Bridging Ordinary-Label Learning and Complementary-Label Learning

Yasuhiro Katsura   KATSURA_ANI@RURI.WASEDA.JP
Masato Uchida   M.UCHIDA@WASEDA.JP
Waseda University, Tokyo, Japan

Abstract

A supervised learning framework has been proposed for the situation where each training data is provided with a *complementary label* that represents a class to which the pattern does not belong. In the existing literature, *complementary-label learning* has been studied independently from *ordinary-label learning*, which assumes that each training data is provided with a label representing the class to which the pattern belongs. However, providing a complementary label should be treated as equivalent to providing the rest of all the labels as the candidates of the one true class. In this paper, we focus on the fact that the loss functions for one-versus-all and pairwise classification corresponding to ordinary-label learning and complementary-label learning satisfy certain *additivity* and *duality*, and provide a framework which directly bridge those existing supervised learning frameworks. Further, we derive classification risk and error bound for any loss functions which satisfy additivity and duality.

Keywords: statistical learning theory, supervised classification

1. Introduction

In the most common problem setting in supervised learning, the class to which the training data belongs is provided as a label, namely an *ordinary label*. There is increasing interest in generalizing the concept of ordinary labels (Chapelle et al., 2010; Luo and Orabona, 2010; Cour et al., 2011; Natarajan et al., 2013; Ishida et al., 2018). In this paper, we consider a problem setting where each training data is provided with a label, that is, a *complementary label*, which specifies *one* class that the pattern does not belong to. The ordinary and complementary labels respectively reflect the information regarding a class to which the training data is likely and unlikely to belong.

The learning framework was first proposed by (Ishida et al., 2017). Assuming that a single complementary label is provided to each training data, the corresponding loss function considering one-versus-all and pairwise classification was defined, and also the classification risk and error were analyzed. The classification risk studied by (Ishida et al., 2019) depends on a a more generalized loss function where the classification strategy is not necessarily restricted to be one-versus-all or pairwise classification.

Although choosing the correct class of a pattern involves considerable work for the annotators, the use of complementary labels can alleviate the annotation cost. However, in this framework, loss of generalities can occur if annotators choose only *one* class as a

complementary label. In other words, it is reasonably assumed that multiple labels can be chosen as classes to which a pattern does not belong.

A framework where multiple complementary labels are provided to each training data was analyzed by (Cao and Xu, 2020; Feng et al., 2020). The number of complementary labels was formulated as a constant value by (Cao and Xu, 2020), while it was formulated as a random variable by (Feng et al., 2020). In these works, the existing framework in (Ishida et al., 2019) where single complementary label is provided to each training data is directly applied. Therefore, there is no essential difference with the framework of (Ishida et al., 2019) on the structure and properties of the loss functions or the derivation of classification risk and error.

In contrast, this paper focuses on equivalency of ordinary label and complementary label and provides essential generalization of learning framework from single ordinary/complementary label. The finding that loss functions for one-versus-all and pairwise classification satisfy certain *additivity* and *duality* plays a significant role in the generalization, and the classification risk and error bound with a new expression not found in (Ishida et al., 2019; Cao and Xu, 2020; Feng et al., 2020) are derived. This enables us to bridge ordinary-label learning and complementary-label learning and to understand them from a unified perspective. To be more specific, the introduced loss functions satisfying *additivity* and *duality* allow a straightforward comparison of the proposed approach and those shown in the existing literature. The properties also allow our classification risk to be in a more simple form than that of (Ishida et al., 2017, 2019; Cao and Xu, 2020; Feng et al., 2020). Further, the derived classification error monotonically decreases corresponding to the number of complementary labels. This property is also shown in the experiment, which supports that our framework bridges learning from ordinary/complementary labels.

The remaining paper is organized as follows. Section 2 provides a review of the existing work regarding the generalization of supervised learning from ordinary labels. Section 3 introduces several key formulations pertaining to ordinary-label and complementary-label learning. Section 4 presents the generation probability model of data provided with multiple complementary labels and the loss functions for the learning conducted from such data. Furthermore, this section describes the natural generalization of the proposed loss function from those for one-versus-all and pairwise classification defined by (Ishida et al., 2017) and the evaluation of the classification risk. The error bound of the one-versus-all and pairwise classification is derived in Section 5. Section 6 describes the experimental investigation to validate the classification error, and Section 7 presents the conclusions.

2. Related Work

This section provides a review of the existing work regarding the generalization of supervised learning from the ordinary labels provided to training data.

Semi-supervised learning, in which the classification algorithm is provided with some training data labeled but not necessarily for all, has been investigated (Grandvalet and Bengio, 2005; Mann and McCallum, 2007; Chapelle et al., 2010; Niu et al., 2013; Kipf and Welling, 2017; Laine and Aila, 2017). Although the prediction accuracy of such learning is less than that of fully supervised learning, this technique requires less labeled training data, thereby reducing the annotation cost.

As another type of generalization, there exists a method of learning from training data, in which each data is provided with multiple labels, only one of which specifies the *one* true class. Such labels are generally known as *partial labels* (or *candidate labels*) (Raykar et al., 2010; Luo and Orabona, 2010; Cour et al., 2011; Cid-sueiro, 2012). This approach can potentially be applied when the true class of a given training data is not clear or when the label information requires to be kept private.

Framework of learning from labels, namely *complementary labels*, specifying classes that the pattern does not belong to has been also formulated. Providing a complementary label to a training data is equivalent to considering the other labels as candidate labels. In other words, a set of candidate labels is a comprehensive concept involving both ordinary and complementary labels as special cases. The complementarily labeled data-generation probability model and classification risk were provided by (Ishida et al., 2017). In addition, when learning from *noisy labels*, which stochastically represent incorrect information, providing complementary label to prevent the specification of a noisy label as a true label has been noted to be effective (Kim et al., 2019).

However, the problem setting in (Ishida et al., 2017) encounters a limitation in cases in which only one complementary label is provided to each training data, leading to a loss of generality. To overcome the limitation, a problem setting where each training data is provided to multiple complementary labels was considered in (Cao and Xu, 2020; Feng et al., 2020). The existing framework in (Ishida et al., 2019) is directly applied in these works, and thus new insights into the relationship between ordinary-label learning and complementary-label learning or the properties satisfied by the loss functions are not obtained.

