Variation-Bounded Loss for Noise-Tolerant Learning

Jialiang Wang[1,2*], **Xiong Zhou**[1*], **Xianming Liu**[1], **Gangfeng Hu**[1]
Deming Zhai[1†], **Junjun Jiang**[1], **Haoliang Li**[2]

[1]Harbin Institute of Technology, China
[2]City University of Hong Kong, China

Abstract

Mitigating the negative impact of noisy labels has been a perennial issue in supervised learning. Robust loss functions have emerged as a prevalent solution to this problem. In this work, we introduce the *Variation Ratio* as a novel property related to the robustness of loss functions, and propose a new family of robust loss functions, termed *Variation-Bounded Loss* (VBL), which is characterized by a bounded variation ratio. We provide theoretical analyses of the variation ratio, proving that a smaller variation ratio would lead to better robustness. Furthermore, we reveal that the variation ratio provides a feasible method to relax the symmetric condition and offers a more concise path to achieve the asymmetric condition. Based on the variation ratio, we reformulate several commonly used loss functions into a variation-bounded form for practical applications. Positive experiments on various datasets exhibit the effectiveness and flexibility of our approach.

Code — https://github.com/cswjl/variation-bounded-loss

Introduction

In recent years, Deep Neural Networks (DNNs) have demonstrated outstanding performance in a wide range of fields (Dong, Wang, and Abbas 2021). However, achieving strong performance typically relies on large-scale, high-quality annotated datasets. In practice, real-world datasets often contain substantial label noise due to human error or carelessness (Song et al. 2022). DNNs may suffer a considerable decline in performance when they overfit to noisy labels (Zhang et al. 2017). As a result, ensuring robust generalization in the presence of noisy labels remains a major challenge. Among the many approaches proposed to address this issue, robust loss functions have remained a popular solution, owing to their simplicity and flexibility (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Ma et al. 2020; Zhou et al. 2023; Ye et al. 2025).

Cross Entropy (CE) is the standard loss function for classification tasks due to its strong fitting capability; however, it is also prone to overfitting noisy labels. Ghosh, Manwani, and Sastry (2015); Ghosh, Kumar, and Sastry (2017) proved that if a loss function satisfies the symmetric condition, it can achieve noise-tolerance. A classic example of a symmetric loss function is Mean Absolute Error (MAE). However, due to the strict symmetric condition, symmetric loss functions are usually difficult to optimize (Zhang and Sabuncu 2018; Ma et al. 2020; Zhou et al. 2021). Therefore, a popular way is to balance the fitting ability and robustness of the loss function by taking an intermediate value between the CE loss and the MAE loss, such as GCE (Zhang and Sabuncu 2018), SCE (Wang et al. 2019), and JS (Englesson and Azizpour 2021). However, this approach to improving fitting ability comes at the expense of robustness due to the retained CE property. More specifically, the absolute values of the gradients of these loss functions tend to approach infinity at low prediction probabilities. This character leads them to still pay too much attention to particularly low-confidence samples which are most likely noisy examples (Wei et al. 2023). As a result, they do not achieve full noise-tolerance and may overfit to a portion of the noisy labels. Beyond the symmetric condition (Ghosh, Kumar, and Sastry 2017), Zhou et al. (2021, 2023) proposed Asymmetric Loss Functions (ALFs), which focus on the loss term with maximum weight in the optimization. In this way, the impact of noisy components is mitigated, making the loss function inherently noise-tolerant. Although they proposed the asymmetric condition, they did not provide a straightforward method for its implementation. To date, no design guidelines have been established to help create simpler and more efficient asymmetric loss functions.

In this work, we introduce a new property of loss functions, called *variation ratio*, which involves the robustness of loss functions. Then, we propose a new family of robust loss functions, namely *Variation-Bounded Loss* (VBL), whose variation ratio is bounded. We perform rigorous theoretical analyses of the variation ratio. Firstly, from the perspective of the symmetric condition, we prove that a bounded variation ratio can achieve a relaxed symmetric condition. Moreover, we show that a smaller variation ratio leads to a tighter excess risk bound across various types of label noise. Secondly, from the perspective of the asymmetric condition, we build a path from the variation ratio to the asymmetric condition. We prove that if the variation ratio is below a certain threshold determined by the noise rate,

[*]These authors contributed equally.
[†]Corresponding author.

the variation-bounded loss becomes asymmetric and, consequently, completely noise-tolerant. This gives us a simpler and more efficient way to achieve the asymmetric condition. Our comprehensive analyses demonstrate that the variation ratio is critical for both symmetric and asymmetric conditions. This suggests that the variation ratio can serve as a valuable tool for designing more effective and robust loss functions. Furthermore, we modify several commonly used loss functions to a variation-bounded form for practical applications. The main contributions of our work are highlighted as follows:

- We introduce a novel property related to the robustness of loss functions, namely *variation ratio*, and propose a new family of robust loss functions, termed *Variation-Bounded Loss* (VBL), which have a bounded variation ratio.
- We provide comprehensive theoretical analyses of our variation-bounded loss, demonstrating that a small variation ratio is essential for achieving noise-tolerant learning.
- The concise variation ratio can serve as a valuable tool for designing more effective robust loss functions. We develop a series of practical variation-bounded losses. The results of extensive experiments underscore the superiority of our method.

Preliminary

Supervised Classification. For supervised classification tasks, we have a labeled dataset $\mathcal{S} = (\mathbf{x}_n, y_n)_{n=1}^N$ for training models. Each pair $(\mathbf{x}_n, y_n)$ is i.i.d. drawn from a joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the sample space, $\mathcal{Y} = [K] = \{1, 2, ..., K\}$ is the label space, and K is the number of classes. The classifier f is a model with a softmax layer, mapping inputs from the sample space to the probability simplex; thus, the predicted label is given by $\hat{y} = \arg\max_k f(\mathbf{x})_k$. We consider a loss function $L : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$, where $\arg\min_{\mathbf{u}} L(\mathbf{u}, \mathbf{e}_y) = \mathbf{e}_y$ and $\mathbf{e}_y$ is the one-hot vector corresponding to class y. The loss function $L(\mathbf{u}, \mathbf{e}_y)$ is monotonically decreasing on the prediction probability u_y of class y. For brevity, we abbreviate $L(\mathbf{u}, \mathbf{e}_k)$ as $L(\mathbf{u}, k)$ in this paper. Given a loss function $L \in \mathcal{L}$ and a classifier $f \in \mathcal{F}$, the expected risk (Bartlett, Jordan, and McAuliffe 2006) is defined as $\mathcal{R}_L(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[L(f(\mathbf{x}), y)]$. The objective of supervised learning is to find a classifier $f^* \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_L(f)$ that minimizes the expected risk.

