Adversarial Training May Induce Deteriorating Distributions

Runzhi Tian[1] **Yongyi Mao**[1]

[1]Dept of EECS, University of Ottawa, Canada

Abstract

The interactions between the update of model parameters and the update of perturbation operators complicate the dynamics of adversarial training (AT). This paper reveals a surprising behavior in AT, namely that the distribution induced by adversarial perturbations during AT becomes progressively more difficult to learn. We derive a generalization bound to theoretically attribute this behavior to the increasing of a quantity associated with the perturbation operator, namely, its local dispersion. We corroborate this explanation with concrete experimental validations and show that this deteriorating behavior of the induced distributions is correlated with robust overfitting of AT. Code is available at `https://github.com/rzTian/AT-Deteriorating-Distributions`.

1 INTRODUCTION

Despite their outstanding performance, deep neural networks (DNNs) are known to be vulnerable to adversarial attacks, where a carefully designed perturbation of input may cause the network to make a wrong prediction [Szegedy et al., 2014, Goodfellow et al., 2015]. Such perturbed inputs are termed adversarial examples. The existence of adversarial examples raises great concerns when DNNs are applied to decision-critical tasks such as autonomous driving and facial recognition [Eykholt et al., 2018, Sharif et al., 2016]. Many methods have been proposed to improve the robustness of DNNs against adversarial perturbations [Madry et al., 2019, Zhang et al., 2019, Croce et al., 2020], among which the framework, known as adversarial training (AT) [Madry et al., 2019], is arguably the most effective [Athalye et al., 2018, Dong et al., 2020].

In a nutshell, AT may be regarded as stochastic gradient descent (SGD) on an adversarially perturbed version of the training set at each iteration. Specifically, at each gradient descent iteration, each input instance in a training batch is first perturbed to maximize the training loss with respect to the current model parameter, and then gradient descent is performed to update the model parameter. The maximization of the training loss prior to gradient descent is constrained on a maximum allowable perturbation radius; in other words, this maximization is equivalent to an adversarial attack to the model with current parameter setting. The most popular method to solve this maximization problem is the Projected Gradient Descent (PGD) [Madry et al., 2019].

Despite that AT have been shown to have greatly improved the robustness of the learned model against adversarial attacks on the training set, a recent work in Rice et al. [2020] has however revealed that models trained by AT may still be vulnerable to adversarial attacks on the unseen data. Specifically, after training, even though the robust error (i.e., error probability in the predicted label for adversarially perturbed instances) is nearly zero on the training set, it may remain very high on the testing set. For example, on the testing set of CIFAR-10 [Krizhevsky et al., 2009], the robust error of AT trained model can be as large as 44.19%. This significantly contrasts the typical observations in standard training: on CIFAR-10, when the standard error (i.e., the error probability in the predicted label for non-perturbed instances) is nearly zero on the training set, its value on the testing set is only about 4%. This unexpected phenomenon is often referred to as robust overfitting.

Since its discovery, a great deal of research effort has been spent on understanding the cause of robust overfitting. Various perspectives have been exploited in this research direction. For instance, Wu et al. [2020], Stutz et al. [2021], Chen et al. [2021], Kanai et al. [2023] study the properties of the landscape of the adversarial loss; the authors of Singla et al. [2021] investigate the curvature of the activation functions used in the neural networks; Dong et al. [2021] attempt to relate robust overfitting to potential label noises in AT; Xing et al. [2021], Xiao et al. [2022b] look into the training

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:4228–4246.

trajectories of AT through the lens of algorithmic stability.

Despite partial answers provided by these works, the cause of robust overfitting remains largely elusive. Arguably this is due to the significant challenges posed by the complex dynamics of AT. In particular, this complexity arises from the convoluted interaction between the update of model parameter along AT iterations and the update of the adversarial perturbations in the inner maximization step. More concretely, when the model parameter gets updated, the adversarial perturbation is updated to one that attacks the updated model, and the updated adversarial perturbation in turn governs the next update of the model parameter. It is then conceivable that understanding the generalization behavior of AT requires a deep understanding of the interaction between the model updates and perturbation updates, even "untangling" the convoluted interaction along the training trajectory. This philosophy is behind the motivation of this work.

A key observation of this paper is the recognition that in each AT iteration, the perturbation operator effectively induces a new data distribution and that the model update may be viewed as the standard training on data drawn from this induced distribution. Since perturbation in each AT iteration has a small magnitude, the induced distribution is provably close to the original data distribution. However, a surprising finding in this work is that these induced distributions behave distinctively from the original distribution: as AT progresses, they may become increasingly more difficult to learn. The experiments supporting this finding were conducted as follows: for a check point of AT, we extract the perturbation operator and use it to perturb both the training set and test set; we then train a model from scratch on the perturbed training set, using standard training, until the (standard) training loss is effectively zero; we then evaluate the learned model on the perturbed testing set to obtain its classification error. We call such an experiment as an "induced distribution experiment" or IDE. When conducting IDE on datasets such as CIFAR-10, we usually observe large testing errors, particularly when the check point is near the end of AT. In fact, on such datasets, the generalization gap for models learned from the induced distribution appears to progressively increase as AT proceeds.

To understand the deteriorating behavior of the induced distribution along AT, we derive a uniform-convergence upper bound of the generation gap for models learned on the induced distributions. The key quantity in the bound is a term we call "local dispersion" of the perturbation operator. Our bound suggests that only when the perturbation operator has small local dispersion, a good generalization guarantee can be obtained for models learned on the distribution induced by the operator. Through experiments, we show that local dispersion is indeed indicative to the generalization gap of models learned on the induced distribution and can be used to explain the deteriorating behavior of the induced distribution along the AT trajectories, as observed in our IDE experiments.

In summary, in this work we discover an interesting phenomenon in AT, namely, that the induced distributions by the perturbation operator in AT are progressively more difficult to learn. We prove a generalization bound as a theoretical explanation for this phenomenon and corroborate it with experimental validations. Our results shed new lights in understanding the complex AT dynamics and the interaction therein between model updates and perturbation updates. Although there have been previous works examining AT trajectories, very few actually zoom into the properties of the perturbation operator. The only work that we are aware of in this direction is a recent paper of Tian and Mao [2025], where a notion of expansiveness is introduced for the perturbation operator and subsequently used to analyze robust generalization via algorithmic stability. Like that work, this paper highlights the importance of investigations in this angle in paving ways towards understanding robust generalization. This importance is further manifested by our additional experimental observation presented at the end of the paper, where we show that the deteriorating behavior of the induced distributions correlates with robust overfitting.

2 RELATED WORKS

Adversarial examples Existing studies have uncovered intriguing properties of adversarial examples, such as their transferability across different models [Goodfellow et al., 2015, Papernot et al., 2016, Tramèr et al., 2017] and their distinct geometric characteristics compared to clean examples [Ma et al., 2018, Fawzi et al., 2018]. The work in Ilyas et al. [2019] reveals that adversarial examples generated w.r.t a model trained via standard training may still contain useful features. Specifically, they demonstrate that a classifier trained on mislabeled adversarial examples can achieve remarkable generalization performance on unseen clean data. Theoretical explanations for this finding are then provided in Kumano et al. [2024, 2025]. Additionally, the work in Zhang et al. [2022] presents another intriguing finding that adversarial perturbations for two-layer neural networks with random weights are linearly separable, suggesting structural properties of adversarial perturbations exist.

Unlike Ilyas et al. [2019] and Zhang et al. [2022], who focus on adversarial examples for models trained via standard training or with random weights, our work explores adversarial examples along AT trajectories, providing new insights into how features of adversarial examples evolve throughout the training process.

Adversarial Robustness A growing body of work has investigated the underlying causes of adversarial vulnerability, especially in linear and high-dimensional settings. Tanay and Griffin [2016] offered a geometric perspective,

suggesting that adversarial examples arise when the decision boundary extends beyond the data manifold; in such regions, the boundary may lie close to data points, even if it remains distant within the manifold itself. Tanner et al. [2024] analyzed adversarial training for margin-based linear classifiers in high dimensions, highlighting how the interplay between data geometry and attack direction influences robustness. Ribeiro et al. [2023] examined adversarial training in linear regression, showing that it induces different forms of implicit regularization depending on whether the model is overparameterized or underparameterized. Similarly, Javanmard et al. [2020] studied the trade-off between robustness and standard accuracy using linear regression with Gaussian features, providing precise theoretical characterizations in the high-dimensional regime.

Robust generalization Different from standard generalization, robust generalization for deep neural networks —especially on high-dimensional data —appears significantly more challenging. Various work have attempted to understand the reason behind. Schmidt et al. [2018] proves that in simple data models such as the Gaussian and Bernoulli models, robust generalization requires significantly higher sample complexity than standard generalization. The sample complexity of robust generalization has been further analyzed using classical statistical learning tools, including Rademacher complexity [Khim and Loh, 2019, Yin et al., 2018, Awasthi et al., 2020, Xiao et al., 2022a, Attias et al., 2018], VC dimension [Montasser et al., 2019] and algorithmic stability analysis [Xing et al., 2021, Xiao et al., 2022b], as well as the PAC learning frameworks [Cullina et al., 2018, Diochnos et al., 2019].

