Asymmetric Loss Functions for Learning with Noisy Labels

Xiong Zhou 1 2  Xianming Liu 1 2  Junjun Jiang 1 2  Xin Gao 3 2  Xiangyang Ji 4

Abstract

Robust loss functions are essential for training deep neural networks with better generalization power in the presence of noisy labels. Symmetric loss functions are confirmed to be robust to label noise. However, the symmetric condition is overly restrictive. In this work, we propose a new class of loss functions, namely asymmetric loss functions, which are robust to learning with noisy labels for various types of noise. We investigate general theoretical properties of asymmetric loss functions, including classification calibration, excess risk bound, and noise tolerance. Meanwhile, we introduce the asymmetry ratio to measure the asymmetry of a loss function. The empirical results show that a higher ratio would provide better noise tolerance. Moreover, we modify several commonly-used loss functions and establish the necessary and sufficient conditions for them to be asymmetric. Experimental results on benchmark datasets demonstrate that asymmetric loss functions can outperform state-of-the-art methods. The code is available at https://github.com/hitcszx/ALFs

1. Introduction

The success of deep neural networks based supervised learning largely relies on massive high-quality labeled data. However, in practice, the annotation process inevitably introduces wrong labels, due to the lack of experts involved or data from public crowdsourcing platforms (Liu et al., 2011; Arpit et al., 2017). Empirical studies show that over-parameterized deep networks can even fit random labels (Zhang et al., 2017). When samples are mis-labeled, the network would memorize wrong patterns, leading to impaired performance in the subsequent inference tasks. Accordingly, robust learning of classifier in the presence of label noise has received a lot of attention.

To alleviate the impact of label noise to classifier learning, one popular research line is to design noise-tolerant loss functions. This approach has been pursued in a large body of work (Long & Servedio, 2008; Wang et al., 2019a; Liu & Guo, 2020; Lyu & Tsang, 2020; Menon et al., 2020; Feng et al., 2020) that embraces new losses, especially symmetric loss functions and their variants (Manwani & Sastry, 2013; van Rooyen et al., 2015; Ghosh et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019b; Ma et al., 2020).

Symmetric loss functions were proposed as a sufficient condition such that the risk minimization with respect to the loss becomes noise-tolerant for binary classification (Manwani & Sastry, 2013). Subsequently, the unhinged loss (van Rooyen et al., 2015), which is equivalent to a scaled Mean Absolute Error (MAE) (Ghosh et al., 2017), was proved to be the only convex loss function that is strongly robust for symmetric label noise (SLN). Ghosh et al. (Ghosh et al., 2017) theoretically demonstrated that a loss function would be inherently tolerant to SLN as long as it satisfies the symmetric condition. The sufficient condition was then extended for multi-class classification (Ghosh et al., 2017) and was emphasized in the BER minimization and AUC maximization from corrupted labels (Charoenphakdee et al., 2019). However, MAE treats every sample equally, leading to significantly longer training time before convergence. This drawback motivates some works to improve MAE, which follows the principle of combining the robustness of MAE and the fast convergence of Cross Entropy (CE). For instance, (Zhang & Sabuncu, 2018) advocated the use of a more general class of noise-robust loss functions, called Generalized Cross Entropy (GCE), which encompasses both MAE and CE. Inspired by the symmetric KL-divergence, the symmetric cross entropy (SCE) (Wang et al., 2019b) was proposed to combine CE with a noise tolerance term, namely Reverse Cross Entropy (RCE). Ma et al. (Ma et al., 2020) theoretically demonstrated that by applying a simple normalization, any loss can be made robust to noisy labels. However, the normalized loss functions are not sufficient to train accurate DNNs and are prone to encounter the gradient explosion problem.

From the above review, it can be found that the fitting ability of the existing symmetric loss functions is restricted by the symmetric condition (Zhang & Sabuncu, 2018; Charoenphakdee et al., 2019). However, the symmetric condition...
Asymmetric Loss Functions for Learning with Noisy Labels

is too stringent to find a convex loss function (Plessis et al., 2015; Ghosh et al., 2015; van Rooyen et al., 2015), leading to difficulties in optimization. Thus, learning with symmetric loss function usually suffers from underfitting issues.

In this paper, we propose a new class of robust loss functions, namely asymmetric loss functions, which are tailored to satisfy that the Bayes-optimal prediction under the loss is a point-mass on the highest scoring label, i.e., the loss has Bayes-optimal prediction that matches that of the 0-1 loss. Specifically, our scheme is based on a reasonable assumption that in a training dataset samples have higher probability to be annotated with true semantic labels than any other class labels. According to this clean-labels-domination assumption, the proposed asymmetric loss is derived. It indicates that, minimizing the $L$-risk under noisy case, which can be formulated as the weighted form, would make the optimization direction shift to the loss term with the maximum weight. In this way, the contribution of noisy labels in the process of classifier learning is eliminated, and thus the proposed asymmetric loss functions are inherently noise-tolerant. Furthermore, we offer a complete theoretical analysis about the properties of asymmetric loss functions, including classification calibration, excess risk bound, noise-tolerance, and asymmetry ratio. We show that several commonly-used loss functions can be modified to be asymmetric, and establish the corresponding necessary and sufficient conditions for them. The main contributions of our work are highlighted as follows:

- We propose a new family of robust loss functions, namely asymmetric loss functions, which are noise-tolerant with an appropriate model for various types of noise. We theoretically prove that completely asymmetric losses, which include symmetric losses as a special case, are classification-calibrated, and have an excess risk bound when they are strictly asymmetric.

