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.

[1]Harbin Institute of Technology [2]Peng Cheng Laboratory [3]King Abdullah University of Science and Technology [4]Tsinghua University. Correspondence to: Xianming Liu <csxm@hit.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

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 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 $\mathcal{Y} = [k] = \{1, ..., k\}$ 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, $\mathcal{S} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, where $(\mathbf{x}_i, y_i)$ 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 $h(\mathbf{x}) = \arg\max_i f(\mathbf{x})_i$, where $f : \mathcal{X} \to \mathcal{C}, \mathcal{C} \subseteq [0,1]^k, \forall\, \mathbf{c} \in \mathcal{C}, \mathbf{1}^T\mathbf{c} = 1$. $f(\mathbf{x})$ denotes an approximation of $p(\cdot|\mathbf{x})$, which is considered as a neural network ending with a softmax layer in this work.

We define a loss function as a mapping $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$, where $\arg\min_{\mathbf{u}\in\mathcal{C}} L(\mathbf{u}, y) = \mathbf{e}_y$, $L(\mathbf{u}, y)$ is monotonically decreasing on the prediction probability u_y of class y, and $\mathbf{e}_y$ denotes a one-hot vector. The L-risk for the hypothesis f is defined as

$$R_L(f) = \mathbb{E}_{\mathcal{D}}[L(f(\mathbf{x}), y)] = \mathbb{E}_{\mathbf{x},y}[L(f(\mathbf{x}), y)], \tag{1}$$

where $\mathbb{E}$ is denoted as expectation operator. Under the risk minimization framework, our objective is to learn a optimal classifier, f^*, which is a global minimum of $R_L(f)$.

2.2. Label Noise Model

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

$$\tilde{y}_n = \begin{cases} i,\ i \in [k],\ i \neq y_n & \text{with probability } \eta_{\mathbf{x}_n,i} \\ y_n & \text{with probability } (1 - \eta_{\mathbf{x}_n}) \end{cases}, \tag{2}$$

where $\eta_{\mathbf{x}_n,i}$ denotes the probability of flipping the true label y_n into i for $\mathbf{x}_n$, and $\eta_{\mathbf{x}_n} = \sum_{i\neq y_n} \eta_{\mathbf{x}_n,i}$ denotes the noise ratio of $\mathbf{x}_n$. 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.*, $\eta_{\mathbf{x}_n}$ and $\eta_{\mathbf{x}_n,i}$ are only dependent on the class labels, which can be then represented as a label transition matrix. If $\eta_{\mathbf{x}_n,i} = \eta$, $\forall \mathbf{x}_n, i$, 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 $\forall n,\ \eta_{\mathbf{x}_n} = \eta$ and $\exists i \neq y_n,\ \forall j \neq y_n \& i,\ \eta_{\mathbf{x}_n,i} > \eta_{\mathbf{x}_n,j}$, *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

$$R_L^\eta(f) = \mathbb{E}_{\mathcal{D}}\big[(1 - \eta_{\mathbf{x}})L(f(\mathbf{x}), y) + \sum_{i\neq y} \eta_{\mathbf{x},i} L(f(\mathbf{x}), i)\big].$$

It can be found that, due to the presence of noisy labels, the classifier learning process is influenced by $\sum_{i\neq y}\eta_{\mathbf{x},i}L(f(\mathbf{x}),i)$, *i.e.*, noisy labels would degrade the generalization performance of deep neural networks. Define f_η^* be the global minimum of $R_L^\eta(f)$, then L is noise-tolerant if f_η^* 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

$$\sum_{i=1}^{k} L(f(\mathbf{x}),i) = C,\ \forall x \in \mathcal{X}, \forall f, \tag{3}$$

