Machine Learning (2022) 111:2489–2513
https://doi.org/10.1007/s10994-022-06186-9

Re-thinking model robustness from stability: a new insight to defend adversarial examples

Shufei Zhang[1,2,3] · **Kaizhu Huang**[4] · **Zenglin Xu**[5]

Received: 8 July 2021 / Revised: 8 April 2022 / Accepted: 22 April 2022 /
Published online: 14 June 2022

Abstract
We study the model robustness against adversarial examples, referred to as small perturbed input data that may however fool many state-of-the-art deep learning models. Unlike previous research, we establish a novel theory addressing the robustness issue from the perspective of stability of the loss function in the small neighborhood of natural examples. We propose to exploit an energy function to describe the total variation in a small neighborhood and prove that reducing such energy guarantees the robustness against adversarial examples. We also show that the traditional training methods including adversarial training and virtual adversarial training tend to minimize the lower bound of our proposed energy function. Importantly, we prove that minimizing the energy function can obtain a better generalization bound than traditional adversarial training approaches. Through a series of experiments, we demonstrate the superiority of our model on different datasets for defending adversarial attacks. In particular, our proposed adversarial framework achieves the best performance compared with previous adversarial training methods on benchmark datasets CIFAR-10, CIFAR-100 and SVHN and they demonstrate much better robustness against adversarial examples than all the other comparison methods.

Keywords Model robustness · Adversarial examples · Adversarial training · Energy · Adversarial generalization

1 Introduction

Deep Neural Networks (DNN) have achieved great success in various tasks, such as speech recognition, image classification, and object detection (LeCun et al. 2015; He et al. 2017; Huang et al. 2017). However, recent research shows that certain small perturbations over the input samples, called adversarial examples, may fool many powerful deep learning models (Goodfellow et al. 2014). Adversarial examples have been shown to be ubiquitous

Editor: Zhi-Hua Zhou.

✉ Kaizhu Huang
kaizhu.huang@dukekunshan.edu.cn

Extended author information available on the last page of the article

in different fields such as image recognition, natural language processing, image generation, and data mining (Fischer et al. 2017; Eykholt et al. 2018).

There have been many seminal works studying how to generate more invasive adversarial examples (Goodfellow et al. 2014; Kurakin et al. 2016; Liu and Nocedal 1989; Lyu et al. 2015; Shaham et al. 2015; Madry et al. 2017).

Meanwhile, several defense methods are proposed to improve the adversarial robustness (Kos et al. 2018; Miyato et al. 2018, 2015; Papernot et al. 2016; Xu et al. 2017; Madry et al. 2017; Mao et al. 2019; Willetts et al. 2019; Pang et al. 2020; Zhang et al. 2020). Most of them are mainly based on adversarial training, i.e., in the manner of replacing natural examples with adversarial examples during the training process.

In parallel to studying how to defend adversarial examples, researchers also made great efforts in thinking about the theory and principles underlying the adversarial examples. In particular, Ma et al. have shown that adversarial examples are not isolated points but a dense region of the input space (Ma et al. 2018). Fawzi et al. studied the model robustness against adversarial examples by establishing a general upper bound (Fawzi et al. 2018, 2016). Finlay et al. and Lyu et al. have demonstrated that FGSM and their extended general cases can be interpreted as a form of regularization (Finlay et al. 2018; Lyu et al. 2015). Similarly, Cisse et al. showed that the sensitivity to adversarial examples can be controlled by the Lipschitz constant of the network and proposed a new structure of network which is insensible to adversarial examples (Cisse et al. 2017). CLEVER score is proposed to meausure the model robustness (Weng et al. 2018).

The above-mentioned seminal studies have got interesting and important results for trying to understand adversarial examples. Although some theoretical robustness bounds have been proposed, most are practically difficult to be used or optimized for improving the performance. Moreover, both previous adversarial training methods and theoretical robustness bounds ignore the robust generalization, an important factor affecting the performance of models on adversarial examples of unseen data. Namely, they fail to describe how well the robustness of adversarial training generalizes on unseen data (Stanforth et al. 2019; Carmon et al. 2019; Song et al. 2019; Wu et al. 2020).

Distinguished from these existing work, in this paper, we re-think the model robustness from the perspective of stability and establish a novel theoretical framework which is able to address the robustness issue mathematically and rigorously. We also analyze our framework from the perspective of robust generalization. In more details, inspired from the stability of the loss function in the small neighborhood of natural examples, we propose to exploit an energy function to describe the total variations within a small region, and we prove that reducing such energy guarantees the robustness against adversarial examples. We also prove that many traditional adversarial training methods (including both supervised and semi-supervised adversarial training) are essentially equivalent to minimizing the lower bound of the proposed energy function. This may therefore gain a new insight of the limitations of these current methods, since minimization of the lower bound unnecessarily minimizes the energy. From the perspective of robust generalization, we have showed that our method obtains a better robust generalization bound than traditional adversarial training methods. Furthermore, we propose a simple approach to approximate the energy function and design a more rational and practical method with the energy regularization which proves to achieve better robustness than previous methods. Finally, to verify the performance of the proposed method, we have conducted a series of experiments on several datasets with different adversarial attacks. Experimental results have shown that our proposed adversarial framework can achieve the best performance compared with previous adversarial methods benchmarked on CIFAR-10, CIFAR-100 and SVHN.

Importantly, they demonstrate much better robustness against adversarial examples than all the other comparison methods.

2 Notations and backgrounds

We denote by D_{train} a training set containing N samples, namely $D_{train} = \{x_i, y_i | i = 1, ..., N\}$, where $x_i \in \mathbb{R}^I$ indicates an input sample (or natural sample) and $y_i \in \mathbb{R}^O$ denotes its corresponding label (with I and O representing the dimension of the input space and the output space, respectively). We also define $B(x_i, \epsilon)$ as an I-dimensional small ball around each x_i with the radius ϵ.

Given a specific type of DNN, we let $f(x, \theta) : R^I \rightarrow R^O$ denote its mapping function (implicitly or explicitly), $L(x, y, \theta)$ be the loss function used by the DNN, and θ be a set of parameters which is to be optimized over D_{train} for the DNN. For simplicity, $L(x, y, \theta)$ could be written in short as $L(x, \theta)$ or even $L(x)$, so do some other similar notations. Moreover, we assume in this paper that the last layer of the DNN be a softmax layer, but it should be noted that other functions can also be used.

2.1 Adversarial training with the l_2 norm constraint

The adversarial training method with the l_2 norm constraint (AT) is a supervised method, which attempts to find the worst perturbed example in the neighborhood of a natural example to mislead the classification. Such perturbed examples are then augmented into the training set for training a better DNN. The objective of this adversarial training method can be written as:

$$\min_{\theta} \max_{x \in B(x_0, \epsilon)} L(x, y_0, \theta), \tag{1}$$

where x indicates the perturbed version of a natural example x_0 (with the label y_0) within a small neighborhood $B(x_0, \epsilon)$ (which is defined by $\|x - x_0\|_2 \leq \epsilon$).

2.2 Virtual adversarial training

Different from the adversarial training method with the l_2 norm constraint (AT), Virtual Adversarial Training (VAT) does not require the label information. It tends to find the worst perturbed example near a natural example so that the output of DNN $f(x, \theta)$ can be altered. The corresponding objective is defined as:

$$\min_{\theta} \max_{x \in B(x_0, \epsilon)} D(f(x_0, \theta), f(x, \theta)), \tag{2}$$

where $D(f(x_0, \theta), f(x, \theta))$ denotes the divergence between the outputs $f(x_0, \theta)$ and $f(x, \theta)$. For simplicity, $D(f(x_0, \theta), f(x, \theta))$ is defined in this paper as the Euclidean distance between the outputs, i.e., $\|f(x_0, \theta) - f(x, \theta)\|_2$, but it is straightforward to extend the Euclidean distance to other divergence measures.

3 Main methodology

We first present a reasonable assumption.

Assumption 1: Given a sensible loss function $L(x, y, \theta) : R^I \rightarrow R$ for a specific learning task, we assume that, there exists a small threshold σ_{th}, such that those inputs x satisfying $L(x, y, \theta) \leq \sigma_{th}$ can be correctly classified.

Note that such an assumption generally holds for common loss functions such as the cross entropy and the square error. A detailed analysis on the assumption can be seen in the appendix of the supplementary materials. With the above notations and assumptions, the adversarial training problem can be described as follows.