Beyond sample complexity, several theoretical perspectives have been explored. The work of Li et al. [2022] analyze robust generalization through the lens of neural network's expressive power, showing that practical models may lack sufficient capacity to achieve low robust test error. The authors in Li et al. [2019] investigate inductive bias of gradient descent for AT, while another line of research connects AT with distributionally robust optimization (DRO) [Kuhn et al., 2019, Sinha et al., 2020] . The works of Staib and Jegelka [2017] and Bui et al. [2022] demonstrate that different AT schemes can be reformulated as special cases in DRO. Bennouna et al. [2023] further show that, under a saddle-point assumption, AT inevitably leads to a larger generalization gap than directly solving empirical risk minimization using adversarially perturbed data. Numerous endeavors have been undertaken to address the challenge of robust overfitting with various empirical training algorithms proposed. Bai et al. [2021] and Qian et al. [2022] provide a comprehensive overview of the latest developments in empirical research in this field.

3 PRELIMINARIES AND PROBLEM SETUP

We consider a classification setting with input space $\mathcal{X} \subseteq \mathbb{R}^d$ and label space $\mathcal{Y} := \{1, 2, \cdots, K\}$. We use $\mathcal{D}$ to denote a distribution on $\mathcal{X} \times \mathcal{Y}$ and denote $\mathcal{D}^{\mathcal{X}}$ as the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$. Let Θ be the parameter space of a parameterized model of interest, and for each $\phi \in \Theta$, let $f_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ denote a model which consists of a loss function (e.g, the cross-entropy loss or 0-1 loss) and a classifier h_ϕ with parameter ϕ.

For any data distribution $\mathcal{D}$ and any model f_ϕ, we define the model's *standard population risk* $R_{\mathcal{D}}(\phi)$ as

$$R_{\mathcal{D}}(\phi) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f_\phi(x, y)\right] \tag{1}$$

For a set of m samples $S := \{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$, we define the model's standard empirical risk $R_S(\phi)$ as

$$R_{\mathcal{S}}(\phi) := \frac{1}{m}\sum_{i=1}^{m} f_\phi(x_i, y_i) \tag{2}$$

The standard generalization performance of the model f_ϕ is then measured by the *standard generalization gap*:

$$\mathrm{GG}_m(\phi, S; \mathcal{D}) := |R_{\mathcal{D}}(\phi) - R_S(\phi)| \tag{3}$$

Adversarial perturbations Let $\mathbb{B}_\infty(x, \epsilon)$ denote a ∞-norm ball centered at x with radius ϵ, or $\mathbb{B}_\infty(x, \epsilon) := \{t \in \mathbb{R}^d : \|t - x\|_\infty \le \epsilon\}$. Given any instance-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and a target model f_ϕ parameterized by ϕ, we define the *ϵ-adversarial perturbation of x with respect to f_ϕ* as

$$\mathcal{Q}_\phi(x, y) := \arg\max_{v\in\mathbb{B}_\infty(x,\epsilon)} f_\phi(v, y) \tag{4}$$

Clearly the operator $\mathcal{Q}_\phi$ also depends on the allowable perturbation magnitude ϵ, but we suppress such dependency in our notations throughout the paper for simplicity.

Adversarial risks Given a data distribution $\mathcal{D}$ and its i.i.d samples S, we define the *adversarial population risk* $R^{\mathrm{adv}}_{\mathcal{D}}(\phi)$ and the *adversarial empirical risk* $R^{\mathrm{adv}}_{S}(\phi)$ of a model f_ϕ respectively as

$$R^{\mathrm{adv}}_{\mathcal{D}}(\phi) := \mathbb{E}_{(x,y)\sim\mathcal{D}} f_\phi(\mathcal{Q}_\phi(x, y), y) \tag{5}$$

and

$$R^{\mathrm{adv}}_{S}(\phi) := \frac{1}{m}\sum_{i=1}^{m} f_\phi(\mathcal{Q}_\phi(x_i, y_i), y_i) \tag{6}$$

Adversarial training Given a training set S, at the t^{th} iteration of adversarial training (AT), where the model parameter is ϕ_t, the model parameter is updated, with learning rate η, by

$$\phi_{t+1} = \phi_t - \eta\nabla_{\phi_t}\left[\frac{1}{n}\sum_{i=1}^{n} f_{\phi_t}\left(\mathcal{Q}_{\phi_t}(x_i, y_i), y_i\right)\right] \tag{7}$$

We note that when optimizing $f_\phi(\mathcal{Q}_\phi(x,y),y)$ using gradient descent, despite $\mathcal{Q}$ is also a function of ϕ, the gradient does not propagate through the perturbation operator $\mathcal{Q}$, an option consistent with the standard AT implementation as in Madry et al. [2019], Rice et al. [2020].

Notably, the update equation (7) of AT results in a complex dynamics, namely, the update of ϕ causes the update of the perturbation operator $\mathcal{Q}_\phi$, and the update of $\mathcal{Q}_\phi$ in turn influences the next update of ϕ. This complex interaction between the model parameter and the perturbation operator makes analyzing AT trajectories very difficult.

One key perspective of this work is recognizing that at training iteration t, the perturbation operator $\mathcal{Q}_{\phi_t}$ essentially induces a different distribution and that the AT step in (7) may be seen as a one-step gradient descent on the standard empirical risk of training data drawn from this induced distribution. We next make this precise.

Perturbation induced distribution Let (X,Y) be drawn from $\mathcal{D}$. Given an adversarial perturbation $\mathcal{Q}_\phi$, the *perturbation induced distribution* (or simply induced distribution) is defined as the joint distribution of $(\mathcal{Q}_\phi(X,Y),Y)$ and is denoted by $\tilde{\mathcal{D}}_\phi$. For a given training set $S=\{(x_i,y_i)\}_{i=1}^m$, denote $\tilde{S}_\phi := \{(v_i,y_i)\}_{i=1}^m$, where $v_i := \mathcal{Q}_\phi(x_i,y_i)$. It is clear that the samples $\tilde{S}_\phi$ are drawn from the induced distribution $\tilde{\mathcal{D}}_\phi$.

Since each perturbed instances $\mathcal{Q}_\phi(x,y)$ lies within a small neighborhood of x (i.e., $\|\mathcal{Q}_\phi(x,y)-x\|_\infty \le \epsilon$), it follows immediately that for any ϕ, the Wasserstein p-distance (denoted by $\mathcal{W}_p(\cdot,\cdot)$) between $\mathcal{D}$ and $\tilde{\mathcal{D}}_\phi$ satisfies

$$\mathcal{W}_p(\tilde{\mathcal{D}}_\phi,\mathcal{D}) \le \epsilon \tag{8}$$

for any $p\in[1,+\infty]$. Here the metric, say d, on $\mathcal{X}\times\mathcal{Y}$ by which the Wasserstein distance is defined, is

$$d((x,y),(x',y')) := \|x-x'\|_\infty + d_{\mathcal{Y}}(y,y')$$

where $d_{\mathcal{Y}}$ is an arbitrary metric on $\mathcal{Y}$.

Notably, in the context of adversarial training, the maximum perturbation magnitude ϵ is usually small. Then by equation (8), the distribution $\tilde{\mathcal{D}}_\phi$ induced by the perturbation operator $\mathcal{Q}_\phi$ during AT is very close to the original data distribution $\mathcal{D}$. However, a surprising observation in this work is that models trained (via standard training) on $\mathcal{D}$ and on $\tilde{\mathcal{D}}_\phi$ may have very different behaviors.

It is also worth noting that $R^{\mathrm{adv}}_{\mathcal{D}}(\phi) = R_{\tilde{\mathcal{D}}_\phi}(\phi)$ and $R^{\mathrm{adv}}_{S}(\phi) = R_{\tilde{S}_\phi}(\phi)$ —the adversarial risks of ϕ can be treated as the standard population (resp. empirical) risk of ϕ measured on the induced distribution (resp. the samples drawn from the induced distribution) generated by ϕ.

Following the definition of generalization gap in (3), the notations $\mathrm{GG}_m(\phi,\tilde{S}_\phi;\tilde{\mathcal{D}}_\phi)$ and $\mathrm{GG}_m(\theta,\tilde{S}_\phi;\tilde{\mathcal{D}}_\phi)$ are both well defined, where the former is the *robust generalization gap* of an arbitrary model f_ϕ and the latter is the standard generalization gap of an arbitrary model f_θ measured with respect to a given induced distribution $\tilde{\mathcal{D}}_\phi$ and its samples $\tilde{S}_\phi$.

4 LEARNING ON THE INDUCED DISTRIBUTIONS

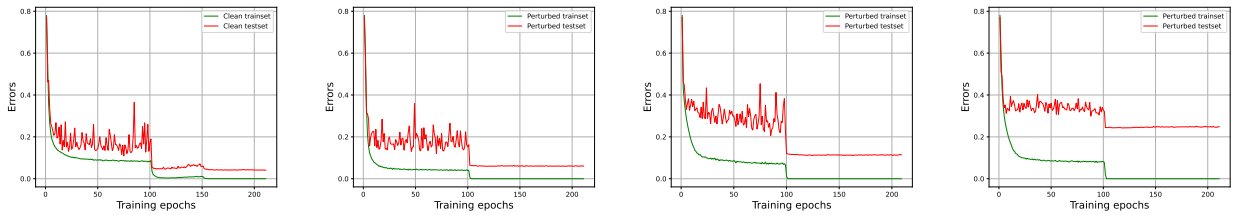
In this section, we experimentally study the problem of learning on the induced distribution $\tilde{\mathcal{D}}_\phi$, where ϕ is the parameter of a model being trained during AT.

