Can Cross Entropy Loss Be Robust to Label Noise?

Lei Feng[1] , **Senlin Shu**[2*] , **Zhuoyi Lin**[1] , **Fengmao Lv**[3] , **Li Li**[2] , **Bo An**[1]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]College of Computer and Information Science, Southwest University, Chongqing, China
[3]Center of Statistical Research, Southwestern University of Finance and Economics, China
{feng0093,zhuoyi001,boan}@ntu.edu.sg, {ssl2018,lily}@email.swu.edu.cn

Abstract

Trained with the standard cross entropy loss, deep neural networks can achieve great performance on correctly labeled data. However, if the training data is corrupted with label noise, deep models tend to overfit the noisy labels, thereby achieving poor generation performance. To remedy this issue, several loss functions have been proposed and demonstrated to be robust to label noise. Although most of the robust loss functions stem from Categorical Cross Entropy (CCE) loss, they fail to embody the intrinsic relationships between CCE and other loss functions. In this paper, we propose a general framework dubbed Taylor cross entropy loss to train deep models in the presence of label noise. Specifically, our framework enables to weight the extent of fitting the training labels by controlling the order of Taylor Series for CCE, hence it can be robust to label noise. In addition, our framework clearly reveals the intrinsic relationships between CCE and other loss functions, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). Moreover, we present a detailed theoretical analysis to certify the robustness of this framework. Extensive experimental results on benchmark datasets demonstrate that our proposed approach significantly outperforms the state-of-the-art counterparts.

1 Introduction

Deep Neural Networks (DNNs) have achieved great advances over the past years. With proper training, DNNs can easily achieve great classification performance. However, the success of DNNs relies on a large number of high-quality samples during the training process. Unfortunately, incorrect labels in large-scale datasets are often inevitable. In most scenarios, it can be more beneficial to have datasets with more but noisier labels than less but more accurate labels [Khetan *et al.*, 2017]. Therefore, training a robust classifier in the presence of label noise is an increasingly valued task.

In general, (softmax) Categorical Cross Entropy (CCE) loss is the standard loss function used to train deep models. However, past studies [Ghosh *et al.*, 2017; Patrini *et al.*, 2017] show that using the standard CCE to train deep models leads to serious over-fitting (on noisy labels) and results in poor generalization ability. To mitigate this issue, increasing interests have been drawn in exploiting robust loss functions for training deep models against label noise. For example, the works of [Ghosh *et al.*, 2015; Ghosh *et al.*, 2017] reveal that *symmetric* loss functions, in which the sum of the risks over all categories is equivalent to a constant for each arbitrary example, can be robust to label noise. Representative symmetric loss functions include Ramp Loss [Ghosh *et al.*, 2015] and (softmax) Mean Absolute Error (MAE) [Ghosh *et al.*, 2017]. By both theoretical and empirical analysis, symmetric loss functions are demonstrated to be robust to label noise. Recently, several other loss functions that do not satisfy the symmetry condition strictly, including Generalized Cross Entropy (GCE) loss [Zhang and Sabuncu, 2018], Partial Huberised Cross Entropy (PHuber-CE) loss [Menon *et al.*, 2020] and Symmetric Cross Entropy (SCE) loss [Wang *et al.*, 2019], have also been proposed to be robust against label noise when training deep neural networks. Although these loss functions achieve robustness to label noise in different ways, one thing in common is that these loss functions are derived from the standard CCE. In spite of their effectiveness, they all fail to embody the intrinsic relationships between CCE and other loss functions.

Motivated by the above observations, we wonder whether CCE can be further exploited to design a general framework that embody the intrinsic relationships between CCE and other loss functions, for robust learning with label noise. To answer this question, this paper proposes a general robust learning framework to train deep models in the presence of label noise. Specifically, we apply Taylor Series to derive an alternative representation of CCE. Moreover, we can flexibly adjust the order of Taylor Series to approximate CCE, and we call the approximated CCE Taylor Cross Entropy (TCE). In this framework, the order of Taylor Series reflects the extent of how TCE approximates to CCE. Furthermore, by varying the order of Taylor Series, we are able to reveal the intrinsic relationships between CCE and other loss functions. For example, MAE can be considered as the first-order Taylor Series approximation of CCE. The second-order Taylor Series approximation of CCE is an average combination of MAE and a lower bound of Mean Squared Error (MSE). The sufficient-

*Corresponding author.

ly large-order Taylor Series approximation of CCE recovers CCE. In addition, we show that our proposed TCE is upper-bounded with finite order (which means the risk would never be infinite), and TCE is always an upper bound of MAE. In other words, minimizing TCE naturally enables to minimize MAE to some degree.

To sum up, our main contributions are three-fold:

- We propose a general robust learning framework dubbed Taylor Cross Entropy (TCE) loss. Taylor Series is applied to obtain a representation of CCE for training deep models in the presence of label noise.
- We present a detailed theoretical analysis to certify the robustness of TCE against label noise.
- We conduct extensive experiments for learning from symmetric and asymmetric label noise, and compare with a number of loss functions. Extensive experimental results on benchmark datasets demonstrate that our proposed approach significantly outperforms the state-of-the-art counterparts.

2 Related Work

In this section, we briefly review existing works on learning in the presence of label noise.

Noise rate estimation. Some of the early works [Natarajan *et al.*, 2013; Sukhbaatar and Fergus, 2014; Menon *et al.*, 2015; Patrini *et al.*, 2017] aim to estimate the label transition matrix (sometimes called confusion matrix), and use it to train the target model. For this type of approach, the classification performance hinges on the quality of noise rate estimation [Goldberger and Ben-Reuven, 2017; Hendrycks *et al.*, 2018; Han *et al.*, 2018b; Xia *et al.*, 2019]. However, noise rate estimation is challenging, especially on datasets with a huge number of classes.