Learning with Noisy Labels. In the learning with noisy labels scenario, the available training set $\tilde{\mathcal{S}} = \{(\mathbf{x}_n, \tilde{y}_n)\}_{n=1}^N$ is noisy rather than a clean set $\mathcal{S}$. For a sample $\mathbf{x}$, its true label y will be corrupted into a noisy label $\tilde{y}$ with a conditional probability $\eta_{\mathbf{x},\tilde{y}} = p(\tilde{y}|\mathbf{x}, y)$ (Natarajan et al. 2013). We define the noise rate for $\mathbf{x}$ as $\eta_{\mathbf{x}} = \sum_{k\neq y} \eta_{\mathbf{x},k}$. We mainly consider the following three common types of label noise (Xia et al. 2020; Chen et al. 2021; Ye et al. 2023):

- Instance-Dependent Noise: $\eta_{\mathbf{x},y} = 1 - \eta_{\mathbf{x}}$ and $\sum_{k\neq y} \eta_{\mathbf{x},k} = \eta_{\mathbf{x}}$, where noise rate $\eta_{\mathbf{x}}$ depends on the instance $\mathbf{x}$.
- Symmetric Noise: $\eta_{\mathbf{x},y} = 1 - \eta$ and $\eta_{\mathbf{x},k\neq y} = \frac{\eta}{K-1}$, where noise rate η is a constant.
- Asymmetric Noise: $\eta_{\mathbf{x},y} = 1 - \eta_y$ and $\sum_{k\neq y} \eta_{\mathbf{x},k} = \eta_y$, where noise rate η_y depends on the class y.

For this context, we only have the noisy dataset. Therefore, we focus on minimizing the noisy expected risk as follows:

$$\mathcal{R}_L^\eta(f) = \mathbb{E}_{\mathcal{D}}[(1 - \eta_{\mathbf{x}})L(f(\mathbf{x}), y) + \sum_{k\neq y} \eta_{\mathbf{x},k} L(f(\mathbf{x}), k)], \tag{1}$$

where $\sum_{k\neq y} \eta_{\mathbf{x},k} L(f(\mathbf{x}), k)$ is the noisy portion that usually damages the performance of DNNs. A loss function L is defined to be *noise-tolerant* (Manwani and Sastry 2013) if the noise minimizer $f_\eta^* \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_L^\eta(f)$ also minimizes the clean expected risk, i.e., $R_L(f_\eta) = R_L(f)$.

Variation-Bounded Loss

In this section, we provide comprehensive descriptions and theoretical analyses of our variation ratio and variation-bounded loss (VBL), demonstrating that our method achieves robust and efficient learning in a concise manner. Detailed proofs are included in the Appendix.

Definitions

First, we introduce the proposed variation ratio and variation-bounded loss.

Definition 1 (Variation Ratio)**.** *For a loss function $L(\mathbf{u}, y) = \ell(u_y)$, we define the variation ratio $v(L)$ as*

$$v(L) = \frac{\max_{u\in(0,1)} |\nabla \ell(u)|}{\min_{u\in(0,1)} |\nabla \ell(u)|}, \tag{2}$$

where $|\cdot|$ denotes the absolute value and $\nabla \ell = \frac{\partial \ell(u)}{\partial u}$ is the gradient of ℓ w.r.t. u.

Definition 2 (Variation-Bounded Loss)**.** *If the variation ratio $v(L) < \infty$, the loss function L is variation-bounded. Conversely, if $v(L) = \infty$, the loss function L is variation-unbounded.*

A bounded variation ratio ensures that the loss function does not descend excessively within an interval. We provide some examples of common loss functions to enhance understanding. The variation ratio $v(L)$ of Mean Absolute Error (MAE), $L_{\text{MAE}}(\mathbf{u}, y) = 2(1 - u_y)$, is 1, indicating that MAE is a variation-bounded loss. Another variation-bounded example is the Exponential Loss (EL), $L_{\text{EL}}(\mathbf{u}, y) = e^{-u_y}$, which has a variation ratio of e. In contrast, the variation ratio $v(L)$ of the Cross-Entropy (CE), $L_{\text{CE}}(\mathbf{u}, y) = -\log u_y$, is infinite, indicating that CE is variation-unbounded.

In the following, we will explore the detailed properties of the variation-bounded loss.

Symmetric Condition

Previous works (Ghosh, Manwani, and Sastry 2015; Ghosh, Kumar, and Sastry 2017) theoretically proved that symmetric loss functions are inherently robust to label noise under some mild conditions.

Definition 3 (Symmetric Condition)**.** *A loss function is symmetric if it satisfies*

$$\sum_{k=1}^{K} L(\mathbf{u}, k) = C \tag{3}$$

where C is a constant and $k \in [K]$ is the label corresponding to each class.

Because the symmetric condition in Definition 3 is overly strict, symmetric losses are challenging to optimize (Zhang and Sabuncu 2018; Zhou et al. 2021). A common method to address this issue is to interpolate between the symmetric MAE and the fast-converging CE (Zhang and Sabuncu 2018; Wang et al. 2019; Englesson and Azizpour 2021). Although these robust loss functions can mitigate label noise, we observe that, as they are derived through interpolation with CE, they inevitably inherit a certain property of CE. Specifically, when the predicted probability approaches 0, the absolute value of their gradient approaches ∞. Unfortunately, since noisy samples often have low confidence, this characteristic can result in overfitting to some noisy labels (Wei et al. 2023).

