Improving Adversarial Robustness via Unlabeled Out-of-Domain Data

Zhun Deng [*]
Harvard University

Linjun Zhang [*]
Rutgers University

Amirata Ghorbani
Stanford University

James Zou
Stanford University

Abstract

Data augmentation by incorporating cheap unlabeled data from multiple domains is a powerful way to improve prediction especially when there is limited labeled data. In this work, we investigate how adversarial robustness can be enhanced by leveraging out-of-domain unlabeled data. We demonstrate that for broad classes of distributions and classifiers, there exists a sample complexity gap between standard and robust classification. We quantify the extent to which this gap can be bridged by leveraging unlabeled samples from a shifted domain by providing both upper and lower bounds. Moreover, we show settings where we achieve better adversarial robustness when the unlabeled data come from a shifted domain rather than the same domain as the labeled data. We also investigate how to leverage out-of-domain data when some structural information, such as sparsity, is shared between labeled and unlabeled domains. Experimentally, we augment object recognition datasets (CIFAR-10, CINIC-10, and SVHN) with easy-to-obtain and unlabeled out-of-domain data and demonstrate substantial improvement in the model's robustness against ℓ_∞ adversarial attacks on the original domain.

1 Introduction

Robustness to adversarial attacks has been a major focus in machine learning security (Biggio & Roli, 2018; Dalvi et al., 2004; Lowd & Meek, 2005), and has been intensively studied in the past few years (Goodfellow et al., 2014; Carlini & Wagner, 2017b; Nguyen et al., 2015). However, the theoretical understanding of adversarial robustness is still far from being satisfactory. Recently Schmidt et al. (2018) have demonstrated sample complexity may be one of the obstacles in achieving high robustness under standard learning, which is a large challenge since in many real-world applications, labeled examples are few and expensive. To address this challenge, recent works (Carmon et al., 2019; Stanforth et al., 2019) showed that adversarial robustness can be improved by leveraging unlabeled data that come from the same distribution/domain as the original labeled training samples. Nevertheless, that is still limited due to the difficulty to make sure that the unlabeled data are exactly from the same distribution as the labeled data. For example, gathering a large number of unlabeled images that follow the same distribution as CIFAR-10 is challenging, since one would have to carefully match the same lighting conditions, backgrounds, etc. Meanwhile, out-of-domain unlabeled data can be much easier and cheaper to collect. For instance, we used Bing search engine to query a small number of keywords and, within hours, generated a new 500k dataset of noisy CIFAR-10 categories; we call this Cheap-10 (available at `https://tinyurl.com/mere5j0x`). Despite being fast and easy to collect, we show that using Cheap-10 can substantially improve the adversarial robustness of the original CIFAR-10 classifier.

Our contributions In this paper, we investigate how such widely-available, out-of-domain unlabeled data could improve robustness in the original domain. We analyze the behavior of standard and robust classification under a flexible generative model with Gaussian seeds and a non-linear classifier class. Our model and classifier classes can be viewed as an extension of the Gaussian model and linear classifier class proposed in Schmidt et al. (2018). We show in this more general setting, the sample complexity gap between standard and robust classification still exists. That is, to achieve the same amount of accuracy, the sam-

[*] Equal contribution.

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

ple complexity of robust training is significantly larger than that of standard training. We also demonstrate the necessity of this gap by providing a minimax type lower bound result. Luckily, we show that using unlabeled out-of-domain data can substantially improve robust accuracy as long as the unlabeled domain is not too different from the original domain or if they share some (unknown) structural information, such as similar sparse features. Interestingly, we further show settings where using out-of-domain unlabeled data can produce even better robust accuracy than using in-domain unlabeled data.

We support our theory with experiments on three benchmark image recognition tasks, CIFAR-10, CINIC-10 and SVHN, for empirical ℓ_∞ robustness and certified ℓ_2 robustness. In both CIFAR-10 and CINIC-10, adding our easily generated Cheap-10 unlabeled data produces substantially higher robust accuracy than using just the CIFAR-10 or CINIC-10 data.On SVHN, we systematically characterize the tradeoff between the amount of noise in the unlabeled data and the robustness gain from adding such data.

Related works. After the successful implementation of white-box and black-box adversarial examples (Goodfellow et al., 2014; Biggio & Roli, 2018; Moosavi-Dezfooli et al., 2016), several heuristic defense methods were introduced and broken one after another (Zantedeschi et al., 2017; Athalye et al., 2018; Guo et al., 2017; Biggio et al., 2013; Carlini & Wagner, 2017a; Athalye et al., 2017). A line of work has focused on certified robustness (Cohen et al., 2019; Lecuyer et al., 2019; Raghunathan et al., 2018; Liu et al., 2020; Chiang et al., 2020) which has appealing guarantees but has relatively limited empirical performance. Most recent efforts on training empirically robust models is based on adversarial training (Madry et al., 2017; Zhang et al., 2019; Kurakin et al., 2016; Hendrycks et al., 2019). Theoretically, some works justify the efficiency of adversarial training (Deng et al., 2020b), and to explain why it is difficult to achieve satisfactory performance in robust learning, some works try to explain the obstacles to gain robustness in a perspective of computation cost (Bubeck et al., 2018; Degwekar et al., 2019). Meanwhile, other works demonstrate how to quantify the trade-off of adversarial robustness and standard accuracy (Deng et al., 2020a) and data augmentation such as Mixup could mitigate the trade-off (Zhang et al., 2020). In addition, work such as Schmidt et al. (2018) try to explain the obstacle by showing the sample complexity of robust learning can be significantly larger than that of standard learning. They investigated the Gaussian model, which is a special case of our Gaussian generative model. Some recent works (Carmon et al., 2019; Stanforth et al., 2019) propose using semi-supervised learning method, which has a rich literature (Laine & Aila, 2016; Miyato et al., 2018; Sajjadi et al., 2016), to bridge that sample gap. Their theoretical results all assume the unlabeled data are drawn from the same marginal distribution as the labeled data. We show that to bridge the sample complexity gap, it is sufficient to have well-behaved unlabeled out-of-domain data. We substantially extend the previous results to more general models and classifier classes and also make the first step to quantify when and how unlabeled data coming from a shifted distribution can help in improving adversarial robustness. Experimentally, previous works augmented CIFAR-10 with Tiny images, which is curated and very similar to CIFAR-10. We introduce a new dataset Cheap-10 and obtain comparable results and demonstrate the power of incorporating out-of-domain data. Other related works include Zhai et al. (2019), which demonstrates a PCA-based procedure to incorporate unlabeled data to gain robustness and Najafi et al. (2019), who consider combining distributional robust optimization and semi-supervised learning.