Robust loss functions. Designing loss functions that are robust to label noise has been received increasing attention from researchers. The first work is from [Ghosh *et al.*, 2015], which demonstrates that binary loss functions that satisfy the symmetric condition $\ell(z) + \ell(-z) = c$ (e.g., ramp loss and sigmoid loss) where c is a constant, are robust to label noise for binary classification. Then, for multi-class classification, loss functions that satisfy the symmetric condition $\sum_{j=1}^{k} \mathcal{L}(f(\boldsymbol{x}), j) = C$ (e.g., MAE) where C is a constant, are demonstrated to be robust to label noise for the multi-class classification [Ghosh *et al.*, 2017]. However, a recent study [Zhang and Sabuncu, 2018] shows that MAE is not able to achieve good performance on complicated datasets, due to its optimization issue. To alleviate this problem, Generalized Cross Entropy (GCE) [Zhang and Sabuncu, 2018] adopts the negative Box-Cox transformation strategy, and uses a hyper-parameter q to balance between MAE and CCE. Partial Huberised Cross Entropy (PHuber-CE) [Menon *et al.*, 2020] corrects CCE on hard examples by gradient clipping. Symmetric Cross Entropy (SCE) [Wang *et al.*, 2019] combines CCE and Reverse Cross Entropy (RCE, which is equivalent to MAE) by tuning the regularization parameters. Although the above loss functions stem from CCE and some of them may be able to recover MAE or CCE by tuning the parameters, they all fail to embody the intrinsic relationships between CCE and other loss functions.

Other deep learning methods. There are some other approaches that adopt other solutions [Wei *et al.*, 2020; Tanaka *et al.*, 2018; Han *et al.*, 2018a; Yu *et al.*, 2019; Berthon *et al.*, 2020; Yang *et al.*, 2019] to deal with noisy labels. For example, MentorNet [Jiang *et al.*, 2017] is trained to supervise the training of a StudentNet with a sample weighting scheme. Co-teaching [Han *et al.*, 2018c] trains two networks simultaneously and enables the two networks to learn from each other. PENCIL [Yi and Wu, 2019] trains neural networks using label probability distributions and updates these distributions in each epoch.

3 Taylor Cross Entropy Loss for Robust Learning with Label Noise

In this section, we first briefly review CCE and MAE. Then, we introduce our proposed Taylor cross entropy loss. Finally, we theoretically analyze the robustness of Taylor cross entropy loss.

3.1 Preliminaries

We consider the problem of k-class classification. Suppose the clean data set is represented as $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \mid 1 \leq i \leq m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ ($\mathcal{X} \in \mathbb{R}^d$) is a d-dimensional feature vector and $y_i \in \{1, \ldots, k\}$ is the label associated with $\boldsymbol{x}_i$. A classifier is a function that maps the feature space to the label space $f : \mathcal{X} \to \mathbb{R}^k$. In this paper, we consider the common case where the function f is a DNN with the softmax output layer. In this way, the commonly used loss functions CCE and MAE can be represented as:

$$\mathcal{L}_{\text{CCE}}(f(\boldsymbol{x}), y) = -\boldsymbol{e}_y \log f(\boldsymbol{x}) = -\log f_y(\boldsymbol{x}),$$
$$\mathcal{L}_{\text{MAE}}(f(\boldsymbol{x}), y) = \|\boldsymbol{e}_y - f(\boldsymbol{x})\|_1 = 2 - 2f_y(\boldsymbol{x}),$$

where the $f_y(\boldsymbol{x})$ denotes the y-th element of $f(\boldsymbol{x})$ and $\boldsymbol{e}_y$ is a one-hot vector with $\boldsymbol{e}_{yj} = 1$ if $j = y$, otherwise 0. Besides, the gradients of CCE and MAE can be shown as:

$$\frac{\partial \mathcal{L}_{\text{CCE}}(f(\boldsymbol{x}), y)}{\partial \boldsymbol{\theta}} = -\frac{1}{f_y(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} f_y(\boldsymbol{x}),$$
$$\frac{\partial \mathcal{L}_{\text{MAE}}(f(\boldsymbol{x}), y)}{\partial \boldsymbol{\theta}} = -2\nabla_{\boldsymbol{\theta}} f_y(\boldsymbol{x}),$$

where $\boldsymbol{\theta}$ is the set of parameters of f. As shown in the above equations, when training with CCE, examples with smaller prediction confidences are weighted more than examples with larger prediction confidences for gradient update. In other words, CCE pays more attention to hard examples. On the contrary, MAE treats all the examples equally. Hence CCE is more preferred than MAE when training with clean data. However, in the presence of label noise, the performance of CCE dramatically drops, since the given labels of hard examples may be incorrect. In the extreme case, the weight becomes infinite if $f_y(\boldsymbol{x}) \to 0$, which leads to serious overfitting on noisy labels. In contrast, MAE does not suffer from this problem, since each example is treated equally. Therefore, there arises a question: can cross entropy loss be robust

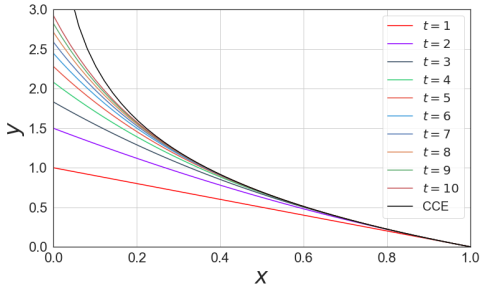


Figure 1: $\mathcal{L}_{t\text{-CE}}$ with different parameters t.

to label noise? To answer this question, we propose a general framework of CCE that endows the robustness to CCE against label noise.