- We introduce the asymmetric ratio to measure the asymmetry of a loss function, which, together with the clean level of labels, can be associated with noise-tolerance. The empirical results show that higher ratio will provide better noise robustness.

- We generalize several commonly-used loss functions, and establish the necessary and sufficient conditions for them to be asymmetric. The experimental results demonstrate that the new loss functions can outperform the state-of-the-art methods.

2. Preliminaries

2.1. Risk Minimization

Define $\mathcal{X} \subset \mathbb{R}^d$ as the feature space from which the samples are drawn, and $\mathcal{Y} = [k] = \{1, \ldots, k\}$ as the class label space, i.e., we consider a $k$-classification problem, where $k \geq 2$. In an ideal classifier learning problem, we are given a clean training set, $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $(x_n, y_n)$ is drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The classifier is a mapping function from feature space to label space $h(x) = \arg \max_i f(x_i)$, where $f : \mathcal{X} \rightarrow \mathcal{C}, \mathcal{C} \subseteq [0, 1]^k, \forall c \in \mathcal{C}, 1^Tc = 1$. $f(x)$ denotes an approximation of $p\left(\cdot \mid x\right)$, which is considered as a neural network ending with a softmax layer in this work.

We define a loss function as a mapping $\eta : \mathcal{X} \times \mathcal{C} \rightarrow [0, \infty)$, and establish the necessary and sufficient conditions, including classification calibration, excess risk bound, metric losses, which include symmetric losses as a special case, are classification-calibrated, and have an asymmetry of a loss function, which, together with its complexity.

2.2. Label Noise Model

The annotation process inevitably introduces label noise, the model of which can be formulated as

$$\hat{y}_n = \begin{cases} i, & i \in [k], i \neq y_n \text{ with probability } \eta_{\kappa_n, i} \\ y_n, & \text{with probability } (1 - \eta_{\kappa_n}) \end{cases}$$

where $\eta_{\kappa_n, i}$ denotes the probability of flipping the true label $y_n$ into $i$ for $x_n$, and $\eta_{\kappa_n} = \sum_{i \neq y_n} \eta_{\kappa_n, i}$ denotes the noise ratio of $x_n$. This noise model shows that a realistic corruption probability is dependent on both data features and class labels (Xiao et al., 2015; Goldberger & Ben-Reuven, 2016), and this kind of noise is called instance- and label-dependent noise (Cheng et al., 2020). However, this modeling approach of label noise has not been investigated extensively yet due to its complexity.

Instead, a popular approach for modeling label noise simply assumes that the corruption process is conditionally independent of data features when the true label is given (Natarajan et al., 2013), i.e., $\eta_{\kappa_n}$ and $\eta_{\kappa_n, i}$ are only dependent on the class labels, which can be then represented as a label transition matrix. If $\eta_{\kappa_n, i} = \eta, \forall x_n, i$, the noise is called symmetric (or uniform) noise, where a true label is flipped into other labels with equal probability. In contrast to symmetric noise, another type of noise is called asymmetric if $\exists i, \eta_{\kappa_n} = \eta$ and $\exists i \neq y_n, \forall j \neq y_n, \eta_{\kappa_n, i} > \eta_{\kappa_n, j}$, i.e., a certain class is more likely to be wrongly annotated into a particular label (Song et al., 2020).

Based on the label noise model, the $L$-risk under noisy case can be formulated as

$$R_L(f) = \mathbb{E}_D[\left(1 - \eta\right)L(f(x), y) + \sum_{i \neq y} \eta_{\kappa_n}L(f(x), i)]$$
It can be found that, due to the presence of noisy labels, the classifier learning process is influenced by \( \sum_{i \neq y} \eta_{x,i} L(f(x), i) \), i.e., noisy labels would degrade the generalization performance of deep neural networks. Define \( f_\eta^* \) be the global minimum of \( R_L^\eta(f) \), then \( L \) is noise-tolerant if \( f_\eta^* \) is also the global minimum of \( R_L(f) \).

2.3. Symmetric Loss Functions

The most popular family of loss functions in robust learning is symmetric loss (Manwani & Sastry, 2013; Ghosh et al., 2017). A loss is called symmetric if it satisfies

\[
\sum_{i=1}^{k} L(f(x), i) = C, \quad \forall x \in X, \forall f,
\]

where \( C \) is a constant value. Ghosh et al. (Ghosh et al., 2017) proved that, for a \( k \)-classification problem, if the loss \( L \) is symmetric and the noise ratio \( \eta < \frac{k - 1}{k} \), then under symmetric noise \( L \) is noise-tolerant. Moreover, if \( R_L(f^*) = 0 \), the loss function is also noise-tolerant under asymmetric noise, where \( f^* \) is a global minimizer of \( R_L \).

One of the most classic symmetric loss functions is MAE (Ghosh et al., 2017), which is defined as \( L(u, i) = ||e_i - u||_1 = 2 - 2u_i \) and obviously satisfies \( \sum_i L(u, i) = 2k - 2 \). Reverse Cross Entropy (RCE) proposed in (Wang et al., 2019b) is also belonging to the kind of symmetric loss, which is actually the variant of MAE. Ma et al. (Ma et al., 2020) proposed the normalized loss functions, which can make any loss symmetric by using a simple normalization operation. However, the symmetric condition is too stringent to find a convex loss function (Plessis et al., 2015; Ghosh et al., 2015; van Rooyen et al., 2015), leading to difficulties in optimization. Thus, learning with symmetric loss function usually suffers from the underfitting effect.

In this paper, we propose a new family of loss functions, asymmetric loss functions, which includes symmetric loss functions as its special case. More importantly, the proposed asymmetric loss family also guarantees some desirable properties and contain many convex loss functions, which facilitate the subsequent optimization process.