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 $L(\mathbf{u},i) = \|\mathbf{e}_i - \mathbf{u}\|_1 = 2-2u_i$ and obviously satisfies $\sum_i L(\mathbf{u},i) = 2k-2$. 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 $\mathcal{S}$, the corresponding observed noisy dataset is $\tilde{\mathcal{S}}$. The i-th class subset of $\tilde{\mathcal{S}}$ is formulated as $\tilde{\mathcal{S}}_i = \{(\mathbf{x},\tilde{y}) : y = i, (\mathbf{x},\tilde{y}) \in \tilde{\mathcal{S}}, (\mathbf{x}, y) \in \mathcal{S}\}$, with $i \in [k]$. We define that the class label i is dominant in $\tilde{\mathcal{S}}_i$ if it satisfies*

$$\sum_{(\mathbf{x},\tilde{y})\in\tilde{\mathcal{S}}_i} \mathbb{I}(\tilde{y} = i) > \max_{j\neq i} \sum_{(\mathbf{x},\tilde{y})\in\tilde{\mathcal{S}}_i} \mathbb{I}(\tilde{y} = j), \tag{4}$$

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

The dataset $\tilde{\mathcal{S}}$ 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 corroboration that $\tilde{\mathcal{S}}$ 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 is clean-labels-dominant, i.e., it satisfies that $\forall \mathbf{x}$, $1 - \eta_{\mathbf{x}} > \max_{j\neq y} \eta_{\mathbf{x},j}$.*

Compared with the symmetric noise assumption behind symmetric losses (Ghosh et al., 2017), *i.e.*, $1 - \eta_{\mathbf{x}} > \frac{1}{k}$, 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 $(\mathbf{x}, y)$ drawn from $\mathcal{D}$, we have the *conditional L-risk* (Bartlett et al., 2006):

$$L^\eta(f(\mathbf{x}),y) = (1-\eta_{\mathbf{x}})L(f(\mathbf{x}),y) + \sum_{i\neq y}\eta_{\mathbf{x},i}L(f(\mathbf{x}),i).$$

The exact values of $\{\eta_{\mathbf{x},i}\}_{i\neq y}$ are usually unknown, and what we only know is $1-\eta_{\mathbf{x}} > \max_{i\neq y}\eta_{\mathbf{x},i}$ 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 $w_1, ..., w_k \geq 0$, where $\exists t \in [k]$, s.t., $w_t > \max_{i\neq t} w_i$, a loss function $L(\mathbf{u}, i)$ is called asymmetric if L satisfies*

$$\arg\min_{\mathbf{u}} \sum_{i=1}^{k} w_i L(\mathbf{u}, i) = \arg\min_{\mathbf{u}} L(\mathbf{u}, t), \tag{5}$$

where we always have $\arg\min_{\mathbf{u}} L(\mathbf{u}, t) = \mathbf{e}_t$.

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

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*.

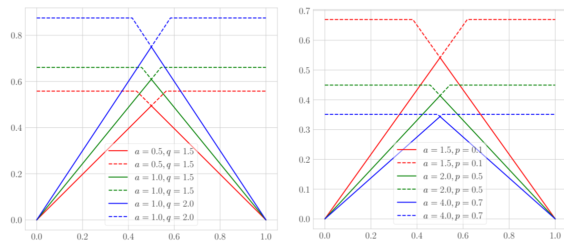
More specifically, according to Definition 2 and Assumption 1, asymmetric loss functions are inherently noise-tolerant, which can eliminate the contribution of noisy labels (*i.e.*, $\sum_{i\neq y}\eta_{\mathbf{x},i}L(f(\mathbf{x}), i)$) in the process of classifier learning. It is desirable since it provides an approach of obtaining the minimum for noise-free case $L(\mathbf{u}, t)$ from the minimization for noisy case $\sum_{i=1}^{k} w_i L(\mathbf{u}, i)$. 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 $L(\mathbf{u}, i)$ be asymmetric on the label noise model which is clean-labels-dominant. According to the asymmetric condition (5), it can be derived that $(1-\eta_{\mathbf{x}})L(\mathbf{u}, y) + \sum_{i\neq y}\eta_{\mathbf{x},i}L(\mathbf{u}, i) \geq (1-\eta_{\mathbf{x}})L(\mathbf{u}^*, y) + \sum_{i\neq y}\eta_{\mathbf{x},i}L(\mathbf{u}^*, i)$, where $\mathbf{u}^* = \mathbf{e}_y$, and the equality holds if and only if $\mathbf{u} = \mathbf{u}^*$. 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 ϕ is classification-calibrated if driving the excess risk over the Bayes-optimal predictor for ϕ 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 $\mathbb{I}(\eta_{\mathbf{x}} > \frac{1}{2})$ (see more in supplementary materials).