3.2 Taylor Cross Entropy Loss

Given a function $g(\boldsymbol{x})$, if $g(\boldsymbol{x})$ is differentiable to order n at $\boldsymbol{x} = \boldsymbol{x}_0$, then $g(\boldsymbol{x})$ can be written as a Taylor Series:

$$g(\boldsymbol{x}) = \sum_{i=0}^{\infty} \frac{g^{(i)}(\boldsymbol{x}_0)}{i!}(\boldsymbol{x} - \boldsymbol{x}_0)^i, \tag{1}$$

where $g^i(\boldsymbol{x}_0)$ denotes the i-th order derivative of $g(\boldsymbol{x})$ at $\boldsymbol{x}_0$. Recall that $\mathcal{L}_{\text{CCE}}(f(\boldsymbol{x}), y) = -\log f_y(\boldsymbol{x})$, we can define $g(f_y(\boldsymbol{x})) = -\log f_y(\boldsymbol{x})$. Then we have

$$g(f_y(\boldsymbol{x})) = \sum_{i=0}^{\infty} \frac{g^{(i)}(f_y(\boldsymbol{x}_0))}{i!}(f_y(\boldsymbol{x}) - f_y(\boldsymbol{x}_0))^i.$$

If we set $f_y(\boldsymbol{x}_0) = 1$, then $\forall i \geq 1$ we have

$$g^{(i)}(f_y(\boldsymbol{x}_0) = 1) = (-1)^i(i-1)!, \tag{2}$$

Thus we can express $\mathcal{L}_{\text{CCE}}$ as

$$\mathcal{L}_{\text{CCE}}(f(\boldsymbol{x}), y) = g(f_y(\boldsymbol{x})) = \sum_{i=1}^{\infty} \frac{(1 - f_y(\boldsymbol{x}))^i}{i}. \tag{3}$$

Obviously, it is unrealistic to take into account all the terms, since $n \to \infty$. Hence we propose to reserve finite terms in Eq. (3) and obtain an approximation of CCE, which is called Taylor Cross Entropy (TCE) loss:

$$\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) = \sum_{i=1}^{t} \frac{(1 - f_y(\boldsymbol{x}))^i}{i}, \tag{4}$$

where $t \in \mathbb{N}_+$ is a hyper-parameter that denotes the order of the Taylor Series, in other words, the proximity to the CCE. Figure 1 illustrates $\mathcal{L}_{t\text{-CE}}$ with t varying in $\{1, \ldots, 10\}$. According to the definition of $\mathcal{L}_{t\text{-CE}}$ and Figure 1, we can obtain some interesting deductions.

Theorem 1. *Our proposed Taylor cross entropy loss has the following properties: 1) When $t = 1$, $\mathcal{L}_{t\text{-CE}} = \frac{1}{2}\mathcal{L}_{MAE}$; 2) When $t = 2$, $\mathcal{L}_{t\text{-CE}}$ is an average combination of $\mathcal{L}_{MAE}$ and a lower bound of $\mathcal{L}_{MSE}$; 3) When $t \to \infty$, $\mathcal{L}_{t\text{-CE}}$ is equivalent to $\mathcal{L}_{CCE}$; 4) $\mathcal{L}_{MAE} \leq 2\mathcal{L}_{t\text{-CE}}$, $\forall t \in \mathbb{N}_+$.*

Proof. 1) When $t = 1$, $\mathcal{L}_{t\text{-CE}}$ is represented as

$$\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) = 1 - f_y(\boldsymbol{x}).$$

It is clear that in this case, $\mathcal{L}_{t\text{-CE}} = \frac{1}{2}\mathcal{L}_{\text{MAE}}$, which means, minimizing $\mathcal{L}_{t\text{-CE}}$ is equivalent to minimizing $\mathcal{L}_{\text{MAE}}$.

2) When $t = 2$, $\mathcal{L}_{t\text{-CE}}$ is represented as

$$\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) = (1 - f_y(\boldsymbol{x})) + \frac{(1 - f_y(\boldsymbol{x}))^2}{2}.$$

For the first term, we have already shown $(1 - f_y(\boldsymbol{x})) = \frac{1}{2}\mathcal{L}_{\text{MAE}}$. For the second term, we have

$$\begin{aligned}\frac{(1 - f_y(\boldsymbol{x}))^2}{2} &= \frac{1}{2}(1 - 2f_y(\boldsymbol{x}) + (f_y(\boldsymbol{x}))^2) \\ &\leq \frac{1}{2}(1 - 2f_y(\boldsymbol{x}) + \|f(\boldsymbol{x})\|_2^2) \\ &= \frac{1}{2}\|f(\boldsymbol{x}) - \boldsymbol{e}_y\|_2^2 = \frac{1}{2}\mathcal{L}_{\text{MSE}}.\end{aligned}$$

Hence the second term is a lower bound of MSE.

3) According to the definition of $\mathcal{L}_{t\text{-CE}}$, it is clear that $\mathcal{L}_{t\text{-CE}}$ will be equivalent to $\mathcal{L}_{\text{CCE}}$ if $t \to \infty$.

4) Based on Property 1), we have

$$\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) = \frac{1}{2}\mathcal{L}_{\text{MAE}} + \sum_{i=2}^{t} \frac{(1 - f_y(\boldsymbol{x}))^i}{i}.$$

Since the second term is always non-negative, $\mathcal{L}_{\text{MAE}} \leq 2\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y)$, and the equality holds when and only when $t = 1$. However, the difference becomes larger as t increases, which also suggests that $\mathcal{L}_{\text{CCE}}$ is not robust from another point of view, since $\mathcal{L}_{\text{MAE}}$ is the standard robust loss with strong theoretic guarantees [Ghosh *et al.*, 2017]. □

Properties 1) and 3) show that $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{CE}}$ can be considered as special cases of $\mathcal{L}_{t\text{-CE}}$. Properties 2) and 4) certify the robustness of $\mathcal{L}_{t\text{-CE}}$ to some degree, since $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{MSE}}$ have been shown to be robust in the presence of label noise [Ghosh *et al.*, 2017].

