. It is obvious that the preparation of complementarily labeled data is much more labor-saving than that of ordinarily labeled data.

The complementary-label learning problem has been investigated in previous studies [11, 25, 12]. In these works, different risk estimators were proposed to recover classification risk only with complementarily labeled data under the empirical risk minimization (ERM) framework. In [11] and [25], the proposed risk estimators had restrictions on loss functions and unbiasedness respectively. [12] overcame the shortcomings by giving an unbiased risk estimator without any restriction on models and loss functions while guaranteeing the superior performance in terms of classification accuracy over the previous two methods.
It is noticeable that in these works, each training sample was given only a single complementary label. However, in quite a few cases, the training samples may be multi-complementarily labeled, namely each training sample is equipped with multiple complementary labels. For example, in the stage of data annotation, an annotator who has no idea of a training sample’s exact label may be able to recognize multiple classes that the sample does not belong to, which results in a sample with multiple complementary labels. In crowdsourcing scenario [8], the quality of crowdsourcing labels is especially crucial [23]. Instead of being ordinarily labeled, a sample can be complementarily labeled to alleviate the effect of low-quality noisy crowdsourcing labels. Since a sample can be complementarily labeled by different crowdworkers, each training samples may have more than one complementary labels. Moreover, compared with the single-complementary-label setting in previous CLL studies, the samples with multiple complementary labels are more informative, e.g., extremely, in $K$-class classification problem, the combination of $K-1$ complementary labels can be seen as an exact class label while a single complementary label only means either of the $K-1$ ordinary labels. To sum up, a framework for learning from data with arbitrary number of complementary labels is in demand.

Furthermore, the information concealed in the easily accessible unlabeled data is proven to be helpful in many other areas of weakly-supervised learning both theoretically and practically [7, 29, 22]. Therefore, it is promising to further enhance the capability of CLL framework by utilizing unlabeled data.

In this paper, we study the multi-complementary label and unlabeled learning (MCUL) problem, where both multi-complementarily labeled data and unlabeled data are leveraged to obtain better classifiers. In our method, we proposed a novel unbiased risk estimator for MCUL problem with no limitation on loss functions and models. By using a mild assumption, we first derive the risk estimator for multi-complementary label learning (MCL) problem. The estimator in [12] is proven to be a special case in the MCL setting. Then we further utilize the unlabeled data to construct the risk estimator for MCUL problem. With no more assumption on loss functions and models, we show that the estimation error bounds of MCL and MCUL are in optimal parametric convergence rate [21]. The usefulness of proposed MCUL is demonstrated through experiments on benchmark and real world datasets.

The main contributions are summarized as follows:

- We propose the novel MCL framework that allows unbiased estimation to the classification risk only from samples with arbitrary number of complementary samples and can be applied on arbitrary losses and models.

- We further propose the MCUL framework to utilize the unlabeled samples and validate the benefits of the incorporation of unlabeled samples both experimentally and theoretically.

- The previous CLL framework and ordinary classification problem are proved to be special cases of the MCUL framework, which shows the comprehensiveness of the MCUL framework as a weakly-supervised leaning framework.

\footnote{1 ‘Arbitrary’ means the samples can be equipped with different numbers of complementary labels.}
2 Review of Complementary-Label Learning

To begin with, we first show the classification risk of learning from ordinary labels and then review how the previous risk estimators of learning from single-complementary labels recover the classification risk under the ERM framework.

2.1 Ordinary Classification Problem

Let’s denote the feature space with \( \mathcal{X} \subseteq \mathbb{R}^d \) and \( \mathcal{Y} = \{1, 2, \ldots, K\} \) is the label space. Suppose the training samples are drawn independently and identically from the unknown distribution \( D \), which is the joint distribution over \( \mathcal{X} \times \mathcal{Y} \) with density \( p(x, y) \). Then the critical work is to find a decision function \( g : \mathcal{X} \rightarrow \mathbb{R}^K \) that minimizes the classification risk with loss function \( \ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+ \):

\[
R(g) := \mathbb{E}_{p(x, y)}[\ell(g(x), y)].
\]

Since the density \( p(x, y) \) is unknown, the classification risk is approximated by the empirical risk:

\[
\hat{R}(g) := \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i).
\]

2.2 Complementary-Label Learning

In the CLL setting, each sample is attached with a complementary label. The complementarily labeled data \( \{(x_i, \overline{y}_i)\}_{i=1}^{n} \) are sampled independently and identically from a joint distribution with density \( p(x, \overline{y}) \).

In [11], an assumption on density \( p(x, \overline{y}) \) is made:

\[
p(x, \overline{y}) = \frac{1}{K - 1} \sum_{y \neq \overline{y}} p(x, y).
\]

Under this assumption, [11] proved that classification risk [1] can be estimated by an unbiased estimator only with complementarily labeled data. However, the loss functions are restricted to one-versus-all and pairwise comparison multi-class loss functions [28]. Moreover, the binary loss functions \( \ell'(z) : \mathbb{R} \rightarrow \mathbb{R}^+ \) used in the two multi-class loss functions are required to fulfill symmetric condition: \( \ell'(z) + \ell'(-z) = 1 \). Obviously, the popular softmax cross-entropy loss and all the other convex loss functions do not meet these conditions. Since the softmax cross-entropy loss is widely used in deep learning, this requirement will be a serious limitation for the application of state-of-the-art deep models.

