Mitigating Memorization of Noisy Labels by Clipping the Model Prediction

Hongxin Wei [1] **Huiping Zhuang** [2] **Renchunzi Xie** [3] **Lei Feng** [3] **Gang Niu** [4] **Bo An** [3] **Yixuan Li** [5]

Abstract

In the presence of noisy labels, designing robust loss functions is critical for securing the generalization performance of deep neural networks. Cross Entropy (CE) loss has been shown to be *not* robust to noisy labels due to its *unboundedness*. To alleviate this issue, existing works typically design specialized robust losses with the symmetric condition, which usually lead to the underfitting issue. In this paper, our key idea is to induce a loss bound at the logit level, thus universally enhancing the noise robustness of existing losses. Specifically, we propose *logit clipping* (**LogitClip**), which clamps the norm of the logit vector to ensure that it is upper bounded by a constant. In this manner, CE loss equipped with our LogitClip method is effectively bounded, mitigating the overfitting to examples with noisy labels. Moreover, we present theoretical analyses to certify the noise-tolerant ability of LogitClip. Extensive experiments show that LogitClip not only significantly improves the noise robustness of CE loss, but also broadly enhances the generalization performance of popular robust losses.

1. Introduction

The success of supervised learning relies heavily on a massive amount of data, where each training instance is labeled by a human annotator. However, labels solicited from humans can often be subject to label noise. The issue of noisy labels has been commonly observed in many real-world scenarios, such as crowdsourcing (Yan et al., 2014) and online queries (Blum et al., 2003). As a result, models trained on such data containing noisy labels suffer from poor generalization performance (Arpit et al., 2017; Zhang et al., 2016). This gives rise to the importance of *noise-robust learning*, where the goal is to train a robust classifier in the presence of noisy and erroneous labels. The learning task thus provides stronger flexibility and practicality than the standard supervised learning, where each training example is provided with clean ground truth.

Despite the most popular loss in classification tasks, CE loss has shown to be non-robust in the presence of label noise (Ghosh et al., 2017; Zhang & Sabuncu, 2018), due to its *unboundedness*. Concerningly, the loss could approach infinity when the observed noisy label mismatches the model's prediction. Consequently, the model would attempt to counteract the large loss by overfitting the label noise, leading to poor generalization performance. To bound the loss value, previous methods typically design robust losses with principal constraint, *e.g.*, symmetric condition (Ghosh et al., 2017; Ma et al., 2020). Despite their theoretical robustness, these specialized losses can cause difficulty in optimization, leading to underfitting issues on complex datasets (Zhang & Sabuncu, 2018; Zhou et al., 2021). This motivates our method, which mitigates the undesirable influence of unbounded loss without modifying the loss function.

In this paper, our key idea is to induce the loss bound at the logit level, which universally enhances the noise robustness of existing losses. Specifically, we propose *logit clipping* (**LogitClip**), which clamps the norm of the logit vector to ensure that it is upper bounded by a constant. Our method can be interpreted as a constrained optimization with an inequality constraint placed on the logit vector. Theoretically, we show that CE loss equipped with our LogitClip method is always bounded. Consequently, the difference between the risks caused by the derived hypotheses under noisy and clean labels is always bounded. More importantly, the two bounds (Theorem 2.2 and Theorem 2.3) depend on the logit norm threshold, where a smaller threshold induces tighter bounds. In this way, our theoretical analyses demonstrate the noise-tolerant ability of LogitClip.

To verify the effectiveness of our method, we conduct thorough empirical evaluations on both simulated and

[1]Southern University of Science and Technology. Work done while working at UW-Madison as a visiting scholar. [2]South China University of Technology [3]Nanyang Technological University [4]RIKEN AIP [5]University of Wisconsin-Madison. Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

real-world noisy datasets, including CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and WebVision (Li et al., 2017) datasets. The results demonstrate that logit clipping can significantly improve the noise-robustness of CE loss, under symmetric, asymmetric, instance-dependent, and real-world label noise. For example, on CIFAR-10 with instance-dependent label noise, LogitClip improves the test accuracy of CE loss from 68.36% to 86.60% – a **18.24**% of direct improvement. More importantly, we show that LogitClip can boost the performance of a wide range of popular robust loss functions, including MAE (Ghosh et al., 2017), PHuber-CE (Menon et al., 2020), SCE (Wang et al., 2019), GCE (Zhang & Sabuncu, 2018), Taylor-CE (Feng et al., 2020), NCE (Ma et al., 2020), AEL, AUL (Zhou et al., 2021), Cores (Cheng et al., 2021), and Active Passive losses (Ma et al., 2020).

We summarize our contributions as follows:

1. We propose LogitClip – a simple and effective method to enhance the noise robustness of existing losses. The key idea is to clamp the norm of the logit vector to bound the loss value, as shown in Proposition 2.1 and Proposition 2.4.

2. We provide theoretical analyses in Theorem 2.2 and Theorem 2.3 to certify the noise-tolerant ability of our LogitClip, where a smaller threshold induces tighter bounds.

3. We conduct extensive evaluation to show that LogitClip can improve the robustness of CE and popular robust losses across various types of label noise. We empirically show that our method is model-agnostic and also applicable in large-scale real-world scenarios.

4. We perform ablation studies that lead to improved understandings of our method. In particular, we contrast with alternative methods (*e.g.*, RELU6 (Howard et al., 2017), Clipping-by-value) and demonstrate the advantages of our method with Clipping-by-norm.

2. Motivation and Method

2.1. Preliminaries: The Unboundedness of CE loss

In this work, we consider the multi-class classification task with K different classes. Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space and $\mathcal{Y} = \{1, \ldots, K\}$ be the label space, we consider a training dataset with N samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the i-th instance sampled *i.i.d.* from an underlying data-generating distribution $\mathcal{P}$ and $y_i \in \mathcal{Y}$ is the observed (and potentially noisy) label. A classifier is a function that maps from the input space to the label space $f : \mathcal{X} \to \mathbb{R}^K$ with trainable parameter $\boldsymbol{\theta} \in \mathbb{R}^p$.

Here, we consider *composite* losses, which are comprised of a base loss function ϕ and an invertible link function $\sigma : \mathbb{R} \to [0, 1]$. For example, the most commonly used composite loss in multi-class classification is Softmax Cross Entropy (CE) loss:

$$\begin{aligned}\mathcal{L}_{\text{CE}}\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right) &= -\sum_{j=1}^{K} \boldsymbol{y}_j \log(\sigma\left(\boldsymbol{z}_j\right)) \\ &= -\sum_{j=1}^{K} \boldsymbol{y}_j \log\left(\frac{e^{\boldsymbol{z}_j}}{\sum_{k=1}^{K} e^{\boldsymbol{z}_k}}\right),\end{aligned} \tag{1}$$

where $\boldsymbol{z}_j = f_j\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ corresponds to the j-th element of model output for the sample $\boldsymbol{x}$, and $\boldsymbol{y}_j$ is the j-th element of one-hot encoded label vector $\boldsymbol{y}$. Here σ denotes the softmax function, which is also the invertible link function. As an *unbounded* loss function, CE is shown to be non-robust in the presence of label noise (Ghosh et al., 2017; Zhang & Sabuncu, 2018), since the observed labels might be incorrect. In particular, the gradients of CE can be shown as:

$$\frac{\partial \mathcal{L}_{\text{CE}}(f(\boldsymbol{x},\boldsymbol{\theta}), y)}{\partial \boldsymbol{\theta}} = -\frac{1}{\sigma_y(\boldsymbol{z}))} \nabla_{\boldsymbol{\theta}} \sigma_y(\boldsymbol{z})),$$

From the equation, we find that CE pays more attention to those examples with lower confidences, i.e., hard examples (or noisy examples). As $\sigma(z) \to 0$, the unbounded loss would approach infinity, leading to severe overfitting issues on noisy labels.

2.2. Our Proposed Method

In this paper, we propose a general strategy that can make the loss function noise-robust, avoiding the inherent drawback of overfitting the label noise. Our key idea is to *bound* the logit value in the link function. Our method is motivated by the following reformulation of the softmax CE loss:

$$\begin{aligned}\mathcal{L}_{\text{CE}}(\boldsymbol{z}, y) &= \log(1 + \sum_{j \neq y} e^{\boldsymbol{z}_j - \boldsymbol{z}_y}) \\ &\leq \log(1 + (K-1) \cdot e^{\boldsymbol{z}_{\max} - \boldsymbol{z}_{\min}}),\end{aligned} \tag{2}$$

where $\boldsymbol{z}^{\max}$ and $\boldsymbol{z}^{\min}$ denote the maximum and minimum value in the logit vector $\boldsymbol{z} = f(\boldsymbol{x};\boldsymbol{\theta})$. The above formulation suggests that, if $\boldsymbol{z}^{\max} - \boldsymbol{z}^{\min}$ is upper bounded, $\mathcal{L}_{\text{CE}}$ could not reach infinity, thereby preventing the model from overfitting to examples with noisy labels.

To enforce the upper bound, we propose to bound the logit values by norm, which preserves the direction of the original logit vector. The training objective can be formalized as a constrained optimization with inequality constraint:

$$\begin{aligned}&\text{minimize} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}}\left[\mathcal{L}_{\text{CE}}\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\right] \\ &\text{subject to} \quad \left\|f(\boldsymbol{x};\boldsymbol{\theta})\right\|_p \leq \alpha,\end{aligned}$$

where $\|\cdot\|_p$ denotes the p-norm (also called ℓ_p-norm, $p \geq 1$) of the logit vector. However, performing constrained optimization in the context of modern neural networks is non-trivial, as explicitly shown in Appendix G.2. To circumvent the issue, we convert the objective into an alternative loss function that can be end-to-end trainable, strictly enforcing an upper bound of vector norm.