(a) AGCE (b) AUL

Figure 1. Verification of classification calibration. Solid and dashed lines denote the curve of $H_\ell(\eta)$ and $H_\ell^-(\eta)$, respectively. As can be observed, the curve of H_ℓ^- is always above that of H_ℓ, *i.e.*, the loss functions are classification-calibrated.

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 $L(\mathbf{u}, i) = \ell(u_i)$ can be expressed as*

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)}, \tag{6}$$

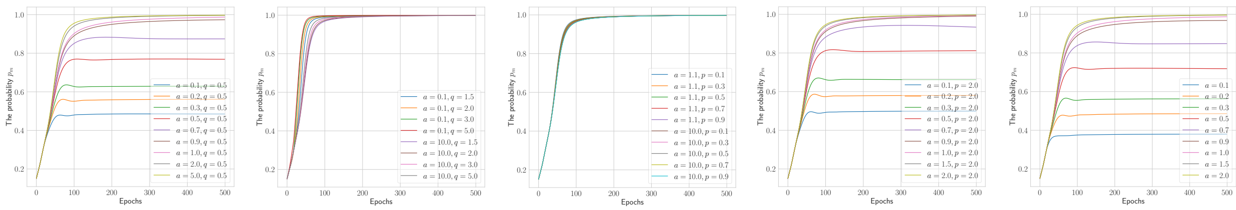
where $R^*_{\ell_{0-1}} = \inf_g R_{\ell_{0-1}}(g)$ and $R^*_\ell = \inf_g R_\ell(g)$.

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 $\mathcal{D}$ perfectly. Here we use deep networks as the hypothesis class to obtain enough fitting ability (Zhang et al., 2017; Zou & Gu, 2019).



(a) AGCE ($q < 1$) (b) AGCE ($q > 1$) (c) AUL ($p < 1$) (d) AUL ($p > 1$) (e) AEL

Figure 2. The validation for necessary and sufficient conditions of AGCE, AUL and AEL, where $m = \arg\max_i w_i$, $n = \arg\max_{i \neq m} w_i$, and a^* is the value such that $\frac{w_m}{w_n} \cdot r(\ell) = 1$ for different loss functions.

Assumption 2. *Given the loss function L and a separable distribution $\mathcal{D}$, we assume that there exists a hypothesis $f : \mathcal{X} \rightarrow \mathcal{C}$, $f \in \mathcal{H}_{net}$, $\forall(\mathbf{x}, y)$ drawn from $\mathcal{D}$, such that f minimizes $L(f(\mathbf{x}), y)$.*

To satisfy this assumption, the hypothesis class $\mathcal{H}_{net}$ 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, $\mathcal{H}_{net}$ 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:

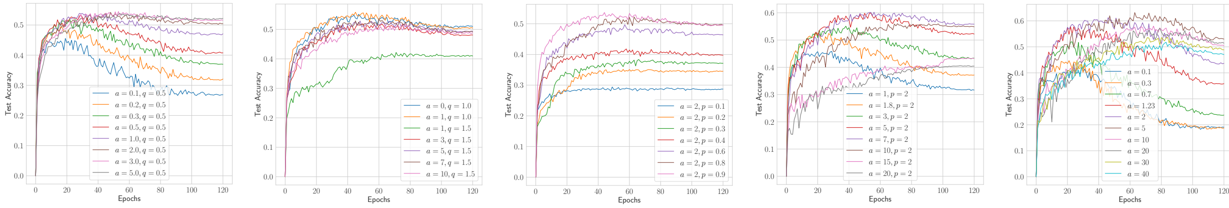
Definition 3. *Consider a loss function $L(\mathbf{u}, i) = \ell(u_i)$, we define the asymmetry ratio $r(\ell)$ as*

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

The asymmetry ratio r denotes the infimum ratio of change in the loss function ℓ 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 $w_1, .., w_k$, where $w_m > w_n$ and $w_n = \max_{i \neq m} w_i$, the loss function $L(\mathbf{u}, i) = \ell(u_i)$ is asymmetric if $\frac{w_m}{w_n} \cdot r(\ell) \geq 1$.*

Remark. Let $c = \min_{(\mathbf{x},y), i \neq y} \frac{1-\eta_{\mathbf{x}}}{\eta_{\mathbf{x},i}}$, 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 ℓ 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 $r(\ell) \geq \frac{1}{1.5}$ 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.