3.3 Theoretical Analysis

Here, we theoretically analyze the robustness of our proposed Taylor cross entropy loss.

By the definition of $\mathcal{L}_{t\text{-CE}}$ (i.e., Eq. (4)), we can easily derive an upper bound and a lower bound of $\mathcal{L}_{t\text{-CE}}$ as follows:

$$1 - f_y(\boldsymbol{x}) \leq \mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) \leq \sum_{i=1}^{t} \frac{1 - f_y(\boldsymbol{x})}{i}, \tag{5}$$

then we have the following lemma.

Lemma 1. *For any $\boldsymbol{x}$ and any positive integer $t < +\infty$, the sum of $\mathcal{L}_{t\text{-CE}}$ with respect to all the classes satisfies:*

$$k - 1 \leq \sum_{y=1}^{k} \mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) \leq (k-1)C_t, \tag{6}$$

where $C_t = \sum_{i=1}^{t} \frac{1}{i}$ is a constant that depends on t.

Proof. Based on Eq. (5), if we consider the sum of $\mathcal{L}_{t\text{-CE}}$ with respect to all the classes, the following equality holds:

$$\sum_{y=1}^{k} 1 - f_y(\boldsymbol{x}) \leq \sum_{y=1}^{k} \mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) \leq \sum_{y=1}^{k}\sum_{i=1}^{t} \frac{1 - f_y(\boldsymbol{x})}{i},$$

Hence

$$k - 1 \leq \sum_{y=1}^{k} \mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}), y) \leq (k-1)C_t,$$

which concludes the proof. □

We can find that $\mathcal{L}_{t\text{-CE}}$ is not always a symmetric loss, which means, $\sum_{y=1}^{k}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),y)$ is not always a constant. However, as shown in Lemma 1, $\sum_{y=1}^{k}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),y)$ is upper-bounded, and the bound gets tighter when t decreases. Specially, when $t=1$, $\sum_{y=1}^{k}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),y)=k-1$. In this case, $\mathcal{L}_{t\text{-CE}}$ becomes a symmetric loss, which can be considered equivalent to MAE.

Based on Lemma 1, we further analyze the robustness of $\mathcal{L}_{t\text{-CE}}$. We assume that the noisy example $(\boldsymbol{x},\tilde{y})$ is drawn from $p_\eta(\boldsymbol{x},\tilde{y})$, and the ordinary example $(\boldsymbol{x},y)$ is drawn from $p(\boldsymbol{x},y)$. Note that this paper follows the most common setting where label noise is *instance-independent*. Then we have $\tilde{y}=i$ $(y=i)$ with probability $\eta_i=(1-\eta)$ and $\tilde{y}=j$ with probability η_{ij} for all $j\neq i$ and $\sum_{j\neq i}\eta_{ij}=\eta$. If $\eta_{ij}=\frac{\eta}{k-1}$ for all $j\neq i$, then the noise is said to be *uniform* or *symmetric*, otherwise, the noise is said to be *class-conditional* or *asymmetric*. Given any classifier f and loss function $\mathcal{L}$, we define the risk of f under clean labels as $\mathcal{R}_{\mathcal{L}}(f)=\mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}),y)]$ and the risk under label noise rate η as $\mathcal{R}^{\eta}_{\mathcal{L}}(f)=\mathbb{E}_{p_\eta(\boldsymbol{x},\tilde{y})}[\mathcal{L}(f(\boldsymbol{x}),\tilde{y})]$. Let $\tilde{f}$ and f^* be the global minimizers of $\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)$ and $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)$ respectively.

Theorem 2. *Under uniform label noise with $\eta\leq 1-\frac{1}{k}$,*

$$0\leq\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})-\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f^*)\leq\frac{\eta(k-1)(C_t-1)}{(1-\eta)k-1},\quad(7)$$

where $C_t=\sum_{i=1}^{t}\frac{1}{i}$ is a constant that depends on t.

Proof. Under uniform label noise, we have

$$\begin{aligned}
&\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)\\
=&\mathbb{E}_{p_\eta(\boldsymbol{x},\tilde{y})}[\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),\tilde{y})]\\
=&\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{p(y|\boldsymbol{x})}\mathbb{E}_{p(\tilde{y}|y)}[\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),\tilde{y})]\\
=&\mathbb{E}_{p(\boldsymbol{x},y)}\Big[(1-\eta)\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),y)\\
&\qquad+\frac{\eta}{k-1}\sum\nolimits_{j\neq y}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big]\\
=&(1-\eta)\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)+\frac{\eta}{k-1}\Big(\sum\nolimits_{j=1}^{k}\mathcal{L}_{t\text{-CE}}\big(f(\boldsymbol{x}),j\big)\\
&\qquad-\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)\Big)\\
=&(1-\frac{\eta k}{k-1})\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)+\frac{\eta}{k-1}\sum\nolimits_{j=1}^{k}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j).
\end{aligned}$$

From Lemma 1 (by Eq. (6)), for all f, we have:

$$\beta\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)+\eta\leq\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)\leq\beta\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)+\eta C_t$$

where $\beta=(1-\frac{\eta k}{k-1})$. On the other hand, we have:

$$\frac{1}{\beta}(\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)-\eta C_t)\leq\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)\leq\frac{1}{\beta}(\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)-\eta)$$

