On robust overfitting: adversarial training induced distribution matters

Runzhi Tian
School of EECS
University of Ottawa
Ottawa, Canada
{rtian081}@uottawa.ca

Yongyi Mao
Department of EECS
University of Ottawa
Ottawa, Canada
{ymao}@uottawa.ca

Abstract

Robust overfitting has been observed to arise in adversarial training. We hypothesize that this phenomenon may be related to the evolution of the data distribution along the training trajectory. To investigate this, we select a set of checkpoints in adversarial training and perform standard training on distributions induced by adversarial perturbation w.r.t the checkpoints. We observe that the obtained models become increasingly harder to generalize when robust overfitting occurs, thereby validating the hypothesis. We show the hardness of generalization on the induced distributions is related to certain local property of the perturbation operator at each checkpoint. The connection between the local property and the generalization on the induced distribution is proved by establishing an upper bound of the generalization error. Other interesting phenomena related to the adversarial training trajectory are also observed.

1 Introduction

Deep neural networks (DNNs) are known to be vulnerable to adversarial attacks where a carefully designed perturbation may cause the network to make a wrong prediction [18, 7]. Many methods have been proposed to improve the robustness of DNNs against adversarial perturbations [13, 24, 3], among which PGD-based adversarial training (PGD-AT) [13] is arguably the most effective [1, 5]. A recent work in Rice et al. [14] however revealed that PGD-AT can cause *robust overfitting*: a very high robust error (i.e., error on the adversarially perturbed instances) appears on the testing set (e.g., 44.19% on CIFAR-10) with a nearly zero robust error achieved on the training set. This is in sharp contrast with the standard classification where a significantly lower standard error (i.e., error on the unperturbed instances) can be achieved on the testing set (e.g., 4% on CIFAR-10).

Since its discovery, robust overfitting has attracted significant research attention. A great deal of research effort has been spent on understanding its cause and various explanations have been proposed. These include correlating robust overfitting with the flatness of the loss landscape [20, 17, 2], the curvature of activation functions [16], the presence of label noise [4], the phenomenon of memorization during adversarial training [6], training examples with small adversarial loss [22] and the non-smoothness loss used in adversarial training [10]. A line of mitigation techniques are also proposed based on the corresponding analysis, although each shown to only reduce the testing robust error by a few percent. The work in Hameed and Buesser [8] also point out that the explanations in Dong et al. [6] and Yu et al. [22] appear to conflict to each other. These indicate that the current understanding of robust overfitting is still arguably far from being conclusive. The robust overfitting can be due to a multitude of sources, the full picture remaining obscure.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

This work aims at further understanding robust overfitting. The inspiration of this work stems from the recognition that along adversarial training, adversarial perturbation effectively induces a new data distribution, say $\tilde{D}_t$, at training step t. This distribution, different from the original data distribution $\mathcal{D}$, continuously evolves in a fashion that depends on the current model parameter θ_t. A question then naturally arises: does robust overfitting have anything to do with this evolution of data distribution? We conducted a set of experiments, in which we inspect whether how well a model trained on a sample drawn from $\tilde{\mathcal{D}}_t$ (under standard training) generalize. Our experimental results suggest that robust overfitting may indeed correlates with generalization difficulty inherent in the induced distribution $\tilde{D}_t$ and our further theoretical analysis reveals that such difficulty of generalization is governed by a local "dispersion property" of the adversarial perturbation that induces $\tilde{D}_t$. The conclusions are validated by empirical observations across different datasets.

2 Adversarial training and induced distributions

Given an input space $\mathcal{X} \subseteq \mathbb{R}^d$, a label space $\mathcal{Y} := \{1, 2, \cdots, K\}$, a data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, a model parameter space $\Theta \subseteq \mathbb{R}^n$ and a loss function $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ parameterized by $\theta \in \Theta$, we define the robust population risk (or robust testing error)[1] as:

$$R^{\text{rob}}_{\mathcal{D}}(\theta) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{v \in \mathbb{B}(x,\epsilon)} l_\theta(v, y) \right] \tag{1}$$

where we have chosen $\mathbb{B}(x, \epsilon) := \{t \in \mathbb{R}^d : \|t - x\|_\infty \leq \epsilon\}$ as the ∞-norm ball around x with the radius ϵ. To find a model parameter θ that minimizes $R^{\text{rob}}_{\mathcal{D}}(\theta)$ but only with access to a training set $S := \{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d from $\mathcal{D}$, in practice, a natural choice is to minimize an empirical version of $R^{\text{rob}}_{\mathcal{D}}(\theta)$, that is to solve

$$\min_{\theta\in\Theta} R^{\text{rob}}_S(\theta), \quad \text{where} \quad R^{\text{rob}}_S(\theta) := \frac{1}{m} \sum_{i=1}^{m} \max_{v_i \in \mathbb{B}(x_i,\epsilon)} l_\theta(v_i, y_i) \tag{2}$$

The most popular adversarial training technique for solving this problem is iterating between solving the inner maximization via k-step projected gradient descend (PGD) and updating θ through stochastic gradient descent. We now give a concise explanation of this procedure.

k-step PGD A k-step PGD can be described by k-fold composition of an one-step PGD mapping. With a fixed choice of $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\theta \in \Theta$, the one-step PGD mapping $\mathcal{A}_{x,y,\theta} : \mathbb{R}^d \to \mathbb{B}(x, \epsilon)$ is defined as

$$\mathcal{A}_{x,y,\theta}(x') := \Pi_{\mathbb{B}(x,\epsilon)} \left[x' + \lambda \text{sgn} \left(\nabla_{x'} l_\theta(x', y) \right) \right] \tag{3}$$

Here $\Pi_{\mathbb{B}(x,\epsilon)} : \mathbb{R}^d \to \mathbb{B}(x, \epsilon)$ denotes the operation of projecting onto the set $\mathbb{B}(x, \epsilon)$ and $\lambda \in \mathbb{R}_+$ is a hyperparameter. The k-step PGD mapping $\mathcal{Q}_{x,y,\theta} : \mathbb{R}^d \to \mathbb{B}(x, \epsilon)$ is then

$$\mathcal{Q}_{x,y,\theta}(x') := (\underbrace{\mathcal{A}_{x,y,\theta} \circ \cdots \circ \mathcal{A}_{x,y,\theta}}_{k \text{ times}})(x') \tag{4}$$

Iterations of Adversarial Training In PGD-AT, the process of generating a perturbed example (v, y) from an example (x, y) w.r.t a model parameter θ can be described as

$$v = \mathcal{Q}_{x,y,\theta}(x + \rho) \tag{5}$$

where ρ is drawn from $\mathcal{U}([-\epsilon, +\epsilon]^d)$, the uniform distribution over the d-dimensional cubic $[-\epsilon, +\epsilon]^d$. At iteration t, where the model parameter is θ_t, the solution of the inner maximization $\max_{v\in\mathbb{B}(x_i,\epsilon)} l_{\theta_t}(x_i, y_i)$ is taken as $l_{\theta_t}(v_i, y_i)$, and the model parameter is updated by