In contrast, we addressed the problem by using an approach different from that of (Cao and Xu, 2020; Feng et al., 2020). Focusing on equivalency of ordinary label and complementary label, we found that additivity and duality satisfied by loss functions plays a significant role in bridging learning from single ordinary/complementary labels; we then derived classification risk and error bound with a new expression not found in the existing literature. The structure and properties inherent in the loss functions enable us to naturally relate and discuss learning from ordinary and complementary labels in a unified manner.

3. Background

This section presents the data-generation probability model and loss functions shown in the existing literature. All these formulations were on the assumption that a single ordinary label or complementary label is provided.

3.1. Supervised Learning from Ordinary Label

In the K (≥ 2) classification problem, supervised learning concerns the efficiency of a classifier $f : \mathcal{X} \to \mathcal{Y}$, which maps a pattern $x \in \mathcal{X}$ to the class to which it belongs, that is, $y \in \mathcal{Y}$, where $|\mathcal{Y}| = K$. Given the discriminant function $g_y : \mathcal{X} \to \mathbb{R}$, which represents the confidence on the 2 class classification about y, the classifier f can be defined as $f(x) = \arg\max_{y \in \mathcal{Y}} g_y(x)$.

Given a pair (x, y), in which x is the pattern and y is an ordinary label representing the belonging class of x, the prediction by f can be evaluated using a loss function $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to$

$\mathbb{R}$. For example, the loss functions for *one-versus-all* and *pairwise classification*, $\mathcal{L}_{\text{OVA}}$ and $\mathcal{L}_{\text{PC}}$, respectively, can be defined as (Zhang, 2004):

$$\mathcal{L}_{\text{OVA}}\left(f(x), y\right) = \ell\left(g_y(x)\right) + \frac{1}{K-1}\sum_{y' \neq y}\ell\left(-g_{y'}(x)\right) \tag{1}$$

$$\mathcal{L}_{\text{PC}}\left(f(x), y\right) = \sum_{y' \neq y}\ell\left(g_y(x) - g_{y'}(x)\right) \tag{2}$$

Note that the function $\ell : \mathbb{R} \to \mathbb{R}_+$ monotonically decreases corresponding to the input.

The 0-1 loss, $\ell_{0\text{-}1}(z) := I(z \leq 0)$, is a standard type of function ℓ, where I is an indicator function. The 0-1 loss is unsuitable for loss optimization, as it is undifferentiable at $z = 0$, and its gradient is always 0 for all other inputs. Consequently, the 0-1 loss is usually surrogated by other functions, such as the sigmoid and ramp losses, all of which satisfy the following condition:

$$\ell\left(z\right) + \ell\left(-z\right) = a \tag{3}$$

where $a \in \mathbb{R}$ is constant. In the remaining study, we assume ℓ to satisfy (3) but not its differentiability.

In addition, we assume that (x, y), involving the pattern x and its belonging class y, is generated from a probability distribution $P(x, y)$. The classification risk $R(f)$ of a classifier f can be defined as follows:

$$R(f) = \mathbb{E}_{P(x,y)}\left[\mathcal{L}\left(f(x), y\right)\right] \tag{4}$$

where $\mathbb{E}_{P(x,y)}$ represents the expectation with respect to $P(x, y)$.

3.2. Supervised Learning from Complementary Label

The formulation considering a label specifying a class that a pattern does not belong to was discussed in (Ishida et al., 2017). Given a pair $(x, \overline{y})$ of a pattern $x \in \mathcal{X}$ and a class $\overline{y} \in \mathcal{Y}$ to which the pattern does not belong, the prediction by $f : \mathcal{X} \to \mathcal{Y}$ can be evaluated by the loss function $\overline{\mathcal{L}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The loss functions for one-versus-all and pairwise classification were defined in (Ishida et al., 2017) as follows:

$$\overline{\mathcal{L}}_{\text{OVA}}\left(f(x), \overline{y}\right) = \frac{1}{K-1}\sum_{y \neq \overline{y}}\ell\left(g_y(x)\right) + \ell\left(-g_{\overline{y}}(x)\right) \tag{5}$$

$$\overline{\mathcal{L}}_{\text{PC}}\left(f(x), \overline{y}\right) = \sum_{y \neq \overline{y}}\ell\left(g_y(x) - g_{\overline{y}}(x)\right) \tag{6}$$

Compared to (1) and (2), the loss functions defined in (5) and (6) are in natural forms to deal with the training data provided a label $\overline{y}$. The extended form of the loss functions in (5) and (6) is as follows (Ishida et al., 2019):

$$\overline{\mathcal{L}}\left(f(x), \overline{y}\right) = -(K-1)\mathcal{L}\left(f(x), \overline{y}\right) + \sum_{y \in \mathcal{Y}}\mathcal{L}\left(f(x), y\right) \tag{7}$$

Considering (3), the function in (7) involves both one-versus-all and pairwise classification losses.

Given a pattern x and its complementary label $\overline{y} \in \mathcal{Y}$, the generation probability $\overline{P}(\overline{y}|x)$ of the labeled data is as follows:

$$\overline{P}(\overline{y}|x) = \frac{1}{K-1} \sum_{y \neq \overline{y}} P(y|x) \tag{8}$$

Here, $\overline{P}(\overline{y}|x)$ is proportional to the sum of probabilities regarding all the other labels.

Subsequently, $(x, \overline{y})$ is assumed to be stochastically generated from the distribution $\overline{P}(x, \overline{y}) = \overline{P}(\overline{y}|x) P(x)$. The $R(f)$ defined in (4) can be expressed as:

$$R(f) = (K-1)\mathbb{E}_{\overline{P}(x,\overline{y})}\left[\overline{\mathcal{L}}(f(x), \overline{y})\right] - \overline{m}_1 + m_2 \tag{9}$$

such that $\overline{m}_1$ and m_2 are constants, where $\overline{m}_1 := \sum_{y \in \mathcal{Y}} \overline{\mathcal{L}}(f(x), y)$ and $m_2 := \mathcal{L}(f(x), y) + \overline{\mathcal{L}}(f(x), y)$.

