Machine Learning
https://doi.org/10.1007/s10994-023-06437-3

Nrat: towards adversarial training with inherent label noise

Zhen Chen[1] · **Fu Wang**[2] · **Ronghui Mu**[3] · **Peipei Xu**[1] · **Xiaowei Huang**[1] · **Wenjie Ruan**[1]

Received: 2 June 2023 / Revised: 21 August 2023 / Accepted: 7 October 2023

Abstract
Adversarial training (AT) has been widely recognized as the most effective defense approach against adversarial attacks on deep neural networks and it is formulated as a min-max optimization. Most AT algorithms are geared towards research-oriented datasets such as MNIST, CIFAR10, etc., where the labels are generally correct. However, noisy labels, e.g., mislabelling, are inevitable in real-world datasets. In this paper, we investigate AT with inherent label noise, where the training dataset itself contains mislabeled samples. We first empirically show that the performance of AT typically degrades as the label noise rate increases. Then, we propose a *Noisy-Robust Adversarial Training* (NRAT) algorithm, which leverages the recent advancements in learning with noisy labels to enhance the performance of AT in the presence of label noise. For experimental comparison, we consider two essential metrics in AT: (i) trade-off between natural and robust accuracy; (ii) robust overfitting. Our experiments show that NRAT's performance is on par with, or better than, the state-of-the-art AT methods on both evaluation metrics. Our code is publicly available at: https://github.com/TrustAI/NRAT.

Keywords Adversarial training · Robust loss functions · Noisy labels

Editors: Vu Nguyen, Dani Yogatama.

✉ Wenjie Ruan
w.ruan@trustai.uk

Zhen Chen
cz97@liverpool.ac.uk

Fu Wang
fw377@exeter.ac.uk

Ronghui Mu
ronghui.mu@lancaster.ac.uk

Peipei Xu
peipei.xu@liverpool.ac.uk

Xiaowei Huang
xiaowei.huang@liverpool.ac.uk

[1] Department of Computer Science, Liverpool University, Liverpool, UK

[2] Department of Computer Science, University of Exeter, Exeter, UK

[3] Department of Computer Science, Lancaster University, Lancaster, UK

Published online: 10 January 2024

1 Introduction

Deep neural networks have achieved considerable success in many fields (He et al. 2016; Devlin et al. 2018; Mnih et al. 2015), such as computer vision, natural language processes, and reinforcement learning, which have emerged as a transformative forces due to their remarkable efficacy and broad applicability. However, these powerful models are vulnerable to imperceptible perturbation (Goodfellow et al. 2014), i.e., adversarial examples (AEs). An AE, denoted as x', can be crafted by adding an adversarial perturbation $\boldsymbol{\delta}$ to a natural example x, *i.e.*, $x' = x + \boldsymbol{\delta}$. Adversarial attacks typically cause the classifier $h_{\boldsymbol{\theta}}$ to make an incorrect prediction. Such a perturbation $\boldsymbol{\delta}$ is often small and imperceptible to human perception, bounded by a L_p - norm ball, that can be written as $\|x' - x\|_p \leq \epsilon$.

AEs were first introduced by Szegedy et al. (2013), which enables the community to be aware of the vulnerability of neural networks and inspires the development of adversarial defenses, including defensive distillation (Papernot et al. 2016), feature squeezing (Xu et al. 2017), and adversarial training (AT) (Goodfellow et al. 2014), etc. Among them, AT has been regarded as the most powerful one (Athalye et al. 2018). The basic idea of AT is to incorporate both natural examples and AEs during the training stage, enabling models to perform better against AEs compared to standard training (ST). Formally, AT can be formulated as a min-max problem, i.e.,

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} L\big(h_{\boldsymbol{\theta}}(\boldsymbol{X}+\boldsymbol{\delta}), y\big) \right], \tag{1}$$

where the inner maximization searches for perturbations that maximize the loss, while the outer minimization optimizes the neural network. A multi-step gradient-based attack known as the PGD attack was proposed by Madry et al. (2017) to solve the inner maximization of AT, which can significantly improve the adversarial robustness of neural networks against various attacks, it has been deemed as the standard and baseline method of AT, referred to as PGD-AT in this paper. Based on their idea, researchers have proposed various variations, such as TRADES (Zhang et al. 2019), MART (Wang et al. 2019), AWP (Wu et al. 2020), and S^2O (Jin et al. 2022). Their ideas are basically based on three directions: objective functions, data augmentation, and weight perturbation. However, most of these methods have not taken into account the presence of noisy labels, whereas real-world datasets are reported to have an inherent noise label rate between 8 to 38.5% (Xiao et al. 2015).

Therefore, designing an effective AT algorithm in the presence of inherent label noise is a nontrivial research challenge, yet this challenge is under-explored by the community. While there is some literature discusses the relationship between AT and noisy labels (Zhu et al. 2021; Zhang et al. 2021; Dong et al. 2021), they are more concerned with *how to strengthen AT through noisy labels while do not consider AT with inherent label noise*, i.e., *noisy labels already exist in the original dataset*, this paper aims to make the first attempt to tackle this research challenge. There are two important metrics for evaluating the performance of AT: (1) natural-robust trade-off: the trade-off between natural accuracy and robust accuracy; (2) the extent of robust overfitting, i.e., the robust accuracy decreases after a certain training epoch, while the natural accuracy for natural examples remains increasing or relatively constant, i.e., robust overfitting thereby hindering the natural-robust trade-off.

First, we empirically evaluate the performance of three recent AT methods on CIFAR-10 with injected inherent noisy labels. We observe that both natural accuracy and robust accuracy decrease significantly with increasing noise rate, across all AT methods. Furthermore, in the presence of inherent label noise, we notice that the natural accuracy exhibits a decline from a specific training epoch, i.e., natural overfitting, in addition to the already observed robust overfitting. This phenomenon, we call "double overfitting" in this paper. Conversely, when there is no inherent label noise, the natural accuracy consistently improves or remains stable throughout the training process. The Cross-Entropy (CE) Loss, although widely used in the aforementioned mainstream AT methods, has been shown to be non-robust to label noise (Feng et al. 2021). This vulnerability may lead to degrades in their generalization performance. To address this issue, we propose incorporating noisy-robust loss functions in AT to enhance generalizability in the presence of label noise. The performance of these methods is shown in Fig. 1.

To accurately assess the true performance of a model trained with inherent label noise, the common practice is to train the model on a training dataset that contains noisy labels and then evaluate its performance on a clean test dataset without any noisy labels. This ensures that the model's performance is correctly evaluated. If the test dataset also contains a proportionate amount of noisy labels, it would not be possible to gauge the model's true performance, as the label noise in the test dataset would confound the evaluation and not accurately reflect its actual effectiveness. The overview figure of AT with inherent label noise is shown in Fig. 2, which can be seen as a general framework in our setting, including the noisy training set, adversarial examples generation, classifier, and prediction on the clean test dataset. Our main contributions are as follows:

• We investigate AT with inherent label noise and observe that it typically be unstable and prone to show poor generalization performance. Furthermore, we empirically identify the occurrence of the "double overfitting" phenomenon, where both the natural accuracy of natural examples and the robust accuracy of AEs start to decline after a certain training stage;

• From the perspective of objective functions for AT with inherent label noise, we replace the non-robust CE loss with a noisy-robust loss function and further propose *Noisy-Robust Adversarial Training* (NRAT);

• Theoretically and empirically, we demonstrate that NRAT achieves compatible performance or outperforms recent AT methods when dealing with inherent label noise.