(a) AGCE ($q = 0.5$) (b) AGCE ($q \geq 1$) (c) AUL ($p < 1$) (d) AUL ($p = 2$) (e) AEL)

Figure 3. Test accuracies of AGCE, AUL and AEL with different parameters on CIFAR-10 under 0.8 symmetric noise.

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 $f^* = \arg\min_{f\in\mathcal{H}} R_L^\eta(f)$. If $\forall \mathbf{x}$, we have $\frac{1-\eta_{\mathbf{x}}}{\eta_{\mathbf{x}}} \cdot r(L) > 1$, then f^* also minimizes a positive weighted L-risk $R_{w,L}(f) = \mathbb{E}[w(\mathbf{x}, y) \cdot L(f(\mathbf{x}), y)]$.*

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

$$r(\ell) \leq \inf_{\substack{0\leq u_1,u_2\leq 1\\ u_1+u_2=1}} \frac{\ell(u_1) - \ell(1)}{\ell(0) - \ell(u_2)} = r_u(\ell). \quad (8)$$

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 ℓ satisfies $r_u(\ell) \geq 1$.

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

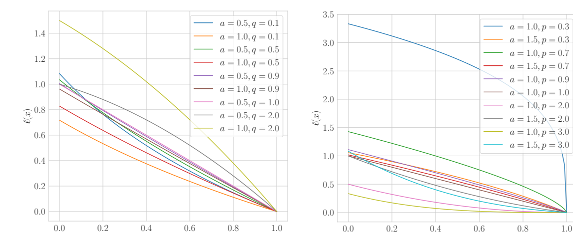
According to Theorem 6 and Theorem 8, when $r(\ell) = r_u(\ell)$, $\frac{w_m}{w_n} \cdot r(\ell) \geq 1$ will become the necessary and sufficient condition for $L(\mathbf{u}, i) = \ell(u_i)$ to be asymmetric. The following corollaries are straightforward from this.

Corollary 1. *On the given weights $w_1, .., w_k$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the loss function $L_q(\mathbf{u}, i) = [(a+1)^q - (a+u_i)^q]/q$ (where $q > 0$, $a > 0$) is asymmetric $\Leftrightarrow \frac{w_m}{w_n} \geq (\frac{a+1}{a})^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1)$.*

Mathematically, the loss function L_q, 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 the r and r_u of AGCE are equal. More specifically, r is equal to $(\frac{a}{a+1})^{1-q}$ 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 $\frac{w_m}{w_n} \geq (\frac{a+1}{a})^{1-q}$, 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 $w_1, .., w_k$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the loss function $L_p(\mathbf{u}, i) = [(a - u_i)^p - (a-1)^p]/p$ (where $p > 0$ and $a > 1$) is asymmetric $\Leftrightarrow \frac{w_m}{w_n} \geq (\frac{a}{a-1})^{p-1} \cdot \mathbb{I}(p > 1) + \mathbb{I}(p \leq 1)$.*

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



(a) AGCEs (b) AULs

Figure 4. Illustration of asymmetric loss functions.

Corollary 3. *On the given weights $w_1, .., w_k$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the exponential loss function $L_a(\mathbf{u}, i) = \exp(-u_i/a)$ (where $a > 0$) is asymmetric $\Leftrightarrow \frac{w_m}{w_n} \geq \exp(1/a)$.*

We call the convex loss function above the Asymmetric Exponential Loss (AEL). According to the Corollary 3, we know that both r and r_u of AEL are equal to $\exp(-1/a) \leq 1$, so AELs will not be completely asymmetric.

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**.