Thus, for $\tilde{f}$,

$$\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})-\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f^*)\leq\frac{1}{\beta}\big(\alpha+\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})-\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f^*)\big),$$

where $\alpha=\eta(C_t-1)$. Since $\eta\leq 1-\frac{1}{k}$, f^* is the global minimizer of $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)$ and $\tilde{f}$ is the global minimizer of $\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)$, we have

$$0\leq\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})-\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f^*)\leq\frac{\eta(k-1)(C_t-1)}{(1-\eta)k-1},$$

which concludes the proof. □

Theorem 3. *Under class-conditional label noise with $\eta_{ij}<1-\eta_i,\forall j\neq i,\forall i,j\in[k]$, where $\eta_{ij}=p(\tilde{y}=j|y=i),\forall j\neq i$ and $(1-\eta_i)=p(\tilde{y}=i|y=i))$, if $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f^*)=0$, then*

$$0\leq\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f^*)-\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})\leq A,\quad(8)$$

where $A=(k-1)(C_t-1)\mathbb{E}_{p(\boldsymbol{x},y)}(1-\eta_i)>0$, $C_t=\sum_{n=1}^{t}\frac{1}{n}$ and f^* is the global minimizer of $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)$ and $\tilde{f}$ is the global minimizer of $\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)$.

Proof. For class-conditional label noise, we have

$$\begin{aligned}
&\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)\qquad(9)\\
=&\mathbb{E}_{p_\eta(\boldsymbol{x},\tilde{y})}[\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),\tilde{y})]\\
=&\mathbb{E}_{p(\boldsymbol{x},y)}[(1-\eta_i)\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),y)]\\
&\qquad+\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq y}\eta_{ij}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big]\\
\leq&\mathbb{E}_{p(\boldsymbol{x},y)}\Big[(1-\eta_i)\Big((k-1)C_t-\sum_{j\neq y}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big)\Big]\\
&\qquad+\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq y}\eta_{ij}\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big]\\
=&\frac{C_tA}{C_t-1}-\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq y}\lambda_j\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big],
\end{aligned}$$

where $A=(k-1)(C_t-1)\mathbb{E}_{p(\boldsymbol{x},y)}(1-\eta_i)$ and $\lambda_j=(1-\eta_i-\eta_{ij})$. On the other, we can obtain

$$\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f)\geq\frac{A}{C_t-1}-\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq i}\lambda_j\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\Big].$$

Then we have

$$\begin{aligned}
&\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f^*)-\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})\leq A\\
&\quad+\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq i}\lambda_j(\mathcal{L}_{t\text{-CE}}(\tilde{f}_j(\boldsymbol{x}),j)-(\mathcal{L}_{t\text{-CE}}(f^*(\boldsymbol{x}),j)))\Big],
\end{aligned}$$

From our assumption that $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f^*)=0$, we have $\mathcal{L}_{t\text{-CE}}(f^*(\boldsymbol{x}),y)=0$. This is only satisfied iff $f^*_j(\boldsymbol{x})=1$ when $j=y$ and $f^*_j(\boldsymbol{x})=0$ when $j\neq y$. According to the definition of $\mathcal{L}_{t\text{-CE}}$, $\mathcal{L}_{t\text{-CE}}(f^*(\boldsymbol{x}),j)=C_t,\forall j\neq y$ and $\mathcal{L}_{t\text{-CE}}(f(\boldsymbol{x}),j)\leq C_t,\forall j\in[k]$. Since f^* is the global minimizer of $\mathcal{R}_{\mathcal{L}_{t\text{-CE}}}(f)$ and $\lambda=(1-\eta_i-\eta_{ij})>0$, we can obtain

$$\mathbb{E}_{p(\boldsymbol{x},y)}\Big[\sum_{j\neq y}\lambda_j(\mathcal{L}_{t\text{-CE}}(\tilde{f}(\boldsymbol{x}),j)-\mathcal{L}_{t\text{-CE}}(f^*(\boldsymbol{x}),j))\Big]\leq 0.$$

Therefore, we have

$$0\leq\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(f^*)-\mathcal{R}^{\eta}_{\mathcal{L}_{t\text{-CE}}}(\tilde{f})\leq A,$$

