CAN NON-LIPSCHITZ NETWORKS BE ROBUST? THE POWER OF ABSTENTION AND DATA-DRIVEN DECISION MAKING FOR ROBUST NON-LIPSCHITZ NETWORKS

Maria-Florina Balcan
Carnegie Mellon University
ninamf@cs.cmu.edu

Avrim Blum
TTIC
avrim@ttic.edu

Dravyansh Sharma
Carnegie Mellon University
dravyans@cs.cmu.edu

Hongyang Zhang
University of Waterloo
hongyang.zhang@uwaterloo.ca

ABSTRACT

Deep networks have been found to be highly susceptible to adversarial attacks. One fundamental challenge is that it is typically possible for small input perturbations to produce large movements in the final-layer feature space of these networks. In this work, we define an attack model that abstracts this challenge, to help understand its intrinsic properties. In our model, the adversary may move data an arbitrary distance in feature space but only in random low-dimensional subspaces. We prove that such adversaries can be quite powerful: defeating any classifier that must output a class prediction on any input it is given. However, by giving the algorithm the ability to abstain, we show that such an adversary can be overcome when classes are reasonably well-separated in feature space and the dimension of the feature space is high, by an algorithm that examines distances of test points to training data in feature space. We further show how data-driven methods can be used to set algorithm parameters to optimize over the accuracy vs. abstention trade-off with strong theoretical guarantees. Our theory can also be viewed as providing new robustness guarantees for nearest-neighbor style algorithms, and has direct applications to the technique of contrastive learning, where we empirically demonstrate the ability of such algorithms to obtain high robust accuracy with only small amounts of abstention. Overall, our results provide insight into the intrinsic vulnerabilities of non-Lipschitz networks and the ways these may be addressed.

1 INTRODUCTION

A substantial body of work has shown that deep networks can be highly susceptible to adversarial attacks, in which minor changes to the input lead to incorrect, even bizarre classifications (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2016; Madry et al., 2018; Su et al., 2019; Brendel et al., 2018). Much of this work has considered bounded ℓ_p-norm attacks, though many other forms of attack are considered as well (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019). What these results have in common is that changes that either are imperceptible or should be irrelevant to the classification task can lead to drastically different network behaviors.

One key reason for this vulnerability to attacks is the non-Lipschitzness of typical neural networks: small but adversarial movements in the input space can produce large perturbations in the feature space. This ability of an adversary to produce large movements in feature space appears to be at the heart of many of the successful attacks to date. If we assume that non-Lipschitzness is important for good performance on natural data, then it is crucial to understand to what extent this property makes a network instrinsically susceptible to attacks.

In this work, we propose and analyze an abstract attack model designed to focus on this question of the intrinsic vulnerability of non-Lipschitz networks, and what might help to make such networks robust. In particular, suppose an adversary, by making an imperceptible change to an input x, can cause its representation $F(x)$ in feature space (the final layer of the network) to move by an arbitrary amount: will such an adversary always win? Clearly if the adversary can modify $F(x)$ by an arbitrary amount *in an arbitrary direction*, then yes, because it can then move $F(x)$ into the classification

region of any other class it wishes. But what if the adversary can modify $F(x)$ by an arbitrary amount but only in a *random* direction or within a random low-dimensional subspace? In this case, we show an interesting dichotomy: if the classifier must output a classification on any input it is given, then indeed the adversary will still win, no matter how well-separated the natural data points from different classes are in feature space and no matter what decision surface the classifier uses. However, if we provide the classifier the ability to abstain, then we show it can defeat such an adversary.

Our contributions. Conceptually, we introduce a new *random feature subspace* threat model to abstract the effect of non-Lipschitzness in deep networks. Technically, we show the power of abstention and data-driven decision-making in this setting, proving that classifiers with the ability to abstain are provably more powerful than those that cannot in this model, and giving formal guarantees for parameterized nearest-neighbor style algorithms. We use data-driven hyperparameter learning to set the algorithm parameters to minimize robust error while keeping abstention on natural data low. Experimentally, we show that our algorithms perform well in this model on representations learned by supervised and self-supervised contrastive learning.