3. Asymmetric Loss Functions

In this section, we introduce in details the proposed asymmetric loss functions. Firstly, we state the clean-labels-domination assumption, which serves as the fundamental basic in the subsequent derivation. Then we introduce the proposed asymmetric loss functions, a new class of robust loss function, which achieve robust learning by keep consistency between the Bayes-optimal prediction of the loss and that of the 0-1 loss. Subsequently, we theoretically explore general properties of asymmetric loss functions, including classification calibration, excess risk bound, noise tolerance, and asymmetry ratio. Finally, we show that several commonly-used loss functions can be modified to be asymmetric and thus robust to label noise. The necessary and sufficient conditions are offered for them. The detailed proofs for theorems and corollaries can be found in the supplementary material.

3.1. Clean-labels-domination Assumption

For robust learning, it is reasonable to assume that in a training dataset samples have higher probability to be annotated with true semantic labels than any other class labels, which is referred to as clean-labels-domination assumption.

In the following, we first provide the formal definition of class-wise clean-labels-domination.

**Definition 1.** Given an underlying clean dataset \( S \), the corresponding observed noisy dataset is \( \hat{S} \). The \( i \)-th class subset of \( \hat{S} \) is formulated as \( \hat{S}_i = \{(x, \hat{y}) : y = i, (x, \hat{y}) \in \hat{S}, (x, y) \in S\} \), with \( i \in [k] \). We define that the class label \( i \) is dominant in \( \hat{S}_i \) if it satisfies

\[
\sum_{(x, \hat{y}) \in \hat{S}_i} \mathbb{I}(\hat{y} = i) > \max_{j \neq i} \sum_{(x, \hat{y}) \in \hat{S}_i} \mathbb{I}(\hat{y} = j),
\]

where \( \mathbb{I}() \) is the identity function.

The dataset \( \hat{S} \) is claimed to be clean-labels-dominant if in all classes correct labels are dominant. In real-world datasets, the noise ratio of noisy labels is reported to range from 8.0% to 38.5% (Song et al., 2020), which serves as the corroborration that \( \hat{S} \) is usually clean-labels-dominant. Based on this empirical observation, we further assume that the label noise model defined in (2) is clean-labels-dominant:

**Assumption 1.** The label noise model in (2) is clean-labels-dominant, i.e., it satisfies that \( \forall x, 1 - \eta_{x} > \max_{j \neq y} \eta_{x,j} \).

Compared with the symmetric noise assumption behind symmetric losses (Ghosh et al., 2017), i.e., \( 1 - \eta_{x} > \frac{1}{k} \), although Assumption 1 is more restrictive, it makes sense in general and applicable to most of the real-world applications. Specifically, without the help of any prior knowledge, if there exists an approach that can help to learn a correct classifier on clean-labels-non-dominant cases, then it would fail in learning a correct classifier on clean-labels-dominant cases since the learned classifier tends to classify a sample into a non-dominant class rather than the corresponding dominant class (i.e., the true class).

3.2. Asymmetric Loss Functions

For a sample \((x, y)\) drawn from \( \mathcal{D} \), we have the conditional \( L \)-risk (Bartlett et al., 2006):

\[
L^0(f(x), y) = (1 - \eta_x)L(f(x), y) + \sum_{i \neq y} \eta_{x,i}L(f(x), i).
\]
The exact values of \( \{\eta_{x,i}\}_{i \neq y} \) are usually unknown, and what we only know is \( 1 - \eta_{x} > \max_{i \neq y} \eta_{x,i} \) according to Assumption 1. Our purpose is to find a simple and elegant formulation of \( L \) such that minimizing the risk leads to a classifier with the same probability of mis-classification as the noise-free case. To this end, in this work, we suggest a new class of loss functions defined as follows:

**Definition 2.** On the given weights \( w_1, \ldots, w_k \geq 0 \), where \( \exists t \in [k], \text{s.t.} \, w_t > \max_{i \neq t} w_i \), a loss function \( L(u, i) \) is called asymmetric if \( L \) satisfies

\[
\arg \min_u \sum_{i=1}^{k} w_i L(u, i) = \arg \min_u L(u, t),
\]

where we always have \( \arg \min_u L(u, t) = e_t \).

We define that \( L \) is asymmetric on the label noise model that satisfies Assumption 1, if \( L \) is asymmetric on \( \{1 - \eta_{x}\} \cup \{\eta_{x,i} \neq y\} \forall (x, y) \) drawn from \( D \). \( L \) is called completely asymmetric, if \( L \) is asymmetric on any weights \( w_1, \ldots, w_k \geq 0 \) that contain a unique maximum. And we call \( L \) strictly asymmetric, if it satisfies \( \sum_{i=1}^{k} w_i L(u, i) < \sum_{i=1}^{k} w_i L(u', i) \), \( \forall w_1, \ldots, w_k \geq 0 \) with a unique maximum \( w_t \), and \( \forall u', u \in C, u'_t > u_t \).

The asymmetry is reflected by the fact that minimizing the weighted risk would make the optimization direction shift to the loss term with the maximum weight. This strategy is referred to as the-largest-takes-all.

More specifically, according to Definition 2 and Assumption 1, asymmetric loss functions are inherently noise-tolerant, which can eliminate the contribution of noisy labels (i.e., \( \sum_{i \neq y} \eta_{x,i} L(f(x), i) \)) in the process of classifier learning. It is desirable since it provides an approach of obtaining the minimum for noise-free case \( L(u, t) \) from the minimization for noisy case \( \sum_{i=1}^{k} w_i L(u, i) \). In other words, the asymmetric loss are tailored to satisfy that the Bayes-optimal prediction under the loss is a point-mass on the highest scoring label, i.e., the loss has Bayes-optimal prediction that matches that of the 0-1 loss.