Datasets	Methods	Clean ($\eta = 0.0$)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.15 ± 0.05	91.62 ± 0.39	73.98 ± 0.27	49.36 ± 0.43	22.66 ± 0.61
	FL	99.13 ± 0.09	91.68 ± 0.14	74.54 ± 0.06	50.39 ± 0.28	22.65 ± 0.26
	GCE	99.27 ± 0.05	98.86 ± 0.07	97.16 ± 0.03	81.53 ± 0.58	33.95 ± 0.82
	NLNL	98.61 ± 0.13	98.02 ± 0.14	97.17 ± 0.09	95.42 ± 0.30	86.34 ± 1.43
	SCE	99.23 ± 0.10	98.92 ± 0.12	97.38 ± 0.15	88.83 ± 0.55	48.75 ± 1.54
	NCE	98.60 ± 0.06	98.57 ± 0.01	98.29 ± 0.05	97.65 ± 0.08	93.78 ± 0.41
	NCE+RCE	99.36 ± 0.05	**99.14 ± 0.03**	98.51 ± 0.06	95.60 ± 0.21	74.00 ± 1.68
	AUL	99.14 ± 0.05	**99.05 ± 0.09**	**98.90 ± 0.09**	**98.67 ± 0.04**	**96.73 ± 0.20**
	AGCE	99.05 ± 0.11	**98.96 ± 0.10**	**98.83 ± 0.06**	**98.57 ± 0.12**	**96.59 ± 0.12**
	AEL	99.03 ± 0.05	98.93 ± 0.06	**98.78 ± 0.13**	**98.51 ± 0.06**	**96.40 ± 0.11**
CIFAR10	CE	90.48 ± 0.11	74.68 ± 0.25	58.26 ± 0.21	38.70 ± 0.53	19.55 ± 0.49
	FL	89.82 ± 0.20	73.72 ± 0.08	57.90 ± 0.45	38.86 ± 0.07	19.13 ± 0.28
	GCE	89.59 ± 0.26	87.03 ± 0.35	82.66 ± 0.17	67.70 ± 0.45	26.67 ± 0.59
	SCE	91.61 ± 0.19	87.10 ± 0.25	79.67 ± 0.37	61.35 ± 0.56	28.66 ± 0.27
	NLNL	90.73 ± 0.20	73.70 ± 0.05	63.90 ± 0.44	50.68 ± 0.47	29.53 ± 1.55
	NCE	75.65 ± 0.26	72.89 ± 0.25	69.49 ± 0.39	62.64 ± 0.18	41.49 ± 0.66
	NCE+RCE	90.87 ± 0.37	**89.25 ± 0.42**	**85.81 ± 0.08**	**79.72 ± 0.20**	**55.74 ± 0.95**
	AUL	91.27 ± 0.12	89.21 ± 0.09	85.64 ± 0.19	78.86 ± 0.66	**52.92 ± 1.20**
	AGCE	88.95 ± 0.22	86.98 ± 0.12	83.39 ± 0.17	76.49 ± 0.53	44.42 ± 0.74
	AEL	86.38 ± 0.19	84.27 ± 0.12	81.12 ± 0.20	74.86 ± 0.22	51.41 ± 0.32
	NCE+AUL	91.10 ± 0.13	**89.31 ± 0.20**	**86.23 ± 0.18**	**79.70 ± 0.08**	**59.44 ± 1.14**
	NCE+AGCE	90.94 ± 0.12	**89.21 ± 0.08**	**86.19 ± 0.15**	**80.13 ± 0.18**	50.82 ± 1.46
	NCE+AEL	90.71 ± 0.04	88.57 ± 0.14	85.01 ± 0.38	77.33 ± 0.18	47.90 ± 1.21
CIFAR100	CE	71.33 ± 0.43	56.51 ± 0.39	39.92 ± 0.10	21.39 ± 1.17	7.59 ± 0.20
	FL	70.06 ± 0.70	55.78 ± 1.55	39.83 ± 0.43	21.91 ± 0.89	7.51 ± 0.09
	GCE	63.09 ± 1.39	61.57 ± 1.06	56.11 ± 1.35	45.28 ± 0.61	17.42 ± 0.06
	SCE	69.62 ± 0.42	52.25 ± 0.14	36.00 ± 0.69	20.14 ± 0.60	7.67 ± 0.63
	NLNL	68.72 ± 0.60	46.99 ± 0.91	30.29 ± 1.64	16.60 ± 0.90	11.01 ± 2.48
	NCE	29.96 ± 0.73	25.27 ± 0.32	19.54 ± 0.52	13.51 ± 0.65	8.55 ± 0.37
	NCE+RCE	68.65 ± 0.40	64.97 ± 0.49	58.54 ± 0.13	45.80 ± 1.02	**25.41 ± 0.98**
	NCE+AUL	68.96 ± 0.16	**65.36 ± 0.20**	**59.25 ± 0.23**	**46.34 ± 0.21**	23.03 ± 0.64
	NCE+AGCE	69.03 ± 0.37	**65.66 ± 0.46**	**59.47 ± 0.36**	**48.02 ± 0.58**	**24.72 ± 0.60**
	NCE+AEL	68.70 ± 0.20	**65.36 ± 0.14**	**59.51 ± 0.03**	**46.94 ± 0.07**	**24.48 ± 0.24**