which concludes the proof. □

Datasets	Methods	symmetric noise				asymmetric noise			
		0.2	0.4	0.6	0.8	0.1	0.2	0.3	0.4
MNIST	CCE	98.38±0.16•	97.38±0.22•	93.07±0.17•	90.48±0.76•	98.35±0.09•	97.23±0.13•	94.32±0.16•	79.67±0.17•
	MAE	98.80±0.09•	98.32±0.08•	97.15±0.07•	90.31±0.49•	98.72±0.09•	88.77±0.12•	88.13±0.18•	78.27±0.18•
	MSE	98.86±0.06•	98.41±0.19	97.28±0.19	90.14±0.19•	98.44±0.11•	96.51±0.14•	92.62±0.19•	86.22±0.19•
	GCE	98.78±0.08•	97.85±0.12•	95.02±0.09•	90.77±0.14•	98.97±0.08	98.61±0.11•	96.95±0.16	89.42±0.20○
	SCE	98.93±0.21	98.44±0.21	96.99±0.21•	90.01±0.21•	98.92±0.09•	98.63±0.15•	**97.74±0.21**○	85.97±0.19•
	PHuber-CE$_{\tau=10}$	98.46±0.08•	97.60±0.11•	95.37±0.07•	90.08±0.15•	98.65±0.07•	96.97±0.11•	93.98±0.09•	87.37±0.19•
	Taylor-CE$_{t=2}$	**98.99±0.14**	**98.46±0.12**	**97.53±0.23**	**91.14±0.18**	**99.01±0.06**	**98.84±0.09**	96.69±0.13	87.68±0.23
Kuzushiji	CCE	82.92±0.15•	79.75±0.39•	73.37±0.26•	36.14±0.84•	91.48±0.32•	89.84±0.43•	85.75±1.06•	79.41±0.89•
	MAE	92.43±0.22•	90.26±0.39	82.61±0.21○	60.46±0.54•	94.11±0.21•	93.18±0.29	89.01±0.25○	82.43±0.67•
	MSE	90.59±0.18•	86.14±0.24•	75.71±0.23•	**69.29±0.38**○	91.21±0.20•	89.29±0.56•	83.96±0.74•	76.72±1.12•
	GCE	92.41±0.19•	**91.48±0.17**○	**87.46±0.28**○	66.70±0.34•	94.18±0.16•	91.58±0.18•	**89.44±0.51**○	79.15±0.44•
	SCE	93.03±0.14•	91.22±0.14○	76.74±0.14•	66.42±0.22•	94.36±0.21•	92.57±0.22•	88.82±0.28	75.15±0.35•
	PHuber-CE$_{\tau=10}$	92.25±0.27•	88.07±0.23•	76.56±0.31•	68.45±0.25•	93.61±0.26•	89.09±0.38•	84.65±0.84•	77.94±1.36•
	Taylor-CE$_{t=2}$	**93.34±0.16**	89.99±0.41	81.88±0.36	67.24±0.33	**94.64±0.19**	**93.34±0.23**	88.53±0.36	**82.89±0.36**
Fashion	CCE	89.15±0.14•	88.04±0.15•	79.15±0.21•	65.70±0.33•	89.75±0.11•	82.91±0.21•	77.77±0.32•	72.45±0.59•
	MAE	89.68±0.12•	88.49±0.23•	85.37±0.37•	70.02±0.42•	90.01±0.09•	89.88±0.14•	89.09±0.23○	85.81±0.49•
	MSE	89.62±0.21•	88.25±0.36•	86.79±0.35	74.11±0.44•	89.89±0.13•	89.38±0.12•	89.11±0.21○	86.53±0.52•
	GCE	89.81±0.08•	87.61±0.11•	86.75±0.29•	78.39±0.64•	89.76±0.11•	88.34±0.09•	86.21±0.19•	84.82±0.35•
	SCE	89.31±0.12•	87.61±0.28•	86.06±0.32•	76.44±0.37•	**90.37±0.06**○	90.26±0.10	**89.69±0.16**○	85.82±0.34•
	PHuber-CE$_{\tau=2}$	89.56±0.21•	87.96±0.20•	85.27±0.26•	77.75±0.28•	90.17±0.12	89.68±0.11•	88.55±0.24○	78.97±0.41•
	Taylor-CE$_{t=4}$	**89.96±0.11**	**88.97±0.28**	**87.07±0.31**	**78.95±0.47**	90.25±0.09	**90.31±0.13**	88.31±0.15	**87.38±0.25**
CIFAR-10	CCE	74.89±0.32•	57.27±0.44•	36.65±0.35•	17.21±1.12•	85.57±0.33•	82.35±0.35•	78.14±0.55•	72.02±0.51•
	MAE	85.53±1.02•	79.28±0.88•	65.70±0.76•	31.07±4.13•	80.38±0.67•	79.06±0.58•	73.42±0.89•	59.78±0.87•
	MSE	75.63±2.36•	55.18±3.52•	35.41±4.12•	16.65±3.22•	86.19±0.22•	81.87±0.34•	76.85±0.32•	72.38±0.39•
	GCE	**86.04±0.11**	75.72±0.09•	48.34±0.12•	18.92±0.21•	87.77±0.07○	83.43±0.07•	77.41±0.10•	71.31±0.11•
	SCE	84.12±0.08•	66.22±0.09•	43.84±0.08•	16.19±0.09•	87.02±0.01•	83.83±0.03•	77.67±0.04•	72.51±0.14
	PHuber-CE$_{\tau=2}$	85.81±0.21	80.25±0.22•	**67.71±0.19**○	32.97±0.32•	**87.91±0.13**○	84.87±0.26	79.01±0.38•	71.87±0.36•
	Taylor-CE$_{t=2}$	85.96±0.09	**80.51±0.11**	66.36±0.32	**33.48±0.44**	87.34±0.12	**85.02±0.11**	**79.37±0.12**	**72.65±0.11**
CIFAR-100	CCE	47.00±0.13•	34.34±0.23•	19.37±0.33•	7.34±0.21•	56.71±0.34•	50.02±0.62•	43.29±0.58•	35.01±0.67•
	MAE	33.33±0.82•	26.56±0.79•	12.26±0.83•	2.01±0.01•	33.74±0.37•	33.01±0.46•	30.25±0.44•	22.66±0.53•
	MSE	47.66±0.63•	32.94±0.58•	18.41±0.73•	7.59±0.79•	56.02±0.47•	48.63±0.45•	40.69±0.52•	34.15±0.51•
	GCE	58.99±0.13	50.37±0.14•	**39.41±0.19**○	15.26±0.16•	60.31±0.09•	**56.49±0.12**○	45.78±0.11	34.99±0.14•
	SCE	47.32±0.09•	33.87±0.19•	18.79±0.26•	7.28±0.38•	56.84±0.12•	49.82±0.16•	43.54±0.22•	**35.98±0.23**○
	PHuber-CE$_{\tau=10}$	58.11±0.11•	50.89±0.13	35.85±0.29•	13.83±0.25•	60.07±0.09•	53.30±0.10•	44.39±0.14•	35.36±0.13
	Taylor-CE$_{t=6}$	**59.11±0.11**	**50.99±0.09**	38.31±0.12	**15.96±0.31**	**60.96±0.21**	55.45±0.12	**45.81±0.19**	35.45±0.25

Table 1: Average test accuracy (%) and standard deviation (over 5 trials) on benchmark datasets with symmetric label noise and asymmetric label noise. The best results are highlighted in bold. In addition, •/○ indicates whether the performance of our approach is statistically superior/inferior to the comparing approach on each dataset (paired t-test at 0.05 significance level).