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \left[\frac{1}{m} \sum_{i=1}^{m} l_{\theta_t} (v_i, y_i) \right] \tag{6}$$

It is worth noting that the stochastic mapping (5) that perturbs (x, y) to (v, y) depends on θ. Thus the distribution of $(v, y) = (\mathcal{Q}_{x,y,\theta_t}(x + \rho), y)$ at iteration t depends on θ_t. We will denote this

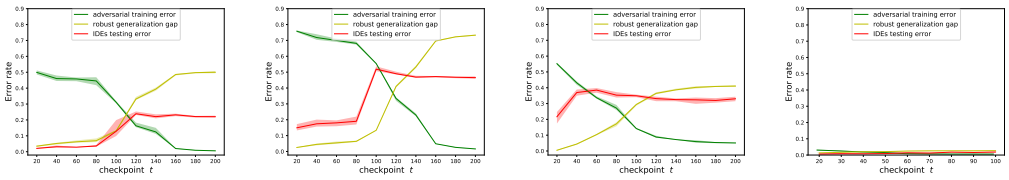
[1] In practice, the robust population risk is estimated by computing the error rate on a given testing set with the inner maximization approximately solved by some choice of adversarial attack algorithm.

distribution by $\tilde{\mathcal{D}}_{\theta_t}$, or simply $\tilde{\mathcal{D}}_t$, and refer to it as the *(adversarial training) induced distribution* at iteration t. Then at iteration t, we may regard the perturbed examples $\{(v_i, y_i)\}$ as an i.i.d. sample from $\tilde{\mathcal{D}}_t$. Note that the distribution $\tilde{\mathcal{D}}_t$ evolves with θ_t during training and in turn affects the update of θ_t. This dynamic interplay implies that the evolution of $\tilde{\mathcal{D}}_t$ may significantly affect the robust generalization perform of θ_t.

3 Training on the induced distributions

The following experiments are conducted. First PGD-AT is performed on a training set S. Along this process, for a prescribed set of training iterations (or "checkpoints") $\{t_j : j = 1, 2, \ldots, N\}$, the model parameter θ_{t_j} at each checkpoint t_j is saved. Then at each checkpoint t_j, each example in the training set S is perturbed according to (5) with $\theta = \theta_{t_j}$, giving rise to the perturbed training set $\tilde{S}_{t_j}$. The testing set T is also similarly perturbed, giving rise to perturbed testing set $\tilde{T}_{t_j}$. The model is then retrained fully on $\tilde{S}_{t_j}$, using standard training (i.e, without perturbation) with random initialization. The resulting model is tested on $\tilde{T}_{t_j}$. Note that in this setting, both $\tilde{S}_{t_j}$ and $\tilde{T}_{t_j}$ are i.i.d. samples from $\tilde{\mathcal{D}}_{t_j}$. We call these experiments "induced distribution experiment" (IDE) for the ease of reference.

The experiments are conducted on MNIST [12], CIFAR10, CIFAR100[11] and a "scaled-down" version of the ImageNet dataset [15] (referred as "Reduced ImageNet"). The detailed experimental setup is introduced in Appendix 6. For each dataset, the experiments are repeated five times with different random seeds. The experimental results are shown in Figure 1 where the green and yellow curves individually plot the evolution of the robust training error $R_S^{\text{rob}}(\theta_t)$ and the robust generalization gap $|R_{\mathcal{D}}^{\text{rob}}(\theta_t) - R_S^{\text{rob}}(\theta_t)|$. The red curves in each figures plot the testing error of each IDEs w.r.t different PGD-AT checkpoints. In Figure 1 (a)-(c), a significant rise in the average IDE testing error is observed.[2] This shift coincides with the onset of robust overfitting, where a substantial increase in robust generalization gap appears. On the other hand, the experiments on MNIST (see Figure 1 (d)) shows that the absence of robust overfitting coincides with the consistently low IDE testing error. We also conduct additional experiments on CIFAR-10 to demonstrate this correlation between the IDE testing error and the robust overfitting (see Appendix 7).



(a) CIFAR-10 (b) CIFAR-100 (c) Reduced ImageNet (d) MNIST

Figure 1: Adversarial training and the corresponding IDE results across different datasets.

At this end, we have established that the increasing difficulty of generalization inherent in the induced distribution plays an important role in robust overfitting. It remains curious what causes $\tilde{\mathcal{D}}_t$ to become harder to generalize in adversarial training. We provide a theoretical explanation in the next section.

4 Generalization properties of the induced distributions

As a start, we first introduce a notion characterizing a local property of the perturbation map $Q_{x,y,\theta}$, through which $\tilde{D}_\theta$ is induced.

[2]Note that $\tilde{\mathcal{D}}_t$ is "not far" from the original data distribution $\mathcal{D}$, since perturbation at every iteration is restricted to a small neighborhood of x. It is interesting to observe that generalization on $\tilde{\mathcal{D}}_t$ can become much harder than $\mathcal{D}$ (e.g., 4% error rate on the original CIFAR-10 testing set can be easily achieved compared to 23.89% error rate from the IDE at the 120th checkpoint) despite $\tilde{\mathcal{D}}_t$ and $\mathcal{D}$ are "close to" each other.

Definition 1. Let $(\mathcal{X}', \|\cdot\|_2)$ be a norm space equipped with the 2-norm. Given a map $T : \mathcal{X} \to \mathcal{X}'$ and an arbitrary bounded measurable subset C of $\mathcal{X}$, we define the C-dispersion of T by

$$\gamma_C(T) := \mathbb{E}_{x,x' \sim \mathcal{U}(C)} \|T(x) - T(x')\|_2^2 \tag{7}$$

where $\mathcal{U}(C)$ denotes a uniform distribution over C. Intuitively, this quantity measures on average how far two random points in C spread after being mapped by T. Now restricting $T = Q_{x,y,\theta}$ and $C = \mathbb{B}(x, \epsilon)$, we have

$$\gamma_{\mathbb{B}(x,\epsilon)}(\mathcal{Q}_{x,y,\theta}) = \mathbb{E}_{\rho,\rho' \sim \mathcal{U}([-\epsilon,+\epsilon]^d)} \|\mathcal{Q}_{x,y,\theta}(x+\rho) - \mathcal{Q}_{x,y,\theta}(x+\rho')\|_2^2.$$

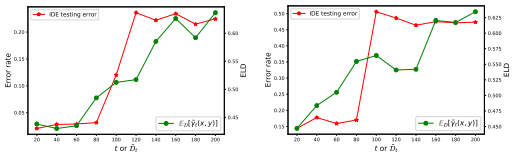
For simplicity we rewrite this quantity as $\tilde{\gamma}_\theta(x, y)$ and refer to it as the *local dispersion* of the perturbation (family)[3] $\mathcal{Q}_\theta$ at (x, y). In our experiments, to estimate $\tilde{\gamma}_\theta(x, y)$, we sample 10 pairs of (ρ, ρ') and approximate the expectation by the sample mean.