4. Formulation

The formulation in (Ishida et al., 2017) assumes that a single complementary label is provided to the training data. We generalize this formulation, assuming that M $(1 \leq M \leq K-1)$ complementary labels are provided to each training data. In this section, we generalize loss functions for one-versus-all and pairwise classification focusing on the fact that those functions satisfy *additivity* and *duality*. Further, we derive classification risk assuming the data-generation probability model defined in (Ishida et al., 2017), and then show that additivity and duality play a significant role in the analysis.

The labels specifying the incorrect classes are equally informative as the remaining labels that specify the classes to which a pattern likely belongs. The latter labels are termed as *candidate labels* and the set of these labels is defined as $Y \in \mathfrak{P}_N(\mathcal{Y})$, where $\mathfrak{P}_N(\mathcal{Y})$ denotes the set of all the N-size subsets of $\mathcal{Y}$. From now on the formulation in this paper is constructed based on the fact that providing M $(= K - N)$ complementary labels to a training data is equivalent to providing N $(= |Y| = K - M)$ candidate labels.

4.1. Generalization of Loss Function

For one-versus-all and pairwise classification from a set of multiple candidate labels Y, the corresponding loss functions $\mathcal{L}_{\text{OVA}}$ and $\mathcal{L}_{\text{PC}}$ can be defined as:

$$\mathcal{L}_{\text{OVA}}(f(x), Y) = \frac{K-N}{K-1} \sum_{y \in Y} \ell(g_y(x)) + \frac{N}{K-1} \sum_{\overline{y} \notin Y} \ell(-g_{\overline{y}}(x)) \tag{10}$$

$$\mathcal{L}_{\text{PC}}(f(x), Y) = \sum_{y \in Y} \sum_{\overline{y} \notin Y} \ell(g_y(x) - g_{\overline{y}}(x)) \tag{11}$$

Note that (10) and (11) are the same as (1) and (2), respectively, if $N = 1$. Similarly, the newly defined equations are the same as (5) and (6) if $N = K - 1$.

As a generalization of (10) and (11), we introduce a loss function $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is defined as the *additive form* of loss functions for ordinary-label learning:

$$\mathcal{L}\left(f(x), Y\right) = \xi_1 \sum_{y \in Y} \mathcal{L}\left(f(x), y\right) + \xi_2 \tag{12}$$

where $\xi_1 \in \mathbb{R}_+$ and $\xi_2 \in \mathbb{R}$ are constants. Assuming $\mathcal{L}\left(f(x), y\right) = \mathcal{L}_{\mathrm{OVA}}\left(f(x), y\right)$, $\xi_1 = 1$, and $\xi_2 = -\frac{aN(N-1)}{K-1}$, (12) becomes the same expression as (10) if ℓ satisfies (3). Under the same condition on ℓ, by assuming $\mathcal{L}\left(f(x), y\right) = \mathcal{L}_{\mathrm{PC}}\left(f(x), y\right)$, $\xi_1 = 1$, and $\xi_2 = -a \cdot {}_NC_2$, (12) and (11) have the same expression (see supplementary materials for complete proofs).

Similar to the loss functions defined in (7), $\mathcal{L}$ in (12) is defined from the ordinary loss function $\mathcal{L}$. In contrast to the loss function defined in (7), which can be applied only if a single complementary label is provided, the loss function pertaining to (12) can be applied for any number of complementary (or candidate) labels. Consequently, $\mathcal{L}$ in (12) can be considered as a generalized form of (5) and (6).

Furthermore, for any x, let the loss function $\mathcal{L}$ satisfy the following condition:

$$\sum_{y \in \mathcal{Y}} \mathcal{L}\left(f(x), y\right) = m_1$$

where $m_1 \in \mathbb{R}$ is a constant. Both $\mathcal{L}_{\mathrm{OVA}}$ and $\mathcal{L}_{\mathrm{PC}}$ satisfy this condition, where $m_1 = aK$ and $m_1 = a \cdot {}_KC_2$, respectively. For any x and y, we then define a function $\overline{\mathcal{L}}$, which satisfies the following condition:

$$\mathcal{L}\left(f(x), y\right) + \overline{\mathcal{L}}\left(f(x), y\right) = m_2$$

where $m_2 \in \mathbb{R}$ is a constant. This condition is satisfied by $\mathcal{L}_{\mathrm{OVA}}$ and $\overline{\mathcal{L}}_{\mathrm{OVA}}$ and also by $\mathcal{L}_{\mathrm{PC}}$ and $\overline{\mathcal{L}}_{\mathrm{PC}}$, where $m_2 = 2a$ and $a(K - 1)$ respectively.

Now, the following equation holds:

$$\mathcal{L}\left(f(x), Y\right) = \xi_1 \sum_{\overline{y} \in \overline{Y}} \overline{\mathcal{L}}\left(f(x), \overline{y}\right) + \overline{\xi}_2 \tag{13}$$

where $\overline{Y} \cup Y = \mathcal{Y}$ and $\overline{Y} \cap Y = \emptyset$. Note that $\overline{\xi}_2 = \xi_1 m_1 + \xi_2 - \xi_1 m_2(K - N)$, and it can be expressed as $\xi_2 = \overline{\xi}_2$ if $m_1 = m_2 = 0$. (12) and (13) express the *duality* of the loss function $\mathcal{L}$.

To the best of our knowledge, we are the first to reveal additivity and duality commonly found in loss functions for one-versus-all and pairwise classification. This property is critical to study the classification risk and error, and it has not been discussed in any of the existing studies (Ishida et al., 2019; Cao and Xu, 2020).

4.2. Assumption of Generation Probability

As stated in Section 3, (8) represents the probability of $\overline{y}$ not being a true label. This aspect can be interpreted in the context of a candidate label; $\overline{P}\left(\overline{y}|x\right)$ represents the probability with which a true label is included in Y. In contrast to (8), which can only be applied when

$M = 1$ (or $N = |\mathcal{Y}| - 1 = K - 1$), our generalized data-generation probability model is defined as:

$$P_N(Y|x) = \frac{1}{_{K-1}C_{N-1}} \sum_{y \in Y} P(y|x) \tag{14}$$

where $P_N(Y|x)$ represents the probability of the true label being included in Y. In the remaining work, we assume that Y is generated from $P_N(Y|x)$ independently; (14) is obviously the same as (8) if $N = K - 1$.