Logit Clipping. We propose *logit clipping* (dubbed LogitClip), which clamps the norm of the logit vector to ensure it is upper bounded by a constant. Formally, the new link function is defined as:

$$\bar{\sigma}_\tau(\boldsymbol{z}) \doteq \sigma(\text{clip}_\tau(\boldsymbol{z})),\ \text{clip}_\tau(\boldsymbol{z}) \doteq \begin{cases} \tau \cdot \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_p} & \text{if } \|\boldsymbol{z}\|_p \geq \tau \\ \boldsymbol{z} & \text{else} \end{cases}, \tag{3}$$

where τ is the upper bound of the norm. Our method ensures that: (1) the norm of the clamped logit vector $\text{clip}_\tau(\boldsymbol{z})$ is bounded by τ, and (2) the clamped logit vector preserves the same direction (and prediction) as the original logit vector. To increase flexibility, one can set the scale factor δ to a value that differs from the threshold τ. In this form, the clipping function can be represented as:

$$\text{clip}_{\tau,\delta}(\boldsymbol{z}) \doteq \begin{cases} \delta \cdot \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_p} & \text{if } \|\boldsymbol{z}\|_p \geq \tau \\ \boldsymbol{z} & \text{else} \end{cases}. \tag{4}$$

Note that LogitClip can be compatible with various loss functions, as we will later demonstrate in Section 3. In other words, we can employ LogitClip in the link function σ of the composite losses, where the base loss function ϕ can be CE or other existing robust loss functions. Taking cross-entropy loss as an example, the new training objective now becomes:

$$\text{minimize} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}}\left[\mathcal{L}_{\text{CE}}\left(\text{clip}_\tau(f(\boldsymbol{x};\boldsymbol{\theta})), y\right)\right].$$

In what follows, we provide a formal analysis on the bound of the new loss function. For convenience, we denote CE loss with LogitClip as $\mathcal{L}_{\text{CE}}^\tau$. Without loss of generality, we use the same value for the scale factor and the threshold for simplicity (as shown in Equation 3). We start from the case of max norm ($p = \infty$), i.e., $-\tau \leq \text{clip}_\tau(\boldsymbol{z}_j) \leq \tau$. Then we can derive an upper bound and a lower bound of $\mathcal{L}_{\text{CE}}^\tau$.

Proposition 2.1 (Upper and Lower Bounds of CE with LogitClip)**.** *For any input $\boldsymbol{x}$ and any positive number $\tau \in \mathbb{R}^+$, CE loss with LogitClip defined in Eq. (3) has a lower bound and an upper bound:*

$$\log(1+(K-1)\cdot e^{-2\tau}) \leq \mathcal{L}_{\text{CE}}^\tau(\boldsymbol{z}, y) \leq \log(1+(K-1)\cdot e^{2\tau}).$$

The proof of Proposition 2.1 is provided in Appendix A. Through Proposition 2.1, we show that cross-entropy loss equipped with LogitClip is bounded. The conclusion can be extended to other norms, since $\|\boldsymbol{z}\|_p \leq \|\boldsymbol{z}\|_q \leq \|\boldsymbol{z}\|_\infty$ for $p \geq q$. To provide a straightforward view, we show in Figure 1 how the hyperparameter τ and the class num K affect the upper and lower bounds of CE with LogitClip. When $\tau \to \infty$, we have $0 \leq \mathcal{L}_{\text{CE}}^\tau \leq \infty$, which is equivalent to the original loss $\mathcal{L}_{\text{CE}}$ in Equation 1. On the other hand, if $\tau \to 0$, the lower bound would be close to the upper bound, which may result in difficulties for loss optimization. We will analyze the effect of τ in detail in Section 3.

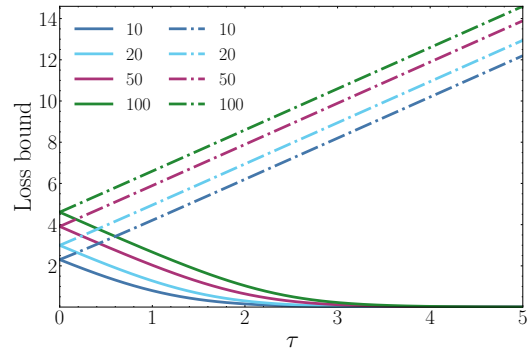


Figure 1: The effect of τ and K on the Loss bound. The dashed lines denote the upper bounds and the solid lines show the lower bounds. Four colors are used to present bounds with various values of K.

Based on proposition 2.1, we further analyze the noise robustness of $\mathcal{L}_{\text{CE}}^\tau$ with LogitClip. We denote the clean ground-truth label of $\boldsymbol{x}$ as $y^\star$. Here, we follow the most common setting where label noise is instance-independent (Ghosh et al., 2017; Feng et al., 2020; Ma et al., 2020). Under this assumption, label noise can be either *symmetric* (i.e., uniform) or *asymmetric* (i.e., class-conditional). Let $\eta \in [0, 1]$ be the overall noise rate and η_{jk} be the class-wise noise rate from ground-truth class j to class k, where $\eta_{jk} = p(y = k \mid y^\star = j)$. For symmetric noise, $\eta_{jk} = \frac{\eta}{K-1}$ for $j \neq k$ and $\eta_{jk} = 1 - \eta$ for $j = k$. For asymmetric noise, η_{jk} is conditioned on both the true class j and the mislabeled class k. Given any classifier f and loss function $\mathcal{L}$, the risk of f under clean labels is defined as: $\mathcal{R}_{\mathcal{L}}(f) = \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}[\mathcal{L}(f(\boldsymbol{x}), y^\star)]$ and the risk under label noise rate η is: $\mathcal{R}_{\mathcal{L}}^\eta(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\text{noisy}}^\eta}[\mathcal{L}(f(\boldsymbol{x}), y)]$. Let $\tilde{f}$ and $f^\star$ be the global minimizers of $\mathcal{R}_{\mathcal{L}_{\text{CE}}^\tau}^\eta(f)$ and $\mathcal{R}_{\mathcal{L}_{\text{CE}}^\tau}(f)$, respectively.

Theorem 2.2. *Under symmetric label noise with $\eta \leq 1 - \frac{1}{K}$,*

$$0 \leq \mathcal{R}_{\mathcal{L}_{\text{CE}}^\tau}(\tilde{f}) - \mathcal{R}_{\mathcal{L}_{\text{CE}}^\tau}(f^\star) \leq \frac{\eta K}{(1-\eta)K - 1} \cdot A_\tau^K,$$

where $A_\tau^K = \log\left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right)$ is a constant that depends on τ and number of classes K.

Theorem 2.3. *Under asymmetric label noise with* $\eta_{ij} < 1-\eta_i, \forall j \neq i, \forall i,j \in [k]$*, where* $\eta_{ij} = p(y=j \mid y^\star = i), \forall j \neq i$ *and* $(1-\eta_i) = p(y=i \mid y^\star = i))$*, then*

$$0 \leq \mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f^*) - \mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(\tilde{f}) \leq B^K_\tau,$$

where $B^K_\tau = K \log\left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right) \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{clean}}(1-\eta_i) > 0$.

The proofs of the above two Theorems are provided in Appendix B and Appendix C, respectively. Theorem 2.2 and Theorem 2.3 show that with LogitClip, the difference of the risks caused by the derived hypotheses $\tilde{f}$ and $f^\star$ under noisy and clean labels is always bounded. More specifically, the two bounds depend on the parameter τ. With a smaller τ, both the A^K_τ in Theorem 2.2 and the B^K_τ in Theorem 2.3 become smaller, indicating tighter bounds. The above analysis provably demonstrates the noise-tolerant ability of cross-entropy loss with LogitClip method. **We extend our analysis to an instance-dependent setting in Appendix E**. We proceed by a general analysis of composite losses with LogitClip.

Proposition 2.4. *Given any base loss* $\phi(x)$ *that satisfies the Lipschitz condition with constant* L *on the domain* $M^K_\tau \leq x \leq N^K_\tau$*, the resulting composite loss with* $\tilde{\sigma}_\tau$ *defined in Equation 3 is bounded:*

$$\left|\mathcal{L}^{\tau}_{\phi}(f(\boldsymbol{x};\boldsymbol{\theta}), y)\right| \leq L\left(N^K_\tau - M^K_\tau\right) + \left|\phi(M^K_\tau)\right|,$$

where $M^K_\tau = \frac{1}{1+(K-1)\cdot e^{2\tau}}$ *and* $N^K_\tau = \frac{1}{1+(K-1)\cdot e^{-2\tau}}$.