We now consider a standard classification problem with $\tilde{\mathcal{D}}_\theta$ as the underlying data distribution over $\mathcal{X} \times \mathcal{Y}$ and $\{(v_i, y_i)\}_{i=1}^m$ as i.i.d sample from $\tilde{\mathcal{D}}_\theta$. Let $\mathcal{F}$ be a hypothesis class for this learning problem, where each member $f \in \mathcal{F}$ is a function mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Note that the hypothesis class $\mathcal{F}$ may have not be related to the model used for adversarial training in any way. In the following, we show that for each $f \in \mathcal{F}$, the *generalization gap* $\left| \frac{1}{m} \sum_{i=1}^m f(v_i, y_i) - \mathbb{E}_{(v,y)\sim\tilde{D}_\theta} f(v, y) \right|$ is related to the level of $\tilde{\gamma}_\theta(x, y)$.

Theorem 1. Let $f \in \mathcal{F}$ and suppose that f satisfies the following conditions: 1. Lipchitzness of f over $\mathcal{X}$: For any $y \in \mathcal{Y}$, $|f(x, y) - f(x', y)| \leq \beta \|x - x'\|_2$ for $\forall x, x' \in \mathcal{X}$. 2. Loss boundedness: $\sup_{x,y \in \mathcal{X}\times\mathcal{Y}} |f(x, y)| = B < \infty$. 3. Boundedness of perturbation-smoothed loss: $\sup_{x,y\in \mathrm{supp}(\mathcal{D})} |\mathbb{E}_\rho f(\mathcal{Q}_{x,y,\theta}(x+\rho), y)| = A < \infty$, where $\mathrm{supp}(\mathcal{D})$ denote the support of distribution $\mathcal{D}$. Then for any $\tau > 0$, with probability at least $1 - \tau$ over the i.i.d. draws of sample $\{(v_i, y_i)\}_{i=1}^m$ from $\tilde{\mathcal{D}}_\theta$,

$$\left|\frac{1}{m}\sum_{i=1}^m f(v_i, y_i) - \mathbb{E} f(v, y)\right| \leq \frac{2\beta}{\sqrt{m}} \sqrt{\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_\theta(x, y)} + \frac{2A}{\sqrt{m}} + 2B\sqrt{\frac{\log \frac{1}{\tau}}{2m}} \tag{8}$$

We leave the proof in Appendix 8. The theorem shows that a small generalization gap of f w.r.t to the distribution $\tilde{\mathcal{D}}_\theta$ can be achieved when the average local dispersion $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_\theta(x, y)$ is small.



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

Figure 2: The evolution of $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$ w.r.t t and the IDE testing error for each $\tilde{\mathcal{D}}_t$.

To verify this result, we inspect the evolution of $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_{\theta_t}(x, y)$ (or $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$ for simplicity) along the adversarial training trajectory and compare it with the IDE results. Figure 2 shows the results evaluated on the testing sets of CIFAR-10 and CIFAR-100, where $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$ is getting larger, correlating with the increasing difficulty of generalization on $\tilde{\mathcal{D}}_t$. The results are consistent with Theorem 1. Similar experimental results are also observed on the other datasets (see Appendix 9).

The theoretical analysis underscores the critical role played by the local properties of $\mathcal{Q}_{x,y,\theta_t}$ in affecting the generalization performance for $\tilde{\mathcal{D}}_t$ (and potentially robust generalization of θ_t). In Appendix 10, we present interesting findings regarding other local properties of $\mathcal{Q}_{x,y,\theta_t}$ beyond local dispersion. This brings various additional insights towards the dynamic of adversarial training.

[3]The perturbation family refers to $\{\mathcal{Q}_{x,y,\theta} : (x, y) \in \mathcal{X} \times \mathcal{Y}\}$, denoted by $\mathcal{Q}_\theta$

5 Conclusion

In this paper, we show that adversarial perturbation induced distribution plays an important role in robust overfitting. In particular, we observe experimentally that the increasing generalization difficulty of the induced distribution along the training trajectory is correlated with robust overfitting. Our theoretical analysis suggests that a key factor governing this difficulty is the local dispersion of the perturbation. The theoretical result is validated by experiments. Remarkably, through this work, we demonstrate that the trajectory of adversarial training plays an important role in robust overfitting. Studying the dynamics of adversarial training is arguably a promising approach to developing deeper understanding of this topic. In particular, we speculate that studying the effect of gradient-based parameter update may provide additional insight.

References

[1] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *CoRR*, abs/1802.00420, 2018. URL `http://arxiv.org/abs/1802.00420`.

[2] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=qZzy5urZw9`.

[3] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *CoRR*, abs/2010.09670, 2020. URL `https://arxiv.org/abs/2010.09670`.

[4] Chengyu Dong, Liyuan Liu, and Jingbo Shang. Double descent in adversarial training: An implicit label noise perspective. *CoRR*, abs/2110.03135, 2021. URL `https://arxiv.org/abs/2110.03135`.

[5] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 321–331, 2020.

[6] Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, and Jun Zhu. Exploring memorization in adversarial training. *CoRR*, abs/2106.01606, 2021. URL `https://arxiv.org/abs/2106.01606`.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[8] Muhammad Zaid Hameed and Beat Buesser. Boundary adversarial examples against adversarial overfitting, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016. URL `http://arxiv.org/abs/1603.05027`.

[10] Sekitoshi Kanai, Masanori Yamada, Hiroshi Takahashi, Yuki Yamanaka, and Yasutoshi Ida. Relationship between nonsmoothness in adversarial training, constraints of attacks, and flatness in the input space. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[14] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. *CoRR*, abs/2002.11569, 2020. URL `https://arxiv.org/abs/2002.11569`.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[16] Vasu Singla, Sahil Singla, David Jacobs, and Soheil Feizi. Low curvature activations reduce overfitting in adversarial training. *CoRR*, abs/2102.07861, 2021. URL `https://arxiv.org/abs/2102.07861`.

[17] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. *CoRR*, abs/2104.04448, 2021. URL `https://arxiv.org/abs/2104.04448`.

[18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[19] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019.

[20] Dongxian Wu, Yisen Wang, and Shutao Xia. Revisiting loss landscape for adversarial robustness. *CoRR*, abs/2004.05884, 2020. URL `https://arxiv.org/abs/2004.05884`.

[21] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. Adversarial robustness through local lipschitzness. *CoRR*, abs/2003.02460, 2020. URL `https://arxiv.org/abs/2003.02460`.

[22] Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond, 2022.

[23] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. URL `http://arxiv.org/abs/1605.07146`.

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *CoRR*, abs/1901.08573, 2019. URL `http://arxiv.org/abs/1901.08573`.