This model represents the situation where the annotator is constrained to provide N labels consistently to any pattern x. This situation can be seen as synthetic; that is, the labelling is constrained to be independent from the annotator's belief. In fact, this model does not consider the bias in labeling by the annotator because the information in y that satisfies $y \in Y$ is treated uniformly. However, this property does not always limit reality of (14). In the following explanation we show that the synthetic property of (14) can be reasonably interpreted in a context of privacy preserving.

Given a data provided with a set of candidate labels which is generated from (14), consider the information about the ground-truth label obtained from the given data. We define a function $Q(\alpha|x) = \Pr\{y_o = \alpha|x\}$ on $\mathcal{Y}$, which represents the degree of confidence where the one given the data believes that the ground-truth label y_o for the pattern x is $\alpha \in \mathcal{Y}$. Now we assume that a set of candidate labels Y is provided to the pattern x by the annotator. Further, we assume that the one given the data does not have any information about the ground-truth label y_o, except the given set of candidate labels Y. That is, we assume that the degree of confidence, where the ground-truth label y_o for the pattern x is $\alpha \in \mathcal{Y}$, depends on whether $\alpha \in Y$ or $\alpha \notin Y$, thus $Q(\alpha|x, \alpha \in Y) = 1/N$ and $Q(\alpha|x, \alpha \notin Y) = 0$ hold. Here, the following equation holds (see supplementary materials for complete proofs).

$$Q(\alpha|x) = \beta P(\alpha|x) + (1 - \beta)\frac{1}{K}, \tag{15}$$

where

$$\beta = \frac{K - N}{N(K - 1)}.$$

According to (15), the degree of confidence $Q(\alpha|x)$ is a mixture of the distribution $P(\alpha|x)$ and the uniform distribution $1/K$; $Q(\alpha|x)$ represents the information about labels that is available to the one given the data, while $P(\alpha|x)$ represents the information about labels originally owned by the annotator. This equation indicates that receiving a data provided with a set of N candidate labels generated from (14) is equivalent to receiving a data provided with a ordinary label to which random noise is added. The level of random noise $(1 - \beta)$ increases as the number of candidate labels N increases. That is, privacy for the generation probability of the label for a given pattern, $P(y|x)$, can be preserved by synthetically generating candidate labels according to (14).

Note that data-generation probability model was defined in the same expression assuming multiple complementary labels are provided in (Cao and Xu, 2020; Feng et al., 2020). Here, the number of labels was assumed as a constant value in (Cao and Xu, 2020) while

that was assumed as a stochastic variable in (Feng et al., 2020). Those probability models are mathematically equivalent to (14), which can be easily shown; assuming $\mathcal{Y} = Y \cup \overline{Y}$ and $\emptyset = Y \cap \overline{Y}$, the number of patterns choosing N candidate labels from the set of all the labels $\mathcal{Y}$ is equal to the number of patterns choosing $K - N$ complementary labels. Despite the fact, both the probability models defined in (Cao and Xu, 2020; Feng et al., 2020) are not in the expression clarifying the relationship between ordinary label and complementary label, as those are straightforward extension of the probability model defined in (Ishida et al., 2017) from a perspective of increasing the number of complementary labels. Further, to the best of our knowledge, this paper is the first to explicitly provide an interpretation of the information obtained from the labels generated by (14); (15) clarifies that the information appears to be obtained from a mixture of the true distribution and uniform distribution. The goal of this paper is to discuss recent studies in complementary labels in a unified manner based on a concept of candidate labels. Thus, we use the expression form in (14) to naturally relate ordinary-label learning and complementary-label learning.

4.3. Classification Risk with Multiple Candidate Labels

If we define loss function $\mathcal{L}$ as in (12), the following theorem holds for any type of $\mathcal{L}$. Similar to (4) and (9), the theorem allows the expression of the classification risk from multiple complementary labels in terms of the expectation of loss. The complete proof is available in the supplementary materials.

Theorem 1 *Given a pattern x, a set of candidate labels Y, and loss function $\mathcal{L}$ defined by (12), the classification risk $R(f)$ can be expressed as follows:*

$$R(f) = \frac{K-1}{\xi_1 (K-N)} \mathbb{E}_{P_N(x,Y)} \left[\mathcal{L} \left(f(x), Y \right) \right] + C \tag{16}$$

where

$$C = -\frac{N-1}{K-N} \sum_{y \in \mathcal{Y}} \mathcal{L} \left(f(x), y \right) - \frac{\xi_2 (K-1)}{\xi_1 (K-N)} \tag{17}$$

Some conditions can be used to simplify the expression in (16) for one-versus-all and pairwise classification, without losing generalities. Note that redefining (3) as $\ell(z) := \ell(z) - \ell(0) = \ell(z) - \frac{a}{2}$ does not affect the loss minimization when learning. If we shift ℓ to satisfy $a = 0$, $m_1 = 0$ holds for both $\mathcal{L}_{\text{OVA}}$ and $\mathcal{L}_{\text{PC}}$; therefore, the first term in (17) can be eliminated.

Similarly, ξ_1 and ξ_2 in (12) do not affect the loss minimization. Considering $\xi_1 = 1$ and $\xi_2 = 0$, we can simplify (12) as follows without loss of generalities:

$$\mathcal{L} \left(f(x), Y \right) = \sum_{y \in Y} \mathcal{L} \left(f(x), y \right) \tag{18}$$

Consequently, by assuming that $a = 0$, $\xi_1 = 1$, and $\xi_2 = 0$, the following expression holds for the classification risk.

$$R(f) = \frac{K-1}{K-N} \mathbb{E}_{P_N(x,Y)} \left[\mathcal{L} \left(f(x), Y \right) \right] \tag{19}$$

5. Statistical Analysis

This section discusses the error bound for one-versus-all and pairwise classification. For simplicity, we assume that ℓ satisfies (3) with $a = 0$, and the conditions $\inf_z \ell(z) = -1/2$ and $\sup_z \ell(z) = 1/2$. Further, we assume that ℓ is Lipschitz continuous. In addition, we assume $\xi_1 = 1$ and $\xi_2 = 0$. For the rest of this paper, we define $\mathcal{L}$ and $R(f)$ by using (18) and (19), respectively.

