PERTURBATION DIVERSITY CERTIFICATES ROBUST GENERALISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whilst adversarial training has been proven the most effective defending method against adversarial attacks for deep neural networks, it suffers from overfitting on unseen adversarial data and thus may not guarantee robust generalisation. It is possibly due to the fact that the conventional adversarial training methods generate adversarial perturbations usually in a supervised way, so that the adversarial samples are highly biased towards the decision boundary, resulting in an inhomogeneous data distribution. To mitigate this limitation, we propose a novel adversarial training method from a perturbation diversity perspective. Specifically, we generate perturbed samples not only adversarially but also diversely, so as to certificate significant robustness improvement through a homogeneous data distribution. We provide both theoretical and empirical analysis which establishes solid foundation to well support the proposed method. To verify our methods' effectiveness, we conduct extensive experiments over different datasets (e.g., CIFAR-10, CIFAR-100, SVHN) with different adversarial attacks (e.g., PGD, CW). Experimental results show that our method outperforms other state-of-the-arts (e.g., PGD and Feature Scattering) in robust generalisation performance. (Source codes are available in the supplementary material.)

1 INTRODUCTION

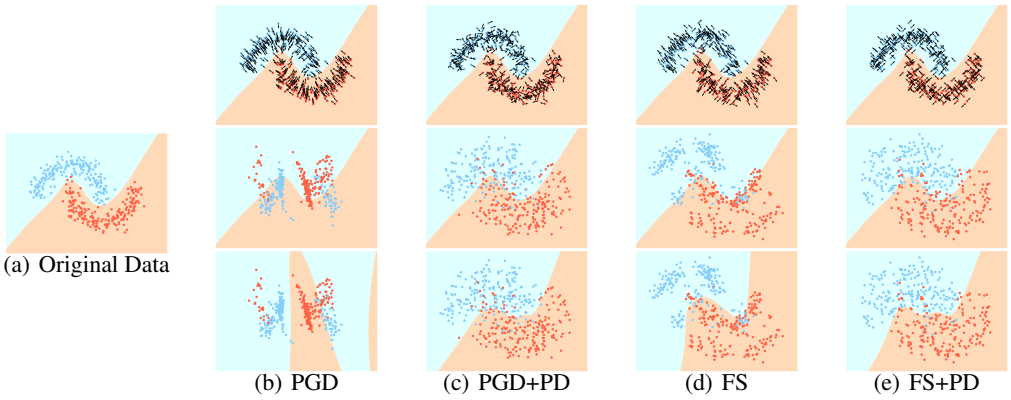
Whilst Deep Neural Networks (DNNs) have attained breakthroughs in recent decades, the robustness issue of these models has arisen as one major concern in many applications. Typically, DNNs appear vulnerable and/or easily obtain unexpected outputs on adversarial examples which are perturbated by crafted imperceptible noise (LeCun et al., 2015; He et al., 2016; Gers et al., 1999). Adversarial examples have been shown ubiquitous in a variety of tasks such as image classification (Goodfellow et al., 2014), segmentation (Fischer et al., 2017), speech recognition (Carlini & Wagner, 2018), and text classification (Yang et al., 2020a). The model robustness has emerged as one of the most challenge tasks and has drawn enormous attention recently.

To defend against adversarial examples, great efforts have been made, such as denoise-based methods (Lamb et al., 2018; Liao et al., 2018; Yang et al., 2019), detecting-based methods (Metzen et al., 2017; Feinman et al., 2017; Xu et al., 2017; Pang et al., 2017), and adversarial training (Kannan et al., 2018; You et al., 2019; Wang & Zhang, 2019; Zhang & Wang, 2019). In the arms race between attacks and defences, adversarial training has been shown as the most promising techniques, for which the models are trained with adversarial samples rather than clean data (Goodfellow et al., 2014; Madry et al., 2017). The adversarial training is a min-max game between the adversarial perturbations and classifier. Namely, the imperceptible adversarial perturbations are crafted to maximise the probability of mis-classification, while the classifier is trained to minimise the loss due to such perturbed data.

Such a robust optimisation approach has been proven effective, significantly improving the model robustness. However, the generated perturbations by most of adversarial training methods such as FGSM (Goodfellow et al., 2014) and PGD (Madry et al., 2017) are restrictively and typically biased towards the decision boundary. This monotonic orientation can cause the close samples to move into the same area, resulting in redundant-density areas where the excessive of adjacent data blocks the optimisation of the decisive boundary and leads to a non-smooth decisive boundary. On the other hand, the centralised data also contribute to the corresponding low-density areas where rare data

could support to develop an optimal decisive boundary which would be easily broken during the later training process. These over-concentrated data result in an inhomogeneous data distribution and thus hinder the training of the model.

To illustrate this point, we present a toy example in Fig. 1, which shows the gradient direction of different methods (top-line), adversarial examples generated from corresponding methods (middle-line), and the decision boundary trained with corresponding adversarial examples (bottom-line). As clearly observed, the adversarial samples generated by PGD, one of the most successful adversarial attacks, are highly biased to the decision boundary and mostly concentrated in the crest and trough of the decision boundary. Under the support of such data distribution, the decision boundary originally located in the crest and trough will be severely distorted and overfitted, while decision boundary previously located at other locations will also be severely distorted in the subsequent training due to lack of data. Unfortunately, this phenomenon is not specific to supervised methods such as PGD, but is also present in other unsupervised methods that consider the inter-samples relation such as Feature Scattering (FS) (Zhang & Wang, 2019).



(a) Original Data (b) PGD (c) PGD+PD (d) FS (e) FS+PD

Figure 1: Illustrative example of the gradient direction (Top), adversarial samples (middle) and trained decision boundary (Bottom) of different adversarial perturbation schemes. (a) Original data without perturbation; perturbed data using (b) PGD, a supervised adversarial attack method; (c) PGD with the proposed perturbation diversity (d) Feature Scattering, and (e) Feature Scattering with the proposed perturbation diversity.

To address the above limitation, we propose to improve adversarial robustness from a new perspective by promoting the diversity among the generated perturbations, called *Perturbation Diversity* (PD). Technically, we first develop a new method of perturbation generation in the adversarial setting, which is significantly different from the previous attempts in both the supervised and unsupervised setting. Namely, previous work has defined that the adversarial perturbations have their own fixed objectives, either to maximise the loss function or to maximise the difference in data distribution. However, these objectives somehow have their own bias which leaves biased perturbed samples and is inappropriate for an optimal training procedure. Instead, we define the novel adversarial samples that are not only adversarial enough, but also have to be as diverse as possible, as implemented by the proposed adversarial perturbation that should be orthogonal to the each other as much as possible while moving towards the original objectives.

On the empirical front, first, we illustrate in Fig. 1 that our proposed adversarial examples can boost the previous baseline such as PGD-AT and FS, which fills the data space as homogeneously as possible. After training with such boosted adversarial examples, the model can preserve more information of the original distribution and learn a better decision boundary than the existing adversarial training methods. From the figure, we can see that both the AT and the FS alter the original decision boundary significantly. Moreover, it can be observed that the adversarial training with PGD corrupts the data manifold completely. On the other hand, FS appears able to retain partially the data manifold information since it considers the inter-sample relationship locally. In contrast, our proposed method considers to maximise perturbation diversity which potentially retains the global information of data manifold after the data are adversarially perturbed, and thus obtains a nearly optimal

decision boundary. This may explain why our proposed perturbation diversity could outperform the other approaches.