6 Detailed Experimental Setup

Reduced ImageNet Given that adversarial training is known to be significantly challenging and computationally expensive on the full-scale ImageNet dataset, we draw inspiration from the approach presented in Tsipras et al. [19] and made a Reduced ImageNet by aggregating several semantically similar subsets of the original ImageNet, resulting in a total of 66594 images. This dataset is then partitioned into a training set containing 5,000 images per class and a testing set containing approximately 1,000 images per class. Compared to the restricted ImageNet in Tsipras et al. [19], our dataset has a more balanced sample size across each classes. Table 1 illustrates the specific classes from the original ImageNet that have been aggregated in our dataset.

Classes in the reduced ImageNet	Classes in ImageNet
"dog"	86 to 90
"cat"	(8,10,55,95,174)
"truck"	279 to 283
"car"	272 to 276
"beetles"	623 to 627
"turtle"	458 to 462
"crab"	612 to 616
"fish"	450 to 454
"snake"	477 to 481
"spider"	604 to 608

Table 1: The left column presents the classes within our reduced ImageNet dataset, with each class being an aggregation of the corresponding classes from the full-scale ImageNet dataset, as depicted in the right column.

Settings for adversarial training We use the following settings for adversarial training: For MNIST, following the settings in Madry et al. [13], we train a small CNN model using 40-step PGD with step size $\lambda = 0.01$ and perturbation radius $\epsilon = 0.3$. For the other three datasets, we train the pre-activation ResNet (PRN) model [9] and the Wide ResNet (WRN) model [23]. We use 5-step PGD with $\epsilon = 4/255$ for the Reduced ImageNet and 10-step PGD with $\epsilon = 8/255$ for CIFAR-10 and CIFAR-100 according to Rice et al. [14] in adversarial training. We set $\lambda = 2/255$ on CIFAR10 and CIFAR100, $\lambda = 0.9/255$ on the reduced ImageNet. The settings on different datasets are summarized in Table 2. Data augmentation is performed on these datasets during the training except for MNIST. For CIFAR-10 and CIFAR-100 we follow the data augmentation setting in Rice et al. [14]. For our reduced ImageNet, we adopt the same data augmentation scheme that is used on the restricted ImageNet in Yang et al. [21].

	MNIST	CIFAR-10	CIFAR-100	Reduced ImageNet
model	small CNN	PRN18&WRN-34	WRN-34	PRN-50
optimizer	Adam	SGD	SGD	SGD
weight deacy	None	5×10^{-4}	5×10^{-4}	None
batch size	128	128	128	128
ϵ	0.3	8/255	8/255	4/255
λ	0.01	2/255	2/255	0.9/255
number of PGD	40	10	10	5

Table 2: Settings in adversarial training across different datasets

Settings for IDE For the IDEs on each datasets, the settings are outlined in Table 3. It is important to note that for each of the individual IDEs conducted on the same dataset, we maintain consistent training settings. This includes using the same model architecture with identical model size and the same level of regularization. This ensures a fair comparison of the IDE results obtained from the same dataset. Furthermore, the model is trained to achieve zero training error in all the IDEs, excluding the situation that the degeneration in model performance could be attributed to inadequate training procedures.

	MNIST	CIFAR-10	CIFAR-100	Reduced ImageNet
model	small CNN	WRN-34	WRN-34	PRN-50
optimizer	Adam	SGD	SGD	SGD
weight deacy	None	5×10^{-4}	5×10^{-4}	5×10^{-4}
batch size	128	128	128	128

Table 3: Settings in the IDE across different datasets

7 Omitted IDE Results

Figure 3 shows results from additional experiments on CIFAR-10. In these experiments, we perform adversarial training with different level of weight decay to control the level of robust overfitting. Subsequently, IDEs are conducted for each such variant of adversarial training. In Figure 3, each distinct color corresponds to a different weight decay factor utilized in adversarial training. Within each color category, the dashed curves and the corresponding solid lines represent, respectively, the robust generalization gaps and the IDE results associated with that specific adversarial training variant. As anticipated, increasing the weight decay factor results in a notable reduction in the robust generalization gap, while conversely, decreasing the weight decay factor leads to the opposite effect. This is shown by the downward shift in the dashed curves across the three color categories. Additionally, a clear synchronization can be observed between each pair of dashed and solid curves, with lower dashed curves consistently corresponding to lower solid curves in the same color category.

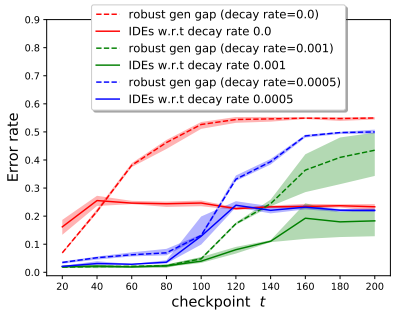


Figure 3: The outcomes of additional experiments conducted on CIFAR-10. In the experiments, we perform adversarial training with various weight decay rates and conduct IDEs for each of the adversarial training variant. The blue curves are reproduced from Figure1 (a), serving as a reference for a clear comparison. The results further solidify the correlation between the robust overfitting and the IDE testing error.

8 Proof for Theorem 1

We use the notations introduced in the main text. For shorter notations, let $z = (x, y)$, $u = (v, y)$ and $f(u) := f(v, y)$. We write $\mathcal{Q}_{x,y,\theta}$ as $\mathcal{Q}_z$, since our derivation does not explicitly depend on the choice of θ.

Denote by $g(u_1 \cdots u_m) := \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \mathbb{E}f(u) \right|$. We have for any $1 \leq j \leq m$

$$\sup_{u_1,\cdots,u_m,u_j'} \left| g(u_1,\cdots,u_m) - g(u_1,\cdots,u_j',u_{j+1},\cdots u_m) \right| \tag{9}$$

$$= \sup_{u_1,\cdots,u_m,u_j'} \left| \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \mathbb{E}f(u) \right| - \left| \frac{1}{m} \left(\sum_{i=1,i\neq j}^{m} f(u_i) + f(u_j') \right) - \mathbb{E}_u f(u) \right| \right| \tag{10}$$

$$\leq \sup_{u_1,\cdots,u_m,u_j'} \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \mathbb{E}_u f(u) - \frac{1}{m} \left(\sum_{i=1,i\neq j}^{m} f(u_i) + f(u_j') \right) + \mathbb{E}_u f(u) \right| \tag{11}$$

$$= \sup_{u_j,u_j'} \frac{1}{m} \left| f(u_j) - f(u_j') \right| \tag{12}$$

$$\leq \frac{1}{m} \sup_{u_j} |f(u_j)| + \frac{1}{m} \sup_{u_j'} \left| f(u_j') \right| \tag{13}$$

$$\leq \frac{2B}{m} \tag{14}$$

where the inequality (11) follows from the inverse triangle inequality. The inequality (13) and (14) make use of the triangle inequality and the boundedness condition of f.