2 Set-up

Consider the classification task of mapping the input $x \in \mathcal{X} \subseteq \mathbb{R}^{s_1}$ to the label $y \in \{\pm 1\}$. We have n labeled training data from an original domain $\mathcal{D}$, with a joint distribution $\mathcal{P}_{x,y}$ over (x, y) pairs and marginal distribution $\mathcal{P}_x$ over x. Meanwhile, we have another $\tilde{n}$ unlabeled samples from a different domain $\tilde{\mathcal{D}}$, with a distribution $\tilde{\mathcal{P}}_x$ over x.

In this work, we focus on studying the possible advantages and limitations by performing semi-supervised learning with data from $\mathcal{D}$ and $\tilde{\mathcal{D}}$ to train a classifier for the domain $\mathcal{D}$. Specifically, we apply the pseudo-labeling approach used in Carmon et al. (2019) as follows. First, we perform supervised learning on the labeled data from domain $\mathcal{D}$ to obtain a classifier f_0. We then apply this classifier on $\tilde{\mathcal{D}}$ and generate pseudo-labels for the unlabeled data: $\{(x, f_0(x)) | x \in \tilde{D}\}$, which are further used to train a final model. The classification error metrics we consider are defined as the following.

Definition 1 ((Robust) classification error)**.** *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$. The classification error β of a classifier $g : \mathbb{R}^{s_1} \mapsto \{\pm 1\}$ is defined as $\beta_g = \mathbb{P}_{(x,y)\sim\mathcal{P}}(g(x) \neq y)$ and robust classification error $\beta_g^{\mathcal{R}} = \mathbb{P}_{(x,y)\sim\mathcal{P}}(\exists u \in \mathcal{C}(x) : g(u) \neq y)$, for some constraint set $\mathcal{C}$.*

Throughout the paper, we consider the constraint set $\mathcal{C}$ to be the ℓ_p-ball $\mathbb{B}_p(x, \varepsilon) := \{u \in \mathcal{X} | \|u - x\|_p \leqslant \varepsilon\}$

with $p=\infty$. In addition, we consider a certain type of data generating process for domain $\mathcal{D}$ — the Gaussian generative model, which is frequently used in generative models in machine learning. This model is more general than the one analyzed in Schmidt et al. (2018), which only considered symmetric Gaussian mixtures. Our Gaussian generative model takes a sample from a Gaussian mixture as input, and then pass it through a nonlinear (possibly high-dimensional) mapping.

Gaussian generative model. For a function $\rho : \mathbb{R}^{s_1} \mapsto \mathbb{R}^{s_2}$, given $z \in \mathbb{R}^{s_1}$, the samples from $\mathcal{D}$ are drawn i.i.d. from a distribution over $(x, y) \in \mathbb{R}^{s_2} \times \{\pm 1\}$, such that

$$x = \rho(yz) \tag{1}$$

where $z \sim \mathcal{N}(\mu, \sigma^2 I_{s_1})$, $y \sim Bern(\frac{1}{2})$ for $\mu \in \mathbb{R}^{s_1}$, $\sigma \in \mathbb{R}$.

Remark 1. *The Gaussian generative model is very flexible and includes many of the recent machine learning models. For example, many common deep generative models such as VAE and GAN are Gaussian generative models: in their case, the input is a Gaussian sample z and ρ is parametrized by a neural network. Therefore our results are quite generally applicable.*

Classifier class. The classifier class we consider in this paper is in the following form:

$$\mathcal{G} = \left\{g | g(x) = sgn\big(w^\top(\vartheta(x) - b)\big), (w, b) \in \mathbb{R}^d \times \mathbb{R}^d\right\}, \tag{2}$$

where ϑ is a basis function and $\vartheta : \mathbb{R}^{s_2} \mapsto \mathbb{R}^d$. We remark here that this classifier class is more general than the linear classifier class considered in Schmidt et al. (2018). For a broad class of kernels, by Mercer's theorem, the corresponding kernel classification belongs to $\mathcal{G}$ with a certain basis function ϑ. Throughout the paper, we use $\theta = (w, b)$ to denote the parameters.

Remark 2. *The Gaussian generative model and the classifier class we considered in this paper forms a hierarchy structure, where a random seed $z \in \mathbb{R}^{s_1}$ is mapped by a generative function ρ to the input space $x \in \mathbb{R}^{s_2}$, and it is further mapped to $\mathbb{R}^d$ by $\vartheta(x)$ when implementing classification.*

