IEICE TRANS. INF. & SYST., VOL.E105–D, NO.2 FEBRUARY 2022

PAPER

Learning from Noisy Complementary Labels with Robust Loss Functions

Hiroki ISHIGURO[†a)], **Takashi ISHIDA**[†,††b)], ***Nonmembers***, ***and*** **Masashi SUGIYAMA**[†,††c)], ***Fellow***

SUMMARY It has been demonstrated that large-scale labeled datasets facilitate the success of machine learning. However, collecting labeled data is often very costly and error-prone in practice. To cope with this problem, previous studies have considered the use of a complementary label, which specifies a class that an instance does not belong to and can be collected more easily than ordinary labels. However, complementary labels could also be error-prone and thus mitigating the influence of label noise is an important challenge to make complementary-label learning more useful in practice. In this paper, we derive conditions for the loss function such that the learning algorithm is *not* affected by noise in complementary labels. Experiments on benchmark datasets with noisy complementary labels demonstrate that the loss functions that satisfy our conditions significantly improve the classification performance.

key words: *complementary label, label noise, robust loss function, loss correction*

1. Introduction

Training of deep neural networks (DNNs) often requires large-scale datasets such as the ImageNet dataset [1], which contains 1.2 million images. However, when the dataset is very large, labeling training instances through a data collecting system (e.g., crowdsourcing [2]) becomes more laborious and consequently error-prone. Previous studies have focused on the use of a *complementary label* [3]–[6] as one of the solutions, which specifies a class that an instance does *not* belong to. Some examples of complementary labels are given in Fig. 1 (a). Although complementary labels are less informative than ordinary labels, they can significantly reduce the burden on labelers. This is because it is much easier to distinguish if an instance does not belong to a particular class than to choose the correct class from all the candidates. Complementary-label learning was originally conceived in Ishida et al. [3] and they derived an unbiased estimator of the classification risk for one-versus-all (OVA) and pairwise comparison (PC) loss functions [7]. Along this line, Yu et al. [4] proposed another solution to complementary-label learning by modifying the softmax output of the model. Although their method does not provide an unbiased risk estimator, it is theoretically guaranteed that the obtained solution is statistically consistent with the optimal risk minimizer. Later Ishida et al. [5] generalized the unbiased risk estimator for arbitrary losses and models, allowing one to choose losses that are widely used in DNN training, including the categorical cross entropy (CCE) loss. Chou et al. [6] proposed a more comprehensive framework of complementary-label learning by introducing the *complementary 0-1 loss* and its surrogates. Within this framework, various extensions have been proposed such as *biased complementary labels* [4] and *multiple complementary labels* [8]. More recently, complementary labels have been attracting attention for a variety of purposes, such as *generative-discriminative learning* [9] and *noisy data filtering* [10].

Fig. 1 Examples of (a) complementary labels and (b) noisy complementary labels. True and actually labeled classes are shown above and below the images. These images are taken from the CIFAR-10 dataset [11].

Even though complementary labels are much easier to obtain than ordinary labels, they still suffer from annotation errors during their collection process. For example, a labeler can mistakenly tie the true class to an instance as a complementary label, as shown in Fig. 1 (b). In such a case, one needs to train a classifier from data that contains *noisy (incorrect) complementary labels*. In fact, it has been pointed out that DNNs can easily fit noisy samples, which cases significant degradation in generalization performance [12], [13]. This motivates us to develop a complementary-label learning method that is robust against noise. In this paper,

Manuscript received February 15, 2021.
Manuscript revised July 14, 2021.
Manuscript publicized November 1, 2021.
[†]The authors are with the University of Tokyo, Tokyo, 113–0033 Japan.
[††]The authors are with RIKEN, Tokyo, 103–0027 Japan.
a) E-mail: ishiguro@ms.k.u-tokyo.ac.jp
b) E-mail: ishi@k.u-tokyo.ac.jp
c) E-mail: sugi@k.u-tokyo.ac.jp
DOI: 10.1587/transinf.2021EDP7035

within the complementary-label learning framework of Ishida et al. [5], we derive an unbiased risk estimator for noisy complementary labels. In our risk estimator, we regard complementary labels as noisy complementary labels from ordinary labels. We model this noise process with the *noise transition matrix* [14]–[20], which describes the probability of label flipping. Then we model label flipping in complementary labels as the misspecification of the noise transition matrix. Our discovery is that it is possible to obtain a statistically consistent classifier even for a misspecified noise transition matrix if we choose losses that satisfy certain conditions. Moreover, we relax the derived conditions to allow the use of a wider choice of losses while certain robustness is still guaranteed. Experiments on benchmark datasets demonstrate that losses that satisfy our conditions perform better than those that do not.

2. Background

In this section, we first briefly review *ordinary-label learning*, *noisy-label learning*, and *complementary-label learning*. We also give a detailed explanation of *noisy complementary-label learning*.

2.1 Ordinary-Label Learning

Let $\mathcal{X}$ be the instance space and $\mathcal{Y} := \{1, 2, \ldots, K\}$ be the label space, the integer $K \geq 2$ is the number of classes. We assume that a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ follows an unknown probability distribution $\mathcal{D}$. Given a loss function $\ell : \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}_+$, the goal of this problem is to learn a classifier $f : \mathcal{X} \to \mathbb{R}^K$ that minimizes the following risk:

$$R(f; \ell) := \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\ell(Y, f(X))\right], \tag{1}$$

where $\mathbb{E}_{(X,Y)\sim\mathcal{D}}$ denotes the expectation over $\mathcal{D}$. Suppose that training samples $\{(x_i, y_i)\}_{i=1}^n$ drawn independently and identically from $\mathcal{D}$ are available. Then the expectation over unknown $\mathcal{D}$ in Eq. (1) can be approximated by the empirical average over these samples.

2.2 Noisy-Label Learning with Loss Correction

Next, we formulate the problem of learning from noisy labels. Suppose we observe noisy training samples $\{(x_i, \widetilde{y}_i)\}_{i=1}^n$ drawn independently and identically from a noisy distribution $\widetilde{\mathcal{D}}$. We denote $(X, \widetilde{Y}) \in \mathcal{X} \times \mathcal{Y}$ as a pair of noisy random variables that follows $\widetilde{\mathcal{D}}$. In our work, we commonly assume that $\widetilde{Y}$ is independent of instance X and depends only on true label Y, i.e.,

$$\begin{aligned} &\mathbb{P}(\widetilde{Y} = j | Y = i, X = x) = \mathbb{P}(\widetilde{Y} = j | Y = i), \\ &\forall x \in \mathcal{X}, \forall i, j \in \mathcal{Y}, \end{aligned} \tag{2}$$

where $\mathbb{P}$ denotes the probability. The above probabilities are summarized into a transition matrix $\boldsymbol{T}$, where $(\boldsymbol{T})_{ij} = \mathbb{P}(\widetilde{Y} = j | Y = i)$, $\forall i, j \in \mathcal{Y}$. In general, we cannot know the ground-truth transition matrix, and thus it needs to be estimated. The estimated transition matrix $\widehat{\boldsymbol{T}}$ can be obtained by learning the noisy class posterior from noisy samples using DNNs and exploiting it to calculate $\widehat{\boldsymbol{T}}$ [14]–[20] or by making some assumptions on $\widetilde{\mathcal{D}}$ and designing $\widehat{\boldsymbol{T}}$ to be consistent with the assumptions [3], [5], [6].

The transition matrix plays a key role in building statistically consistent algorithms. Patrini et al. [15] provided two procedures with loss correction: *backward correction* and *forward correction*. Suppose $\widehat{\boldsymbol{T}}$ is invertible. Given a loss function ℓ, the backward corrected loss is defined as

$$\ell^{\leftarrow}(y, f(x)) := \sum_{j=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{yj} \ell(j, f(x)). \tag{3}$$

If $\widehat{\boldsymbol{T}} = \boldsymbol{T}$, it holds that (Patrini et al. [15], Theorem 1)

$$\begin{aligned} \widetilde{R}(f; \ell^{\leftarrow}) &= \mathbb{E}_{(X,\widetilde{Y})\sim\widetilde{\mathcal{D}}} \left[\ell^{\leftarrow}(\widetilde{Y}, f(X))\right] \\ &= \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\ell(Y, f(X))\right] = R(f; \ell), \ \forall f, \end{aligned} \tag{4}$$

where $\widetilde{R}$ denotes the noisy risk. Therefore, we can learn a minimizer of the original risk by minimizing the noisy risk:

$$\arg\min_f \widetilde{R}(f; \ell^{\leftarrow}) = \arg\min_f R(f; \ell). \tag{5}$$

Although backward correction does not impose any constraints on losses and models, it leads to severe overfitting due to the *negative risk* issues [5], [6], [21] when the models are complex like DNNs.

While backward correction directly modifies the values of a loss function itself, forward correction makes a change to the model predictions. Let Δ^{K-1} be the probability simplex in $\mathbb{R}^K$, $F : \mathbb{R}^K \to \Delta^{K-1}$ be an *invertible link*, and $\varphi : \mathcal{Y} \times \Delta^{K-1} \to \mathbb{R}_+$ be a *proper loss* that is particularly suitable for probability estimation [22], [23]. Given a *proper composite loss* $\ell(y, f(x)) := \varphi(y, F(f(x)))$, the forward corrected loss is defined as

$$\ell^{\rightarrow}(y, f(x)) := \varphi(y, \widehat{\boldsymbol{T}}^{\top} F(f(x))). \tag{6}$$

If $\widehat{\boldsymbol{T}} = \boldsymbol{T}$, it holds that (Patrini et al. [15], Theorem 2)

$$\arg\min_f \widetilde{R}(f; \ell^{\rightarrow}) = \arg\min_f R(f; \ell). \tag{7}$$

CCE may be practically employed in forward correction, since it is a widely used loss that is proper composite.