With the result derived above, by McDiarmid inequality, we have for all $\mu > 0$

$$\Pr\left[g(u_1 \cdots u_m) - \mathbb{E}_U g(u_1 \cdots u_m) \geq \mu \right] \leq \exp\left(\frac{-m\mu^2}{B} \right)$$

This is equivalent to saying that with probability $1 - \tau$, we have

$$g(u_1 \cdots u_m) \leq \mathbb{E}_U g(u_1 \cdots u_m) + 2B \sqrt{\frac{\log \frac{1}{\tau}}{2m}} \tag{15}$$

Given this, the following parts aim at constructing an upper bound for the term $\mathbb{E}_U g(u_1 \cdots u_m)$.

For shorter notation, let $U := (u_1, \cdots, u_m)$, $Z := (z_1, \cdots, z_m)$, $\Gamma := (\rho_1, \cdots, \rho_m)$, $F(Z, \Gamma) := \frac{1}{m} \sum_{i=1}^{m} f(\mathcal{Q}_{z_i}(x_i + \rho_i), y_i)$. We have

$$\mathbb{E}_U g(u_1 \cdots u_m) \tag{16}$$

$$= \mathbb{E}_U \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \mathbb{E}f(u) \right| \tag{17}$$

$$= \mathbb{E}_U \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \mathbb{E}_{\hat{U}} \left[\frac{1}{m} \sum_{i=1}^{m} f(\hat{u}_i) \right] \right| \tag{18}$$

$$\leq \mathbb{E}_U \mathbb{E}_{\hat{U}} \left| \frac{1}{m} \sum_{i=1}^{m} f(u_i) - \frac{1}{m} \sum_{i=1}^{m} f(\hat{u}_i) \right| \tag{19}$$

$$= \mathbb{E}_Z \mathbb{E}_\Gamma \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \left| \frac{1}{m} \sum_{i=1}^{m} f(\mathcal{Q}_{z_i}(x_i + \rho_i), y_i) - \frac{1}{m} \sum_{i=1}^{m} f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i + \hat{\rho}_i), \hat{y}_i) \right| \tag{20}$$

$$= \mathbb{E}_Z \mathbb{E}_\Gamma \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \left| F(Z,\Gamma) - \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) + \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) - F(\hat{Z},\hat{\Gamma}) + \mathbb{E}_{\tilde{\Gamma}} F(\hat{Z},\tilde{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F(\hat{Z},\tilde{\Gamma}) \right| \tag{21}$$

$$\leq \mathbb{E}_Z \mathbb{E}_\Gamma \left| F(Z,\Gamma) - \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) \right| + \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \left| F(\hat{Z},\hat{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F(\hat{Z},\tilde{\Gamma}) \right| + \mathbb{E}_Z \mathbb{E}_{\hat{Z}} \left| \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F(\hat{Z},\tilde{\Gamma}) \right| \tag{22}$$

$$= \underbrace{2\mathbb{E}_Z \mathbb{E}_\Gamma \left| F(Z,\Gamma) - \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) \right|}_{①} + \underbrace{\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \left| \mathbb{E}_{\bar{\Gamma}} F(Z,\bar{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F(\hat{Z},\tilde{\Gamma}) \right|}_{②} \tag{23}$$

where (19) follows from Jensen's inequality and (22) is by the triangle inequality. We now individually construct upper bounds for the term ① and ②.

For the term ①, we have

$$
\begin{aligned}
& 2\mathbb{E}_Z\mathbb{E}_\Gamma \left|F(Z,\Gamma) - \mathbb{E}_{\bar{\Gamma}}F(Z,\bar{\Gamma})\right| && (24)\\
\leq & 2\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \left|F(Z,\Gamma) - F(Z,\bar{\Gamma})\right| && (25)\\
= & 2\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \left|\frac{1}{m}\sum_{i=1}^{m} f(\mathcal{Q}_{z_i}(x_i+\rho_i), y_i) - \frac{1}{m}\sum_{i=1}^{m} f(\mathcal{Q}_{z_i}(x_i+\bar{\rho}_i), y_i)\right| && (26)\\
= & \frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}}\mathbb{E}_\Sigma \left|\sum_{i=1}^{m} \sigma_i \left(f(\mathcal{Q}_{z_i}(x_i+\rho_i), y_i) - f(\mathcal{Q}_{z_i}(x_i+\bar{\rho}_i), y_i)\right)\right| && (27)\\
\leq & \frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sqrt{\sum_{i=1}^{m} \left|f(\mathcal{Q}_{z_i}(x_i+\rho_i), y_i) - f(\mathcal{Q}_{z_i}(x_i+\bar{\rho}_i), y_i)\right|^2} && (28)\\
\leq & \frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sqrt{\sum_{i=1}^{m} \beta^2 \|\mathcal{Q}_{z_i}(x_i+\rho_i) - \mathcal{Q}_{z_i}(x_i+\bar{\rho}_i)\|^2} && (29)\\
\leq & \frac{2\beta}{m}\mathbb{E}_Z \sqrt{\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \left[\sum_{i=1}^{m} \|\mathcal{Q}_{z_i}(x_i+\rho_i) - \mathcal{Q}_{z_i}(x_i+\bar{\rho}_i)\|^2\right]} && (30)\\
= & \frac{2\beta}{m}\mathbb{E}_Z \sqrt{\sum_{i=1}^{m} \mathbb{E}_\rho\mathbb{E}_{\bar{\rho}} \|\mathcal{Q}_{z_i}(x_i+\rho) - \mathcal{Q}_{z_i}(x_i+\bar{\rho})\|^2} && (31)\\
= & \frac{2\beta}{m}\mathbb{E}_Z \sqrt{\sum_{i=1}^{m} \gamma(x_i, y_i)} && (32)\\
\leq & \frac{2\beta}{m} \sqrt{\mathbb{E}_Z \left[\sum_{i=1}^{m} \gamma(x_i, y_i)\right]} && (33)\\
= & \frac{2\beta}{m} \sqrt{\sum_{i=1}^{m} \mathbb{E}_{z_i}\gamma(x_i, y_i)} && (34)\\
= & \frac{2\beta}{\sqrt{m}} \sqrt{\mathbb{E}_z\gamma(x, y)} && (35)
\end{aligned}
$$

Again, we apply Jensen's inequality to get (25). In (27), we introduce Rademacher variables $\Sigma := (\sigma_1, \cdots, \sigma_m)$ (i.e., each random variable σ_i takes values in $\{-1, +1\}$ independently with equal probability 0.5). The Rademacher variables introduces a random exchange of the corresponding difference term. Since Γ and $\hat{\Gamma}$ are independently sampled from the same distribution, such a swap gives an equally likely configuration. Therefore, the equality (27) holds. The inequality (28) is given by Khintchine's inequality. The inequality (29) makes use of the lipschitz condition of f. (30) is derived from Jensen's inequality and due to that square root is a concave function. (32) is by the definition of the local dispersion of $\mathcal{Q}_z$. Again, we apply Jensen's inequality to obtain (33). Equation (34) and (35) follow from the settings that each $z_i = (x_i, y_i)$ is i.i.d.