### 3.3. Properties of Asymmetric Loss Functions

Let \( L(u, i) \) be asymmetric on the label noise model which is clean-labels-dominant. According to the asymmetric condition (5), it can be derived that \( (1 - \eta_{x}) L(u, y) + \sum_{i \neq y} \eta_{x,i} L(u, i) \geq (1 - \eta_{x}) L(u^*, y) + \sum_{i \neq y} \eta_{x,i} L(u^*, i) \), where \( u^* = e_y \), and the equality holds if and only if \( u = u^* \). This inequality reveals a beautiful property for binary classification as follows:

**Theorem 1** (Classification calibration). Completely asymmetric loss functions are classification-calibrated.

Classification calibration is known to be a minimal requirement of a loss function for the binary classification task (Tong, 2003; Bartlett et al., 2006). We say that \( \phi \) is classification-calibrated if driving the excess risk over the Bayes-optimal predictor for \( \phi \) to zero also drives the excess risk for 0-1 loss to zero. Actually, the conditional risk minimizer of \( L \) is equivalent to the Bayes-optimal classifier \( \hat{y}(\eta_x > \frac{1}{2}) \) (see more in supplementary materials).

![Figure 1. Verification of classification calibration. Solid and dashed lines denote the curve of \( H_t(\eta) \) and \( H_t^c(\eta) \), respectively. As can be observed, the curve of \( H_t^c \) is always above that of \( H_t \), i.e., the loss functions are classification-calibrated.](image)

Another essential property is excess risk bound (Bartlett et al., 2006), which provides a relationship between the excess risk of minimizing the mis-classification risk w.r.t. the 0-1 loss and the surrogate loss. The following theorem indicates an excess bound for any strictly and completely asymmetric loss functions.

**Theorem 2** (Excess risk bound). An excess risk bound of a strictly and completely asymmetric loss function \( L(u, i) = \ell(u_i) \) can be expressed as

\[
R_{\ell_{0-1}}(f) - R_{\ell_{0-1}}^*(f) \leq \frac{2(R_{\ell}(f) - R_{\ell}^*)}{\ell(0) - \ell(1)},
\]

where \( R_{\ell_{0-1}}^* = \inf_g R_{\ell_{0-1}}(g) \) and \( R_{\ell}^* = \inf_g R_{\ell}(g) \).

The result suggests that the excess risk bound of any strictly and completely asymmetric loss function is controlled only by the difference of \( \ell(0) - \ell(1) \). Intuitively, the excess risk bound shows that if the hypothesis \( f \) minimizes the surrogate risk \( R_{\ell}(f) = R_{\ell}^* \), then \( f \) must also minimize the mis-classification risk \( R_{\ell_{0-1}}(f) = R_{\ell_{0-1}}^* \).

As aforementioned, symmetric loss functions are well-studied with general properties (Manwani & Sastry, 2013; Ghosh et al., 2017; Charoenphakdee et al., 2019). Here we reveal the relationship between symmetric loss functions and asymmetric loss functions.

**Theorem 3**. Symmetric loss functions are completely asymmetric.

An important condition for symmetric loss functions to be noise-tolerant under asymmetric noise is \( R_L(f^*) = 0 \), i.e., there exists a hypothesis can fit the distribution \( D \) perfectly. Here we use deep networks as the hypothesis class to obtain enough fitting ability (Zhang et al., 2017; Zou & Gu, 2019).
Assumption 2. Given the loss function $L$ and a separable distribution $\mathcal{D}$, we assume that there exists a hypothesis $f: \mathcal{X} \rightarrow \mathcal{Y}$, $f \in \mathcal{H}_{\text{act}}$, $\forall (x, y)$ drawn from $\mathcal{D}$, such that $f$ minimizes $L(f(x), y)$.

To satisfy this assumption, the hypothesis class $\mathcal{H}_{\text{act}}$ should be as universal as possible to approximate complex functions. According to the universal approximation theorem (Cybenko, 1989; Martin & Peter L., 1999), if a certain deep network model is employed, $\mathcal{H}_{\text{act}}$ will be a universal hypothesis class and thus contains the optimal function.

Theorem 4 (Noise tolerance). In a multi-classification problem, given an appropriate neural network class $\mathcal{H}$ which satisfies Assumption 2, the loss function $L$ is noise-tolerant if $L$ is asymmetric on the label noise model.

This theorem shows that noise tolerance can be obtained without knowing the exact noise rates when the loss is asymmetric on the label noise model. This conclusion does not depend on the data distribution. We just require that the label noise model is clean-labels-dominant and there is a neural network which is as universal as possible. Therefore, the key question becomes how to design a loss function being asymmetric on the label noise model. Moreover, if a loss is completely asymmetric, then it is robust to any label noise model. In the next subsection, we will provide a comprehensive analysis.

Inspired by the benefit of symmetric (Wang et al., 2019b) or complementary learning (Kim et al., 2019), the Active Passive Loss (APL) framework was proposed (Ma et al., 2020) for both robust and sufficient learning. The following theorem indicates that the asymmetric loss functions are also suitable for the APL framework. In our experiments, we also employ the framework to achieve better or at least comparable performance.

Theorem 5. $\forall \alpha, \beta > 0$, if $L_1$ and $L_2$ are asymmetric, then $\alpha L_1 + \beta L_2$ is asymmetric.