2.3 Complementary-Label Learning

Lastly, we consider learning from complementary labels. Suppose that complementary labels are uniformly chosen from classes other than the true class. This setting, called the *uniform assumption* [3], [5], [6], can be justified by properly designing the distribution, e.g., by forcing the data collecting system to ask labelers if an instance belongs to a randomly obtained class. If the answer is yes, an ordinary label

is given; otherwise, a uniformly chosen complementary label is given. Let $\overline{\mathcal{D}}$ be the distribution under this uniform assumption and $(X, \overline{Y}) \in \mathcal{X} \times \mathcal{Y}$ be a pair of random variables that follows $\overline{\mathcal{D}}$. Then, the process of generating complementary labels can be expressed by designing the transition matrix as follows:

$$(\overline{\boldsymbol{T}})_{ij} = \mathbb{P}(\overline{Y} = j | Y = i) = \begin{cases} 0 & \text{if } j = i, \\ \frac{1}{K-1} & \text{if } j \neq i. \end{cases} \tag{8}$$

We call this the *complementary matrix*. Ishida et al. [5] obtained an unbiased risk estimator by using backward correction with the complementary matrix, while Yu et al. [4] used forward correction. Both are among the most promising methods for learning a statistically consistent classifier from complementary labeled data.

2.4 Noisy Complementary-Label Learning

As we saw in Sect. 2.2, as long as the transition matrix $\widehat{\boldsymbol{T}}$ is correctly estimated, a minimizer of the original risk can be obtained with corrected losses. In contrast, this is not the case when $\widehat{\boldsymbol{T}} \neq \boldsymbol{T}$. In this paper, we mainly study the case of complementary-label learning with label noise. In Sect. 2.3, we implicitly assumed that the complementary distribution $\overline{\mathcal{D}}$ is equal to the actual data distribution $\widetilde{\mathcal{D}}$, i.e., the uniform assumption is correct and therefore $\overline{\boldsymbol{T}} = \boldsymbol{T}$ holds. However, if complementary labels are affected by class-dependent label noise, i.e., each complementary label $\overline{y}$ is flipped to $\widetilde{y}$ with probability $\mathbb{P}(\widetilde{Y} = \widetilde{y} | \overline{Y} = \overline{y})$, $\overline{\boldsymbol{T}} = \boldsymbol{T}$ no longer holds in general and thus statistical consistency is lost.

3. Loss Correction with Robust Loss Functions

In this section, we provide two novel conditions on loss functions so that the discrepancy between true and estimated transition matrices can be tolerated. We also present theoretical analysis on their robustness under noise. In addition, as a case study, we show that they can be applied to noisy complementary-label learning by replacing the estimated transition matrix with the complementary matrix.

3.1 Weighted Symmetric Condition

When $\widehat{\boldsymbol{T}}$ is an identity matrix and $\boldsymbol{T}$ is different from $\widehat{\boldsymbol{T}}$, i.e., in the situation of ordinary-label learning from corrupted samples, several studies have focused on the loss function that is robust to label noise [24]–[28]. One of the criteria for measuring the robustness of loss functions is *noise tolerance*. Let f^* be the global minimizer of $R(f; \ell)$. A loss function ℓ is said to be *noise tolerant* [24]–[26] if f^* is also the global minimizer of $\widetilde{R}(f; \ell)$. In the following, let us consider loss functions that satisfy this property. A loss function ℓ is said to be *symmetric* [25], [26] if it satisfies the following condition for some constant $C > 0$:

$$\sum_{i=1}^{K} \ell(i, f(x)) = C, \quad \forall x \in \mathcal{X}, \ \forall f. \tag{9}$$

It has been proved that the symmetric loss ℓ is noise tolerant if $\forall j \neq i$, $(\boldsymbol{T})_{ii} > (\boldsymbol{T})_{ij} = \frac{\eta}{K-1}$ for $\eta \in \left[0, \frac{K-1}{K}\right)$ (Ghosh et al. [26], Theorem 1). Furthermore, if $R(f^*; \ell) = 0$ and $0 \leq \ell(i, f(x)) \leq \frac{C}{K-1}$, $\forall i \in \mathcal{Y}$, ℓ is noise tolerant when $(\boldsymbol{T})_{ii} > (\boldsymbol{T})_{ij}$, $\forall j \neq i$ (Ghosh et al. [26], Theorem 3).

Recall that our goal is to achieve robust learning using a corrected loss, i.e., to learn f^* by minimizing $\widetilde{R}(f; \ell^{\leftarrow})$ or $\widetilde{R}(f; \ell^{\rightarrow})$. In our work, we only consider backward correction. This is because backward correction allows us to choose arbitrary losses, and thus there is a large potential to apply existing losses. Taking a cue from Ghosh et al. [26], let us consider the following condition of ℓ for the backward corrected loss to be symmetric: given a *weight vector* $\boldsymbol{w} \in \mathbb{R}^K$ that satisfies $w_i > 0, \forall i = 1, \ldots, K$, and $\sum_{i=1}^{K} w_i = K$, it holds that for some constant $C > 0$,

$$\sum_{i=1}^{K} w_i \ell(i, f(x)) = C, \quad \forall x \in \mathcal{X}, \ \forall f. \tag{10}$$

Since the weighted sum is a constant, the loss function ℓ is said to be *weighted symmetric*. The following lemma shows the relationship between the weighted sum of the loss and the sum of the backward corrected loss (see Appx. A.1 for its proof):

Lemma 1. *Let* $\boldsymbol{w} = \left(\sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{iK}\right)^{\top}$ *be the weight vector. Suppose* $w_j > 0$, $\forall j = 1, \ldots, K$, *is satisfied and the loss function* ℓ *is weighted symmetric given* $\boldsymbol{w}$*. Then, the backward corrected loss* $\ell^{\leftarrow}$ *is symmetric.*

Next, we derive conditions for robust learning in the presence of an estimation error of the transition matrix. Let f^* be the global minimizer of $R(f; \ell)$ and $\widetilde{f}^*$ be the global minimizer of $\widetilde{R}(f; \ell^{\leftarrow})$. In the following theorems, we prove that the weighted symmetric loss can be used to obtain an optimal classifier when certain conditions hold for the estimation error of the transition matrix (see Appxs. A.2 and A.3 for their proofs):

Theorem 1. *Let* $\boldsymbol{w} = \left(\sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{iK}\right)^{\top}$ *be the weight vector. Suppose* $w_j > 0$, $\forall j = 1, \ldots, K$, *is satisfied and the loss function* ℓ *is weighted symmetric given* $\boldsymbol{w}$*. Also, suppose* $\forall j \neq i$, $(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ij} = \frac{\eta}{K-1}$ *for* $\eta < \frac{K-1}{K}$*. Then,* f^* *is also the minimizer of* $\widetilde{R}(f; \ell^{\leftarrow})$*.*

Theorem 2. *Let* $\boldsymbol{w} = \left(\sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{iK}\right)^{\top}$ *be the weight vector. Suppose* $w_j > 0$, $\forall j = 1, \ldots, K$, *is satisfied and the loss function* ℓ *is weighted symmetric given* $\boldsymbol{w}$*. Also, suppose* $w_j(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ii} > w_i(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ij}$, $\forall j \neq i$*. If* $R(f^*; \ell) = 0$ *and* $\ell(i, \boldsymbol{u}) = 0$ *implies* $\ell(j, \boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbb{R}^K} \ell(j, \boldsymbol{v})$ $(< \infty)$, $\forall j \neq i$, $\forall \boldsymbol{u} \in \mathbb{R}^K$*, then* f^* *is also the minimizer of* $\widetilde{R}(f; \ell^{\leftarrow})$*.*

The main difference between Theorems 1 and 2 is the relationship between the estimated transition matrix $\widehat{\boldsymbol{T}}$ and the ground-truth transition matrix $\boldsymbol{T}$. Theorem 1 imposes the constraint $(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ij} = \frac{\eta}{K-1}$, $\forall j \neq i$. Note that it is possible to rewrite this constraint as $\boldsymbol{T} = \widehat{\boldsymbol{T}}\boldsymbol{U}$, where $\boldsymbol{U} \in \mathbb{R}^{K \times K}$

is a matrix that takes $1-\eta$ on the diagonals and $\frac{\eta}{K-1}$ on the non-diagonals. From this, we can intuitively see that this theorem refers to the case where $\boldsymbol{T}$ is "symmetrically" shifted when viewed from $\widehat{\boldsymbol{T}}$. Theorem 2, on the other hand, imposes the constraint $w_j(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ii} > w_i(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ij}$, $\forall j \neq i$. If we further assume $\widehat{\boldsymbol{T}} = \boldsymbol{T}$ on top of this constraint, we can obtain $w_j > 0$, $\forall j \neq i$, which always holds. With $\widehat{\boldsymbol{T}} = \boldsymbol{T}$ as the starting point, this theorem shows that it does not matter if $\boldsymbol{T}$ is shifted "asymmetrically" when viewed from $\widehat{\boldsymbol{T}}$, as long as the degree of shift is within a certain range. Compared to Theorem 1, Theorem 2 allows more flexibility in dealing with the discrepancy between matrices. However, the conditions on the global minimum of the risk and the loss function might be restrictive in practical cases.