Problem Formulation: Assume that a natural example x_0 satisfies $L(x_0, y_0, \theta) \leq \sigma_1$ where $\sigma_1 << \sigma_{th}$, i.e., the example x_0 can be classified correctly with a high confidence. An adversarial example x_{ad} is then defined as the worst perturbed sample within $B(x_0, \epsilon)$, the small neighborhood of x_0, such that $L(x_{ad}, y_0, \theta) > \sigma_{th}$, i.e., x_{ad} will be mis-classified. The objective of adversarial training for a specific x_0 can be reformulated as

$$\min_{\theta} \max_{x \in B(x_0, \epsilon)} |L(x, y_0, \theta) - L(x_0, y_0, \theta)|.$$

3.1 Robustness against adversarial examples

Before we interpret our robustness analysis against adversarial examples, we set out Lemma 1 as follows:

Lemma 1 *Given a natural example x_0 satisfying $L(x_0, y_0, \theta) \leq \sigma_1$ (where $0 \leq \sigma_1 << \sigma_{th}$), if $\forall x \in B(x_0, \epsilon), \exists \sigma_2 : 0 \leq \sigma_2 \leq \sigma_{th} - \sigma_1$, it holds that*

$$|L(x, y_0, \theta) - L(x_0, y_0, \theta)| \leq \sigma_2 \ , \tag{3}$$

then, all the data points in $B(x_0, \epsilon)$ can be classified correctly.

The proof is provided in the appendix.

Lemma 1 states that, if the loss of data points nearby x_0 is sufficiently close to that of x_0, then all these data points can be classified correctly, since the natural example x_0 has been already classified correctly with a high confidence. In other words, whether the nearby points around x_0 can be classified correctly is affected by the stability of the loss function $L(x, y, \theta)$ in the region $B(x_0, \epsilon)$. We also say that $L(x, y, \theta)$ is robust in the region $B(x_0, \epsilon)$, and thus there exist no adversarial examples in this region, since all the data in this region are classified into the same category.

Remark Previous research studies the adversarial examples mainly through considering whether the adversarial perturbation can guide the natural example to cross the classification boundary in a less rigorous way. Moreover, it would be difficult to investigate the shape of the classification boundary when data lie in a high dimensional space. In comparison, we consider in this paper the robustness against adversarial examples from the perspective of the loss function stability, which would lead to strict analysis as follows.

In order to describe the stability of $L(x, y, \theta)$ in the neighborhood of x, we propose the following novel energy function as given in Definition 1.

Definition 1 Let $L : R^I \rightarrow R$ be a differential and integral function and $B(x_0, \epsilon)$ be a small neighborhood of x_0 with radius ϵ. Then, the energy of $L(x)$ in this neighborhood is defined as:

$$E_B(\theta) = \int_B ||\nabla_x L(x, \theta)||_2 dV, \tag{4}$$

where V denotes the volume.

This energy describes a metric measuring the stability of a function, i.e., how a function would change within a small region defined by $B(x_0, \epsilon)$.

More precisely, the integral of the l_2 norm of the gradient of the loss with respect to the input x measures how the loss function changes at each point in $B(x_0, \epsilon)$. Intuitively, if the variation on each point is not large, the loss function would not change dramatically in this neighborhood of each point. This means that the loss function would be more stable. Importantly, we will prove that minimizing such energy function can guarantee the robustness for adversarial examples in $B(x_0, \epsilon)$. Before that, we provide Lemma 2.

Lemma 2 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0, and x be an arbitrary point within $B(x_0, \epsilon)$. If the value of energy $E_B(\theta) = \int_B \|\nabla_x L(x, \theta)\|_2 dV$ decreases, then the number of examples classified correctly in $B(x_0, \epsilon)$ increases. When the energy goes to zero, the number of adversarial examples in $B(x_0, \epsilon)$ goes to zero.*

Proof of Lemma 2 is provided in the appendix of the supplementary materials.

Lemma 2 shows that decreasing the energy function leads to increasing the number of points x such that $|L(x) - L(x_0)| \leq \sigma_{th}$ in $B(x_0, \epsilon)$. Namely, a more number of points in $B(x_0, \sigma)$ would be correctly classified according to Lemma 1. When the energy function is small enough, there would be no adversarial examples gradually. Therefore, this novel energy function can be used to measure the robustness against adversarial examples in $B(x_0, \epsilon)$.

3.2 New insight to traditional adversarial methods

In this subsection, using our proposed stability measure, we provide interpretations as well as new insight to the previous traditional adversarial training methods including both supervised and semi-supervised version (Adversarial Training with l_2 norm constraint and VAT). Moreover, we prove that these traditional adversarial training methods are just to minimize the lower bound of the proposed energy along the radius,

First, we set out Definition 2 to describe the notion of the energy function along the radius.

Definition 2 Let the spherical coordinate of $x \in B(x_0, \epsilon)$ be (r, ϕ) where $r \in [0, \epsilon]$ and $\phi \in [-\pi, \pi]^{I-1}$. Then, the energy along radius on ϕ is defined by

$$E_\epsilon(\phi) = \int_0^\epsilon ||\nabla_x L(r, \phi)||_2 dr. \tag{5}$$

The energy E_ϵ is defined in the spherical coordinate system and describes the total variation of the function $L(r, \phi)$ along the radius at angle ϕ. We present Lemma 3 for a further explanation.

Lemma 3 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of the natural example x_0 with label y_0 and $x_{ad} \in B(x_0, \epsilon)$ such that $|L(x_{ad}) - L(x_0)| \geq |L(x) - L(x_0)|$ for all $x \in B(x_0, \epsilon)$. Suppose that x_{ad} is on the boundary of $B(x_0, \epsilon_1)$ $(\epsilon_1 \leq \epsilon)$ and the spherical coordinate of point x_{ad} can be expressed by (ϵ_1, ϕ_1) where $\phi_1 \in [-\pi, \pi]^{I-1}$. Then, we have*

$$\int_0^\epsilon ||\nabla_x L(r, \phi_1)||_2 dr \geq |L(x_{ad}) - L(x_0)|. \tag{6}$$

Proof of this lemma is provided in the appendix.

It is easy to reformulate the adversarial training method with l_2 norm constraint (AT) as follows (Lyu et al. 2015):

$$\min_\theta \max_{x \in B(x_0, \epsilon)} |L(x, \theta) - L(x_0, \theta)| = \min_\theta |L(x_{ad}, \theta) - L(x_0, \theta)|. \tag{7}$$

Remark If we compare Eq. (7) with Inequality (6), it can be noted that this traditional adversarial training method with l_2 norm constraint (AT) is equivalent to minimizing the lower bound of the energy $E_\epsilon(\phi_1)$. Only when the adversarial example is on the boundary of $B(x_0, \epsilon)$ and the function $L(r, \phi_1)$ is monotonically increasing w.r.t r, the traditional adversarial training method can be equivalent to minimizing the energy E_ϵ itself.

Similarly, we can also prove VAT is equivalent to minimizing a lower bound of the energy along the radius at a certain angle ϕ. Before the proof, we present Lemma 4 as follows.

Lemma 4 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0 and $x_{va} \in B(x_0, \epsilon)$ such that $\|f(x_{va}) - f(x_0)\|_2 \geq \|f(x) - f(x_0)\|_2$ for all $x \in B(x_0, \epsilon)$. Suppose that x_{va} is on the boundary of $B(x_0, \epsilon_1)$ $(\epsilon_1 \leq \epsilon)$ and the spherical coordinate of point x_{va} can be expressed by (ϵ_1, ϕ_2) where $\phi_2 \in [-\pi, \pi]^{I-1}$. Then, we have*

$$\int_0^\epsilon ||\nabla_x f(r, \phi_2)||_2 dr \geq \|f(x_{va}) - f(x_0)\|_2 \tag{8}$$

(*Proof is provided in the supplementary material*).

On the other hand, we can readily reformulate the VAT as (Miyato et al. 2018):

$$\min_\theta \max_{x \in B(x_0, \epsilon)} \|f(x, \theta) - f(x_0, \theta)\|_2 = \min_\theta \|f(x_{va}, \theta) - f(x_0, \theta)\|_2 \tag{9}$$

Remark If we compare Eq. (9) with Inequality (8), it can be noted that VAT is equivalent to minimizing the lower bound of energy $E_\epsilon(\phi_2)$. In implementing virtual adversarial training (VAT), there are two versions of the loss function: mean square error loss (MSE loss)

and KL-divergence loss. We mainly consider the MSE loss in this paper, which facilitates the proof that the energy function can bound the stability of the loss function. When we consider the KL-divergence loss which contains uncertainty information, it may not guarantee (or at least it is difficult to prove) the theoretical bound for the energy function.