We know that all symmetric loss functions are also asymmetric according to Theorem 3. Is there a new asymmetric loss function? However, Definition 2 is too abstract to find a new specific form. In the following, we will provide a comprehensive theoretical analysis about designing asymmetric loss functions and propose several specific ones.

3.4. Asymmetry Ratio

As we can see, the asymmetry or direction of minimization is dependent on which one is the maximum weight, but how to measure the asymmetry of the function and select an asymmetric enough loss? We give the following definition:

Definition 3. Consider a loss function $L(u, i) = \ell(u_i)$, we define the asymmetry ratio $r(\ell)$ as

$$r(\ell) = \inf_{0 \leq u_1, u_2 \leq 1, \frac{u_1}{u_2} \leq \frac{1}{1 + \Delta u}} \frac{\ell(u_1) - \ell(u_1 + \Delta u)}{\ell(u_2) - \ell(u_2 - \Delta u)}.$$  (7)

The asymmetry ratio $r$ denotes the infimum ratio of change in the loss function $\ell$ when we increase the value of $u_1$ to $u_1 + \Delta u$, and correspondingly decrease the value of $u_2$ to $u_2 - \Delta u$. For example, the asymmetry ratio of MAE is 1, and the asymmetric ratio of GCE is 0 ($q < 1$). Based on the definition, we obtain the sufficient condition that $L$ is asymmetric on some weights.

Theorem 6 ( Sufficiency). On the given weights $w_1, \ldots, w_k$, where $w_m > w_n$ and $w_m = \max_{i \neq m} w_i$, the loss function $L(u, i) = \ell(u_i)$ is asymmetric if $\frac{w_m}{w_n} \cdot r(\ell) \geq 1$.

Remark. Let $c = \min_{(x, y), i \neq y} \frac{1 - p_{xy}}{m_x}$, which can be regarded as a measure of the clean level for the label noise mode in (2). The larger the $c$, the cleaner the labels. Moreover, we usually have $c \geq 1$ in accordance with Assumption 1, then $\ell$ will be completely asymmetric if $r(\ell) \geq 1$. On the other hand, we can estimate $c$ to design a loss that satisfies $r(\ell) \geq \frac{1}{c}$, i.e., being asymmetric on the label noise model, which leads to noise tolerance in accordance with Theorem 4. In fact, a loss satisfying $r(\ell) \geq \frac{1}{c}$ can be verified to be asymmetric to handle all synthetic noises regardless of on MNIST or CIFAR-10/100. For a real-world dataset, the clean level is usually higher than the synthetic case. In a sense, Theorem 6 associates the clean level and the asymmetry ratio with noise tolerance. In Section 4, we empirically show that larger $c \cdot r(\ell)$ would provide more noise tolerance. In the following, we shows that if $c \cdot r(\ell) \geq 1$, asymmetric losses will produce at least a positive weighted optimization rather than negative effects for any hypothesis class.
Asymmetric Loss Functions for Learning with Noisy Labels

Theorem 7. In a binary classification problem, we assume that \( L \) is strictly asymmetric on the label noise model which is clean-labels-dominant, for any hypothesis class \( \mathcal{H} \), let \( f^* = \arg \min_{f \in \mathcal{H}} R^2_{w, L}(f) \). If \( \forall x \), we have \( 1 - a \cdot r(L) > 1 \), then \( f^* \) also minimizes a positive weighted L-risk \( R_{w, L}(f) = \mathbb{E}[w(x, y) \cdot L(f(x), y)] \).

According to the definition of the asymmetric ratio, we can easily obtain an upper bound of \( r(\ell) \) when modifying the half-space constraint \( u_1 + u_2 \leq 0 \) to the hyperplane \( u_1 + u_2 = 1 \) and setting \( \Delta u = u_2 \), i.e.,

\[
r(\ell) \leq \inf_{0 \leq u_1, u_2 \leq 1} \ell(0) - \ell(u_2) = r_u(\ell).
\]

(8)

In some cases, the equality will hold, for example, both \( r \) and \( r_u \) of MAE are equal to 1. Actually, a completely asymmetric loss \( \ell \) satisfies \( r_u(\ell) \geq 1 \).

Theorem 8 (Necessity). On the given weights \( w_1, \ldots, w_k \), where \( w_m \geq w_k \) and \( w_m = \max_{i \neq m} w_i \), the loss function \( L(u, i) = \ell_u(u_i) \) is asymmetric only if \( \frac{w_m}{w_n} \cdot r_u(\ell) \geq 1 \).

According to Theorem 6 and Theorem 8, when \( r(\ell) = r_u(\ell) \cdot \frac{w_m}{w_n} \cdot r(\ell) \geq 1 \) will become the necessary and sufficient condition for \( L(u, i) = \ell(u_i) \) to be asymmetric. The following corollaries are straightforward from this.

Corollary 1. On the given weights \( w_1, \ldots, w_k \), where \( w_m > w_n \) and \( w_m = \max_{i \neq m} w_i \), the loss function \( L_{\text{AGCE}}(u, i) = \left( (a+1)^q - (a+u_i)^q \right) / q \) (where \( q > 0, a > 0 \)) is asymmetric if \( \frac{w_m}{w_n} \geq \left( \frac{a+1}{a} \right)^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1) \).