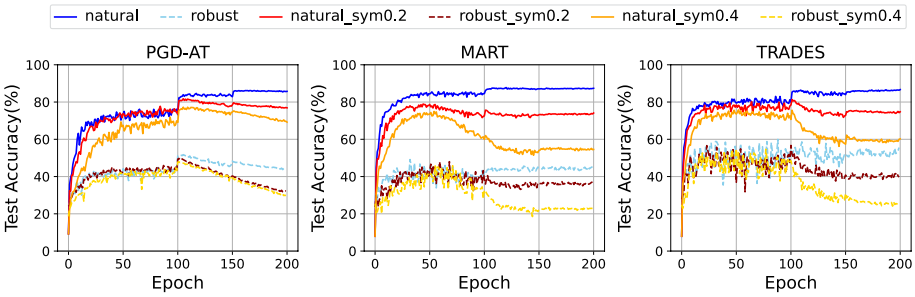

Fig. 1 The learning curves of natural accuracy and PGD robust accuracy for PGD-AT, MART, and TRADES under 0% (natural/robust), 20% and 40% inherent symmetric label noise on CIFAR-10 with ℓ_∞ threat model

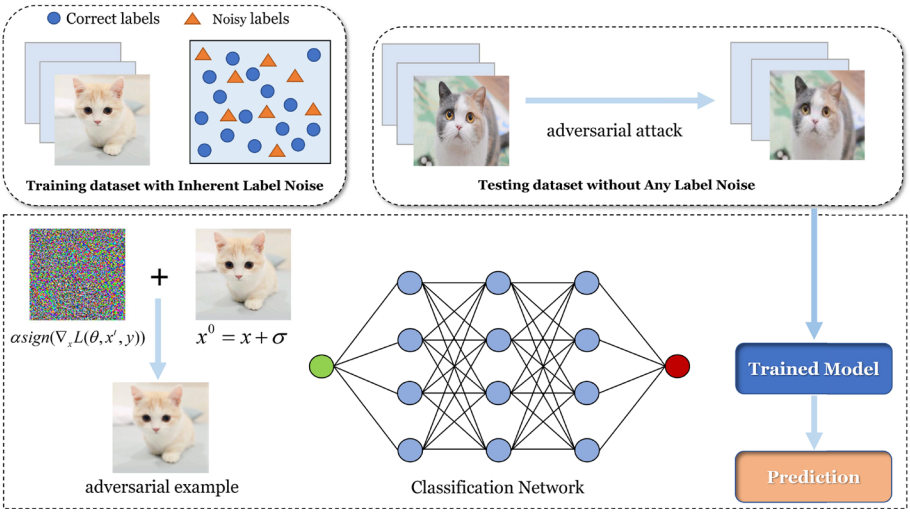

Fig. 2 Overview of AT with inherent label noise

2 Related works

2.1 Adversarial training algorithms

This section introduces three widely recognized AT algorithms, which are PGD-AT (Madry et al. 2017), TRADES (Zhang et al. 2019), and MART (Wang et al. 2019) (We use h_θ to denote the classifier with model parameter $\boldsymbol{\theta}$). First, we elaborate on their objective functions employed during the training process. Subsequently, we present the objective function of our NRAT and conduct a comparative analysis.

PGD-AT The idea of PGD-AT is straightforward, as it first generates AEs and then directly optimizes them. Despite its intuitive and simplistic formulation, it has been empirically shown to achieve excellent performance in terms of adversarial robustness. The training loss function is given by

$$\ell^{PGD-AT}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = \mathrm{CE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i\big). \tag{2}$$

TRADES TRADES aimed to trade off natural accuracy and robust accuracy by employing the CE loss for natural examples and incorporating a KL-divergence as the regularization term for adversarial examples. Its objective function can be written as

$$\ell^{TRADES}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \mathrm{CE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\big) + \beta \cdot \max_{\mathbf{x}_i' \in \mathcal{S}} \mathrm{KL}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i) \| h_{\boldsymbol{\theta}}(\mathbf{x}_i')\big) \right\}, \tag{3}$$

where the first term aims to maximize natural accuracy, and the second term aims to improve robust accuracy by minimizing the distance between the predictions of natural examples and adversarial examples, thereby encouraging the outputs to be smooth. The hyperparameter β controls the trade-off between natural accuracy and robust accuracy.

MART The fundamental idea behind MART is to treat misclassified and correctly classified examples as distinct instances and assign different optimization directions for them. The training loss can be formulated as follows

$$\begin{aligned}\ell^{MART}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) &= \mathrm{BCE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i\big) \\ &+ \lambda \cdot \mathrm{KL}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i) \| h_{\boldsymbol{\theta}}(\mathbf{x}_i')\big) \cdot \big(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i)\big),\end{aligned} \tag{4}$$

BCE is the boosted cross-entropy that can be written as

$$\mathrm{BCE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i\big) = \mathrm{CE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i\big) - \log\left(1 - \max_{k \neq y_i}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i\big)\right), \tag{5}$$

where k is the predicted class of $h_{\boldsymbol{\theta}}$. Empirically, BCE can mitigate insufficient learning of CE to some extent. The hyperparameter λ balances the influence of misclassified and correctly classified examples.

2.2 Interactions between adversarial training and noisy labels

Noisy labels are unavoidable in real-world datasets due to errors in manual annotation or in annotation platforms (Xiao et al. 2015). Accordingly, research on learning with noisy labels has also emerged. While the analysis of noisy labels in ST has been extensively explored (Li et al. 2017; Natarajan et al. 2013), researchers have recently started investigating the relationship between noisy labels and AT, i.e., AT with noisy labels. Zhu et al. (2021) explored the distinctions between AT and ST in the presence of noisy labels from the perspective of the smoothing effects of AT and the loss landscape. Zhang et al. (2021) proposed NoiLIn, which gradually injects noisy labels in both the inner maximization and outer minimization stages of AT to improve adversarial robustness. Dong et al. (2021) studied AT under random labels (almost 100% noisy labels) and identified that the remembering of one-hot labels as the cause of robust overfitting. They then adopted the Temporal Ensemble to mitigate this overfitting. Basically, these works explore the properties of noisy labels in AT and how to enhance AT's performance on clean datasets by introducing noisy labels. Inherent label noise, instead refers to the situation where the dataset itself already contains noisy labels.

2.3 Robust loss functions and learning with noisy labels

It has been demonstrated that a trained DNN with a suitable adjusted loss function $\mathcal{L}$, namely, a noisy-robust loss function (referred to as robust loss function hereafter), can approach the best classifier $h_{\boldsymbol{\theta}}$ under some mild assumptions with symmetric and asymmetric label noise (Ghosh et al. 2017). These robust loss functions satisfy the following equation (for a K-class classification problem, $K>1$ and any training example x)

$$\sum_{j=1}^{K} \mathcal{L}(h(\boldsymbol{x}), j) = C, \tag{6}$$

where C is a constant. Equation(6) indicates that these loss functions are symmetric and considered to be noise-tolerant according to the definitions in Ghosh et al. (2017). Even though there are many loss functions that satisfy this symmetry, the most commonly used CE does not possess this symmetry. Recent studies (Zhang and Sabuncu 2018; Amid et al. 2019) have demonstrated that adopting robust loss functions is the most straightforward and generic approach for effectively training deep neural networks with inherent label noise. Specifically, Ma et al. (2020) proposed the robust loss function NCE +RCE following an active-passive loss (APL) framework, which currently achieves state-of-the-art

performance on ST. For a K-class classification task, NCE (Ma et al. 2020) represents the normalized version of CE, and RCE (Wang et al. 2019) is the reversed version of CE. We will provide a more detailed explanation of both NCE and RCE in the next section.

In addition to the noisy-robust loss function, there are several other methods available for learning under label noise, including label correction (Zheng et al. 2021) and collaborative learning (Han et al. 2018), etc. However, these methods often involve too complicated procedures, making their application in AT quite challenging. Considering that AT already requires significant computational resources and may incur additional performance costs. Therefore, in this paper, we specifically focus on implementing the core concept of AT under inherent label noise using a robust loss function.

3 Noisy-robust adversarial training

In the previous sections, we discussed robust loss functions which exhibit the symmetric property. Based on this, we present a novel perspective on AT with inherent label noise, i.e., replacing the non-robust loss function with a robust counterpart to enhance the performance of AT in the presence of inherent label noise. Finally, we will conduct a comprehensive comparison between our proposed NRAT with existing approaches.