To make deep models available, [25] proposed another risk estimator limited to softmax cross-entropy loss. Though the risk estimator is not necessarily unbiased, the method is ensured to identify the optimal classifier that minimizes classification risk [1] by minimizing its learning objective. The method also introduces bias into the choice of complementary labels. However, in the stages of bias estimation, ordinarily labeled data are required. The severe requirement might not align with the motivation of complementary-label learning.

The limitations above were removed in [12]. An unbiased risk estimator with only complementarily labeled data was deduced by a different approach than [11]. With the same assumption [3] adopted, the risk formulation is valid for arbitrary losses and models.
Experiments on both linear and deep models show the superiority of the estimator in [12] than those in previous works [11, 25]. Nevertheless, the estimator is still confined within single-complementary-label setting, where each sample is given merely one complementary label. The unlabeled data are also neglected in previous CLL studies, which prevents the CLL from being a more general framework.

The MCUL framework proposed in this paper further enables learning from both multi-complementarily labeled samples and unlabeled samples. Figure 1 describes the differences between previous CLL setting and the proposed MCUL setting.

3 Proposed Frameworks

In this section, we propose our framework to enable unbiased estimation of classification risk from both multi-complementarily labeled data and unlabeled data.

We first prove that the classification risk can be recovered from multi-complementarily labeled data under a mild assumption by employing the risk rewrite technique [16]. Then we further present the risk formulation of MCUL and show the estimation error bounds of the two methods.

Notations and Settings: Denote by $\overline{\mathcal{Y}} = \{1, \ldots, K\}$ the complementary label space. $\overline{\mathcal{Y}}_c$ is the collection of all the possible combinations of $c$ different complementary labels, e.g. $\overline{\mathcal{Y}}_c = \{1, \ldots, c\} \subset \overline{\mathcal{Y}}_c$. $\overline{\mathcal{Y}}_c$ is referred to as complementary-label set in the following sections. Suppose the training samples are sampled as follows:

$$
\mathcal{S}_u := \{ x_i^u \}_{i=1}^{n_u} \overset{i.i.d.}{\sim} p(x), \\
\mathcal{S}_c := \{ (x_i^c, \overline{y}_i^c) \}_{i=1}^{n_c} \overset{i.i.d.}{\sim} p_u(x, \overline{y}), \quad c = 1, \ldots, K-1.
$$

Figure 1: The demonstration of previous CCL setting and the proposed MCL&MCUL settings.
where \( p(x) \) is the marginal density and \( \overline{p}_c(x, \overline{Y}) \) is the density on \( X \times \overline{Y}_c \). \( S_c \) are the unlabeled data and \( S_c \) are the multi-complementarily labeled data with complementary-label sets of size \( c \). The size of complementary-label set \( \overline{Y} \) is denoted by \(|\overline{Y}|\) and \(|S_c| = n_c, \ c = 1, \ldots, K - 1.\)

### 3.1 Multi-Complementary Label Learning (MCL)

In this section, we give an account of the risk minimization framework of multi-complementary label learning.

As in the previous works, we first make assumptions on the relation between density \( \overline{p}_c(x, \overline{Y}) \) and \( p(x, y) \).

\[
\overline{p}_c(x, \overline{Y}) = \frac{1}{(K-1)c} \sum_{y \in \overline{Y}} p(x, y). \tag{4}
\]

The assumption implies each combination of \( c \) complementary labels are selected uniformly.

Under this assumption, we prove that the multi-complementary loss allows unbiased estimation of classification risk \( R \) from samples with complementary-label sets:

**Lemma 1.** Suppose the density \( \overline{p}_c(x, \overline{Y}) \) and \( p(x, y) \) follow the assumption (4). For any loss \( \ell \) and decision function \( g \), the classification risk \( R \) is equal to the risk formulation below:

\[
R_c(g) = \mathbb{E}_{\overline{p}_c(x, \overline{Y})}[\overline{\ell}(g(x), \overline{Y})]. \tag{5}
\]

where \( \overline{\ell} \) is the multi-complementary loss:

\[
\overline{\ell}(x, \overline{Y}) := \sum_y \ell(g(x), y) - \frac{K-1}{|\overline{Y}|} \sum_{y \in \overline{Y}} \ell(g(x), y). \tag{6}
\]

where \( \mathbb{E}_{\overline{p}_c(x, \overline{Y})}[\cdot] \) is the expectation on the distribution with density \( \overline{p}_c(x, \overline{Y}) \). The proof can be found in the appendix.

Though the risk formulation in Lemma 1 shows that the classification risk can be recovered only from samples with complementary-label sets of fixed size \( c \), the complementary-label sets of samples are not limited to a certain size. To completely remove the limitation on the size of complementary-label set, we consider the convex combination of \( R_c(g) \) called multi-complementary risk.

**Definition 1.** (Multi-Complementary Risk) For any decision function \( g \), its MCL risk is defined as:

\[
R_{MCL}(g) = \sum_{c=1}^{K-1} \alpha_c R_c(g), \tag{7}
\]

where \( \alpha \) is any vector in \( \{\alpha | \sum_{c=1}^{K-1} \alpha_c = 1, \alpha \geq 0\} \).