Mathematically, the loss function \( L_{\text{AGCE}} \) shown in Figure 4(a), is the negative shifted Box-Cox transformation, which we name as the Asymmetric Generalized Cross Entropy (AGCE) because when \( 0 < q \leq 1 \) and \( a = 0 \), the loss function is called GCE (Zhang & Sabuncu, 2018) which can be seen as a generalized mixture of CCE (when \( q \to 0 \)) and MAE (when \( q = 1 \)). Like MAE, both \( r \) and \( r_u \) of AGCE are equal. More specifically, \( r \) is equal to \( \left( \frac{a+1}{a} \right)^{1-q} \) when \( q \leq 1 \), and 1 when \( q \geq 1 \). As a consequence, AGCE is completely asymmetric when \( q \geq 1 \). Corollary 1 shows that if \( q > 1 \), the loss function beyond the range of \( q \) in GCE is asymmetric, or if \( q \leq 1 \) and \( \frac{w_m}{w_n} \geq \left( \frac{a+1}{a} \right)^{1-q} \), the convex loss function is also asymmetric, but when \( q < 1 \) and \( a = 0 \), the conventional GCE is not asymmetric.

Corollary 2. On the given weights \( w_1, \ldots, w_k \), where \( w_m > w_n \) and \( w_m = \max_{i \neq m} w_i \), the loss function \( L_{\text{AUL}}(u, i) = \left( [(a-u_i)\cdot \exp(1)] \right) / p \) (where \( p > 0 \) and \( a > 1 \)) is asymmetric if \( \frac{w_m}{w_n} \geq \left( \frac{a-1}{a} \right)^{p-1} \cdot \mathbb{I}(p > 1) + \mathbb{I}(p \leq 1) \).

We call the loss function above the Asymmetric Unhinged Loss (AUL) shown in Figure 4(b), because it is derived from the unhinged loss \( (a = 1 \text{ and } p = 1) \). Both the \( r \) and \( r_u \) of AUL are also equal, more specifically, the value of \( r \) is \( \left( \frac{a-1}{a} \right)^{p-1} \) when \( p \geq 1 \), and 1 when \( p < 1 \). Similar to AGCE, AUL is completely asymmetric when \( p \leq 1 \).

4. Experiments

In this section, we empirically investigate asymmetric loss functions on benchmark datasets, including MNIST (Lecun et al., 1998), CIFAR-10/-100 (Krizhevsky & Hinton, 2009), and a real-world noisy dataset WebVision (Li et al., 2017).

4.1. The Robustness of Asymmetric Loss Functions

Validation of Classification Calibration. We first conduct an experiment to validate the classification calibration in Theorem 1. As a corroboration, we plot the curves of \( H_{\ell}(\eta) \) and \( H_{\ell}(\eta) \) (the definitions can be found in the supplemen-
### Table 1. Test accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in \{0.2, 0.4, 0.6, 0.8\}$). The results (mean±std) are reported over 3 random runs and the top 3 best results are **boldfaced**.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>Clean ($\eta = 0.0$)</th>
<th>0.2</th>
<th>Symmetric Noise Rate ($\eta$)</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>CE</td>
<td>99.15 ± 0.05</td>
<td>91.62 ± 0.39</td>
<td>73.98 ± 0.27</td>
<td>49.36 ± 0.43</td>
<td>22.66 ± 0.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FL</td>
<td>99.13 ± 0.09</td>
<td>91.68 ± 0.14</td>
<td>74.54 ± 0.06</td>
<td>50.39 ± 0.28</td>
<td>22.65 ± 0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCE</td>
<td>99.27 ± 0.05</td>
<td>98.86 ± 0.07</td>
<td>97.16 ± 0.03</td>
<td>81.53 ± 0.58</td>
<td>33.95 ± 0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NLNL</td>
<td>98.61 ± 0.13</td>
<td>98.02 ± 0.14</td>
<td>97.17 ± 0.09</td>
<td>95.42 ± 0.30</td>
<td>86.34 ± 1.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCE</td>
<td>99.23 ± 0.10</td>
<td>98.92 ± 0.12</td>
<td>97.38 ± 0.15</td>
<td>88.53 ± 0.85</td>
<td>48.75 ± 1.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCE</td>
<td>98.60 ± 0.06</td>
<td>98.57 ± 0.01</td>
<td>98.29 ± 0.05</td>
<td>97.65 ± 0.08</td>
<td>93.78 ± 0.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCE+RCE</td>
<td>99.36 ± 0.05</td>
<td><strong>99.14 ± 0.03</strong></td>
<td>98.51 ± 0.06</td>
<td>95.60 ± 0.21</td>
<td>74.00 ± 1.68</td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td>AUL</td>
<td>99.14 ± 0.05</td>
<td><strong>99.05 ± 0.09</strong></td>
<td><strong>98.79 ± 0.09</strong></td>
<td><strong>98.67 ± 0.04</strong></td>
<td><strong>96.73 ± 0.20</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGCE</td>
<td>99.05 ± 0.11</td>
<td><strong>98.96 ± 0.10</strong></td>
<td><strong>98.83 ± 0.06</strong></td>
<td><strong>98.57 ± 0.12</strong></td>
<td><strong>96.59 ± 0.12</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AEL</td>
<td>99.03 ± 0.05</td>
<td>98.93 ± 0.06</td>
<td><strong>98.78 ± 0.13</strong></td>
<td><strong>98.51 ± 0.06</strong></td>
<td><strong>96.40 ± 0.11</strong></td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>AUL</td>
<td>91.27 ± 0.12</td>
<td>82.41 ± 0.09</td>
<td>85.64 ± 0.19</td>
<td>78.86 ± 0.66</td>
<td>52.92 ± 1.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGCE</td>
<td>88.95 ± 0.22</td>
<td>86.98 ± 0.12</td>
<td>83.39 ± 0.17</td>
<td>76.49 ± 0.53</td>
<td>44.42 ± 0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AEL</td>
<td>86.38 ± 0.19</td>
<td>84.27 ± 0.12</td>
<td>81.12 ± 0.20</td>
<td>74.86 ± 0.22</td>
<td>51.41 ± 0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCE+AUL</td>
<td>91.10 ± 0.13</td>
<td><strong>90.31 ± 0.20</strong></td>
<td><strong>86.23 ± 0.18</strong></td>
<td><strong>79.70 ± 0.08</strong></td>
<td><strong>59.44 ± 1.14</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCE+AGCE</td>
<td>90.94 ± 0.12</td>
<td><strong>89.21 ± 0.08</strong></td>
<td><strong>86.19 ± 0.15</strong></td>
<td><strong>80.13 ± 0.18</strong></td>
<td><strong>50.82 ± 1.46</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCE+AEL</td>
<td>90.71 ± 0.04</td>
<td>88.57 ± 0.14</td>
<td>85.01 ± 0.38</td>
<td>77.33 ± 0.18</td>
<td>47.90 ± 1.21</td>
<td></td>
</tr>
</tbody>
</table>