For the term ②, we have

$$\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left|\mathbb{E}_{\bar{\Gamma}}F(Z,\bar{\Gamma})-\mathbb{E}_{\tilde{\Gamma}}F(\hat{Z},\tilde{\Gamma})\right| \tag{36}$$

$$=\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left|\mathbb{E}_{\bar{\Gamma}}\left[\frac{1}{m}\sum_{i=1}^{m}f(\mathcal{Q}_{z_i}(x_i+\bar{\rho}_i),y_i)\right]-\mathbb{E}_{\tilde{\Gamma}}\left[\frac{1}{m}\sum_{i=1}^{m}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\tilde{\rho}_i),\hat{y}_i)\right]\right| \tag{37}$$

$$=\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\bar{\rho}_i}\left[f(\mathcal{Q}_{z_i}(x_i+\bar{\rho}_i),y_i)\right]-\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\tilde{\rho}_i}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\tilde{\rho}_i),\hat{y}_i)\right]\right| \tag{38}$$

$$=\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right]-\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right]\right| \tag{39}$$

$$=\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\mathbb{E}_{\Sigma}\left|\sum_{i=1}^{m}\sigma_i\left(\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right]-\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right]\right)\right| \tag{40}$$

$$\leq\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\sqrt{\sum_{i=1}^{m}\left|\left(\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right]-\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right]\right)\right|^2} \tag{41}$$

where equation (38) and (39) are due to each $\hat{\rho}_i$ and $\tilde{\rho}_i$ is i.i.d. Again, we introduce Rademacher variables at (40) and apply Khintchine's inequality to get (41). For the term $\left|\left(\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right]-\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right]\right)\right|^2$, we have

$$\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)-\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|^2 \tag{42}$$

$$\leq\left(\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right|+\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|\right)^2 \tag{43}$$

$$\leq 2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right|^2+2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|^2 \tag{44}$$

where inequality (44) is derived by the inequality $(a+b)^2\leq 2(a^2+b^2)$. Returning to (41), we then have

$$\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\sqrt{\sum_{i=1}^{m}\left|\left(\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right]-\mathbb{E}_{\rho}\left[f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right]\right)\right|^2}$$

$$\leq\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\sqrt{\sum_{i=1}^{m}2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right|^2+\sum_{i=1}^{m}2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|^2} \tag{45}$$

$$\leq\frac{1}{m}\sqrt{\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left[\sum_{i=1}^{m}2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right|^2+\sum_{i=1}^{m}2\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|^2\right]} \tag{46}$$

$$=\frac{1}{m}\sqrt{\sum_{i=1}^{m}2\mathbb{E}_{z_i}\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{z_i}(x_i+\rho),y_i)\right|^2+\sum_{i=1}^{m}2\mathbb{E}_{\hat{z}_i}\left|\mathbb{E}_{\rho}f(\mathcal{Q}_{\hat{z}_i}(\hat{x}_i+\rho),\hat{y}_i)\right|^2} \tag{47}$$

$$=\frac{2}{\sqrt{m}}\sqrt{\mathbb{E}_z\left|\mathbb{E}_{\rho}f(\mathcal{Q}_z(x+\rho),y)\right|^2} \tag{48}$$

$$\leq\frac{2}{\sqrt{m}}\sqrt{\sup_{z\in\mathrm{supp}(\mathcal{D})}\left|\mathbb{E}_{\rho}f(\mathcal{Q}_z(x+\rho),y)\right|^2} \tag{49}$$

$$=\frac{2A}{\sqrt{m}} \tag{50}$$

The final line is derived by the condition that $\sup_{z\in\mathrm{supp}(\mathcal{D})}\left|\mathbb{E}_{\rho}f(\mathcal{Q}_z(x+\rho),y)\right|=A$. This gives the final result

$$\mathbb{E}_U g(u_1\cdots u_m)\leq\frac{2\beta}{\sqrt{m}}\sqrt{\mathbb{E}_z\gamma(x,y)}+\frac{2A}{\sqrt{m}}$$

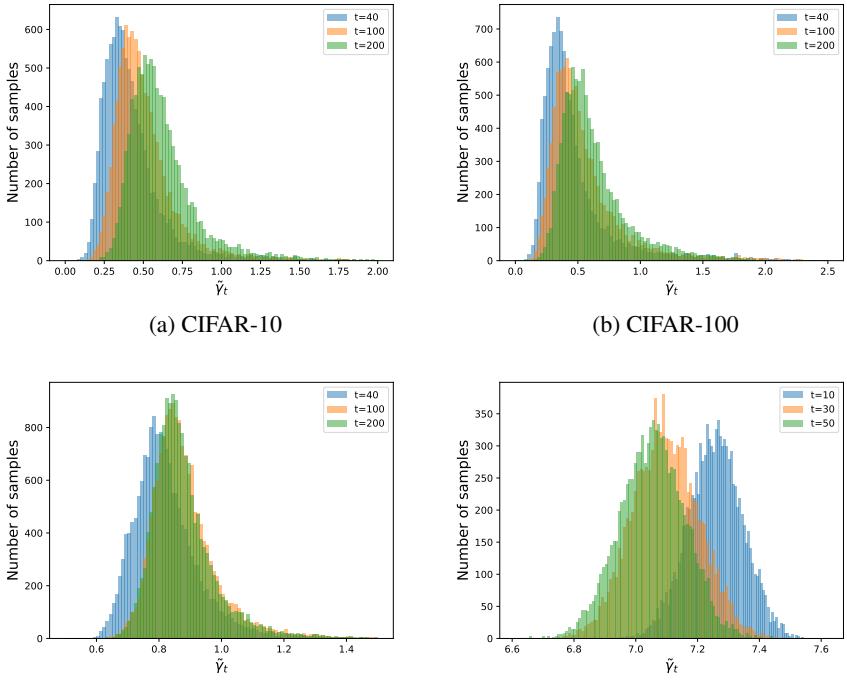
Plugging back to (15), we derive the bound in Theorem 1. This completes the proof. □

Lastly we want to remark that the bound is not trivial, since we have

$$\begin{aligned} A = & \sup_{z \in \mathrm{supp}(\mathcal{D})} |\mathbb{E}_\rho f(\mathcal{Q}_z(x+\rho), y)| \\ \leq & \sup_{z \in \mathrm{supp}(\mathcal{D})} \mathbb{E}_\rho |f(\mathcal{Q}_z(x+\rho), y)| \\ \leq & \sup_{z \in \mathcal{X} \times \mathcal{Y}} \sup_{\|\rho\|_\infty \leq \epsilon} |f(\mathcal{Q}_z(x+\rho), y)| \\ \leq & \sup_{v, y \in \mathcal{X} \times \mathcal{Y}} |f(v, y)| = B \end{aligned}$$

In fact, A could be much smaller than B, meaning the bound is tight.