In this work, we propose a novel method to relax the symmetric condition instead of interpolating between the MAE and CE. This method avoids the inherent drawback of CE, which tends to overfit to label noise. Some studies (Ghosh, Kumar, and Sastry 2017; Li, Socher, and Hoi 2020; Wei et al. 2023) have indicated that bounded losses are more robust than unbounded losses. The bounded property of the loss, $C_L \leq L(\mathbf{u}, k) \leq C_U$, ensures that $|\sum_{k=1}^{K} L(\mathbf{u}, k) - \sum_{k=1}^{K} L(\mathbf{v}, k)| \leq K(C_U - C_L)$, where $\mathbf{u}$ and $\mathbf{v}$ are arbitrary vectors in the domain. This property brings the bounded loss closer to meeting the symmetric condition and enhances its robustness compared to unbounded losses such as CE. Here, we derive a more efficient bounded property based on the variation ratio.

Lemma 1. *For a loss function $L(\mathbf{u}, y) = c \cdot \ell(u_y)$, we have*

$$\left| \sum_{k=1}^{K} L(\mathbf{u}, k) - \sum_{k=1}^{K} L(\mathbf{v}, k) \right| \leq v(L) - 1. \tag{4}$$

where $c = \frac{1}{\min_u |\nabla \ell(u)|}$ is a normalization constant.

c in Lemma 1 is used to normalize the minimum absolute value of the gradient to 1, so that different loss functions can be compared on the same scale. Lemma 1 shows that the variation ratio constitutes a sufficient condition for the bounded property and a smaller $v(L)$ must result in better symmetry. Specifically, when $v(L)$ reaches its minimum value of 1, the loss is symmetric. And this loss is essentially a linear function, representing a scaled MAE.

Based on Lemma 1, we derive excess risk bounds (Bartlett, Jordan, and McAuliffe 2006) under various types of label noise. First, we prove the situation of symmetric noise.

Theorem 1 (Excess Risk Bound under Symmetric Noise)**.** *In a multi-class classification problem, if the loss function L satisfies $|\sum_{k=1}^{K} L(\mathbf{u}, k) - \sum_{k=1}^{K} L(\mathbf{v}, k)| \leq v(L) - 1$, then for symmetric noise satisfying $\eta < 1 - \frac{1}{K}$, the excess risk bound for f can be expressed as*

$$\mathcal{R}_L(f_\eta^*) - \mathcal{R}_L(f^*) \leq c(v(L) - 1), \tag{5}$$

where $c = \frac{\eta}{(1-\eta)K-1}$ is a constant, f_η^* and f^* denote the global minimum of $\mathcal{R}_L^\eta(f)$ and $\mathcal{R}_L(f)$, respectively.

Next, we address the more complex situations of asymmetric and instance-dependent noise.

Theorem 2 (Excess Risk Bound under Asymmetric and Instance-Dependent Noise)**.** *In a multi-class classification problem, if the loss function L satisfies $|\sum_{k=1}^{K} L(\mathbf{u}, k) - \sum_{k=1}^{K} L(\mathbf{v}, k)| \leq v(L) - 1$, then for label noise $1 - \eta_{\mathbf{x}} > \max_{k \neq y} \eta_{\mathbf{x},k}$, $\forall \mathbf{x}$, if $\mathcal{R}_L(f^*)$ is minimum, the excess risk bound for f can be expressed as*

$$\mathcal{R}_L(f_\eta^*) - \mathcal{R}_L(f^*) \leq (1 + \frac{c}{a})(v(L) - 1), \tag{6}$$

where $c = \mathbb{E}_{\mathcal{D}}(1 - \eta_{\mathbf{x}})$ and $a = \min_{\mathbf{x},k}(1 - \eta_{\mathbf{x}} - \eta_{\mathbf{x},k})$ are constants, f_η^* and f^* denote the global minimum of $\mathcal{R}_L^\eta(f)$ and $\mathcal{R}_L(f)$, respectively. For asymmetric noise, $\eta_{\mathbf{x}} = \eta_y$, and for instance-dependent noise, $\eta_{\mathbf{x}} = \eta_{\mathbf{x}}$.

Theorem 1 and 2 demonstrate that a smaller variation ratio $v(L)$ results in more robust to label noise. Additionally, the excess risk bound can be controlled by the $v(L)$.

Asymmetric Condition

Previous works (Zhou et al. 2021, 2023) proposed asymmetric loss functions, which are noise-tolerant to label noise. However, the proposed asymmetric loss functions are too complex with many hyperparameters, and easily produce the underfitting problem. In this subsection, we revisit the asymmetric condition through the variation ratio.

Definition 4 (Asymmetric Condition)**.** *On the given weights $w_1, \ldots, w_k \geq 0$, where $\exists t \in [K]$, s.t., $w_t > \max_{k \neq t} w_k$, a loss function $L(\mathbf{u}, k)$ is called asymmetric if L satisfies*

$$\arg\min_{\mathbf{u}} \sum_{k=1}^{K} w_k L(\mathbf{u}, k) = \arg\min_{\mathbf{u}} L(\mathbf{u}, t), \tag{7}$$

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

Asymmetric loss functions are noise-tolerant under clean-label-dominant noise, i.e., $1 - \eta_{\mathbf{x}} > \max_{k \neq y} \eta_{\mathbf{x},k}$, $\forall \mathbf{x}$ (Zhou et al. 2021). Here, we revisit the way to achieve the asymmetric condition and prove that when the variation ratio $v(L)$ is less than a specific constant related to the label distribution, the variation-bounded loss is asymmetric.