Importantly, on the theoretical part, we have proved that the robust generalisation gap of adversarial training can be upper bounded by a standard generalisation gap and a term related to the proposed perturbation diversity. Namely, maximising perturbation diversity tends to reduce a smaller robust generalisation gap, which establishes the theoretical foundation for the proposed method.

We test our method on three different datasets: CIFAR-10, CIFAR-100, and SVHN with the most commonly used PGD, CW and FGSM attacks. Our method can be applied to any baseline. Without sacrificing the accuracy of the original samples, it outperforms the state-of-the-art baselines by a large margin. For example, PGD+PD improves over PGD-AT (Madry et al., 2017; Rice et al., 2020) by 30.3% and 27.92% on SVHN for PGD20 and CW20 attacks.

2 Background and Related Work

Among many other adversarial defence techniques, adversarial training has been proven as the most efficient one to improve the model robustness (Goodfellow et al., 2014; Madry et al., 2017). Owing to its persistence to increasingly powerful adversarial attacks, it has been drawing more and more attention in the past years, e.g., (Zhang et al., 2021; Wang & Zhang, 2019; Zhang & Wang, 2019; Zhang et al., 2020; Mao et al., 2019; Shafahi et al., 2019; Zhang et al., 2019a; Zhu et al., 2019; Wong et al., 2020). By directly generating adversarial perturbations of the inputs, the philosophy of adversarial training is to learn robust DNN models that could be tenable even under some adversarial attacks. In this section, we give a brief introduction to adversarial training and robust generalisation.

2.1 Conventional Adversarial Training

In the conventional adversarial training, the main idea is to train the model with the adversarial perturbations which could possibly lead to a worst situation, e.g., a wrong prediction. With the model parameters $\theta \in \mathbb{R}^m$ of the DNN, let $\mathcal{L}(\cdot)$ be the loss function. Then, adversarial training can be formulated as a minimax optimisation problem as follows:

$$\min_{\theta} \quad \{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\max_{\boldsymbol{x}'\in S_x} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}; \theta)]\}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{N}$ denote the clean data samples and the corresponding labels, respectively, drawn from a data distribution $\mathcal{D}$; $\boldsymbol{x}' \in \mathbb{R}^d$ is the adversarially perturbed copy of $\boldsymbol{x}$ with the perturbation constrained in a feasible region $S_x \triangleq \{\boldsymbol{z} : \boldsymbol{z} \in B(\boldsymbol{x}, \epsilon) \cap [-1.0, 1.0]^d\}$; and $B(\boldsymbol{z}, \epsilon) \triangleq \{\boldsymbol{z} : \|\boldsymbol{x} - \boldsymbol{z}\|_\infty \leq \epsilon\}$ specifies the ℓ_∞-norm ball at the centre $\boldsymbol{x}$ with radius ϵ.

The above minimax optimisation problem is known to be difficult to solve directly. A feasible solution is the alternating optimisation (e.g., Madry et al. (2017)), which iteratively updates between the outer minimisation via SGD training and the inner maximisation via adversarial attacks (e.g., PGD, FGSM).

The conventional adversarial training aims to obtain a robust model by identifying the potential attacks and eliminating the effect of the perturbation during the training process. Nevertheless, as the data distribution is not considered in these methods, the adversarial perturbations generated from the conventional adversarial training ignore the global data structure information and therefore are highly biased, which inevitably degrades the robust generalisation performance.

2.2 Adversarial Training with Inter-samples Relation

Recent methods (Sinha et al., 2017; Miyato et al., 2017; Zhang & Wang, 2019) argue that the perturbations generated in the supervised way are restrictive and produced individually within the local region, in which way the adversarial examples may corrupt the underlying data structure. To exploit the data manifold, recent works have started to consider the inter-sample relationship during adversarial perturbation generation. Of particular relevance is the adversarial training method named Feature Scattering (FS) (Zhang & Wang, 2019), which is deemed as one of the most promising methods in the literature. The most appealing factor of FS is that the inter-sample relationships of the inputs are considered when generating adversarial examples. Specifically, it perturbs

the local neighborhood structure through the maximisation of the optimal transport (OT) distance $c(\boldsymbol{x}_i, \boldsymbol{x}'_j) = 1 - \frac{f_\theta(\boldsymbol{x}_i)^\top f_\theta(\boldsymbol{x}'_j)}{\|f_\theta(\boldsymbol{x}_i)\|_2 \|f_\theta(\boldsymbol{x}'_j)\|_2}$ between natural examples and perturbed examples. Therefore, FS not only considers the worst-case samples, but also other weakly perturbed samples that are critical to the robustness of the model.

Different from FS measuring two distributions with one fixed metric, Adversarial Training with Latent Distribution (Qian et al., 2021) utilises a learnable discriminator to distinguish the distribution of natural examples and perturbed examples. This method generates adversarial examples to maximise the divergence between the latent distributions of clean data and their adversarial counterparts.

These adversarial examples aim to make their latent feature as far away from their clean counterparts as possible. Due to the unsupervised way of generating perturbations, the correlation between the perturbations and the decision boundary is relatively low, which makes the potential label leaking problem somewhat avoided (Zhang & Wang, 2019).

However, although these methods consider the data structure when generating adversarial perturbation, the perturbations still have a strong bias, which could be observed in Fig. 1 and the toy example in (Qian et al., 2021). As discussed previously in Section 1, such bias would hinder the optimisation of the models and make models less robust.

2.3 Robust generalisation

Analogously to standard generalisation with respect to unseen clean data, robust generalisation measures the performance of robust models on unseen adversarial data. It has been reported in (Schmidt et al., 2018; Zhai et al., 2019) that learning a model with good robust generalisation is particularly difficult because of the significantly higher (adversarial) data complexity. This attracted a new line of research on the interplay between robustness and generalisation. For instance, by decomposing the robust error due to adversarial examples into the natural classification error and the boundary error, Zhang et al. (2019b) proposed to make the trade-off between the robustness and the accuracy via adversarial training. It is in sharp contrast to Yang et al. (2020b), which argued that both accuracy and robustness are achievable if the local Lipschitzness can be maintained to some extent.

In addition to the bounding techniques of robust generalisation, another thread of this research aims to devise new regularisation techniques to promote robust generalisation for adversarial training. Of particular relevance is the work in Yin et al. (2019), which demonstrated that ℓ_1 norm of weight matrices affects the robust generalisation performance, and the work in Wu et al. (2020), which made both weight and input sample perturbations to enhance both generalisation and robustness. Pang et al. (2019a) proposed the Max-Mahalanobis center (MMC) loss to explicitly induce dense feature regions in order to benefit robustness. More recently, Roth et al. (2020) established a link between adversarial training and the operator norm regularisation, and Rice et al. (2020) demonstrated that the conventional PGD-AT (Madry et al., 2017) suffered from the adversarial overfitting and improved the PGD-AT by early-stop simply. Such overfitting phenomenon can also be found in many other methods.