3.1 Basic notation of AT with inherent label noise

For a K-class classification task, let $X = \{(x_i, y_i)\}_{i=1,...n}$ be the training dataset drawn from an input distribution $\mathcal{D}$ with n training instances, where $x_i \in \mathbb{R}^d$ represents a natural example and $y_i \in \{1, ..., K\}$ denotes its annotated label, which may be incorrect, therefore we denote y_i^* as the true label for x_i. We use $\boldsymbol{q}(k \mid \boldsymbol{x})$ to represent the distribution of sample x of label $k \in K$ and $\sum_{k=1}^{K} \boldsymbol{q}(k \mid \boldsymbol{x}) = 1$. We consider two types of label noise: *symmetric* and *asymmetric* label noise with an overall noise rate $\eta \in [0, 1]$. For each class j flipped to k, we denote its class-wise noise rate by η_{jk}. Symmetric label noise, which means that each label has the same probability of flipping to any other class, i.e., $\eta_{jk} = \frac{\eta}{K-1}, j \neq k$; While asymmetric noise refers to labels being flipped between similar classes, e.g., the class "truck" being flipped to "car".

Given a classifier $h_{\boldsymbol{\theta}}$ with model parameter $\boldsymbol{\theta}$ (For simplicity, we may omit the $\boldsymbol{\theta}$ in the subsequent content), it predicts the class of an input example as

$$h_{\boldsymbol{\theta}}(x) = \arg\max \boldsymbol{p}_k(x, \boldsymbol{\theta}), \quad \text{where } \boldsymbol{p}_k(x, \boldsymbol{\theta}) = \frac{e^{z_k(x,\boldsymbol{\theta})}}{\sum_{j=1}^{K} e^{z_j(x,\boldsymbol{\theta})}}, \tag{7}$$

where $z_k(x, \boldsymbol{\theta})$ denotes the logits output of a network and $\boldsymbol{p}_k(x, \boldsymbol{\theta})$ represents the softmax output of x. Then we denote x' as the AE, X' and $\mathcal{D}'$ be the adversarial set and distribution, respectively. We perform PGD attack to produce the AEs, i.e.,

$$x^0 = x + \sigma, \quad \text{where } \sigma \sim \mathcal{N}(0, 1), \tag{8}$$

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \alpha sign(\nabla_x \mathcal{L}(\theta, x^t, y)), \tag{9}$$

where x denotes the natural example and x^0 is obtained by perturbing x with random noise σ sampled from the normal distribution $\mathcal{N}(0, 1)$, t denotes the current time step, α is the step size, Π denotes the projection function, $\mathcal{S} \subseteq \mathbb{R}^d$ denotes the perturbation set of AEs.

Based on the definition, the adversarial risk to be optimized for the given dataset and classifier h_θ is defined as follows:

$$\mathcal{R}_{adv}(h_\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{x}_i' \in \mathcal{S}} \ell\left(h_\theta(\mathbf{x}_i') \neq y_i\right). \tag{10}$$

3.2 Interactions between adversarial training and robust loss functions

In Sect. 2.3, we introduced that loss functions with symmetric properties are robust to inherent label noise. In this section, we will mainly delve into their theoretical details. Following the works in Ghosh et al. (2017) and Ma et al. (2020), we first demonstrate that a symmetric loss function exhibits noise tolerance under both symmetric and asymmetric label noise with certain mild assumptions.

Lemma 1 *In a multi-class classification problem, let a loss function L satisfy Eq.* (6)*. Then L is noisy-robust under symmetric label noise if noise rate* $\eta < \frac{K-1}{K}$*.* (*Ghosh et al.* 2017)

Lemma 2 *In a multi-class classification problem, suppose L satisfies Eq.* (6) *and* $0 \leq L(h(x), k) \leq \frac{C}{(K-1)}, \forall k \in K$*. If* $R(h^*) = 0$*, then L is noisy-robust under asymmetric label noise if noise rate* $\eta_{jk} < \eta_{jj}$*.* (*Ghosh et al.* 2017) *where C is a constant,* h^* *denotes the global minimizer. Ghosh et al.* (2017) *provide detailed proofs of these two lemmas. Lemma* 2 *is not easy to understand, we also provide our proof in "Appendix A". Given the above conditions on the label noise rate* η*, then the learning risk under both clean labels* $R(h)$ *and under noisy labels with noise rate* η *:* $R^\eta(h)$ *shares the same global minimizer* h^**, i.e., the loss function L is noisy-robust.*

The above discussion focuses on the robust loss functions for ST, *i.e.*, natural examples. Referring to Eqs. (8) and (9), we know that AEs are generated at the input level. Therefore, when confronted with symmetric label noise, the loss function L remains noisy-robust for AEs if it is already noisy-robust for natural examples, since there are no additional conditions regarding the inputs, as stated in Lemma 1. However, when considering asymmetric label noise, first note that the condition $0 \leq L(h(x), k) \leq \frac{C}{(K-1)}, \forall k \in K$ can be easily satisfied by a typical loss function (Ma et al. 2020). However, when replacing natural examples with AEs, it is intuitive that the value of L will increase significantly, potentially exceeding the upper bound. Another condition $R(h^*) = 0$ is a restrictive condition for the noisy-robust theory to hold, which means that an h^* can achieve 100% classification accuracy. In experiments, it has been observed that satisfactory performance can still be achieved as long as $R(h^*)$ is close to 0 (Ma et al. 2020). However, as $R(h^*)$ increases, the corresponding performance tends to decline. Currently, the SOTA robust accuracy for AEs on CIFAR-10 is below 70%, indicating a high value of $R(h^*)$ and low utility of noisy-robust loss functions for AEs. Based on the above analysis, in the case of asymmetric label noise, AEs may not satisfy the two conditions for the noisy-robust properties to hold. We can conclude that:

Proposition 1 *In a multi-class classification problem, suppose L is a noisy-robust loss function, then L remains noisy-robust for AEs under symmetric label noise while may be non-robust for AEs under asymmetric label noise.*

Given the analysis presented above, the mainstream AT algorithms currently rely on the CE loss, some of them direct optimizing AEs, as seen in PGD-AT in Eq. (2) and MART in Eq. (4). However, replacing the CE with a robust loss function may not be effective for them under asymmetric label noise referring to Proposition 1. In contrast, TRADES in Eq. (3) optimizes natural examples and incorporates a regularization term to approximate the distribution of AEs and natural examples, it is mathematically well-suited for the application of a robust loss function. We will also empirically verify this proposition in the experiments.

3.3 Noisy robust cross entropy loss

First, we demonstrate why a simple mean absolute error (MAE) is symmetric while CE does not. For a K-class classification, recall Eq. (6), it is obvious that for MAE, $\sum_j^K \mathcal{L}(f(\boldsymbol{x}),j) = \sum_j^K (2 - 2\sum_j^K \boldsymbol{p}(k \mid \boldsymbol{x})) = 2 * (K-1)$ which is a constant. While for CE, the $\sum_j^K \mathcal{L}(f(\boldsymbol{x}),j) = -log\boldsymbol{p}(y \mid \boldsymbol{x})$ where y is the ground truth, it is obvious that the value of $-log\boldsymbol{p}(y \mid \boldsymbol{x})$ may vary for different x, *i.e.*, the value of $\sum_j^K \mathcal{L}(f(\boldsymbol{x}),j)$ is not a constant for CE. However, the simplicity of MAE makes it susceptible to underfitting on large datasets. While CE is a widely used and effective loss function, it lacks the property of noisy-robust. Thus, it is intuitive to consider combining the advantages of both MAE and CE.

In Ma et al. (2020), they divide the loss function into active and passive components according to whether it solely counts on the value of $\boldsymbol{p}(k = y|x)$. Specifically, CE is considered as an active loss function while MAE is regarded as a passive loss function. In Ma et al. (2020), they argue that combining an active loss function with a passive loss function can benefit from complementary learning, as demonstrated by Kim et al. (2019). This combination is referred to as Active-Passive Loss (APL) framework. Furthermore, by using noisy-robust versions of both active and passive loss functions, a noisy-robust APL loss can be obtained.