9 Other Results towards Local Dispersion



(a) CIFAR-10

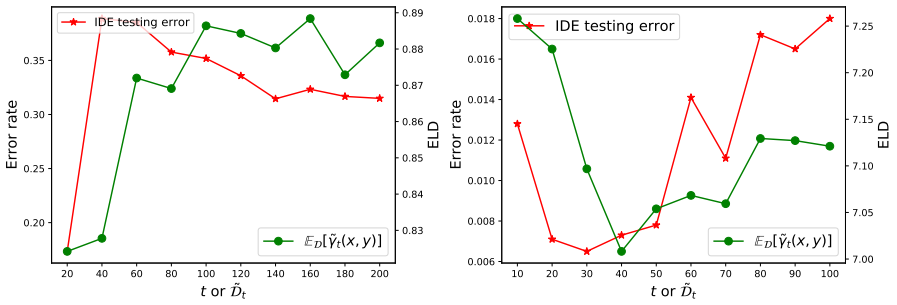
(b) CIFAR-100

(c) Reduced ImageNet

(d) MNIST

Figure 4: histograms of $\tilde{\gamma}_t$ on the CIFAR-10, CIFAR-100, Reduced ImageNet and MNIST testing set. On CIFAR-10, CIFAR-100 and the Reduced ImageNet, the mode of the histogram shifts towards a larger number, indicating the level of $\tilde{\gamma}_t$ increases along adversarial training. By sharp contrast, on MNIST, the mode of the histogram shifts toward a smaller value. This behaviour matches the IDE results and the generalization bound derived in Theorem 1.

We observed how the distribution of $\tilde{\gamma}_t(x, y)$ evolves along the adversarial training trajectory. Figure 4 plots the histogram of $\tilde{\gamma}_t(x, y)$ at three different training checkpoints on the testing sets of different datasets. In CIFAR-10, CIFAR-100 and the Reduced ImageNet, where robust overfitting appears, it is the clear that the distribution shifts to the right as adversarial training proceeds, indicating that perturbation map $\mathcal{Q}_{\theta_t}$ becomes more locally dispersed as adversarial training goes on. On the other hand, on MNIST the distribution of $\tilde{\gamma}_t(x, y)$ shifts to the left with the absence of robust overfitting and constantly low IDE testing error observed. Figure 5 further demonstrate the correlation between $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$ and the IDE testing results (i.e., the generalization performance of model trained on $\tilde{\mathcal{D}}_t$). The experimental results match the conclusion in Theorem 1.



(a) Reduced ImageNet (b) MNIST

Figure 5: The expectation $\mathbb{E}_{\mathcal{D}}\tilde{\gamma}_t(x,y)$ evaluated on the Reduced ImageNet and MNIST testing set (green curves) and the corresponding IDE results (red curves). In each figure, the green and red curves are tightly correlated. This supports the conclusion in Theorem 1.

10 Other Implications

Our preceding theoretical analysis underscores the critical role played by the local properties of $\mathcal{Q}_{x,y,\theta_t}$ in affecting the generalization performance for $\tilde{\mathcal{D}}_t$. This, in turn, inspires our curiosity to investigate whether additional local properties, beyond local dispersion, also posses critical influences on the generalization of $\tilde{\mathcal{D}}_t$. As such, we inspect the expected distance between the adversarial examples generated by PGD and its clean counterparts, defined as

$$d_\theta(x,y) := \mathbb{E}_{\rho\sim\mathcal{U}([-\epsilon,\epsilon]^d)}\|\mathcal{Q}_{x,y,\theta}(x+\rho)-x\|_2 \tag{51}$$

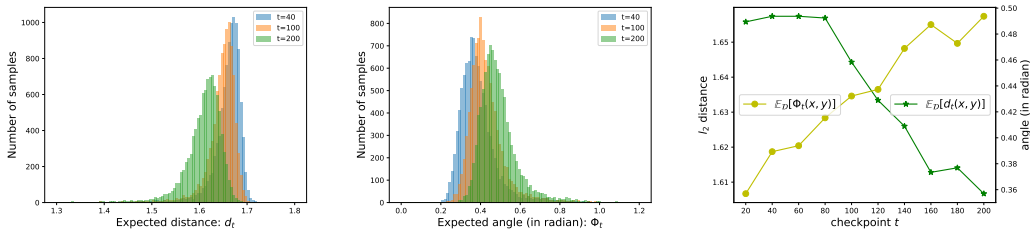
By triangle inequality, we notice that

$$\begin{aligned}
&\mathbb{E}_{\rho,\rho'\sim\mathcal{U}([-\epsilon,+\epsilon]^d)}\|\mathcal{Q}_{x,y,\theta}(x+\rho)-\mathcal{Q}_{x,y,\theta}(x+\rho')\|_2 &(52)\\
=&\mathbb{E}_{\rho,\rho'\sim\mathcal{U}([-\epsilon,+\epsilon]^d)}\|\mathcal{Q}_{x,y,\theta}(x+\rho)-x+x-\mathcal{Q}_{x,y,\theta}(x+\rho')\|_2 &(53)\\
\leq&2d_\theta(x,y) &(54)
\end{aligned}$$

The term (52) is related to the local dispersion of $\mathcal{Q}_{x,y,\theta}$ despite that the definition of the local dispersion computes the square of the l_2-distance. Recall that we have observed an increase in the level of local dispersion along the adversarial training trajectory. According to the inequality (54), one might logically expect that the level of $d_\theta(x,y)$ should also increase, meaning that the perturbed data generated by x are getting not only more "dispersed" around x but also move farther from x. However, our experimental findings present a contradictory result. Instead of an increase, we observe a decrease in the level of $d_\theta(x,y)$ during adversarial training. This unexpected trend suggests that the perturbed data generated by x are, in fact, moving closer to the original data point x.

In our experiments, we estimate $d_\theta(x,y)$ by computing the sample mean with 10 samples of ρ drawn from $\mathcal{U}([-\epsilon,\epsilon]^d)$. We analyze the dynamic behavior of $d_{\theta_t}(x,y)$, which we refer to as $d_t(x,y)$ for simplicity, along the adversarial training trajectory. In Figure 6 (a), we present histogram of $d_t(x,y)$ for the CIFAR-10 testing set at three distinct training checkpoints. Notably, the histogram exhibits a notable mode shift towards a smaller value, indicating a trend that as adversarial training proceeds, the generated adversarial examples progressively approach their clean counterparts. The reduction in the level of $d_t(x,y)$ along adversarial training is further observed by evaluating the expectation $\mathbb{E}_{\mathcal{D}}d_t(x,y)$ on the testing set (see Figure 6 (c), green curve), where a clear drop in $\mathbb{E}_{\mathcal{D}}d_t(x,y)$ is exhibited.