The proofs of Proposition 2.4 are provided in Appendix D. From Proposition 2.4, we show that composite losses equipped with LogitClip are bounded if their base losses are locally Lipschitz continuous (Sohrab, 2003) on the domain, which depends on τ and class number K. If $\tau \to \infty$, the domain turns to be $0 \leq x \leq 1$. In this case, CE and Focal loss (Lin et al., 2017) do not satisfy the Lipschitz condition and thus are unbounded. Otherwise, with a constant $\tau < \infty$, CE and Focal loss are locally Lipschitz continuous on the domain so that they can be bounded with our LogitClip. Similarly, this conclusion is also applicable to existing robust losses, which generally satisfy the Lipschitz condition. Based on the loss bound in Proposition 2.4, we can easily derive the bounds of $\mathcal{R}_{\mathcal{L}^{\tau}_{\phi}}(\tilde{f}) - \mathcal{R}_{\mathcal{L}^{\tau}_{\phi}}(f^\star)$ under symmetric and asymmetric label noise, following our proofs for Theorem 2.2 and Theorem 2.3. Overall, the above analysis provably shows that LogitClip can enable the resulting composite loss to be noise-tolerant, with the locally Lipschitz condition for the base loss. We will further verify our analysis experimentally in Section 3.

3. Experiments

In this section, we validate the effectiveness of our method on three benchmarks, including simulated and real-world datasets under various types of label noise. We show that LogitClip not only significantly improves the robustness of CE loss, but also broadly enhances the performance of popular robust losses. In addition, we perform a sensitivity analysis to validate the effect of τ.

3.1. Setups

Datasets. To verify the efficacy of LogitClip, we comprehensively consider four different types of label noise, including **(1)** symmetric noise, **(2)** asymmetric noise (Zhang & Sabuncu, 2018), **(3)** instance-dependent noise (Chen et al., 2020), and **(4)** real-world noise on CIFAR-10/100 (Krizhevsky et al., 2009) and WebVision (Li et al., 2017) datasets. For symmetric noise, each label can be flipped to any other class with the same probability. In our experiments, we uniformly flip the label to other classes with a probability of 20% and 50%, respectively. For asymmetric noise, the labels might be only flipped to similar classes (Patrini et al., 2017; Zhang & Sabuncu, 2018). In our CIFAR-10 experiments, we generate asymmetric noisy labels by mapping TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, and CAT $\leftrightarrow$ DOG with probability 40%. For CIFAR-100, we flip each class into the next circularly with a probability of 40%. For instance-dependent noise, we assume the mislabeling probability of each instance is dependent on the corresponding input features (Chen et al., 2020; Xia et al., 2020b). In the experiments, we use the instance-dependent noise from PDN (Xia et al., 2020b) with a noisy rate of 40%, where the noise is synthesized based on the DNN prediction error. For real-world noisy labels on the CIFAR datasets, we use the "Worst" label set of CIFAR-10N and the "Fine" label set of CIFAR-100N (Wei et al., 2022d), respectively.

Training details. We perform training with WRN-40-2 (Zagoruyko & Komodakis, 2016) on CIFAR-10 and CIFAR-100. In particular, we train the network for 200 epochs using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. We set the initial learning rate as 0.1, and reduce it by a factor of 10 after 80 and 140 epochs. For our LogitClip in all experiments, we set $\delta = 1/\tau$ (see Equation 4) and use Euclidean norm, i.e., $p = 2$. We use 5k noisy samples as the validation dataset to tune the hyperparameter $1/\tau$ in $\{0.1, 0.5, 1, 1.5, \ldots, 4.5, 5\}$, then train the model on the full training set and report the average test accuracy in the last 10 epochs. We repeat all experiments 5 times with different random seeds. More training details are described in Appendix F.

3.2. CIFAR-10 and CIFAR-100

On CIFAR-10 and CIFAR-100, we validate that LogitClip can enhance the noise robustness of existing loss functions. In particular, we consider the following loss func-

Table 1: Average test accuracy (%) with standard deviation on CIFAR-10 under various types of noisy labels (over 5 runs). The bold indicates the improved results by integrating LogitClip (LC).

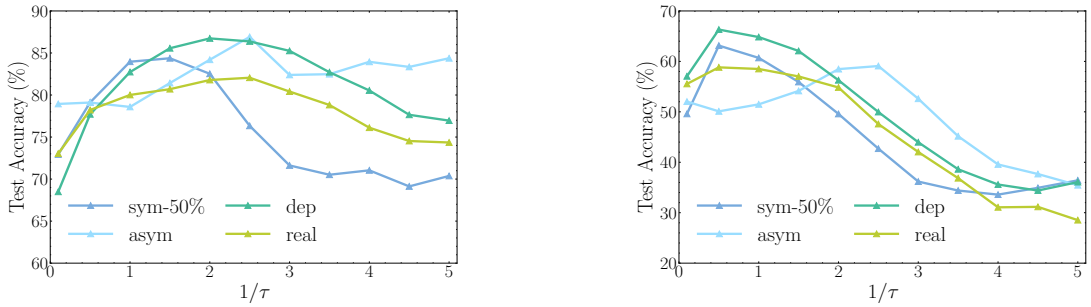
Method	Sym-20%	Sym-50%	Asymmetric	Dependent	Real
CE	86.73±0.72	70.88±0.46	78.34±0.54	68.26±0.21	72.85±0.32
+ LC (Ours)	**91.62±0.16**	**84.37±0.34**	**86.91±0.68**	**86.74±0.55**	**82.06±0.70**
Focal	87.17±0.68	70.61±0.59	79.61±0.40	69.40±0.55	72.29±0.41
+ LC (Ours)	**91.91±0.46**	**84.65±0.74**	**85.42±0.80**	**87.02±0.22**	**81.78±0.37**
MAE	88.92±0.36	75.73±1.15	56.74±0.71	53.92±1.07	53.26±0.67
+ LC (Ours)	**90.84±0.20**	**86.06±0.52**	**83.74±0.46**	**87.76±0.79**	**82.83±0.38**
PHuber-CE	90.92±0.93	74.07±0.41	81.26±0.65	75.07±0.26	76.61±0.58
+ LC (Ours)	**91.90±0.65**	**84.64±0.19**	**85.54±0.38**	**86.96±0.72**	**82.41±0.82**
SCE	91.48±0.78	85.38±0.47	78.65±0.60	87.05±0.58	81.65±0.35
+ LC (Ours)	91.56±0.14	**86.18±0.35**	**84.47±1.04**	**87.61±0.19**	**82.43±0.37**
GCE	90.58±0.66	85.51±0.45	79.35±0.58	87.64±0.62	81.38±0.70
+ LC (Ours)	**91.21±0.52**	85.90±0.15	**84.24±0.84**	87.69±0.21	**82.44±0.11**
Taylor-CE	90.44±0.40	85.71±0.26	80.92±1.37	87.20±0.98	82.32±1.12
+ LC (Ours)	**91.37±0.30**	**86.31±0.18**	**84.57±0.45**	**88.05±0.72**	**82.86±0.55**
NCE	90.87±0.94	68.45±0.43	83.68±0.85	73.53±0.93	79.96±0.25
+ LC (Ours)	**91.70±0.27**	**85.88±0.89**	**88.44±0.53**	**87.59±1.21**	**82.11±0.64**
AEL	88.59±1.03	77.48±0.88	60.90±1.27	84.55±0.71	69.40±0.38
+ LC (Ours)	**90.58±0.36**	**86.07±0.43**	**82.12±0.89**	**87.77±0.48**	**82.17±0.71**
AUL	76.73±0.33	75.27±0.93	59.80±0.75	73.66±0.45	63.96±1.15
+ LC (Ours)	**91.31±0.85**	**85.98±0.74**	**84.38±0.53**	**87.93±0.86**	**82.46±0.66**
Cores	91.56±0.23	85.32±0.36	85.30±0.63	86.65±0.83	82.28±0.28
+ LC (Ours)	91.72±0.15	**86.06±0.28**	**86.36±0.33**	**87.75±0.09**	**82.83±0.40**

tions: **(1)** Cross-Entropy (CE), which is the most commonly used classification loss. **(2)** Focal loss, which is originally proposed for dense object detection and also an unbounded classification loss function, $\mathcal{L}_{\text{Focal}}(\boldsymbol{p}, y) = -\sum_{j=1}^{K} \boldsymbol{y}_j (1-\boldsymbol{p}_j)^{\gamma} \log(\boldsymbol{p}_j)$. We set $\gamma = 0.5$ in our experiments. **(3)** Mean absolute error (MAE) (Ghosh et al., 2017), a symmetric loss function $\mathcal{L}_{\text{MAE}}(\boldsymbol{p}, y) = \left\|\boldsymbol{y}_j - \boldsymbol{p}_j\right\|_1$ that has been demonstrated to be robust to label noise. **(4)** PHuber-CE (Menon et al., 2020), a loss variant of gradient clipping for learning with noisy labels. **(5)** SCE (Wang et al., 2019), which boosts CE symmetrically with a noise-robust counterpart Reverse Cross Entropy (RCE). **(6)** GCE (Zhang & Sabuncu, 2018), a bounded loss function that uses a hyperparameter q to balance between MAE and CE. Following the recommended setting in the corresponding paper, we set the hyperparameter q as 0.7. **(7)** Taylor-CE (Feng et al., 2020), which controls the order of the Taylor Series to balance between MAE and CE. **(8)** NCE (Ma et al., 2020), which employs loss normalization to boost the robustness of CE loss. **(9)** AEL, AUL, and AGCE (Zhou et al., 2021), which are asymmetric loss functions. **(10)** Cores (Cheng et al., 2021), a robust loss that is guaranteed to be robust to instance-dependent label noise. We also consider the Active Passive Loss (Ma et al., 2020) by including NCE+MAE and NCE+AGCE.