3.2 Relaxation of Weighted Symmetric Condition

Theorems 1 and 2 showed that weighted symmetric losses can be robust against the estimation error of the transition matrix. However, it is unclear whether the network parameters can be learned stably when using these losses to perform empirical risk minimization by a stochastic optimization algorithm. Indeed, Zhang et al. [27] demonstrated that the mean absolute error (MAE), an example of the symmetric loss (although not weighted symmetric in general), suffers from instability in training DNNs because of gradient saturation. This issue can occur in the weighted symmetric condition as well, and thus *more relaxed conditions* are needed so that a wider range of losses can be selected.

Here let us consider the condition that the weighted sum of losses is bounded, i.e., given a weight vector $\boldsymbol{w} \in \mathbb{R}^K$ that satisfies $w_i > 0, \forall i = 1, \ldots, K$, and $\sum_{i=1}^K w_i = K$, it holds that for some constants $C_1 > 0, C_2 > 0$,

$$C_1 \leq \sum_{i=1}^{K} w_i \ell(i, f(x)) \leq C_2, \quad \forall x \in \mathcal{X},\ \forall f. \tag{11}$$

Under this condition, we can derive the following theorems (see Appxs. A.4 and A.5 for their proofs):

Theorem 3. *Let* $\boldsymbol{w} = \left(\sum_{i=1}^K (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^K (\widehat{\boldsymbol{T}}^{-1})_{iK}\right)^\top$ *be the weight vector. Suppose* $w_j > 0$, $\forall j = 1, \ldots, K$, *is satisfied and the weighted sum of* ℓ *is bounded as* $C_1 \leq \sum_{i=1}^K w_i \ell(i, f(x)) \leq C_2$. *Also, suppose* $\forall j \neq i$, $(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ij} = \frac{\eta}{K-1}$ *for* $\eta < \frac{K-1}{K}$. *Then, the following inequality holds:*

$$(0 \leq)\ \widetilde{R}(f^*; \ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*; \ell^{\leftarrow}) \leq \frac{|\eta|}{K-1}(C_2 - C_1). \tag{12}$$

Theorem 4. *Let* $\boldsymbol{w} = \left(\sum_{i=1}^K (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^K (\widehat{\boldsymbol{T}}^{-1})_{iK}\right)^\top$ *be the weight vector. Suppose* $w_j > 0$, $\forall j = 1, \ldots, K$, *is satisfied and the weighted sum of* ℓ *is bounded as* $C_1 \leq \sum_{i=1}^K w_i \ell(i, f(x)) \leq C_2$. *Also, suppose* $w_j(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ii} > w_i(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ij}$, $\forall j \neq i$. *If* $R(f^*; \ell) = 0$ *and* $\ell(i, \boldsymbol{u}) = 0$ *implies* $\ell(j, \boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbb{R}^K} \ell(j, \boldsymbol{v})$ $(< \infty)$, $\forall j \neq i$, $\forall \boldsymbol{u} \in \mathbb{R}^K$, *then the following inequality holds:*

$$(0 \leq)\ \widetilde{R}(f^*; \ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*; \ell^{\leftarrow})$$
$$\leq \mathbb{E}_{\mathcal{D}}\left[\left|(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right|\right](C_2 - C_1). \tag{13}$$

Theorems 3 and 4 show that the difference $\widetilde{R}(f^*; \ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*; \ell^{\leftarrow})$ is bounded. As we can see from Ineqs. (12) and (13), the closer $C_2 - C_1$ is to zero, the tighter the bound becomes.

3.3 Robustness Conditions for Learning from Noisy Complementary Labels

By substituting the complementary matrix for $\widehat{\boldsymbol{T}}$ in the theorems introduced earlier, the following corollary can be derived.

Corollary 1 (Noisy complementary-label learning)**.** *Suppose* $\widehat{\boldsymbol{T}}$ *is the* $K \times K$ *complementary matrix, i.e.,* $\widehat{\boldsymbol{T}} = \overline{\boldsymbol{T}}$. *Then, the following four properties hold:*

(i) *Let* ℓ *be a symmetric loss. Suppose* $\forall j \neq i$, $(\boldsymbol{T})_{ii} < (\boldsymbol{T})_{ij} = \frac{K-1-\eta}{(K-1)^2}$ *for* $\eta \in \left[0, \frac{K-1}{K}\right)$. *Then,* f^* *is also the minimizer of* $\widetilde{R}(f; \ell^{\leftarrow})$.

(ii) *Let* ℓ *be a symmetric loss. Suppose* $(\boldsymbol{T})_{ii} < (\boldsymbol{T})_{ij}$, $\forall j \neq i$. *If* $R(f^*; \ell) = 0$ *and* $\ell(i, \boldsymbol{u}) = 0$ *implies* $\ell(j, \boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbb{R}^K} \ell(j, \boldsymbol{v})$ $(< \infty)$, $\forall j \neq i$, $\forall \boldsymbol{u} \in \mathbb{R}^K$, *then* f^* *is also the minimizer of* $\widetilde{R}(f; \ell^{\leftarrow})$.

(iii) *Suppose the sum of* ℓ *is bounded as* $C_1 \leq \sum_{i=1}^K \ell(i, f(x)) \leq C_2$. *Also, suppose* $\forall j \neq i$, $(\boldsymbol{T})_{ii} < (\boldsymbol{T})_{ij} = \frac{K-1-\eta}{(K-1)^2}$ *for* $\eta \in \left[0, \frac{K-1}{K}\right)$. *Then, the following inequality holds:*

$$(0 \leq)\ \widetilde{R}(f^*; \ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*; \ell^{\leftarrow}) \leq \frac{\eta}{K-1}(C_2 - C_1). \tag{14}$$

(iv) *Suppose the sum of* ℓ *is bounded as* $C_1 \leq \sum_{i=1}^K \ell(i, f(x)) \leq C_2$. *Also, suppose* $(\boldsymbol{T})_{ii} < (\boldsymbol{T})_{ij}$, $\forall j \neq i$. *If* $R(f^*; \ell) = 0$ *and* $\ell(i, \boldsymbol{u}) = 0$ *implies* $\ell(j, \boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbb{R}^K} \ell(j, \boldsymbol{v})$ $(< \infty)$, $\forall j \neq i$, $\forall \boldsymbol{u} \in \mathbb{R}^K$, *then the following inequality holds:*

$$(0 \leq)\ \widetilde{R}(f^*; \ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*; \ell^{\leftarrow})$$
$$\leq \mathbb{E}_{\mathcal{D}}\left[1 - (K-1)(\boldsymbol{T})_{YY}\right](C_2 - C_1). \tag{15}$$

Given that $\widehat{\boldsymbol{T}}$ is the complementary matrix, i.e., $\widehat{\boldsymbol{T}} = \overline{\boldsymbol{T}}$, we can rewrite the constraints on $\boldsymbol{T}$ as above. In addition, since it holds that $\sum_{i=1}^K (\overline{\boldsymbol{T}}^{-1})_{ij} = 1$, $\forall j = 1, \ldots, K$, we can use symmetric losses instead of weighted symmetric losses. Ineqs. (14) and (15) are the rewritten forms from Ineqs. (12) and (13), respectively. The absolute values disappeared in Ineqs. (14) and (15) because both η and $1-(K-1)(\boldsymbol{T})_{YY}$ are non-negative. Overall, Corollary 1 points out that symmetric losses (or losses with more relaxed conditions) make the corrected loss less sensitive to label noise in complementary labels as long as the diagonals of $\boldsymbol{T}$ is smaller than the non-diagonals, i.e., the proportion of noisy complementary labels is relatively small.

4. Experiments

In this section, we show some experimental results that evaluate the performance on noisy complementary-label learning.

4.1 Experimental Setup

Datasets and Network Architectures. We used four well-known benchmark datasets: MNIST [29], Fashion-MNIST [30], Kuzushiji-MNIST [31], and CIFAR-10 [11]. For training, we used only noisy training samples that were artificially corrupted by $\boldsymbol{T}$. The test data were not affected by label noise and were all assumed to have true labels. For MNIST, Fashion-MNIST, and Kuzushiji-MNIST, a multilayer perceptron (MLP) model ($d-500-K$) was used. For CIFAR-10, we used the network architecture that contains 4 convolutional layers with 32 filters whose filter size is 3×3. All the convolutional layers were followed by 2×2 max pooling layers, except for the last convolutional layer, which was followed by a global average pooling layer [32]. Batch normalization [33] was adopted right after each convolution, followed by the rectified linear unit (ReLU) activation.

Robust Loss Functions. We used five different loss functions in the backward correction procedure. All loss functions receive the softmax outputs of the model. Let $F_i(\boldsymbol{v}) = \exp(v_i)/\sum_{j=1}^{K}\exp(v_j)$ be the softmax function. The definitions and parameter settings of each loss function are listed as follows.