Similarly, only when the adversarial example is on the boundary of $B(x_0, \epsilon)$ and the function $f(r, \phi_2)$ monotonically increases, the VAT would be exactly equivalent to minimizing the energy $E_\epsilon(\phi_2)$.

Now we start to introduce our proposed method and we first present Theorem 5.

Theorem 5 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0 and x be an arbitrary point in $B(x_0, \epsilon)$. If the value of the energy $E_B(\theta) = \int_B \|\nabla_x L(x, \theta)\|_2 dV$ decreases, the value of the energy $E_\epsilon(\phi, \theta) = \int_0^\epsilon ||\nabla_x F(r, \phi, \theta)||_2 dr$ decreases almost everywhere in $[-\pi, \pi]^{I-1}$. When the energy $E_B(\theta)$ goes to zero, the energy $E_\epsilon(\phi, \theta)$ goes to zero almost everywhere in $[-\pi, \pi]^{I-1}$.*

Proof of this theorem is provided in the appendix.

In the above, for a measurable set E, we say that a property holds **almost everywhere** on E, or it holds for almost all $x \in E$, provided there is a subset E_0 of E for which $m(E_0) = 0$ ($m(E_0)$ denotes the measure for E_0) and the property holds for all $x \in E - E_0$.

Remark In this paper, we mainly establish the energy function with ℓ_2-norm. When the energy function is established with ℓ_2-norm, the stability of the loss function can be bounded. However, with ℓ_∞-norm, we find that it would be difficult to bound such stability. Therefore, it is theoretically more appealing and more direct to minimize the energy function with ℓ_2-norm for improving the stability of the loss function. Additionally, in the experimental part, we actually evaluate our method with ℓ_∞-norm attack. The results show that our method outperforms the baseline method on such attack. It means that our method can also work on ℓ_∞-norm attacks. The investigation about the energy function with ℓ_∞ -norm can be viewed as future work.

Theorem 5 states that decreasing the total energy E_B can lead to a reduction of the energy along the radius E_ϵ. Therefore, we can reduce all of E_ϵ by penalizing the total energy E_B. According to Lemma 1 and Theorem 5, it is natural to propose a new method:

$$\min_\theta \int_B ||\nabla_x L(x, y, \theta)||_2 dV. \tag{10}$$

Intuitively, optimizing the total variations of the function $L(x, y, \theta)$ can help avoid the dramatic fluctuation of the loss function.

However, the optimization problem (10) is impractical due to the integration. For convenient optimization, we approximate the energy function with its lower and upper bound and reformulate the problem (10) as:

$$\min_\theta \left[\max_{x \in B} L(x, \theta) + \lambda \max_{x \in B} ||\nabla_x L(x, y, \theta)||_2 \right] \tag{11}$$

where the first term and the second term can be viewed as the lower and upper bound of energy function. Here we optimize the upper and lower bound of energy function rather

than itself. λ is a positive trade-off hyper-parameter. This method can be also extended to VAT:

$$\min_{\theta}\left[\max_{x\in B} D(f(x_i,\theta),f(x_i+\epsilon_{vat},\theta))+\lambda\max_{x\in B}||\nabla_x f(x,\theta)||_2\right] \tag{12}$$

Relevant proof and details can be seen in the appendix.

3.3 Robust generalization

In this section, we analyze the generalization for our proposed frame and prove our proposed method can achieve a better generalization bound than the traditional adversarial training methods. We first introduce relevant notations and definitions. The ϵ-neighborhood of a training set is defined as $\mathcal{M}_t=\bigcup_{x_i\in D_{train}} B(x_i,\epsilon)$ and the whole set of natural examples $\mathcal{M}=\bigcup_{x\in D} B(x,\epsilon)$. ρ_t and ρ are the probability density supported on an m-dimensional manifolds $\mathcal{M}_t$ and $\mathcal{M}$ respectively. Then, the robust generalization can be defined as the difference between the expected loss over the training set and the whole data set: $|\mathbb{E}_{x\backsim\rho_t}L(x)-\mathbb{E}_{x\backsim\rho}L(x)|$. When the difference is small, it means that the model can perform similarly well on unseen data on the training set. We now provide the upper bound of robust generalization as Theorem 13.

Theorem 6 *Suppose that* $\inf_{\mathcal{M}}\rho>0$, $dim(\mathcal{M})=m$ *and* $\|\nabla_x\|\nabla_x L\|_2\|_2\le K$ *for* $x\in\mathcal{M}$. *Then for any* $t>0$ *and proper constants* C_1 *and* C_2, *we have*

$$|\frac{1}{n}\sum_{i=1}^{n}L(x_i')-\int_{\mathcal{M}}L\rho dVol(x)|)\le C_1(\max_{x\in\mathcal{M}_t}\|\nabla_x L\|_2+C_2K)(\frac{tlog(n)}{n})^{\frac{1}{m+2}} \tag{13}$$

with the probability at least $1-2t^{-\frac{m}{m+2}}n^{-(ct-1)}$, *where* $x_i'\in\mathcal{M}_t$.

According to Theorem 13, the upper bound of the robust generalization is decided by $\max_{x\in\mathcal{M}_t}\|\nabla_x L\|_2$,[1] which is exactly minimized in our framework as seen in (11). In comparison, the other traditional adversarial training methods did not minimize such a term, leading that the associated generalization upper bound would be looser than our proposed framework.

3.4 Practical optimization algorithm

We design practical optimization algorithms for our proposed new framework, which basically extends the previous methods with the novel energy regularization. For convenience, we start with the problem (11), while the problem (12) can be solved in a similar way. In the problem (11), the first term can be solved with the traditional adversarial training method. The second term can be divided into two problems: inner maximization problem and outer minimization problem. However, for the inner problem, since $\|\nabla_x L(x,y,\theta)\|_2$ is a non-convex function, it is difficult to evaluate the maximizer of function $\|\nabla_x L(x,y,\theta)\|_2$. Following many similar approaches (Lyu et al. 2015), we relax it to the convex problem with the first order Taylor expansions:

[1] All the other terms in the right-hand side are constants.

$$\max_{x \in B(x_0,\epsilon)} ||\nabla_x L(x_0, y_0, \theta)||_2 + \nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2^T (x - x_0). \tag{14}$$

The problem (14) is now a convex problem w.r.t x and can be solved by Lagrangian multiplier method. The maximizer can be calculated as:

$$x_{max} = \epsilon \overline{\nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2} + x_0, \tag{15}$$

where $\div$ represents the normalized operator. The gradient of $\|\nabla_x L(x_0, y_0, \theta)\|_2$ w.r.t x is difficult to compute. We can use the finite difference method to approximate it:

$$\begin{aligned} x_{max} =& \epsilon \overline{\nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2} + x_0 \\ =& \epsilon \overline{H(x_0) \nabla_x L(x_0)} + x_0 \\ \approx& \epsilon \overline{\frac{\nabla_x L(x_0 + \xi \nabla_x L(x_0)) - \nabla_x L(x_0)}{\xi}} + x_0, \end{aligned} \tag{16}$$

where ξ is a small value. In this paper, we set $\xi = 10^{-6}$. More details of derivation of (16) are provided in the appendix. After computing the maximizer x_{max} of inner problem, the outer problem can be solved by the gradient decent method. The whole algorithm of Adversarial Training with Energy Regularization we called in short ATER is shown in Algorithm 1. We also develop the VAT with Energy Regularization (in short VATER).