Can LogitClip improve the noise-robustness of existing loss functions? Table 1 and Table 2 present the average test accuracy of models trained with different noise-robust loss functions on CIFAR-10 and CIFAR-100, under various types of noisy labels. A salient observation is that our method drastically improves the noise-robustness performance of CE by employing LogitClip. For example, on the CIFAR-10 with instance dependent label noise, our approach improves the test accuracy of CE loss from 68.36% to 86.60% – a **18.24**% of direct improvement. On CIFAR-100, our method also improves performance by a significant margin. More importantly, we show that the LogitClip can boost performance for a wide range of loss functions, including non-robust and robust losses. For example, we observe that, on the CIFAR-10 with asymmetric label noise, the test accuracy of the NCE loss is improved to 88.44% when employing LogitClip, establishing strong robustness against all types of label noise. In addition to loss functions, we show our method can also enhance other deep learning methods in Appendices G.1 and G.4.

Table 2: Average test accuracy (%) with standard deviation on CIFAR-100 under various types of noisy labels (over 5 runs). The bold indicates the improved results by integrating our method.

Method	Sym-20%	Sym-50%	Asymmetric	Dependent	Real
CE	64.81±1.10	47.07±1.07	47.68±0.93	52.49±0.79	55.68±0.81
+ LC (Ours)	**71.59±0.76**	**63.16±0.74**	**59.04±0.18**	**66.24±0.71**	**58.61±0.35**
Focal	64.76±0.14	47.06±0.56	48.59±0.73	52.87±0.57	55.01±0.65
+ LC (Ours)	**71.39±0.79**	**62.91±0.25**	**59.53±0.76**	**66.38±0.30**	**58.76±0.23**
PHuber-CE	71.47±0.29	60.52±0.67	47.26±0.44	64.33±0.41	56.18±0.59
+ LC (Ours)	71.89±0.31	**61.46±0.75**	**53.95±0.38**	**65.08±0.17**	**58.64±0.49**
SCE	70.11±0.31	58.56±0.78	44.91±0.62	62.86±0.74	58.27±0.88
+ LC (Ours)	**71.19±0.23**	**60.11±0.15**	**58.67±0.84**	**64.76±0.28**	**59.23±0.69**
GCE	63.30±0.48	9.10±0.72	40.40±0.45	27.45±0.50	49.54±0.58
+ LC (Ours)	**70.22±0.52**	**62.14±1.20**	**54.41±0.64**	**66.25±0.85**	**58.77±0.76**
Cores	69.97±0.56	55.37±0.84	50.24±0.38	59.85±0.61	56.49±0.53
+ LC (Ours)	**71.67±0.26**	**62.67±0.33**	**63.32±0.70**	**66.31±0.25**	**59.23±0.34**
NCE+MAE	70.55±0.83	61.01±0.94	53.68±0.18	65.02±0.42	59.27±0.12
+ LC (Ours)	**71.59±0.65**	**62.85±0.52**	**54.51±0.73**	**66.58±0.18**	**60.08±0.34**
NCE+AGCE	69.69±0.30	58.13±0.43	58.17±0.25	64.35±0.39	58.64±0.65
+ LC (Ours)	**71.30±0.48**	**63.55±0.45**	**59.27±0.32**	**65.51±0.47**	**59.57±0.55**



(a) CIFAR-10 (b) CIFAR-100

Figure 2: The Effect of τ in LogitClip with CIFAR-10 and CIFAR-100 across various noise types.

How does the logit norm threshold τ affect the noise-robustness of LogitClip? In Figure 2a and Figure 2b, we ablate how the parameter τ in our method (*cf.* Eq. 3) affects the noise robustness performance. The analysis is based on CIFAR-10 and CIFAR-100 with four types of noisy labels, including symmetric-50%, asymmetric, dependent, and real-world noisy labels. Our results echo the analysis in Proposition 2.1 and Theorem 2.2, where a smaller τ would lead to a tighter bound on the difference of the risks between using noisy and clean labels. On the other hand, too small of τ causes a large lower bound on the loss, which is less desirable from the optimization perspective. In Appendix G.3, We clearly validate the underfitting issue caused by a small τ with experiments on a clean dataset.

Is LogitClip effective with different architectures? To show our proposed method is model-agnostic, we conduct experiments on a diverse collection of model architectures and present the results in Table 3. From the results, we observe that LogitClip consistently improves the test performance on CIFAR-10 when using SqueezeNet (Iandola et al., 2016), ResNet (He et al., 2016), DenseNet (Huang et al., 2017) architectures. For instance, with DenseNet, using LogitClip boosts the test accuracy of CE from 59.34% to 81.29%, a **21.99%** of direct improvement on CIFAR-10 with Symmetric-50% noisy labels.

3.3. WebVision

Going beyond CIFAR benchmarks, we verify the effectiveness of LogitClip on a large-scale real-world noisy dataset – WebVision (Li et al., 2017). In the WebVision dataset, there are 2.4 million images with real-world noisy labels, crawled from the web (e.g., Flickr and Google) based upon the 1,000 classes of ImageNet ILSVRC12 (Deng et al., 2009). Following the "Mini" setting used in previous works (Jiang et al.,

Table 3: Average test performance comparison on noisy CIFAR-10 with different network architectures: SqueezeNet (Iandola et al., 2016), ResNet (He et al., 2016), DenseNet (Huang et al., 2017). All values are percentages. The results are shown as CE / +LC (ours).

Architecture	Sym-20%	Sym-50%	Asymmetric	Dependent	Real
SqueezeNet	80.77 / **81.05**	54.73 / **71.64**	75.67 / **78.98**	75.43 / **76.74**	61.28 / **75.38**
ResNet-34	74.16 / **84.88**	54.48 / **74.81**	73.94 / **79.69**	58.03 / **76.51**	63.24 / **75.25**
DenseNet	80.40 / **90.85**	59.34 / **81.29**	76.35 / **84.35**	62.27 / **84.20**	63.13 / **78.93**

Table 4: Top-1 validation accuracy (%) on the clean ILSVRC12 validation set of ResNet-18 models trained on WebVision using different loss functions, under the Mini setting (Jiang et al., 2018). The bold indicates the best results. Here, "Ours" denotes CE equipped with LogitClip and "Ours+" denotes NCE+AGCE (latest state-of-the-art (Zhou et al., 2021)) equipped with LogitClip.

Method	CE	PHuber-CE	GCE	SCE	NCE+MAE	NCE+AGCE	**Ours**	**Ours+**
best	62.6	61.6	57.32	59.52	64.08	63.80	**65.12**	64.92
last	60.84	59.76	53.26	58.47	62.85	62.46	63.75	**64.50**

2018; Ma et al., 2020; Zhou et al., 2021), we take the first 50 classes of the Google resized image subset. For evaluation, we test the trained networks on the same 50 classes of the ILSVRC12 validation set, which can be seen as a clean validation. For each loss, we train a ResNet-18 network using SGD for 120 epochs with an initial learning rate of 0.1, Nesterov momentum 0.9, weight decay 5×10^{-4}, and batch size 128. The learning rate is reduced by a factor of 10 after 40 and 80 epochs. We resize the images to 224×224 and apply the standard data augmentations, including random cropping and random horizontal flip. As shown in Table 4, *best* denotes the score of the epoch where the validation accuracy is optimal, and *last* denotes the scores at the average accuracy in the last 10 epochs. As shown in the table, LogitClip not only outperforms but also enhances existing loss functions by a meaningful margin. The results verify that our method is effective for improving noise-robustness in large-scale real-world scenarios.

4. Discussion

Relations to existing clipping methods. In the literature, clipping-based methods have been studied in the context of deep learning (Bengio et al., 1994; Zhang et al., 2020; Abadi et al., 2016; Howard et al., 2017; Sun et al., 2021). One of the most classic clipping-based methods is gradient clipping, a widely used technique in recurrent neural networks (Bengio et al., 1994), optimization (Hazan et al., 2015; Levy, 2016; Zhang et al., 2020), and privacy (Abadi et al., 2016; Pichapati et al., 2019). In the simplest form, gradient clipping is designed to constrain the global parameter gradient norm at a specified threshold. With a loss function ℓ_θ, the *clipped* gradient with a user-specified threshold $\tau > 0$ can be computed as:

$$\bar{g}_\tau(\theta) \doteq \mathrm{clip}_\tau(g(\theta)),\ \mathrm{clip}_\tau(w) \doteq \begin{cases} \frac{\tau \cdot w}{\|w\|_2} & \text{if } \|w\|_2 \geq \tau \\ w & \text{else} \end{cases},$$

where $g(\theta)$ denotes the gradient for a mini-batch: $g(\theta) \doteq \frac{1}{b}\sum_{n=1}^{b} \nabla \ell_\theta(x_n, y_n)$. Different from gradient clipping which limits the norm of the parameter gradient, our LogitClip method places the constraint directly on the *model output*, i.e., the logit vector. Our method is thus designed to have a *direct and explicit* effect in bounding the loss (*cf.* Theorem 2.2 and Theorem 2.3) and preventing overfitting to examples with noisy labels.

Indeed, recent work (Menon et al., 2020) has shown that gradient clipping alone does not endow label noise robustness to neural networks. They instead proposed a noise-robust variant, composite loss-based gradient clipping and the resulting partially Huberised loss (PHuber-CE). The results in Section 3 have shown that LogitClip not only outperforms but also enhances the performance of PHuber-CE loss. Our results overall demonstrate the superiority and complementarity of LogitClip to gradient clipping on noise robustness.