**Asymmetric Loss Functions for Learning with Noisy Labels**

- **Methods**: Clean (\(\eta = 0.0\)) and has an obvious gap from 1 when \(\eta > 0\). Symmetric and has an obvious gap from 1 when \(\eta = 1\) when \(\eta = 0.2\). Test accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise (\(\eta \in \{0.2, 0.4, 0.6, 0.8\}\)). The results (mean±std) are reported over 3 random runs and the top 3 best results are **boldfaced**.

- **Validation of Corollaries**: We also design a simple experiment to validate the necessary and sufficient conditions in Corollaries 1, 2 and 3, where we randomly generate a positive weight vector \(w \in \mathbb{R}^k\) (\(k\) is set to 10), and initialize a random variable \(z \in \mathbb{R}^k\). Our goal is to optimize \(z\) by minimizing \(\sum_{i=1}^{k} w_i L(\sigma(z), i)\), where \(\sigma(\cdot)\) denotes the softmax function.

- **About Hyper-parameters**: We then run a set of experiments on CIFAR-10 to verify the robustness of asymmetric loss functions AGCE, AUL, and AEL with different hyperparameter settings. The label noise is set to be symmetric and the noise rate is set to 0.8. We use an 8-layer CNN as the model to be learned.

- **One of the advantages of asymmetric losses** is that we do not need to know the exact values of noise rates, especially for completely asymmetric losses. For the symmetric noise with rate 0.8, the clean level \(c = \min_{i \in S} \sum_{j \not\in S} \frac{1}{n} = 0\). To make the AGCE asymmetric with \(q = 0.5\), we need to guarantee \(\frac{9}{4} \cdot \left(\frac{a}{a+1}\right)^{1-0.5} \geq 1\), i.e., \(a \geq 16\). To make the AUL asymmetric with \(p = 2\), we need to guarantee \(\frac{9}{4} \cdot \left(\frac{a-1}{a}\right)^{2-1} \geq 1\), i.e., \(a \geq 9\). To make the AEL asymmetric, we need to guarantee \(\frac{9}{4} \geq \exp(1/a)\), i.e., \(a \geq 1/\ln 2\).

- **As shown in Figures 3(a) and 3(b)**, when \(q = 1.5\), all the curves remain robust, and when \(q = 0.5\), the AGCE whose asymmetry ratio is smaller than \(16\) exhibits significant overfitting after epoch 20. In Figure 3(a), although the curve is not robust on \(a = 0.3 > \frac{16}{15}\), it will be more and more
robust as \( \sigma \) increases gradually. Our understanding is that the data is not ideal enough such that the optimization is a trade-off between sample separability and the asymmetry of loss. Similar experimental phenomena have occurred on AUL and AEL. According to Figures 3(c) and 3(d), when \( p < 1 \), AUL always remains robust, and when \( p = 2 \), AUL becomes more and more robust as the asymmetric ratio \( r(\ell) = \frac{a-2}{a} \) gets larger, which is similar to AEL.

**Remark.** An important experimental conclusion is that as \( \sigma \) or the asymmetric ratio \( r \) increases, whether it is for AGCE (\( q < 1 \)), AUL (\( p > 1 \)), or AEL, \( c \cdot r(L) \) is becoming larger, and the training process shows more robust results, but may lead to less fitting ability. Therefore, we roughly follow a principled approach for hyper-parameter tuning: for simple datasets, we prefer the hyperparameters with a higher asymmetry ratio to obtain robustness, while for complicated datasets we tend to use hyper-parameters with a lower asymmetry ratio to obtain better fitting ability.

### 4.2. Evaluation on Benchmark Datasets

**Baselines.** We consider several state-of-the-art methods: Generalized Cross Entropy (GCE) (Zhang & Sabuncu, 2018), Negative Learning for Noisy Labels (NLNL) (Kim et al., 2019), Symmetric Cross Entropy (SCE) (Wang et al., 2019b), Normalized Cross Entropy (NCE), the weighting of NCE and Reverse Cross Entropy (RCE), as well as our proposed AUL, AGCE and AEL. Inspired by the Active Passive Loss (Ma et al., 2020), we combine the proposed AGCE, AUL, and AEL with NCE, then we obtain NCE+ALFs, i.e., NCE+AGCE, NCE+AUL and NCE+AEL. We also train networks using the commonly-used losses Cross Entropy and Focal Loss (Lin et al., 2017).

**Experimental Details.** The noise generation, networks, training details, hyper-parameter settings and more experimental results can be found in the supplementary material.