Algorithm 1 Algorithm for ATER.

for number of training iterations **do**
  Sample a batch of labeled data (x_i, y_i) with size N.
  for i in $1...N$ **do**
    compute adversarial examples x^i_{adv} with PGD
    $d^e_i \leftarrow \overline{\frac{\nabla_x L(x_i + \xi \nabla_x f(x_i)) - \nabla_x L(x_i)}{\xi}}$
    $x^i_{max} = \epsilon d^e_i + x_i$
  end for
  Update the parameters of neural network with stochastic gradient:
  $-\nabla_\theta \frac{1}{N} \sum_{i=1}^N log\mathcal{L}(y_i, x_i, \theta) - \nabla_\theta \frac{1}{N} \sum_{i=1}^N log\mathcal{L}(y_i, x^i_{adv}, \theta) - \nabla_x \frac{1}{N} \sum_{i=1}^N \|\nabla_\theta L(x^i_{max}, y_i, \theta)\|_2$
end for

Algorithm 2 Algorithm for VATER.

for number of training iterations **do**
  Sample a batch of labeled data (x_l, y_l) with size N_l and a batch of unlabeled data (x_{ul}) with size N_{ul}. Z_{l_i} denotes the matrix of latent features of data with label i in a batch.
  for j in $1...n$ **do**
    $d \leftarrow \overline{\nabla_x D(f(x_i, \theta), f(x_i + r, \theta))|_{\epsilon=\xi d}}$
  end for
  $\epsilon_{vat} = \xi d$
  $d_i^e \leftarrow \frac{\nabla_x f(x_i + \xi \nabla_x f(x_i)) - \nabla_x f(x_i)}{\xi}$
  $x_{max}^i = \epsilon d_i^e + x_i$
  Update the parameters of neural network with stochastic gradient:

$$-\nabla_\theta \{ \frac{1}{N} \sum_{i=1}^{N} L(x_i, y_i, \theta) - D(f(x_i, \theta), f(x_i + \epsilon_{vat}, \theta)) - \lambda \| \nabla_x f(x_{max}^i, \theta) \|_2 \}$$

end for

4 Experiments

We evaluate the robustness of our proposed method Adversarial Training with Energy Regularization (ATER) on datasets CIFAR-10, CIFAR-100 and SVHN. We also extend the Virtual Adversarial Training (VAT) model with the proposed energy regularization termed as VATER in the setting of semi-supervised learning (SSL) to check further the generalization. Specifically, taking the wide resnet (Zagoruyko and Komodakis 2016) as the baseline model, we mainly conduct the experiment of ATER to examine the robustness, i.e., the accuracy on different adversarial attacks. We further apply large-covnet with the same setting as Miyato et al. (2018) to check how well the VATER performs in the semi-supervised setting on CIFAR10 and SVHN. We set the hyper parameter λ in Eq. (11) (ATER), to 1.0 for both CIFAR-10 and CIFAR-100 and 10.0 for SVHN empirically, which were the best one chosen from $\{1000.0, 100.0, 10.0, 1.0, 0.1\}$. For the hyper parameter λ in Eq. (12) (VATER), we set it 0.1 for both CIFAR-10 and MNIST chosen from $\{0.1, 0.2, 0.5, 0.7, 1.0\}$. For the training time, we generate the adversarial examples for training through PGD method with 20 iteration and attack radius is 8/255. The training epoch is 60 and the learning rate is 0.1 for CIFAR-10 and CIFAR-100 and 0.01 for SVHN. For the test time, we generate the adversarial examples through different attack methods including PGD, CW and FGSM with different attack iterations 10, 20, 40, 100. The attack radius is also 8/255.

4.1 Performance on adversarial examples

In this section, we compare our proposed ATER with the recent adversarial training methods on different datasets for defending adversarial attacks. Tables 1, 2 and 3 show the performance of different models on CIFAR-10, CIFAR-100 and SVHN respectively under different adversarial attacks including FGSM, PGD and CW. In both the tables, it is noted that except AT and our proposed ATER, all the other results were copied from the related work (Zhang et al. 2019; Song et al. 2019; Madry et al. 2017; Kannan et al. 2018; Mao et al. 2019). Moreover, for the PGD and CW attacks, adversarial examples are generated with different iterations for fair comparisons, i.e., 10, 20, 40, 100 iterations. We compare our proposed ATER with several recent competitive methods such as

Table 1 Accuracy under different attacks on CIFAR-10

Methods	Clean	FGSM	PGD10	PGD20	PGD40	PGD100	CW10	CW20	CW40	CW100
Original	**95.6**	36.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AT (Madry et al. 2017)	85.7	54.9	45.1	44.9	44.8	44.8	45.9	45.7	45.6	45.4
TLA (Mao et al. 2019)	86.21	58.88	53.87	51.59	–	–	–	–	–	–
TRADES (Song et al. 2019)	80.35	–	–	50.95	–	–	–	49.8	–	–
RLFAT-p (Song et al. 2019)	84.77	–	–	53.97	–	–	–	52.40	–	–
ATER	86.72	**68.42**	**56.7**	**55.75**	**55.22**	**55.12**	**54.71**	**54.42**	**53.47**	**53.38**

Table 2 Accuracy under different attacks on CIFAR-100

Methods	CIFAR-100					
	Clean	FGSM	PGD20	PGD100	CW20	CW100
Original	**79.0**	10.0	0.0	0.0	0.0	0.0
AT (Madry et al. 2017)	59.9	28.5	22.6	22.3	23.2	23.0
TRADES (Song et al. 2019)	52.13	–	27.26	–	24.66	–
RLFAT-p (Song et al. 2019)	56.70	–	31.99	–	29.04	–
ATER	61.47	**42.02**	**33.75**	**33.52**	**29.94**	**28.62**

Table 3 Accuracy under different attacks on SVHN

Methods	SVHN					
	Clean	FGSM	PGD20	PGD100	CW20	CW100
Original	**97.2**	53.0	0.3	0.1	0.3	0.1
AT	93.9	68.4	47.9	46.0	48.7	47.3
ALP (Kannan et al. 2018)	96.2	–	–	46.9	–	–
ATER	94.79	**72.47**	**54.36**	**52.82**	**50.23**	**49.92**

Table 4 Accuracy under black-box attack on CIFAR-10

Defense models	Attacked models								
	Vanilla training			Adversarial training			Ours		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
AT	84.62	84.89	84.83	72.20	63.77	63.27	68.94	61.89	60.31
Ours	**86.64**	**87.27**	**87.32**	**74.05**	**68.63**	**67.72**	**71.99**	**65.08**	**64.33**

AT, TLA, TRADES, and RLFAT$_p$. It can be observed that our proposed method ATER consistently achieves the best performance on these three datasets over all the other methods. This validates the necessities of applying the proposed energy regularization, which guarantees a better generalization as proved in the paper.

To further verify the robustness of ATER, we conduct transfer-based black-box attack experiments on CIFAR-10. Three different models are used for generating attacks including the Vanilla Training model, the Adversarial Training and our model. As demonstrated by the results in Table 4, our proposed approach can achieve better performance in all the cases.

Finally, taking CIFAR-10 as one illustrative example, we examine our proposed method on stronger attack, Auto-Attack (AA) in Table 5. We also report experimental results on large datasets, i.e. Imagenet and Tiny - Imagenet under PGD-LL attack (LL means attacking least likely class) in Table 6. As can be seen in these two tables, our proposed method AFTER outperforms the baseline method AT on both stronger attack and large datasets.

Table 5 The performance of different methods on CIFAR-10 under Auto-Attack (AA)

Method	CIFAR-10 AA
AT-Free (Shafahi et al. 2019)	41.47
ATES (Sitawarin et al. 2020)	50.72
TLA (Mao et al. 2019)	47.41
AT	44.04
ATER	**53.82**

Table 6 The performance of ATER on large datasets under PGD-LL attack

Defense models	Tiny-imagenet		Imagenet	
	Clean	PGD-LL	clean	PGD-LL
AT	70.2	39.2	53.1	25.1
ALP (Kannan et al. 2018)	72.0	41.3	**55.7**	27.9
ATER	**72.6**	**43.6**	55.2	**29.8**

4.2 Further analysis

We conduct some further analysis to validate the effectiveness of our proposed method in this section.

We first examine the performance of the proposed ATER on the adversarial examples with different budgets. Specifically, we generate in the test datasets of CIFAR-10 10,000 adversarial examples according to PGD with 20 iterations. We increase the level of adversarial noises gradually from 4 to 20 in CIFAR-10 (with the step size as 4). We then test the performance of ATER and the traditional adversarial training (AT) on these adversarial examples. These results are plotted in Fig. 2. As clearly observed, the proposed ATER shows much better robustness against adversarial examples. Particularly, when the adversarial noises are heavier, the proposed ATER still demonstrates clearly better performance, verifying their significant robustness.

Next, we show that our proposed ATER can indeed obtain a smaller energy compared with baseline method AT and the original wide resnet. Here, we approximate the energy with its upper bound: the maximum gradient within the ϵ-neighborhood of the training set $\max_{x\in B(x_0,\epsilon)} ||\nabla_x L(x, y, \theta)||_2$. Specifically, we search the norm of the largest gradient within the 6 steps for 10 different batches of training data. Then we average the norm of gradient at the last step over 10 batches. These results are plotted in Fig. 3. It can be noted that our proposed ATER obtains a much lower maximum gradient, thereby leading to a stable model with a better generalization according to Theorem 13. We also plot the distribution for the L_2-norm of the gradient of loss with respect to inputs on different datasets (CIFAR-10, CIFAR-100 and SVHN). As shown in Fig. 4, more gradients of our method distribute at very low value than the baseline method, indicating that our proposed method can lead to a better stability around more data points.