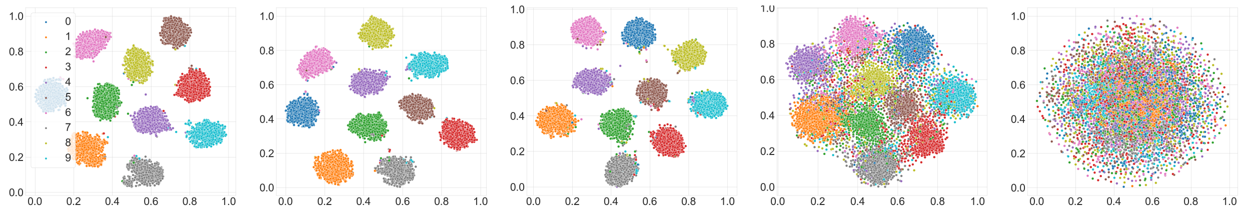
tary material) for the proposed losses AGCE with $q > 1$ and AUL with $p < 1$, which are completely asymmetric according to Corollaries 1 and 2. As shown in Fig. 1, under the same parameter setting, the curve of $H_\ell^-(\eta)$ (dashed ones) is always above the corresponding of $H_\ell(\eta)$ (solid ones) when $\eta \neq 1/2$, *i.e.*, the loss functions are classification-calibrated.

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 $\mathbf{w} \in \mathbb{R}_+^k$ (k is set to 10), and initialize a random variable $\mathbf{z} \in \mathbb{R}^k$. Our goal is to optimize $\mathbf{z}$ by minimizing $\sum_{i=1}^k w_i L(\sigma(\mathbf{z}), i)$, where $\sigma(\cdot)$ denotes the softmax function. As aforementioned, asymmetric loss functions will optimize $\mathbf{p} = \sigma(\mathbf{z})$ as a one-hot vector. The experimental results are shown in Figure 2. Let $m = \arg\max_i w_i$, we can see that when $\frac{w_m}{w_n} \cdot r(\ell) \geq 1$, p_m is very close to 1, and has an obvious gap from 1 when $\frac{w_m}{w_n} \cdot r(\ell) < 1$. The consequence holds regardless of AGCE, AUL, or AEL, and an important phenomenon is that the curve is more and more asymmetric as a or $r(\ell)$ gets bigger and bigger.

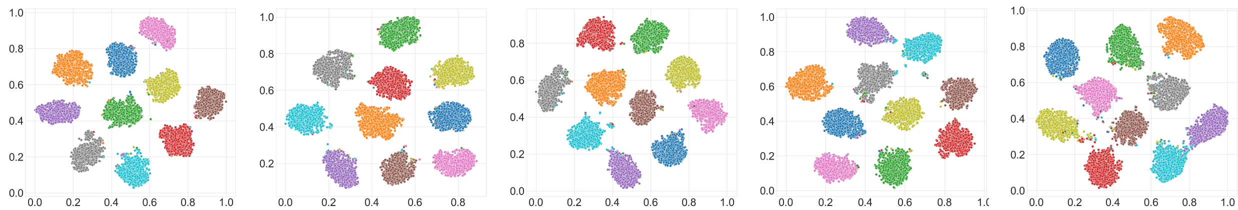
About Hyper-parameters. We then run a set of experiments on CIFAR-10 to verify the robustness of asymmetric loss functions AGCE, AUL, and AEL with different hyper-parameter settings. The label noise is set to be symmetric and the noise rate is set to 0.8. We use an 8-layer CNN as the model to be learned.