**Theorem 2.** The MCL risk is equal to classification risk \( R(g) \):

\[
R_{MCL}(g) = R(g). \tag{8}
\]
Proof. Due to Lemma 1, we can get $R_c(g) = R(g)$. Then the following equations holds:

$$ R_{\text{MCL}}(g) = \sum_{c=1}^{K-1} \alpha_c R_c(g) = \sum_{c=1}^{K-1} \alpha_c R(g) = R(g). $$

The empirical MCL risk is as below:

$$ \hat{R}_{\text{MCL}}(g) = \sum_{c=1}^{K-1} \frac{\alpha_c}{n_c} \sum_{i=1}^{n_c} \left( \sum_y \ell(g(x_i^c), y) - \frac{K-1}{c} \sum_{y \in \mathcal{Y}_i} \ell(g(x_i^c), y) \right). $$

Then the following work is to find the minimizer $\hat{g}_{\text{MCL}}$ of empirical MCL risk:

$$ \hat{g}_{\text{MCL}} = \arg \min_{g \in \mathcal{G}^K} \hat{R}_{\text{MCL}}(g). $$

where $\mathcal{G} = \{g(x)\}$ is the function class of each component of the $K$-dimension decision function $g = [g_1, \ldots, g_K]^T$.

In [9], all the samples are taken into consideration regardless of the size of complementary-label sets. Since there is no restriction on loss function $\ell$ and classifier $g$, any loss and model is available for the multi-complementary learning framework.

Remark 1. There are some special cases in the multi-complementary label learning setting. If $\alpha_1 = 1$, the proposed estimator reduces to the estimator in single-complementary-label setting [12]. If $\alpha_{K-1} = 1$, the proposed estimator is the same with that in ordinary classification problem [2]. According to the special cases, the proposed multi-complementary label learning proved to be a comprehensive framework in weakly-supervised learning.

3.2 Multi-Complementary and Unlabeled Learning (MCUL)

To utilize both multi-complementarily labeled data and unlabeled data, we further rewrite the risk formulation and propose the MCUL framework. Based on Lemma 1, we can incorporate the unlabeled data to construct an unbiased estimator of classification risk [11]:

Lemma 3. The classification risk [11] is equal to the risk formulation below:

$$ R_u(g) = \gamma \mathbb{E}_{p(x)} \left[ \sum_y \ell(g(x), y) \right] + \mathbb{E}_{p_n(x, Y)} \left[ (1 - \gamma) \sum_y \ell(g(x), y) - \frac{K-1}{c} \sum_{y \in \mathcal{Y}} \ell(g(x), y) \right] $$

where $\gamma \in [0, 1]$ is the trade-off coefficient.

Proof. The term $\sum_y \ell(g(x), y)$ is independent $\mathcal{Y}$, and thus:

$$ \mathbb{E}_{p(x)} \left[ \sum_y \ell(g(x), y) \right] = \mathbb{E}_{p_n(x, Y)} \left[ \sum_y \ell(g(x), y) \right] $$
According to the equation above and Lemma 1, we can obtain:

\[ R_u^c(g) = R_c(g) = R(g) \]

In the same manner as in the derivation of (7), we can derive the multi-complementary and unlabeled risk:

**Definition 2.** (Multi-Complementary and Unlabeled Risk) For any decision function \( g \), its MCUL risk is defined as:

\[
R_{MCUL}(g) = \sum_{c=1}^{K-1} \alpha_c R_u^c(g). \tag{12}
\]

where \( \alpha \) is any vector in \( \{ \alpha \mid \sum_{c=1}^{K-1} \alpha_c = 1, \alpha \geq 0 \} \).

When the trade-off coefficient \( \gamma \) is set to 0, the MCUL risk is the same with MCL risk (7). The following theorem allows unbiased estimation with both unlabeled data and multi-complementarily labeled data:

**Theorem 4.** The MCUL risk is equal to classification risk (7):

\[
R_{MCUL}(g) = R(g). \tag{13}
\]

The theorem can be proved in the same way as in Theorem 7. We can approximate the MCUL risk by the empirical MCUL risk below:

\[
\hat{R}_{MCUL}(g) = \frac{\gamma}{n_u} \sum_{y} \sum_{i=1}^{n_u} \ell(g(x_u^i), y) + \sum_{c=1}^{K-1} \frac{\alpha_c(1-\gamma)}{n_c} \sum_{y} \sum_{i=1}^{n_c} \ell(g(x_c^i), y) \]

\[
- \sum_{c=1}^{K-1} \frac{\alpha_c(K-1)}{c\cdot n_c} \sum_{y \in Y_c^i} \ell(g(x_c^i), y). \tag{14}
\]

Then the following work is to find the minimizer \( \hat{g}_{MCUL} \) of empirical MCUL risk:

\[
\hat{g}_{MCUL} = \arg \min_{g \in \mathcal{G}^K} \hat{R}_{MCUL}(g). \tag{15}
\]

Compared with the empirical MCL risk (9), the empirical MCUL risk (12) further incorporates the unlabeled data into the risk formulation. With the incorporation of easily accessible unlabeled data, the estimation error bound will be tighter, which indicates a better decision function \( g \). The claim is further validated in the following sections.
3.3 Estimation Error Bounds of MCL and MCUL

In this section, we give the estimation error bounds of the proposed MCL and MCUL frameworks.