Finally, we conduct the sensitive analysis for hyper parameter λ. Specifically, we plot the performance (robust accuracy) of the proposed method ATER on PGD20 attack with different λ: $\{1000, 100, 10, 1.0, 0.1\}$ in Fig. 1. As can be seen, our proposed method

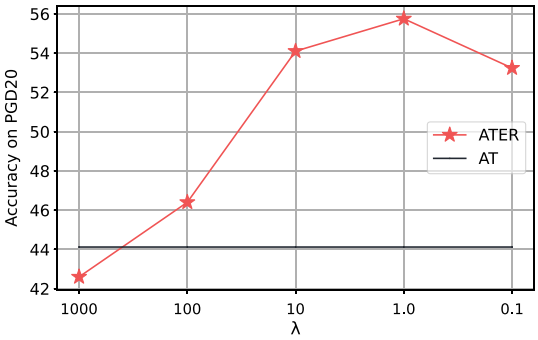

Fig. 1 Performance of ATER under PGD20 attack on CIFAR-10 with different λ

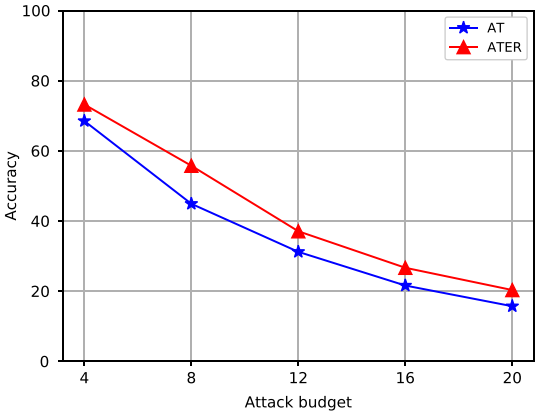

Fig. 2 Performance on adversarial examples of CIFAR-10 with different attack budgets. Better viewed in color

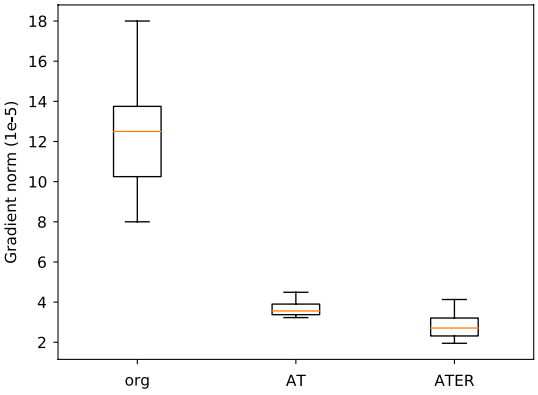

Fig. 3 The norm of the maximum gradient within the $B(x, \epsilon)$ of natural examples on CIFAR10

ATER achieves the best performance when $\lambda = 1.0$. Except for $\lambda = 1000$, the proposed method ATER outperforms the baseline AT on all the other values of λ.

Table 7 Test performance on SVHN in semi-supervised learning

Method	SVHN (1000 labeled) Test error rate (%)
SWWAE (Zhao et al. 2015)	23.56
Skip generative model (Maaløe et al. 2016)	16.30
GAN with feature matching (Salimans et al. 2016)	8.11
Π model (Laine and Aila 2016)	5.43
RPT	8.41(±0.24)
VAT	5.77(±0.32)
VATER	**4.92**(±**0.10**)

Table 8 Test performance on CIFAR-10 in semi-supervis'ed learning

Method	CIFAR-10 (4000 labeled) Test error rate (%)
Ladder networks, Γ model (Rasmus et al. 2015)	20.40
CatGAN (Springenberg 2015)	19.58
GAN with feature matching (Salimans et al. 2016)	18.63
Π model (Laine and Aila 2016)	16.55
RPT	18.56(±0.29)
VAT	14.82(±0.38)
VATER	**12**.53(±0.23)

4.3 Performance on semi-supervised learning

We also conduct experiments on semi-supervised learning to further validate the proposed general energy regularization term. Specifically, we extend VAT to the VATER model in the semi-supervised setting. VAT basically takes an unsupervised adversarial training strategy and can be readily used in SSL. It is noted that AT is a supervised learning method and hence is not implemented in this evaluation. Tables 7 and 8 demonstrate the results of different models on SVHN and CIFAR-10 in this scenario. As observed, on both the datasets, our model attains the best performance, which is significantly more accurate than VAT and all the other algorithms. This shows that the proposed method could truly lead to a better generalization especially than VAT in SSL.

5 Conclusion

In this paper, we investigate the model robustness against adversarial examples from the perspective of the function stability. We develop a novel energy function to describe the stability in the small neighborhood of natural examples and prove that reducing such energy can guarantee the robustness for adversarial examples. We also offer new insights to traditional adversarial methods (AT and VAT) showing that such traditional

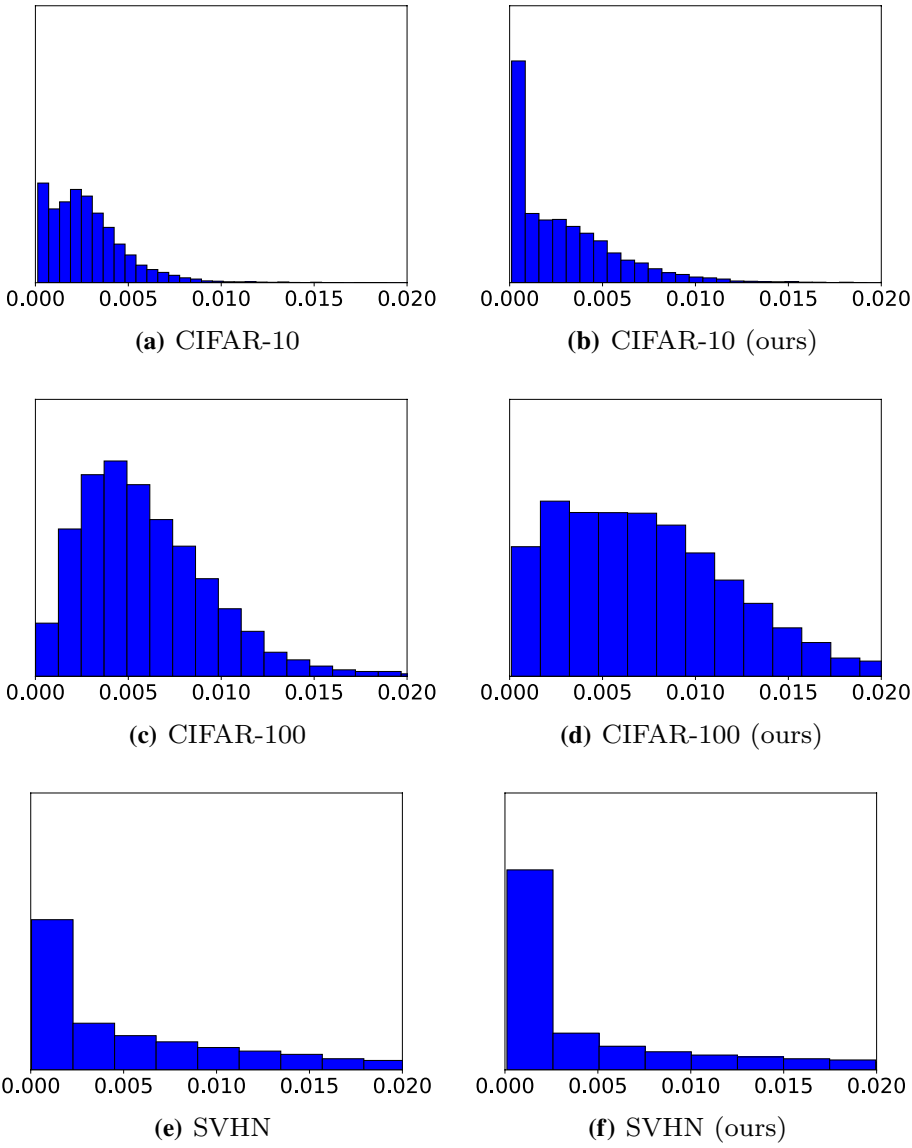

Fig. 4 Distribution for the L_2-norm of the gradient of loss with respect to inputs on CIFAR-10, CIFAR-100 and SVHN (the first, second, and third columns respectively)

methods merely decrease certain lower bounds of the energy function. We analyze the disadvantage of the traditional methods and propose accordingly more rational methods to minimize both the upper bound and lower bound of the energy function. We implement our methods on both supervised and semi-supervised tasks and achieve superior performance on benchmark datasets.