Theorem 3. *On the given weights $w_1, \ldots, w_k \geq 0$, where $\exists t \in [K]$ and $w_t > \max_{i \neq t} w_i$, a loss function $L(\mathbf{u}, k) = \ell(u_k)$ is asymmetric if (1) $\frac{\partial^2 \ell(u)}{\partial u^2} \leq 0$ or (2) $v(L) \leq \frac{w_t}{w_i}$ for any $i \neq t$.*

Condition (1) in Theorem 3 is not favorable for optimization. Specifically, when $\frac{\partial^2 \ell(u)}{\partial u^2} > 0$, we have a convex loss function, such as CE, which is generally favorable for optimization. When $\frac{\partial^2 \ell(u)}{\partial u^2} = 0$, we have a linear loss function,

such as MAE. When $\frac{\partial^2 \ell(u)}{\partial u^2} < 0$, we have a concave loss function, the loss function would be even harder to optimize than linear MAE. Therefore, in practice, concave loss functions are generally not considered. Instead, we primarily focus on condition (2) in Theorem 3.

For a variation-bounded loss described in Theorem 3, if it satisfies $v(L) \leq \frac{1-\eta_{\mathbf{x}}}{\max_{k \neq y} \eta_{\mathbf{x},k}}$, i.e., condition (2) in Theorem 3 for the context of learning with noisy labels, then the loss function is asymmetric. For instance, about a 10-classes dataset with 0.8 symmetric noise, if $v(L) \leq \frac{1-\eta_{\mathbf{x}}}{\max_{k \neq y} \eta_{\mathbf{x},k}} = \frac{0.2}{0.8/9} \approx 2.25$, then the loss function is asymmetric and therefore noise-tolerant. Notably, this constitutes a more relaxed condition compared to the symmetric condition, because it only requires that $v(L) \leq 2.25$, whereas symmetric MAE requires $v(L)$ to equal the minimum value of 1. Thus, variation-bounded losses have better fitting ability than symmetric losses, enabling them to achieve both complete robust and efficient learning simultaneously.

Variation-Bounded Loss

In this subsection, we concisely generalize several commonly used loss functions to a variation-bounded form. We use $\mathbf{u} = f(\mathbf{x})$ to denote the predicted probability after the softmax layer, and u_y is the predicted probability for the label.

Variation Cross Entropy (VCE):

$$L_{\text{VCE}} = -\log(u_y + a), \tag{8}$$

where $a \geq 0$ is a hyperparameter. VCE is modified from the CE loss. The gradient of VCE is $-\frac{1}{u_y+a}$. If $a > 0$, The variation ratio $v(L_{\text{VCE}}) = \frac{1+a}{a}$. If $a = 0$, the variation ratio $v(L_{\text{VCE}}) = \infty$, which recovers the CE loss.

Variation Exponential Loss (VEL):

$$L_{\text{VEL}} = a^{-u_y}, \tag{9}$$

where $a > 1$ is a hyperparameter. VEL is modified from the Exponential Loss (EL). The gradient of VEL is $-a^{-u_y} \log a$. The variation ratio $v(L_{\text{VEL}}) = a$, and if $a = e$, it recovers to the Exponential Loss (EL).

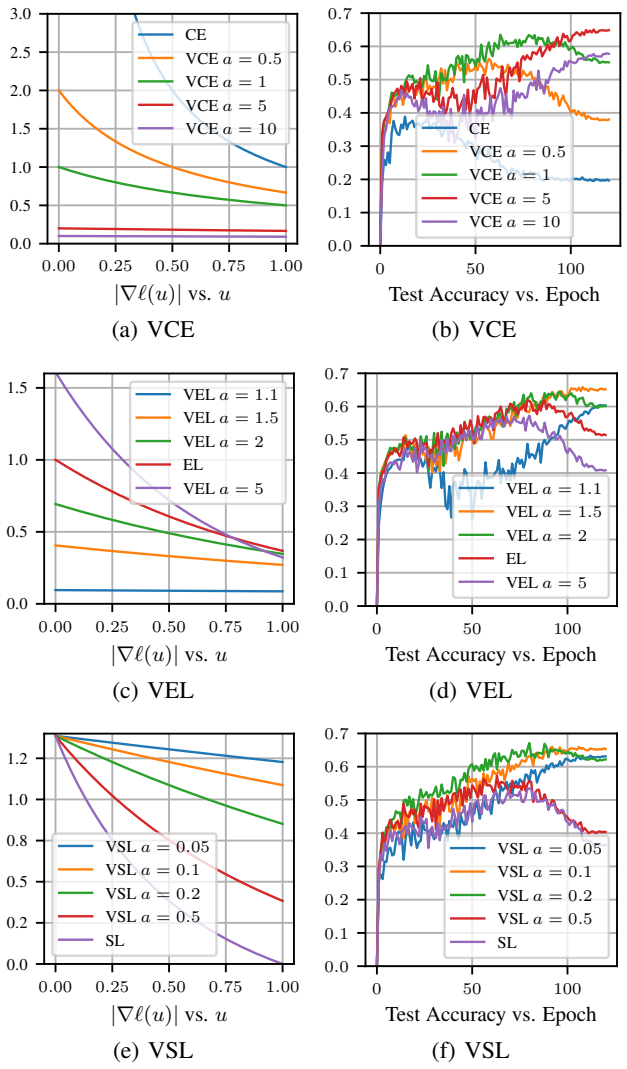
Variation Square Log (VSL):

$$L_{\text{VSL}} = [\log(a \cdot u_y + 1) - \log 2]^2 / a, \tag{10}$$

where $0 < a \leq 1$ is a hyperparameter. VSL is modified from the Square Log Loss. The gradient of VSL is $2[\log(a \cdot u_y + 1) - \log 2] \cdot \frac{1}{a \cdot u_y + 1}$. If $0 < a < 1$, the variation ratio $v(L_{\text{VSL}}) = \frac{(a+1) \cdot \log 2}{\log 2 - \log(a+1)}$. If $a = 1$, the variation ratio $v(L_{\text{VSL}}) = \infty$, which recovers the Square Log (SL).