5.1. Notations

Let us consider that a set of n training data $\mathcal{S} = \{(x_i, Y_i)\}_{i=1}^n$ is given, and each training data is generated with a probability of $P_N(x, Y)$ independently. Based on (19), the empirical classification risk $\hat{R}(f)$ for the set $\mathcal{S}$ is:

$$\hat{R}(f) = \frac{K-1}{n(K-N)} \sum_{i=1}^n \mathcal{L}(f(x_i), Y_i)$$

We define the ideal classifier that minimizes the generalization error (*Bayes classifier*) and the empirically ideal classifier as $f^* := \arg\min_f R(f)$ and $\hat{f} := \arg\min_f \hat{R}(f)$, respectively. We define the classification error $\mathcal{E}_N$ for the classifier $\hat{f}$ as follows.

$$\mathcal{E}_N = R(\hat{f}) - R(f^*) \tag{20}$$

In the literature, the *Rademacher complexity* for a set of discriminant functions $\mathcal{G}$ over the input space $\mathcal{X}$ is usually defined as follows:

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\sigma} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) \right]$$

where $\sigma = \{\sigma_1, \cdots, \sigma_n\}$ is a set of independent stochastic variables, which take one value of $\{-1, +1\}$ with the same probability. In addition, $\mathbb{E}_{\mathcal{S}}$ and $\mathbb{E}_{\sigma}$ represent the expectation for each element of $\mathcal{S}$ and σ, respectively.

5.2. Evaluation of Error Bound

The following lemmas are introduced to derive the classification error bound.

Lemma 2 *We express the supremum of the difference in loss $\|\mathcal{L}\|_\infty$ in accordance with the change in a set of candidate labels, i.e., given any $Y, Y' \in \mathfrak{P}_N(\mathcal{Y})$,*

$$\|\mathcal{L}\|_\infty = \sup_{g_1, \cdots, g_K \in \mathcal{G}} \left(\sum_{y \in Y} \mathcal{L}(f(x), y) - \sum_{y' \in Y'} \mathcal{L}(f(x), y') \right)$$

Then, the following holds for one-versus-all classification.

$$\|\mathcal{L}_{OVA}\|_\infty = \begin{cases} \dfrac{KN}{K-1}, & if\ N \le \dfrac{K}{2} \\ \dfrac{K(K-N)}{K-1}, & otherwise \end{cases}$$

Similarly, for pairwise classification,

$$\|\boldsymbol{\mathcal{L}}_{PC}\|_{\infty} = N(K - N)$$

Lemma 3 *Define a function set $\mathcal{H}_{OVA}$, $\mathcal{H}_{PC}$ as follows:*

$$\mathcal{H}_{OVA} = \{(x, Y) \mapsto \boldsymbol{\mathcal{L}}_{OVA}\left(f(x), Y\right) \mid g_1, \cdots, g_K \in \mathcal{G}\}$$
$$\mathcal{H}_{PC} = \{(x, Y) \mapsto \boldsymbol{\mathcal{L}}_{PC}\left(f(x), Y\right) \mid g_1, \cdots, g_K \in \mathcal{G}\}$$

Then, if ℓ is L_ℓ (≥ 0) Lipschitz continuous, the following holds for $\mathcal{H}_{OVA}$.

$$\mathfrak{R}_n\left(\mathcal{H}_{OVA}\right) \leq \begin{cases} \dfrac{K(K + N)}{K - 1} L_\ell \mathfrak{R}_n\left(\mathcal{G}\right), & \text{if } N \leq \dfrac{K}{2} \\ \dfrac{K(2K - N)}{K - 1} L_\ell \mathfrak{R}_n\left(\mathcal{G}\right), & \text{otherwise} \end{cases}$$

Similarly, for $\mathcal{H}_{PC}$,

$$\mathfrak{R}_n\left(\mathcal{H}_{PC}\right) \leq 2K(K - 1)L_\ell \mathfrak{R}_n\left(\mathcal{G}\right)$$

Based on these lemmas, the error bounds for the one-versus-all and pairwise classification can be defined as follows. The complete proofs are provided in the supplementary materials.

Theorem 4 *Assume that function ℓ satisfies the stated condition in the beginning of this section, and is L_ℓ Lipschitz continuous. Then, for any δ (> 0), the following equation for one-versus-all classification holds with a probability of at least $1 - \delta$.*

$$\mathcal{E}_N \leq \begin{cases} \dfrac{4K(K + N)}{K - N} L_\ell \mathfrak{R}_n\left(\mathcal{G}\right) + \dfrac{KN}{K - N}\sqrt{\dfrac{2\ln(2/\delta)}{n}}, & \text{if } N \leq \dfrac{K}{2} \\ \dfrac{4K(2K - N)}{K - N} L_\ell \mathfrak{R}_n\left(\mathcal{G}\right) + K\sqrt{\dfrac{2\ln(2/\delta)}{n}}, & \text{otherwise} \end{cases} \tag{21}$$

Similarly, for pairwise classification, the following equation holds with a probability of at least $1 - \delta$.

$$\mathcal{E}_N \leq \dfrac{8K(K - 1)^2}{K - N} L_\ell \mathfrak{R}_n\left(\mathcal{G}\right) + N(K - 1)\sqrt{\dfrac{2\ln\left(2/\delta\right)}{n}} \tag{22}$$