Induced distribution experiment Let S and T be the training set and testing set of a classification task. We perform AT for a neural network model using S. Let $\mathrm{AT}(t)$ denote that model's parameter obtained by performing AT for t epochs. For some choice of t, we obtain model parameter $\phi=\mathrm{AT}(t)$. We then perturb S and T using $\mathcal{Q}_\phi$, and obtain the perturbed training and testing datasets $\tilde{S}_\phi$ and $\tilde{T}_\phi$ respectively. A new model (with the same architecture) is then trained from scratch (namely, starting from random initialization of its parameters) on $\tilde{S}_\phi$ **using standard training** and denote the learned model parameter by θ. This model θ is evaluated on $\tilde{T}_\phi$. For the ease of reference we call such an experiment the "induced distribution experiment" (IDE).

In our IDE experiments, $\mathcal{Q}_\phi$ is taken as the Projected Gradient Descend (PGD) attack [Madry et al., 2019], which is used both for AT and for generating the perturbed datasets. Other details of the experiments are given below.

Datasets The experiments are conducted on CIFAR10 and CIFAR100 [Krizhevsky et al., 2009]. We also conduct experiments on a "scaled-down" version of the ImageNet dataset [Russakovsky et al., 2015], which we call Reduced ImageNet, drawing inspiration from a similar approach in Tsipras et al. [2019] for reduced training complexity. Reduced ImageNet aggregates several subsets of the original ImageNet and comprises 10 classes, each containing 5000 training samples and approximately 1000 testing samples per class. More details concerning this dataset are given in Appendix A.

Settings for AT and PGD On CIFAR-10 and Reduced ImageNet we perform AT to train the pre-activation ResNet (PRN) model [He et al., 2016] with 18 and 50 layers respectively. On CIFAR-100 we train the Wide ResNet (WRN) model with 34 layers [Zagoruyko and Komodakis, 2016]. We use 5-step PGD with $\epsilon=4/255$ for Reduced ImageNet and 10-step PGD with $\epsilon=8/255$ for CIFAR-10 and CIFAR-100 according to Rice et al. [2020]. We set $\lambda=2/255$ on CIFAR10 and CIFAR100, $\lambda=0.9/255$ on Reduced ImageNet. More details concerning the hyper-parameter settings are given in Appendix A.



(a) Clean dataset (b) ϕ=AT(0) (c) ϕ=AT(80) (d) ϕ=AT(200)

Figure 1: Learning curves of standard training on the clean CIFAR-10 dataset and IDEs w.r.t various ϕ. In each training, the learning rate is decayed at the 100^{th} epoch.

Experimental results Let $\phi = \text{AT}(0)$ denote a randomly initialized model. Figure 1(b)-(d) presents the learning curves of IDEs conducted on the CIFAR-10 datasets for ϕ obtained after AT for different numbers of epochs, while Figure 1(a) shows the learning curves of standard training on the clean CIFAR-10 dataset for comparison. The green and red curves respectively represent the training and testing error recorded along the training process. In all cases, the model is trained to achieve zero training error. However, the testing error varies significantly in different IDEs. On the clean dataset, the model attains a testing error as low as 4.13%; A similar performance is observed on the IDE with $\phi = \text{AT}(0)$, where the testing error reaches around 6.06%. In contrast, for $\phi = \text{AT}(80)$, the learned model shows a reduced generalization performance, with the testing error increasing to 11.38%. A more significant rise on the testing error occurs when a model is trained on the perturbed dataset generated by $\phi = \text{AT}(200)$, where the testing error increases to 24.89%. Similar results are also observed on CIFAR-100 and Reduced ImageNet (see Appendix C Figure 5 and 6).

For IDE with $\phi = \text{AT}(200)$, a large generalization gap —the gap between the red and green curves —emerges in the early phase of the training (around the 20^{th} training epoch). After the drop of learning rate (at the 100^{th} training epoch), the training error quickly reduces to zero, yet the generalization gap remains nearly unchanged, resulting in a high final testing error. This is in contrast to the learning behavior observed on the clean dataset and the IDE with $\phi = \text{AT}(0)$, where a small generalization gap is established at the early phase of training and is consistently preserved along the training.

These experiments reveal a rather surprising phenomenon: despite $\tilde{\mathcal{D}}_\phi$ being very close to $\mathcal{D}$, the model's learning performance on the induced distribution $\tilde{\mathcal{D}}_\phi$ can be significantly different from that on $\mathcal{D}$. In particular, as AT proceeds, the induced distribution $\tilde{\mathcal{D}}_\phi$ may deteriorate, in the sense that it becomes increasingly more difficult to generalize, as signified by the increasing generalization gap.

5 THEORETICAL ANALYSIS

In this section, we provide a theoretical analysis to explain the deteriorating learning behavior of the induced distribution along AT. Specifically, we derive an upper bound for the "worst-case" generalization gap $\sup_{\theta\in\Theta} \text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$.

Assumption 5.1 (Anchored data model)**.** We assume that underlying the data distribution $\mathcal{D}$, there is a latent distribution, or "anchor distribution", $\mathcal{D}_*$ on $\mathcal{X}\times\mathcal{Y}$. $\mathcal{D}_*$ is specified by its marginal $\mathcal{D}_*^{\mathcal{X}}$ on $\mathcal{X}$ and a classifier $h^* : \mathcal{X} \to \mathcal{Y}$ (which assigns every sample drawn from $\mathcal{D}_*^{\mathcal{X}}$ a label in $\mathcal{Y}$). The data distribution $\mathcal{D}$ of interest is a "smoothed" version of $\mathcal{D}_*$ as follows: Draw an "anchor variable" T from $\mathcal{D}_*^{\mathcal{X}}$. Then draw a noise ρ independent of T from a distribution π (on $\mathbb{R}^d$) with zero mean and a finite variance in each dimension (recall that $\mathcal{X}$ is a subset of $\mathbb{R}^d$) —we assume the variance in each dimension is small. The distribution of $(T+\rho, h^*(T))$ is then the distribution $\mathcal{D}$.

Remark 5.2. In this anchored data model, the true input variable X is treated as a noise-perturbed version of an anchor variable $T \sim \mathcal{D}_*^{\mathcal{X}}$. Such an assumption is widely used in various machine learning contexts, for example, in the VAE model [Kingma and Welling, 2019] where the reconstruction loss adopts the square error loss .

On the other hand, the assumption that $X = T + \rho$ share the same label as T is sensible, since one expects that within small neighborhood of T, the class label remains unchanged.

Given a model class $\mathcal{F} := \{f_\theta : \theta \in \Theta\}$, we now study its generalization performance w.r.t the induced distributions. Specifically, we will derive an upper bound for the generalization gap $\text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$ for all $\theta \in \Theta$. As it turns out, a key quantity governing the upper bound is a local property of the perturbation map $\mathcal{Q}_\phi$ that induces $\tilde{\mathcal{D}}_\phi$.

Definition 5.3 (Local dispersion)**.** For any $(x, y) \in \mathcal{X}\times\mathcal{Y}$, we define the local dispersion $\tilde{\gamma}_\phi(x, y)$ of the perturbation mapping $\mathcal{Q}_\phi$ at (x, y) as

$$\tilde{\gamma}_\phi(x, y) := \mathbb{E}_{\rho,\rho'} \|\mathcal{Q}_\phi(x+\rho, y) - \mathcal{Q}_\phi(x+\rho', y)\|_2^2. \quad (9)$$

where ρ and ρ' are drawn independently from π.

Remark 5.4. We refer to this quantity as the *local dispersion* of $\mathcal{Q}_\phi$, as it measures how far apart the operator $\mathcal{Q}_\phi$ disperses two noise-perturbed versions of (x, y). In fact, one may verify that $\tilde{\gamma}_\phi(x, y)$ can be expressed as

$$\tilde{\gamma}_\phi(x, y) = 2 \cdot \text{Trace}\left(\text{COV}_\rho(\mathcal{Q}_\phi(x + \rho, y))\right) \tag{10}$$

where ρ is drawn from π and $\text{COV}_\rho(\mathcal{Q}_\phi(x+\rho, y))$ denotes the covariance matrix. That is, $\tilde{\gamma}_\phi(x, y)$ also measures the how far $\mathcal{Q}_\phi$ spreads a randomly perturbed version of (x, y). We defer the proof of (10) to Appendix B.1.

One may argue intuitively that smaller local dispersion of $\mathcal{Q}_\phi$ may allow the model to generalize better when learning on the distribution $\tilde{\mathcal{D}}_\phi$: consider an instance (T, Y) drawn from the anchor distribution $\mathcal{D}_*$, and two observed data points $(T + \rho, Y)$ and $(T + \rho', Y)$ (with ρ and ρ' drawn independently from π). Suppose that $(T+\rho, Y)$ is included in the training set and $(T+\rho', Y)$ is included in the testing set. When the local dispersion is small, the perturbed version of the training point $(\mathcal{Q}_\phi(T + \rho, Y), Y)$ and that of the testing point $(\mathcal{Q}_\phi(T + \rho', Y), Y)$ (both of which are realizations from $\tilde{\mathcal{D}}_\phi$) are close, allowing the model's prediction on the latter to behave similarly as that on the former.