Notations and terminology. We let $\phi = \vartheta \circ \rho$ and denote $f_{w,b}(x) = sgn(w^\top(\vartheta(x) - b))$. In Section 3, the results will be mainly described in terms of ϕ. Besides, let $\beta(w, b) = \mathbb{P}_{(x,y)\sim\mathcal{P}}(f_{w,b}(x) \neq y)$ and $\beta^{\mathcal{R}}(w, b) = \mathbb{P}_{(x,y)\sim\mathcal{P}}(\exists u \in \mathcal{C}(x) : f_{w,b}(u) \neq y)$ for a constraint set $\mathcal{C}$. In particular, we use $\beta^{\varepsilon,\infty}$ when $\mathcal{C}$ is the ℓ_∞-ball with radius ε. Meanwhile, we use $\|\cdot\|_{\psi_2}$ for sub-gaussian norm[1]. We call the conditional distribution of x on $y=1$ as positive distribution while for $y=-1$ as negative distribution. For distribution $\mathcal{P}_1$ and $\mathcal{P}_2$ over x, we call a distribution $\mathcal{P}$ over x is a *uniform mixture* of $\mathcal{P}_1$ and $\mathcal{P}_2$ if it equals to $\mathcal{P}_1$ and $\mathcal{P}_2$ with probability $1/2$ respectively. For a sequence of random variables $\{X_n\}$ and a sequences of positive numbers $\{a_n\}$, we write $X_n = O_{\mathbb{P}}(a_n)$ if there exists a constant C, such that $\mathbb{P}(X_n \leq Ca_n) \to 1$ when $n \to \infty$. For real-valued sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \leqslant cb_n$ for some universal constant $c \in (0, \infty)$, and $a_n \gtrsim b_n$ if $a_n \geqslant c' b_n$ for some universal constant $c' \in (0, \infty)$. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. In this paper, $c, C, c_0, c_1, c_2, \cdots$, refer to universal constants, and their specific values may vary from place to place.

3 Theoretical Results

We demonstrate for Gaussian generative models that combining unlabeled data from a reasonably well-behaved shifted domain leads to a classifier with better robust accuracy on the original domain $\mathcal{D}$ compared to the achievable robust accuracy using only the labeled data from $\mathcal{D}$. We further analyze the tradeoff between how different the shifted domain can be from $\mathcal{D}$ before the unlabeled data hurts the robust accuracy on $\mathcal{D}$. Finally, we show that if the data from a shifted domain share certain unknown sparsity structure with the data from original domain, performing semi-supervised learning also helps in obtaining a classifier of higher robust accuracy on the original domain.

Assumptions. Throughout this section, our theories are based on the following assumptions unless we state otherwise explicitly. 1). $\vartheta(\cdot)$ is L_1-Lipchitz continuous in ℓ_2-norm, i.e. $\|\vartheta(a) - \vartheta(b)\| \leqslant L_1\|a - b\|$, and L_1'-Lipchitz continuous in ℓ_∞-norm; 2). $\rho(\cdot)$ is L_2-Lipchitz continuous in ℓ_2-norm and L_2'-Lipchitz continuous in ℓ_∞-norm; 3). $\|\mathbb{E}\phi(z) - \mathbb{E}\phi(-z)\| = 2\alpha\sqrt{d}$ for $z \sim \mathcal{N}(\mu, \sigma^2 I_{s_1})$ and some constant $\alpha > 0$. The last condition on the magnitude of the separation is added for the simplicity of presentation. Such a magnitude choice is also used in Schmidt et al. (2018); Carmon et al. (2019).

3.1 Supervised learning in Gaussian generative models

We first consider the supervised setting where only the labeled data are used. In this setting, we prove the following two theorems demonstrating the sample complexity gap when one considers standard error and robust error respectively. Analogous results for Gaussian mixture models was shown in Schmidt et al. (2018); our results cover the more general Gaussian generative model setting.

[1] Due to the limit of space, we present the rigorous definition of the sub-gaussian norm in the appendix.

Supervised learning algorithm: in this section, for the simplicity of presentation, we use $2n$ to denote the size of labeled training data. For Gaussian generative models, we focus on the following method. We first estimate w and b by $\hat{w} = 1/n\sum_{i=1}^{n} y_i\vartheta(x_i)$ and $\hat{b} = 1/n\sum_{i=n+1}^{2n}\vartheta(x_i)$. The final classifier is then constructed as $f_{\hat{w},\hat{b}}(x) = sgn(\hat{w}^\top(\vartheta(x) - \hat{b}))$. Here half of the labeled data, n, is used to fit $\hat{w}$ and the other half used to fit $\hat{b}$, so that their estimation errors are independent, which simplifies the analysis. The following theorem shows that this method achieves high standard accuracy.

Theorem 1 (Standard accuracy)**.** *For a Gaussian generative model with $\sigma \lesssim d^{1/4}$, the method described above obtains a classifier $f_{\hat{w},\hat{b}}$ such that for d sufficiently large, with high probability, the classification error $\beta(\hat{w},\hat{b})$ is at most 1% even with $n = 1$.*

Meanwhile, we have the following lower bound to show the essentiality of the increased sample complexity if we are interested in the robust error.

Theorem 2 (Sample complexity gap for robust accuracy)**.** *Let $\mathcal{A}_n$ be any learning algorithm, i.e. a function from n samples to a binary classifier g_n. Let $\sigma \asymp d^{1/4}$, $\varepsilon \geqslant 0$, and $\mu \in \mathbb{R}^{s_1}$ be drawn from a prior distribution $\mathcal{N}(0, I_{s_1})$. We draw $2n$ samples from (μ,σ)-Gaussian generative model. Then, the expected robust classification error $\beta_{g_n}^{\varepsilon,\infty}$ is at least $(1-1/d)/2$ if*

$$n \lesssim \frac{\varepsilon^2\sqrt{d}}{\log d}.$$

Taken together, these two Theorems demonstrate that a substantial larger number of labeled samples (from the same domain) are necessary in order to achieve a decent robust accuracy in that domain.

3.2 Improving learning via out-of-domain data

We next investigate how to improve the robust accuracy of a classifier via incorporating unlabeled out-of-domain data.

Semi-supervised learning on out-of-domain data. Let us denote the samples from the shifted domain as $\{\tilde{x}_i\}_{i=1}^{2\tilde{n}}$, which is incorporated via the following semi-supervised learning algorithm.