- Categorical Cross Entropy (CCE):
$$\ell_{\mathrm{cce}}(y, f(x)) = -\log(F_y(f(x))). \tag{16}$$
- Mean Absolute Error (MAE) [26]:
$$\ell_{\mathrm{mae}}(y, f(x)) = 2 - 2F_y(f(x)). \tag{17}$$
- Weighted Mean Absolute Error (WMAE): Let $\boldsymbol{w}$ be the weight vector. Suppose $w_i > 0$, $\forall i = 1,\dots,K$, and $\sum_{i=1}^{K} w_i = K$. WMAE is defined as follows:
$$\ell_{\mathrm{wmae}}(y, f(x)) = \frac{1}{w_y}(2 - 2F_y(f(x))). \tag{18}$$
In all experiments, we set $w_y = \sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{iy}$, $\forall y = 1,\dots,K$.
- Generalized Cross Entropy (GCE) [27]:
$$\ell_{\mathrm{gce}}(y, f(x)) = \frac{(1 - F_y(f(x))^q)}{q}, \tag{19}$$
where $q \in (0, 1]$ is a hyper-parameter. In all experiments, we set $q = 0.7$.
- Symmetric Cross Entropy (SL) [28]: SL is composed by adding an extra term called *reverse cross entropy* (RCE) to CCE. RCE is defined as
$$\ell_{\mathrm{rce}}(y, f(x)) = -\sum_{i\neq y} AF_i(f(x)), \tag{20}$$
where $A < 0$. Using this, the SL loss is defined as
$$\ell_{\mathrm{sl}}(y, f(x)) = \alpha\ell_{\mathrm{cce}}(y, f(x)) + \beta\ell_{\mathrm{rce}}(y, f(x)), \tag{21}$$
where α and β are hyper-parameters. In all experiments, we set $A = -4$. For MNIST, Fashion-MNIST, and Kuzushiji-MNIST, we set $\alpha = 0.01, \beta = 1.0$. For CIFAR-10, we set $\alpha = 0.1, \beta = 1.0$.

Table 1 Summary of loss functions and their properties. Robustness to noise is determined by whether each loss can have the property of absorbing the estimation error of the transition matrix compared to CCE. $*$ indicates that it satisfies the condition only if the weights are all 1.

Losses	(10)	(11)	Upper-bounded	Robust to noise?
CCE	×	×	×	×
MAE	$*$	✓	✓	✓
WMAE	✓	✓	✓	✓
GCE	×	✓	✓	✓
SL	×	×	×	✓

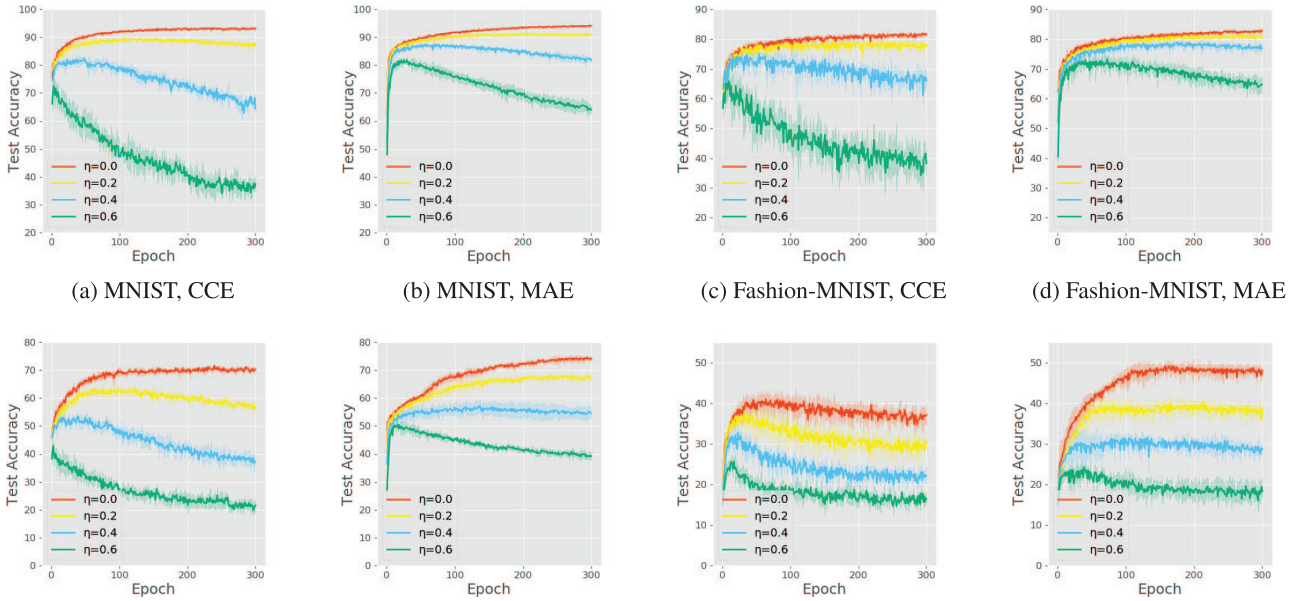
Since CCE and SL are not upper-bounded, they do not satisfy both the weighted symmetric condition (10) and the more relaxed condition (11). Nevertheless, the RCE term satisfies (11), and thus SL can be more robust than the original CCE. MAE and GCE satisfy (11). WMAE, which can be regarded as an extension of MAE, satisfies both (10) and (11). Table 1 summarizes the properties of each loss function. The bounds for their weighted sums are provided in Appx. B.

Noise Settings. Suppose $\boldsymbol{T} = \widehat{\boldsymbol{T}}\boldsymbol{U}$, where $\boldsymbol{U}$ is the matrix that compensates for the discrepancy between $\widehat{\boldsymbol{T}}$ and $\boldsymbol{T}$. Then, we can adjust the noise setting by determining the matrix $\boldsymbol{U}$ appropriately. Let η be a noise rate. In the experiments, the following matrices were used.

- Symmetric noise setting:
$$\boldsymbol{U}(\eta) = \begin{pmatrix} 1-\eta & & & \frac{\eta}{K-1} \\ & 1-\eta & & \\ & & \ddots & \\ \frac{\eta}{K-1} & & & 1-\eta \end{pmatrix}. \tag{22}$$
- Asymmetric noise setting:
$$\boldsymbol{U}(\eta) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-\eta & 0 & 0 & 0 & 0 & \eta & 0 & 0 \\ 0 & 0 & 0 & 1-\eta & 0 & 0 & 0 & 0 & \eta & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-\eta & \eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \eta & 1-\eta & 0 & 0 & 0 \\ 0 & \eta & 0 & 0 & 0 & 0 & 0 & 1-\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{23}$$

Implementation with Gradient Ascent.

Compared with forward correction, backward correction has the benefit of having less constraints on losses and models. However, when calculated with limited data and flexible models such as DNNs, the backward corrected risk can go negative [5], [6], [21] (which does not occur with forward correction), and thus may cause an overfitting issue. This phenomenon occurs because the matrix $\widehat{\boldsymbol{T}}^{-1}$ can contain nagative elements, i.e., in training, nagative terms in (3) are well below zero and non-negative terms approach



(a) MNIST, CCE (b) MNIST, MAE (c) Fashion-MNIST, CCE (d) Fashion-MNIST, MAE

(e) Kuzushiji-MNIST, CCE (f) Kuzushiji-MNIST, MAE (g) CIFAR-10, CCE (h) CIFAR-10, MAE

Fig. 2 Experimental results for symmetric noise setting with different noise rates (η = 0.0, 0.2, 0.4, 0.6). Dark and light colors show the mean accuracy and the standard deviation of 4 trials, respectively.

zero. Similar issues have occurred in other problem settings such as *positive-unlabeled learning* (e.g., Kiryo et al. [35]) and *unlabeled-unlabeled learning* (e.g., Lu et al. [36]), all of which rewrite the original risk in different ways. Furthermore, it has been pointed out that the degree of overfitting differs depending on whether the losses are bounded or not [5], [8], [35], making a fair comparison across losses difficult. To avoid these issues, we perform an optimization using the *gradient ascent* technique, an extention of the method proposed by Ishida et al. [5]. Let $\widetilde{\pi}_j \coloneqq \mathbb{P}(\widetilde{Y} = j)$ and $\widetilde{P}_j \coloneqq \mathbb{P}(X|\widetilde{Y} = j)$. By using them, we can rewrite $\widetilde{R}(f; \ell^{\leftarrow})$ as

$$\widetilde{R}(f; \ell^{\leftarrow}) = \underbrace{\sum_{k=1}^{K} \sum_{j=1}^{K} \widetilde{\pi}_j (\widehat{\boldsymbol{T}}^{-1})_{jk} \mathbb{E}_{\widetilde{P}_j} \left[\ell(k, f(X))\right]}_{\text{(A)}}. \tag{24}$$

Note that if $\widehat{\boldsymbol{T}} = \boldsymbol{T}$, it is also possible to rewrite (A) as $\mathbb{E}_{\mathcal{D}} \left[\mathbb{1}_{[Y=k]} \ell(k, f(X))\right] (\geq 0)$, where $\mathbb{1}_{[\cdot]}$ is the indicator function. Since (A) is non-negative when $\widehat{\boldsymbol{T}}$ is correctly estimated, the empirical version of (A) should not be much less than zero during training. As an idea to make the empirical risk non-negative, when an estimate of (A) is below a threshold, we perform a gradient ascent instead of the usual gradient descent. The procedure of the stochastic optimization with mini-batches is shown in detail in Algorithm 1. In all experiments, we fixed $\beta = 0$ and $\gamma = 1$ for simplicity. Adam [34] was used for optimization and the classifier was trained for 300 epochs with mini-batch size of 256.