To transfer CE to a noisy-robust APL loss form, we require the noisy-robust active and passive versions of CE. The normalized CE, shown in Eq. (11) which is obtained by dividing by $\sum_j^K \mathcal{L}(h(\boldsymbol{x}),j)$, is proven to be a noisy-robust active loss function in Ma et al. (2020), while the reversed CE in Eq. (12) is a noisy-robust passive loss function. Currently, the SOTA version of robust loss functions is a combination of NCE+RCE.

$$\begin{aligned} NCE = \frac{CE}{\sum_{j=1}^K \mathcal{L}(h(x),j)} &= \frac{-\sum_{k=1}^K \boldsymbol{q}(k \mid \boldsymbol{x}) \log \boldsymbol{p}(k \mid \boldsymbol{x})}{-\sum_{j=1}^K \sum_{k=1}^K \boldsymbol{q}(y=j \mid \boldsymbol{x}) \log \boldsymbol{p}(k \mid \boldsymbol{x})} \\ &= \log_{\prod_k^K \boldsymbol{p}(k|\boldsymbol{x})} \boldsymbol{p}(y \mid \boldsymbol{x}), \end{aligned} \tag{11}$$

$$RCE = -\sum_{k=1}^{K} \boldsymbol{p}(k \mid \boldsymbol{x}) \log \boldsymbol{q}(k \mid \boldsymbol{x}), \tag{12}$$

By definition, both NCE and RCE satisfy the symmetry in Eq. (6). Therefore, NCE+RCE can be regarded as a robust variant of CE when noisy labels are inhered in the training datasets.

3.4 Noisy robust adversarial training

In Sect. 3.2, we have analyzed why TRADES is suitable for incorporating a robust loss function while PGD-AT and MART are not as compatible in this regard. Our NRAT is formally based on TRADES, with enhancements made to both of its components. These enhancements aim to make NRAT more effective for datasets with inherent label noise. We rewrite the original objective function of TRADES first

$$\ell^{TRADES}(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \mathrm{CE}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\big) + \beta \cdot \max_{\mathbf{x}_i' \in \mathcal{S}} \mathrm{KL}\big(h_{\boldsymbol{\theta}}(\mathbf{x}_i) \| h_{\boldsymbol{\theta}}(\mathbf{x}_i')\big) \right\}, \tag{13}$$

We already know that CE is non-robust for inherent label noise. Now, we demonstrate that the KL-divergence can also be replaced by a more robust alternative. The KL-divergence, given by

$$\mathrm{KL}(p(x_i, \boldsymbol{\theta}) || p(x_i', \boldsymbol{\theta})) = \sum_{k=1}^{K} p_k(x_i, \boldsymbol{\theta}) log \frac{p_k(x_i, \boldsymbol{\theta})}{p_k(x_i', \boldsymbol{\theta})}, \tag{14}$$

the KL-divergence is an asymmetric measure that treats two distributions unequally. In learning with noisy labels, it becomes apparent that the ground truth distribution $\boldsymbol{q}(k|x)$ for x may not accurately reflect the true distribution, while the predicted distribution $\boldsymbol{p}(k|x)$ may better represent the true distribution to some extent (Wang et al. 2019). Therefore, in learning with noisy labels, it will be more robust to utilize both $\mathrm{KL}(p_1||p_2)$ and $\mathrm{KL}(p_2||p_1)$ to obtain a symmetric divergence measure

$$\mathrm{KL}_{sym}(p(x_i, \boldsymbol{\theta}) || p(x_i', \boldsymbol{\theta})) = \frac{1}{2} \big\{ \mathrm{KL}(p(x_i, \boldsymbol{\theta}) || p(x_i', \boldsymbol{\theta})) + (\mathrm{KL}(p(x_i', \boldsymbol{\theta}) || p(x_i, \boldsymbol{\theta})) \big\}. \tag{15}$$

Another theory that supports the use of symmetric KL-divergence is the memorization effects (Dong et al. 2021) of neural networks. These effects indicate that neural networks have the capability to fit training data well, even in the presence of noisy labels. However, mislabeled examples tend to incur larger losses compared to correctly labeled examples (Song et al. 2019), leading to increased uncertainty in the output probabilities $p(x_i, \boldsymbol{\theta})$ and $p(x_i', \boldsymbol{\theta})$ (assuming x_i as a mislabelled example). Consequently, as the label noise rate increases, the inequality of KL-divergence is magnified throughout the dataset. We also analyze the robust risk of symmetric KL-divergence under symmetric label noise in Appendix B and show that symmetric KL-divergence tends to have a tighter bound compared with KL-divergence. Based on this analysis, we replace the two terms in the original TRADES with more robust alternatives that are more suitable for datasets with inherent noisy labels, the training objective function of NRAT is defined as followed

$$\begin{aligned} \ell^{NRAT}(\mathrm{x}_i, y_i, \boldsymbol{\theta}) := \sum_{i=1}^{n} \Big\{ & L_{apl}\big(h_{\boldsymbol{\theta}}(\mathrm{x}_i), y_i\big) \\ & + \lambda \cdot \max_{\mathbf{x}_i' \in \mathcal{S}} \mathrm{KL}_{\mathrm{sym}}\big(h_{\boldsymbol{\theta}}(\mathrm{x}_i) \| h_{\boldsymbol{\theta}}(\mathrm{x}_i')\big) \Big\}, \end{aligned} \tag{16}$$

where L_{apl} denotes the robust loss functions NCE+RCE following the APL framework in Eqs. (11) and (12), KL_{sym} denotes the symmetric KL-divergence in Eq. (15). The pseudocode of the training algorithm for NRAT is given below.

Input: Network h_θ, training set $X = \{(x_i, y_i)\}_{i=1,\ldots n}$, training epochs T, mini-batch B, perturb steps P, step size α, learning rate β, noise rate $\eta \in [0, 1]$.
Output: Adversarially robust model h_θ against label noise.
Initialization: Randomly inject noisy labels to training set X with a noisy rate η; Random initialization of h_θ;
while $t \leq T$ **do**
  Sample a mini-batch B from training set X;
  $x'_B = x_B + \sigma$, with $\sigma \sim \mathcal{N}(0, 1)$;
  for $p = 1$ *to* P **do**
    $x'_B = \Pi_{x'_B + \mathcal{S}}(x'_B + \alpha sign(\nabla_x CE(\theta, x'_B, y))$;
  end
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \sum_{\mathbf{x}'_B} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$. $\#\mathcal{L}$ followed Eq. (16);
end

3.5 Relation to existing work

So far, PGD-AT, MART, and TRADES have commonly been used as baselines for newly proposed AT algorithms, like in Wu et al. (2020), we have introduced our NRAT as an enhanced and more noisy-robust objective function compared to TRADES in Sect. 3.4. In this section, we will mainly focus on the distinctions between our NRAT and MART, since they also share a similar formulation of objective functions.

MART divides the training dataset into correctly classified examples and misclassified examples which are similar to correctly labeled examples and mislabeled examples. When facing noisy labels, we can also divide the natural training set $\mathcal{S}$ into two subsets, that is, examples with correct labels as $\mathcal{S}^+$ and examples with noisy labels as $\mathcal{S}^-$, given a classifier h^*_θ that satisfies $\mathcal{R}(h^*_\theta) = 0$, then we get:

$$\begin{aligned} \mathcal{S}^+_{h_\theta} &= \left\{ i : i \in [n], h^*_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i = y^*_i \right\}; \\ \mathcal{S}^-_{h_\theta} &= \left\{ i : i \in [n], h^*_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i \neq y^*_i \right\}. \end{aligned} \tag{17}$$

However, in learning with noisy labels, we do not know which label is incorrect or correct in advance. Hence, it becomes necessary to minimize the overall risk $\mathcal{R}(h_{\boldsymbol{\theta}})$ instead of dividing it into subsets, as done in MART. Therefore, our NRAT algorithm follows PGD-AT and TRADES by minimizing the risk of the whole dataset. MART, on the other hand, it utilizes different objective functions for correctly classified and misclassified examples. MART achieves optimal performance when applied to clean datasets, while the presence of noisy labels often results in misclassifications being actually correct, and vice versa. This leads to the possibility of its different objective functions being applied to inappropriate examples, thereby diminishing its performance.

4 Experiments

In this section, we empirically evaluate the performance of the proposed NRAT on CIFAR-10 dataset against two types of injected inherent label noise: symmetric noise and asymmetric noise. We compare our method with three existing AT methods and their variants on the noisy dataset with varying label noise rates.