As a reminder, we previously noted that adversarial examples tend to become more dispersed around their clean counterparts as training progresses. The experimental findings presented here shed light on this phenomenon, suggesting that the growing dispersion is likely to be a result of the perturbation angles expanding, while the perturbation magnitudes seem to have a lesser impact on the level of dispersion.



(a) (b) (c)

Figure 6: Experiments on the CIFAR-10 testing set. (a) and (b): histograms of $d_t(x, y)$ and $\Phi_t(x, y)$ at different adversarial training epochs t. (c): The evolution of $\mathbb{E}_{\mathcal{D}} d_t(x, y)$ and $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ along adversarial training trajectory. Combined with the results in Figure 2, an interesting phenomenon in adversarial training is revealed: as the adversarial proceeds, the perturbed data generated by x are getting closer to x and in the meanwhile getting more dispersed potentially due to the spreading of perturbation angles.

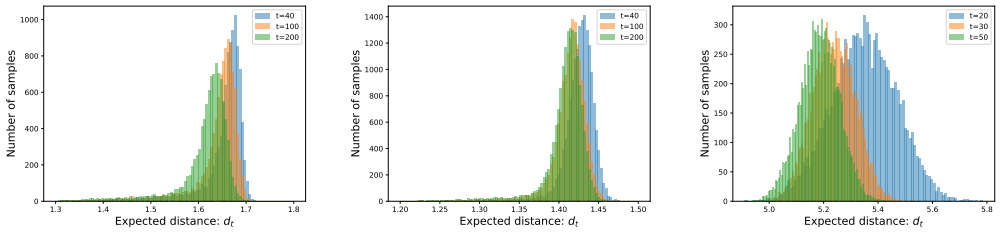
To verify this conjecture, we evaluate the expected angle between a pair of perturbations generated from (x, y), defined as

$$\Phi_\theta(x, y) := \mathbb{E}_{\rho,\rho' \sim \mathcal{U}([-\epsilon,+\epsilon]^d)} \cos^{-1} \left(\frac{(\mathcal{Q}_{x,y,\theta}(x+\rho) - x)^T (\mathcal{Q}_{x,y,\theta}(x+\rho') - x)}{\|\mathcal{Q}_{x,y,\theta}(x+\rho) - x\|_2 \|\mathcal{Q}_{x,y,\theta}(x+\rho') - x\|_2} \right) \tag{55}$$

with the expectation estimated by computing the sample mean of 10 pairs of ρ, ρ' drawn from $\mathcal{U}([-\epsilon, +\epsilon]^d)$. Figure 6 (b) plots the histograms of $\Phi_{\theta_t}(x, y)$ (or $\Phi_t(x, y)$) at three distinct checkpoint t and Figure 6 (c) illustrate the evolution of $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ with the yellow curve. The results present an increase in the level of $\Phi_t(x, y)$, indicating a spreading of perturbation angles and less "aligned" perturbations generated by each x during adversarial training. Similar experimental results have been observed across other datasets. (see Figure 7, 8 and 9).

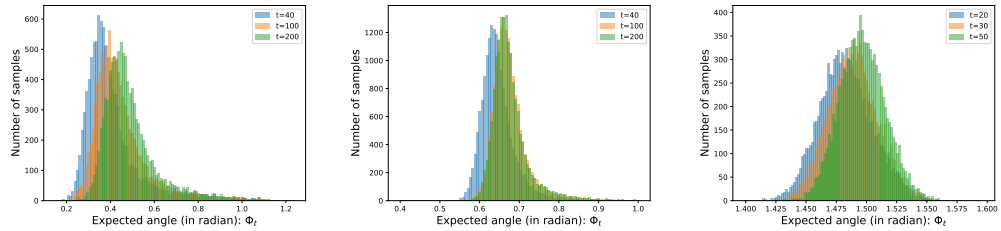
We conjecture that this wider spread of angles in adversarial perturbations is a consequence of an intricate or "ragged" shape in the model's decision boundary. In essence, the shape of the decision boundary has a substantial influence on the direction of perturbations. For instance, the perturbations generated by linear classifiers are always aligned due to that the decision boundary is "smooth". Conversely, one would expect that a jagged or irregular decision boundary could result in perturbations that are both more dispersed and less aligned. We speculate that the presented dynamics of $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ and $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$ can be explained as: during the early stages of adversarial training, the adversarial perturbations generated by the data (x, y) exhibit a higher degree of alignment due to the initial smoothness of the model's decision boundary. This results in a smaller level of $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ and $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$. However, as training progresses, the decision boundary is twisted, in order to fit or "memorize" the training data, causing the perturbations to become less aligned and more dispersed, leading to the rise in the level of $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ and $\mathbb{E}_{\mathcal{D}} \tilde{\gamma}_t(x, y)$. Consequently, the increasing dispersion or spread of angles may serve as an indicative measure for the degree of irregularity present in the decision boundary.

Our observation of the increasing dispersion and spreading angles of adversarial perturbations along adversarial training is, to our best knowledge, a novel finding. This discovery may provide valuable insights into comprehending the dynamic of adversarial training and the phenomenon of robust overfitting.



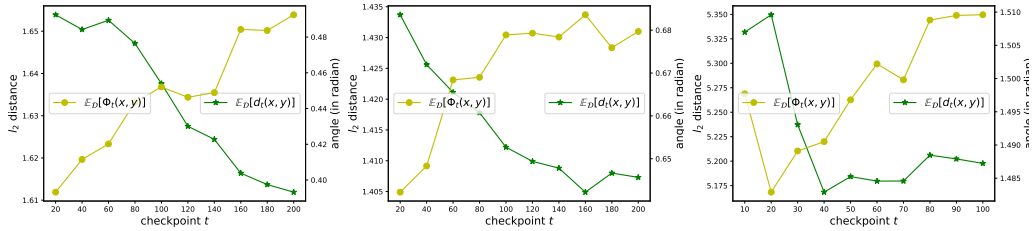
(a) CIFAR-100 (b) Reduced ImageNet (c) MNIST

Figure 7: histograms of d_t on the CIFAR-100, Reduced ImageNet and MNIST testing set. The reduction in the level of d_t along adversarial training is shown in the figures.



(a) CIFAR-100 (b) Reduced ImageNet (c) MNIST

Figure 8: histograms of Φ_t on the CIFAR-100, Reduced ImageNet and MNIST testing set. It shows an increment in the level of Φ_t along adversarial training.



(a) CIFAR-100 (b) Reduced ImageNet (c) MNIST

Figure 9: The evolution of $\mathbb{E}_{\mathcal{D}} d_t(x, y)$ and $\mathbb{E}_{\mathcal{D}} \Phi_t(x, y)$ along adversarial training evaluated on the testing set of CIFAR-100, Reduced ImageNet and MNIST. The behaviours of the two quantities are similar with those on CIFAR-10.