Suppose the loss function $\ell$ does not exceed $C_\ell$ on feature space $\mathcal{X}$ and let $L_\ell$ be the Lipschitz constant of $\ell$. $\mathcal{R}_n(\mathcal{G})$ is the Rademacher complexity [18] of function class $\mathcal{G}$ with sample size of $n$ from $p(x)$ and we suppose that it decays in the rate of $O(1/\sqrt{n})$. We have the following estimation error bounds, which show the convergence of $\hat{g}_{MCL}$ and $\hat{g}_{MCUL}$ to the optimal decision function $g^* = \arg\min_{g \in \mathcal{G}} R(g)$:

**Theorem 5.** (*Estimation error bound of MCL*) For any $\delta > 0$, with probability at least $1 - \delta$:

$$R(\hat{g}_{MCL}) - R(g^*) \leq \sum_{c=1}^{K-1} 2K \left( \frac{K - 1}{c} + 1 \right) \alpha_c \left( 2KL_\ell \mathcal{R}_{n_c}(\mathcal{G}) + C_\ell \sqrt{\frac{\ln(2K/\delta)}{2n_c}} \right).$$

(16)

which indicates $R(\hat{g}_{MCL}) - R(g^*) = O(\sum_{c=1}^{K-1} 1/\sqrt{n_c})$.

**Theorem 6.** (*Estimation error bound of MCUL*) For any $\delta > 0$, with probability at least $1 - \delta$:

$$R(\hat{g}_{MCUL}) - R(g^*) \leq 2K \left[ \gamma \left( 2KL_\ell \mathcal{R}_{n_u}(\mathcal{G}) + C_\ell \sqrt{\frac{\ln(2K/\delta)}{2n_u}} \right) + \left( \frac{K - 1}{c} + 1 - \gamma \right) \sum_{c=1}^{K-1} \alpha_c \left( 2KL_\ell \mathcal{R}_{n_c}(\mathcal{G}) + C_\ell \sqrt{\frac{\ln(2K/\delta)}{2n_c}} \right) \right].$$

which indicates $R(\hat{g}_{MCUL}) - R(g^*) = O(1/\sqrt{n_u} + \sum_{c=1}^{K-1} 1/\sqrt{n_c})$.

The proof is given in the appendix.

**Remark 2.** From Theorem 5 and 6, we can learn the estimation error bounds of the proposed methods are in the optimal convergence rate without any additional assumption[21]. Moreover, with increasing number of unlabeled data, the error bound of MCUL will get tighter, which implies the helpfulness of utilizing unlabeled data.

4 Experiments

In this section, we conduct comparative experiments on several datasets including PENDIGITS, LETTER, SATIMAGE, USPS\footnote{The first five benchmark datasets are available on https://www.csie.ntu.edu.tw/~cjlin/libsvm/}, MNIST[13], Fashion-MNIST[24], Kuzushi-MNIST[2] and EMNIST-balanced[3]. We compare with three complementary-label learning baseline methods: Pairwise Comparison (PC) with sigmoid loss from [11], Forward correction (Fwd) from [25] and Gradient Ascent (GA) from [12].
The details of the datasets is shown in the following sections. The implementation is based on Pytorch.

4.1 Experimental Setup

In the experiments, MCL, MCUL, GA and Fwd was trained with softmax cross-entropy loss and PC was trained with pairwise-comparison loss. Adam [13] was applied to optimize the models. All the datasets were split into train/validation sets with a 9:1 ratio.

To ensure that the assumption (4) is satisfied, each complementary-label set of size $c$ was generated by randomly choosing $c$ labels from the candidates labels other than the true label. Though we can simply split a sample with $c$ complementary labels into $c$ samples with one complementary label each, it’s obvious that the $c$ samples are not independent from each other. Therefore this approach will leads to serious violation of the fundamental i.i.d. assumption. For fair comparison in these experiments, the complementary labels were generated in the same way as in [12].

For PENDIGITS, LETTER, SATIMAGE, USPS and MNIST, a linear-in-input model with a bias term was used. For MNIST, the learning rate was fixed to 1e-4; weight decay 1e-4; maximum iterations 60000; and batch size was set to 100. For the rest datasets, the learning rate was selected from {1e-1, 1e-2, 1e-3, 1e-4} and maximum iterations was changed to 5000.

For Fashion-MNIST, Kuzushi-MNIST and EMNIST-balanced, a MLP model($d$-500-$K$) was trained for 300 epochs. The learning rate and weight decay was fixed to 1e-4 and the batch size was 256. In the experiments for flexible models, both MCL and MCUL were optimized with the same strategy as GA in [12].

Experiments on datasets with or without unlabeled samples were both conducted. When the unlabeled samples were incorporated, we randomly set 99% of training samples to be unlabeled for datasets with less than 50000 samples. The fraction was further increased to 99.5% for datasets with more than 50000 samples.