2 POWER OF ABSTENTION

In principle, an adversarial example for a given labeled data point $(\mathbf{x}, y)$ is a small perturbation $\mathbf{x}'$ of $\mathbf{x}$ with the same true label y but that causes the classifier to make a mistake. One of the most popular models for adversarial examples is that of norm-bounded perturbations in the input space. Despite a large literature devoted to defending against such adversaries by improving the Lipschitzness of neural networks as functions mapping from input space to feature space (Zhang et al., 2019; Yang et al., 2020), it is typically not true that small perturbations in the input space necessarily imply small modifications in the feature space. Motivated by this fact, in this paper we study a threat model where an adversary can modify the data by a large amount in the feature space. Note that because this large modification in feature space is assumed to come from a small perturbation in input space, we always assume that the *true correct label y is the same for the modified point and the original point*.

Our threat model. Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{n_1}$ be a test input, embedded into feature space $\mathcal{F} \subseteq \mathbb{R}^{n_2}$ using a deep neural network F. The adversary may corrupt $F(\mathbf{x})$ such that the modified feature vector is any point in a random n_3-dimensional affine subspace denoted by $\mathcal{S} + \{F(\mathbf{x})\}$. For example, if $n_3 = 2$ then $\mathcal{S} + \{F(\mathbf{x})\}$ is a random 2-dimensional plane through $F(\mathbf{x})$, and the adversary may select an arbitrary point in that plane. Conceptually, we are viewing the network as "squashing" the adversarial ball in input space into a random infinitely thin and infinitely wide n_3-dimensional pancake in feature space. The adversary is given access to everything including the training data, F, $\mathbf{x}$, $\mathcal{S}$ and the true label of $\mathbf{x}$. We will use *adversary* and *adversarial example* throughout to refer to this threat model.

We first present a hardness result showing that no matter how nicely data is distributed in feature space, any classifier that is not allowed to abstain will fail even if the adversary can perturb points in a single random direction ($n_3 = 1$).

Theorem 2.1. *For any classifier that partitions $\mathbb{R}^{n_2}$ into two or more classes, any data distribution $\mathcal{D}$, any $\delta > 0$ and any feature embedding F, there must exist at least one class y^*, such that for at least a $1 - \delta$ probability mass of examples $\mathbf{x}$ from class y^* (i.e., $\mathbf{x}$ is drawn from $\mathcal{D}_{\mathcal{X}|y^*}$), for a random unit-length vector $\mathbf{v}$, with probability at least $1/2 - \delta$ for some $\Delta_0 > 0$, $F(\mathbf{x}) + \Delta_0 \mathbf{v}$ is not labeled y^* by the classifier. In other words, there must be at least one class y^* such that for at least $1 - \delta$ probability mass of points $\mathbf{x}$ of class y^*, the adversary wins with probability at least $1/2 - \delta$.*

Theorem 2.1 gives a hardness result for robust classification without abstention. We will now give positive results for a nearest-neighbor style classifier (Algorithm 1) that has the power to abstain.

Algorithm 1 ROBUSTCLASSIFIER(τ, σ)

1: **Input:** A test feature $F(\mathbf{x})$ (potentially an adversarial example), a set of training features $F(\mathbf{x}_i)$ and their labels y_i, $i \in [m]$, a threshold parameter τ, a separation parameter σ.
2: **Preprocessing:** Delete training examples $F(\mathbf{x}_i)$ if $\min_{j \in [m], y_i \neq y_j} \mathsf{dist}(F(\mathbf{x}_i), F(\mathbf{x}_j)) < \sigma$
3: **Output:** A predicted label of $F(\mathbf{x})$, or "don't know".
4: **if** $\min_{i \in [m]} \mathsf{dist}(F(\mathbf{x}), F(\mathbf{x}_i)) < \tau$ **then**
5: **return** $y_{\mathrm{argmin}_{i \in [m]} \mathsf{dist}(F(\mathbf{x}), F(\mathbf{x}_i))}$
6: **return** "don't know"

Denote by $\mathcal{E}^{\mathbf{x}}_{\text{adv}}(f) := \mathbb{E}_{S\sim\mathcal{S}}\mathbf{1}\{\exists \mathbf{e} \in S + F(\mathbf{x}) \text{ s.t. } f(\mathbf{e}) \neq \mathbf{y} \text{ and } f(\mathbf{e}) \text{ does not abstain}\}$ the robust error of a given classifier f for classifying instance $\mathbf{x}$. The following theorem states that so long as the threshold τ in Algorithm 1 is sufficiently small compared to the distance r between classes, and the dimension n_2 of the ambient feature space is sufficiently large compared to the dimension n_3 of the adversarial subspace S, the algorithm will have low robust error.