We now rigorously formalize this intuition, under the following assumptions.

- (Lipchitzness of f_θ over $\mathcal{X}$) For any $y \in \mathcal{Y}$ and any $\theta \in \Theta$, $|f_\theta(x, y) - f_\theta(x', y)| \leq \beta \|x - x'\|_2$ for $\forall x, x' \in \mathcal{X}$.
- (Boundedness) $\sup_{x,y \in \mathcal{X} \times \mathcal{Y}} |f_\theta(x, y)| = B < \infty$ for any $\theta \in \Theta$.

The generalization gap (3) then has the following uniform convergence result:

Lemma 5.5. *Consider the model class $\mathcal{F}$ where each $f_\theta \in \mathcal{F}$ satisfies the above boundedness condition. For any ϕ (or $\tilde{\mathcal{D}}_\phi$), with probability $1 - \tau$ over drawing $\tilde{S}_\phi$ from $\tilde{\mathcal{D}}_\phi$, we have*

$$\sup_{\theta \in \Theta} \text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$$
$$\leq \mathbb{E}_{\tilde{S}_\phi \sim \tilde{\mathcal{D}}_\phi^m} \sup_{\theta \in \Theta} \text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi) + 2B\sqrt{\frac{\log \frac{1}{\tau}}{2m}} \tag{11}$$

The proof of the lemma is deferred to Appendix B.2. Building upon lemma 5.5, we now derive an upper bound for $\mathbb{E}_{\tilde{S}_\phi \sim \tilde{\mathcal{D}}_\phi^m} \sup_{\theta \in \Theta} \text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$ where the local dispersion of $\mathcal{Q}_\phi$ plays a role.

Theorem 5.6. *Consider the model class $\mathcal{F}$ where each $f_\theta \in \mathcal{F}$ satisfies the above Lipchitzness and boundedness conditions. Consider the data distribution $\mathcal{D}$ which satisfies the assumptions 5.1. Let $\tilde{\mathcal{D}}_\phi$ denote the induced distribution of $\mathcal{D}$, generated by a perturbation $\mathcal{Q}_\phi$. We have*

$$\mathbb{E}_{\tilde{S}_\phi \sim \tilde{\mathcal{D}}_\phi^m} \sup_{\theta \in \Theta} \text{GG}_m(\theta, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$$
$$\leq \frac{2\beta}{\sqrt{m}} \sqrt{\mathbb{E}_{(x,y) \sim \mathcal{D}_*} \tilde{\gamma}_\phi(x, y)} + \frac{2(\beta\sqrt{d}\epsilon + B)}{\sqrt{m}} \tag{12}$$

We leave the proof of the Theorem in Appendix B.3. Combining (12) with (11) immediately gives an upper bound for the generalization gap (3) that applies for any $\theta \in \Theta$.

Remark 5.7. The derivation of Theorem 5.6 is based on a modification of the Rademacher complexity analysis. It worth noting that any direct application of Rademacher complexity to establish a learning bound requires certain restriction on the hypothesis class $\mathcal{F}$, thus suffering from a loss of generality.

The theorem suggests that the generalization gap of any f_θ w.r.t to the distribution $\tilde{\mathcal{D}}_\phi$ is affected by the expected local dispersion (ELD) $\mathbb{E}_{\mathcal{D}_*} \tilde{\gamma}_\phi(x, y)$ of $\mathcal{Q}_\phi$ and that a small generalization gap can be uniformly attained—for every $f_\theta \in \mathcal{F}$ —with high probability when ELD $\mathbb{E}_{\mathcal{D}_*} \tilde{\gamma}_\phi(x, y)$ is small.

An interpretation of this theorem is that the learning difficulty of the induced distribution $\tilde{\mathcal{D}}_\phi$ may be attributed to the ELD $\mathbb{E}_{\mathcal{D}_*} \tilde{\gamma}_\phi(x, y)$ of the perturbation operator $\mathcal{Q}_\phi$. But since the theorem only provides an upper bound, such an interpretation is only valid if the upper bound in the theorem is indicative of the true generalization gap. We next report experimental measurements to show this is indeed the case.

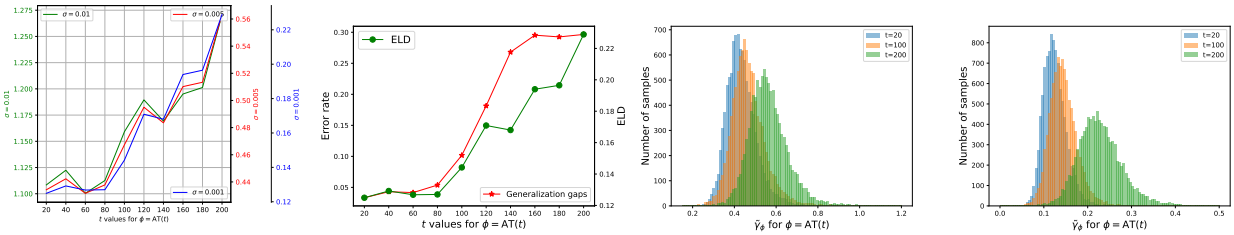
6 EXPERIMENTAL VALIDATION

We conducted experiments to estimate the ELD of $\mathcal{Q}_\phi$ for $\phi = \text{AT}(t)$ with various t values along the AT trajectory. Note that the expectation here is over the distribution $\mathcal{D}_*$, from which no samples are available. However, due to the relationship between $\mathcal{D}^\mathcal{X}$ and $\mathcal{D}_*^\mathcal{X}$, namely that $\mathcal{D}^\mathcal{X}$ is merely a slightly smoothed version of $\mathcal{D}_*^\mathcal{X}$ (since π has small variances), one expects that when we draw x from $\mathcal{D}^\mathcal{X}$, $\mathcal{D}^\mathcal{X}(x) \approx \mathcal{D}_*^\mathcal{X}(x)$ with high probability. As a consequence, $\mathbb{E}_{\mathcal{D}_*} \tilde{\gamma}_\phi(x, y) \approx \mathbb{E}_\mathcal{D} \tilde{\gamma}_\phi(x, y)$ with high probability. But the latter can be estimated using the i.i.d. samples from $\mathcal{D}$. This gives us the following estimation formula for ELD:

$$\mathbb{E}_{\mathcal{D}^*} \tilde{\gamma}_\phi(x, y) \approx \frac{1}{m} \sum_{i=1}^{m} \tilde{\gamma}_\phi(x_i, y_i),$$

where $\{(x_i, y_i)\}_{i=1}^m$ are drawn from $\mathcal{D}$.

Estimating the local dispersion $\tilde{\gamma}_\phi(x_i, y_i)$ requires the knowledge of π, which is unfortunately unavailable to us. In our experiments, we take π as a spherical Gaussian, with variance in each dimension equal to σ^2. Various values of



(a) ELDs for different σ (b) $\sigma = 0.001$ (c) $\sigma = 0.005$ (d) $\sigma = 0.001$

Figure 2: Local dispersion measured on the CIFAR-10 test set. (a) ELDs estimated using different σ values. For different choice of σ, the estimated ELDs fall within different ranges. To clearly compare the trends of ELD across different σ, we plot all estimations in the same graph and position their respective vertical axes on the sides of the figure. (b) ELD (green curve) of $\mathcal{Q}_\phi$ for different ϕ in comparison to the generalization gap achieved on $\tilde{\mathcal{D}}_\phi$. (c) and (d): histograms of $\tilde{\gamma}_\phi(x, y)$ for three distinct ϕ.

σ^2 are considered in our experiments. The estimation of each $\tilde{\gamma}_\phi(x_i, y_i)$ is done by Monte-Carlo approximation via sampling 250 pairs of (ρ, ρ') from π. The expectation in (9) is then approximated using the sample mean.

Same trend of ELD estimated from different σ Figure 2(a) show that the estimated ELD values with $\phi = \mathrm{AT}(t)$ using $\sigma = 0.001, 0.005, 0.01$ respectively. In the figure, the three curves, each corresponding to a different σ value, have very similar trend. In fact, when adjusting the range of vertical axes, the three curves closely align with each other.

ELD as an indicator of generalization gap Figure 2 (b) presents the generalization gaps of the models learned on various $\tilde{\mathcal{D}}_\phi$ (red curve) and the estimated ELD values of the corresponding $\mathcal{Q}_\phi$ (green curve). In the experiments, we set $\sigma = 0.01$ for ELD estimation. In each IDE, the model is trained to achieve zero training error, hence the generalization gaps in Figure 2 (b) correspond directly to the testing errors of the learned models. As shown in the figure, when the ELD of $\mathcal{Q}_\phi$ is small, the model learned on the corresponding $\tilde{\mathcal{D}}_\phi$ tends to achieve a smaller generalization gap. This empirical observation aligns with the theoretical findings in Theorem 5.6. The positive correlation between the red and green curve in 2(b) suggests that the local dispersion of the perturbation operator significantly affects the generalization performance of the models learned on the induced distribution. This also validates the usefulness of Theorem 5.6, corroborating ELD as an indicator of the generalization gap for the induced distributions.