Semi-supervised learning algorithm: we use $\hat{w}$ and $\hat{b}$ obtained in supervised learning to label $\{\tilde{x}_i\}_{i=1}^{2\tilde{n}}$ via $\hat{g}(x) = sgn\big(\hat{w}^\top(\vartheta(x) - \hat{b})\big)$ and obtain the corresponding pseudo-labels $\{\tilde{y}_i\}_{i=1}^{2\tilde{n}}$. We denote sample sizes for each label class by $\tilde{n}_1 = \sum_{i=1}^{2\tilde{n}}1(\tilde{y}_i = 1)$ and $\tilde{n}_2 = \sum_{i=1}^{2\tilde{n}}1(\tilde{y}_i = -1)$ respectively. Then we estimate w and b respectively by

$$\tilde{w} = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\vartheta(\tilde{x}_i) - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}^{\tilde{n}}\vartheta(\tilde{x}_i)$$
$$\tilde{b} = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\vartheta(\tilde{x}_i) + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\vartheta(\tilde{x}_i).$$

Given the pseudo-labels, these two estimators only depend on the shifted domain data. They are slightly different than those in the supervised setting, since the shifted domain data is not necessarily mixed uniformly. The classifier is then constructed as $f_{\tilde{w},\tilde{b}}(x) = sgn(\tilde{w}^\top(\vartheta(x) - \tilde{b}))$. For the simplicity of theoretical analysis, we don't merge the original and out-of-domain datasets to get $\tilde{w}$ and $\tilde{b}$. However, as we show in Section 4, merging both datasets for robust training lead to better empirical performance.

Recall $\phi = \vartheta \circ \rho$, and the semi-supervised learning algorithm only involves y and $\vartheta(x)$, we can equivalently view the input distribution as $\phi(yz)$ for $z \sim \mathcal{N}(0, I_{s_1})$, $y \sim Bern(1/2)$, and the classifier class as $\mathcal{G}' = \big\{g \big| g(x) = sgn\big(w^\top x - b\big)\big), (w,b) \in \mathbb{R}^d \times \mathbb{R}^d\big\}$ (such linearization is the common purpose of kernel tricks). For the simplicity of description, our later statements will use this equivalent setting and simply consider the distributions of $\vartheta(\tilde{x})$.

Theorem 3 (Robust accuracy)**.** *Recall in Gaussian generative model, the marginal distribution of the input x of labeled domain is a uniform mixture of two distributions with mean $\mu_1 = \mathbb{E}[\phi(z)]$ and $\mu_2 = \mathbb{E}[\phi(-z)]$ respectively, where $z \sim \mathcal{N}(0, \sigma^2 I_{s_1})$. Suppose the marginal distribution of the input of unlabeled domain is a mixture of two sub-gaussian distributions with mean $\tilde{\mu}_1$ and $\tilde{\mu}_2$ with mixing probabilities q and $1-q$ and $\big\|\mathbb{E}\left[\vartheta(\tilde{x}_i) - \mathbb{E}[\vartheta(\tilde{x}_i)] \mid a^T\vartheta(\tilde{x}_i) = b\right]\big\| \lesssim \sqrt{d} + |b|$ for fixed unit vector a. Assuming the sub-gaussian norm for both labeled and unlabeled data are upper bounded by a universal quantity $\sigma_{\max} \asymp d^{1/4}$, $\|\tilde{\mu}_1 - \tilde{\mu}_2\|_2 \asymp \sqrt{d}$, $c < q < 1-c$ for some constant $0 < c < 1/2$, and*

$$d_\nu = \max\left\{\frac{\|\tilde{\mu}_1 - \mu_1\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}, \frac{\|\tilde{\mu}_2 - \mu_2\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}\right\} < c_0,$$

for some constant $c_0 \leq 1/4$, then the robust classification error is at most 1% when d is suffciently large, $n \geq C$ for some constant C (not depending on d and ϵ) and

$$\tilde{n} \gtrsim \varepsilon^2 \log d\sqrt{d}.$$

Remark 3. *We remark here that $\sigma_{\max}$, the upper bound of the sub-gaussian norm of the Gaussian generative model, is upper bounded by $L_1L_2\sigma$. Comparing to the Theorem 2, which shows that the sample complexity of order $\varepsilon^2\sqrt{d}/\log d$ is necessary to achieve* *small robust error, the above theorem shows that, by incorporating the same order of similar unlabeled data (up to a logarithm factor), which is generally cheaper, one can achieve the same robust accuracy. We further note that the sub-Gaussian assumption in Theorem 3 is quite relaxed. For example, any dataset where the feature values are bounded are automatically sub-Gaussian. This includes all image data since the pixel values are bounded.*

Remark 4. *Moreover, by using the same technique, we can extend our theoretical results to a much more general family of distributions in $\mathbb{R}^d$ whose tails are bounded by any strictly decreasing function g. Define*

$$D_g(\mu,\sigma^2)=\{X\in\mathbb{R}^d:\forall v\in\mathbb{R}^d,\|v\|_2=1,\, Var(X_j)\le\sigma^2 \\ \mathbb{P}(|v^T(X-\mu)|>\sigma\cdot t)\le g(t)\}.$$

For example letting $g(t)=C_1e^{-C_2t^2}$ reduces D_g to the family of sub-Gaussian distributions. The in-domain distribution is now assumed to be $x_1,...,x_n\overset{i.i.d.}{\sim}1/2\cdot D_g(\mu_1,\sigma^2)+1/2\cdot D_g(\mu_2,\sigma^2)$ and the out-of-domain distribution is assumed to be $\tilde{x}_1,...,\tilde{x}_{\tilde{n}}\overset{i.i.d.}{\sim}q\cdot D_g(\tilde{\mu}_1,\tilde{\sigma}^2)+(1-q)\cdot D_g(\tilde{\mu}_2,\tilde{\sigma}^2)$ where $q\in(0,1/2)$. We remark here that this extension allows the in-domain and out-of-domain distributions to be very different as long as they are all in the family $D_g(\cdot,\cdot)$. We present the following proposition for this extension.