Theorem 2.2. *Let $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ be a test instance, m be the number of training examples and r be the shortest distance between $F(\mathbf{x})$ and $F(\mathbf{x}_i)$ where $\mathbf{x}_i$ is a training point from a different class. Suppose $\tau = o\left(r\sqrt{1-\frac{n_3}{n_2}}\right)$. The robust error of Algorithm 1, $\mathcal{E}^{\mathbf{x}}_{\text{adv}}(\text{ROBUSTCLASSIFIER}(\tau, 0))$, is at most $m\left(\frac{c\tau}{r\sqrt{1-\frac{n_3}{n_2}}}\right)^{n_2-n_3} + mc_0^{n_2-n_3}$, where $c > 0$ and $0 < c_0 < 1$ are absolute constants. For the case $n_3 = 1$, the robust error is at most $m\left(\frac{\tau}{r}\right)^{n_2-1}$.*

Theorem 2.2 states that the robust error of Algorithm 1 is small if the number of labeled examples m is sub-exponential in $n_2 - n_3$. We also extend Theorem 2.2 to a more general class of adversaries, which can have any bounded distribution over the space of linear subspaces of a fixed dimension n_3 and the adversary can perturb a test feature vector arbitrarily in the sampled adversarial subspace.

Theorem 2.3. *Consider the setting of Theorem 2.2, with an adversary having a κ-bounded[1] distribution over the space of linear subspaces of a fixed dimension n_3 for perturbing the test point. If $\mathbf{E}(\tau, r)$ denotes the bound on error rate in Theorem 2.2 for $\text{ROBUSTCLASSIFIER}(\tau, 0)$ in Algorithm 1, then the error bound of the same algorithm against the κ-bounded adversary is $\mathcal{O}(\kappa\mathbf{E}(\tau, r))$.*

We relax the well-separateness assumption in Theorems 2.2 and 2.3 by using a separation parameter $\sigma > 0$ in Algorithm 1. We further show that we can control the frequency of outputting "don't know", when the data are nicely distributed according to the following assumption.

Assumption 1. *We assume that at least $1 - \delta$ fraction of mass of the marginal distribution $\mathcal{D}_{F(\mathcal{X})|y}$ over $\mathbb{R}^{n_2}$ can be covered by N balls $\mathbb{B}_1$, $\mathbb{B}_2$, ... $\mathbb{B}_N$ of radius $\tau/2$ and of mass $\Pr_{\mathcal{D}_{F(\mathcal{X})}}[\mathbb{B}_k] \geq \frac{C_0}{m}\left(n_2 \log m + \log\frac{4N}{\beta}\right)$, where $C_0 > 0$ is an absolute constant and $\delta, \beta \in (0, 1)$.*

Intuitively, it says that for every label class one can cover most of the distribution of the class with (potentially overlapping) balls of a fixed radius, each having a small lower bound on the density contained. This holds for well-clustered datasets (as is typical for feature data) for a sufficiently large radius. Our analysis leads to the following guarantee on the abstention rate.

Theorem 2.4. *Suppose that $F(\mathbf{x}_1), ..., F(\mathbf{x}_m)$ are m training instances i.i.d. sampled from marginal distribution $\mathcal{D}_{F(\mathcal{X})}$. Under Assumption 1, with probability at least $1 - \beta/4$ over the sampling, we have $\Pr(\cup_{i=1}^{m}\mathbb{B}(F(\mathbf{x}_i), \tau)) \geq 1 - \delta$.*

Theorem 2.4 implies that when $\Pr[\mathbb{B}_k] \geq \frac{\beta}{N}$ and $m = \Omega(\frac{n_2 N}{\beta}\log\frac{n_2 N}{\beta})$, with probability at least $1 - \beta/4$ over the sampling, we have $\Pr(\cup_{i=1}^{m}\mathbb{B}(F(\mathbf{x}_i), \tau)) \geq 1 - \delta$. Therefore, with high probability, the algorithm will output "don't know" only for an δ fraction of natural data.