Increasing dispersiveness along AT Since in our experiments ϕ is obtained at different AT epochs, the upward trend in the green curve of Figure 2(b) and that of all the three curves in 2(a) suggest that performing AT for more iterations tends to make the perturbation operator $\mathcal{Q}_\phi$ increasingly dispersive. To further illustrate this trend, Figure 2 (c) and (d) respectively plot the histograms of $\tilde{\gamma}_\phi(x, y)$ for $\phi = \mathrm{AT}(20), \mathrm{AT}(100), \mathrm{AT}(200)$, estimated using different σ values. As shown on both figures, the histograms shift progressively to the right as AT is performed for more iterations, indicating that the perturbation operator $\mathcal{Q}_\phi$ becomes more locally dispersive as ϕ evolves in AT. Similar experimental results are also observed on CIFAR-100 and Reduced ImageNet (see Appendix C Figure 7 and 8).

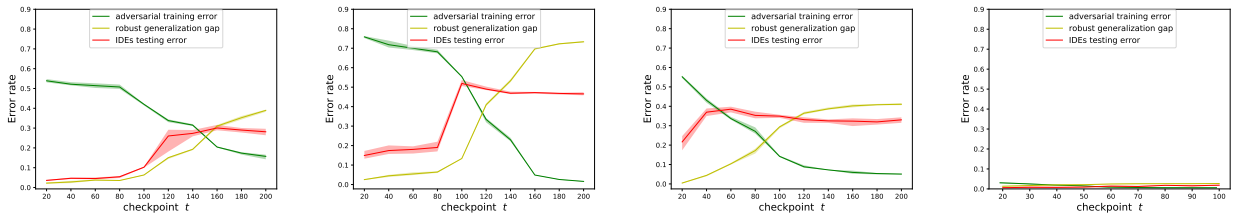
Summary From Theorem 5.6 and these experiments, one may conclude that the deteriorating learning performance on the induced distribution along the AT trajectory can be attributed to the progressive increase of local dispersions of the perturbation operators. It remains unclear what causes perturbation operators in AT to become increasingly dispersive. Nonetheless, this study may shed new lights in understanding the complex dynamics of AT. In particular, we show next that the induced distribution deteriorating along the AT trajectory is correlated with robust overfitting.

7 CORRELATION WITH ROBUST GENERALIZATION

We now explore if the (standard) generalization performance of models learned on the induced distribution $\tilde{\mathcal{D}}_\phi$ along the AT trajectory has any connection to the *robust generalization* performance of ϕ on original data distribution $\mathcal{D}$.

We conduct extra IDEs for ϕ collected along AT at various epochs and compare the IDE testing errors with the robust generalization performance of the corresponding ϕ. AT and each IDE are repeated five times with different random seeds.

The experimental results on CIFAR-10 and CIFAR-100 are shown in Figure 3(a) and (b), where the green and yellow curves respectively report the adversarial training error and the robust generalization gap of ϕ (i.e., $R_S^{\mathrm{adv}}(\phi)$ and $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$). The two curves illustrate a phenomenon known as robust overfitting [Rice et al., 2020]: after a certain point in AT, the robust generalization gap steadily increases



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

Figure 3: Robust generalization gap of $\phi = \mathrm{AT}(t)$ in comparison to the IDE test error w.r.t ϕ. Note that since models in each IDE are trained to achieve zero training error, the IDE test error effectively represents the standard generalization gap achieved on the induced distribution. The trend of the red curves matches that of the yellow curves in each sub-figures, demonstrating a compelling correlation between these two quantities.

while the adversarial training error constantly decreases. The red curves in the figures depict the standard testing errors achieved in each IDEs (i.e., $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ with θ learned on $\tilde{S}_\phi$). Notably, a significant rise in the IDE testing error is observed when ϕ is taken between $\mathrm{AT}(80)$ and $\mathrm{AT}(120)$, increasing from 3.6% to 27.68% for CIFAR-10 and from 19% to 48.99% for CIFAR-100. Furthermore, this shift coincides with the onset of robust overfitting, where a significant rise in $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$) is also observed.

These results further demonstrate that $\tilde{\mathcal{D}}_\phi$ becomes harder to learn as AT progresses. More importantly, it shows that the appearance of this deteriorating induced distribution is closely linked to the onset of the robust overfitting phenomenon, revealing a correlation between the two. This correlation is further demonstrated by experimental results on Reduced ImageNet (see Figure 3 (c)), where robust overfitting emerges at an earlier training stage and simultaneously a rise in $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ occurs. This increment in $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ is also substantial, with an averaged error of 21.65% at $\mathrm{AT}(20)$ elevating to 38.52% at $\mathrm{AT}(60)$.

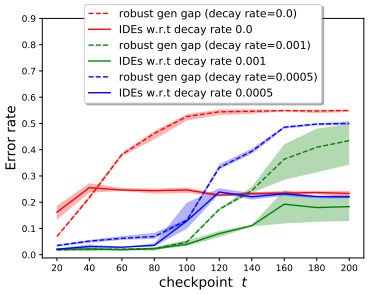


Figure 4: AT with various weight decay rates and the test error achieved in IDEs for each of the AT variants. The blue curves are reproduced from Figure 3(a), serving as a reference for a clear comparison. The results further solidify the correlation between the robust generalization and the generalization performance on the induced distribution.

Our experiments on MNIST [LeCun et al., 1998] (see Figure 3 (d)) exhibits a scenario where a good robust generalization is achieved. [1] Interestingly, a small testing error $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ is maintained throughout the evolution of $\tilde{\mathcal{D}}_\phi$ with the absence of robust overfitting. Figure 4 shows results from additional experiments on CIFAR-10. In these experiments, we perform AT with different levels of weight decay to control the robust generalization gap. Subsequently, IDEs are conducted for each such variant of AT. In Figure 4, each distinct color corresponds to a different weight decay factor utilized in AT. Within each color category, the dashed curves and the corresponding solid lines represent, respectively, $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$ and $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ with ϕ trained by that specific AT variant. From these results, we see that increasing the weight decay factor results in a notable reduction in the $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$, 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. More noteworthy is a clear synchronization observed between each pair of dashed and solid curves (of the same color), with lower dashed curves consistently corresponding to lower solid curves in the same color category.

All these results suggest a strong correlation between $R_{\tilde{\mathcal{D}}_\phi}(\theta)$ and the robust generalization gap $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi)$. Although by construction, the robust generalization gap is written by $\mathrm{GG}_m(\phi, \tilde{S}_\phi; \tilde{\mathcal{D}}_\phi) = |R_{\tilde{\mathcal{D}}_\phi}(\phi) - R_{\tilde{S}_\phi}(\phi)|$ due to that $R^{\mathrm{adv}}_{\mathcal{D}}(\phi) = R_{\tilde{\mathcal{D}}_\phi}(\phi)$ and $R^{\mathrm{adv}}_{S}(\phi) = R_{\tilde{S}_\phi}(\phi)$, such a correlation is still quite surprising. This is because the learning of the parameter θ has been started from a completely random initialization and one would not expect the resulting parameter θ is linked to the parameter ϕ in any obvious way, despite that the latter contributes to shaping the distribution $\tilde{\mathcal{D}}_\phi$.

A novel observation in this work, this correlation is certainly curious in its own right and deserves further investigation. At this point, it has at least highlighted the impact of the dynamics of AT on robust overfitting, beyond the static

[1] Experimental settings on MNIST are shown in Appendix A

quantities, such as loss landscape, while also paving a way for developing deeper understanding of how AT results in robust overfitting.

8 CONCLUSION

In this paper, we show that the distribution induced by the perturbation operator in AT may deteriorate along the trajectory of AT. In particular, we observe experimentally that as AT progresses, the induced distribution may become harder to learn. Our theoretical analysis suggests that a key factor governing this increasing difficulty of learning is the local dispersion of the perturbation operator that induces the distribution. Experimental results confirm that as AT proceeds, the perturbation becomes more dispersive, validating our theoretical results. Additionally, we empirically observed a correlation between the deteriorating behavior of the induced distributions with robust overfitting.

The novel observations and our theoretical explanation presented in this paper contribute to better understanding the complex dynamics of AT. Unraveling this complexity is arguably essential to understanding robust generalization in AT.

Limitations & Future Works While this paper establishes a connection between local dispersion and the learning difficulty of the induced data distribution, the theoretical framework does not fully explain the underlying causes of increased local dispersion during AT. Nor does it provide improved AT algorithms based on the theoretical insights.

Understanding the mechanism that increases local dispersions during AT remains an open and intriguing direction. Any progress in this direction is likely to improve the practical design of AT algorithms. Not having a concrete answer at present, we speculate that this might be related to the increased complexity of classifier decision boundaries during AT: when the boundaries become more complex, the perturbations pointing to the boundaries are more scattered, thereby increasing the local dispersion. Formalizing this intuition with rigorous analysis is a promising avenue for future research.

Another promising direction is to explore ways to mitigate the deterioration of induced distributions, such as through regularization of perturbation operator to control local dispersion.

References

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.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. *CoRR*, abs/1810.02180, 2018. URL http://arxiv.org/abs/1810.02180.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. *CoRR*, abs/2004.13617, 2020. URL https://arxiv.org/abs/2004.13617.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness, 2021.

Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. *arXiv preprint arXiv:2303.02251*, 2023.

Tuan Anh Bui, Trung Le, Quan Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. *arXiv preprint arXiv:2202.13437*, 2022.

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.

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.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries, 2018.

Dimitrios I. Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust PAC learning. *CoRR*, abs/1906.05815, 2019. URL http://arxiv.org/abs/1906.05815.

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.

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.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models, 2018. URL https://arxiv.org/abs/1707.08945.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier, 2018. URL https://arxiv.org/abs/1802.08686.

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

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.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. URL https://arxiv.org/abs/1905.02175.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4732–4741. PMLR, 2020.

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

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation, 2019.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

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

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Theoretical understanding of learning from adversarial perturbations, 2024. URL https://arxiv.org/abs/2402.10470.

Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Wide two-layer networks can learn from adversarial perturbations, 2025. URL https://arxiv.org/abs/2410.23677.

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.

Binghui Li, Jikai Jin, Han Zhong, John E. Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power, 2022.

Yan Li, Ethan X. Fang, Huan Xu, and Tuo Zhao. Inductive bias of gradient descent based adversarial training on separable data. *CoRR*, abs/1906.02931, 2019. URL http://arxiv.org/abs/1906.02931.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality, 2018. URL https://arxiv.org/abs/1801.02613.

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

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. *CoRR*, abs/1902.04217, 2019. URL http://arxiv.org/abs/1902.04217.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016. URL https://arxiv.org/abs/1605.07277.

Zhuang Qian, Kaizhu Huang, Qiu-Feng Wang, and Xu-Yao Zhang. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies, 2022.

Antônio H. Ribeiro, Dave Zachariah, Francis Bach, and Thomas B. Schön. Regularization properties of adversarially-trained linear regression. *arXiv preprint arXiv:2310.10807*, 2023.

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.

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.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *CoRR*, abs/1804.11285, 2018. URL http://arxiv.org/abs/1804.11285.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016.

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.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020.

Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017.

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.

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

Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. A high dimensional statistical model for adversarial training: Geometry and trade-offs. *arXiv preprint arXiv:2402.05674*, 2024.

Runzhi Tian and Yongyi Mao. Algorithmic stability based generalization bounds for adversarial training. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=2GwMazl9ND.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples, 2017. URL https://arxiv.org/abs/1704.03453.

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

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.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, and Zhi-Quan Luo. Adversarial rademacher complexity of deep neural networks, 2022a.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training, 2022b.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=xz80iPFIjvG.

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.

Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. *CoRR*, abs/1810.11914, 2018. URL http://arxiv.org/abs/1810.11914.

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

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.

Huishuai Zhang, Da Yu, Yiping Lu, and Di He. Adversarial noises are linearly separable for (nearly) random neural networks, 2022. URL https://arxiv.org/abs/2206.04316.

Appendices

Runzhi Tian[1] **Yongyi Mao**[1]

[1]Dept of EECS, University of Ottawa, Canada

A DETAILED EXPERIMENTAL SETUP

Our Reduced ImageNet is made 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 Russakovsky et al. [2015], 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.

For adversarial training (AT), the settings on different datasets are summarized in Table 2. Data augmentation is performed on these datasets except for MNIST during the training. For CIFAR-10 and CIFAR-100 we follow the data augmentation setting in Rice et al. [2020]. For our reduced ImageNet, we adopt the same data augmentation scheme that is used on the restricted ImageNet in Yang et al. [2020].

For the induced distribution experiments (IDEs) on each datasets, the settings are outlined in Table 3. It is important to note that for each of the individual IDEs that is 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	PRN18	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
PGD step size	0.01	2/255	2/255	0.9/255
number of PGD	40	10	10	5

Table 2: Settings in PGD and AT across different datasets

	MNIST	CIFAR-10	CIFAR-100	Reduced ImageNet
model	small CNN	PRN-18	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

B PROOFS

B.1 PROOF OF (10)

We have that

$$\begin{aligned}
\tilde{\gamma}_\phi(x,y) &:= \mathbb{E}_{\rho,\rho'}\|\mathcal{Q}_\phi(x+\rho,y)-\mathcal{Q}_\phi(x+\rho',y)\|_2^2 \\
&= \mathbb{E}_{\rho,\rho'}\left(\mathcal{Q}_\phi(x+\rho,y)-\mathcal{Q}_\phi(x+\rho',y)\right)^T\left(\mathcal{Q}_\phi(x+\rho,y)-\mathcal{Q}_\phi(x+\rho',y)\right) && (13)\\
&= \mathbb{E}_{\rho,\rho'}\left[\|\mathcal{Q}_\phi(x+\rho,y)\|_2^2+\|\mathcal{Q}_\phi(x+\rho',y)\|_2^2-2\mathcal{Q}_\phi(x+\rho',y)^T\mathcal{Q}_\phi(x+\rho,y)\right] && (14)\\
&= 2\mathbb{E}_\rho\|\mathcal{Q}_\phi(x+\rho,y)\|_2^2-2\|\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)\|_2^2 && (15)
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\tilde{\gamma}_\phi(x,y) &:= \mathbb{E}_{\rho,\rho'}\|\mathcal{Q}_\phi(x+\rho,y)-\mathcal{Q}_\phi(x+\rho',y)\|_2^2 \\
&= \mathbb{E}_{\rho,\rho'}\|\mathcal{Q}_\phi(x+\rho,y)-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)-\left(\mathcal{Q}_\phi(x+\rho',y)-\mathbb{E}_{\rho'}\mathcal{Q}_\phi(x+\rho',y)\right)\|_2^2 && (16)\\
&= 2\mathbb{E}_\rho\|\mathcal{Q}_\phi(x+\rho,y)-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)\|_2^2-2\|\mathbb{E}_\rho\left[\mathcal{Q}_\phi(x+\rho,y)-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)\right]\|_2^2 && (17)\\
&= 2\mathbb{E}_\rho\|\mathcal{Q}_\phi(x+\rho,y)-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)\|_2^2 && (18)\\
&= 2\mathbb{E}_\rho\left[\sum_{i=1}^d\left(\mathcal{Q}_\phi(x+\rho,y)[i]-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)[i]\right)^2\right] && (19)\\
&= 2\sum_{i=1}^d\mathbb{E}_\rho\left(\mathcal{Q}_\phi(x+\rho,y)[i]-\mathbb{E}_\rho\mathcal{Q}_\phi(x+\rho,y)[i]\right)^2 && (20)\\
&= 2\mathrm{Trace}\left(\mathrm{COV}_\rho(\mathcal{Q}_\phi(x+\rho,y))\right) && (21)
\end{aligned}$$

where equality (17) is derived by applying the results of (15). We use $\mathcal{Q}_\phi(x+\rho,y)[i]$ to denote the i^{th} coordinate of the vector $\mathcal{Q}_\phi(x+\rho,y)$.

□

B.2 PROOF OF LEMMA 5.5

With a little abuse of notation, let (t,y) denote an instance drawn from $\mathcal{D}_*$ and let (v,y) denote an instance drawn from the induced distribution $\tilde{\mathcal{D}}_\phi$ associate with a perturbation $\mathcal{Q}_\phi$. For shorter notations, we will denote $z := (t,y)$, $u := (v,y)$ and

$f(u) := f(v, y)$ and simply write $\mathcal{Q}_\phi$ as $\mathcal{Q}$.

Denote by $g(u_1 \cdots u_m) := \sup_{\theta \in \Theta} \mathrm{GG}(\theta; \tilde{S}_\phi, \tilde{\mathcal{D}}_\phi) = \sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E} f_\theta(u) \right|$. We have for any $1 \leq j \leq m$

$$\sup_{u_1, \cdots, u_m, u_j'} \left| g(u_1, \cdots, u_m) - g(u_1, \cdots, u_j', u_{j+1}, \cdots u_m) \right| \tag{22}$$

$$= \sup_{u_1, \cdots, u_m, u_j'} \left| \sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E} f_\theta(u) \right| - \sup_{\theta \in \Theta} \left| \frac{1}{m} \left(\sum_{i=1, i \neq j}^{m} f_\theta(u_i) + f_\theta(u_j') \right) - \mathbb{E}_u f_\theta(u) \right| \right| \tag{23}$$

$$\leq \sup_{u_1, \cdots, u_m, u_j'} \sup_{\theta \in \Theta} \left| \left| \frac{1}{m} \sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E} f_\theta(u) \right| - \left| \frac{1}{m} \left(\sum_{i=1, i \neq j}^{m} f_\theta(u_i) + f_\theta(u_j') \right) - \mathbb{E}_u f_\theta(u) \right| \right| \tag{24}$$

$$\leq \sup_{u_1, \cdots, u_m, u_j'} \sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E}_u f_\theta(u) - \frac{1}{m} \left(\sum_{i=1, i \neq j}^{m} f_\theta(u_i) + f_\theta(u_j') \right) + \mathbb{E}_u f_\theta(u) \right| \tag{25}$$

$$= \sup_{\theta \in \Theta} \sup_{u_j, u_j'} \frac{1}{m} \left| f_\theta(u_j) - f_\theta(u_j') \right| \tag{26}$$

$$\leq \frac{1}{m} \sup_{\theta \in \Theta} \sup_{u_j} \left| f_\theta(u_j) \right| + \frac{1}{m} \sup_{\theta \in \Theta} \sup_{u_j'} \left| f_\theta(u_j') \right| \tag{27}$$

$$\leq \frac{2B}{m} \tag{28}$$

where the inequality (25) follows from the inverse triangle inequality. The inequality (27) and (28) make use of the triangle inequality and the boundedness condition of f.