Proposition 1. *Under the similar assumptions to those in Theorem 3.3, that is, $\left\|\mathbb{E}\left[\tilde{x}_i-\mathbb{E}[\tilde{x}_i]\mid a^T\tilde{x}_i=b\right]\right\|\lesssim\sqrt{d}+|b|$ for fixed unit vector a, $\tilde{\sigma}\le\sigma_{\max}\asymp d^{1/4}$, $\|\tilde{\mu}_1-\tilde{\mu}_2\|_2\asymp\sqrt{d}$, $c<q<1-c$ for some constant $0<c<1/2$, and*

$$d_\nu=\max\left\{\frac{\|\tilde{\mu}_1-\mu_1\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|},\frac{\|\tilde{\mu}_2-\mu_2\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}\right\}<c_0,$$

for some constant $c_0\le 1/4$, then the robust classification error is at most 1% when d is sufficiently large, $n\ge C$ for some constant C (not depending on d and ε) and

$$\tilde{n}\gtrsim\varepsilon^2\cdot(g^{-1}(1/d\log d))^2\cdot\sqrt{d}.$$

Connections to statistical measures. A key quantity in Theorem 3 is d_ν, which quantifies the difference between the labeled and unlabeled domain. In this section, we make connections between some commonly used statistical measures and d_ν via some more specific examples.We establish connections to Wasserstein Distance, Maximal Information and $\mathcal{H}$-Divergence. Due to the limit of space, we only demonstrate the result for Wasserstein Distance here and put the other results in the appendix. Throughout this paragraph, we consider the distribution of the labeled domain with positive distribution $\mathcal{P}_1$, negative distribution $\mathcal{P}_2$, and $y\sim Bern(1/2)$. The marginal distributions of shifted domain is assumed to be a uniform mixtures of $\tilde{\mathcal{P}}_1$ and $\tilde{\mathcal{P}}_2$.

Wasserstein Distance: the Wasserstein Distance induced by metric ρ between distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathbb{R}^d$ is defined as

$$W_\rho(\mathcal{P}_1,\mathcal{P}_2)=\sup_{\|f\|_{\text{Lip}}\leqslant 1}\left[\int fd\mathcal{P}_1-fd\mathcal{P}_2\right],$$

where $\|f\|_{\text{Lip}}\leqslant 1$ indicates the class of $f:\mathbb{R}^d\mapsto\mathbb{R}$ such that for any $x,x'\in\mathbb{R}^d$, $|f(x)-f(x)|\leqslant\rho(x,x')$. Let us consider $\rho(x,x')=\|x-x'\|$.

Proposition 2. *Under the assumption that $\max\{W_\rho(\mathcal{P}_1,\tilde{\mathcal{P}}_1),W_\rho(\mathcal{P}_2,\tilde{\mathcal{P}}_2)\}\leqslant\tau$, for $\tau\geqslant 0$, then we have $\|\mu_i-\tilde{\mu_i}\|\leqslant\tau$ for $i=1,2$. As a result,*

$$d_\nu\leqslant\frac{\tau}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}.$$

If we further have $\tau\leqslant\|\mu_1-\mu_2\|/2$, we will then have $d_\nu\leqslant\tau/(\|\mu_1-\mu_2\|-2\tau)$.

As we can see, when the Wasserstein distance get smaller, the quantity d_ν decreases.

Data from a shifted domain can work even better. Theorem 3 demonstrates the sample complexity gap in Section 3.1 can be bridged via out-of-domain data. Next, we show that in certain settings, one can achieve even better adversarial robustness when the unlabeled data comes from a shifted domain rather than the same domain as the labeled data. To illustrate this phenomenon, let us analyze a specific example of our model — the Gaussian model proposed in Schmidt et al. (2018).

Theorem 4. *Suppose the distribution of the labeled domain has positive distribution $\mathcal{N}(\mu,\sigma^2I_d)$ and negative distribution $\mathcal{N}(-\mu,\sigma^2I_d)$ with $\mu\in\mathbb{R}^d$ and $y\sim Bern(1/2)$. Samples $(x_1,y_1),\cdots,(x_n,y_n)$ are i.i.d. drawn from the labeled domain. Suppose we have unlabeled inputs from the same domain $x_{n+1},\cdots,x_{n+\tilde{n}}$, and also have unlabeled shifted domain inputs $\tilde{x}_1,\cdots,\tilde{x}_{\tilde{n}}$, which are drawn from a uniform mixture of $\mathcal{N}(\tilde{\mu},\sigma^2I_d)$ and $\mathcal{N}(-\tilde{\mu},\sigma^2I_d)$ with $\tilde{\mu}\in\mathbb{R}^d$. Denote the parameter $\theta=(w,b)$ of the classifier obtained through semi-supervised algorithm by $\hat{\theta}_{same}$ and $\hat{\theta}_{shifted}$, when we use $\{x_i\}_{i=n+1}^{n+\tilde{n}}$ and $\{\tilde{x}_i\}$ respectively. If we let $\mu=2\varepsilon 1_d$ and $\tilde{\mu}=\varepsilon 1_d$, where 1_d is a d-dimensional vector with every entry equals to 1, when $(1/d+1/\tilde{n})\cdot\sigma^2/\epsilon^2\to 0$ as d, $\tilde{n}\to\infty$, we then have*

$$\beta^{\varepsilon,\infty}(\hat{\theta}_{shifted})\le\beta^{\varepsilon,\infty}(\hat{\theta}_{same}).$$

This seemingly surprising result can be explained intuitively. Heuristically, when one tries to minimize the robust error, the robust optimizer will behave similarly to a regularized version of the standard optimizer. In our semi-supervised setting, the shifted domain data also act as regularization. Such an intuition is rigorously justified in the proof. Further, we illustrate the results in Theorem 4 by experiments with synthetic data, the experiment set-up and results are presented in the appendix, where we find that the robust error by incorporating the out-of-domain unlabeled data is smaller than than incorporating the same amount of unlabeled data from the same domain as the labeled data.

Too irrelevant unlabeled data hurts robustness. In the results above, we demonstrate that incorporating unlabeled data from a shifted domain can improve robust accuracy in the original domain, if the shifted domain is not too different from the original (as measured by d_ν). Here we show that there is no free lunch; if the shifted domain is too different from the original, then incorporating its unlabeled data through pseudo-labeling could decrease the robust accuracy in the original domain.