Dataset	# Train	# Test	# Feature	# Class	Model
MNIST	60000	10000	784	10	LeNet-5
Fashion	60000	10000	784	10	LeNet-5
Kuzushiji	60000	10000	784	10	LeNet-5
CIFAR-10	50000	10000	3072	10	ResNet-34
CIFAR-100	50000	10000	3072	100	ResNet-34

Table 2: Summary of benchmark datasets and models.

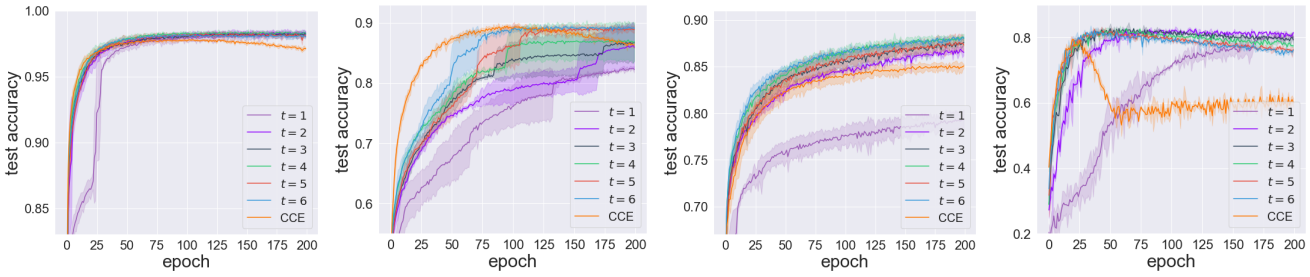
Theorem 2 and Theorem 3 show that using TCE, the difference of the risks caused by the derived hypotheses $\tilde{f}$ and $f^\star$ under noisy labels and clean labels are always bounded. Besides, the two bounds are related to the parameter t. Smaller t results in smaller C_t, hence both bounds in Theorem 2 and Theorem 3 would be tighter if t gets smaller. The above analysis clearly demonstrates the noise-tolerant ability of TCE. Specially when $t = 1$, TCE has the same theoretical guarantees as MAE.

4 Experiments

4.1 Experimental Settings

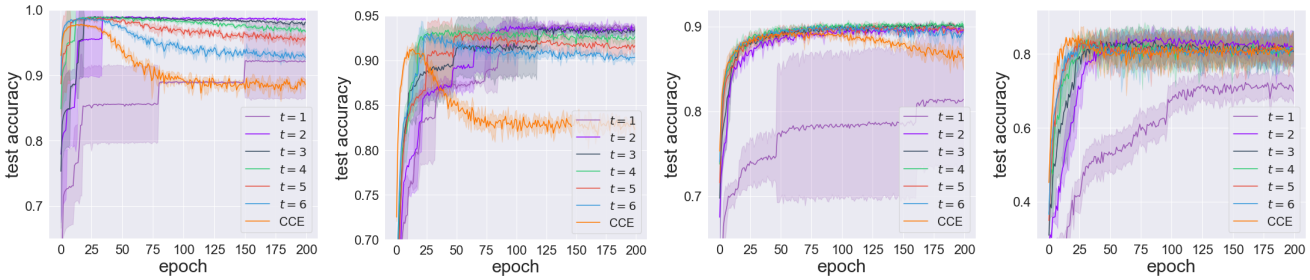
Baselines. We compare our proposed TCE with CCE and multiple state-of-the-art robust loss functions. All the used loss functions in this paper are listed as follows. 1) CCE: The standard categoriccal cross entropy loss. It is neither symmetric nor bounded. 2) MAE: A symmetric loss function that has been demonstrated [Ghosh *et al.*, 2017] to be robust to label noise. 3) MSE: The mean squared error. It is not symmetric but bounded. 4) GCE [Zhang and Sabuncu, 2018]: A bounded loss function that uses a hyper-parameter q to balance between MAE and CCE. The hyper-parameter q is set to 0.7, as this is the recommended setting by the corresponding paper [Zhang and Sabuncu, 2018]. 5) PHuber-CE: A Partially Huberised loss of CCE that corrects CCE on hard examples by gradient clipping. The hyper-parameter τ is selected from $\{2, 10\}$. 6) SCE [Wang *et al.*, 2019]: The approach boosts CE symmetrically with a noise robust counterpart Reverse Cross Entropy (RCE). The regularization parameters α and β are configured with the suggested values on different datasets. 7) TCE: The loss function proposed in our paper, is based on the Taylor Series of CCE. The hyper-parameter t is selected from $\{2, \ldots, 6\}$. For all the methods, learning rate is selected from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

Datasets. Our experiments are conducted on MNIST [LeCun *et al.*, 1998], Fashion-MNIST (Fashion in short) [Xiao *et al.*, 2017], Kuzushiji-MNIST (Kuzushiji in short) [Clanuwat *et al.*, 2018], CIFAR-10 [Krizhevsky *et al.*, 2009] and CIFAR-100 [Krizhevsky *et al.*, 2009] with two types of label noise: symmetric noise and asymmetric noise. We use appropriate networks to train different datasets, and all networks are trained using the Adam optimizer [Kingma and Ba, 2014] with the number of epochs set to 200 and the batch size set to 256. We test different noise rates, with $\eta \in \{20\%, 40\%, 60\%, 80\%\}$ for symmetric label noise and $\eta \in \{10\%, 20\%, 30\%, 40\%\}$ for asymmetric label noise.