Note that the upper-bound of $\mathcal{E}_N$ increases monotonically corresponding to N in accordance with (21) and (22). This aspect is in agreement with the fact that a decrease in the number of candidate labels leads to less ambiguous supervision of the training data. The upper-bound in (21) breaks subject to $N \leq K/2$; this property is due to the equivalency satisfied by ordinary labels and complementary labels, as discussed in Section 4.2. Note that taking the expectation of the error bounds in the theorem 4 assuming that N as a random variable, the derived formulations specify the error bounds under the problem setting where the number of candidate labels provided to each training data stochastically fluctuates.
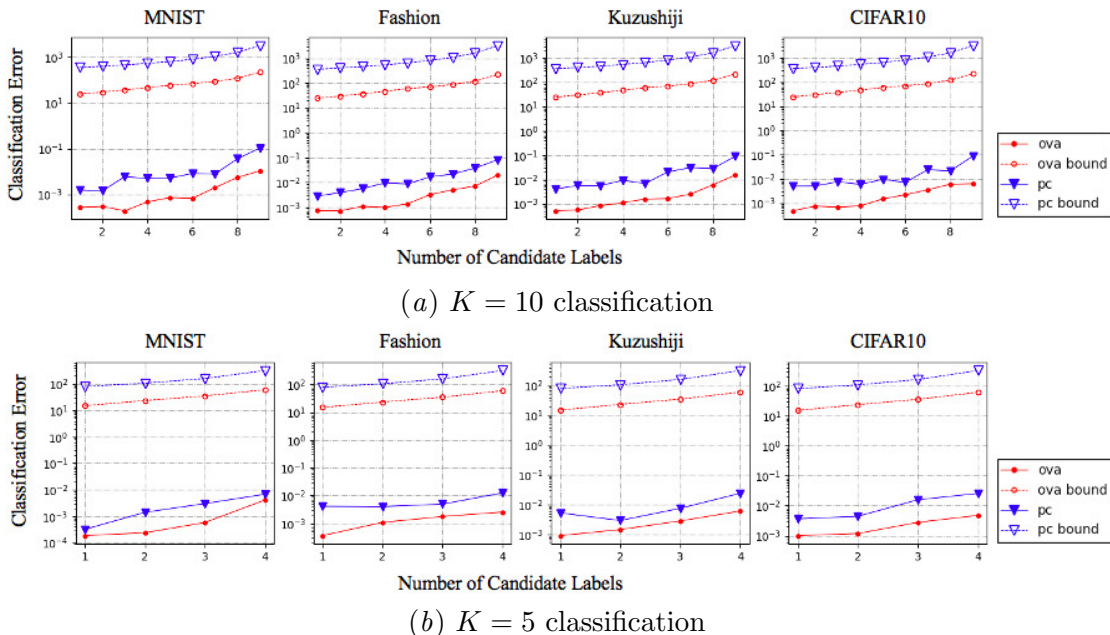
(a) $K = 10$ classification

(b) $K = 5$ classification

Figure 1: Error for 10 classification for different numbers of complementary labels N. The red and blue closed plots represent the experimental results for the one-versus-all and pairwise classification, respectively. The red and blue open plots represent the theoretical error bounds for the one-versus-all and pairwise classification, respectively.

6. Experiment

This section describes the evaluation of the accuracy of one-versus-all and pairwise classification and the validation of the formulation discussed in Section 5.2. Understanding the exact behavior of the classification error only from the derived equations in Theorem 4 is difficult. Therefore, we attempt to quantitively discuss the error in a real-world classification problem. The source code for the described experiment is available online[1].

6.1. Dataset Generation

The generation probability model of the training data provided with multiple candidate labels is as defined in (14). We generate the experimental dataset according to this definition, and the annotation of the training data is performed as follows. First, we prepare pretrained K class classifiers, namely, *annotators*, for the ordinarily labeled training data. The annotators are multi-layer perceptrons (MLP) with softmax as the output layer activation function. The annotators are trained on the MNIST[2], Fashion-MNIST[3], Kuzushiji-MNIST[4], and CIFAR-10[5] datasets. Table 1 summarizes the datasets. Because all these

---

1. https://github.com/YasuhiroKatsura/ord-comp
2. http://yann.lecun.com/exdb/mnist/
3. https://github.com/zalandoresearch/fashion-mnist
4. https://github.com/rois-codh/kmnist
5. https://www.cs.toronto.edu/~kriz/cifar.html

Table 1: Datasets used in the experiment.

	MNIST	Fashion	Kuzushiji	CIFAR-10
Number of classes	10	10	10	10
Number of dimensions	28×28	28×28	28×28	$3 \times 32 \times 32$
	(gray)	(gray)	(gray)	(RGB)
Number of data	60,000	60,000	60,000	60,000

cases correspond to the 10 class ($y \in \{0, \cdots, 9\}$) classification problem, the data belonging to $y = 5, \cdots, 9$ is eliminated when we assume $K = 5$ in the following experiment. Because the discriminant functions of the annotators $g_1, \cdots, g_K$ are normalized using softmax, each of the functions can be treated as having a data generation probability $P(y|x)$. Therefore, assuming that the number of candidate labels is N, we generate Y in accordance with $P_N(f(x), Y)$, calculated using the discriminant functions. Similarly, for the test dataset, the true label $y \in \mathcal{Y}$ is provided in accordance with $g_1, \cdots, g_K$ of the annotators.

In the experiment, we pretrained the annotators for $K = 10$ and $K = 5$. When $K = 10$, the classification accuracies of the annotators for the MNIST, Fashion-MNIST, Kuzushiji-MNIST, and CIFAR-10 datasets are 99.17%, 92.67%, 95.1%, and 81.62%, respectively. For $K = 5$, the corresponding accuracies are 99.90%, 95.16%, 96.56%, and 85.60%. As all the accuracies are reasonably high, the generation probability $P(y|x)$ for each class $y \in \mathcal{Y}$ is not uniform. As discussed in section 4.2, providing a set of candidate labels generated from (14) is equivalent to providing a single ordinary label generated from the true probability distribution $P(y|x)$ with random noise from uniform distribution. Therefore, generating synthetic data as in the experiment is natural in the context of preserving privacy of training data.

Here, we show the data augmentation algorithm used in the experiment in Algorithm 1. Algorithm 1 requires n patterns $\{x_i\}_{i=1}^{n}$ as inputs and returns $\{x_i, Y_i\}_{i=1}^{n}$ by providing a set of candidate labels to each of input patterns. Further, we define `Annotator` as a pretrained multi-layer perceptron which returns degree of classification confidence for each patterm x_i and `ChooseCandidates` as a stochastic function which returns a set of candidate labels based on (14).