Theorem 5. *Suppose that the distribution of labeled domain's positive and negative distribution area uniform two symmetric sub-gaussian distribution with means $\mu_2 = -\mu_1$, and $y \sim Bern(1/2)$. The distribution of unlabeled domain $\mathcal{P}'$ is a mixture of two sub-gaussian distributions with mean $\tilde{\mu}_1$ and $\tilde{\mu}_2$. Let $\Xi = \{\tilde{\mu}_1, \tilde{\mu}_2 : d_\nu \geq \frac{1}{2}\}$. Then for $\mathcal{P}'$ with $(\tilde{\mu}_1, \tilde{\mu_2}) \in \Xi$, with high probability, the worst case robust misclassification error $\beta^{\varepsilon,\infty}(\tilde{w}, \tilde{b})$ via the previous semi-supervised learning satisfies*

$$\sup_{\mathcal{P}':(\tilde{\mu}_1, \tilde{\mu_2}) \in \Xi} \beta^{\varepsilon,\infty}(\tilde{w}, \tilde{b}) \geq 49\%.$$

Shifted domain with unknown sparsity. In Theorem 3, we show how reasonably close shifted domain data helps in improving adversarial robustness. However, sometimes, the shifted domain is not so close to the original domain in terms of d_ν, but they still share some similarity. For instance, both domains can share some structural commonness. Here, we consider the case where the two distributions have common salient feature set; that is, the labeled and unlabeled domains share discriminant features, though the corresponding coefficients can be far apart. This setting is common in practice. For example, when one tries to classify images of different kinds of cats, the discriminant features include the eyes, ears, shapes etc. These discriminant features also applies when one aims to classify dogs, though the weights on these features might be very different.

Specifically, we consider the distributions of the labeled domain's positive and negative parts are $\mathcal{N}(\mu_1, \sigma^2 I_d)$ and $\mathcal{N}(\mu_2, \sigma^2 I_d)$, and the labels $y \sim Bern(1/2)$. The samples $(x_1, y_1), ..., (x_n, y_n)$ are drawn i.i.d. from this labeled domain. Suppose we have unlabeled samples $\tilde{x}_1, ..., \tilde{x}_{\tilde{n}}$, which are drawn from a uniform mixture of $\mathcal{N}(\tilde{\mu}, \sigma^2 I_d)$ and $\mathcal{N}(-\tilde{\mu}, \sigma^2 I_d)$ with $\tilde{\mu} \in \mathbb{R}^d$. Here, we assume the two domains share the support information, that is, $\text{supp}(\mu_1 - \mu_2) = \text{supp}(\tilde{\mu}_1 - \tilde{\mu}_2)$, though the distance between $\mu_1 - \mu_2$ and $\tilde{\mu}_1 - \tilde{\mu}_2$ is not necessarily small. For such a case, we propose to use the following algorithm to help improving adversarial robustness.

Algorithm of unknown sparsity: we first apply the high-dimensional EM algorithm (Cai et al., 2019) to estimate $\text{supp}(\tilde{\mu}_1 - \tilde{\mu}_2)$ from the unlabeled data. This high-dimensional EM algorithm is an extension of the traditional EM algorithm with the M-step being replaced by a regularized maximization. The detailed description can be found in the Appendix. After implementing the high-dimensional EM to estimate the support $\hat{S}$ from the unlabeled data, we then project the labeled data to this support $\hat{S}$ and therefore reduce the dimension. Finally, we apply the algorithm in the supervised setting on the labeled samples with reduced dimensionality to get the estimated $\hat{w}_{sparse} = 1/n \sum_{i=1}^{n} y_i [\vartheta(x_i)]_{\hat{S}}$ and $\hat{b}_{sparse} = 1/n \sum_{i=n+1}^{2n} [\vartheta(x_i)]_{\hat{S}}$. The following theorem provides theoretical guarantee for the robust classification error for $\hat{\theta}_{\text{sparse}}$ by this algorithm.

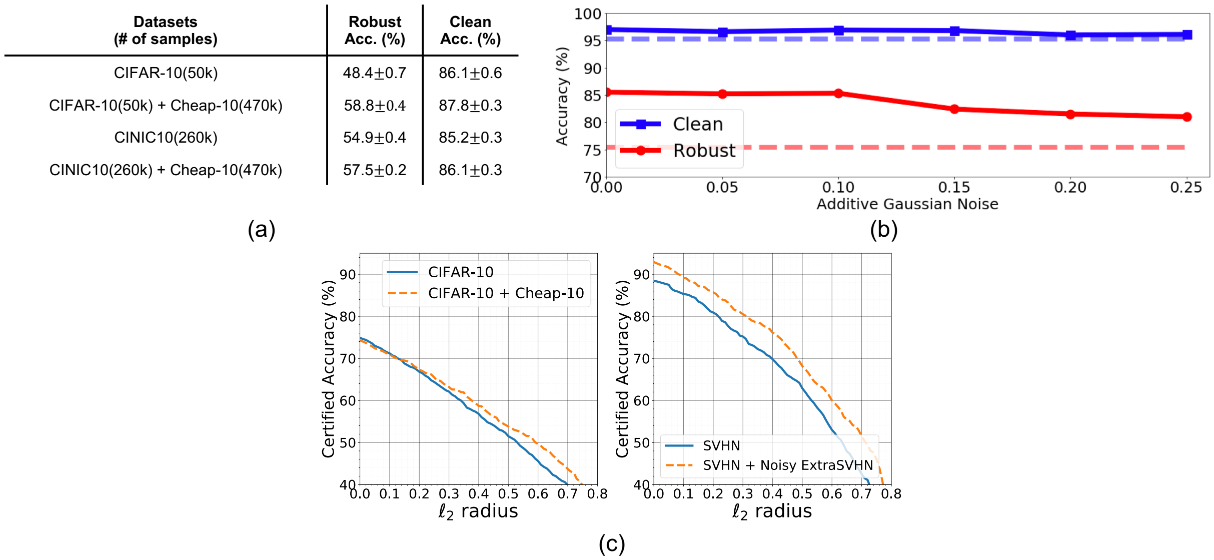
Theorem 6. *Under the conditions of Theorem 3.3 on parameters $\mu_1, \mu_2, \tilde{\mu}_1, \tilde{\mu}_2, \sigma$ and q. Suppose $|\,\text{supp}(\mu_1 - \mu_2)| = |\,\text{supp}(\tilde{\mu}_1 - \tilde{\mu}_2)| = m$, and $\min_{\tilde{\mu}_{1,j} - \tilde{\mu}_{2j} \neq 0} |\tilde{\mu}_{1j} - \tilde{\mu}_{2,j}| \geq \sigma \sqrt{2m \log d / \tilde{n}}$, where $\tilde{\mu}_{i,j}$ is the j-th entry in vector $\tilde{\mu}_i$. If $n \gtrsim \varepsilon^2 \log d \sqrt{m}$, we have*