3 LEARNING DATA-SPECIFIC OPTIMAL THRESHOLDS

Given an embedding function F and a classifier f_τ which outputs either a predicted class if the nearest neighbor is within distance τ of a test point or abstains from predicting, we want to evaluate the performance of f_τ on a test set $\mathcal{T}$ against an adversary which can perturb a test feature vector in a random subspace $S \sim \mathcal{S}$. To this end, we define $\mathcal{E}_{\text{adv}}(\tau) := \mathbb{E}_{S\sim\mathcal{S}}\frac{1}{|\mathcal{T}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}}\mathbf{1}\{\exists \mathbf{e} \in S + F(\mathbf{x}) \subseteq \mathbb{R}^{n_2} \text{ s.t. } f(\mathbf{e}) \neq \mathbf{y} \text{ and } f_\tau(\mathbf{e}) \text{ does not abstain}\}$ as the robust error on the test set $\mathcal{T}$, and $\mathcal{D}_{\text{nat}}(\tau) := \frac{1}{|\mathcal{T}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}}\mathbf{1}\{f_\tau(F(\mathbf{x})) \text{ abstains}\}$ as the abstention rate on the natural data. $\mathcal{E}_{\text{adv}}(\tau)$ and $\mathcal{D}_{\text{nat}}(\tau)$ are monotonic in τ. We can capture the trade-off between abstention rate and accuracy in a single objective $g(\tau) := \mathcal{E}_{\text{adv}}(\tau) + c\mathcal{D}_{\text{nat}}(\tau)$, where c is a positive constant and denotes the *cost of abstention* (adversarial version of Chow's objective (Chow, 1970)). We can optimize $g(\tau)$ in a data-driven fashion and obtain theoretical guarantee on the convergence to a global optimum. In Theorem 3.1, we show no regret can be achieved for online learning of the threshold τ using test batches of size b. Using online-to-batch conversion, our results imply a uniform convergence bound for objective $g(\tau)$ in the supervised setting.

[1] A distribution is κ-bounded if the corresponding probability density $f(x)$ satisfies, $\sup_x f(x) \leq \kappa$.

Table 1: Natural error $\mathcal{E}_{\text{nat}}$ and robust error $\mathcal{E}_{\text{adv}}$ on the CIFAR-10 dataset (Szegedy et al., 2015) when $n_3 = 1$ and the 512-dimensional representations are learned by contrastive learning, where $\mathcal{D}_{\text{nat}}$ represents the fraction of each algorithm's output of "don't know" on the natural data.

	Contrastive	Linear Protocol		Ours ($\tau = 3.0$)			Ours ($\tau = 2.0$)		
		$\mathcal{E}_{\text{nat}}$	$\mathcal{E}_{\text{adv}}$	$\mathcal{E}_{\text{nat}}$	$\mathcal{E}_{\text{adv}}$	$\mathcal{D}_{\text{nat}}$	$\mathcal{E}_{\text{nat}}$	$\mathcal{E}_{\text{adv}}$	$\mathcal{D}_{\text{nat}}$
$(\sigma = 0)$	Self-supervised	8.9%	100.0%	15.4%	40.7%	2.2%	14.3%	26.2%	28.7%
	Supervised	5.6%	100.0%	5.7%	60.5%	0.0%	5.7%	33.4%	0.0%
$(\sigma = 0.9\tau)$	Self-supervised	8.9%	100.0%	7.2%	9.4%	12.9%	10.0%	17.7%	29.9%
	Supervised	5.6%	100.0%	6.2%	18.9%	0.0%	5.6%	22.0%	0.1%
$(\sigma = \tau)$	Self-supervised	8.9%	100.0%	1.1%	1.2%	33.4%	2.1%	3.1%	49.9%
	Supervised	5.6%	100.0%	1.9%	2.8%	10.6%	4.1%	4.8%	3.3%

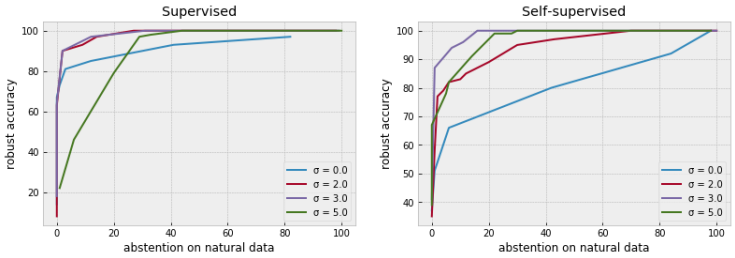

Figure 1: Adversarial accuracy (i.e., rate of adversary failure) vs. abstention rate as threshold τ varies for $n_3 = 1$ and different outlier removal thresholds σ.