ReLU6 (Howard et al., 2017) is a modification of the rectified linear unit (ReLU) to facilitate the learning of sparse features. In particular, ReLU6 clamps the activation of the intermediate layers to a maximum value of 6, $\mathrm{ReLU6}(\boldsymbol{x}) = \min(\max(0, \boldsymbol{x}), 6)$. Although ReLU6 provides a constraint on the outputs of the intermediate layers, the resulting loss is unbounded yet since the final output can be multiplied through the last linear layer. The empirical results in Figure 3 show that ReLU6 cannot enhance the robustness to noisy labels while our method (LC-N) achieves a significant improvement.

Clipping-by-value vs. Clipping-by-norm. While our logit clipping has demonstrated strong promise in the manner of Clipping-by-norm, one may also ask: can a similar effect be achieved by clipping the logit vector by value? In this ablation, we show that directly constraining the maximum and minimum values of the logit vector does not work well as our method. In particular, we consider the link function as softmax with Clipping-by-value:

$$\bar{\sigma}'_\lambda(\boldsymbol{z}) \doteq \sigma(\text{clip}'_\lambda(\boldsymbol{z})) \quad \text{clip}'_\lambda(\boldsymbol{z}_j) \doteq \begin{cases} \lambda & \text{if } \boldsymbol{z}_j \geq \lambda \\ -\lambda & \text{if } \boldsymbol{z}_j \leq -\lambda \\ \boldsymbol{z}_j & \text{else} \end{cases},$$

where λ denotes the constant threshold. For convenience, we set the maximum and minimum values as λ and $-\lambda$, respectively. We search the best λ in $\{0.1, 0.5, 1, 1.5, \ldots, 4.5, 5\}$.

Figure 3 presents the performance comparison between our method and the variant of Clipping-by-value, denoted as LC-N and LC-V, respectively. While both the two logit clipping methods improve the robustness of CE against noisy labels, LC-V obtains inferior performance compared to our proposed method, and the gaps become remarkably significant under complex noise settings, *e.g.*, instance-dependent and real-world label noise. From a theoretical perspective, CE with LC-V also satisfies the bound in Proposition 2.1 and is hence applicable for the two bounds in Theorem 2.2 and Theorem 2.3, which indicates the noise-robustness of this method. Nevertheless, LC-V is suboptimal as the clipping operation may diminish the gradients on the clipped components of the logit vector. Besides, LC-V would also modify the direction of the input vector and even change the final prediction. Overall, we demonstrate that our method is superior to the variant of clipping-by-value.

Relations to LogitNorm. A concurrent work (Wei et al., 2022a) employs logit normalization (LogitNorm) to improve the OOD detection and calibration performance. For all training inputs, the logit vector is normalized to be a unit vector with a constant norm and the resulting loss is defined as: $\mathcal{L}_{\text{logit_norm}}(f(\boldsymbol{x};\theta), y) = -\log \frac{e^{f_y/(\tau\|\boldsymbol{f}\|)}}{\sum_{i=1}^{k} e^{f_i/(\tau\|\boldsymbol{f}\|)}}$. Our work bears three critical differences, in terms of the problem setting, methodology, and theory.

(1) *Problem setting*: LogitNorm focuses on improving the performance of detecting out-of-distribution (OOD) examples during the test time, while our work aims to enhance the robustness against noisy labels in the training stage. The learning tasks are fundamentally different.

(2) *Methodology*: We propose to clamp the logit vector to ensure it is *upper bounded by a constant*, while LogitNorm enforces the norm of logit vectors to be an *exact constant* for all samples. From a constrained optimization perspective,

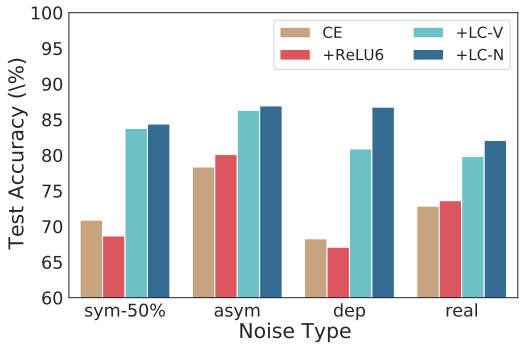


Figure 3: Performance comparison among ReLU6, LC-V, and our method (LC-N) on noisy CIFAR-10.

Table 5: Comparison between LogitClip and LogitNorm on the CIFAR-10 dataset with various noise settings.

Method	Sym-50%	Asymmetric	Dependent	Real
LogitNorm	83.97	81.81	84.56	80.10
LogitClip	84.37	**86.91**	**86.74**	**82.06**

LogitNorm enforces *equality constraint* on the L_2 norm of logit vector, whereas LogitClip enforces *inequality constraint*. Hence, LogitClip enforces a less strict objective than LogitNorm. Referring to the relationship between Gradient Clipping (Abadi et al., 2016; Zhang et al., 2020) and Gradient Normalization (NGD) (Hazan et al., 2015; Murray et al., 2019), LogitClip is a unique method that differs from LogitNorm.

In Table 5, we present the performance comparison of LogitClip and LogitNorm in learning with noisy labels. From the comparison, we find that LogitClip is superior to the LogitNorm in this task, especially in those complicated settings. For example, in the asymmetric setting, LogitClip outperforms the LogitNorm method by a large margin of 5.1%. Intuitively, LogitNorm can improve the robustness to label noise because it also induces a loss bound. However, it enforces a more strict constraint on all training examples, which may make LogitNorm suboptimal in this task.

(3) *Theoretical insight*: LogitNorm aims to decouple the influence of logits' magnitude from network optimization, and they empirically show that their method leads to more meaningful information to differentiate in-distribution and OOD samples. In contrast, our LogitClip is designed to enforce an upper bound of the resulting loss, as shown in Proposition 2.1. Furthermore, we provide a theoretical interpretation to further understand why LogitClip introduces the noise-tolerant ability and which kinds of loss functions LogitClip can work with. In summary, our analysis builds a connection between LogitClip and noise robustness, which is novel to the best of our knowledge.

5. Related Work

Robust loss functions. Designing loss functions that are robust to noisy labels has attracted a surge of interest. Ghosh et al. first shows that, for multi-class classification, loss functions that satisfy the symmetric condition, $\sum_{i=1}^{k} \mathcal{L}(f(\mathbf{x}), i) = C, \forall x \in \mathcal{X}, \forall f$, where C is a constant, can be robust to label noise. One of the most classic symmetric loss functions is Mean Absolute Error (MAE): $\mathcal{L}_{\text{MAE}}(\boldsymbol{z}, y) = \|\boldsymbol{y} - \sigma(\boldsymbol{z})\|_1$, where $\sigma(\boldsymbol{z})$ denotes the softmax output. Besides, NCE (Ma et al., 2020) makes any loss function to be symmetric by loss normalization. Despite its theoretical robustness, symmetric losses have been shown to exhibit extremely slow convergence on complicated datasets, since the symmetric condition is too stringent to find a convex loss function (Zhang & Sabuncu, 2018; Ma et al., 2020). To alleviate this issue, Generalized Cross Entropy (GCE) (Zhang & Sabuncu, 2018) uses a hyperparameter q to balance between MAE and CE, adopting the negative Box-Cox transformation strategy. Taylor Cross Entropy (Taylor-CE) (Feng et al., 2020) balances between MAE and CE by controlling the order of the Taylor Series. Partial Huberised Cross Entropy (PHuber-CE) (Menon et al., 2020) enhances the noise robustness of CE with a loss variant of gradient clipping. Symmetric Cross Entropy (SCE) (Wang et al., 2019) boosts CE symmetrically with a noise-robust counterpart, Reverse Cross Entropy (RCE). Recent work (Zhou et al., 2021) proposes a new family of robust loss functions, termed asymmetric loss functions, which provides a global clean weighted risk when minimizing the noisy risk for any hypothesis class. In this work, our focus is complementary to existing robust loss functions — we propose a strategy that can universally enhance the noise robustness of existing losses, across various types of label noise.

Other deep learning methods for noise-robustness. In addition to robust loss functions, some other solutions are also applied to learn with noisy labels (Xia et al., 2019; Li et al., 2022b;a; Wu et al., 2021a; Chen et al., 2023; Shu et al., 2023; 2021; Wu et al., 2021b; Ding et al., 2023; Zhu et al., 2021b; 2022b;a; Cheng et al., 2023; Wei et al., 2023; Liu et al., 2023; Wei et al., 2022c; Wei & Liu, 2021; Wei et al., 2022b), including: 1) Some methods aim to design sample weighting schemes that give higher weights on clean samples (Jiang et al., 2018; Liu & Tao, 2015; Ren et al., 2018; Shu et al., 2019; Wei et al., 2020b). 2) Some methods propose to train on selected samples, using small-loss selection (Han et al., 2018; Wei et al., 2020a; Yu et al., 2019; Xia et al., 2022), GMM distribution (Arazo et al., 2019; Li et al., 2020) or (dis)agreement between two models (Malach & Shalev-Shwartz, 2017; Wei et al., 2020a; Yu et al., 2019). 3) Loss correction is also a popular direction based on an estimated noise transition matrix (Hendrycks et al., 2018; Patrini et al., 2017), or the model's predictions (Arazo et al., 2019; Chen et al., 2020; Reed et al., 2014; Tanaka et al., 2018; Zheng et al., 2020). 4) Some methods apply regularization techniques to improve generalization under the settings of label noise (Fatras et al., 2021; Hu et al., 2019; Xia et al., 2020a; Liu & Guo, 2020; Bai et al., 2021; Liu et al., 2020; Zhu et al., 2021a; Liu et al., 2022b), such as label smoothing (Lukasik et al., 2020; Szegedy et al., 2016), temporal ensembling (Laine & Aila, 2016), and virtual adversarial training (Miyato et al., 2018). 5) Some training strategies for combating noisy labels are built based upon semi-supervised learning methods (Li et al., 2020; Nguyen et al., 2020) or self-supervised learning (Li et al., 2022a). Compared to the above deep learning methods, designing robust loss function are generally a more straightforward and arguably more generic solution with theoretical guarantees.