4.1 Experimental setup

Baselines We consider three well-known AT algorithms as baselines: (1) PGD-AT; (2) TRADES; (3) MART. To evaluate the effectiveness of robust loss functions in these algorithms, we also replace the CE loss used in these algorithms with NCE+RCE as three additional baselines, i.e., (4) PGD-AT-APL; (5) TRADES-APL; (6) MART-APL.

Generation of label noise To simulate real-world datasets that may contain inherent label noise, we introduce two types of research-oriented label noise to the original CIFAR-10 dataset. Symmetric label noise refers to each label having an equal probability of being flipped to any other class; In contrast, asymmetric noise involves label flipping between similar classes, which is more representative of real-world scenarios. For asymmetric label noise, we flip labels between $TRUCK \leftrightarrow AUTOMOBILE$, $BIRD \leftrightarrow AIRPLANE$, $DEER \leftrightarrow HORSE$, and $CAT \leftrightarrow DOG$, following (Zhang and Sabuncu 2018). We consider noise rates ranging from 20% and 40% to simulate the noise rate in real-world datasets, and we also report the performance on the clean dataset without any label noise (0%). We also provide the results of NRAT on MNIST and FashionMNIST via Table 4 in the "Appendix C".

Adversarial training settings For AT, we train ResNet18 on all algorithms, we basically follow the standard settings in Rice et al. (2020) with some improvements made to be more suitable for AT with noisy labels. Specifically, we use stochastic gradient descent (SGD) with momentum 0.9, the total training epochs is 200, with weight decay 5e-4, we used standard data augmentation, *i.e.*, random crops and random horizontal flips, we also implement data normalization for all methods. For the training attack, we use PGD-10 with random initialization and perturbation limit $\epsilon = 8/255$, step size 2/255. For the initial learning rate, the standard default value is 0.1, while we choose different smaller initial learning rates from [0.01, 0.05, 0.1] for different noisy rates since is prone to show gradient collapse when AT with inherent label noise, the general principle is to choose the largest possible learning rate without encountering gradient collapse. (In "Appendix D", we provide an additional experiment by replacing the CE in PGD attack with our proposed loss function.)

We use ℓ_∞ threat models for all methods. We do not train any WideResNet since it usually shows a similar trend with ResNet18, while it is much more time-consuming. For NCE+RCE in NRAT, we follow the setting in Rice et al. (2020) for CIFAR10, i.e., both the coefficients before the two terms are 1. The hyperparameters of the baselines are consistent with their original papers: $\lambda = 5$ for MART and $\beta = 6$ for TRADES. For our NRAT, we try $\lambda = [4, 6, 8, 10]$ and find that $\lambda = 6$ yields the best empirical results across different noise rates, we report the best natural-robust trade-off performance for all the methods. All experiments are implemented on a server with an Intel i7-12700F CPU and an RTX3090 GPU. *Note* that we do not perform any training tricks like gradient clipping, label smoothing, etc., to accurately compare the performance between different objective strategies.

4.2 Performance evaluation

Adversarial attacks We conduct two different typical white-box attacks: PGD-20, CW-20 (Carlini and Wagner 2017) (the ℓ_∞ version of CW loss optimized by PGD-20), and one more powerful auto attack (Croce et al. 2020) to evaluate the baselines as well as NRAT. Auto attack contains an ensemble of parameter-free attacks, which can serve as a reliable metric for assessing the robustness performance of a model. While some of the attacks here may not be a white-box attack in the noisy labels setting, like CW-20, as it may easily be swayed by gradient obfuscation caused by the random label flipping, we believe they can reflect the robustness performance of the model to a certain extent.

To evaluate the performance, we report "natural" and "robust" which denote the accuracy of natural test images and adversarial test images using different attacks, respectively. From it, we can see the natural-robust trade-off of different methods. Another metric of measuring AT is the degree of robust overfitting, so we also report the "Best" (highest accuracy) and "Last" (accuracy at the last training epoch) natural and robust accuracy to see the gap between them, the smaller the gap, the lower the degree of overfitting. Results are shown in Tables 1 and 2 for learning with symmetric/asymmetric label noises respectively. Recall Fig. 1, we find that even natural accuracy is overfitting when there is label noise for the baselines. From our results, this double overfitting can be largely mitigated by our method.

Remark for Tables 1 and 2. Under the symmetric label noise, Table 1 shows that NRAT can outperform the baselines considering the best robust accuracy when facing symmetric label noise. While for the clean dataset, TRADES exhibits superior robust performance. Comparing the performance of MART with MART-APL and TRADES with TRADES-APL, we observe that the robust performance both improved under 20% and 40% symmetric label noise. Particularly, MART-APL demonstrates a significant improvement, these results indicate that in the presence of noisy labels, a robust loss function can be considered as a more robust alternative to the CE loss.

While MART and TRADES exhibit significant robust overfitting, considering the gap between the last performance and best performance (around 11% to 16% for 20% symmetric label noise and 18% to 28% for 40% asymmetric label noise), the APL versions can significantly mitigate the double overfitting issues (around 6% to 10% for 20% symmetric label noise and 8% to 10% for 40% asymmetric label noise. These demonstrate the effectiveness of robust loss functions in addressing the double overfitting issues. We provide the learning curves for MART-APL and TRADES-APL in Fig. 3 below:

Under the asymmetric label noise, Table 2 further demonstrates that NRAT achieves the highest robust performance under 20% and 40% asymmetric label noise. Another noteworthy observation is that the best robust performance of MART-APL consistently falls below that of MART, which aligns with our Proposition 1 that the robust loss functions may be non-robust for AEs under asymmetric label noise, as the CE is non-robust and the performance of NCE+RCE is even lower than CE around 1% to 3.5%. Another notable phenomenon is that the robust overfitting observed in PGD-AT, MART, and TRADES is not as pronounced under asymmetric label noise compared to symmetric label noise. This suggests that, in AT, the CE loss is relatively more robust for asymmetric label noise compared to symmetric label noise. Conversely, in ST, asymmetric label noise is generally more challenging.

The effectiveness of symmetric KL-divergence The key difference between TRADE-APL and our NRAT is whether to use a symmetric KL-divergence or not. Considering

Table 1 Best and last robustness performance(%) on CIFAR-10 with inherent symmetric label noise with 0%, 20% and 40% noise rate

Methods	Clean				Sym0.2				Sym0.4			
	Natural	PGD-20	CW-20	AA	Natural	PGD-20	CW-20	AA	Natural	PGD-20	CW-20	AA
	Best natural and robust accuracy											
PGD-AT	86.18	51.76	51.74	47.40	81.77	49.91	49.23	46.04	77.25	47.96	47.29	44.38
MART	86.72	50.13	50.86	42.88	78.96	48.01	47.92	42.51	74.97	45.82	45.25	34.49
MART-APL	85.93	53.60	53.42	44.72	83.22	54.35	54.13	43.15	79.37	53.65	53.01	45.66
TRADES	86.67	**59.95**	**59.79**	**54.77**	81.37	56.71	56.23	54.67	75.80	54.80	54.92	45.68
TRADES-APL	**87.73**	56.12	55.02	46.42	**84.52**	56.80	56.25	54.89	**79.40**	55.26	55.20	47.02
NRAT	87.11	56.97	55.93	46.32	83.88	**56.95**	**56.79**	**55.38**	79.00	**56.46**	**56.23**	**49.00**
	Last natural and robust accuracy											
PGD-AT	85.73	44.25	43.87	41.75	76.88	32.24	30.92	29.18	69.21	29.77	28.04	26.67
MART	86.69	45.28	44.92	43.99	74.11	37.07	37.23	34.84	54.54	22.89	21.97	19.93
MART-APL	85.68	45.39	45.01	44.67	83.07	44.47	42.98	43.20	**78.07**	44.13	43.99	42.27
TRADES	86.67	**55.57**	**53.92**	**54.77**	74.76	39.74	39.10	37.99	60.23	26.12	26.09	24.47
TRADES-APL	**87.42**	49.81	49.07	48.50	**83.89**	**50.46**	**50.06**	**49.35**	77.91	45.93	45.78	44.09
NRAT	86.75	47.94	47.58	47.27	82.72	49.97	49.25	48.70	75.87	**48.88**	**47.90**	**47.15**