Theorem 3.1. *Assume τ is $o\left(\min\{m^{-1/n_2}, r\}\right)$, and the data distribution is continuous, κ-bounded, positive and has bounded partial derivatives. If τ is set using a continuous version of the multiplicative updates algorithm, then with probability at least $1-\delta$, the expected regret for optimizing $g(\tau)$ in T rounds is bounded by $O\left(\sqrt{n_2 T \log\left(\frac{\kappa R T m b}{\delta r^{n_2 - n_3}}\right)}\right)$, where R is a bound on the largest distance between any two training points, b is the batch size, and r is the smallest distance between points of different labels.*

4 Experiments on Contrastive Learning

We verify the robustness of Algorithm 1 when the representations are learned by contrastive learning. Given a embedding function F and a classifier f which outputs either a predicted class or abstains from predicting, recall that we define the natural and robust errors, respectively, as $\mathcal{E}_{\text{nat}}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\mathbf{1}\{f(F(\mathbf{x})) \neq \mathbf{y}$ and $f(F(\mathbf{x}))$ does not abstain$\}$, and $\mathcal{E}_{\text{adv}}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D},S\sim\mathcal{S}}\mathbf{1}\{\exists \mathbf{e} \in S + F(\mathbf{x}) \subseteq \mathbb{R}^{n_2}$ s.t. $f(\mathbf{e}) \neq \mathbf{y}$ and $f(\mathbf{e})$ does not abstain$\}$, where $S \sim \mathcal{S}$ is a random adversarial subspace of $\mathbb{R}^{n_2}$ with dimension n_3. $\mathcal{D}_{\text{nat}}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\mathbf{1}\{f(F(\mathbf{x}))$ abstains$\}$ is the abstention rate on the natural examples. Note that the robust error is always at least as large as the natural error.

Our self-supervised and supervised contrastive learning setups follow (Chen et al., 2020) and (Khosla et al., 2020) respectively. In both the setups, we compare the robustness of the linear protocol with that of our defense protocol in Algorithm 1 under exact computation of adversarial examples using a convex optimization program in n_3 dimensions and m constraints.

Experimental results. (Table 1.) Compared with the standard linear protocol, our algorithms have much lower robust error. Note that even if abstention is added based on distance from the linear boundary, sufficiently large perturbations will ensure the adversary can always succeed. For an approximate adversary which can be efficiently implemented for large n_3.

The threshold parameter τ captures the trade-off between the robust accuracy $\mathcal{A}_{\text{adv}} := 1 - \mathcal{E}_{\text{adv}}$ and the abstention rate $\mathcal{D}_{\text{nat}}$ on the natural data. We report both metrics for different values of τ for supervised and self-supervised constrastive learning. The supervised setting enjoys higher adversarial accuracy and a smaller abstention rate for fixed τ's due to the use of extra label information. We plot $\mathcal{A}_{\text{adv}}$ against $\mathcal{D}_{\text{nat}}$ for Algorithm 1 as hyperparameters vary (Figure 1). For small τ, both accuracy and abstention rate approach 1.0. As the threshold increases, the abstention rate decreases rapidly and our algorithm enjoys good accuracy even with small abstention rates. For $\tau \to \infty$ (i.e. the nearest

neighbor search), both $\mathcal{D}_{\text{nat}}$ and $\mathcal{A}_{\text{adv}}$ are nearly 0%. Increasing σ (for small σ) gives us higher robust accuracy for the same abstention rate. Too large σ also leads to degraded performance.

ACKNOWLEDGMENTS

This work was supported in part by an NSERC Discovery Grant, the National Science Foundation under grants CCF-1815011, CCF-1535967, IIS-1901403, CCF-1910321, SES-1919453, the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003, an AWS Machine Learning Research Award, a Microsoft Research Faculty Fellowship, and a Bloomberg Research Grant. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

REFERENCES

Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson. ADef: an iterative algorithm to construct adversarial deformations. In *International Conference on Learning Representations*, 2019.

Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018.

Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

CK Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. Adversarial robustness through local Lipschitzness. In *Advances in neural information processing systems*, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.