Algorithm 1 Learning with backward corrected loss and gradient ascent (an extension of Algorithm 1 in Ishida et al. [5])

Input: noisy labeled training data $\{\mathcal{X}_k\}_{k=1}^{K}$, where $\mathcal{X}_k$ denotes the samples noisy labeled as class k;
Output: model parameter $\boldsymbol{\theta}$ for $f(x; \boldsymbol{\theta})$

Let $\mathcal{A}$ be an external SGD-like stochastic optimization algorithm such as [34];
while no stopping criterion has been met **do**
  Shuffle $\{\mathcal{X}_j\}_{j=1}^{K}$ into B mini-batches;
  for b = 1 to B **do**
    Denote $\{\mathcal{X}_j^b\}$ as the b-th mini-batch for noisy class j;
    Denote $r_k^b(\boldsymbol{\theta}) = \sum_{j=1}^{K} \widehat{\pi}_j (\widehat{\boldsymbol{T}}^{-1})_{jk} \widehat{\mathbb{E}}_{\widetilde{P}_j} \left[\ell(k, f); \mathcal{X}_j^b\right]$;
    if $\min_k [r_1^b(\boldsymbol{\theta}), \ldots, r_k^b(\boldsymbol{\theta}), \ldots, r_K^b(\boldsymbol{\theta})] > -\beta$ **then**
      Denote $L^b(\boldsymbol{\theta}) = \sum_{k=1}^{K} r_k^b(\boldsymbol{\theta})$;
      Set gradient $\nabla_{\boldsymbol{\theta}} L^b(\boldsymbol{\theta})$;
      Update $\boldsymbol{\theta}$ by $\mathcal{A}$ with its current step size ϵ;
    else
      Denote $\overline{L}^b(\boldsymbol{\theta}) = \sum_{k=1}^{K} \min\{-\beta, r_k^b(\boldsymbol{\theta})\}$;
      Set gradient $-\nabla_{\boldsymbol{\theta}} \overline{L}^b(\boldsymbol{\theta})$;
      Update $\boldsymbol{\theta}$ by $\mathcal{A}$ with a discounted step size $\gamma\epsilon$;
    end if
  end for
end while

4.2 Results

Test Accuracy for Each Epoch. We set $\widehat{\boldsymbol{T}}$ to the complementary matrix $\overline{\boldsymbol{T}}$ and demonstrated noisy complementary-label learning with symmetric noise. The models were trained with CCE and MAE on benchmark datasets following Algorithm 1 and the learning rate was set to 10^{-4}. Note that the roles of MAE and WMAE are exactly the same because all weights $w_j = \sum_{i=1}^{K} (\overline{\boldsymbol{T}}^{-1})_{ij}$ are 1 in this problem setting. Figure 2 shows the mean and standard deviation of the test accuracy for 4 trials on test data. We can see from Fig. 2 that MAE works significantly better than CCE under

Table 2 Mean test accuracy (4 trials) of different methods on benchmark datasets under symmetric and asymmetric noise settings. All methods use the backward correction by complementary matrix together with gradient ascent technique, and differ only in the loss function. Bold face denotes the best and comparable methods according to the paired t-test at the significance level 5%.

Datasets	Losses	Symmetric Noise			Asymmetric Noise	
		Noise Rate η			Noise Rate η	
		0.0	0.2	0.6	0.2	0.4
MNIST	CCE	89.10 ± 0.36	85.65 ± 1.15	72.31 ± 3.37	84.06 ± 3.17	75.11 ± 0.79
	MAE	**93.05 ± 0.39**	**90.36 ± 0.82**	**80.24 ± 0.93**	**92.02 ± 0.50**	**86.41 ± 1.74**
	GCE	**92.83 ± 1.03**	**90.83 ± 0.81**	78.01 ± 1.86	**91.80 ± 1.05**	**86.02 ± 1.31**
	SL	91.56 ± 0.50	89.34 ± 0.53	**81.08 ± 0.75**	89.81 ± 1.51	**86.82 ± 1.26**
Fashion MNIST	CCE	79.77 ± 0.68	77.95 ± 1.13	68.69 ± 1.10	78.06 ± 0.51	72.75 ± 1.91
	MAE	**81.66 ± 1.13**	**81.12 ± 0.34**	**73.05 ± 1.90**	**81.58 ± 0.38**	**77.56 ± 0.74**
	GCE	**81.50 ± 0.87**	**80.53 ± 0.35**	70.45 ± 1.35	**81.45 ± 0.52**	75.42 ± 1.72
	SL	**81.75 ± 0.49**	**80.31 ± 0.74**	67.06 ± 2.54	**80.96 ± 0.85**	**78.73 ± 1.65**
Kuzushiji MNIST	CCE	61.59 ± 2.48	57.88 ± 0.94	40.95 ± 4.55	59.48 ± 4.00	52.25 ± 2.89
	MAE	**70.53 ± 1.37**	**64.46 ± 1.70**	**48.62 ± 2.36**	61.02 ± 3.28	**59.74 ± 2.44**
	GCE	63.91 ± 2.70	**64.60 ± 1.52**	41.31 ± 5.01	63.50 ± 3.08	**58.46 ± 2.68**
	SL	**70.74 ± 2.97**	**65.89 ± 2.49**	**46.54 ± 6.67**	**68.41 ± 1.25**	**59.95 ± 2.00**
CIFAR-10	CCE	44.78 ± 1.62	39.11 ± 1.62	17.29 ± 1.18	40.58 ± 2.96	33.28 ± 2.68
	MAE	**52.06 ± 0.94**	**45.01 ± 1.06**	23.02 ± 4.97	**48.53 ± 0.83**	**40.12 ± 2.02**
	GCE	50.52 ± 1.49	43.62 ± 0.81	**25.91 ± 3.11**	46.35 ± 1.17	**39.81 ± 2.34**
	SL	**52.04 ± 0.27**	**44.23 ± 1.63**	18.26 ± 0.93	**48.41 ± 1.11**	**41.78 ± 2.35**

Table 3 Mean test accuracy (4 trials) of different methods on benchmark datasets under symmetric and asymmetric noise settings. All methods use the backward correction by the matrix given in (25) together with gradient ascent technique, and differ only in the loss function. Bold face denotes the best and comparable methods according to the paired t-test at the significance level 5%.

Datasets	Losses	Symmetric Noise			Asymmetric Noise	
		Noise Rate η			Noise Rate η	
		0.0	0.2	0.6	0.2	0.4
MNIST	CCE	84.27 ± 0.48	80.31 ± 2.18	60.26 ± 4.67	77.86 ± 2.70	64.06 ± 4.33
	MAE	**91.21 ± 0.07**	86.80 ± 1.28	70.73 ± 5.67	**89.40 ± 1.05**	82.01 ± 2.40
	WMAE	**91.66 ± 0.68**	**88.81 ± 0.84**	**76.42 ± 1.69**	**89.54 ± 0.89**	**85.83 ± 1.36**
Fashion MNIST	CCE	72.11 ± 3.11	72.75 ± 1.49	62.68 ± 1.34	74.38 ± 2.37	64.95 ± 3.20
	MAE	**80.62 ± 1.06**	78.74 ± 0.83	**70.35 ± 1.20**	**80.10 ± 0.50**	**75.71 ± 2.33**
	WMAE	**81.59 ± 0.24**	**80.06 ± 0.59**	**71.24 ± 2.31**	**80.34 ± 0.49**	**76.47 ± 1.08**
Kuzushiji MNIST	CCE	54.51 ± 3.13	46.39 ± 2.66	38.11 ± 2.84	47.00 ± 2.20	41.97 ± 1.77
	MAE	**58.15 ± 0.54**	**55.43 ± 2.33**	**44.90 ± 0.72**	**58.40 ± 1.54**	**57.77 ± 1.57**
	WMAE	**57.69 ± 2.01**	**56.23 ± 3.71**	**45.13 ± 2.52**	**59.23 ± 1.25**	**57.64 ± 1.46**
CIFAR-10	CCE	39.75 ± 0.81	32.46 ± 1.49	16.14 ± 2.50	34.45 ± 4.57	27.07 ± 3.31
	MAE	46.64 ± 1.60	**42.22 ± 1.47**	**20.59 ± 1.45**	**44.74 ± 1.51**	37.37 ± 2.02
	WMAE	**50.19 ± 1.35**	**41.01 ± 1.39**	**21.43 ± 4.44**	**45.27 ± 1.37**	**39.56 ± 1.30**

symmetric noise. In the case of MNIST, Fashion-MNIST, and Kuzushiji-MNIST, test accuracies of CCE and MAE under noise-free conditions are almost the same. Despite this, as the noise rate increases, CCE becomes more sensitive to noise than MAE. These results support the properties derived in the previous section (especially Corollary 1). In the case of CIFAR-10, CCE suffers from overfitting even in noise-free situations due to the lack of an upper bound. As a result, MAE provides better performance, even though it is known to give worse results for DNNs on challenging datasets [27].

Performance Comparison. We first set $\widehat{\boldsymbol{T}}$ to the complementary matrix $\overline{\boldsymbol{T}}$ and compared the loss functions introduced in Sect. 4.1. During the training, 10% of the original training data was reserved for validation. The models were trained following Algorithm 1 and the learning rate was selected from $\{10^{-2}, 10^{-3}, \ldots, 10^{-6}\}$ so that the validation loss is minimized. Table 2 shows the experimental results for different loss functions on MNIST, Fashion-MNIST, Kuzushiji-MNIST, and CIFAR-10. MAE, GCE and SL performed better than CCE under both symmetric noise and asymmetric noise, which is consistent with our theoretical analysis. Overall, MAE in particular seems to work better than other losses. Note that our experiments exploy optimization using the gradient ascent technique for the backward corrected losses, i.e., the theoretical guarantees derived in the previous section might be missing. As an ablation study, we trained them without gradient ascent. Although GCE and SL still outperformed CCE, performance degradation was observed especially in CCE and MAE. Detailed results are shown in Appx. C.1. We also tried to apply forward correction for these losses. Detailed results are shown in Appx. C.2. In the case of forward correction, the bene-

fits of learning with the losses other than CCE seems to be small, especially under asymmetric noise.