Table 2 Best and Last Robustness performance(%) on CIFAR-10 with inherent asymmetric label noise with 20% and 40% noise rate

Methods	asym0.2				asym0.4			
	Natural	PGD-20	CW-20	AA	Natural	PGD-20	CW-20	AA
	Best natural and robust accuracy							
PGD-AT	83.72	50.98	51.12	46.40	**79.29**	48.49	49.20	42.83
MART	83.04	54.19	54.09	50.27	76.85	47.97	47.28	44.87
MART-APL	84.43	53.19	52.90	48.63	70.36	44.44	44.29	42.98
TRADES	82.46	54.12	54.80	50.68	77.44	50.46	50.02	47.67
TRADES-APL	**85.15**	55.64	55.71	51.74	76.67	51.64	50.99	48.89
NRAT	84.08	**57.84**	**57.67**	**53.08**	75.49	**51.86**	**51.21**	**49.62**
	Last natural and robust accuracy							
PGD-AT	80.61	40.35	39.80	38.99	**73.66**	38.27	38.02	35.68
MART	80.81	42.62	42.70	41.98	71.88	39.68	39.37	38.10
MART-APL	84.28	44.53	43.01	42.72	69.08	38.91	37.56	36.97
TRADES	77.71	49.31	**49.34**	48.10	70.87	**40.62**	39.89	**38.91**
TRADES-APL	**84.52**	46.91	46.28	45.76	72.74	40.28	**40.42**	38.53
NRAT	83.36	**49.81**	49.03	**48.55**	71.01	39.47	40.09	37.42

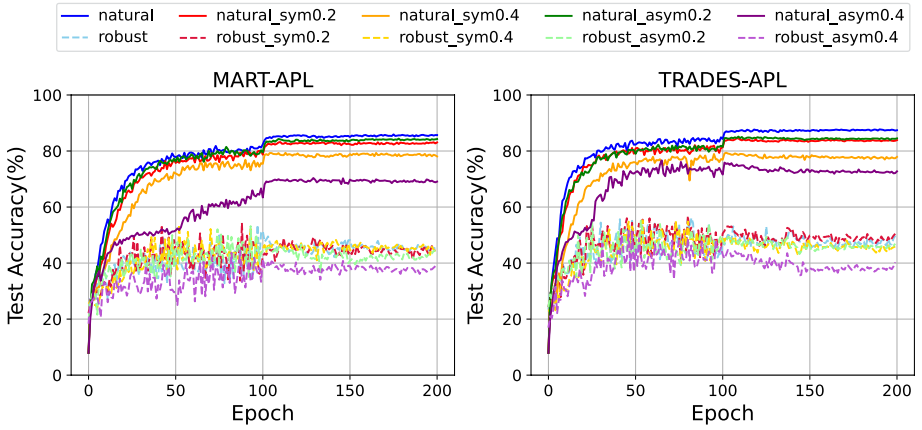

Fig. 3 The learning curves of natural accuracy and PGD robust accuracy for MART-APL, and TRADES-APL under 0% (natural/robust), 20% and 40% inherent symmetric/asymmetric label noise on CIFAR-10 with ℓ_∞ threat model

their performance shown in Tables 1 and 2, it is evident that NRAT consistently achieves higher robust performance but lower natural performance compared to TRADES-APL. This highlights the role of symmetric KL divergence, as it serves to bridge the performance gap between natural and robust, albeit at the cost of some natural performance. Given that robust performance is the primary focus of AT, indeed, this trade-off is considered an appropriate compromise.

The performance of PGD-AT-APL As we have analyzed in Sect. 3.2, PGD-AT is not well-suited for APL. Empirically, the training process for PGD-AT-APL exhibits a peculiar

tendency, with significantly low natural accuracy (less than 30%) and high robust accuracy (more than 60%). Therefore we do not show the results of PGD-AT-APL in the above tables. However, the underlying reasons for such performance remain an open issue that requires further investigation.

Further discussions with TRADES We make a full comparison between our NRAT and TRADES in this part. First, under symmetric label noise, NRAT gets a higher best robust accuracy, and with the noise rate increasing from 20 to 40%, the improvement becomes apparent at around 2% to 4% under different attacks; another improvement is that NRAT can mitigate the double overfitting, from the second part of Table 1, TRADES shows a significant double overfitting issue, that the gap between the last accuracy and best accuracy is quite large, while the last performance of NRAT is much higher than that of TRADES. Second, for asymmetric label noise which shows an opposite phenomenon that although NRAT still outperformance TRADES at the best robust accuracy, with the noise rate increase, the gap becomes closer. This is related to the condition $R(h^* = 0)$ in Lemma 2, with the noise rate increase, $R(h^*)$ tends to be far away from 0 which limit the performance of the NCE+RCE loss.

4.3 Mitigating double overfitting

Although NRAT partially mitigates the issue of double overfitting, there is still significant robust overfitting, resulting in a substantial best and last performance gap. The gap is around 7% to 10% for symmetric label noise and 8% to 12% for asymmetric label noise. To further address this, we introduce using weight perturbation. Adversarial weight perturbation (AWP) (Wu et al. 2020) aims to adversarially perturb both the inputs and weights during the training stage. The input perturbation is produced via PGD attack, while the weight perturbation can be written as

$$\mathbf{v} \leftarrow \Pi_{\gamma}\left(\mathbf{v} + \eta \frac{\nabla_{\mathbf{v}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{f}_{\mathbf{w}+\mathbf{v}}\left(\mathbf{x}_i'\right), y_i\right)}{\left\|\nabla_{\mathbf{v}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{f}_{\mathbf{w}+\mathbf{v}}\left(\mathbf{x}_i'\right), y_i\right)\right\|} \|\mathbf{w}\|\right), \tag{18}$$

where $\mathbf{v}$ denotes the weight perturbation, which can be solved by multi-step methods like PGD, and n is the batch size. Combining x' and $\mathbf{v}$ for adversarial training has been shown to enhance adversarial robustness, as well as alleviate robust overfitting. Furthermore, we empirically demonstrate that NRAT is compatible with AWP and can effectively mitigate the issue of double overfitting in the presence of label noise. The comparison between NRAT and NRAT-AWP is shown in Fig. 4 and Table 3.

It is clear that NRAT-AWP achieves higher robust accuracy and significantly mitigates robust overfitting. The performance gap is less than 5% across all label noise rates.

4.4 AT with generated data

Currently, one of the most effective approaches in AT is leveraging additional data. For instance, Wang et al. (2023) used the elucidating diffusion model (EDM) (Turkeltaub et al. 2023) to generate millions of additional data for AT, leading to the state-of-the-art performance on the RobustBench (Croce et al.. 2020) leaderboard. However, it is worth noting that these augmented datasets may also contain an unknown proportion of noisy labels.

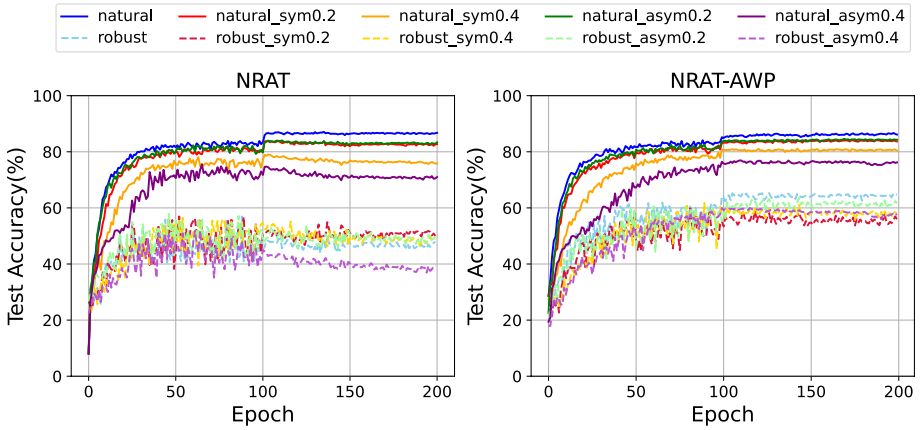

Fig. 4 The learning curves of natural accuracy and PGD robust accuracy for NRAT and NRAT-AWP under 0% (natural/robust), 20% and 40% inherent symmetric/asymmetric label noise on CIFAR-10 with ℓ_∞ threat model