3 Main Methodology

As discussed in the previous sections, conventional adversarial training methods generate adversarial examples for training by focusing on maximising the loss (typically the cross-entropy), which makes the direction of perturbations purely monotonous and biased to the decision boundary. Therefore, these methods lead to adversarial data which are over-centralised and generate redundant-density and low-density areas as shown in Fig. 1(b). In redundant-density areas, excessive data play the same role for the training, causing models to be overfitted and locally non-smooth. However, in low-density areas, the model could not obtain enough data to support training, which makes the decision boundary in such areas hard to generalise. Obviously, this is less desirable as it might neglect other directions that are crucial for learning robust models (Ilyas et al., 2019; Etmann et al., 2019). While several recent methods consider the inter-sample relationships during generating adversarial perturbations, which somewhat increases the diversity of the sample and alleviates the monotonous perturbation, they are still biased and lead to undesirable low-density areas as shown in Fig. 1(d).

In this section, we first introduce the training strategies for adversarial training with perturbation diversity. Then we provide empirical and theoretical analysis on the solutions of our proposed method.

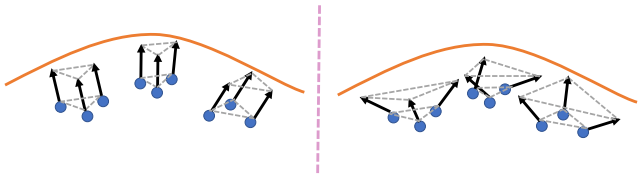


Figure 2: Illustration of perturbation diversity. **Left**: Directions of conventional adversarial examples. **Right**: Directions of adversarial examples with Perturbation Diversity. When the perturbations are orthogonal to each other, their diversity can be guaranteed, hence promoting a more smooth robust decision boundary. Details can be seen in Fig. 1 and Section 3.1.

3.1 Adversarial Training with Perturbation Diversity

Motivated by the determinant point process (DPP) theory (Kulesza & Taskar, 2012; Pang et al., 2019b), we first define the perturbation diversity for n samples within a batch as:

$$\mathbb{PD} = \det(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}), \tag{2}$$

where $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathbb{R}^{d\times n}$ is the perturbations of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{d\times n}$ which are sampled from the training set consisting of N samples. To avoid trivial solution, we require the batch size $n \leq d$ which can be typically satisfied in practice. Each perturbation of $\boldsymbol{x}_i$ is within a feasible region being the ℓ_∞-ball at center $\boldsymbol{x}_i$ with radius ϵ as defined in Eq. (1) where $i = 1, 2, \ldots, n$.

The DPP (Kulesza & Taskar, 2012) method assigns higher probability to sets of items that are diverse in its own original purpose. As shown in Fig. 2, we make a simple demonstration illustrating how the perturbation diversity can be attained. Conventional adversarial examples (left in Fig. 2) are highly biased to the decision boundary, which makes the closer samples have similar perturbation directions. However, the proposed Perturbation Diversity (right in Fig. 2) aims to generate perturbations as orthogonal as possible to each other (whilst maximising the loss). In this way, as theoretically justified shortly, we can encourage diverse perturbations, i.e., the resulting adversarial examples can maintain diverse directions so as to avoid inhomogeneous perturbation distributions (see Fig. 1 (c) and (e)).

According to the definition of volume with respect to matrix determinant (Sain, 2007), we have:

$$\det(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}) = \mathrm{Vol}^2(\{\boldsymbol{z}_i\}_{i\in[n]}), \tag{3}$$

where $[n] = \{1, 2...n\}$ and $\mathrm{Vol}(\cdot)$ is the volume of the polyhedron spanned by the vectors of the input. With such concept of volume, we could interpret the perturbation diversity in a geometrical way. Note that each vector $\boldsymbol{z}_i \in \mathbb{R}^d$ is obtained under L_p-norm normalization due to the definition of adversarial perturbation. With normalized $\{\boldsymbol{z}_i\}$ such that $||\boldsymbol{z}_i||_p = \epsilon$, the perturbation diversity $\mathbb{PD}$ achieves its maximal value ϵ^n if and only if $\{\boldsymbol{z}_i\}_{i\in[n]}$ are mutually orthogonal.

Therefore, to promote perturbation diversity, we propose the adversarial training with perturbation diversity as the regularisation as follows:

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_\theta(\boldsymbol{x}_i^{adv}, \boldsymbol{y}_i, \theta), \quad \text{s.t.} \quad \boldsymbol{x}_i^{adv} = \arg\max_{\boldsymbol{x}_i'\in S_x} (\mathcal{L}'(\boldsymbol{x}_i', \boldsymbol{y}_i, \theta) + \log(\mathbb{PD})), \tag{4}$$

where $\boldsymbol{x}_i^{adv}$ is the generated adversarial example, $\boldsymbol{x}_i'$ is the intermediate adversarial input during the iterative perturbation, $\mathcal{L}_\theta$ is the final objective for training the model parameterized by θ. $\mathcal{L}'$ is the objective for generating perturbations, which can be treated as any other adversarial training method. In PGD-AT (Madry et al., 2017), $\mathcal{L}'$ is the cross-entropy loss which is the same as $\mathcal{L}_\theta$; in FS (Zhang & Wang, 2019), $\mathcal{L}'$ is the optimal transport distance. Our proposed perturbation diversity acts as a regularisation when generating the perturbations, making it readily used in most existing adversarial

(a) PGD (b) PGD+PD (c) FS (d) FS+PD (e) Consumed time

Figure 3: (a) to (d): TSNE manifold of the training and test data. All test data are attacked by PGD, the training data are attacked by a) PGD, b) PGD+PD, c) FS, and d) FS+PD. (e): Training time consumed at per epoch vs. Batch size.

training methods. Again, when we conduct adversarial training, $\mathbb{PD}$ is calculated in a batch fashion (with the size 60 or 120 in this paper as discussed in the appendix).

In the back-propagation procedure, the proposed PD requires the separate computation of two matrix operations $\det(\mathbb{PD})$ and $\mathbb{PD}^{-1}$. In particular, both operations have computational complexity of $\mathcal{O}(n^3)$ (Kulesza & Taskar, 2012), which may incur the curse of dimensionality if n is very large. However, empirically speaking, models trained with large batch size n would not be better than those trained with a small batch size. This property enables our method to scale to most modern machine learning tasks. On the other hand, as experimentally shown in Fig. 3(e), compared to the excessive and iterative gradient computation in adversarial training, the proposed PD seems to require less computation, which makes our method also compatible with other defense methods, e.g., adversarial training (Madry et al., 2017; Zhang & Wang, 2019), and thus our method could be a plug-and-play component to further improve model robustness with perturbation diversity.