Additionally, in order to achieve a comparison between WMAE and MAE, we conducted experiments using a matrix other than the complementary matrix. We set $\widehat{\boldsymbol{T}}$ as follows:

$$\widehat{\boldsymbol{T}} = \begin{pmatrix} 0.01 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.00 & 0.11 & 0.12 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.02 & 0.11 & 0.11 & 0.11 & 0.10 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.12 & 0.00 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.01 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.01 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.00 & 0.11 & 0.12 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.09 & 0.11 & 0.11 & 0.03 & 0.11 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.01 & 0.11 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.01 \end{pmatrix}. \quad (25)$$

$\widehat{\boldsymbol{T}}$ can be seen as a complementary matrix with some noise. The weight vector can be calculated as

$$\begin{aligned} \boldsymbol{w} &= \left(\sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{i1}, \ldots, \sum_{i=1}^{K} (\widehat{\boldsymbol{T}}^{-1})_{iK} \right)^{\top} \\ &= (1.000, 0.909, 1.221, 0.992, 0.750, \\ &\quad 1.000, 0.798, 1.250, 1.080, 1.000)^{\top}, \end{aligned} \quad (26)$$

and therefore we can construct WMAE separately from MAE. The experimental results are shown in Table 3. From the results, we can see that WMAE and MAE are both superior for many datasets, and that WMAE performs slightly better than MAE.

5. Conclusion

In this paper, we discussed the problem setting where complementary labels may be affected by label noise. We chose backward correction as the learning algorithm for complementary labels and showed that noise in complementary labels can be interpreted as an estimation error of the transition matrix. To mitigate the adverse effects of it, we obtained noise robustness by selecting losses which satisfy the weighted symmetric condition or a more relaxed condition. It was experimentally shown that losses that satisfy our conditions work better than those that do not.

Acknowledgments

TI was supported by JSPS KAKENHI 20J11937. MS was supported by KAKENHI 20H04206.

References

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp.1097–1105, Curran Associates, Inc., 2012.
[2] J. Howe, Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, 1 ed., Crown Publishing Group, USA, 2008.
[3] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," Advances in Neural Information Processing Systems, pp.5639–5649, Curran Associates, Inc., 2017.
[4] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," Proc. European Conference on Computer Vision, 2018.
[5] T. Ishida, G. Niu, A. Menon, and M. Sugiyama, "Complementary-label learning for arbitrary losses and models," Proc. 36th International Conference on Machine Learning, pp.2971–2980, PMLR, 2019.
[6] Y.T. Chou, G. Niu, H.T. Lin, and M. Sugiyama, "Unbiased risk estimators can mislead: A case study of learning with complementary labels," Proc. 37th International Conference on Machine Learning, pp.1929–1938, PMLR, 2020.
[7] T. Zhang, "Statistical analysis of some multi-category large margin classification methods," J. Mach. Learn. Res., vol.5, pp.1225–1251, 2004.
[8] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," Proc. 37th International Conference on Machine Learning, pp.3072–3081, PMLR, 2020.
[9] Y. Xu, M. Gong, J. Chen, T. Liu, K. Zhang, and K. Batmanghelich, "Generative-discriminative complementary learning," Proc. Thirty-Fourth AAAI Conference on Artificial Intelligence, pp.6526–6533, 2020.
[10] Y. Kim, J. Yim, J. Yun, and J. Kim, "NLNL: Negative learning for noisy labels," IEEE International Conference on Computer Vision, pp.101–110, 2019.
[11] A. Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, University of Toronto, 2009.
[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," International Conference on Learning Representations, pp.1–15, 2017.
[13] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," Proc. 34th International Conference on Machine Learning, pp.233–242, PMLR, 2017.
[14] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," IEEE Trans. Pattern Anal. Mach. Intell., vol.38, no.3, pp.447–461, 2016.
[15] G. Patrini, A. Rozza, A.K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," IEEE Conference on Computer Vision and Pattern Recognition, pp.2233–2241, 2017.
[16] B. Han, J. Yao, G. Niu, M. Zhou, I. Tsang, Y. Zhang, and M. Sugiyama, "Masking: A new perspective of noisy supervision," Advances in Neural Information Processing Systems, pp.5836–5846, Curran Associates, Inc., 2018.
[17] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," Advances in Neural Information Processing Systems, pp.10456–10465, Curran Associates, Inc., 2018.
[18] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama, "Are anchor points really indispensable in label-noise learning?," Advances in Neural Information Processing Systems, pp.6838–6849, Curran Associates, Inc., 2019.
[19] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama, "Dual T: Reducing estimation error for transition matrix in label-noise learning," Advances in Neural Information Processing Systems, pp.7260–7271, Curran Associates, Inc., 2020.
[20] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama, "Part-dependent label noise: Towards instance-dependent label noise," Advances in Neural Information Processing Systems, pp.7597–7610, Curran Associates, Inc., 2020.
[21] B. Han, G. Niu, X. Yu, Q. Yao, M. Xu, I. Tsang, and M. Sugiyama, "SIGUA: Forgetting may make learning with noisy labels more robust," Proc. 37th International Conference on Machine Learning, pp.4006–4016, PMLR, 2020.
[22] M.D. Reid and R.C. Williamson, "Composite binary losses," J.

Mach. Learn. Res., vol.11, pp.2387–2422, 2010.
[23] R.C. Williamson, E. Vernet, and M.D. Reid, "Composite multiclass losses," J. Mach. Learn. Res., vol.17, no.1, pp.7860–7911, 2016.
[24] N. Manwani and P.S. Sastry, "Noise tolerance under risk minimization," IEEE Trans. Cybern., vol.43, no.3, pp.1146–1151, 2013.
[25] A. Ghosh, N. Manwani, and P.S. Sastry, "Making risk minimization tolerant to label noise," Neurocomputing, vol.160, pp.93–107, 2015.
[26] A. Ghosh, H. Kumar, and P.S. Sastry, "Robust loss functions under label noise for deep neural networks," Proc. Thirty-First AAAI Conference on Artificial Intelligence, pp.1919–1925, 2017.
[27] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," Advances in Neural Information Processing Systems, pp.8778–8788, Curran Associates, Inc., 2018.
[28] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," IEEE International Conference on Computer Vision, pp.322–330, 2019.
[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol.86, no.11, pp.2278–2324, 1998.
[30] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.
[31] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," Neural Information Processing Systems Workshop on Machine Learning for Creativity and Design, pp.1–8, 2018.
[32] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. 32nd International Conference on Machine Learning, pp.448–456, PMLR, 2015.
[34] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, pp.1–15, 2015.
[35] R. Kiryo, G. Niu, M.C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," Advances in Neural Information Processing Systems, pp.1675–1685, Curran Associates, Inc., 2017.
[36] N. Lu, T. Zhang, G. Niu, and M. Sugiyama, "Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach," International Conference on Artificial Intelligence and Statistics, pp.1115–1125, PMLR, 2020.

Appendix A: Proofs

A.1 Proof of Lemma 1

Proof. Using $\sum_{k=1}^{K}(\widehat{\boldsymbol{T}})_{jk} = 1, \forall j = 1, \ldots, K$, we obtain

$$\begin{aligned}\sum_{j=1}^{K} w_j &= \sum_{j=1}^{K}\sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{ij} = \sum_{j=1}^{K}\sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{ij}\sum_{k=1}^{K}(\widehat{\boldsymbol{T}})_{jk} \\ &= \sum_{i=1}^{K}\sum_{k=1}^{K}(\boldsymbol{I}_K)_{ik} = K,\end{aligned}$$

where $\boldsymbol{I}_K$ is a $K \times K$ identity matrix. In addition,

$$\begin{aligned}\sum_{i=1}^{K} w_i \ell(i, f(x)) &= \sum_{i=1}^{K}\sum_{j=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{ji}\ell(i, f(x)) \\ &= \sum_{j=1}^{K}\ell^{\leftarrow}(j, f(x)).\end{aligned}$$

Thus, if ℓ satisfies the definition of weighted symmetric, i.e., the weighted sum is a constant, then $\ell^{\leftarrow}$ becomes symmetric. □

A.2 Proof of Theorem 1

Proof. Let C be the weighted sum of ℓ. Then, we have

$$\begin{aligned}\widetilde{R}(f;\ell^{\leftarrow}) &= \mathbb{E}_{\widetilde{\mathcal{D}}}\left[\ell^{\leftarrow}(\widetilde{Y}, f(X))\right] = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\widetilde{Y}|Y}\left[\ell^{\leftarrow}(\widetilde{Y}, f(X))\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}(\boldsymbol{T})_{Yi}\ell^{\leftarrow}(i, f(X))\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ji}\ell^{\leftarrow}(i, f(X))\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\Big\{(1-\eta)\ell^{\leftarrow}(j, f(X)) \right. \\ &\quad \left. + \frac{\eta}{K-1}\sum_{i\neq j}\ell^{\leftarrow}(i, f(X))\Big\}\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\Big\{(1-\eta)\ell^{\leftarrow}(j, f(X)) \right. \\ &\quad \left. + \frac{\eta}{K-1}(C - \ell^{\leftarrow}(j, f(X)))\Big\}\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\left\{\left(1 - \frac{K\eta}{K-1}\right)\ell^{\leftarrow}(j, f(X)) + \frac{\eta C}{K-1}\right\}\right] \\ &= \left(1 - \frac{K\eta}{K-1}\right)\mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\ell^{\leftarrow}(j, f(X))\right] + \frac{\eta C}{K-1} \\ &= \left(1 - \frac{K\eta}{K-1}\right)\mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\widehat{\boldsymbol{T}})_{Yj}\sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{ji}\ell(i, f(X))\right] \\ &\quad + \frac{\eta C}{K-1} \\ &= \left(1 - \frac{K\eta}{K-1}\right)\mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}(\boldsymbol{I}_K)_{Yi}\ell(i, f(X))\right] + \frac{\eta C}{K-1} \\ &= \left(1 - \frac{K\eta}{K-1}\right)R(f;\ell) + \frac{\eta C}{K-1},\end{aligned}$$

where the fifth equality holds because $(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ji} = \frac{\eta}{K-1}, \forall i \neq j$ and $\sum_{i=1}^{K}(\widehat{\boldsymbol{T}}^{-1}\boldsymbol{T})_{ji} = 1, \forall j = 1, \ldots, K$; the sixth equality holds because of Lemma 1. Thus, for any f,

$$\begin{aligned}&\widetilde{R}(f^*;\ell^{\leftarrow}) - \widetilde{R}(f;\ell^{\leftarrow}) \\ &= \left(1 - \frac{K\eta}{K-1}\right)(R(f^*;\ell) - R(f;\ell)) \leq 0,\end{aligned}$$

because $\eta < \frac{K-1}{K}$ and f^* is the minimizer of $R(f;\ell)$. This proves f^* is also the minimizer of $\widetilde{R}(f;\ell^{\leftarrow})$. □