Table 3 Robustness performance(%) on CIFAR-10 of NRAT-AWP and NRAT with 20% and 40% symmetric/asymmetric label noise

Methods		sym0.2				sym0.4			
		Natural	PGD-20	CW-20	AA	Natural	PGD-20	CW-20	AA
	Natural and robust accuracy of NRAT and NRAT_AWP								
NRAT_AWP	Best	84.18	59.82	59.51	57.80	81.1	61.77	61.55	60.68
	Last	83.70	55.98	55.77	54.45	80.41	57.96	57.80	56.55
NRAT	Best	83.88	56.95	56.79	55.38	79.00	56.46	56.23	49.00
	Last	82.72	49.97	49.25	48.70	75.87	48.88	47.90	47.15
NRAT_AWP	Best	84.59	63.54	63.44	62.24	76.91	60.03	60.10	58.39
	Last	84.31	61.61	61.45	60.35	76.23	57.84	57.98	56.07
NRAT	Best	84.08	57.84	57.67	53.08	75.49	51.86	51.21	49.62
	Last	83.36	49.81	49.03	48.55	71.01	39.47	40.09	37.42

Out of curiosity, we also trained NRAT on these additional data. (We refer to their method as DM_AT in this section.)

Settings for this part We use the 1 M generated data provided in Wang et al. (2023), following most of the settings outlined in Sect. 4.1. While for each method (DM_AT and NRAT), we employed the WideResNet-28–10 model to train this large dataset. Additionally, as per Wang et al. (2023), we apply label smoothing with a value of 0.1 and separate the first 1024 images of the training set to create a fixed validation set to replace the test data in CIFAR-10, since the distribution of generated dataset is still different from the distribution of the test set of CIFAR-10 dataset, a fixed validation set sampled from the generated dataset can be seen as a more fair comparison to eliminate the impact of distribution distance. We train each method for 150 epochs to observe the training tendency. The performance on the validation set is shown in Fig. 5.

Although the exact number of noisy labels in the generated dataset is unknown, it is clear from Fig. 5 that NRAT exhibits higher clean accuracy on the validation set

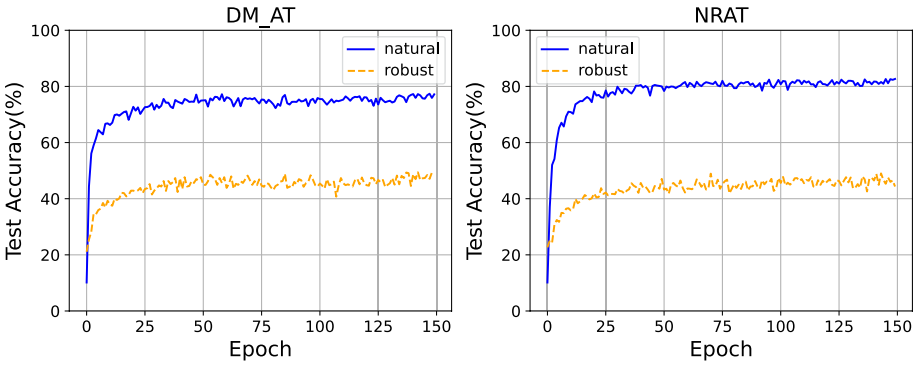

Fig. 5 The learning curves of natural accuracy and PGD robust accuracy on the validation set for DM_AT and NRAT using 1 M generated data with ℓ_∞ threat model

compared to DM_AT. However, the robust accuracy of NRAT appears slightly lower than that of DM_AT. The best natural accuracy achieved is 77.4% for DM_AT and 82.66% for NRAT, while the best robust accuracy is 49.41% for DM_AT and 49.02% for NRAT.

5 Conclusion

In this paper, we first investigate the performance of existing AT methods when confronted with inherent label noise. We observe that these methods exhibit poor generalization on inherent label noise. To address this issue, we propose a novel noisy robust adversarial training algorithm, i.e., NRAT, by incorporating a robust loss function and a more robust regularization term to enhance adversarial robustness in the presence of inherent label noise. This work is a combination of technologies in the field of noisy labels and AT, aiming to improve the performance of adversarial robustness on more realistic datasets. Comprehensive experiments show that, with inherent label noise, NRAT achieves comparable or superior performance compared to existing AT algorithms in terms of robust accuracy and robust overfitting. Furthermore, we empirically show that NRAT is well-suited for training with large generated datasets, which is the state-of-the-art practice for improving adversarial training.

Appendix A: Proof for Lemma 2

Lemma 2 mainly shows the two conditions for a loss function L to be noise tolerant under asymmetric label noise, which means given a classifier h, if h^* is the global minimizer of $R(h)$, then h^*_η is the minimizer under the asymmetric label noise, *i.e.* $R^\eta\left(h^*_\eta\right) - R^\eta(h^*) \leq 0$.

Proof For asymmetric label noise, the risk of a loss function L is:

$$\begin{aligned} R^{\eta}(h) &= \mathbb{E}_{\boldsymbol{x},y} L(h(\boldsymbol{x}), y) = \mathbb{E}_{x}\mathbb{E}_{y^*|x}\mathbb{E}_{y|x,y^*} L(h(\boldsymbol{x}), y) \\ &= \mathbb{E}_{x}\mathbb{E}_{y^*|x}\left[\eta_{y^*} L(h(\boldsymbol{x}), y^*) + \sum_{k\neq y^*} \frac{\left(1-\eta_{y^*}\right) L(h(\boldsymbol{x}), k)}{K-1}\right] \\ &= \mathbb{E}_{x}\mathbb{E}_{y^*|x}\eta_{y^*} L(h(\boldsymbol{x}), y^*) + \mathbb{E}_{x}\mathbb{E}_{y^*|x}\frac{1-\eta_{y^*}}{K-1}(C - L(h(\boldsymbol{x}), y^*)) \\ &= \mathbb{E}_{x}\mathbb{E}_{y^*|x}\frac{C}{K-1}\left(1-\eta_{y^*}\right) + \mathbb{E}_{x}\mathbb{E}_{y^*|x}\left(\left(1 - \frac{K\left(1-\eta_{y^*}\right)}{K-1}\right) L(h(\boldsymbol{x}), y^*)\right) \end{aligned} \tag{19}$$

where y^* denotes true labels, η_{y^*} means the rate of label y^* being the true labels, therefore $1-\eta_y^*$ is the noisy label rate, then

$$R^{\eta}(h^*) - R^{\eta}(h) = \mathbb{E}_{x}\mathbb{E}_{y^*|x}\left\{\left(1 - \frac{K\left(1-\eta_{y^*}\right)}{K-1}\right)(L(h^*(\boldsymbol{x}), y^*) - L(h(\boldsymbol{x}), y^*))\right\} \tag{20}$$

If $R(h^*) = 0$ and L is a non-negative robust loss function, then $(L(h^*(\boldsymbol{x}), y^*) = 0$, since $\eta_{y^*} < 1$, then $1 - \frac{K\left(1-\eta_{y^*}\right)}{K-1} > 0$ and we have $R^{\eta}(h^*) - R^{\eta}(h) \leq 0$, which means the h^* for clean dataset is also the minimizer for asymmetric noisy dataset, thus completes the proof. □

Appendix B: Robust risk of symmetric KL-divergence

For a dataset with the label y which contains the symmetric label noise with the noise rate η, we denote the robust risk of $\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}'))$ under the noise rate of η as $R^{\eta}(\mathcal{D}, y)$, where $\mathcal{D}$ is any distance metrics, $\boldsymbol{x}$ and $\boldsymbol{x}'$ represent natural examples and adversarial examples respectively:

$$\begin{aligned} R^{\eta}(\mathcal{D}, y) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',y}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}')) = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\mathbb{E}_{\boldsymbol{x}'}\mathbb{E}_{y|\boldsymbol{x}'}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}')) \\ &= \mathbb{E}\left[(1-\eta)\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}')) + \frac{\eta}{K-1}\sum_{k\neq y}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}'))\right] \\ &= (1-\eta)R(\mathcal{D}, y) + \frac{\eta}{K-1}\left(\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}'))\right] - R(\mathcal{D}, y)\right) \\ &= R(\mathcal{D}, y)\left(1 - \frac{\eta K}{K-1}\right) + \frac{\eta}{K-1}\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}'))\right], \end{aligned} \tag{21}$$