In respect of model parameter setting, the setting of baseline methods follow the previous work [12]. For MCUL, we used $\gamma = 0.1$. The parameter $\alpha$ was set to fulfill the equations below:

$$\begin{align*}
\alpha_i : \alpha_j &= \frac{n_i}{(K-i)^2} : \frac{n_j}{(K-j)^2}, \\
\sum_{i=1}^{K-1} \alpha_1 &= 1.
\end{align*}$$

In our methods, we made no assumption on distribution of the size of complementary-label sets. To generate the multi-complementarily labeled samples as close to the reality, we supposed that samples with too few or too many complementary labels appear in a small possibility. Then the $n_i$ follows the formulation below:

$$n_i : n_j = e^{-(i-\mu)^2} : e^{-(j-\mu)^2}.$$ 

In this paper, $\mu = \frac{K}{2}$ is used.

4.2 Experiments for Linear Model and Flexible Models

The experiment results for linear and flexible models are summarized in Table 1 and Table 2 respectively. The experiment results under the presence of unlabeled samples are shown in the second row corresponding to each dataset.

https://pytorch.org
Table 1: Test mean and standard deviation of the classification accuracy of linear model for 10 trials. The best one is emphasized in bold. \#n, \#f and \#c denote the number of samples, features and classes of each dataset.

| Datasets    | \#n | \#f | \#c | PC  | Fwd | GA  | MCL | MCUL |
|-------------|-----|-----|-----|-----|-----|-----|-----|------|
| PENDIGITS   | 10092 | 16 | 10  | 62.98±5.09 | 77.91±2.83 | 84.32±1.09 | 84.32±1.09 | 84.32±1.09 |
| LETTER      | 20000 | 16 | 26  | 9.17±2.44 | 9.37±2.52 | 11.05±5.05 | 11.44±5.68 | 11.44±5.68 |
| SATIMAGE    | 6435  | 36 | 6   | 74.05±4.65 | 77.45±5.43 | 82.10±1.82 | 82.10±1.82 | 82.10±1.82 |
| USPS        | 9298  | 256| 10  | 41.75±5.45 | 46.15±13.10 | 82.31±2.32 | 82.31±2.32 | 82.31±2.32 |
| MNIST       | 70000 | 784| 10  | 51.21±4.87 | 52.17±4.88 | 77.36±1.27 | 77.36±1.27 | 77.36±1.27 |

Table 2: Test mean and standard deviation of the classification accuracy of flexible model for 4 trials. The best one is emphasized in bold. \#n, \#f and \#c denote the number of samples, features and classes of each dataset.

| Datasets       | \#n | \#f | \#c | PC  | Fwd | GA  | MCL | MCUL |
|----------------|-----|-----|-----|-----|-----|-----|-----|------|
| Fashion-MNIST  | 70000 | 784| 10  | 77.34±0.88 | 83.49±0.18 | 84.43±0.22 | 84.43±0.22 | 84.43±0.22 |
| Kurushi-MNIST  | 70000 | 784| 10  | 59.31±1.07 | 66.46±0.17 | 78.27±0.41 | 78.27±0.41 | 78.27±0.41 |
| EMNIST-balanced| 131600 | 784| 47  | 14.28±1.18 | 18.21±2.93 | 66.98±0.32 | 66.98±0.32 | 66.98±0.32 |

Results of MCL: First we compare the proposed MCL framework with the three baseline methods. From the experiment results, we can see that MCL framework outperforms the baseline methods on all the datasets regardless of which model was applied. The superiority of MCL is especially apparent when the datasets have a large number of classes. Due to the experiment results on LETTER and EMNIST-balanced, it can be seen that the baseline methods can hardly generate effective classifiers. Compared with the baseline methods, MCL remains valid owing to the capability of utilizing multi-complementarily labeled samples.

Results of MCUL: From the tables, we can obtain that under the presence of a great percentage of unlabeled samples, the baseline methods suffered from the lack of complementarily labeled samples, while MCUL can still enhance its performance by incorporating unlabeled samples. Moreover, in the case that only a small number of complementarily labeled samples were available, the performance of baseline methods GA was seriously degraded due to the imprecise estimation of class prior, which is one of the reason that the performance of GA is relatively weak with a small fraction of complementarily labeled samples.

5 Conclusion
We first derived the MCL framework to learn from samples with any number of complementary labels for arbitrary losses and models. Then we incorporated unlabeled data into the risk formulation and proposed the MCUL framework to learn from multi-complementarily labeled data and unlabeled data simultaneously. We further showed that the estimation error bounds of the proposed methods are in the optimal parametric convergence rate. Finally,
we conducted experiments and showed our methods outperform the current state-of-the-art methods for both linear and deep model. A promising direction is applying our methods on crowdsourcing and other weakly-supervised classification scenarios, e.g. active learning\cite{10}, which is our future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 11671010) and Beijing Natural Science Foundation (No. 4172035).
References

[1] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

[2] Tarin Clauuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018.

[3] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: extending MNIST to handwritten letters. In *IJCNN 2017*, pages 2921–2926, 2017.

[4] Timothée Cour, Benjamin Sapp, and Ben Taskar. Learning from partial labels. *J. Mach. Learn. Res.*, 12:1501–1536, 2011.

[5] Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML 2013*.

[6] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *KDD 2008*.

[7] Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya O. Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *COLT 2019*.

[8] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, 2009.

[9] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *ICML 2019*.