6.2. Evaluation of Classification Error Bound

Experimental Setup: Assuming that the number of classes K is 10 or 5, the classification accuracy is compared under different numbers of candidate labels $N \in \{1, \cdots, K-1\}$. The loss functions $\mathcal{L}_{\text{OVA}}$ and $\mathcal{L}_{\text{PC}}$ are defined using (18), and the function ℓ is an origin-symmetric sigmoid loss, i.e., $\ell(z) = (1 + e^z)^{-1} - 1/2$. We set the number of training data for each class as $1,000$, and the data were randomly selected from the complete dataset. We set the batch size as 64 for the rest of the experiment. We conducted the experiment using MLP for the MNIST, Fashion-MNIST, and Kuzushiji-MNIST datasets. The number of epochs was 300, weight decay was 10^{-4}, learning rate was 5×10^{-4}, and *Adam* was used as the optimization algorithm (Kingma and Ba, 2015). For CIFAR-10, the experiment was performed using DenseNet (Huang et al., 2017). The number of epochs was 300, weight decay was 5×10^{-4}, momentum was 0.9, and the optimization algorithm was the *stochastic*

Algorithm 1: Data Augmentation

Input: the number of the total classes K, the number of candidate labels N, set of n patterns $\{x_i\}_{i=1}^{n}$

Output: set of training data with candidate labels $\{x_i, Y_i\}_{i=1}^{n}$

Function `Annotator(x)`:

    Calculate classification confidence $p_i \in [0,1]$ for each class $(i = 1, \cdots, K)$.

    return $(p_1, \cdots, p_K)$

Function `ChooseCandidates(p)`:

    Calculate generation probability by (14) for all the possible combinations of N candidate labels.

    Then choose a set of candidate labels from the combinations randomly.

    return $(y_1, \cdots, y_N)$

for $i \leftarrow 1$ **to** n **do**

    $p \leftarrow$ `Annotator`(x_i)

    $Y_i \leftarrow$ `ChooseCandidates`(p)

end

return $\{x_i, Y_i\}_{i=1}^{n}$

Table 2: Experimental classification accuracies for 10 and 5 class classification (%). The experiments were performed 5 times for each case; the mean accuracy and standard deviation are presented by the upper and lower values, respectively. The highest accuracy is boldfaced.

N	K=10									K=5			
	1	2	3	4	5	6	7	8	9	1	2	3	4
MNIST	**92.28**	88.22	89.44	86.86	80.72	85.99	77.28	72.27	62.13	**98.06**	97.77	97.06	95.40
(OVA)	(±3.45)	(±6.97)	(±4.23)	(±4.16)	(±6.05)	(±4.82)	(±5.02)	(±6.42)	(±11.03)	(±0.12)	(±0.18)	(±0.20)	(±0.18)
MNIST	**92.38**	91.58	85.94	87.38	84.93	83.93	84.79	71.73	67.07	**98.12**	97.61	97.12	95.80
(PC)	(±3.42)	(±3.57)	(±6.57)	(±4.26)	(±6.53)	(±3.57)	(±3.94)	(±4.63)	(±4.66)	(±0.13)	(±0.14)	(±0.16)	(±0.29)
Fashion	**81.07**	80.66	80.43	80.62	79.64	78.99	75.10	74.89	71.41	**86.42**	85.63	84.94	83.37
(OVA)	(±0.09)	(±0.32)	(±1.05)	(±1.15)	(±1.03)	(±1.49)	(±5.58)	(±2.61)	(±2.59)	(±0.26)	(±0.83)	(±0.41)	(±0.37)
Fashion	**81.47**	81.51	80.41	81.23	79.24	79.88	78.45	75.64	73.11	**86.52**	85.74	84.74	83.14
(PC)	(±0.16)	(±0.95)	(±0.16)	(±1.32)	(±0.38)	(±1.29)	(±0.68)	(±3.79)	(±2.30)	(±0.33)	(±0.44)	(±0.37)	(±0.41)
Kuzushiji	**64.19**	60.83	58.96	57.45	52.20	51.28	49.43	41.27	37.09	**80.49**	79.02	75.78	70.55
(OVA)	(±2.45)	(±2.33)	(±5.90)	(±3.55)	(±2.96)	(±1.60)	(±5.65)	(±3.46)	(±3.76)	(±0.33)	(±0.61)	(±0.25)	(±1.08)
Kuzushiji	**66.15**	63.67	58.53	57.20	55.96	54.28	49.74	43.62	34.82	**80.97**	78.78	76.53	70.23
(PC)	(±0.48)	(±2.06)	(±5.82)	(±4.09)	(±3.44)	(±1.95)	(±4.02)	(±0.66)	(±3.07)	(±0.22)	(±0.47)	(±0.65)	(±1.79)
CIFAR-10	**57.17**	53.44	51.16	45.88	37.95	34.85	25.54	27.43	24.18	**65.72**	63.76	61.30	52.65
(OVA)	(±3.30)	(±4.95)	(±2.01)	(±4.42)	(±5.96)	(±4.39)	(±3.80)	(±3.36)	(±3.63)	(±0.37)	(±0.36)	(±0.56)	(±4.83)
CIFAR-10	**58.30**	49.04	41.63	35.68	33.29	27.62	22.26	22.36	17.32	**69.11**	66.98	62.10	56.29
(PC)	(±2.89)	(±3.56)	(±6.24)	(±4.64)	(±3.10)	(±3.43)	(±3.58)	(±4.35)	(±1.20)	(±2.59)	(±3.72)	(±1.47)	(±2.36)

gradient descent. The initial learning rate was set as 10^{-2}, and it was halved every 30 epochs. The range of discriminant functions g_y was $[-1/2, 1/2]$ for both MLP and DenseNet.

Error Computation: The classification error was calculated according to (20). Because the exact $\hat{f}$ defined in Section 5.1 was not available, we surrogated it with a classifier that could minimize the loss of the training data in the experiment. Similarly, we substituted f^* with a classifier that could minimize the loss of the test data.

Additionally, we computed the theoretical classification error bounds according to (21) and (22). We set $\delta = 0.1$ because Theorem 4 holds with a high probability if δ is relatively small. Furthermore, we set L_ℓ as 1.0 because it is the minimum Lipschitz constant of

Table 3: Experimental classification errors for 10 and 5 class classification ($\times 10^{-3}$). The experiments were performed 5 times for each case; the mean error and standard deviation are presented by the upper and lower values, respectively. The highest error is boldfaced.