**Results.** Tables 1 and 2 report the test accuracy results of each loss function on the benchmark datasets with symmetric label noise and asymmetric label noise, respectively. As we can see, our proposed AGCE, AUL and AEL have a significant improvement in most label noise settings for MNIST and CIFAR-10. For example, compared with GCE, SCE, NLNL and NCE, AUL achieves better test accuracy on MNIST and CIFAR-10 for symmetric noise with any noise rate and asymmetric label noise, respectively. As we can see, our proposed AGCE, AUL and AEL have a significant improvement in most label noise settings for MNIST and CIFAR-10. For example, compared with GCE, SCE, NLNL and NCE, AUL achieves better test accuracy on MNIST and CIFAR-10 for symmetric noise with any noise rate and asymmetric label noise with noise rate \( \eta \in \{0.1, 0.2, 0.3\} \). However, in our limited parameter tuning, ALFs suffer from underfitting with asymmetric label noise with noise rate \( \eta = 0.4 \). According to Theorems 3 and 5, the proposed asymmetric loss functions can be applied to the APL framework (Ma et al., 2020). And our NCE+ALFs, especially NCE+AGCE and NCE+AUL, achieve the top three best results in most test scenarios across all datasets. In several cases, our method are better than all baseline methods. The results demonstrate that asymmetric loss functions can be robust enough to get the outstanding performance for both symmetric and asymmetric label noise.

**Visualization.** We further investigate the feature representations learned by AGCE compared to that learned by GCE. We first extract the high-dimensional features at the second last layer, then project all features of test samples in to 2D embeddings by t-SNE (Van der Maaten & Hinton, 2008). The projected representations on MNIST with different symmetric noise (\( \eta \in \{0.0, 0.2, 0.4, 0.6, 0.8\} \)) by t-SNE (Van der Maaten & Hinton, 2008) 2D embeddings of deep features.

Figure 5. Visualization for GCE (top) and AGCE (bottom) on MNIST with different symmetric noise (\( \eta \in \{0.0, 0.2, 0.4, 0.6, 0.8\} \)) by t-SNE (Van der Maaten & Hinton, 2008) 2D embeddings of deep features.
Table 2. Test accuracies (%) of different methods on benchmark datasets with asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The results (mean±std) are reported over 3 random runs and the top 3 best results are boldfaced.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>Asymmetric Noise Rate ((\eta))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>MNIST</td>
<td>CE</td>
<td>97.57 ± 0.22</td>
</tr>
<tr>
<td></td>
<td>FL</td>
<td>97.58 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>GCE</td>
<td>99.01 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>NLNL</td>
<td>98.63 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>SCE</td>
<td>99.14 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>NCE</td>
<td>98.94 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>NCE+RCE</td>
<td>99.35 ± 0.03</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>AUL</td>
<td>99.15 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>AGCE</td>
<td>99.10 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>AEL</td>
<td>98.99 ± 0.05</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>CE</td>
<td>87.55 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>FL</td>
<td>86.43 ± 0.30</td>
</tr>
<tr>
<td></td>
<td>GCE</td>
<td>88.33 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>NCE</td>
<td>89.77 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>NCE+RCE</td>
<td>96.06 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>NCE+AGCE</td>
<td>90.06 ± 0.13</td>
</tr>
<tr>
<td></td>
<td>AUL</td>
<td>90.19 ± 0.16</td>
</tr>
<tr>
<td></td>
<td>AGCE</td>
<td>88.08 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>AEL</td>
<td>85.22 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>NCE+AGCE</td>
<td>90.05 ± 0.20</td>
</tr>
<tr>
<td></td>
<td>NCE+AGCE</td>
<td>90.35 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>NCE+AEL</td>
<td>89.95 ± 0.04</td>
</tr>
</tbody>
</table>

4.3. Evaluation on Real-world Noisy Label

To evaluate the effectiveness of asymmetric loss functions, we test on the real-world noisy dataset WebVision (Li et al., 2017), where we follow the “Mini” setting in (Jiang et al., 2018; Ma et al., 2020) that only takes the first 50 concepts of the Google resized image subset as the training dataset and further evaluate the trained ResNet-50 (He et al., 2016) on the same 50 concepts of the corresponding validation set.

Table 3. Top-1 validation accuracies (%) on WebVision validation set using different loss functions.

<table>
<thead>
<tr>
<th>Loss</th>
<th>CE</th>
<th>GCE</th>
<th>SCE</th>
<th>NCE+RCE</th>
<th>NCE+AGCE</th>
<th>AGCE</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>66.96</td>
<td>61.76</td>
<td>66.92</td>
<td>66.32</td>
<td>67.12</td>
<td>69.40</td>
<td></td>
</tr>
</tbody>
</table>

The top-1 validation accuracies under different loss functions on the clean WebVision validation set are reported in Table 3. More experimental details and results can be found in supplementary materials. As shown in Table 3, the proposed loss functions AGCE and NCE+AGCE outperform the existing loss functions GCE, SCE, and NCE+RCE. The results demonstrate that asymmetric loss functions can help the trained model against real-world label noise.

5. Conclusion

This paper introduces asymmetric loss functions, which allow training a noise-tolerant classifier with noisy labels as long as clean labels dominate. We then prove that completely asymmetric losses are classification-calibrated, and have an excess risk bound when the asymmetry is strict. Furthermore, we introduce the asymmetric ratio to measure the asymmetry. The empirical results demonstrate that the larger ratio will provide better robustness. We also prove asymmetric loss functions will provide a global clean weighted-risk when minimizing the noisy risk for any hypothesis class. The experiments on benchmark datasets show the advantage of using the modified loss functions.

Acknowledgement

This work was supported by National Key R&D Program of China under Grant 2018AAA0102801 and 2019YFE0109600, and National Natural Science Foundation of China under Grants 61922027 and 61932022.
References