To simplify, we denote $\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\mathbb{E}_{\boldsymbol{x}'}\mathbb{E}_{y|\boldsymbol{x}'}$ in the first line as $\mathbb{E}$ in the following lines. Given the assumption that in clean dataset, the $R(\mathcal{D}, y)$ of KL divergence and symmetric KL divergence are close, then the $R^{\eta}(\mathcal{D}, y)$ is only related to $\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{D}(h(\boldsymbol{x}), h(\boldsymbol{x}'))\right]$, which means in the case of symmetric (random) labels, the prediction distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$, this is not a tight bound, however in the case of random labels it becomes evident that both $h(\boldsymbol{x})$ and $h(\boldsymbol{x}')$ will be less robust compared to clean labels, take the KL divergence into $\mathcal{D}$, we are

not sure $\mathbb{E}\left[\sum_{k=1}^{K} \mathrm{KL}(h(\boldsymbol{x}), h(\boldsymbol{x}'))\right]$ and $\mathbb{E}\left[\sum_{k=1}^{K} \mathrm{KL}h(\boldsymbol{x}'), (h(\boldsymbol{x}))\right]$ which one may have a boarder impact because of the uncertainy of each sample $\boldsymbol{x}_i$, therefore we consider the symmetric KL divergence which utilize both the direction as the more robust and fair counterpart compared with KL divergence which in general the bound of $R^{\eta}(\mathcal{D}, y) - R(\mathcal{D}, y)$ is lower than KL divergence.

Appendix C: Performance of NRAT with MNIST and Fashion MNIST

For MNIST and FashionMNIST, we use a small CNN network with 4 layers as the defense model. The total training epochs are 100 and the initial learning rate is 0.01 which is divided by 10 at the 55-th, 75-th, and 90-th epochs. The perturbation $\boldsymbol{\delta} = 0.3$. Results are shown in Table 4, For both symmetric and asymmetric label noise except for the asymmetric label noise 40%, NRAT achieves similar results with the clean dataset which shows a good generability under noisy labels in these two datasets. We omit the results of the CW attack as they are similar to PGD.

Table 4 Robustness performance(%) on MNIST and FashionMNIST of NRAT with 0%(clean) 20% and 40% symmetric/asymmetric label noise

	Clean			Sym0.2			Sym0.4		
	Natural	PGD-20	AA	Natural	PGD-20	AA	Natural	PGD-20	AA
MNIST									
Best	99.49	98.72	98.65	99.44	98.76	98.68	99.44	98.67	98.61
Last	99.47	98.70	98.58	99.40	98.69	98.64	99.42	98.66	98.60
Best	99.49	98.72	98.65	99.49	98.75	98.67	82.87	71.72	69.89
Last	99.47	98.70	98.58	99.47	98.68	98.61	82.72	71.67	69.87
FashionMNIST									
Best	82.01	73.43	71.66	81.16	73.45	71.73	80.77	73.42	71.68
Last	81.88	73.22	71.57	81.04	73.32	71.72	80.69	73.21	71.67
Best	82.01	73.43	71.66	81.35	73.59	71.93	66.64	62.56	61.48
Last	81.88	73.22	71.57	81.24	73.46	71.84	66.56	62.42	61.46

Appendix D: Performance of NRAT with AE generation using the proposed loss function

In Algorithm 1, we follow the original PGD attack to use the CE in the AE generation, while in the training optimization, we choose the robust loss function NCE+RCE as the main part, therefore if we replace the CE with our proposed loss function in the AE generation would give a better intuition and a fair performance evaluation. Results are shown in Table 5, which indicate a slight improvement with the original PGD attack.

Table 5 Best and last robustness performance(%) of NRAT using our proposed loss function in AE generation

		Natural and robust accuracy of NRAT							
		sym0.2				sym0.4			
		Natural	PGD-20	CW-20	AA	Natural	PGD-20	CW-20	AA
NRAT	Best	83.52	57.76	57.42	56.54	79.84	57.03	56.79	52.92
	Last	82.72	52.25	51.98	50.48	77.43	50.03	49.47	48.22
NRAT	Best	83.80	57.99	57.70	53.42	76.48	52.32	52.01	50.92
	Last	82.19	50.34	50.09	49.27	72.19	41.77	41.19	40.13

Author contributions ZC contributed to the idea, algorithm, theoretical analysis, writing, and experiments. FW, RM, PX and XH contributed to the algorithm and writing. WR contributed to the idea and writing. All the co-authors participated in the discussion and contributed to refining the manuscript.

Funding This work was conducted without any specific funding.

Availability of data and materials We used the publicly available CIFAR-10 dataset.

Code availability All the codes is publicly available on GitHub at https://github.com/TrustAI/NRAT.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This work does not involve any human or animal subjects so has no ethical concerns.

Consent to participate Not applicable

Consent for publication Not applicable

References

Amid, E., Warmuth, M.K., Anil, R. & Koren, T. (2019). Robust bi-tempered logistic loss based on bregman divergences. *Advances in Neural Information Processing Systems 32*

Athalye, A., Carlini, N. & Wagner, D.A. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, pp. 274–283. PMLR

Carlini, N. & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)* (pp. 39–57). https://doi.org/10.1109/SP.2017.49.

Croce, F. & Hein, M. (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning* (pp. 2206–2216). PMLR.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P. & Hein, M. (2020). Robustbench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H. & Zhu, J. (2021) Exploring memorization in adversarial training. arXiv preprint arXiv:2106.01606.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L. & An, B. (2021). Can cross entropy loss be robust to label noise?. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 2206–2212).

Ghosh, A., Kumar, H. & Sastry, P.S.: (2017). Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 31).

Goodfellow, I.J., Shlens, J. & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems 31*

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

Jin, G., Yi, X., Huang, W., Schewe, S. & Huang, X. (2022). Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15273–15283).

Kim, Y., Yim, J., Yun, J. & Kim, J. (2019). Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 101–110).

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J. & Li, L.-J. (2017). Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Natarajan, N., Dhillon, I.S., Ravikumar, P.K., & Tewari, A. (2013). Learning with noisy labels. *Advances in Neural Information Processing Systems* **26** (2013)

Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582–597). IEEE

Rice, L., Wong, E. & Kolter, Z. (2020). Overfitting in adversarially robust deep learning. In *International conference on machine learning* (pp. 8093–8104). PMLR

Song, H., Kim, M., Park, D. & Lee, J.-G. (2019). How does early stopping help generalization against label noise?. arXiv preprint arXiv:1911.08059.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199

Turkeltaub, T., Mannheim, R., Furman, A., & Weisbrod, N. (2023). Elucidating the relationship between gaseous o2 and redox potential in a soil aquifer treatment system using data driven approaches and an oxygen diffusion model. *Journal of Hydrology, 618*, 129168.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X. & Gu, Q. (2019). Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

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

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W. & Yan, S. (2023). Better diffusion models further improve adversarial training. arXiv preprint arXiv:2302.04638.

Wu, D., Xia, S.-T., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems, 33*, 2958–2969.

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

Xu, W., Evans, D. & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L. & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning* (pp. 7472–7482). PMLR.

Zhang, J., Xu, X., Han, B., Liu, T., Niu, G., Cui, L. & Sugiyama, M. (2021). Noilin: Do noisy labels always hurt adversarial training? arXiv preprint arXiv:2105.14676.

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

Zheng, G., Awadallah, A.H., Dumais, S. (2021). Meta label correction for noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, (vol. 35, pp. 11053–11061).

Zhu, J., Zhang, J., Han, B., Liu, T., Niu, G., Yang, H., Kankanhalli, M., & Sugiyama, M. (2021). Understanding the interaction of adversarial training with noisy labels. arXiv preprint arXiv:2102.03482.