6. Conclusion

In this paper, we propose Logit Clipping (**LogitClip**), a general strategy that can universally enhance the noise robustness of existing losses, across various types of label noise. Specifically, we propose to clamp the norm of the logit vector to ensure that it is upper bounded by a constant. In this manner, CE loss equipped with our LogitClip method is effectively bounded, alleviating overfitting to examples with noisy labels. As a result, our method could mitigate the undesirable influence of unbounded loss without modifying the loss function. Moreover, we present theoretical analyses to certify the noise-tolerant ability of this method. Extensive experiments show that LogitClip not only significantly improves the noise robustness of CE loss, but also broadly enhances the generalization performance of popular robust losses. This method is straightforward to implement with existing losses and can be easily adopted in various practical settings. We hope that our method inspires future theoretical research to explore robust loss from the logit perspective.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Prepositioning (IAF-PP) Funding Initiative. Lei Feng was supported by the National Natural Science Foundation of China (Grant No. 62106028), Chongqing Overseas Chinese Entrepreneurship and Innovation Support Program. Li is supported in part by the AFOSR Young Investigator Award under No. FA9550-23-1-0184; and faculty research awards/gifts from Google, Meta, and Amazon. We gratefully acknowledge the support of Center for Computational Science and Engineering at Southern University of Science and Technology for our research. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the sponsors.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on Computer and Communications Security*, pp. 308–318, 2016.

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pp. 312–321. PMLR, 2019.

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 *Proceedings of the 34th International Conference on Machine Learning*, pp. 233–242, 2017.

Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., and Liu, T. Understanding and improving early stopping for learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2021.

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Blum, A., Kalai, A., and Wasserman, H. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

Chen, H., Shah, A., Wang, J., Tao, R., Wang, Y., Xie, X., Sugiyama, M., Singh, R., and Raj, B. Imprecise label learning: A unified framework for learning with various imprecise label configurations. *arXiv preprint arXiv:2305.12715*, 2023.

Chen, P., Ye, J., Chen, G., Zhao, J., and Heng, P.-A. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. *arXiv preprint arXiv:2012.05458*, 2020.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. Learning with instance-dependent label noise: A sample sieve approach. *International Conference on Learning Representations*, 2021.

Cheng, H., Zhu, Z., Sun, X., and Liu, Y. Mitigating memorization of noisy labels via regularization between representations. In *International Conference on Learning Representations (ICLR)*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Ding, K., Shu, J., Meng, D., and Xu, Z. Improve noise tolerance of robust loss via noise-awareness. *arXiv preprint arXiv:2301.07306*, 2023.

Fatras, K., Damodaran, B. B., Lobry, S., Flamary, R., Tuia, D., and Courty, N. Wasserstein adversarial regularization for learning with label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. Can cross entropy loss be robust to label noise? In *International Joint Conference on Artificial Intelligence*, pp. 2206–2212, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 1919–1925, 2017.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pp. 8527–8537, 2018.

Hazan, E., Levy, K., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

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.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *arXiv preprint arXiv:1802.05300*, 2018.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Hu, W., Li, Z., and Yu, D. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *arXiv preprint arXiv:1905.11368*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 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.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Levy, K. Y. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Li, S., Xia, X., Ge, S., and Liu, T. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022a.

Li, S., Xia, X., Zhang, H., Zhan, Y., Ge, S., and Liu, T. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *NeurIPS*, 2022b.

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, M., Wei, J., Liu, Y., and Davis, J. Do humans and machines have the same eyes? human-machine perceptual differences on image classification. *arXiv preprint arXiv:2304.08733*, 2023.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020.

Liu, S., Zhu, Z., Qu, Q., and You, C. Robust training under label noise by over-parameterization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 14153–14172. PMLR, 2022a.

Liu, S., Zhu, Z., Qu, Q., and You, C. Robust training under label noise by over-parameterization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14153–14172. PMLR, 2022b.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

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.

Lukasik, M., Bhojanapalli, S., Menon, A., and Kumar, S. Does label smoothing mitigate label noise? In *Proceedings of International Conference on Machine Learning*, pp. 6448–6458. PMLR, 2020.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Malach, E. and Shalev-Shwartz, S. Decoupling "when to update" from "how to update". *Advances in Neural Information Processing Systems*, 30, 2017.

Menon, A. K., Rawat, A. S., Kumar, S., and Reddi, S. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations (ICLR)*, 2020.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 41 (8):1979–1993, 2018.

Murray, R., Swenson, B., and Kar, S. Revisiting normalized gradient descent: Fast evasion of saddle points. *IEEE Transactions on Automatic Control*, 64(11):4818–4824, 2019.

Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*, 2020.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Pichapati, V., Suresh, A. T., Yu, F. X., Reddi, S. J., and Kumar, S. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, pp. 1919–1930, 2019.

Shu, J., Meng, D., and Xu, Z. Learning an explicit hyperparameter prediction policy conditioned on tasks. *arXiv preprint arXiv:2107.02378*, 2021.

Shu, J., Yuan, X., Meng, D., and Xu, Z. Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Sohrab, H. H. *Basic real analysis*, volume 231. Springer, 2003.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5552–5560, 2018.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

Wei, H., Feng, L., Chen, X., and An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13726–13735, 2020a.

Wei, H., Feng, L., Wang, R., and An, B. Metainfonet: Learning task-guided information for sample reweighting. *arXiv preprint arXiv:2012.05273*, 2020b.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*. PMLR, 2022a.

Wei, H., Xie, R., Feng, L., Han, B., and An, B. Deep learning from multiple noisy annotators as a union. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2022b.

Wei, J. and Liu, Y. When optimizing f-divergence is robust with label noise. In *International Conference on Learning Representations*, 2021.

Wei, J., Liu, H., Liu, T., Niu, G., Sugiyama, M., and Liu, Y. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning*, pp. 23589–23614. PMLR, 2022c.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022d.

Wei, J., Zhu, Z., Luo, T., Amid, E., Kumar, A., and Liu, Y. To aggregate or not? learning with separate noisy labels. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, pp. 11285–11295, 2021a.

Wu, Y., Shu, J., Xie, Q., Zhao, Q., and Meng, D. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10388–10396, 2021b.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., and Chang, Y. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020a.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020b.

Xia, X., Liu, T., Bo, H., Mingming, G., Jun, Y., Gang, N., and Masashi, S. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2022.

Yan, Y., Rosales, R., Fung, G., Subramanian, R., and Dy, J. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327, 2014.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and Sugiyama, M. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pp. 7164–7173. PMLR, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *BMVC*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *Proceedings of International Conference on Learning Representations*, 2016.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

Zheng, S., Wu, P., Goswami, A., Goswami, M., Metaxas, D., and Chen, C. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pp. 11447–11457. PMLR, 2020.

Zhou, X., Liu, X., Jiang, J., Gao, X., and Ji, X. Asymmetric loss functions for learning with noisy labels. In *International Conference on Machine Learning*, pp. 12846–12856. PMLR, 2021.

Zhu, Z., Liu, T., and Liu, Y. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10113–10123, 2021a.

Zhu, Z., Song, Y., and Liu, Y. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning (ICML)*, 2021b.

Zhu, Z., Dong, Z., and Liu, Y. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning (ICML)*, 2022a.

Zhu, Z., Wang, J., and Liu, Y. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning (ICML)*. PMLR, 17–23 Jul 2022b.

A. Proof of Proposition 2.1

Proof. Give a logit vector $z = f(\boldsymbol{x}; \boldsymbol{\theta})$, for any class pair j and k, we have:

$$\boldsymbol{z}_{\min} - \boldsymbol{z}_{\max} \leq \boldsymbol{z}_j - \boldsymbol{z}_k \leq \boldsymbol{z}_{\max} - \boldsymbol{z}_{\min}$$

Recall that $-\tau \leq \text{clip}_\tau(\boldsymbol{z}_j) \leq \tau$, then:

$$-2\tau \leq \boldsymbol{z}_j - \boldsymbol{z}_k \leq 2\tau$$

Based on Equation (2), we have:

$$\log(1 + (K-1) \cdot e^{-2\tau}) \leq \mathcal{L}_{\text{CE}}^{\tau}\left(f(\boldsymbol{x}; \boldsymbol{\theta}), y\right) \leq \log(1 + (K-1) \cdot e^{2\tau}).$$