3.2 Empirical Analysis

In this subsection, we visualise the TSNE embedding of the output features to show that the test adversarial features are away from training data features of the same class in conventional adversarial training methods, meaning that they usually overfit the training data and suffer from poor robust generalisation. However, with the proposed PD, this feature shift could be largely alleviated as shown in Fig. 3. In this figure, all the test data are attacked by PGD, while the training data are PGD attacked, PGD+PD attacked, FS attacked, and FS+DP attacked from Fig. 3(a) to 3(d) respectively.

Concretely, we present the T-SNE embedding of training data attacked by PGD as well as test data in Fig. 3(a) where for each class, the training data feature distribution is far away from the test counterpart. In other words, even for the same class, the test adversarial data feature distribution appears very different from the training one. Therefore, the classifier learned with training data would perform unsatisfactorily on test data and leads to poor robust generalisation. Differently, as seen in Fig. 3(b), as certificated by PD, the test adversarial data features stay close to the training adversarial features of the same class. Apparently, the classes are more clustered than the conventional PGD-AT. Consequently, the classifier learned with the PD can perform well on test adversarial examples.

Moreover, with the baseline FS, we also plot the T-SNE embedding of adversarial samples for both the training and test data in Fig. 3(c) and 3(d) where the training data are FS attacked and FS+PD attacked respectively, and the test data are both PGD attacked. Comparing Fig. 3(c) and 3(d), one can note that due to the good generalisation ability of FS, the clusters are more clear than PGD-AT. However, since the FS attack is different from PGD, test data feature distribution are quite different from training ones. In contrast, with the proposed PD, the test data feature distribution in Fig. 3(d) again stays closer to the training one when we compare it to FS training (especially on the dark red and blue classes). Moreover, FS+PD shows a better clustering property than the conventional FS.

3.3 Theoretical Analysis

With the empirical analysis at hand, we now proceed to analyse the theoretical aspects of the proposed method with focus on the robust generalisation bound.

Before proceeding further, we first introduce the concepts of standard and robust generalisation errors. The (standard) generalisation error is defined as the difference between the expected loss

over data distribution $(\boldsymbol{x}, \boldsymbol{y}) \sim (\mathbb{S}, \mathbb{Y})$ and the empirical loss over the training data $(\boldsymbol{x}_d, \boldsymbol{y}_d) \in (\mathbb{S}_d, \mathbb{Y}_d)$ (Xu & Mannor, 2012; Neyshabur et al., 2017). By letting $\mathbb{S}_d, \mathbb{Y}_d$ be the training data and the corresponding labels, respectively, and $\mathbb{S}, \mathbb{Y}$ be the underlying data and label distributions, respectively, we have

$$\begin{aligned} &\mathrm{GE} \triangleq |l(f_\theta(\mathbb{S}), \mathbb{Y}) - \hat{l}(f_\theta(\mathbb{S}_d), \mathbb{Y}_d)| \text{ , where} \\ &l(f_\theta(\mathbb{S}), \mathbb{Y}) \triangleq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim(\mathbb{S},\mathbb{Y})}[l(f_\theta(\boldsymbol{x}), \boldsymbol{y})] \\ &\hat{l}(f_\theta(\mathbb{S}_d), \mathbb{Y}_d) \triangleq \frac{1}{|\mathbb{S}_d|} \sum_{(\boldsymbol{x}_d,\boldsymbol{y}_d)\in\mathbb{S}_d} l(f_\theta(\boldsymbol{x}_d), \boldsymbol{y}_d). \end{aligned} \tag{5}$$

Further, the robust generalisation error can be defined in a similar way, with the only difference that the clean samples are replaced by the adversarial data samples and distribution (Schmidt et al., 2018; Zhai et al., 2019; Wu et al., 2020) , i.e.,

$$\mathrm{RGE} \triangleq |l(f_\theta(\mathbb{S}^{adv}), \mathbb{Y}) - \hat{l}(f_\theta(\mathbb{S}_d^{adv}), \mathbb{Y}_d)| \tag{6}$$

where $\mathbb{S}_d^{adv}$ and $\mathbb{S}^{adv}$ are similarly defined as the set of adversarial examples for the training data and its underlying distribution, respectively. Equipped with these definitions, we now are ready to bound the robust generalisation errors as shown in Theorem 3.1.

Theorem 3.1 *Given the training dataset $\mathbb{S}_d = \{\boldsymbol{x}_i\}_{i=1}^N$ with N samples independently drawn from the distribution $\mathbb{S}$ with K subsets $\{\mathbb{C}_j\}_{j=1}^K$ where we assume that adversarial perturbations of each subset share the similar direction, and the set of adversarial perturbations of training set is $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$, if the loss function $l(\cdot)$ of DNN f_θ is k-Lipschitz, then for any $\delta > 0$, with the probability at least $1 - \delta$, we have*

$$\mathrm{RGE} \leq \mathrm{GE} + \frac{k}{N} \sum_{j=1}^{K} \sum_{i\in N_j} \|\boldsymbol{W}_{ij}(\boldsymbol{\epsilon}_i - \hat{\boldsymbol{\epsilon}}_j)\|_2^2 + M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{N}} \qquad where \tag{7}$$

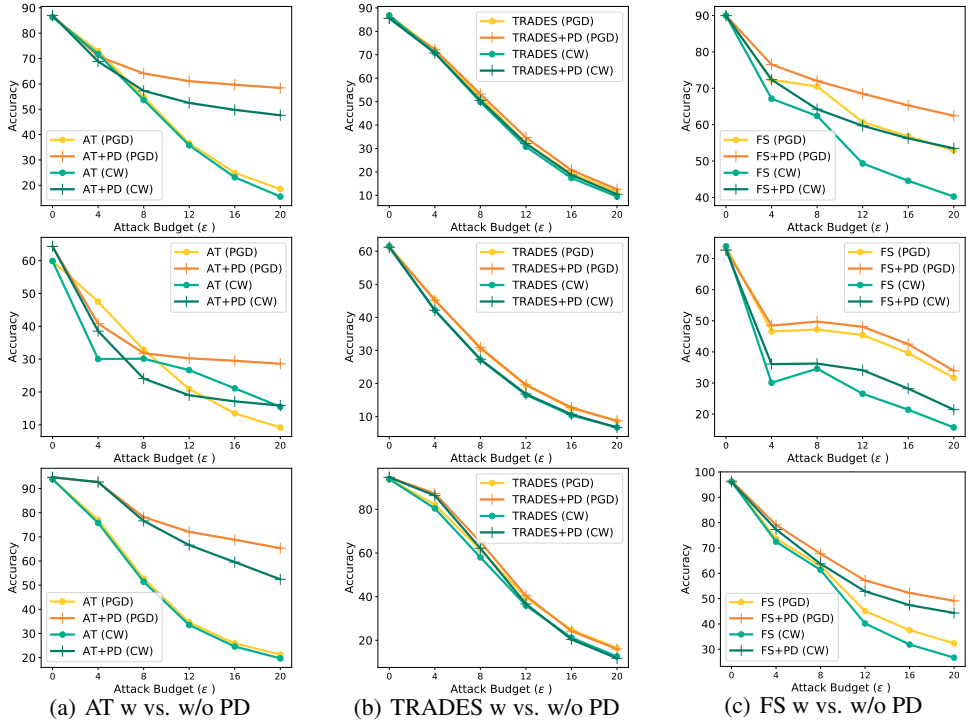
$$\hat{\boldsymbol{\epsilon}}_j = \mathbb{E}[\boldsymbol{z}^{adv} - \boldsymbol{z} | \boldsymbol{z} \in \mathbb{C}_j] \tag{8}$$

where $\boldsymbol{z}$ is the data sampled from $\mathbb{C}_j$ with the corresponding adversarial example $\boldsymbol{z}^{adv}$, N_i denotes the set of index of training data which belongs to $\mathbb{C}_i$, $\boldsymbol{W}_{ij}$ is the transformation matrix, and M is the upper bound of loss of the whole data manifold $\mathbb{S}$. We also assume that the diversity of $\{\hat{\boldsymbol{\epsilon}}_j\}_{j=1}^K$ is greater than $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$.