$$\beta^{\varepsilon,\infty}(\hat{w}_{sparse}, \hat{b}_{sparse}) \leq 10^{-3} + O_{\mathbb{P}}(\frac{1}{n} + \frac{1}{d}).$$

Comparing to the result in Theorem 2, which shows that the sample complexity $O(\varepsilon^2 \sqrt{d}/(\log d))$ is necessary to achieve small robust error, the above theorem suggests that by utilizing the shared structural information from the unlabeled domain, one can reduce the sample complexity from $O(\sqrt{d})$ to $O(\sqrt{m})$. Corresponding simulation results are put in the Appendix.

4 Experiments

In this section, we provide empirical support for our theory and show that using unlabeled data from shifted domains can consistently improve robust accuracy for three widely-used benchmark datasets: CIFAR-10 (Krizhevsky et al., 2009), CINIC-10 (Darlow et al., 2018) and SVHN (Netzer et al., 2011).



Datasets (# of samples)	Robust Acc. (%)	Clean Acc. (%)
CIFAR-10(50k)	48.4±0.7	86.1±0.6
CIFAR-10(50k) + Cheap-10(470k)	58.8±0.4	87.8±0.3
CINIC10(260k)	54.9±0.4	85.2±0.3
CINIC10(260k) + Cheap-10(470k)	57.5±0.2	86.1±0.3

(a)

(b)

(c)

Figure 1: (a) ℓ_∞ **robustness.** Each row shows the test accuracy on clean and adversarially perturbed images ($\epsilon = 8/255$) when the original datasets are used versus when there is additional unlabeled source of data (Cheap-10). The robustness performance when we use the out-of-domain dataset is significantly better than the original training set, in agreement with our theory. (b) **SVHN dataset.** Dashed lines stand for the baseline performance of using only the labeled data. Each point on the x-axis shows a different model that is robustly trained using the original data and the unlabeled set of images with additive Gaussian noise with std (σ) equal to the axis value. The dashed lines indicate the clean and robust accuracy achieved using only the SVHN data. Adding noisy data improved robust accuracy. (c) ℓ_2 **certified robustness.** Each point in the plot shows the percentage of test images that are certified to be classified correctly at that ℓ_2 radius. Adding out-of-domain datasets consistently improves the certified raidii.

Datasets. The CIFAR-10 dataset has a training set of $50k$ images and test set of $10k$ images. The CINIC-10 dataset is a subset of ImageNet Russakovsky et al. (2015) of objects that are similar to CIFAR-10 objects; it has $260k$ images[2]. As our source of unlabeled data, we use the Cheap-10 dataset that we created to be a benchmark for using very cheap unlabeled out-of-domain data (avaiable at `https://tinyurl.com/mere5j0x`). We created Cheap-10 by searching keywords related to CIFAR-10 objects on the Bing image search engine[3]. A more detailed pipeline for creating Cheap-10 is described the Appendix. The important thing to note about Cheap-10 is that it is very fast to generate (hours) and can be quite noisy due to the lack of expert curation. Therefore it is a good illustration of the power of cheap, out-of-domain data.

A model trained on original CIFAR-10 data has a 68% accuracy on predicting Cheap-10 labels. The number is 75% for a model trained on CINIC-10 data. Both results mean that Cheap-10 is a related out of domain datasets with respect to both Cheap-10 and CIFAR-10. The SVHN dataset had $73k$ training and $26k$ test images. For SVHN task, the original dataset contains an extra $513k$ set of training images. We use this extra images as our source of unlabeled data and synthetically push the data out of domain by adding random Gaussian noise to it.

Methods. For each task, we first train a classification model on the original labeled data using cross-entropy loss function. We then use the trained model to assign pseudo-labels to unlabeled images. We next aggregate the two datasets to train a robust model using robust training. Following Carmon et al. (2019), we sample half of each batch from the original data and the other half from the additional pseudo-labeled data during training. We use the robustness regularization loss introduced in Zhang et al. (2019). For a maximum allowed ℓ_p-norm perturbation of size ϵ, we use the training loss function:

$$\mathscr{L}(x, y; \theta) = -\log p_\theta(y|x) \\ + \beta \max_{\hat{x} \in \mathbb{B}_p(x,\varepsilon)} D_{KL}(p_\theta(y|x)||p_\theta(\hat{y}|\hat{x}))$$

where the regularization parameter β balances the loss between accurate classification and stability within the ε ℓ_p-norm ball. We approximate the maximization in the second term as follows:

- Similar to Madry et al. (2017), for ℓ_∞ pertur-

[2] After removing CIFAR-10 test images that are in CINIC-10

[3] `https://www.bing.com/images/`

bations, we focus on empirical robustness of the models and use an inner loop of projected gradient descent for the maximization.