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 $c = \min_{(\mathbf{x},y)\in\mathcal{S}, i\neq y} \frac{1-\eta_\mathbf{x}}{\eta_{\mathbf{x},i}} = \frac{9}{4}$. To make the AGCE asymmetric with $q = 0.5$, we need to guarantee $\frac{9}{4} \cdot (\frac{a}{a+1})^{1-0.5} \geq 1$, *i.e.*, $a \geq \frac{16}{65}$. To make the AUL asymmetric with $p = 2$, we need to guarantee $\frac{9}{4} \cdot (\frac{a-1}{a})^{2-1} \geq 1$, *i.e.*, $a \geq \frac{9}{5}$. To make the AEL asymmetric, we need to guarantee $\frac{9}{4} \geq \exp(1/a)$, *i.e.*, $a \geq 1/\ln\frac{9}{4}$.

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 $\frac{16}{65}$ exhibits significant overfitting after epoch 20. In Figure 3(a), although the curve is not robust on $a = 0.3 > \frac{16}{65}$, it will be more and more



(a) GCE with $\eta = 0.0$ (b) GCE with $\eta = 0.2$ (c) GCE with $\eta = 0.4$ (d) GCE with $\eta = 0.6$ (e) GCE with $\eta = 0.8$



(f) AGCE with $\eta = 0.0$ (g) AGCE with $\eta = 0.2$ (h) AGCE with $\eta = 0.4$ (i) AGCE with $\eta = 0.6$ (j) AGCE with $\eta = 0.8$

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.

robust as a 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 $r(\ell) = \frac{a-1}{a}$ gets larger, which is similar to AEL.

Remark. An important experimental conclusion is that as a 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 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 $\eta = 0.4$. 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 label noise are illustrated in Fig. 5. As can be observed, GCE encounters obvious overfitting with label noise, and the embeddings look completely mixed together when $\eta = 0.8$. On the contrary, AGCE learns good representations with more separated and clearly bounded clusters in all noisy cases.

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**.

Datasets	Methods	Asymmetric Noise Rate (η) 0.1	0.2	0.3	0.4
MNIST	CE	97.57 ± 0.22	94.56 ± 0.22	88.81 ± 0.10	82.27 ± 0.40
	FL	97.58 ± 0.09	94.25 ± 0.15	89.09 ± 0.25	82.13 ± 0.49
	GCE	99.01 ± 0.04	96.69 ± 0.12	89.12 ± 0.24	81.51 ± 0.19
	NLNL	98.63 ± 0.06	98.35 ± 0.01	97.51 ± 0.15	95.84 ± 0.26
	SCE	**99.14 ± 0.04**	98.03 ± 0.05	93.68 ± 0.43	85.36 ± 0.17
	NCE	98.49 ± 0.06	98.18 ± 0.12	96.99 ± 0.17	94.16 ± 0.19
	NCE+RCE	**99.35 ± 0.03**	98.99 ± 0.22	97.23 ± 0.20	90.49 ± 4.04
	AUL	**99.15 ± 0.09**	**99.15 ± 0.02**	**98.98 ± 0.05**	**98.62 ± 0.09**
	AGCE	99.10 ± 0.02	**99.07 ± 0.09**	**98.95 ± 0.03**	**98.44 ± 0.11**
	AEL	98.99 ± 0.05	**99.06 ± 0.07**	**98.90 ± 0.15**	**98.34 ± 0.08**
CIFAR10	CE	87.55 ± 0.14	83.32 ± 0.12	79.32 ± 0.59	74.67 ± 0.38
	FL	86.43 ± 0.30	83.37 ± 0.07	79.33 ± 0.08	74.28 ± 0.44
	GCE	88.33 ± 0.05	85.93 ± 0.23	80.88 ± 0.38	74.29 ± 0.43
	SCE	89.77 ± 0.11	86.20 ± 0.37	81.38 ± 0.35	75.16 ± 0.39
	NLNL	88.54 ± 0.25	84.74 ± 0.08	81.26± 0.43	76.97 ± 0.52
	NCE	74.06 ± 0.27	72.46 ± 0.32	69.86 ± 0.51	65.66 ± 0.42
	NCE+RCE	**90.06 ± 0.13**	**88.45 ± 0.16**	**85.42 ± 0.09**	**79.33 ± 0.15**
	AUL	**90.19 ± 0.16**	88.17 ± 0.11	84.87 ± 0.04	56.33 ± 0.07
	AGCE	88.08 ± 0.06	86.67 ± 0.14	83.59 ± 0.15	60.91 ± 0.20
	AEL	85.22 ± 0.15	83.82 ± 0.15	82.43 ± 0.16	58.81 ± 3.62
	NCE+AUL	90.05 ± 0.20	**88.72 ± 0.26**	**85.48 ± 0.18**	**79.26 ± 0.05**
	NCE+AGCE	**90.35 ± 0.15**	**88.48 ± 0.16**	**85.96 ± 0.24**	**80.00 ± 0.44**
	NCE+AEL	89.95 ± 0.04	87.93 ± 0.06	84.81 ± 0.26	77.27 ± 0.11
CIFAR100	CE	64.85 ± 0.37	58.11 ± 0.32	50.68 ± 0.55	40.17 ± 1.31
	FL	64.78 ± 0.50	58.05 ± 0.42	51.15 ± 0.84	**41.18 ± 0.68**
	GCE	63.01 ± 1.01	59.35 ± 1.10	53.83 ± 0.64	40.91 ± 0.57
	SCE	61.63 ± 0.84	53.81 ± 0.42	45.63 ± 0.07	36.43 ± 0.20
	NLNL	59.55 ± 1.22	50.19 ± 0.56	42.81 ± 1.13	35.10 ± 0.20
	NCE	27.59 ± 0.54	25.75 ± 0.50	24.28 ± 0.80	20.64 ± 0.40
	NCE+RCE	66.38 ± 0.16	**62.97 ± 0.24**	**55.38 ± 0.49**	**41.68 ± 0.56**
	NCE+AUL	**66.62 ± 0.09**	**63.86 ± 0.18**	50.38 ± 0.32	38.59 ± 0.48
	NCE+AGCE	**67.22 ± 0.12**	**63.69 ± 0.19**	**55.93 ± 0.38**	**43.76 ± 0.70**
	NCE+AEL	**66.92 ± 0.22**	62.50 ± 0.23	**52.42 ± 0.98**	39.99 ± 0.12