More Analyses

Loss Function Visualization. To better analyze the properties of the variation-bounded loss, we visualize the absolute values of gradients and perform experiments on CIFAR-10 with 0.8 symmetric noise, as shown in Figure 1. Two common scenarios of variation-unbounded losses are observed. The first scenario, exemplified by CE, is shown in



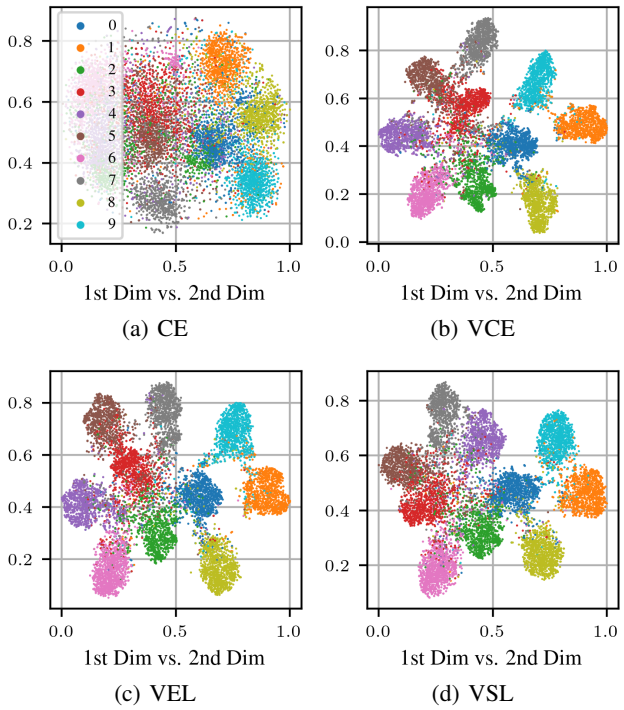
(a) VCE (b) VCE

(c) VEL (d) VEL

(e) VSL (f) VSL

Figure 1: **Left:** Absolute values of gradients, i.e., $|\nabla \ell|$. **Right:** Test accuracies on CIFAR-10 with 0.8 symmetric noise.

Figure 1(a). As depicted, its gradient approaches 1 for high-confidence samples and approaches ∞ for low-confidence samples. The second scenario, represented by the Square Log (SL), is shown in Figure 1(e). Here, the gradient decreases to 0 for high-confidence samples and approaches $2 \log 2$ for low-confidence samples. In both cases, the variation ratio becomes ∞. Intuitively, during optimization, the gradient contribution of low-confidence (noisy) samples becomes disproportionately large, while the contribution of high-confidence (clean) samples becomes too small. Consequently, as training progresses, loss functions like CE and SL overfit to some noisy labels, leading to a decrease in test accuracies, as shown in Figures 1(b) and 1(f). In contrast, our variation-bounded losses constrain the variation ratio, yielding a more balanced gradient trade-off between low and high-confidence samples. With a simple modification, VCE,



(a) CE (b) VCE (c) VEL (d) VSL

Figure 2: Visualizations of learned features on CIFAR-10 with 0.4 symmetric noise by t-SNE.

VEL, and VSL achieve significantly greater robustness compared to vanilla CE, EL, and SL.

Hyperparameter Analysis. For variation-bounded losses, as can be seen from test accuracies (Figure 1(b), 1(d) and 1(f)), a smaller variation ratio ($a \uparrow$ for VCE; $a \downarrow$ for VEL and VSL) can enhance robustness and achieve noise-tolerant learning. However, a too small variation ratio may reduce the fitting ability. Therefore, it is suggested to choose a moderate variation ratio to achieve both robust and efficient learning.

Feature Visualization. We further compare the robustness of variation-bounded losses and vanilla CE in learning representations. We train models on CIFAR-10 with 0.4 symmetric noise and extract the learned features from the test set using t-SNE (Van der Maaten and Hinton 2008), as shown in Figure 2. For the hyperparameter, we utilize VCE ($a = 5$), VEL ($a = 1.5$), and VSL ($a = 0.1$), refer to the experiment in Figure 1. As can be seen, embeddings generated by CE exhibit evident overfitting to label noise, as evidenced by the blending of embeddings from distinct classes. In contrast, embeddings generated by variation-bounded losses consistently form clear, well-separated clusters. This demonstrates their superior capability to learn robust and distinct representations under label noise.

Combination of NCE and VBL. Recently, the most advanced robust loss functions often combine their proposed methods with Normalized Cross Entropy (NCE) (Ma et al. 2020). Notable examples include Active Passive Loss (APL) (Ma et al. 2020), Asymmetric Loss Functions (ALFs) (Zhou et al. 2021), and Active Negative Loss (ANL) (Ye et al. 2023). The combination of two different robust loss functions can mutually enhance the optimization processes of each other, thus improving the overall fitting ability of the model (Ma et al. 2020). To obtain better performance and ensure a fair comparison with other combined methods, we also combine the proposed VBL with NCE. We simply formulate the combination of NCE and VBL as follows:

$$L_{\text{NCE+VBL}} = \alpha \cdot L_{\text{NCE}} + \beta \cdot L_{\text{VBL}} \tag{11}$$

Previous works (Zhou et al. 2021, 2023) have proved that the combination of the symmetric loss and the asymmetric loss remains asymmetric. Because NCE is symmetric, and our VBL is asymmetric. Thus, NCE+VBL is still asymmetric and, therefore, noise-tolerant.