Thus Proposition 2.1 is proved. □

B. Proof of Theorem 2.2

Proof. Recall that for symmetric label noise with noise rate η, we have: $\eta_{jk} = 1 - \eta$ for $j = k$, and $\eta_{jk} = \frac{\eta}{K-1}$. Then, for any model output $f(\boldsymbol{x})$,

$$\begin{aligned}
\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(f) =& \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\text{noisy}}^{\eta}}\left[\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), y)\right] \\
=& \mathbb{E}_{\mathcal{P}_{\boldsymbol{x}}} \mathbb{E}_{\mathcal{P}_{y^\star|\boldsymbol{x}}} \mathbb{E}_{\mathcal{P}_{y|y^\star}}\left[\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), y)\right] \\
=& \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[(1-\eta)\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), y^\star) + \frac{\eta}{K-1}\sum_{j\neq y^\star}\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), j)\right] \\
=& (1-\eta)\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f) + \frac{\eta}{K-1}\left(\sum_{j=1}^{K}\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), j) - \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f)\right) \\
=& \left(1 - \frac{\eta K}{K-1}\right)\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f) + \frac{\eta}{K-1}\sum_{j=1}^{K}\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), j).
\end{aligned}$$

From proposition 2.1, we have:

$$K\log(1 + (K-1) \cdot e^{-2\tau}) \leq \sum_{j}^{K}\mathcal{L}_{\text{CE}}^{\tau}(f(\boldsymbol{x}), j) \leq K\log(1 + (K-1) \cdot e^{2\tau}).$$

Thus,

$$\beta\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f) + \frac{\eta K}{K-1}\log(1 + (K-1)e^{-2\tau}) \leq \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(f) \leq \beta\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}} + \frac{\eta K}{K-1}\log(1 + (K-1)e^{2\tau}).$$

where $\beta = (1 - \frac{\eta K}{K-1})$. We can also write the inequality in terms of $\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(f)$:

$$\frac{1}{\beta}\left(\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(f) - \frac{\eta K}{K-1}\log(1 + (K-1)e^{2\tau})\right) \leq \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f) \leq \frac{1}{\beta}\left(\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta} - \frac{\eta K}{K-1}\log(1 + (K-1)e^{-2\tau})\right).$$

For $\tilde{f}$, we have:

$$\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(\tilde{f}) - \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f^\star) \leq \frac{1}{\beta}\left(\frac{\eta K}{K-1}\log A_\eta^K + \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(\tilde{f}) - \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(f^\star)\right),$$

where $A_\eta^K = \left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right)$. Since $\eta \leq 1 - \frac{1}{K}$, $f^\star$ is the global minimizer of $\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(\tilde{f})$ and $\tilde{f}$ is the global minimizer of $\mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}^{\eta}(\tilde{f})$, we have

$$0 \leq \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(\tilde{f}) - \mathcal{R}_{\mathcal{L}_{\text{CE}}^{\tau}}(f^\star) \leq \frac{\eta K}{(1-\eta)K - 1} \cdot A_\tau^K,$$

which concludes the proof. □

C. Proof of Theorem 2.3

Proof. For asymmetric label noise, we have

$$\begin{aligned}
\mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f) =& \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}^{\eta}_{\text{noisy}}}\left[\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),y)\right] \\
=& \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[(1-\eta_i)\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),y^\star)\right] + \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\eta_{ij}\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),j)\right] \\
\leq& \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[(1-\eta_i)\left(K\log(1+(K-1)\cdot e^{2\tau}) - \sum_{j\neq y^\star}\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),j)\right)\right] \\
&+ \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\eta_{ij}\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),j)\right] \\
=& K\log(1+(K-1)\cdot e^{2\tau})\cdot\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}(1-\eta_i) - \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y}\lambda_j\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),j)\right],
\end{aligned}$$

where $\lambda_j = (1-\eta_i-\eta_{ij})$. On the other hand, we have

$$\mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f) \geq K\log(1+(K-1)\cdot e^{-2\tau})\cdot\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}(1-\eta_i) - \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\lambda_j\mathcal{L}^{\tau}_{\mathrm{CE}}(f(\boldsymbol{x}),j)\right]$$

Hence,

$$\mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f^\star) - \mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(\tilde{f}) \leq B^K_\tau + \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\lambda_j\left(\mathcal{L}^{\tau}_{\mathrm{CE}}\left(\tilde{f}_j(\boldsymbol{x}),j\right) - \left(\mathcal{L}^{\tau}_{\mathrm{CE}}\left(f^*(\boldsymbol{x}),j\right)\right)\right)\right]$$

where $B^K_\tau = K\log\left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right)\mathbb{E}_{p(\boldsymbol{x},y)}(1-\eta_i)$. Let q^- and q^+ denote the lower bound and the upper bound in Proposition 2.1. We assume $\mathcal{R}^{\eta}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f^\star) = q^-$, i.e., $\mathcal{L}^{\tau}_{\mathrm{CE}}(f^\star(\boldsymbol{x}),y^\star) = q^-$, which is only satisfied iff $f^\star_j(\boldsymbol{x}) = \tau$ when $j = y^\star$ and $f^\star_j(\boldsymbol{x}) = -\tau$ when $j \neq y^\star$. In this case, $\mathcal{L}^{\tau}_{\mathrm{CE}}(f^\star(\boldsymbol{x}),j) = q^+, \forall j \neq y^\star$ and $\mathcal{L}^{\tau}_{\mathrm{CE}}(f^\star(\boldsymbol{x}),j) \leq q^+, \forall j \in [K]$. Since $f^\star$ is the global minimizer of $\mathcal{R}_{\mathcal{L}^{\tau}_{\mathrm{CE}}}(f)$ and $\lambda = (1-\eta_i-\eta_{ij}) > 0$, we have

$$\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\lambda_j\left(\mathcal{L}^{\tau}_{\mathrm{CE}}(\tilde{f}(\boldsymbol{x}),j) - \mathcal{L}^{\tau}_{\mathrm{CE}}(f^\star(\boldsymbol{x}),j)\right)\right] \leq 0.$$

Thus, we have

$$0 \leq \mathcal{R}^{\eta}_{\mathcal{L}}(f^\star) - \mathcal{R}^{\eta}_{\mathcal{L}}(\tilde{f}) \leq B^K_\tau,$$

which concludes the proof. □

D. Proof of Proposition 2.4

Proof. Let $\boldsymbol{p} = \sigma'_\tau(\boldsymbol{z})$. Recall that $-\tau \leq \boldsymbol{z}_j \leq \tau, \forall j \in [K]$, we have:

$$\frac{1}{1+(K-1)\cdot e^{2\tau}} \leq \boldsymbol{p}_j \leq \frac{1}{1+(K-1)\cdot e^{-2\tau}}, \forall j \in [K],$$

Let M^K_τ and N^K_τ denote the lower and upper bound of $\boldsymbol{p}_j$. Given that the base loss $\phi(\boldsymbol{p}_y)$ satisfies the Lipschitz condition with constant L on the domain. For any $\boldsymbol{p}_y \in [M^K_\tau, N^K_\tau]$, we have

$$|\phi(\boldsymbol{p}_y) - \phi(M^K_\tau)| \leq L|\boldsymbol{p}_y - M^K_\tau| \leq L|N^K_\tau - M^K_\tau|,$$

and

$$\begin{aligned}|\phi(\boldsymbol{p}_y)| &= |\phi(\boldsymbol{p}_y) - \phi(M_\tau^K) + \phi(M_\tau^K)| \\ &\leq |\phi(\boldsymbol{p}_y) - \phi(M_\tau^K)| + |\phi(M_\tau^K)| \\ &\leq L|N_\tau^K - M_\tau^K| + |\phi(M_\tau^K)|.\end{aligned}$$

Since $N_\tau^K - M_\tau^K \geq 0, \forall \tau > 0$, we have:

$$\left|\mathcal{L}_\phi^\tau\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\right| \leq L\left(N_\tau^K - M_\tau^K\right) + \left|\phi(M_\tau^K)\right|,$$

which concludes the proof.

□

E. Theoretical analysis under instance-dependent setting

In Theorem 2.2 and Theorem 2.3, we have provably shown the noise-tolerant ability of cross-entropy loss with LogitClip. Here, we extend the theoretical analysis to an *instance-dependent setting*, where the noise rate $\eta_{\boldsymbol{x}}$ is a function of instance $\boldsymbol{x}$ and $\eta_{\boldsymbol{x}j}$ may vary across classes j.

Theorem E.1. *Under instance-dependent label noise with* $1 - \eta_{\boldsymbol{x}} > \eta_{\boldsymbol{x}j}, \forall \boldsymbol{x}, j \neq y_{\boldsymbol{x}}^\star$*, where* $\eta_{\boldsymbol{x}j} = p(y = j \mid \boldsymbol{x}), \forall j \neq i$ *and* $(1 - \eta_{\boldsymbol{x}}) = p(y = i \mid \boldsymbol{x}, y^\star = i)$*, then*

$$0 \leq \mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(f^*) - \mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(\tilde{f}) \leq C_\tau^K$$

where $C_\tau^K = K\log\left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right)\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{clean}}(1-\eta_{\boldsymbol{x}}) > 0.$