[10] Sheng-Jun Huang, Nengneng Gao, and Songcan Chen. Multi-instance multi-label active learning. In *IJCAI 2017*, pages 1886–1892, 2017.

[11] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *NIPS 2017*.

[12] Takashi Ishida, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *ICML 2019*.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[15] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[16] Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *ICLR 2019*.

[17] Colin McDiarmid. *On the method of bounded differences*. 1989.
[18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. Adaptive computation and machine learning. MIT Press, 2012.

[19] Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In ICML 2017.

[20] Emanuele Sansone, Francesco G. B. De Natale, and Zhi-Hua Zhou. Efficient training for positive unlabeled learning. IEEE Trans. Pattern Anal. Mach. Intell., 41(11):2584–2598, 2019.

[21] Vladimir Vapnik. An overview of statistical learning theory. IEEE Trans. Neural Networks, 10(5):988–999, 1999.

[22] Qian-Wei Wang, Yu-Feng Li, and Zhi-Hua Zhou. Partial label learning with unlabeled data. In IJCAI 2019.

[23] Wei Wang and Zhi-Hua Zhou. Crowdsourcing label quality: a theoretical analysis. SCIENCE CHINA Information Sciences, 58(11):1–12, 2015.

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR, abs/1708.07747, 2017.

[25] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In ECCV 2018.

[26] Min-Ling Zhang, Fei Yu, and Cai-Zhi Tang. Disambiguation-free partial label learning. IEEE Trans. Knowl. Data Eng., 29(10):2155–2167, 2017.

[27] Teng Zhang and Zhi-Hua Zhou. Semi-supervised optimal margin distribution machines. In IJCAI 2018.

[28] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. J. Mach. Learn. Res., 5:1225–1251, 2004.

[29] Zhen-yu Zhang, Peng Zhao, Yuan Jiang, and Zhi-Hua Zhou. Learning from incomplete and inaccurate supervision. In KDD 2019.
Appendix

A. Proof of Lemma 1

Proof. Suppose \( \bar{Y} \in Y_c \), then we can obtain \( |\bar{Y}| = c \). Denote \( \{ \bar{Y} : \bar{Y} \in Y_c, y \notin \bar{Y} \} \) with \( Y^c_c \).

Due to assumption (4), the equations below hold:

\[
\begin{align*}
\sum_{\bar{Y} \in Y^c_c} p_c(x, \bar{Y}) &= p(x, y) + \frac{K - c - 1}{c} \sum_{y \neq y} p(x, \hat{y}), \\
\sum_{\bar{Y} \notin Y^c_c} p_c(x, \bar{Y}) &= \frac{c}{K - 1} \sum_{y \neq y} p(x, \hat{y}).
\end{align*}
\]

We can get the following equation by substituting the second equation above into the first equation:

\[
p(x, y) = \sum_{\bar{Y} \in Y^c_c} p_c(x, \bar{Y}) - \frac{K - c - 1}{c} \sum_{\bar{Y} \in Y^c_c} p_c(x, \bar{Y}) = \sum_{\bar{Y} \in Y^c_c} p_c(x, \bar{Y}) - \frac{K - 1}{c} \sum_{\bar{Y} \notin Y^c_c} p_c(x, \bar{Y}).
\]

Then we can rewrite the classification risk (1) due to the equation above:

\[
R(g) = \sum_{y} \int \ell(g(x), y) p(x, y) dx = \sum_{y} \int \ell(g(x), y) \sum_{\bar{Y} \in Y^c_c} p_c(x, \bar{Y}) dx - \frac{K - 1}{c} \sum_{y} \int \ell(g(x), y) \sum_{\bar{Y} \notin Y^c_c} p_c(x, \bar{Y}) dx.
\]

By exchanging the order of summation, we can get an equal version of the last equation above and finally conclude the proof:

\[
R(g) = \sum_{\bar{Y} \in Y^c_c} \int \sum_{y} \ell(g(x), y) p_c(x, \bar{Y}) dx - \frac{K - 1}{c} \sum_{\bar{Y} \in Y^c_c} \int \sum_{y} \ell(g(x), y) p_c(x, \bar{Y}) dx
\]

\[
= \sum_{\bar{Y} \in Y^c_c} \int \left( \sum_{y} \ell(g(x), y) - \frac{K - 1}{|\bar{Y}|} \sum_{y \in \bar{Y}} \ell(g(x), y) \right) p_c(x, \bar{Y}) dx
\]

\[
= \mathbb{E}_{p(x, \bar{Y})} [\ell(x, \bar{Y})] = R_c(g).
\]

\( \square \)
B. Proof of Theorem 6

The proof of Theorem 6 is omitted since it is the special case of Theorem 5 by setting \( \gamma \) to 0.