CIFAR-10	Instance-Dependent Noise		
	20%	40%	60%
CE	75.22 ± 0.09	57.33 ± 0.16	37.84 ± 0.32
GCE	86.86 ± 0.22	82.80 ± 0.20	64.84 ± 1.04
SCE	86.72 ± 0.14	74.44 ± 0.39	51.15 ± 0.95
NCE+RCE	89.14 ± 0.15	85.08 ± 0.39	71.55 ± 0.52
NCE+AGCE	88.97 ± 0.18	84.89 ± 0.23	72.75 ± 0.34
LC	82.61 ± 0.23	67.82 ± 0.39	43.32 ± 0.99
NCE+NNCE	89.70 ± 0.21	85.76 ± 0.28	70.61 ± 1.00
OGC	86.71 ± 0.22	83.33 ± 0.29	64.73 ± 3.48
NCE+VCE	**89.77 ± 0.01**	**86.85 ± 0.23**	**73.95 ± 0.26**
NCE+VEL	**89.80 ± 0.20**	**86.93 ± 0.49**	**73.85 ± 0.40**
NCE+VSL	**89.84 ± 0.21**	**86.46 ± 0.36**	**75.60 ± 0.03**
CIFAR-100	Instance-Dependent Noise		
	20%	40%	60%
CE	56.86 ± 1.08	41.66 ± 0.33	24.47 ± 1.28
GCE	60.93 ± 0.92	56.81 ± 1.13	41.82 ± 0.62
SCE	55.70 ± 1.56	40.19 ± 0.43	23.04 ± 0.88
NCE+RCE	64.63 ± 0.44	56.68 ± 0.22	41.64 ± 0.77
NCE+AGCE	65.51 ± 0.24	58.40 ± 0.85	42.64 ± 0.11
LC	56.36 ± 0.25	37.68 ± 0.21	19.28 ± 0.50
NCE+NNCE	66.63 ± 0.53	61.14 ± 0.61	47.42 ± 0.62
OGC	64.36 ± 1.74	56.46 ± 0.44	40.04 ± 0.23
NCE+VCE	**69.33 ± 0.24**	**64.54 ± 0.46**	**54.04 ± 0.58**
NCE+VEL	**69.99 ± 0.31**	**65.31 ± 0.39**	**54.87 ± 0.43**
NCE+VSL	**69.97 ± 0.20**	**65.44 ± 0.26**	**54.43 ± 0.40**

Table 1: Last epoch test accuracies on instance-dependent noise. Top-3 best results are highlighted in **bold**.

Experiments

In this section, we provide extensive experiments to evaluate the effectiveness of our variation-bounded loss with various datasets, including benchmark datasets: CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009); real-world datasets: WebVision (Li et al. 2017), ILSVRC12 (Deng et al.

CIFAR-10	Clean	Symmetric Noise				Asymmetric Noise		Human
	0%	20%	40%	60%	80%	20%	40%	Worst
CE	90.47 ± 0.22	75.25 ± 0.43	58.51 ± 0.52	39.21 ± 0.38	19.04 ± 0.23	83.04 ± 0.17	73.70 ± 0.19	62.04 ± 0.64
GCE	89.08 ± 0.21	87.04 ± 0.33	83.68 ± 0.17	76.30 ± 0.14	42.69 ± 0.24	86.71 ± 0.15	69.12 ± 0.78	77.80 ± 0.38
SCE	91.48 ± 0.16	87.62 ± 0.26	79.37 ± 0.40	61.38 ± 0.84	27.75 ± 0.55	86.01 ± 0.21	74.17 ± 0.41	73.70 ± 0.06
NCE+RCE	91.20 ± 0.15	89.17 ± 0.15	85.75 ± 0.24	79.93 ± 0.30	54.04 ± 2.59	88.30 ± 0.13	77.78 ± 0.19	80.09 ± 0.21
NCE+AGCE	91.05 ± 0.28	89.12 ± 0.24	86.19 ± 0.19	80.28 ± 0.35	45.28 ± 4.09	88.62 ± 0.09	78.50 ± 0.37	80.03 ± 0.47
LC	90.03 ± 0.15	83.47 ± 0.53	70.27 ± 0.42	46.61 ± 0.90	19.88 ± 0.83	83.31 ± 0.44	73.38 ± 0.32	70.07 ± 0.22
NCE+NNCE	**91.71 ± 0.32**	**90.01 ± 0.06**	87.18 ± 0.42	81.05 ± 0.36	61.31 ± 2.62	88.83 ± 0.29	77.97 ± 0.12	80.52 ± 0.24
OGC	88.86 ± 0.12	87.17 ± 0.22	83.82 ± 0.19	77.37 ± 0.26	48.09 ± 0.60	86.83 ± 0.05	66.41 ± 3.20	76.31 ± 3.59
NCE+VCE	**91.75 ± 0.26**	**90.03 ± 0.19**	**87.73 ± 0.40**	**82.33 ± 0.51**	**64.49 ± 0.99**	**89.72 ± 0.19**	**79.16 ± 0.44**	**81.28 ± 0.19**
NCE+VEL	91.60 ± 0.27	**90.24 ± 0.43**	**87.35 ± 0.09**	**82.42 ± 0.19**	**64.29 ± 1.81**	**89.77 ± 0.28**	**80.20 ± 0.39**	**81.38 ± 0.26**
NCE+VSL	**91.75 ± 0.27**	89.97 ± 0.28	**87.31 ± 0.04**	**82.00 ± 0.38**	**62.96 ± 1.06**	**89.33 ± 0.45**	**79.23 ± 0.22**	**81.08 ± 0.40**
CIFAR-100	Clean	Symmetric Noise				Asymmetric Noise		Human
	0%	20%	40%	60%	80%	20%	40%	Noisy
CE	71.00 ± 1.21	55.73 ± 0.73	38.03 ± 2.49	23.34 ± 0.95	8.02 ± 0.22	58.03 ± 0.49	41.53 ± 0.50	49.25 ± 0.38
GCE	64.07 ± 0.83	62.02 ± 1.88	58.03 ± 1.50	46.29 ± 0.43	19.76 ± 0.82	58.57 ± 1.28	41.94 ± 0.72	50.13 ± 0.57
SCE	70.36 ± 0.48	55.47 ± 1.14	40.04 ± 0.25	22.81 ± 0.47	8.00 ± 0.18	57.40 ± 0.87	41.32 ± 0.64	48.51 ± 0.07
NCE+RCE	68.54 ± 0.11	64.63 ± 0.70	58.32 ± 0.34	46.40 ± 1.25	25.57 ± 0.28	63.06 ± 0.13	42.29 ± 0.12	54.48 ± 0.56
NCE+AGCE	68.95 ± 0.27	65.32 ± 0.29	59.40 ± 0.83	47.97 ± 0.45	24.96 ± 0.42	64.21 ± 0.50	44.95 ± 0.36	55.73 ± 0.17
LC	71.04 ± 0.25	57.37 ± 0.30	37.51 ± 0.66	17.39 ± 0.52	6.84 ± 0.18	56.17 ± 0.42	39.40 ± 0.18	48.15 ± 0.31
NCE+NNCE	70.27 ± 0.28	67.07 ± 0.42	61.74 ± 0.20	51.50 ± 0.88	28.09 ± 0.60	66.01 ± 0.25	45.92 ± 0.26	56.39 ± 0.11
OGC	67.99 ± 1.04	63.41 ± 1.96	57.24 ± 0.60	44.71 ± 3.10	14.55 ± 0.73	62.90 ± 1.11	37.56 ± 0.64	53.28 ± 0.57
NCE+VCE	**72.48 ± 0.43**	**69.32 ± 0.30**	**64.79 ± 0.46**	**57.55 ± 0.37**	**30.19 ± 0.30**	**68.94 ± 0.19**	**51.19 ± 0.33**	**57.44 ± 0.60**
NCE+VEL	**73.17 ± 0.33**	**69.97 ± 0.28**	**65.43 ± 0.49**	**58.09 ± 0.54**	**31.41 ± 0.95**	**68.99 ± 0.24**	**48.27 ± 0.40**	**58.31 ± 0.17**
NCE+VSL	**72.48 ± 0.15**	**70.30 ± 0.29**	**65.54 ± 0.10**	**57.57 ± 0.26**	**31.37 ± 1.20**	**69.17 ± 0.03**	**46.58 ± 0.41**	**58.62 ± 0.20**