Appendix

Analysis for Assumption 1

In this section, we prove that for common loss functions (e.g., cross entropy and square error) we can find a small constant σ_{th} such that if $L(x, y, \theta) < \sigma_{th}$, then the input x can be classified correctly. In this paper, we assume the last layer the softmax layer, then we have $\sum_i y_i = 1$ and $y_i \in [0, 1]$.

Cross entropy loss

The cross entropy loss is defined as $L_{ce} = -\sum_i l_i log(y_i)$ where $l = \{l_i\}_{i=1}^{O}$ (O is the output dimension) is an one hot label vector for input x_1. We assume $l_a = 1$ and others are zeros which means x belongs to class a. Then, we can reformulate the cross entropy loss as $L_{ce} = -logy_a$. If $L_{ce}(x_1, l) < \sigma_{th}$ for all $0 < \sigma_{th} < -log0.5$, x_1 can be classified correctly.

$$L_{ce} < \sigma_{th} < -log0.5 \tag{17}$$

Proof
then we have

$$-logy_a < \sigma_{th} < -log0.5 \Rightarrow y_a > e^{-\sigma_{th}} > 0.5 \tag{18}$$

Since $\sum_i y_i = 1$ and $y_a > e^{-\sigma_{th}} > 0.5$, x_1 can be classified correctly. □

Square error loss

The square error loss can be formulated as $L_{se} = \sum_i (y_i - l_i)^2$. Similarly, for the square error loss L_{se}, If $L_{se}(x_1, l) < \sigma_{th}$ for all $\sigma_{th} < 0.25$, x_1 can be classified correctly.

$$\sum_i (y_i - l_i)^2 < \sigma_{th} \tag{19}$$

Proof
Since $l_a = 1$ and others are zeros, then we have

$$\sum_{i/a} (y_i)^2 + (y_a - 1)^2 < \sigma_{th} \Rightarrow (y_a - 1)^2 < \sigma_{th} < 0.25 \Rightarrow y_a > \sigma_{th} > 0.5 \tag{20}$$

Since $\sum_i y_i = 1$ and $y_a > 1 - \sqrt{\sigma_{th}} > 0.5$, x_1 can be classified correctly. □

Proof for lemmas and theorem

In this section, we prove the lemmas and theorems in the main submission. For convenience, we first set out Theorem A.1 and Lemma A.2 which will be used in the proof.

Theorem A.1 *Let $B \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are integrable and continues functions, then there exists a constant c such that*

$$\int_B f(x)g(x)dV = c\int_B g(x)dV \tag{21}$$

where V is the volume and $\int_B g(x)dV \neq 0$.

Proof We can directly find constant c:

$$\frac{\int_B f(x)g(x)dV}{\int_B g(x)dV} = c \tag{22}$$

□

Lemma A.2 *Let us define* $g : \mathbb{R}^m \to \mathbb{R}$ *by* $g(\theta) = \int_B f(x,\theta)dV$ *where* $f : \mathbb{R}^I \to \mathbb{R}$ *is differential and integrable. Then, if* $g(\theta)$ *decreases and goes to zero,* $f(x,\theta)$ *decreases and goes to zero almost everywhere in B.*

(*Def. For a measurable set E, we say that a property holds* **almost everywhere** *on E, or it holds for almost all* $x \in E$*, provided there is a subset* E_0 *of E for which* $m(E_0) = 0$ ($m(E_0)$ *denotes the measure for* E_0) *and the property holds for all* $x \in E - E_0$).

Proof Let $N_0 = \{x | f(x) \geq n_0\}$ where n_0 is a sufficiently small value.

$$\begin{aligned} g(\theta) &= \int_B f(x,\theta)dV \\ &= \int_{N_0} f(x,\theta)dV + \int_{B-N_0} f(x,\theta)dV \\ &\geq \int_{N_0} f(x,\theta)dV \end{aligned} \tag{23}$$

When $g(\theta)$ decreases and goes to zero, $\int_{N_0} f(x,\theta)dV$ decreases and goes to zero. Then the measure of N_0 ($m(\{x|f(x) \geq n_0\})$) decreases and goes to zero which means the measure of $B - N_0$ ($\mathrm{m}(\{x|f(x) < n_0\})$) increases and goes to $m(B)$. Therefore, $f(x,\theta)$ decreases and goes to zero almost everywhere in B. □

Proof for Lemma 1

Lemma 1 *Given a natural example* x_0 *satisfying* $L(x_0, y_0, \theta) \leq \sigma_1$ (*where* $0 \leq \sigma_1 << \sigma_{th}$), *if* $\forall x \in B(x_0, \epsilon), \exists \sigma_2 : 0 \leq \sigma_2 \leq \sigma_{th} - \sigma_1$, *it holds that*

$$|L(x, y_0, \theta) - L(x_0, y_0, \theta)| \leq \sigma_2 \ , \tag{24}$$

then, all the data points in $B(x_0, \epsilon)$ *can be classified correctly.*

Proof In this paper, we have proved in the previous section that there exists a σ_{th} such that if $L(x, y, \theta) \leq \sigma_{th}$, x can be classified correctly. Additionally, we assume that the natural examples can be classified correctly with a high confidence ($L(x_0, y_0, \theta) \leq \sigma_1 << \sigma_{th}$). Then, if $L(x, y_0, \theta) < L(x_0, y_0, \theta)$,

$$L(x, y_0, \theta) < \sigma_1 < \sigma_{th} \tag{25}$$

which means x can be classified correctly.

If $L(x, y_0, \theta) > L(x_0, y_0, \theta)$

$$|L(x, y_0, \theta) - L(x_0, y_0, \theta)| = L(x_0, y_0, \theta) - L(x_0, y_0, \theta) < \sigma_2 < \sigma_{th} - \sigma_1 \Rightarrow L(x_0, y_0, \theta) < L(x_0, y_0, \theta) + \sigma_2 < \sigma_{th} \tag{26}$$

Therefore, if $|L(x_0, y_0, \theta) - L(x_0, y_0, \theta)| \leq \sigma_2$, x_0 can be classified correctly. □

Proof for Lemma 3 and Lemma 4

Here, we just prove Lemma 4 since Lemma 3 is a special case of Lemma 4.

Lemma 4 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0 and $x_{va} \in B(x_0, \epsilon)$ such that $\|f(x_{va}) - f(x_0)\|_2 \geq \|f(x) - f(x_0)\|_2$ for all $x \in B(x_0, \epsilon)$. Suppose that x_{va} is on the boundary of $B(x_0, \epsilon_1)$ $(\epsilon_1 \leq \epsilon)$ and the spherical coordinate of point x_{va} can be expressed by (ϵ_1, ϕ_2) where $\phi_2 \in [-\pi, \pi]^{I-1}$. Then, we have*

$$\int_0^{\epsilon} ||\nabla_x f(r, \phi_2)||_2 dr \geq \|f(x_{va}) - f(x_0)\|_2 \tag{27}$$

$$\begin{aligned} \int_0^{\epsilon} ||\nabla_x f(r, \phi_2)||_2 dr &\geq \int_0^{\epsilon_1} ||\nabla_x f(r, \phi_2)||_2 dr \\ &\geq \int_0^{\epsilon_1} ||\nabla_x f(r, \phi_2) \cdot \vec{d}||_2 dr \\ &\geq || \int_0^{\epsilon_1} \nabla_x f(r, \phi_2) \cdot \vec{d} dr||_2 \\ &= \|f(x_{va}) - f(x_0)\|_2 \end{aligned} \tag{28}$$

Proof

where, $\vec{d}$ is the unit vector pointing from x_0 to x_{va}. In the same way, we can prove Lemma 3. □

Proof for Lemma 2

Lemma 2 *Let $B(x_0, \epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0 and x_{ar} be arbitrary point in $B(x_0, \epsilon)$. If the value of energy $E_B(\theta) = \int_B \|\nabla_x L(x, \theta)\|_2 dV$ decreases, the number of examples classified correctly in $B(x_0, \epsilon)$ increases. When the energy goes to zero, the number of adversarial examples in $B(x_0, \epsilon)$ goes to zero.*

Proof we reformulate the energy in spherical coordinate:

$$E_B = \int_B ||\nabla_x L(x)||_2 dV = \int_{S^{I-1}} \int_0^{\epsilon} ||\nabla_x L(r, \phi)||_2 r^{I-1} dr d\phi \tag{29}$$

According to Theorem A.1, there exists a constant r_1 such that

$$\int_B ||\nabla_x L(r,\phi)||_2 dV = r_1 \int_{S^{I-1}} \int_0^{\epsilon} ||\nabla_x L(r,\phi)||_2 dr d\phi \tag{30}$$