The proof of Theorem 3.1 is delegated to Appendix. Theorem 3.1 says that, the robust generalisation error (RGE) can be upper bounded by the sum of standard generalisation error (GE), a term of the perturbation variance (the second term), and a constant part (the last term). Specifically, the perturbation variance term is leveraged to measure the difference between the adversarial perturbations of training set and whole unknown underlying data. It can be noted that the gap between the robust and standard generalisation is mainly caused by the perturbation variance term. In the perturbation variance term, if the diversity of $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ is small, all the perturbations of training data $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ tend to share quite similar directions and hence would fall into a very small number of of subsets $\{\mathbb{C}_j\}_{j=1}^K$ (according to the subset definition); this consequently leads to a large value of perturbation variance and poor robust generalisation bound.

Intuitively, the adversarial perturbations of the whole underlying data usually have a greater diversity than the training set, thus there are a considerable portion of unseen adversarial perturbations beyond the training process and it is hard for the adversarially trained model to generalise well on such unseen perturbations. To tackle such the problem, we try to promote the diversity of adversarial perturbations during the training process such that the resulting space spanned by training adversarial perturbations could be uniform or homogeneous. As such, the adversarial trained model can generalise well on unseen perturbations.

It should be noted that directly minimising the second term of Eq. (7) is not possible as the adversarial perturbations of the underlying data are unknown. Though in this paper, we manage to enlarge the perturbation diversity as a simple yet effective regularisation in the training set so as to reduce the robust generalisation gap and achieve encouraging results, it keeps interesting if a better way can be sought to further promote the robust generalisation. We will leave this as one open problem.



(a) AT w vs. w/o PD (b) TRADES w vs. w/o PD (c) FS w vs. w/o PD

Figure 4: Attack Budget (ϵ) vs. Robust Accuracy on the three baseline. **Top Line:** CIFAR-10; **Middle Line:** CIFAR-100; **Bottom Line:** SVHN.

4 EXPERIMENTS

In this part, we perform extensive experiments to evaluate our proposed PD on several baselines, including PGD-AT, TRADES and FS in defending against various adversarial attacks.

4.1 EXPERIMENTAL SETTING

To verify the effectiveness of our proposed method, we compare the robustness performance with the state-of-the-art adversarial training methods on CIFAR-10, CIFAR-100, and SVHN against white/black box adversarial attacks. For the three benchmark methods, i.e. FS, AT, and TRADES, we adopt WideResNet-28-10 as the basic model structure, following the settings in (Zhang & Wang, 2019; Madry et al., 2017). Specifically, on all three datasets CIFAR-10, CIFAR-100, and SVHN, we train the models using SGD with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1. The learning rate decays at epoch 60 and 90 with the rate 0.1. The attack iteration number during the training period follows the baselines, such as 7 for PGD-AT and TRADES, 1 for FS, and the attack step size is $2/255$ for PGD and TRADES, $8/255$ for FS. The attack budget ϵ is set to $8/255$ for all the methods, following the practice of man previous methods (Zhang & Wang, 2019; Madry et al., 2017). In both training and test, all attacks are computed with l_∞ norm. For saving space, the detailed experimental settings are listed in Appendix. The *source codes* are included in the supplementary material.

4.2 ROBUSTNESS TO ADVERSARIAL ATTACKS

Improvement over Different Baselines: To clearly see how the proposed PD can improve various models, we compare the three baselines (including AT (Madry et al., 2017; Rice et al., 2020), Feature Scattering adversarial training (FS) (Zhang & Wang, 2019), and TRADES (Zhang et al., 2019b)) with and without the proposed PD by varying the perturbation magnitude ϵ under different adversarial attacks (including FGSM, PGD20, and CW20) as shown in Fig. 4. As observed, the

proposed PD could generally increase the robustness, particularly it significantly improves the robustness against large attack budget (ϵ) especially on AT and FS models, although all the models are trained under the same $\epsilon = 8$. On the other hand, though the proposed PD just marginally improves Trades, such improvement appears consistent on all the attack budgets. (**Note:** More detailed comparison experiments with the three baselines including black-box attack and the more sophisticated AutoAttack (AA) (Croce & Hein, 2020) can be seen in the Appendix.)

Table 1: Accuracy under white-box attacks on CIFAR-10

MODELS	CLEAN	ACCURACY UNDER WHITE-BOX ATTACK (ϵ = 8)						
		FGSM	PGD20	PGD40	PGD100	CW20	CW40	CW100
STANDARD	**95.60**	36.90	0.00	0.00	0.00	0.00	0.00	0.00
AT	85.70	54.90	44.90	44.80	44.80	45.70	45.60	45.40
TLA	86.21	58.88	51.59	-	-	-	-	-
LAT	87.80	-	53.84	-	53.04	-	-	-
BILATERAL	91.20	70.70	57.50	–	55.20	56.20	–	53.80
TRADES	86.07	67.25	55.16	52.97	54.94	52.87	54.88	52.83
FS	90.00	78.40	70.50	70.30	68.60	62.40	62.10	60.60
RST-AWP	88.25	67.94	63.73	-	63.58	61.62	-	-
RLFAT$_T$	82.72	-	58.75	-	-	51.94	-	-
RLFAT$_P$	84.77	-	53.97	-	-	52.40	-	-
FS+PD	89.99	**79.37**	**72.02**	**70.53**	**69.57**	**64.26**	**62.76**	**61.70**

Comparison with SOTAs: Taking the FS as the typical baseline, we conduct comparisons between the proposed FS+PD and the current state-of-the-art adversarial training methods including 1) AT Madry et al. (2017), 2) TLA (Mao et al., 2019), 3) LAT (Sinha et al., 2019), 4) Bilateral (Wang & Zhang, 2019), 5) FS (Zhang & Wang, 2019). Moreover, other recent methods proposed to promote the robust generalisation are also included, such as 6) RST/AT-AWP (Wu et al., 2020) and 7) RLFAT$_{T/P}$ Song et al. (2019). We demonstrate the accuracy of these different methods in Table 1 and Table 2 on CIFAR-10, CIFAR-100 and SVHN respectively. For CIFAR-10, it can be seen that our proposed FS+PD has achieved the best performance under all the adversarial attacks without further sacrificing the standard accuracy compared to the baseline FS. In addition, it can be observed that our method also outperform the recent robust generalisation methods such as RST/AT-AWP, and RLFAT$_{T/P}$ by a large margin. For CIFAR-100 and SVHN, our proposed method also demonstrates consistently higher accuracy than almost all the other models, except that a slightly inferior to FS under PGD100 and CW100 on SVHN.