Table 2: Last epoch test accuracies on symmetric, asymmetric, and human noise. Top-3 best results are highlighted in **bold**.

Method	CIFAR-10 Symmetric Noise				CIFAR-100 Symmetric Noise			
	20%	40%	60%	80%	20%	40%	60%	80%
CE	75.25 ± 0.43	58.51 ± 0.52	39.21 ± 0.38	19.04 ± 0.23	55.73 ± 0.73	38.03 ± 2.49	23.34 ± 0.95	8.02 ± 0.22
NCE	73.22 ± 0.35	69.37 ± 0.22	62.47 ± 0.85	41.20 ± 1.25	25.43 ± 0.91	20.26 ± 0.25	14.66 ± 1.04	8.82 ± 0.47
VCE	**90.44 ± 0.51**	87.29 ± 0.22	82.28 ± 0.29	63.77 ± 2.13	65.42 ± 0.81	60.39 ± 0.84	49.72 ± 1.26	**33.01 ± 0.83**
NCE+VCE	90.03 ± 0.19	**87.73 ± 0.40**	**82.33 ± 0.51**	**64.49 ± 0.99**	**69.32 ± 0.30**	**64.79 ± 0.46**	**57.55 ± 0.37**	30.19 ± 0.30

Table 3: Last epoch test accuracies of ablation experiment on symmetric noise. Best results are highlighted in **bold**.

2009), and Clothing1M (Xiao et al. 2015). Detailed experiment settings are included in the Appendix.

Benchmark Noisy Datasets

Baselines. We experiment with various state-of-the-art methods, including (1) Cross Entropy (CE); (2) Generalized Cross Entropy (GCE) (Zhang and Sabuncu 2018); (3) Symmetric Cross Entropy (SCE) (Wang et al. 2019); (4) Active Passive Loss (APL) (Ma et al. 2020), including NCE+RCE; (5) Asymmetric Loss Functions (ALFs) (Zhou et al. 2021, 2023), including NCE+AGCE; (6) LogitClip (LC) (Wei et al. 2023); (7) Active Negative Loss (ANL) (Ye et al. 2023), including NCE+NNCE; (8) Optimized Gradient Clipping (OGC) (Ye et al. 2025). To avoid confusion and improve performance, we use the NCE+VBL on benchmark datasets. We follow the same setting as in (Ma et al. 2020; Zhou et al. 2021; Ye et al. 2023), training an 8-layer CNN (LeCun et al. 1989) for 120 epochs on CIFAR-10 and a ResNet-34 (He et al. 2016) for 200 epochs on CIFAR-100, respectively. The experiment results are reported as "mean ± std" over 3 independent runs.

Results. Table 1 and 2 showcase the test accuracies of different methods under various types of label noise, including instance-dependent symmetric, asymmetric, and human-annotated (Wei et al. 2021) noise. Notably, our introduced variation-bounded losses, NCE+VCE, NCE+VEL, and NCE+VSL, exhibit exceptional performance, consistently ranking among the top-3 in most noise types. In

Method	CIFAR-10N					CIFAR-100N
	Aggregate	Random 1	Random 2	Random 3	Worst	Noisy
DivideMix	95.01 ± 0.71	95.16 ± 0.19	95.23 ± 0.07	95.21 ± 0.14	92.56 ± 0.42	71.13 ± 0.48
DivideMix+VCE	$\mathbf{95.97 \pm 0.14}$	$\mathbf{96.24 \pm 0.15}$	$\mathbf{96.07 \pm 0.12}$	$\mathbf{96.00 \pm 0.19}$	$\mathbf{93.70 \pm 0.29}$	$\mathbf{71.92 \pm 0.36}$
Negative-LS	$\mathbf{91.97 \pm 0.46}$	90.29 ± 0.32	$\mathbf{90.37 \pm 0.12}$	90.13 ± 0.19	82.99 ± 0.36	58.59 ± 0.98
Negative-LS+VCE	91.72 ± 0.16	$\mathbf{90.75 \pm 0.22}$	90.30 ± 0.27	$\mathbf{90.34 \pm 0.26}$	$\mathbf{84.17 \pm 0.42}$	$\mathbf{61.93 \pm 0.22}$

Table 4: Best epoch test accuracies on CIFAR-N. Better results are highlighted in **bold**.