According to Lemma 3.4, we have

$$r_1 \int_{S^{I-1}} \int_0^{\epsilon} ||\nabla_x L(r,\phi)||_2 dr d\phi \geq r_1 \int_{S^{I-1}} |L(\epsilon_1,\phi) - L(x_0)| d\phi \tag{31}$$

where $\epsilon_1 \leq \epsilon$ and (ϵ_1, ϕ) is the spherical coordinate of arbitrary point x. Since E_B is the upper bound of $\int_{S^{I-1}} |L(\epsilon_1,\phi) - L(x_0)| d\phi$, when E_B decreases and goes to zero, $\int_{S^{I-1}} |L(\epsilon_1,\phi) - L(x_0)| d\phi$ decreases and goes to zero. According to Lemma A.2, for almost all $x_{ar} \in B$, $|L(x_{ar}) - L(x_0)|$ decreases and goes to zero which means the number of adversarial examples in B decreases and goes to zero (according to Lemma 1). □

Proof for Theorem 5

Theorem 5 *Let $B(x_0,\epsilon) \in \mathbb{R}^I$ be a small neighborhood of natural example x_0 with label y_0. and x_{ar} be arbitrary point in $B(x_0,\epsilon)$. If the value of energy $E_B(\theta) = \int_B \|\nabla_x L(x,\theta)\|_2 dV$ decreases, the value of energy $E_\epsilon(\phi,\theta) = \int_0^{\epsilon} ||\nabla_x F(r,\phi,\theta)||_2 dr$ decreases almost everywhere in $[-\pi,\pi]^{I-1}$. When the energy $E_B(\theta)$ goes to zero, the energy $E_\epsilon(\phi,\theta)$ goes to zero almost everywhere in $[-\pi,\pi]^{I-1}$.*

Proof Similar to Lemma A.1, there exists a constant r_1 such that

$$\int_B ||\nabla_x L(r,\phi)||_2 dV = r_1 \int_{S^{I-1}} \int_0^{\epsilon} ||\nabla_x L(r,\phi)||_2 dr d\phi = r_1 \int_{S^{I-1}} E_\epsilon(\phi) d\phi \tag{32}$$

According to Lemma A.2, when $E_B(\theta)$ decreases and goes to zero, for almost all $\phi \in [-\pi,\pi]^{I-1}$, $E_\epsilon = \int_0^{\epsilon} ||\nabla_x F(r,\phi)||_2 dr$ deceases and goes to zero. □

Proof for Theorem 6

Theorem 6 *Suppose that $\inf_{\mathcal{M}} \rho > 0$, $dim(\mathcal{M}) = m$ and $\|\nabla_x \|\nabla_x L\|_2\|_2 \leq K$ for $x \in \mathcal{M}$. Then for any $t > 0$ and proper constants C_1 and C_2, we have*

$$|\frac{1}{n}\sum_{i=1}^{n} L(x_i') - \int_{\mathcal{M}} L\rho dVol(x)| \leq C_1 (\max_{x \in \mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K)(\frac{tlog(n)}{n})^{\frac{1}{m+2}} \tag{33}$$

with probability at least $1 - 2t^{-\frac{m}{m+2}} n^{-(ct-1)}$, where $x_i' \in \mathcal{M}_t$.

Proof Before proving the generalization bound, we first introduce the Bernstein's inequality:

Bernstein's inequality Let $x_1, x_2, ..., x_n$ be independent bounded random variables such that $|x_i| \leq M$ with probability 1 and let $\sigma^2 = \mathbb{E}[(x_i - [x_i])^2]$. Then for any $\alpha > 0$, we have

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}[x_i]| > \alpha) \leq 2exp(-\frac{n\alpha^2}{2\sigma^2 + 4M\alpha/3}) \tag{34}$$

Since $\|\nabla_x \|\nabla_x L\|_2\|_2 \leq K$ for arbitrary $x \in \mathcal{M}$, we have $\max_{x\in\mathcal{M}} \|\nabla_x L\|_2 \leq \max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K$ and $\max_{x\in\mathcal{M}} |L(x)| \leq C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K)$ for proper constant C_1 and C_2.

We assume $\mathcal{M} = [0,1]^m$ and partition $\mathcal{M}$ into hyper cubes $B_1, B_2, ..., B_N$ with side length $r > 0$ and $N = r^{-m}$. Let S_j be the number of $x'_1, x'_2, ..., x'_n$ falling in B_j (x'_i is the perturbed example in $B(x_i, \epsilon)$). Then S_j is a Binomial random variable with parameters n and $p_j = \int_{B_j} \rho dx \geq cr^m$. According to Bernstein inequality, we have

$$\mathbb{P}(|\frac{1}{n}S_j - \int_{B_j} \rho dx| > \alpha) \leq 2exp(-cnh^{-m}\alpha^2) \tag{35}$$

for any j. For $0 < \alpha \leq h^m$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} L(x'_i) &\leq \frac{1}{n}\sum_{j=1}^{N} S_j \max_{B_j} L \leq \underbrace{\sum_{j=1}^{N}(\int_{B_j} \rho dx + \alpha)\max_{B_j} L}_{According\ to\ (35)} \\
&\leq \sum_{j=1}^{N} \max_{B_j} L \int_{B_j} \rho dx + C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K) r^{-m}\alpha \\
&\leq \sum_{j=1}^{N} (\min_{B_j} L + C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K) r) \int_{B_j} \rho dx \\
&\quad + C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K) r^{-m}\alpha \\
&\leq \sum_{j=1}^{N} \int_{B_j} L\rho dx + C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K) r^{-m}(\alpha + r^{m+1}) \\
&= \int_{\mathcal{M}} L\rho dx + C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K)(\alpha h^{-m} + r)
\end{aligned}$$

with probability at least $1 - 2r^{-m}exp(-cnr^{-m}\alpha^2)$.

For $\alpha = r^m$, we have

$$|\frac{1}{n}\sum_{i=1}^{n} L(x'_i) - \int_{\mathcal{M}} L\rho dVol(x)| \leq C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K) r \tag{36}$$

with probability at least $1 - 2r^{-m}exp(-cnr^{m+2})$. By selecting $nr^{m+2} = tlog(n)$

$$|\frac{1}{n}\sum_{i=1}^{n} L(x'_i) - \int_{\mathcal{M}} L\rho dVol(x)| \leq C_1(\max_{x\in\mathcal{M}_t} \|\nabla_x L\|_2 + C_2 K)(\frac{tlog(n)}{n})^{\frac{1}{m+2}} \tag{37}$$

with probability at least $1 - 2t^{-\frac{m}{m+2}} n^{-(ct-1)}$. □

Details of practical algorithm

In this paper, we minimize both the upper bound and lower bound of energy E_ϵ. The algorithm to minimize the lower bound is the same as the traditional adversarial training. Here, we only give the relevant proof and algorithm for the upper bound of E_ϵ and E_B:

The upper bound for E_B:

$$E_B = \int_B ||\nabla_x L(x)||_2 dV \leq \int_B \max_{x\in B} \|\nabla L(x)\|_2 dV = \max_{x\in B} \|\nabla L(x)\|_2 \cdot Vol(B) \tag{38}$$

The upper bound for E_ϵ:

$$\begin{aligned} E_\epsilon &= \int_0^\epsilon ||\nabla_x L(r,\phi)||_2 dr \\ &\leq \int_0^\epsilon \max_{x\in B} \|\nabla L(x)\|_2 dr \\ &= \max_{x\in B} \|\nabla L(x)\|_2 \cdot \epsilon \end{aligned} \tag{39}$$

Since $Vol(B)$ and ϵ are constants, reducing $\max_{x\in B} \|\nabla L(x)\|_2$ is equivalent to decreasing the upper bound of E_ϵ and E_B.