A.3 Proof of Theorem 2

Proof. Let C be the weighted sum of ℓ. Then, we have

$$\begin{aligned}
\widetilde{R}(f;\ell^{\leftarrow}) &= \mathbb{E}_{\widetilde{\mathcal{D}}}\left[\ell^{\leftarrow}(\widetilde{Y},f(X))\right] = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\widetilde{Y}|Y}\left[\ell^{\leftarrow}(\widetilde{Y},f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}(\boldsymbol{T})_{Yi}\ell^{\leftarrow}(i,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}(\boldsymbol{T})_{Yi}\sum_{j=1}^{K}(\widehat{\boldsymbol{T}}^{-1})_{ij}\ell(j,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{j=1}^{K}(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\ell(j,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\ell(Y,f(X)) + \sum_{j\neq Y}(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\ell(j,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\left(C - \sum_{j\neq Y}w_j\ell(j,f(X))\right)\right. \\
&\quad \left. + \sum_{j\neq Y}(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\ell(j,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{C}{w_Y}(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\right. \\
&\quad \left. - \sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\ell(j,f(X))\right].
\end{aligned}$$

As $\widetilde{f}^*$ is the minimizer of $\widetilde{R}(f;\ell^{\leftarrow})$, $\widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow}) - \widetilde{R}(f^*;\ell^{\leftarrow}) \leq 0$. Therefore from the above equality, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[-\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\right. \\
&\left.\left(\ell(j,\widetilde{f}^*(X)) - \ell(j,f^*(X))\right)\right] \leq 0.
\end{aligned}$$

Since we are given $R(f^*;\ell) = 0$, we have $\ell(Y,f^*(X)) = 0$. From this and the constraint on ℓ in the theorem, we have $\ell(j,f^*(X)) = \sup_{\boldsymbol{v}\in\mathbb{R}^K}\ell(j,\boldsymbol{v})$ $(<\infty)$, $\forall j \neq Y$. Moreover, since we are given $w_j(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ii} > w_i(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ij}$, $\forall j \neq i$, the above inequality holds iff $\ell(j,\widetilde{f}^*(X)) = \sup_{\boldsymbol{v}\in\mathbb{R}^K}\ell(j,\boldsymbol{v})$ $(<\infty)$, $\forall j \neq Y$, which means $\widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow}) = \widetilde{R}(f^*;\ell^{\leftarrow})$. This proves f^* is also the minimizer of $\widetilde{R}(f;\ell^{\leftarrow})$. □

A.4 Proof of Theorem 3

Proof. First, we consider the case where $0 \leq \eta < \frac{K-1}{K}$. In a similar way to the proof of Theorem 1, it can be shown that

$$\begin{aligned}
\widetilde{R}(f;\ell^{\leftarrow}) &= \left(1 - \frac{K\eta}{K-1}\right)R(f;\ell) \\
&\quad + \frac{\eta}{K-1}\mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{K}\ell^{\leftarrow}(i,f(X))\right].
\end{aligned}$$

Using $C_1 \leq \sum_{i=1}^{K}\ell^{\leftarrow}(i,f(X)) \leq C_2$, we obtain

$$\begin{aligned}
&\left(1 - \frac{K\eta}{K-1}\right)R(f;\ell) + \frac{\eta}{K-1}C_1 \leq \widetilde{R}(f;\ell^{\leftarrow}) \\
&\leq \left(1 - \frac{K\eta}{K-1}\right)R(f;\ell) + \frac{\eta}{K-1}C_2.
\end{aligned}$$

Therefore from the above inequality, we have

$$\begin{aligned}
(0 \leq)\ &\widetilde{R}(f^*;\ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow}) \\
&\leq \frac{\eta}{K-1}(C_2 - C_1) + \left(1 - \frac{K\eta}{K-1}\right)\left(R(f^*;\ell) - R(\widetilde{f}^*;\ell)\right) \\
&\leq \frac{\eta}{K-1}(C_2 - C_1),
\end{aligned}$$

because $\eta < \frac{K-1}{K}$ and f^* is the minimizer of $R(f;\ell)$. Similarly, when $\eta < 0$, we can prove that

$$(0 \leq)\ \widetilde{R}(f^*;\ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow}) \leq -\frac{\eta}{K-1}(C_2 - C_1).$$

Thus, for $\eta < \frac{K-1}{K}$, $(0 \leq)\ \widetilde{R}(f^*;\ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow}) \leq \frac{|\eta|}{K-1}(C_2 - C_1)$. □

A.5 Proof of Theorem 4

Proof. In a similar way to the proof of Theorem 2, it can be shown that

$$\begin{aligned}
\widetilde{R}(f;\ell^{\leftarrow}) &= \mathbb{E}_{\mathcal{D}}\left[(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\sum_{i=1}^{K}w_i\ell(i,f(X))\right] \\
&\quad - \mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\ell(j,f(X))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\max\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\sum_{i=1}^{K}w_i\ell(i,f(X))\right] \\
&\quad + \mathbb{E}_{\mathcal{D}}\left[\min\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\sum_{i=1}^{K}w_i\ell(i,f(X))\right] \\
&\quad - \mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\ell(j,f(X))\right].
\end{aligned}$$

Using $C_1 \leq \sum_{i=1}^{K}w_i\ell(i,f(X)) \leq C_2$, we obtain

$$\begin{aligned}
\widetilde{R}(f;\ell^{\leftarrow}) &\leq C_2\mathbb{E}_{\mathcal{D}}\left[\max\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right] \\
&\quad + C_1\mathbb{E}_{\mathcal{D}}\left[\min\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right] \\
&\quad - \mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\ell(j,f(X))\right],
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{R}(f;\ell^{\leftarrow}) &\geq C_1\mathbb{E}_{\mathcal{D}}\left[\max\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right] \\
&\quad + C_2\mathbb{E}_{\mathcal{D}}\left[\min\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right] \\
&\quad - \mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y} - (\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\ell(j,f(X))\right].
\end{aligned}$$

Therefore from the above inequalities, we have

$$(0 \leq)\ \widetilde{R}(f^*;\ell^{\leftarrow}) - \widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow})$$

Table A·1 Mean test accuracy (4 trials) of different methods on benchmark datasets under symmetric and asymmetric noise settings. All methods use the backward correction by complementary matrix without gradient ascent techniques, and differ only in the loss function. Bold face denotes the best and comparable methods according to the paired t-test at the significance level 5%.

Datasets	Losses	Symmetric Noise			Asymmetric Noise	
		Noise Rate η			Noise Rate η	
		0.0	0.2	0.6	0.2	0.4
MNIST	CCE	29.95 ± 10.28	77.61 ± 2.22	64.26 ± 8.74	71.39 ± 2.56	55.25 ± 1.20
	MAE	**94.22 ± 0.46**	**89.99 ± 0.68**	70.86 ± 4.38	71.00 ± 10.96	30.50 ± 3.44
	GCE	92.34 ± 0.17	**89.89 ± 0.37**	**74.13 ± 1.79**	84.75 ± 3.38	**67.57 ± 0.92**
	SL	93.20 ± 0.23	**90.52 ± 0.74**	69.91 ± 4.31	**88.68 ± 1.50**	**68.38 ± 0.98**
Fashion MNIST	CCE	59.79 ± 12.85	68.55 ± 1.26	64.81 ± 7.64	51.63 ± 8.14	48.05 ± 3.47
	MAE	**83.24 ± 0.53**	75.91 ± 3.63	61.56 ± 4.45	53.77 ± 1.45	19.29 ± 9.21
	GCE	82.21 ± 0.45	**80.92 ± 0.53**	**72.18 ± 3.12**	71.00 ± 6.34	52.10 ± 2.39
	SL	**82.48 ± 0.58**	**80.75 ± 1.32**	65.94 ± 3.02	**77.70 ± 3.80**	**58.71 ± 1.74**
Kuzushiji MNIST	CCE	57.25 ± 1.40	55.02 ± 1.53	41.39 ± 3.16	48.52 ± 3.14	32.77 ± 1.23
	MAE	64.88 ± 2.95	60.23 ± 0.80	37.59 ± 5.25	40.08 ± 3.44	27.10 ± 2.65
	GCE	69.52 ± 1.01	61.84 ± 2.03	**46.99 ± 3.44**	58.98 ± 1.66	44.12 ± 1.17
	SL	**71.34 ± 0.95**	**65.57 ± 0.51**	43.89 ± 2.56	**61.52 ± 3.91**	**47.13 ± 2.66**
CIFAR-10	CCE	36.90 ± 1.44	28.15 ± 2.44	**21.14 ± 2.91**	**33.19 ± 0.93**	15.56 ± 2.66
	MAE	41.19 ± 2.34	30.93 ± 4.82	17.02 ± 1.92	18.67 ± 4.05	13.72 ± 1.49
	GCE	**49.03 ± 1.31**	**42.77 ± 1.63**	19.09 ± 3.55	**32.59 ± 1.10**	**19.74 ± 0.84**
	SL	**48.57 ± 0.33**	**41.60 ± 0.51**	**22.26 ± 0.25**	27.82 ± 2.64	**20.22 ± 1.75**

$$
\begin{aligned}
&\le \mathbb{E}_{\mathcal{D}}\left[\max\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right](C_2-C_1)\\
&\quad+\mathbb{E}_{\mathcal{D}}\left[\min\left\{0,(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right\}\right](C_1-C_2)\\
&\quad+\mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y}-(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\right.\\
&\qquad\left.\left(\ell(j,\widetilde{f}^*(X))-\ell(j,f^*(X))\right)\right]\\
&=\mathbb{E}_{\mathcal{D}}\left[\left|(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right|\right](C_2-C_1)\\
&\quad+\mathbb{E}_{\mathcal{D}}\left[\sum_{j\neq Y}\left\{(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{w_j}{w_Y}-(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{Yj}\right\}\right.\\
&\qquad\left.\left(\ell(j,\widetilde{f}^*(X))-\ell(j,f^*(X))\right)\right].
\end{aligned}
$$