Table 2: Accuracy under different attack on CIFAR-100 and SVHN

MODELS	CIFAR-100(ϵ = 8)						SVHN(ϵ = 8)					
	CLEAN	FGSM	PGD20	PGD100	CW20	CW100	CLEAN	FGSM	PGD20	PGD100	CW20	CW100
STANDARD	**79.00**	10.00	0.00	0.00	0.00	0.00	**97.20**	53.00	0.30	0.10	0.30	0.10
AT	59.90	28.50	22.60	22.30	23.20	23.00	93.90	68.40	47.90	46.00	48.70	47.30
LAT	60.94	-	27.03	26.41	-	-	91.65	-	60.23	**59.97**	-	-
BILATERAL	68.20	60.80	26.70	25.30	-	22.10	94.10	69.80	53.90	50.30	-	48.90
FS	73.90	61.00	47.20	46.20	34.60	30.60	96.20	83.50	62.90	52.00	61.30	**50.80**
AT-AWP	-	-	30.71	-	-	-	-	-	59.12	-	-	-
RLFAT$_T$	58.96	-	31.63	-	27.54	-	-	-	-	-	-	-
RLFAT$_P$	56.70	-	31.99	-	29.04	-	-	-	-	-	-	-
FS+PD	72.72	**74.77**	**49.75**	**49.35**	**36.25**	**36.19**	96.54	**97.42**	**67.72**	53.13	**63.75**	49.72

5 CONCLUSION

We have developed a novel adversarial training method from a perturbation diversity perspective in this paper. While existing adversarial training typically focuses the perturbation objective only, which generates inhomogeneous data distribution and limits the model's generalisation, our proposed novel regularisation can certificate to generate adversarial perturbations as diverse as possible to obtain better robust generalisation. We have provided theoretical and empirical investigations which validate our perturbation diversity can lead to performance gains in a number of baseline models.

REPRODUCIBILITY STATEMENT

For empirical experiments, we have described our experiment settings including but not limited to Operating systems, Pytorch Version, Graphics card model. We also put more details in the Appendix including the parameters used in the experiments. We also upload our source codes as the additional supplementary material with anonymous download link for our reference model. For theoretical analysis, we have provided detailed proof and derivation in the Appendix.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation. In *arXiv preprint arXiv:1703.01101*, 2017.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. IET, 1999.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. In *arXiv:1803.06373*, 2018.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

Alex Lamb, Jonathan Binas, Anirudh Goyal, Dmitriy Serdyuk, Sandeep Subramanian, Ioannis Mitliagkas, and Yoshua Bengio. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. In *arXiv:1804.02485*, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. In *nature*, volume 521, pp. 436. Nature Publishing Group, 2015.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *arXiv:1706.06083*, 2017.

Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 478–489, 2019.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *arXiv preprint arXiv:1702.04267*, 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. In *arXiv:1704.03976*, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *arXiv preprint arXiv:1706.08947*, 2017.

Tianyu Pang, Chao Du, and Jun Zhu. Robust deep learning via reverse cross-entropy training and thresholding test. *arXiv preprint arXiv:1706.00633*, 3, 2017.

Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019a.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979. PMLR, 2019b.

Zhuang Qian, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Improving model robustness with latent distribution locally and globally. *arXiv preprint arXiv:2107.04401*, 2021.

Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. *arXiv preprint arXiv:2002.11569*, 2020.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33, 2020.

Michael K Sain. Matrix mathematics: Theory, facts, and formulas with application to linear systems theory [book review; ds berstein]. *IEEE Transactions on Automatic Control*, 52(8):1539–1540, 2007.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Abhishek Sinha, Mayank Singh, Nupur Kumari, Balaji Krishnamurthy, Harshitha Machiraju, and VN Balasubramanian. Harnessing the vulnerability of latent layers in adversarially trained models. In *arXiv:1905.05186*, 2019.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Chubiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E Hopcroft. Robust local features for improving the generalization of adversarial training. In *arXiv preprint arXiv:1909.10147*, 2019.

Aad W. Van, der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 2000.

Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Huan Xu and Shie Mannor. Robustness and generalization. In *Machine learning*, volume 86, pp. 391–423. Springer, 2012.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *Journal of Machine Learning Research*, 21(43):1–36, 2020a.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33, 2020b.

Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *arXiv preprint arXiv:1905.11971*, 2019.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

Zhonghui You, Jinmian Ye, Kunming Li, Zenglin Xu, and Ping Wang. Adversarial noise layer: Regularize neural network by adding noise. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 909–913. IEEE, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. In *arXiv preprint arXiv:1906.00555*, 2019.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *arXiv preprint arXiv:1905.00877*, 2019a.

Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, pp. 1829–1839, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *arXiv preprint arXiv:2002.11242*, 2020.

Shufei Zhang, Kaizhu Huang, Jianke Zhu, and Yang Liu. Manifold adversarial training for supervised and semi-supervised learning. *Neural Networks*, 140:282–293, 2021. doi: 10.1016/j.neunet.2021.03.031. URL https://doi.org/10.1016/j.neunet.2021.03.031.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

A Appendix

A.1 Detailed Experiment Setting

All the experiments are conducted on Ubuntu 20.04.1 LTS, Python 3.7.9, Pytorch 1.8.0, CUDA 11.1.1, and the GPUs are Nvidia-RTX3090. All the models adopt WideResNet-28-10 as the basic model structure. Specifically, on all three datasets CIFAR-10, CIFAR-100 and SVHN, we train

Algorithm 1 PGD-AT with Perturbation Diversity

Input: dataset S, training epochs K, batch size n, learning rate γ, budget ϵ, attack iterations T
Parameter: model parameter θ

for $k = 1$ to K **do**
  for random batch $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n} \sim S$ **do**
    initialization: $\boldsymbol{x}' \leftarrow \boldsymbol{x} + U(\boldsymbol{x}, \epsilon)$, where $U(\boldsymbol{x}, \epsilon)$ is the uniform random vector at center $\boldsymbol{x}$ with radius $\pm\frac{\epsilon}{2}$
    PGD attack with PD (maximising the cross-entropy loss and perturbation diversity):
    for $t = 1$ to T **do**
      Calculate the diverse adversarial perturbation on a mini-batch of data:
      $\boldsymbol{Z} = \text{Concat}_{i=1}^{n}(\boldsymbol{x}_i' - \boldsymbol{x}_i)$
      $\mathbb{PD} = \det(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})$
      $x_i' \leftarrow \mathcal{P}_{S_x}(x_i' + \boldsymbol{\epsilon} \cdot \text{sign}(\nabla_{x_i'}(\mathcal{L}_\theta(x_i', y_i, \theta) + \log(\mathbb{PD}))))$
    end for
    adversarial training (updating model parameters):
    $\theta \leftarrow \theta - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathcal{L}(x_i', y_i, \theta)$
  end for
end for
return model parameter θ.

models using SGD with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1. The learning rate decays at epoch 60 and 90 with the rate 0.1. The attack iteration number during the training period follows the baselines, such as 7 for PGD-AT and TRADES, 1 for FS, and the attack step size is $2/255$ for PGD and TRADES, $8/255$ for FS. The attack budget ϵ is set to $8/255$ for all the methods, following many other related work.