First we introduce the Talagrand’s contraction lemma 15:

**Lemma 7.** Let \( \mathcal{G} \) be a class of real functions and \( \mathcal{G}^K = [\mathcal{G}]_{i=1}^K \) be a K-dimensional function class and \( \ell : \mathbb{R}^K \to \mathbb{R} \) a Lipschitz function with constant \( L_\ell \) and \( \ell(0) = 0 \). Then \( \mathcal{R}_n(\ell \circ \mathcal{G}^K) \leq KL_\ell \mathcal{R}_n(\mathcal{G}). \)

To apply Talagrand’s contraction lemma, we use the shifted loss \( \tilde{\ell}(z) = \ell(z) - \ell(0) \) instead of \( \ell(z) \). Then we abbreviate some complex terms in these forms:

\[
\begin{align*}
A &= \gamma E_{p(x)}[\sum_y \tilde{\ell}(g(x), y)] \\
B_c &= E_{p_c(x, \mathcal{Y})}[(1 - \gamma) \sum_y \tilde{\ell}(g(x), y) - \frac{K-1}{c} \sum_{y \in \mathcal{Y}} \tilde{\ell}(g(x), y)] \\
\tilde{\ell}_A(g(x)) &= \gamma \sum_y \tilde{\ell}(g(x), y) \\
\tilde{\ell}_B(g(x), \mathcal{Y}) &= (1 - \gamma) \sum_y \tilde{\ell}(g(x), y) - \frac{K-1}{c} \sum_{y \in \mathcal{Y}} \tilde{\ell}(g(x), y).
\end{align*}
\]

We give the following conclusions:

**Lemma 8.** Denote \( p_c(x, \mathcal{Y}) \) with \( p_c \):

\[
\begin{align*}
\mathcal{R}_n(\tilde{\ell}_A \circ \mathcal{G}^K) &\leq \gamma K^2 L_\ell \mathcal{R}_n(\mathcal{G}), \\
\mathcal{R}_n, p_c(\tilde{\ell}_B^c \circ \mathcal{G}^K) &\leq \left( \frac{K-1}{c} + 1 - \gamma \right) K^2 L_\ell \mathcal{R}_n(\mathcal{G}).
\end{align*}
\]

**Proof.** The first inequality can be deduced from Lemma 7 directly. By definition and the sub-additivity of supremum:

\[
\begin{align*}
\mathcal{R}_n, p_c(\tilde{\ell}_B^c \circ \mathcal{G}^K) &= E_{\mathcal{S}, \mathcal{E}_\sigma} \left[ \sup_{g \in \mathcal{G}^K} \frac{1}{n} \sum_{(x_i, \mathcal{Y}_i) \in \mathcal{S}_c} \sigma_i \tilde{\ell}_B(g(x_i), \mathcal{Y}_i) \right] \\
&\leq (1 - \gamma) E_{\mathcal{S}, \mathcal{E}_\sigma} \left[ \sup_{g \in \mathcal{G}^K} \frac{1}{n} \sum_{(x_i, \mathcal{Y}_i) \in \mathcal{S}_c} \sigma_i \sum_y \tilde{\ell}(g(x_i), y) \right] \\
&\quad + \frac{K-1}{c} E_{\mathcal{S}, \mathcal{E}_\sigma} \left[ \sup_{g \in \mathcal{G}^K} \frac{1}{n} \sum_{(x_i, \mathcal{Y}_i) \in \mathcal{S}_c} \sigma_i \sum_y \tilde{\ell}(g(x_i), y) \right]
\end{align*}
\]

The \( \tilde{\ell}_B \) is a fixed loss function and the first equation holds. Since \( \Sigma_y \tilde{\ell} \) is independent of \( \mathcal{Y}_i \), we can get:

\[
E_{\mathcal{S}, \mathcal{E}_\sigma} \left[ \sup_{g \in \mathcal{G}^K} \frac{1}{n} \sum_{(x_i, \mathcal{Y}_i) \in \mathcal{S}_c} \sigma_i \sum_y \tilde{\ell}(g(x_i), y) \right] = E_{\mathcal{X}, \mathcal{E}_\sigma} \left[ \sup_{g \in \mathcal{G}^K} \frac{1}{n} \sum_{x_i \in \mathcal{X}} \sigma_i \sum_y \tilde{\ell}(g(x_i), y) \right]
\]
Let \( I(\cdot) \) be the indicator function and \( \alpha_i = 2I(y \in Y_i) - 1 \). Then we have the conclusion below:

\[
\begin{align*}
\mathbb{E}_{S_c} \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{n} \sum_{(x_i, Y_i) \in S_c} \sigma_i \sum_{y \in Y_i} \tilde{\ell}(g(x_i), y) \right] &= \mathbb{E}_{S_c} \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{2n} \sum_{(x_i, Y_i) \in S_c} \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y)(\alpha_i + 1) \right] \\
&\leq \mathbb{E}_{S_c} \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{2n} \sum_{(x_i, Y_i) \in S_c} \alpha_i \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y) \right] \\
&+ \mathbb{E}_{S_c} \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{2n} \sum_{(x_i, Y_i) \in S_c} \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y) \right] \\
&= \mathbb{E}_{X} \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{n} \sum_{x_i \in X} \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y) \right]
\end{align*}
\]

Then the inequalities hold:

\[
\mathcal{R}_{n, p_c}(\tilde{\ell}_B \circ G^K) \leq (1 - \gamma) \mathbb{E}_X \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{n} \sum_{x_i \in X} \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y) \right] \\
+ \frac{(K - 1)}{c} \mathbb{E}_X \mathbb{E}_\sigma \left[ \sup_{g \in G^K} \frac{1}{n} \sum_{x_i \in X} \sigma_i \sum_{y} \tilde{\ell}(g(x_i), y) \right] \\
= \left( \frac{(K - 1)}{c} + \gamma \right) \mathcal{R}_n(\ell_A \circ G^K)
\]

\[
\leq \left( \frac{K - 1}{c} + 1 - \gamma \right) K^2 L_{\ell} \mathcal{R}_n(\mathcal{G})
\]