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.

Loss	CE	GCE	SCE	NCE+RCE	**NCE+AGCE**	**AGCE**
Acc	66.96	61.76	66.92	66.32	**67.12**	**69.40**

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

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Bartlett, P. L., Jordan, M. I., and Mcauliffe, J. D. Convexity, classi cation, and risk bounds. *Journal of the American Statistical Association*, 101, 2006.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 961–970, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance- and label-dependent label noise. In *ICML*, 2020.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. Can cross entropy loss be robust to label noise? In *IJCAI*, 2020.

Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93 – 107, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.09.081.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *ICLR (Poster)*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Kim, Y., Yim, J., Yun, J., and Kim, J. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 101–110, 2019.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1, 01 2009.

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

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Liu, W., Jiang, Y.-G., Luo, J., and Chang, S.-F. Noise resistant graph ranking for improved web image search. In *CVPR 2011*, pp. 849–856. IEEE, 2011.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pp. 6226–6236. PMLR, 2020.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 608–615, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054.

Lyu, Y. and Tsang, I. Curriculum loss: Robust learning and generalization against label corruption. *ICLR*, 2020.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

Manwani, N. and Sastry, P. S. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3): 1146–1151, 2013. doi: 10.1109/TSMCB.2012.2223460.

Martin, A. and Peter L., B. Neural network learning: Theoretical foundations. 1999.

Menon, K. A., Rawat, S. A., Reddi, J. S., and Kumar, S. Can gradient clipping mitigate label noise. *ICLR*, 2020.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26, pp. 1196–1204. Curran Associates, Inc., 2013.

Plessis, M. D., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1386–1394, Lille, France, 07–09 Jul 2015. PMLR.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

Tong, Z. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2003.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 10–18. Curran Associates, Inc., 2015.

Wang, X., Kodirov, E., Hua, Y., and Robertson, N. Improving mae against cce under label noise. *ArXiv*, abs/1903.12141, 2019a.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision*, 2019b.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representation*, 2017.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 8778–8788. Curran Associates, Inc., 2018.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 2055–2064. Curran Associates, Inc., 2019.