Other training details are listed in Table 3. For the batch size, we tune empirically that 60 or 120 can lead to good performance. The hyperparameter means the ratio of the original objective and the proposed PD. Note that though originally TRADES suggested to perform training on clean data, we found in our experiments such practice achieved very poor performance on CIFAR-100 and SVHN. Therefore, we train the TRADES baseline and TRADES+PD with adversarial examples on these two datasets instead. The early stopping (Rice et al., 2020) is implemented in all the baseline methods and the proposed PD methods. All the reported results are obtained with the most robust models for each method.

Table 3: Detailed experiment setting

Methods	Dataset	Batch Size	Hyperparameter	Training Epoch
AT+PD	CIFAR-10	120	1	600
	CIFAR-100	60	50	600
	SVHN	120	1	600
FS+PD	CIFAR-10	60	1	600
	CIFAR-100	60	0.1	300
	SVHN	60	0.1	600
TRADES+PD	CIFAR-10	120	1	60
	CIFAR-100	120	10	60
	SVHN	120	1	60

The training procedure is provided in Alg. 1

A.2 Detailed model robustness on different baseline models

We show the robust accuracy on the three baseline frameworks under several white-box attacks on CIFAR-10, CIFAR-100, and SVHN in this section with the attack iterations $T = 20, 100$ for PGD (Madry et al., 2017) and CW (Carlini & Wagner, 2017).

As observed from Table 4, Table 5, and Table 6, overall, our proposed PD achieves a clear improvement over all the three baseline models on the adversarial examples. Even though our method may

reduce the standard accuracy very slightly, for the adversarial samples, our proposed PD improves the baselines by a large margin. **Particularly, with the implementation of PD, our approach AT+PD is** 30.3% **and** 27.92% **higher than the baseline AT under PGD20 and CW20 attack on SVHN respectively**.

We also evaluate our proposed PD on the more sophisticated AutoAttack (AA) (Croce & Hein, 2020) as shown in the last column of Table 4. It can be observed that the proposed PD performs fair under AA: PD wins in TRADES, loses in AT, and ties (or slightly loses) in FS. Recent studies show that approaches considering sample relationships actually fail to defend against AA attack such as FS and ATLD (Qian et al., 2021). We attribute this to the reason that attacks considering sample relationships are usually weak. While weakly attacked samples can support the model to learn a more smooth decision boundary in the high-dimensional space, they usually do not benefit finding the most precise gradient, resulting in vulnerabilities in the decision boundary that can be found and harnessed by more sophisticated attacks such as AA during test. We leave this phenomenon for future research.

Table 4: Robust accuracy on different baseline models on CIFAR-10

MODELS	CLEAN	ACCURACY UNDER WHITE-BOX ATTACK ($\epsilon = 8$)							
		FGSM	PGD20	PGD40	PGD100	CW20	CW40	CW100	AA
AT	86.35	68.15	54.66	54.39	54.32	53.66	53.45	53.42	44.04
AT+PD	86.97↑	71.96↑	64.14↑	63.15↑	62.42↑	57.28↑	56.28↑	55.81↑	42.15↓
TRADES	86.76	66.12	51.80	51.60	51.54	49.86	49.77	49.70	48.54
TRADES+PD	85.55↓	66.04↓	53.26↑	53.09↑	53.12↑	50.54↑	50.47↑	50.50↑	49.63↑
FS	90.00	78.40	70.50	70.30	68.60	62.40	62.10	60.60	36.64
FS+PD	89.99↓	79.37↑	72.02↑	70.53↑	69.57↑	64.26↑	62.76↑	61.70↑	36.37↓

Table 5: Robust accuracy on different baseline models on CIFAR-100

MODELS	CLEAN	ACCURACY UNDER WHITE-BOX ATTACK ($\epsilon = 8$)				
		FGSM	PGD20	PGD100	CW20	CW100
AT	59.90	28.50	22.60	22.30	23.20	23.00
AT+PD	64.35↑	39.76↑	31.80↑	31.66↑	24.08↑	23.42↑
TRADES	61.46	39.45	30.54	30.55	27.19	27.12
TRADES+PD	61.23↓	39.48↑	30.86↑	30.71↑	27.35↑	27.22↑
FS	73.90	61.00	47.20	46.20	34.60	30.60
FS+PD	72.72↓	74.77↑	49.75↑	49.35↑	36.25↑	36.19↑

Table 6: Robust accuracy on different baseline models on SVHN

MODELS	CLEAN	ACCURACY UNDER WHITE-BOX ATTACK ($\epsilon = 8$)				
		FGSM	PGD20	PGD100	CW20	CW100
AT	93.90	68.40	47.90	46.00	48.70	47.30
AT+PD	94.66↑	90.25↑	78.20↑	75.31↑	76.62↑	72.46↑
TRADES	93.90	77.42	61.54	60.75	58.05	57.66
TRADES+PD	94.70↑	85.47↑	65.06↑	62.22↑	62.19↑	60.38↑
FS	96.20	83.50	62.90	52.00	61.30	50.80
FS+PD	96.54↑	97.42↑	67.72↑	53.13↑	63.75↑	49.72↓

A.3 EFFECT ON BLACK-BOX ATTACK

We further examine the effects of PD on AT and AT+PD under transfer-based black-box attack. We take CIFAR-10 as one typical example to illustrate such results. Four different models are used for generating test time attacks including the Vanilla Training model, AT model, FS model, and our

AT+PD model. As shown in Table 7, our proposed PD can improve AT in 9 cases, and is only slightly inferior to the baseline AT in 3 cases.

Table 7: Accuracy under transfer-based black-box attack on CIFAR-10

DEFENSE MODELS	ATTACKED MODELS (CIFAR-10)											
	VANILLA TRAINING			AT			FS			AT+PD		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
AT	85.35	85.53	85.35	73.06	64.49	63.82	82.97	81.30	80.12	85.24	84.80	84.79
AT+PD	85.81	86.09	86.11	73.80	65.69	65.41	82.36	79.98	78.11	85.85	85.29	85.49

A.4 PROOF FOR THEORY 3.1

In this section, we provide the detailed proof for Theorem 3.1.