N					$K=10$							$K=5$	
	1	2	3	4	5	6	7	8	9	1	2	3	4
MNIST	0.27	0.30	0.19	0.47	0.72	0.67	1.96	5.47	**10.69**	0.18	0.24	0.57	**4.07**
(OVA)	(±0.19)	(±0.19)	(±0.18)	(±0.29)	(±0.40)	(±0.34)	(±0.99)	(±3.92)	(±5.80)	(±0.19)	(±0.18)	(±0.89)	(±1.57)
MNIST	1.56	1.41	6.00	4.94	5.15	8.22	7.78	35.94	**106.35**	0.32	1.40	2.99	**6.74**
(PC)	(±1.33)	(±1.07)	(±3.79)	(±2.72)	(±3.41)	(±4.78)	(±6.75)	(±32.09)	(±92.07)	(±0.23)	(±0.94)	(±1.15)	(±2.20)
Fashion	0.72	0.71	1.08	0.98	1.36	3.19	4.98	7.01	**19.51**	0.36	1.07	1.75	**2.53**
(OVA)	(±0.48)	(±0.37)	(±0.27)	(±0.73)	(±1.26)	(±1.06)	(±3.64)	(±2.30)	(±10.60)	(±0.25)	(±0.61)	(±1.17)	(±1.76)
Fashion	2.78	3.86	5.54	9.55	8.59	16.56	21.51	37.45	**81.23**	4.09	3.99	4.88	**12.46**
(PC)	(±2.30)	(±3.70)	(±2.21)	(±3.70)	(±8.09)	(±10.85)	(±11.12)	(±28.52)	(±50.82)	(±2.64)	(±4.05)	(±3.60)	(±6.29)
Kuzushiji	0.52	0.56	0.85	1.11	1.57	1.66	2.49	6.01	**15.26**	0.91	1.46	2.85	**6.10**
(OVA)	(±0.37)	(±0.45)	(±0.58)	(±0.45)	(±0.30)	(±0.63)	(±1.85)	(±3.20)	(±7.67)	(±0.74)	(±0.96)	(±1.45)	(±2.54)
Kuzushiji	4.06	5.73	5.60	9.27	6.88	20.39	31.05	27.94	**89.97**	5.31	2.94	7.54	**24.13**
(PC)	(±3.79)	(±7.44)	(±4.07)	(±6.34)	(±4.26)	(±16.62)	(±20.92)	(±14.73)	(±17.50)	(±1.61)	(±2.95)	(±6.36)	(±20.90)
CIFAR-10	0.48	0.74	0.67	0.76	1.47	2.18	3.54	5.98	**6.33**	1.01	1.18	2.79	**4.76**
(OVA)	(±0.25)	(±0.39)	(±0.19)	(±0.17)	(±0.75)	(±1.12)	(±1.99)	(±2.43)	(±1.61)	(±0.66)	(±0.74)	(±1.54)	(±2.97)
CIFAR-10	5.02	5.08	7.51	5.92	9.39	7.15	23.80	21.19	**86.60**	3.65	4.43	15.77	**25.66**
(PC)	(±0.51)	(±2.38)	(±1.46)	(±5.00)	(±4.70)	(±5.09)	(±24.52)	(±11.01)	(±38.95)	(±0.72)	(±2.40)	(±7.76)	(±8.08)

the shifted sigmoid loss. Assuming that both MLP and DenseNet in the experiment had adequate capacity, we set $\mathfrak{R}_n(\mathcal{G}) = 0.5$.

Results: We performed 5 trials for each experiment and observed the classification accuracy, that is, rate of correct classification for the test dataset estimated by a classifier which minimizes loss for training data. The mean and standard deviation for the accuracy are listed in Table 2. The results indicate that the mean of the accuracy tends to monotonically decrease corresponding to N. That is, increase in the number of complementary labels leads to a better performance than the case of $N = K - 1$ ($M = 1$).

The mean and standard deviation of the experimental classification error are listed in Table 3. Furthermore, the mean of the experimental error and theoretical error bounds are shown in Figure 1 as a logarithmic graph. The results indicate that the theoretical error bound for one-versus-all and pairwise classification are about equally tight. The experimental error tends to monotonically increase corresponding to N, which is in accordance with the discussion in Section 5.2. The increase in theoretical error bound and experimental error are similar in shape, indicating that the bound reflects qualitative property found in the experimental error.

Note that all the experiments performed in this section are intended to evaluate the derived formulation in Section 5. Experiments under exhaustive conditions would be more informative; one example is a performance comparison to evaluate the number of training data n required to train a classifier which performs better than a certain threshold.

7. Conclusion

Focusing on the fact that ordinary label and complementary label are essentially equivalent, we naturally related learning framework from single ordinary/complementary label. We found that the loss functions for one-versus-all and pairwise classification satisfy certain *additivity* and *duality* which play a significant role in the analysis of classification risk and error bound. As a result, we made it possible to understand learning from ordi-

nary/complementary label, which are studied independently in a existing literature, as a new expression form in a unified perspective.

The analysis in this paper is under the the assumption that each training data is provided from a data-generation probability model which satisfies a certain uniformity of the assigned labels. All the existing literature studying learning framework from multiple complementary labels are under the same assumption. This assumption is natural in a context of privacy preserving. As we described in this paper, the defined data-generation probability model indicates that providing multiple candidate labels generated from the model helps to prevent leakage of information about true distribution of labels for any patterns. However, it is also interesting to analyze under the problem setting where the assumption does not hold. Studies on the data-generation probability models to expand the range of application is the important future work.

Acknowledgments

References

Yuzhou Cao and Yitian Xu. Multi-complementary and unlabeled learning for arbitrary losses and models. *ArXiv*, abs/2001.04243, 2020.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

Jesús Cid-sueiro. Proper losses for learning from partial labels. In *Advances in Neural Information Processing Systems 25*, pages 1565–1573. 2012.

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

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning from multiple complementary labels. *ArXiv*, abs/1912.12927, 2020.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17*, pages 529–536. 2005.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in Neural Information Processing Systems 30*, pages 5639–5649. 2017.

Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In *Advances in Neural Information Processing Systems 31*, pages 5917–5928. 2018.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2971–2980, 2019.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 101–110, 2019.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *In Proceedings of the International Conference on Learning Representations*, 2015.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2017.

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

Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems 23*, pages 1504–1512. 2010.

Gideon S. Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th International Conference on Machine Learning*, page 593–600, 2007.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, London Mathematical Society Lecture Notes, pages 148–188. 1989.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, pages 1196–1204. 2013.

Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 10–18, 2013.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.