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

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

where we use $U := (u_1, \cdots, u_m)$. This is equivalent to saying that with probability $1 - \tau$, we have

$$g(u_1 \cdots u_m) \leq \mathbb{E}_U g(u_1 \cdots u_m) + 2B \sqrt{\frac{\log \frac{1}{\tau}}{2m}} \tag{29}$$

□

B.3 PROOF OF THEOREM 5.6

Following the notations in the proof of Lemma 5.5, we now derive an upper bound for the term $\mathbb{E}_U g(u_1 \cdots u_m)$.

For shorter notations, let $Z := (z_1, \cdots, z_m)$, $\Gamma := (\rho_1, \cdots, \rho_m)$ and $F_\theta(Z, \Gamma) := \frac{1}{m}\sum_{i=1}^{m} f_\theta(\mathcal{Q}(x_i + \rho_i, y_i), y_i)$. We have

$$\mathbb{E}_U g(u_1 \cdots u_m) \tag{30}$$

$$= \mathbb{E}_U \sup_{\theta\in\Theta} \left| \frac{1}{m}\sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E} f_\theta(u) \right| \tag{31}$$

$$= \mathbb{E}_U \sup_{\theta\in\Theta} \left| \frac{1}{m}\sum_{i=1}^{m} f_\theta(u_i) - \mathbb{E}_{\hat{U}} \left[\frac{1}{m}\sum_{i=1}^{m} f_\theta(\hat{u}_i) \right] \right| \tag{32}$$

$$\leq \mathbb{E}_U \sup_{\theta\in\Theta} \left[\mathbb{E}_{\hat{U}} \left| \frac{1}{m}\sum_{i=1}^{m} f_\theta(u_i) - \frac{1}{m}\sum_{i=1}^{m} f_\theta(\hat{u}_i) \right| \right] \tag{33}$$

$$\leq \mathbb{E}_U \mathbb{E}_{\hat{U}} \sup_{\theta\in\Theta} \left| \frac{1}{m}\sum_{i=1}^{m} f_\theta(u_i) - \frac{1}{m}\sum_{i=1}^{m} f_\theta(\hat{u}_i) \right| \tag{34}$$

$$= \mathbb{E}_Z \mathbb{E}_\Gamma \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \sup_{\theta\in\Theta} \left| \frac{1}{m}\sum_{i=1}^{m} f_\theta(\mathcal{Q}(x_i + \rho_i, y_i), y_i) - \frac{1}{m}\sum_{i=1}^{m} f_\theta(\mathcal{Q}(\hat{x}_i + \hat{\rho}_i, \hat{y}_i), \hat{y}_i) \right| \tag{35}$$

$$= \mathbb{E}_Z \mathbb{E}_\Gamma \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \sup_{\theta\in\Theta} \left| F_\theta(Z, \Gamma) - \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) + \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) - F_\theta(\hat{Z}, \hat{\Gamma}) + \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) \right| \tag{36}$$

$$\leq \mathbb{E}_Z \mathbb{E}_\Gamma \sup_{\theta\in\Theta} \left| F(Z, \Gamma) - \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) \right| + \mathbb{E}_{\hat{Z}} \mathbb{E}_{\hat{\Gamma}} \sup_{\theta\in\Theta} \left| F_\theta(\hat{Z}, \hat{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) \right| + \mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta\in\Theta} \left| \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) \right| \tag{37}$$

$$= \underbrace{2\mathbb{E}_Z \mathbb{E}_\Gamma \sup_{\theta\in\Theta} \left| F_\theta(Z, \Gamma) - \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) \right|}_{①} + \underbrace{\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta\in\Theta} \left| \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) \right|}_{②} \tag{38}$$

where (33) follows from Jensen's inequality and (34) is due to that the supremum of expectation is less than equal to expectation of the supremum. The inequality (37) is derived by the triangle inequality and the fact that supremum of sum is less than equal to sum of supremum. We now individually construct upper bounds for the term ① and ②.

For the term ①, we have

$$
\begin{aligned}
&2\mathbb{E}_Z\mathbb{E}_\Gamma \sup_{\theta\in\Theta}\left|F_\theta(Z,\Gamma)-\mathbb{E}_{\bar{\Gamma}}F_\theta(Z,\bar{\Gamma})\right| \\
\leq&2\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sup_{\theta\in\Theta}\left|F_\theta(Z,\Gamma)-F_\theta(Z,\bar{\Gamma})\right| && (39)\\
=&2\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sup_{\theta\in\Theta}\left|\frac{1}{m}\sum_{i=1}^m f_\theta(\mathcal{Q}(x_i+\rho_i,y_i),y_i)-\frac{1}{m}\sum_{i=1}^m f_\theta(\mathcal{Q}(x_i+\bar{\rho}_i,y_i),y_i)\right| && (40)\\
=&\frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sup_{\theta\in\Theta}\mathbb{E}_\Sigma\left|\sum_{i=1}^m \sigma_i\left(f_\theta(\mathcal{Q}(x_i+\rho_i,y_i),y_i)-f_\theta(\mathcal{Q}(x_i+\bar{\rho}_i,y_i),y_i)\right)\right| && (41)\\
\leq&\frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}} \sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^m\left|f_\theta(\mathcal{Q}(x_i+\rho_i,y_i),y_i)-f_\theta(\mathcal{Q}(x_i+\bar{\rho}_i,y_i),y_i)\right|^2} && (42)\\
\leq&\frac{2}{m}\mathbb{E}_Z\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}}\sqrt{\sum_{i=1}^m\beta^2\|\mathcal{Q}(x_i+\rho_i,y_i)-\mathcal{Q}(x_i+\bar{\rho}_i,y_i)\|^2} && (43)\\
\leq&\frac{2\beta}{m}\mathbb{E}_Z\sqrt{\mathbb{E}_\Gamma\mathbb{E}_{\bar{\Gamma}}\left[\sum_{i=1}^m\|\mathcal{Q}(x_i+\rho_i,y_i)-\mathcal{Q}(x_i+\bar{\rho}_i,y_i)\|^2\right]} && (44)\\
=&\frac{2\beta}{m}\mathbb{E}_Z\sqrt{\sum_{i=1}^m\mathbb{E}_\rho\mathbb{E}_{\bar{\rho}}\|\mathcal{Q}(x_i+\rho_i,y_i)-\mathcal{Q}(x_i+\bar{\rho}_i,y_i)\|^2} && (45)\\
=&\frac{2\beta}{m}\mathbb{E}_Z\sqrt{\sum_{i=1}^m\gamma(x_i,y_i)} && (46)\\
\leq&\frac{2\beta}{m}\sqrt{\mathbb{E}_Z\left[\sum_{i=1}^m\gamma(x_i,y_i)\right]} && (47)\\
=&\frac{2\beta}{m}\sqrt{\sum_{i=1}^m\mathbb{E}_{z_i}\gamma(x_i,y_i)} && (48)\\
=&\frac{2\beta}{\sqrt{m}}\sqrt{\mathbb{E}_z\gamma(x,y)} && (49)
\end{aligned}
$$

The inequality (39) is derived similarly to inequality (33) and (34). In (41), 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 (41) holds. The inequality (42) is given by the Khintchine's inequality. The inequality (43) makes use of the Lipschitz condition of f. (44) is derived from Jensen's inequality and due to that square root is a concave function. (46) is by the definition of the local dispersion of $\mathcal{Q}$. Again, we apply Jensen's inequality to obtain (47). Equation (48) and (49) follow from the settings that each $z_i = (x_i, y_i)$ is i.i.d.