Theorem 3.1 *Given the training set $\mathbb{S}_d = \{\boldsymbol{x}_i\}_{i=1}^N$ that consists of N i.i.d samples drawn from a distribution $\mathbb{S}$ with K subsets $\{\mathbb{C}_j\}_{j=1}^K$ where we assume that adversarial perturbations of each subset share the similar direction, and the set of adversarial perturbations of training set is $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$, if the loss function $l(\cdot)$ of DNN f_θ is k-Lipschitz, then for any $\delta > 0$, with the probability at least $1-\delta$, we have*

$$\mathrm{RGE} \leq \mathrm{GE} + \frac{k}{N}\sum_{j=1}^{K}\sum_{i\in N_j}\|\boldsymbol{W}_{ij}(\boldsymbol{\epsilon}_i - \hat{\boldsymbol{\epsilon}_j})\|_2^2 + M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{N}} \qquad \textit{where}$$
$$\hat{\boldsymbol{\epsilon}}_j = \mathbb{E}[\boldsymbol{z}^{adv} - \boldsymbol{z} | \boldsymbol{z} \in \mathbb{C}_j] \tag{9}$$

where $\boldsymbol{z}$ is data sampled from $\mathbb{C}_j$ with corresponding adversarial example $\boldsymbol{z}^{adv}$, N_i denotes the set of index of training data which belongs to $\mathbb{C}_i$, $\boldsymbol{W}_{ij}$ is the transformation matrix, and M is the upper bound of loss of the whole data manifold $\mathbb{S}$. We also assume that the diversity of $\{\hat{\boldsymbol{\epsilon}_j}\}_{j=1}^K$ is greater than $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$.

Proof: Let N_i be the set of index of points of training set $\mathbb{S}_d = \{\boldsymbol{s}_i\}_{i=1}^N$ that fall into the $\mathbb{C}_i$ and $(|N_1|, ..., |N_K|)$ is an i.i.d multinomial random variable with parameters n and $(\mu(\mathbb{C}_1), ..., \mu(\mathbb{C}_K))$. The following holds by the Breteganolle-Huber-Carol inequality (cf Proposition A6.6 of Van & Wellner (2000)):

$$Pr\left\{\sum_{i=1}^{K}\left|\frac{N_i}{N} - \mu(\mathbb{C}_j)\right| \geq \lambda\right\} \leq 2^K exp(\frac{-N\lambda^2}{2}) \tag{10}$$

Hence, with the probability at least $1-\delta$, we have:

$$\sum_{j=1}^{K}\left|\frac{N_j}{N} - \mu(\mathbb{C}_j)\right| \leq \sqrt{\frac{2Kln2 + 2ln(1/\delta)}{N}} \tag{11}$$

The upper bound of the robust generalisation can be formulated as:

$$
\begin{aligned}
&|l(f_\theta(\mathbb{S}^{adv}),\mathbb{Y}) - \hat{l}(f_\theta(\mathbb{S}_d^{adv}),\mathbb{Y}_d)| = \left|\sum_{j=1}^{K}\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j)\mu(\mathbb{C}_j) - \frac{1}{N}\sum_{i=1}^{N} l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i)\right| \\
&= |\sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) - \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) + \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|z\in\mathbb{C}_j))\mu(\mathbb{C}_j) \\
&- \frac{1}{N}\sum_{i=1}^{N}(l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i) - l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i) + l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i))| \\
&\leq |\sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) - \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j))\mu(\mathbb{C}_j) \\
&- \frac{1}{N}\sum_{i=1}^{N}(l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i) - l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i))| + \left|\sum_{j=1}^{K}\mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})z\in\mathbb{C}_j)(\mathbb{C}_j) - \frac{1}{N}\sum_{i=1}^{N} l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i)\right| \\
&\leq \text{GE} + \left|\sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) - \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j))\frac{|N_j|}{N} - \frac{1}{N}\sum_{i=1}^{N}(l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i) - l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i))\right| \\
&+ |\sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) - \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j))\mu(\mathbb{C}_j) - \sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) \\
&- \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j))\frac{|N_j|}{N}| \\
&\leq \text{GE} + \left|\sum_{j=1}^{K}(\mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j) - \mathbb{E}(l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j))\frac{|N_j|}{N} - \frac{1}{N}\sum_{i=1}^{N}(l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i) - l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i))\right| \\
&+ M\sum_{j=1}^{K}\left|\frac{N_j}{N} - \mu(\mathbb{C}_j)\right| \\
&\leq \text{GE} + \frac{1}{N}\sum_{j=1}^{K}\sum_{i\in N_j}\left|(l(f_\theta(\boldsymbol{x}_i^{adv}),\boldsymbol{y}_i) - l(f_\theta(\boldsymbol{x}_i),\boldsymbol{y}_i)) - \mathbb{E}(l(f_\theta(\boldsymbol{z}^{adv}),\boldsymbol{y}) - l(f_\theta(\boldsymbol{z}),\boldsymbol{y})|\boldsymbol{z}\in\mathbb{C}_j)\right| \\
&+ M\sum_{j=1}^{K}\left|\frac{N_j}{N} - \mu(\mathbb{C}_j)\right|
\end{aligned}
$$

Here, we assume $|l(f_\theta(\boldsymbol{x}_1),\boldsymbol{y}_1) - l(f_\theta(\boldsymbol{x}_2),\boldsymbol{y}_2)| \leq k\|f_\theta(\boldsymbol{x}_1) - f_\theta(\boldsymbol{x}_2)\|_2^2$ and $\boldsymbol{\epsilon}$ is small enough that $f_\theta(\boldsymbol{x}^{adv})$ can be approximated by its first order Taylor expansion. Then we have

$$
\begin{aligned}
&\text{RGE} \triangleq |l(f_\theta(\mathbb{S}^{adv}),\mathbb{Y}) - \hat{l}(f_\theta(\mathbb{S}_d^{adv}),\mathbb{Y}_d)| \\
&\leq \text{GE} + \frac{k}{N}\sum_{j=1}^{K}\sum_{i\in N_i}\|(f_\theta(\boldsymbol{x}_i^{adv}) - f_\theta(\boldsymbol{x}_i)) - \mathbb{E}(f_\theta(\boldsymbol{z}^{adv}) - f_\theta(\boldsymbol{z})|\boldsymbol{z}\in\mathbb{C}_j)\|_2^2 + M\sum_{j=1}^{K}\left|\frac{N_i}{N} - \mu(\mathbb{C}_j)\right| \\
&\leq \text{GE} + \frac{k}{N}\sum_{j=1}^{K}\sum_{i\in N_i}\|\nabla_x f_\theta(\boldsymbol{x})(\boldsymbol{x}_i^{adv} - \boldsymbol{x}_i) - \mathbb{E}(\nabla_z f(z)(\boldsymbol{z}^{adv} - \boldsymbol{z})|\boldsymbol{z}\in\mathbb{C}_j)\|_2^2 + M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{N}}
\end{aligned}
$$

There exists a transformation matrix $\boldsymbol{W}_{ij}$ such that

$$
\text{RGE} \leq \text{GE} + \frac{k}{N}\sum_{j=1}^{K}\sum_{i\in N_i}\|\boldsymbol{W}_{ij}(\boldsymbol{\epsilon}_i - \hat{\boldsymbol{\epsilon}}_j)\|_2^2 + M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{N}}
$$

where $\hat{\boldsymbol{\epsilon}}_j = \mathbb{E}[\boldsymbol{z}^{adv} - \boldsymbol{z} | \boldsymbol{z} \in \mathbb{C}_j]$.

This completes the proof. $\square$