- Following Carmon et al. (2019), for ℓ_2 perturbations, we focus on certified robustness and use the idea of stability training Zheng et al. (2016); Li et al. (2018). We replace the maximization with large additive noise draws: $\mathbb{E}_{\hat{x}\sim\mathcal{N}(,\sigma)} D_{KL}(p_\theta(y|x)||p_\theta(\hat{y}|\hat{x}))$. The idea is to have a model that is robust to large random perturbations. Using Cohen et al.'s method Cohen et al. (2019), in test time, we can find a safe radius of certified robust prediction for each sample.

As our first experiment, we focused on empirical robustness against ℓ_∞ perturbations. We used a Wide ResNet 28-10 architecture Zagoruyko & Komodakis (2016). Following the literature, for ℓ_∞ perturbation, we set $\varepsilon = 8/255$. Results for empirical robustness against l_∞ perturbations are shown in Fig. 1(a). The clean accuracy is the model's performance on non-perturbed images. The robust accuracy is the model's performance on adversarially perturbed images. We use the strongest known adversarial attack methods, iterative projected gradient descent (PGD), to create the l_∞ perturbations. We fine-tuned the attack hyperparameters and found that using 40 iterations results in the smallest robust accuracy. More details are in the Appendix. We find that using Cheap-10 consistently improves the robust accuracy. Note that, for CIFAR-10, while Cheap 10 was created in a few hours, it produced significantly better robust accuracy ($58.8 \pm 0.4\%$) compared to using only the original data, and similarly for CINIC-10. This strategy of using cheap noisy data to improve robustness compares favorably to state-of-the-art existing defenses applied to CIFAR-10: TRADES method (55.4%, Zhang et al. (2019)) and Adversarial Pretraining (57.4%, Hendrycks et al. (2019)).

As our second experiment, we use the SVHN dataset and a Wide ResNet 16-8 as our model architecture. SVHN has a training set of $73k$ real digit images and an extended set of $531k$ images that come with the dataset. The extra set is a synthetically generated set of digits that to mimic the original dataset closely. We use the extended set as our source of unlabeled data. The model we trained (normal training) on the original training set has an accuracy of 96.8% on SVHN test set and an accuracy of 98.4% on the extra training set; this means that the extra data is very similar to the original SVHN dataset. To push the unlabeled data out-of-domain, we add four different levels of additive Gaussian noise to the images. We focus on ℓ_∞ perturbations with $\epsilon = 4/255$. Fig. 1bc) describes the results. The dashed lines are the baselines for not having any additional unlabeled data. They show clean and robust accuracies when only the original training set is used. It can be observed that adding the unlabeled data robustly improves robust accuracy. As the unlabeled data distribution gets more distant from SVHN data, the improvement achieved from adding the extra set of unlabeled images becomes smaller.

As our final experiment, we focus on certified robustness. For stability training, we used $\sigma = 0.25$. Fig. 1(c) shows the percentage of images that are certified to be classified correctly at each ℓ_2 radius. First, use the CIFAR-10 dataset as the labeled data and the Cheap-10 data set as the unlabeled data. Secondly, we use the original SVHN tranining set as the labeled data and the extra set of SVHN images with additive Gaussian noise ($std = 0.15$) as the unlabeled data source. It demonstrates that adding cheap out-of-domain data consistently improves certified robustness compared to only using the original training set. More implementation details are described in the Appendix.

5 Further Discussions

Incorporating cheap unlabeled data is a popular way to improve the prediction performance in machine learning. In this work, we show that this substantially improves adversarial robustness, even when the unlabeled data come from a different domain.

We prove our theoretical results for Gaussian generative models, which are very flexible (e.g. it includes common deep generative models such as GANs and VAEs). Moreover our theory is supported by our experiments using a new dataset Cheap-10. This suggests that the vast amount of noisy out-of-domain data is a relatively untapped resource that could substantially improve the reliability of many machine learning tasks.

In this work, we showed that, in general, the adversarial robustness of a semi-supervised algorithm will be improved when the out-of-domain distribution is similar to the labeled data, and the robustness will be hurt if the out-of-domain distribution is too different. One possible extension of our work is to use the aggregation idea in Li et al. (2020) to deal with the challenging setting where the similarity between the out-of-domain distribution and labeled data distribution is unknown a priori. Such an extension will make the results applied to more general settings. Further, our theoretical results and analysis also lay the foundation of studying the adversarial robustness of other tasks, such as multi-class classification and linear/kernel regression in the semi-supervised setting when the unlabeled data come from a different domain.

The focus of this work is on the effects of out-of-domain unlabeled data, and we use the popular and simple pseudo-labeling method to capture the key insights. An interesting direction of future work is to investigate how to improve robustness with other semi-supervised learning methods. For example, one could apply several iteration of pseudo-labeling to improve label quality.

6 Acknowledgements

The research of Linjun Zhang is supported by NSF DMS-2015378. The research of James Zou is supported by NSF CCF 1763191, NSF CAREER 1942926 and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative. This work is also in part supported by NSF award 1763665.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

T Tony Cai, Jing Ma, Linjun Zhang, et al. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017b.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11190–11201, 2019.

Ping-Yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*, 2020.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. Computational limitations in robust classification and win-win results. *arXiv preprint arXiv:1902.01086*, 2019.

Zhun Deng, Cynthia Dwork, Jialiang Wang, and Linjun Zhang. Interpreting robust optimization via adversarial influence functions. In *International Conference on Machine Learning*, pp. 2464–2473. PMLR, 2020a.

Zhun Deng, Hangfeng He, Jiaoyang Huang, and Weijie Su. Towards understanding the dynamics of the first-order adversaries. In *International Conference on Machine Learning*, pp. 2484–2493. PMLR, 2020b.

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

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019.

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

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. 2018.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*, 2020.

Chizhou Liu, Yunzhen Feng, Ranran Wang, and Bin Dong. Enhancing certified robustness of smoothed classifiers via weighted model ensembling. *arXiv preprint arXiv:2005.09363*, 2020.

Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647, 2005.

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

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pp. 5542–5552, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

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(3):211–252, 2015.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pp. 1163–1171, 2016.

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

Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 39–49, 2017.

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

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

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.