For the term ②, we have

$$
\begin{aligned}
&\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta \in \Theta} \left| \mathbb{E}_{\bar{\Gamma}} F_\theta(Z, \bar{\Gamma}) - \mathbb{E}_{\tilde{\Gamma}} F_\theta(\hat{Z}, \tilde{\Gamma}) \right| \\
=&\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta \in \Theta} \left| \mathbb{E}_{\bar{\Gamma}} \left[\frac{1}{m} \sum_{i=1}^{m} f_\theta(\mathcal{Q}(x_i + \bar{\rho}_i, y_i), y_i) \right] - \mathbb{E}_{\tilde{\Gamma}} \left[\frac{1}{m} \sum_{i=1}^{m} f_\theta(\mathcal{Q}(\hat{x}_i + \tilde{\rho}_i, \hat{y}_i), \hat{y}_i) \right] \right| && (50) \\
=&\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\bar{\rho}_i} \left[f_\theta(\mathcal{Q}(x_i + \bar{\rho}_i, y_i), y_i) \right] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\tilde{\rho}_i} \left[f_\theta(\mathcal{Q}(\hat{x}_i + \tilde{\rho}_i, \hat{y}_i), \hat{y}_i) \right] \right| && (51) \\
=&\mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta \in \Theta} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right] \right| && (52) \\
=&\frac{1}{m} \mathbb{E}_Z \mathbb{E}_{\hat{Z}} \mathbb{E}_\Sigma \sup_{\theta \in \Theta} \left| \sum_{i=1}^{m} \sigma_i \left(\mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right] - \mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right] \right) \right| && (53) \\
\leq&\frac{1}{m} \mathbb{E}_Z \mathbb{E}_{\hat{Z}} \sup_{\theta \in \Theta} \sqrt{\sum_{i=1}^{m} \left| \left(\mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right] - \mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right] \right) \right|^2} && (54)
\end{aligned}
$$

where equation (51) and (52) are due to each $\hat{\rho}_i$ and $\tilde{\rho}_i$ is i.i.d. Again, we introduce Rademacher variables at (53) and apply Khintchine's inequality to get (54). For the term $\left| \left(\mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right] - \mathbb{E}_{\rho} \left[f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right] \right) \right|^2$, we have

$$
\begin{aligned}
&\left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) - \mathbb{E}_\rho f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right|^2 && (55) \\
\leq&\left(\left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right| + \left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right| \right)^2 && (56) \\
\leq&2 \left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(x_i + \rho, y_i), y_i) \right|^2 + 2 \left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(\hat{x}_i + \rho, \hat{y}_i), \hat{y}_i) \right|^2 && (57)
\end{aligned}
$$

where inequality (57) is derived by the inequality $(a+b)^2 \leq 2(a^2 + b^2)$. We also have that

$$
\begin{aligned}
&\left| \mathbb{E}_\rho f_\theta(\mathcal{Q}(x + \rho, y), y) \right|^2 \\
\leq&\left(\mathbb{E}_\rho \left| f_\theta(\mathcal{Q}(x + \rho, y), y) - f_\theta(x + \rho, y) \right| + \left| f_\theta(x + \rho, y) \right| \right)^2 && (58) \\
\leq&\left(\mathbb{E}_\rho \left| f_\theta(\mathcal{Q}(x + \rho, y), y) - f_\theta(x + \rho, y) \right| + B \right)^2 && (59) \\
\leq&\left(\mathbb{E}_\rho \beta \| \mathcal{Q}(x + \rho, y) - (x + \rho) \|_2 + B \right)^2 && (60)
\end{aligned}
$$

The inequalities (58)-(60) respectively make use of the triangle inequality, Jensen's inequality, and the boundedness and lipschitz condition of f.

Returning to (54), we then have

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

$$\leq\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{m}2\left|\mathbb{E}_\rho f_\theta(\mathcal{Q}(x_i+\rho,y_i),y_i)\right|^2+\sum_{i=1}^{m}2\left|\mathbb{E}_\rho f_\theta(\mathcal{Q}(\hat{x}_i+\rho,\hat{y}_i),\hat{y}_i)\right|^2} \tag{61}$$

$$\leq\frac{1}{m}\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\sqrt{\sum_{i=1}^{m}2(\mathbb{E}_\rho\beta\|\mathcal{Q}(x_i+\rho,y_i)-(x_i+\rho)\|_2+B)^2+\sum_{i=1}^{m}2(\mathbb{E}_\rho\beta\|\mathcal{Q}(\hat{x}_i+\rho,\hat{y}_i)-(\hat{x}_i+\rho)\|_2+B)^2} \tag{62}$$

$$\leq\frac{1}{m}\sqrt{\mathbb{E}_Z\mathbb{E}_{\hat{Z}}\left[\sum_{i=1}^{m}2(\mathbb{E}_\rho\beta\|\mathcal{Q}(x_i+\rho,y_i)-(x_i+\rho)\|_2+B)^2+\sum_{i=1}^{m}2(\mathbb{E}_\rho\beta\|\mathcal{Q}(\hat{x}_i+\rho,\hat{y}_i)-(\hat{x}_i+\rho)\|_2+B)^2\right]} \tag{63}$$

$$\leq\frac{2}{\sqrt{m}}\sqrt{\mathbb{E}_z(\mathbb{E}_\rho\beta\|\mathcal{Q}(x+\rho,y)-(x+\rho)\|_2+B)^2} \tag{64}$$

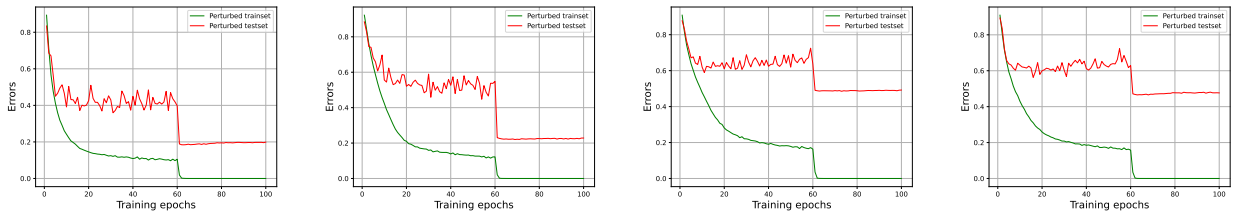
$$\leq\frac{2(\beta\sqrt{d}\epsilon+B)}{\sqrt{m}} \tag{65}$$

The final line is due to that with $\|\mathcal{Q}(x+\rho)-(x+\rho)\|_\infty\leq\epsilon$ we have $\|\mathcal{Q}(x+\rho)-(x+\rho)\|_2\leq\sqrt{d}\epsilon$. This gives the final result

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

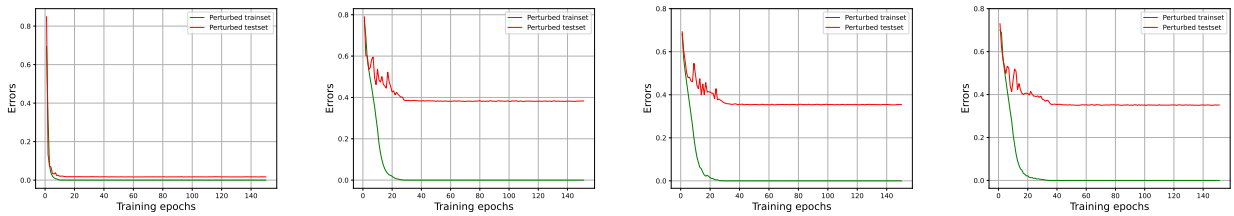
□

C OMITTED FIGURES



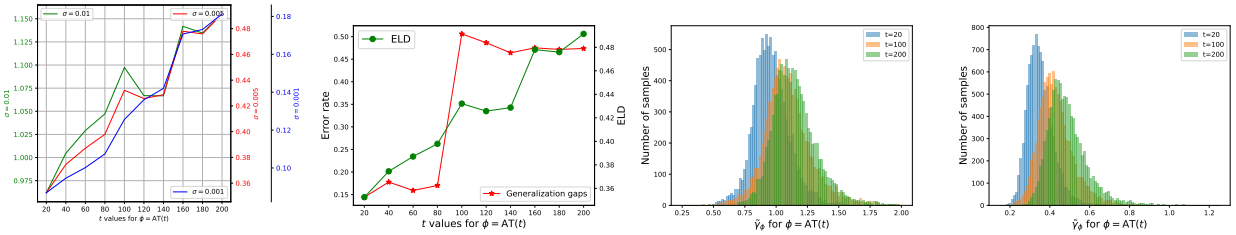
(a) ϕ=AT(0) (b) ϕ=AT(80) (c) ϕ=AT(120) (d) ϕ=AT(200)

Figure 5: Experiments in Figure 1 reproduced on CIFAR-100.



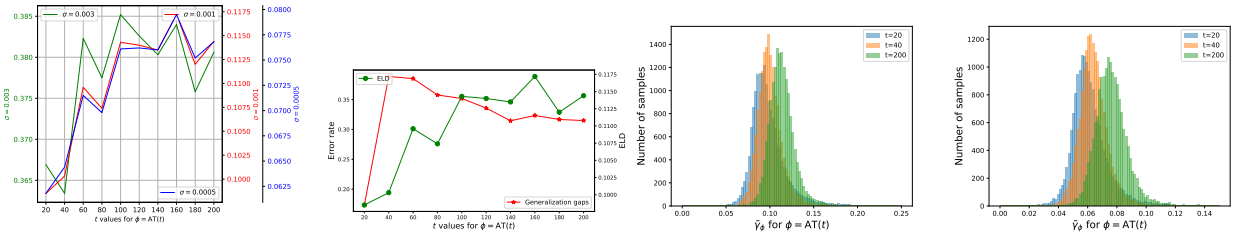
(a) ϕ=AT(0) (b) ϕ=AT(80) (c) ϕ=AT(120) (d) ϕ=AT(200)

Figure 6: Experiments in Figure 1 reproduced on Reduced ImageNet.



(a) ELDs for different σ (b) $\sigma = 0.005$ (c) $\sigma = 0.01$ (d) $\sigma = 0.005$

Figure 7: Experiments in Figure 2 reproduced on CIFAR-100.



(a) ELDs for different σ (b) $\sigma = 0.001$ (c) $\sigma = 0.001$ (d) $\sigma = 0.0005$

Figure 8: Experiments in Figure 2 reproduced on Reduced ImageNet.