We can bound \( \sup_{g \in G^K} |A - \hat{A}| \) and \( \sup_{g \in G^K} |B_c - \hat{B}_c| \) using Mcdiarmid’s inequality [17]:

**Lemma 9.** For a certain \( c \), the inequalities below hold with probability at least \( 1 - \delta \):

\[
\sup_{g \in G^K} |A - \hat{A}| \leq \gamma K \left( 2KL_{\ell} \mathcal{R}_{n_c}(\mathcal{G}) + C_{\ell} \sqrt{\ln \frac{2/\delta}{2n_c}} \right)
\]

\[
\sup_{g \in G^K} |B_c - \hat{B}_c| \leq \left( \frac{K - 1}{c} + 1 - \gamma \right) K \left( 2KL_{\ell} \mathcal{R}_{n_c}(\mathcal{G}) + C_{\ell} \sqrt{\ln \frac{2/\delta}{2n_c}} \right).
\]

**Proof.** We are going to prove the first inequality and the second can be proved in a similar way. Firstly, we consider the single direction \( \sup_{g \in G^K} (A - \hat{A}) \). The \( \ell_A \) will not exceed \( \gamma KC_{\ell} \) due to the definition, then the change of \( \sup_{g \in G^K} (A - \hat{A}) \) will not exceed \( \gamma KC_{\ell}/n \) when we
replace a single $x_i$ with $x'_i$. Due to the McDiarmid’s inequality, the inequality below holds with probability at least $1 - \frac{\delta}{2}$:

$$\sup_{g \in \mathcal{G}^K} (A - \hat{A}) \leq E \left[ \sup_{g \in \mathcal{G}^K} (A - \hat{A}) \right] + \gamma K C \ell \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

Due to the symmetrization inequality [18], we can obtain that:

$$E \left[ \sup_{g \in \mathcal{G}^K} (A - \hat{A}) \right] \leq 2 R_n(\ell \circ \mathcal{G}^K) \leq 2 \gamma K^2 L \ell R_n(\mathcal{G})$$

The other direction is similar.

Now we can prove the Theorem [9].

**Proof.** Notice that $\hat{R}_{\text{MCUL}}(g_{\text{MCUL}}) \leq \hat{R}_{\text{MCUL}}(g^*)$. Due to Lemma [8] we can get:

$$R(\hat{g}_{\text{MCUL}}) - R(g^*) = R_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) - R_{\text{MCUL}}(g^*)$$

$$= R_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) - \hat{R}_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) + \hat{R}_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) - R_{\text{MCUL}}(g^*)$$

$$\leq R_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) - \hat{R}_{\text{MCUL}}(\hat{g}_{\text{MCUL}}) + \hat{R}_{\text{MCUL}}(g^*) - R_{\text{MCUL}}(g^*)$$

$$\leq 2 \sup_{g \in \mathcal{G}^K} \left| R_{\text{MCUL}}(g) - \hat{R}_{\text{MCUL}}(g) \right|.$$

According to the sub-additivity of supremum, we can get the inequality below:

$$\sup_{g \in \mathcal{G}^K} \left| R_{\text{MCUL}}(g) - \hat{R}_{\text{MCUL}}(g) \right| \leq \sum_{c=1}^{K-1} \alpha_c \sup_{g \in \mathcal{G}^K} \left| R_{\text{MCUL}}^u(g) - \hat{R}_{\text{MCUL}}^u(g) \right|$$

$$\leq \sum_{c=1}^{K-1} \alpha_c \left( \sup_{g \in \mathcal{G}^K} \left| A - \hat{A} \right| + \sup_{g \in \mathcal{G}^K} \left| B_c - \hat{B}_c \right| \right)$$

$$= \sup_{g \in \mathcal{G}^K} \left| A - \hat{A} \right| + \sum_{c=1}^{K-1} \alpha_c \sup_{g \in \mathcal{G}^K} \left| B_c - \hat{B}_c \right|$$

Due to the union bound and Lemma [9] the inequality below holds with probability at
least 1 − δ:

\[
R(\hat{g}_{MCUL}) - R(g^*) \leq 2 \sup_{g \in \mathcal{G}} \left| R_{MCUL}(g) - \hat{R}_{MCUL}(g) \right| \\
\leq 2 \left( \sup_{g \in \mathcal{G}} \left| A - \hat{A} \right| + \sum_{c=1}^{K-1} \alpha_c \sup_{g \in \mathcal{G}} \left| B_c - \hat{B}_c \right| \right) \\
\leq 2K \left[ \gamma \left( 2KL\mathcal{R}_{n_c}(G) + C_\ell \sqrt{\frac{\ln(2K/\delta)}{2n_u}} \right) \\
+ \left( \frac{K-1}{c} + 1 - \gamma \right) \sum_{c=1}^{K-1} \alpha_c \left( 2KL\mathcal{R}_{n_c}(G) + C_\ell \sqrt{\frac{\ln(2K/\delta)}{2n_c}} \right) \right].
\]

which concludes the proof. \qed