Method	CE	GCE	SCE	NCE+RCE	NCE+AGCE	NCE+NNCE	NFL+NNFL	**VCE**	**NCE+VCE**
WebVision	61.2	59.44	68	64.92	63.92	67.44	68.32	$\mathbf{69.69 \pm 0.31}$	$\mathbf{69.00 \pm 0.60}$
ILSVRC12	58.64	56.56	62.6	62.4	60.76	65	65.56	$\mathbf{66.05 \pm 0.25}$	$\mathbf{65.85 \pm 0.20}$
Clothing1M	68.07	68.94	-	69.07	-	69.93	-	$\mathbf{69.96 \pm 0.30}$	$\mathbf{70.05 \pm 0.36}$

Table 5: Last epoch test accuracies on WebVision, ILSVRC12, and Clothing1M. Top-2 best results are highlighted in **bold**. Baseline results are obtained from (Ye et al. 2023) with the same setting.

particular, under most difficult 0.6 instance-dependent, 0.8 symmetric, 0.4 asymmetric, and human-annotated noise, our method improves accuracy by 1%~6% over previous state-of-the-art methods. Furthermore, for clean labels, our variation-bounded losses consistently demonstrate superior fitting ability. For instance, on CIFAR-100 clean labels, variation-bounded losses achieve accuracies of 72%~73%, whereas other loss functions reach accuracies around 70%. These results show that our method achieves superior performance compared to latest advance methods.

Ablation Experiments. To further investigate the effect of using variation-bounded loss alone on benchmark datasets, we conducted additional experiments on symmetric noise, as shown in Table 3. As can be seen, a simple modification yields a significant improvement in VCE over vanilla CE. In some cases, VCE outperforms the combination of NCE + VCE. Overall, NCE + VCE exhibit a better performance.

Combination of VBL and Other Methods. Because of its simplicity and convenience, our variation-bounded loss can be easily integrated with other methods by replacing the original CE loss, without any additional overhead. We combine our VCE loss with two other approaches: (1) DivideMix (Li, Socher, and Hoi 2020), a classic method based on semi-supervised learning, and (2) Negative-LS (Wei et al. 2022), a label smoothing technique for mitigating label noise. We conduct experiments on the CIFAR-N datasets, following the same settings as in the original paper (Wei et al. 2021), as shown in Table 4. As can be seen, our method consistently improves the performance of DivideMix across all scenarios, and yields significant gains for Negative-LS in the most challenging CIFAR-10 worst and CIFAR-100 noisy cases. These positive results demonstrate that our variation-bounded loss can readily enhance a variety of existing methods, highlighting the broad applicability of our approach.

Real-World Noisy Datasets

In this subsection, we conduct extensive experiments on massive real-world datasets, including WebVision (Li et al. 2017), ILSVRC12 (ImageNet) (Deng et al. 2009), and Clothing1M (Xiao et al. 2015), following the same setting as in (Ma et al. 2020; Ye et al. 2023). For WebVision, we adopt the "Mini" setting from (Jiang et al. 2018), which utilizes the first 50 classes of the Google subset. We train a ResNet-50 model (He et al. 2016) and evaluate it on the same 50 classes on both the ILSVRC12 and WebVision validation sets. For Clothing1M, we use a ResNet-50 pre-trained on ImageNet. We train it on the noisy dataset of 1 million samples and evaluate it on the clean test set.

Results. Table 5 reports the performances on WebVision, ILSVRC12, and Clothing1M. As shown, our variation-bounded losses, VCE and NCE+VCE, achieve the top-2 best accuracies, surpassing previous advanced methods such as NCE+RCE, NCE+AGCE, and NCE+NNCE. These results demonstrate the effectiveness of our approach in real-world scenarios.

Conclusion

This paper introduces the *Variation Ratio* property of loss functions and proposes a new category of robust loss functions known as *Variation-Bounded Loss* (VBL). We demonstrate that a smaller variation ratio represents better robustness. Moreover, we reveal that variation-bounded losses have a better way to relax the symmetric condition and more straightforwardly achieve the asymmetric condition. Our concise robust loss functions have shown positive results in mitigating label noise across diverse noise types. We believe that these loss functions can be widely applied in scenarios where it is difficult to obtain precise annotations. Additionally, we anticipate that the variation ratio will serve as a valuable tool for designing more effective robust loss functions.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China under Grant 2023YFC2509100, in part by National Natural Science Foundation of China under Grants 62525107 and 632B2031, in part by DON_RMG 9229106 and 9229161, and in part by Fundamental Research Funds for the Central Universities HIT.DZJJ.2025055.

References

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473): 138–156.

Chen, P.; Ye, J.; Chen, G.; Zhao, J.; and Heng, P.-A. 2021. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11442–11450.

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

Dong, S.; Wang, P.; and Abbas, K. 2021. A survey on deep learning and its applications. *Computer Science Review*, 40: 100379.

Englesson, E.; and Azizpour, H. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34: 30284–30297.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomputing*, 160: 93–107.

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

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

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

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*.

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

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

Manwani, N.; and Sastry, P. 2013. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3): 1146–1151.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11): 8135–8153.

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

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 322–330.

Wei, H.; Zhuang, H.; Xie, R.; Feng, L.; Niu, G.; An, B.; and Li, Y. 2023. Mitigating memorization of noisy labels by clipping the model prediction. In *International Conference on Machine Learning*, 36868–36886. PMLR.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; Sugiyama, M.; and Liu, Y. 2022. To Smooth or Not? When Label Smoothing Meets Noisy Labels. In *International Conference on Machine Learning*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2021. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *International Conference on Learning Representations*.

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

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Ye, X.; Li, X.; Dai, S.; Liu, T.; Sun, Y.; and Tong, W. 2023. Active Negative Loss Functions for Learning with Noisy Labels. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ye, X.; Wu, Y.; Zhang, W.; Li, X.; Chen, Y.; and Jin, C. 2025. Optimized Gradient Clipping for Noisy Label Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9463–9471.

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

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhou, X.; Liu, X.; Jiang, J.; Gao, X.; and Ji, X. 2021. Asymmetric loss functions for learning with noisy labels. In *International conference on machine learning*, 12846–12856. PMLR.

Zhou, X.; Liu, X.; Zhai, D.; Jiang, J.; and Ji, X. 2023. Asymmetric loss functions for noise-tolerant learning: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.