Proof. For instance-dependent label noise, we have

$$\begin{aligned}\mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(f) =& \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\text{noisy}}^\eta}\left[\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), y)\right] \\ =& \mathbb{E}_{\mathcal{P}_{\boldsymbol{x}}}\mathbb{E}_{\mathcal{P}_{y^\star|\boldsymbol{x}}}\mathbb{E}_{\mathcal{P}_{y|\boldsymbol{x},y^\star}}\left[\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), y)\right] \\ =& \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[(1-\eta_{\boldsymbol{x}})\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), y^\star)\right] + \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\eta_{\boldsymbol{x}j}\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), j)\right] \\ \leq& \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[(1-\eta_{\boldsymbol{x}})\left(K\log(1+(K-1)\cdot e^{2\tau}) - \sum_{j\neq y^\star}\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), j)\right)\right] \\ &+ \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\eta_{\boldsymbol{x}j}\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), j)\right] \\ =& K\log(1+(K-1)\cdot e^{2\tau})\cdot\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}(1-\eta_{\boldsymbol{x}}) - \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y}\lambda_{\boldsymbol{x}j}\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), j)\right],\end{aligned}$$

where $\lambda_j = (1 - \eta_i - \eta_{\boldsymbol{x}j})$. On the other hand, we have

$$\mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(f) \geq K\log(1+(K-1)\cdot e^{-2\tau})\cdot\mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}(1-\eta_{\boldsymbol{x}}) - \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\lambda_{\boldsymbol{x}j}\mathcal{L}_{\mathrm{CE}}^\tau(f(\boldsymbol{x}), j)\right]$$

Hence,

$$\mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(f^\star) - \mathcal{R}_{\mathcal{L}_{\mathrm{CE}}^\tau}^\eta(\tilde{f}) \leq C_\tau^K + \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}\left[\sum_{j\neq y^\star}\lambda_{\boldsymbol{x}j}\left(\mathcal{L}_{\mathrm{CE}}^\tau\left(\tilde{f}_j(\boldsymbol{x}), j\right) - \left(\mathcal{L}_{\mathrm{CE}}^\tau\left(f^*(\boldsymbol{x}), j\right)\right)\right)\right]$$

where $C_\tau^K = K \log\left(\frac{1+(K-1)e^{2\tau}}{1+(K-1)e^{-2\tau}}\right) \mathbb{E}_{(\boldsymbol{x},y^\star)\sim\mathcal{P}_{\text{clean}}}(1-\eta_{\boldsymbol{x}})$. From the proof of Theorem 2.3, we have $\mathcal{L}_{\text{CE}}^\tau(\tilde{f}(\boldsymbol{x}),j) - \mathcal{L}_{\text{CE}}^\tau(f^\star(\boldsymbol{x}),j) \leq 0, \forall \boldsymbol{x}, j \neq y_{\boldsymbol{x}}^\star$. Recall that $\lambda = (1-\eta_{\boldsymbol{x}} - \eta_{\boldsymbol{x}j}) > 0$, we have

$$\mathbb{E}_{p(\boldsymbol{x},y)}\left[\sum_{j\neq y^\star} \lambda_{\boldsymbol{x}j}\left(\mathcal{L}_{\text{CE}}^\tau(\tilde{f}(\boldsymbol{x}),j) - \mathcal{L}_{\text{CE}}^\tau(f^\star(\boldsymbol{x}),j)\right)\right] \leq 0.$$

Thus, we have

$$0 \leq \mathcal{R}_{\mathcal{L}}^\eta(f^\star) - \mathcal{R}_{\mathcal{L}}^\eta(\tilde{f}) \leq C_\tau^K,$$

which concludes the proof. □

F. More details on experimental setup

Hyperparameter setting. We conduct all the experiments on NVIDIA GeForce RTX 3090, and implement all methods by PyTorch. We tune the hyperparameters for all compared methods and find that the optimal settings basically match those in their original papers. Specifically, for GCE, we set $q = 0.7$. For SCE, we set $\alpha = 0.5$ and $\beta = 1.0$. For AEL loss, we set $a = 2.5$. For AUL loss, we set $a = 5.5$ and $q = 3$. For PHuber-CE, we set $\tau = 10$ for CIFAR-10 and $\tau = 30$ for CIFAR-100 and WebVision. For the experiments of NCE+MAE on CIFAR-100 and WebVision, we set $\alpha = 50$ and $\beta = 1$. For NCE+AGCE on CIFAR-100, we set $\alpha = 50$, $\beta = 0.1$, $a = 1.8$ and $q = 3.0$. On WebVision, we set the hyperparameters of NCE+AGCE as $\alpha = 50$, $\beta = 0.1$, $a = 2.5$, and $q = 3.0$. For the best τ of our LogitClip, it may depend on the dataset, noise type, and the base loss. In Table 6, we present the best values of $1/\tau$ in CE with LogitClip.

Table 6: Best Values of $1/\tau$ for CE+LogitClip on different datasets with various noise settings.

Dataset	Symmetric-20%	Symmetric-50%	Asymmetric	Dependent	Real-world
CIFAR-10	1.0	1.5	2.5	2.0	2.5
CIFAR-100	0.5	0.5	2.5	0.5	0.5
WebVision			1.2		

G. More empirical results

G.1. Can LogitClip improve deep learning methods?

In the experiments shown in Section 3, we show that our method can consistently improve the noise robustness of existing popular losses, including non-robust losses and robust losses. One may raise the question: Can LogitClip improve deep learning methods? Here, we use DivideMix (Li et al., 2020) as the representative method to show the universality of our LogitClip. For the experiments with DivideMix, we use the same setting as those reported in the paper of DivideMix. Specifically, we use an 18-layer PreAct Resnet (He et al., 2016) and train it using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. The network is trained for 300 epochs. The warm-up period is 10 epochs for CIFAR-10. For the hyperparameters, we set $M = 2, T = 0.5$, $\alpha = 4$, and $\tau = 0.5$. For DivideMix + LogitClip, we employ the logit clipping for all the model outputs during training. The test performance for DivideMix on noisy CIFAR-10 is reported in Table 7. From the results, we can observe that our LogitClip can consistently improve the performance of DivideMix with a meaningful margin, which validates the universality of our method in boosting noise robustness.

Table 7: Test performance comparison for DivideMix (Li et al., 2020) on noisy CIFAR-10 with different noisy types. The results show that our method can boost the performance of DivideMix.

Method	Symmetric-50%	Asymmetric	Dependent	Real-world
DivideMix	94.41	92.02	94.11	92.24
+LC(Ours)	**95.15**	**92.73**	**95.25**	**93.16**

G.2. Logit Clipping vs. Norm Regularization

As demonstrated in Subsection 2.2, our training objective can be formalized as a constrained optimization with inequality constraint. Therefore, we may consider an alternative method by simply adding the constraint via the Lagrangian multiplier, termed Norm Regularization:

$$\mathcal{L}_{\text{logit_penalty}}(f(\boldsymbol{x};\theta), y) = \mathcal{L}_{\text{CE}}(f(\boldsymbol{x};\theta), y) + \lambda \|f(\boldsymbol{x};\theta)\|_2.$$

In the experiments, we select the best λ in $\{0.01, 0.05, 0.1, 0.5\}$. Our results in Figure 4 show that Norm Regularization is inferior to our LogitClip across four noise types, while both methods improve the test accuracy compared to Cross Entropy loss. With Norm Regularization, we notice that the trained network can suffer from optimization difficulty and sometimes fail to converge if λ is too large (which is needed to regularize the logit norm effectively). Overall, we show that simply constraining the logit norm during training cannot achieve comparable performance as our LogitClip, which significantly improves the noise robustness.

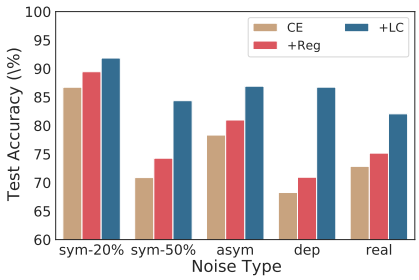


Figure 4: Test performance comparison among CE, Norm Regularization (+Reg), and our method (+LC) on noisy CIFAR-10 across different noise types.

G.3. Performance on Clean datasets

We evaluate the performance of LogitClip on the clean CIFAR-100 dataset with different τ. As discussed in Section 2 (Proposition 3.1), LogitClip is equivalent to vanilla cross-entropy loss if the τ is sufficiently large. Contrastively, a small τ will induce a large lower bound on the loss value, which may lead to difficulty in loss optimization (Underfitting). As shown in Table 8, LogitClip with a large τ achieves comparable performance as the vanilla CE loss. Besides, the performance of LogitClip can be degraded as we decrease the value of τ, which validates the underfitting issue in our analysis.

Table 8: Performance of CE+LogitClip under different τ on the clean CIFAR-100 dataset.

τ	20	10	2	1	0.5	0.25	CE
Accuracy	**75.90**	75.51	74.38	73.23	68.00	46.04	**75.98**

G.4. Logit Clipping can improve the latest SOTA methods

In this paper, we mainly focus on improving existing robust losses. Besides, we show our method can also enhance some other deep learning methods by improving DivideMix (see Appendix G.1). In the tables 9, we provide empirical evidence that our method can meaningfully improve SOP (Liu et al., 2022a) and SAM (Foret et al., 2021). We use the same settings

as Section 3 of our paper for SAM, while adopting the settings reported in their paper for SOP. These results further demonstrate the complementarity of LogitClip with previous techniques.

Table 9: Test performance comparison for SOP (Liu et al., 2022a) and SAM (Foret et al., 2021) on noisy CIFAR-10/100 with different noisy types. The results show that our method can boost the performance of SAM and SOP.

Method	CIFAR10-Sym-50%	CIFAR10-ASym-40%	CIFAR100-Sym-50%	CIFAR100-ASym-40%
SOP	88.31	84.43	61.68	66.89
+LC(Ours)	**89.20**	**86.03**	**63.85**	**70.41**
SAM	86.16	91.60	55.23	51.90
+LC(Ours)	**89.29**	**91.80**	**67.17**	**72.08**