(a) *MNIST* (b) *Kuzushiji* (c) *Fashion* (d) *CIFAR-10*

Figure 2: Test accuracy against number of epochs for training with 40% symmetric label noise.



(a) *MNIST* (b) *Kuzushiji* (c) *Fashion* (d) *CIFAR-10*

Figure 3: Test accuracy against number of epochs for training with 20% asymmetric label noise.

Training Details. For MNIST/Fashion/Kuzushiji, LeNet-5 [LeCun *et al.*, 1998] is used. Since they are MNIST-type datasets, noisy labels are generated in the same way for all of them. Symmetric label noise is generated by mapping a true label to a random label by a given probability $\frac{\eta}{c-1}$. For asymmetric label noise, flipping labels only occur within a specific set of classes [Patrini *et al.*, 2017]. For MNIST, flipping $2 \rightarrow 7$, $3 \rightarrow 8$, $5 \rightarrow 6$ and $7 \rightarrow 2$. On the three datasets, networks are trained with weight decay of 10^{-4}. For CIFAR-10/CIFAR-100, ResNet-34 [He *et al.*, 2016] is used. We perform 32×32 random crops after padding with 4 pixels on each side on the two datasets. Symmetric noise is generated in the same way as that for MNIST-type datasets. Following [Patrini *et al.*, 2017], asymmetric label noise is generated by mapping TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, and CAT↔ DOG with probability η for CIFAR-10. For CIFAR-100, the 100 classes are grouped into 20 super-classes with each has 5 sub-classes, and we flip between two randomly selected sub-classes within each super-class. On the two datasets, networks are trained with weight decay of 0.

4.2 Experimental Results

Table 1 reports the detailed experimental results of each loss function on the benchmark datasets. In Table 1, $\bullet/\circ$ indicates whether the performance of our proposed approach is statistically superior/inferior to other comparing approaches on each dataset (paired t-test at 0.05 significance level). Out of the total 240 cases (with 6 comparing approaches, 5 datasets, and 8 label noise settings), our proposed approach is statistically superior to other comparing approaches in 83.33% cases and inferior to the comparing approaches in only 8.33% cases. Further more, in all cases, $\mathcal{L}_{t\text{-CE}}$ is superior to CE, MAE and MSE. As we showed above, $\mathcal{L}_{t\text{-CE}}$ is derived from CE, with MAE and MSE as average components. Thus, we may infer that $\mathcal{L}_{t\text{-CE}}$ maintains their advantages and surpasses them.

4.3 Parametric Analysis

We also conduct experiments on MNIST, Fashion, Kuzushiji, and CIFAR-10 with 40% symmetric label noise and 20% asymmetric label noise, for parametric analysis. We also include CCE for comparison, and vary the hyper-parameter t in $\{1, \ldots, 6\}$. As shown in Figure 2 and Figure 3, CCE always leads to over-fitting and MAE cannot achieve a high classification accuracy in some cases. In contrast, $\mathcal{L}_{t\text{-CE}}$ avoids over-fitting and achieves good performance by varying t. Obviously, $\mathcal{L}_{t\text{-CE}}$ inherits the advantages of CCE while $\mathcal{L}_{t\text{-CE}}$ is more robust to label noise than CCE.

5 Conclusion

In this paper, we propose a general framework dubbed Taylor cross entropy loss to train deep models in the presence of label noise. Our framework can not only enable to weight the extent of fitting the training labels by controlling the order of Taylor Series for Categorical Cross Entropy (CCE) loss, but also reveals the intrinsic relationships between CCE and other loss functions, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). Moreover, we present a detailed theoretical analysis to certify the robustness of this framework. Experiments on benchmark datasets also show that the proposed framework is superior to the state-of-the-art counterparts. In future work, we will explore if there exist robust loss functions that do not include any hyper-parameters.

Acknowledgements

This research is supported by NSOE-TSS2019-01, AISG-RP-2019-0013, and NTU.

References

[Berthon *et al.*, 2020] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. *arXiv preprint arXiv:2001.03772*, 2020.

[Clanuwat *et al.*, 2018] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[Ghosh *et al.*, 2015] Aritra Ghosh, Naresh Manwani, and P-S Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

[Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P-S Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

[Goldberger and Ben-Reuven, 2017] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[Han *et al.*, 2018a] Bo Han, Gang Niu, Jiangchao Yao, Xingrui Yu, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Pumpout: A meta approach to robust deep learning with noisy labels. *arXiv preprint arXiv:1809.11008*, 2018.

[Han *et al.*, 2018b] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *NuerIPS*, pages 5836–5846, 2018.

[Han *et al.*, 2018c] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NuerIPS*, pages 8527–8537, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NuerIPS*, pages 10456–10465, 2018.

[Jiang *et al.*, 2017] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

[Khetan *et al.*, 2017] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Menon *et al.*, 2015] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pages 125–134, 2015.

[Menon *et al.*, 2020] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *ICLR*, 2020.

[Natarajan *et al.*, 2013] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NuerIPS*, pages 1196–1204, 2013.

[Patrini *et al.*, 2017] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017.

[Sukhbaatar and Fergus, 2014] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

[Tanaka *et al.*, 2018] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019.

[Wei *et al.*, 2020] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. *arXiv preprint arXiv:2003.02752v3*, 2020.

[Xia *et al.*, 2019] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NuerIPS*, pages 6835–6846, 2019.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Yang *et al.*, 2019] Hansi Yang, Quanming Yao, Bo Han, and Gang Niu. Searching to exploit memorization effect in learning from corrupted labels. *arXiv preprint arXiv:1911.02377*, 2019.

[Yi and Wu, 2019] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, pages 7017–7025, 2019.

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173, 2019.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NuerIPS*, pages 8778–8788, 2018.