Since we are given $R(f^*;\ell)=0$, we have $\ell(Y,f^*(X))=0$. From this and the constraint on ℓ in the theorem, we have $\ell(j,f^*(X))=\sup_{\boldsymbol{v}\in\mathbb{R}^K}\ell(j,\boldsymbol{v})\ (<\infty)$, $\forall j\neq Y$. Moreover, since we are given $w_j(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ii}>w_i(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{ij}$, $\forall j\neq i$, we have $(0\le)$ $\widetilde{R}(f^*;\ell^{\leftarrow})-\widetilde{R}(\widetilde{f}^*;\ell^{\leftarrow})\le\mathbb{E}_{\mathcal{D}}\left[\left|(\boldsymbol{T}\widehat{\boldsymbol{T}}^{-1})_{YY}\frac{1}{w_Y}\right|\right](C_2-C_1)$. □

Appendix B: Bounds for Weighted Sums of Losses

We derive the bounds for the weighted sums of WMAE, MAE, and GCE.

- WMAE: Let $\boldsymbol{w}$ be the weight vector. Suppose $w_i>0$, $\forall i=1,\dots,K$, and $\sum_{i=1}^K w_i=K$. Then, because $\sum_{i=1}^K F_i(f(x))=1$, we have

$$
\begin{aligned}
\sum_{i=1}^K w_i\ell_{\mathrm{wmae}}(i,f(x)) &= \sum_{i=1}^K w_i\cdot\frac{1}{w_i}(2-2F_i(f(x)))\\
&= \sum_{i=1}^K(2-2F_i(f(x))=2K-2.
\end{aligned}
$$

- MAE: We have

$$
\begin{aligned}
\sum_{i=1}^K w_i\ell_{\mathrm{mae}}(i,f(x)) &= \sum_{i=1}^K w_i(2-2F_i(f(x)))\\
&= 2\left(K-\sum_{i=1}^K w_iF_i(f(x))\right),
\end{aligned}
$$

and thus the weighted sum is bounded as

$$
2(K-\max_{i\in\mathcal{Y}}w_i)\le\sum_{i=1}^K w_i\ell_{\mathrm{mae}}(i,f(x))\le 2(K-\min_{i\in\mathcal{Y}}w_i).
$$

- GCE: For the upper bound, we have

$$
\begin{aligned}
\sum_{i=1}^K w_i\ell_{\mathrm{gce}}(i,f(x)) &\le \sum_{i=1}^K w_i\frac{(1-F_i(f(x)))}{q}\\
&= \frac{K-\sum_{i=1}^K w_iF_i(f(x))}{q}\\
&\le \frac{K-\min_{i\in\mathcal{Y}}w_i}{q},
\end{aligned}
$$

where the first inequality holds because $(1-F_i(f(x))^q)\le(1-F_i(f(x)))$ for $q\in(0,1]$. Moreover, bacause of Jensen's inequality, the lower bound is

Table A· 2 Mean test accuracy (4 trials) of different methods on benchmark datasets under symmetric and asymmetric noise settings. All methods use the forward correction by complementary matrix, and differ only in the loss function. Bold face denotes the best and comparable methods according to the paired t-test at the significance level 5%.

Datasets	Losses	Symmetric Noise			Asymmetric Noise	
		Noise Rate η			Noise Rate η	
		0.0	0.2	0.6	0.2	0.4
MNIST	CCE	93.74 ± 0.08	88.81 ± 0.37	73.55 ± 2.20	**88.49 ± 2.11**	**69.03 ± 1.76**
	MAE	**94.18 ± 0.23**	**89.89 ± 1.42**	68.94 ± 9.70	65.92 ± 7.06	29.68 ± 4.32
	GCE	**94.36 ± 0.16**	**90.72 ± 0.96**	**75.50 ± 3.18**	82.57 ± 0.05	51.15 ± 5.03
	SL	**94.26 ± 0.35**	**90.87 ± 0.75**	**77.95 ± 4.27**	**90.07 ± 0.59**	57.68 ± 3.95
Fashion MNIST	CCE	**83.27 ± 0.34**	79.40 ± 0.51	**66.59 ± 0.34**	**79.17 ± 1.29**	**49.38 ± 4.76**
	MAE	80.27 ± 6.21	**80.32 ± 0.50**	63.74 ± 5.55	54.60 ± 2.41	26.86 ± 8.24
	GCE	**83.56 ± 0.21**	**81.30 ± 0.68**	**67.93 ± 2.25**	73.79 ± 4.63	30.43 ± 6.22
	SL	**83.50 ± 0.09**	**80.79 ± 1.15**	64.79 ± 4.68	51.12 ± 3.44	23.06 ± 4.50
Kuzushiji MNIST	CCE	**71.17 ± 1.79**	**60.79 ± 1.32**	45.40 ± 2.04	**64.83 ± 2.09**	**39.07 ± 1.71**
	MAE	66.49 ± 1.41	57.69 ± 2.90	40.50 ± 1.81	30.97 ± 4.20	20.81 ± 5.66
	GCE	67.68 ± 1.78	**60.14 ± 0.45**	**47.64 ± 0.97**	59.31 ± 0.82	28.30 ± 0.86
	SL	63.98 ± 3.53	**60.07 ± 3.06**	**45.82 ± 2.59**	**63.71 ± 6.76**	34.44 ± 4.66
CIFAR-10	CCE	**52.04 ± 0.41**	42.12 ± 1.57	**21.79 ± 3.19**	**36.38 ± 1.96**	**18.83 ± 2.08**
	MAE	40.44 ± 2.58	32.38 ± 3.02	18.72 ± 2.43	18.77 ± 2.66	12.78 ± 0.83
	GCE	**50.88 ± 1.73**	**44.87 ± 0.82**	**20.83 ± 7.50**	27.81 ± 3.59	14.00 ± 1.13
	SL	**51.31 ± 1.36**	**44.20 ± 2.16**	**22.18 ± 2.56**	24.56 ± 2.06	13, 70 ± 1.98

$$\begin{aligned}\sum_{i=1}^{K} w_i \ell_{\mathrm{gce}}(i, f(x)) &= \frac{K}{q}\sum_{i=1}^{K}\frac{w_i}{K}(1 - F_i(f(x))^q) \\ &\geq \frac{K}{q}\left\{1 - \left(\sum_{i=1}^{K}\frac{w_i}{K}F_i(f(x))\right)^q\right\} \\ &\geq \frac{K - K^{(1-q)}(\max_{i\in\mathcal{Y}} w_i)^q}{q}.\end{aligned}$$

Appendix C: Additional Information of Experiments

Here we report the results of additional experiments.

C.1 Backward Correction without Gradient Ascent Techniques

We applied backward correction to each loss and trained the models without using gradient ascent techniques. The detailed experimental results are shown in Table A· 1. From the results, we can see that GCE and SL outperformed CCE under noise. However, when compared to the results in Table 2, the overall classification performance is low. In particular, CCE and MAE seem to be greatly affected. Since CCE is not upper-bounded, the empirical risk is not lower-bounded and was consequently most plagued by the negative risk issues [5], [6], [21]. On the other hand, MAE did not work well despite having an upper bound. This might be because MAE can perform poorly with DNNs and challenging datasets due to gradient saturation [27]. This means that this problem did not occur when the parameters were optimized with the gradient ascent technique.

C.2 Forward Correction

The detailed experimental results for forward correction are shown in Table A· 2. From the results, we can see that the losses other than CCE did not work well, especially under asymmetric noise.

Hiroki Ishiguro was born in Tokyo, Japan, in 1997. He received the B.S. degree from Tokyo Institute of Technology in 2019 and the M.S. degree from the University of Tokyo in 2021. His research interests include machine learning and its applications.

Takashi Ishida is a Lecturer at the Graduate School of Frontier Sciences, the University of Tokyo. He received his PhD from the University of Tokyo in 2021. He received the MSc from the University of Tokyo in 2017 and the BEc from Keio University in 2013.

Masashi Sugiyama is Director of RIKEN Center for Advanced Intelligence Project and Professor at the University of Tokyo. His research interests include theories and algorithms of machine learning and their applications in real-world problems. He received the Japan Society for the Promotion of Science Award and the Japan Academy Medal in 2017 for his machine learning research.