The problem (13) in the main paper can be reduced to:

$$\max_{\|r\|_p=\epsilon} \nabla_x \mathcal{F}^T r \tag{40}$$

where, $r = x - x_0$ and $\mathcal{F} = \|\nabla_x L(x_0, y_0, \theta)\|_2$. We solve it with the Lagrangian multiplier method and we have

$$\nabla_x \mathcal{F} r = \lambda(\|r\|_p - \epsilon)$$

Then we make the first derivative with respect to r:

$$\begin{aligned} \nabla_x \mathcal{F} &= \lambda \frac{r^{p-1}}{p(\sum_i r_i^p)^{1-\frac{1}{p}}} \\ \nabla_x \mathcal{F} &= \frac{\lambda}{p}(\frac{r}{\epsilon})^{p-1} \\ (\nabla_x \mathcal{F})^{\frac{p}{p-1}} &= (\frac{\lambda}{p})^{\frac{p}{p-1}}(\frac{r}{\epsilon})^p \end{aligned} \tag{41}$$

If we sum over two sides, we have

$$\begin{aligned} \sum (\nabla_x \mathcal{F})^{\frac{p}{p-1}} &= \sum (\frac{\lambda}{p})^{\frac{p}{p-1}}(\frac{r}{\epsilon})^p \\ \|\nabla_x \mathcal{F}\|_{p^*}^{p^*} &= (\frac{\lambda}{p})^{p^*} * 1 \end{aligned}$$

where p^* is the dual of p. ($\frac{1}{p} + \frac{1}{p^*} = 1$)

$$(\frac{\lambda}{p}) = \|\nabla_x \mathcal{F}\|_{p^*} \tag{42}$$

By combining (41) and (42), we have

$$r^* = \epsilon sign(\nabla \mathcal{F})(\frac{|\nabla \mathcal{F}|}{\|\nabla \mathcal{F}\|_{p^*}})^{\frac{1}{p-1}}$$

In this paper, we set p to 2. Then we have

$$r^* = \epsilon(\frac{\nabla \mathcal{F}}{\|\nabla \mathcal{F}\|_2}) = \epsilon \overline{\nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2}$$

Therefore, the maximizer x_{max} can be calculated as:

$$x_{max} = r^* + x_0 = \epsilon \overline{\nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2} + x_0$$

$\nabla_x \|\nabla_x L(x_0, y_0, \theta)\|_2$ can be calculated as:

$$\begin{aligned}
\nabla_x \|\nabla_x L(x_0)\|_2 =& [\frac{\partial \|\nabla_x L(x_0)\|_2^2}{\partial x_1}, \frac{\partial \|\nabla_x L(x_0)\|_2^2}{\partial x_2}, ..., \frac{\partial \|\nabla_x L(x_0)\|_2^2}{\partial x_I}] \cdot \frac{1}{2\|\nabla_x L(x_0)\|_2} \\
=& \frac{1}{\|\nabla_x L(x_0)\|_2} [\sum_{i=1}^{I} \frac{\partial L(x_0)}{\partial x_i} \frac{\partial L(x_0)}{\partial x_i \partial x_1}, \sum_{i=1}^{I} \frac{\partial L(x_0)}{\partial x_i} \frac{\partial L(x_0)}{\partial x_i \partial x_2}, \\
& ..., \sum_{i=1}^{I} \frac{\partial L(x_0)}{\partial x_i} \frac{\partial L(x_0)}{\partial x_i \partial x_I}] \\
=& \frac{1}{\|\nabla_x L(x_0)\|_2} \cdot H(x_0) \nabla_x L(x_0)
\end{aligned}$$

Then, using the finite difference method, we have

$$\frac{1}{\|\nabla_x L(x_0)\|_2} \cdot H(x_0) \nabla_x L(x_0) \approx \frac{1}{\|\nabla_x L(x_0)\|_2} \cdot \frac{\nabla_x L(x_0 + \xi \nabla_x L(x_0)) - \nabla_x L(x)}{\xi}$$

where ξ is small value ($\xi = 10^{-6}$). Since $\frac{1}{\|\nabla_x L(x_0)\|_2}$ is scalar, we have

$$x_{max} \approx \epsilon \frac{\overline{\nabla_x L(x_0 + \xi \nabla_x L(x_0)) - \nabla_x L(x_0)}}{\xi} + x_0 \tag{43}$$

Acknowledgements The research results of this article are sponsored by the Kunshan Municipal Government research funding.

Author Contributions The contributions of SZ: He derived the theorems, wrote the paper and conceived and designed the experiments. The contributions of KH: He led the project, helped derive the theorems and contributed to drafting the article. The contributions of ZX: He contributed to drafting the article.

Declarations

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

Ethics approval This article does not contain any studies with human participants performed by any of the authors.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Consent for publication was obtained from the participants.

References

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., & Liang, P. S. (2019). Unlabeled data improves adversarial robustness (pp. 11190–11201).

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. arXiv preprint arXiv:1704.08847.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T., & Song, D. (2018). Physical adversarial examples for object detectors. arXiv preprint arXiv:1807.07769.

Fawzi, A., Fawzi, O., & Frossard, P. (2018). Analysis of classifiers robustness to adversarial perturbations. *Machine Learning, 107*(3), 481–508.

Fawzi, A., Moosavi-Dezfooli, S. M., & Frossard, P. (2016). Robustness of classifiers: From adversarial to random noise (pp. 1632–1640).

Finlay, C., Oberman, A., & Abbasi, B. (2018). Improved robustness to adversarial examples using Lipschitz regularization of the loss. arXiv preprint arXiv:1810.00953.

Fischer, V., Kumar, M. C., Metzen, J. H., & Brox, T. (2017). Adversarial examples for semantic image segmentation. arXiv preprint arXiv:1703.01101.

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

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn (pp. 2980–2988). IEEE.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks.

Kannan, H., Kurakin, A., & Goodfellow, I. (2018). Adversarial logit pairing. arXiv preprint arXiv:1803.06373.

Kos, J., Fischer, I., & Song, D. (2018). Adversarial examples for generative models (pp. 36–42). IEEE.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Liu, D. C., & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical Programming, 45*(1–3), 503–528.

Lyu, C., Huang, K., & Liang, H. N. (2015). A unified gradient regularization family for adversarial examples (pp. 301–309). IEEE.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Houle, M. E., Schoenebeck, G., Song, D., & Bailey, J. (2018). Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint arXiv:1801.02613.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., & Winther, O. (2016). Auxiliary deep generative models. arXiv preprint arXiv:1602.05473.

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

Mao, C., Zhong, Z., Yang, J., Vondrick, C., & Ray, B. (2019). Metric learning for adversarial robustness (pp. 478–489).

Miyato, T., Maeda, S. I., Ishii, S., & Koyama, M. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*, 1979.

Miyato, T., Maeda, S. I., Koyama, M., Nakae, K., & Ishii, S. (2015). Distributional smoothing with virtual adversarial training. arXiv preprint arXiv:1507.00677.

Pang, T., Yang, X., Dong, Y., Su, H., & Zhu, J. (2020). Bag of tricks for adversarial training. arXiv preprint arXiv:2010.00467.

Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks (pp. 3546–3554).

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234–2242).
Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! arXiv preprint arXiv:1904.12843.
Shaham, U., Yamada, Y., & Negahban, S. (2015). Understanding adversarial training: Increasing local stability of neural nets through robust optimization. arXiv preprint arXiv:1511.05432.
Sitawarin, C., Chakraborty, S., & Wagner, D. (2020). Improving adversarial robustness through progressive hardening. arXiv preprint arXiv:2003.09347.
Song, C., He, K., Lin, J., Wang, L., & Hopcroft, J. E. (2019). Robust local features for improving the generalization of adversarial training. arXiv preprint arXiv:1909.10147.
Springenberg, J. T. (2015). Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390.
Stanforth, R., Fawzi, A., & Kohli, P., et al. (2019). Are labels required for improving adversarial robustness? arXiv preprint arXiv:1905.13725.
Weng, T. W., Zhang, H., Chen, P. Y., Yi, J., Su, D., Gao, Y., Hsieh, C. J., & Daniel, L. (2018). Evaluating the robustness of neural networks: An extreme value theory approach. arXiv preprint arXiv:1801.10578.
Willetts, M., Camuto, A., Rainforth, T., Roberts, S., & Holmes, C. (2019). Improving vaes' robustness to adversarial attack. arXiv preprint arXiv:1906.00230.
Wu, D., Xia, S. T., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. arXiv preprint arXiv:2004.05884.
Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.
Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:1605.07146.
Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. arXiv preprint arXiv:1901.08573.
Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., & Kankanhalli, M. (2020). Geometry-aware instance-reweighted adversarial training. arXiv preprint arXiv:2010.01736.
Zhao, J., Mathieu, M., Goroshin, R., & Lecun, Y. (2015). Stacked what-where auto-encoders. arXiv preprint arXiv:1506.02351.

Authors and Affiliations

Shufei Zhang[1,2,3] · Kaizhu Huang[4] · Zenglin Xu[5]

Shufei Zhang
zhangshufei@pjlab.org.cn

Zenglin Xu
xuzenglin@hit.edu.cn

[1] Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK

[2] Department of Intelligent Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China

[3] Present Address: Shanghai Artificial Intelligence Laboratory, 37th floor, AI Tower, 701 Yunjin Road, Shanghai, China

[4] Data Science Research Center, Duke Kunshan University, Duke Avenue No. 8, Kunshan, Suzhou 215316, China

[5] Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China