CORRELATING ADVERSARIAL ROBUSTNESS WITH THE ENERGY OF UNTRAINED CLASSIFIERS

LUCA MAUTINO[1], **MUJTABA HUSSAIN MIRZA**[1], **IACOPO MASI**[1]

[1]OmnAI Lab, CS Department, Sapienza University of Rome, Italy
E-MAIL: mautino.1997410@studenti.uniroma1.it mirza,masi@di.uniroma1.it

Abstract:

Neural networks are known to be vulnerable to imperceptible perturbations to their inputs, causing unwarranted behavior in the model's predictions and attracting the scientific community's interest with the aim of designing architectures and training methods that are robust to such attacks. By reinterpreting a discriminant classifier as an energy-based model, the following work further studies the connection between robustness, architectural configuration, and energy landscape. Energy is leveraged as a tool to study a network's intrinsic robustness by analyzing models with randomized weights, suggesting a connection between a classifier's untrained energy landscape and its final robustness after adversarial training. Notably, the energy profile of an untrained network appears to align with its final robustness metric. This correspondence holds true across three distinct model scales, each encompassing a diverse range of architectural configurations, suggesting a universal link independent of specific structural choices.

Keywords:

Adversarial Robustness; Energy Based Models; Discriminant Classifiers; Architecture Search; Implicit Bias

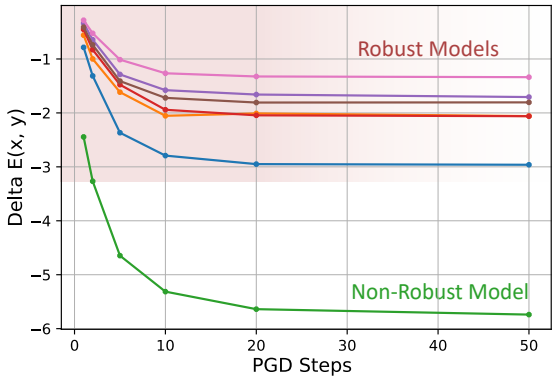


FIGURE 1. We compare average $\Delta E(x, y)$ of robust networks trained with adversarial training under the shaded area with color with a non-robust standard trained model of color .

1. Introduction

The problem of designing and training a machine learning model that is resistant to adversarial perturbations intended to cause errors in its predictions remains, as of today, an open challenge. Such perturbations can be easily generated in both black- and white-box environments, often leading to significant errors in models that would otherwise perform very accurately in an unperturbed context. A plethora of learning paradigms has been developed to increase what is referred to as the *robustness* of a model, but all share the core idea of integrating adversarial attacks into the learning phase, defined as *Adversarial Training* [4]. Later work, however, also highlighted [14] [8] the critical role model architecture can play in the final robustness of a model, exploring a new approach to the issue and hinting at a correlation between a model's intrinsic properties determined by its architecture and its final performances, regardless of learning methodologies. The connection between the two, however, remains hard to investigate as isolating the effects of topology from those of training is not trivial. Following the foundational work in [10], interpretation of ML models through the lens of an energy measure framework allows for a more fine-grained interpretation of model behavior and, in particular, renders feasible and interpretable the study of models with randomized weights, which constitutes an ideal environment to study relations between architecture and intrinsic robustness, given the absence of training.

This paper combines these insights by analyzing model architectures commonly used in image classification through adversarial attacks and documents an apparent connection between sensitivity to perturbations of inputs before any kind

of training and final robustness. Specifically, it studies three model sizes and finds that the more robust model configurations also better resist perturbations when untrained, where resistance to attacks is measured by the magnitude of the difference of marginal energies between perturbed and normal samples $|\Delta E(x, y)|$ which also serves as an indicator of local energy *smoothness* as already seen in [21, 12]. Such metric, defined as the difference in the logits of a sample's normal and perturbed versions, then directly correlates with overall robustness of a model.

If further findings were to substantiate our hypothesis, sensitivity to attacks when untrained could become both an efficient proxy for final robustness, circumventing the intensive task of adversarial training, and an analysis metric for inspecting an architecture's implicit bias.

2. Related Work

2.1. Adversarial Attacks and Defenses

Szegedy et al. [16] first revealed that neural networks are vulnerable to small, carefully crafted perturbations, and Goodfellow et al. [4] provided a linear explanation and introduced the Fast Gradient Sign Method (FGSM), establishing the foundation of gradient-based attacks. This discovery sparked extensive research on both stronger attacks and more effective defenses. Among the latter, adversarial training (AT) [11] emerged as the dominant approach. Numerous refinements followed; in particular, TRADES [20] proposed a theoretically motivated objective function that leverages the Kullback-Leibler (KL) divergence to explicitly balance natural and robust accuracy.

However the role of architectural design in AT has received relatively little attention. Peng et al. [14] conducted the most in depth investigation of the correlation between robustness and architectural components to date, confirming previous hypothesis from [8] and [9] regarding the size of the last stage, while [15] showed that non-parametric activation functions, such as `SiLU` also provide a consistent increase in robustness compared to other choices of activation functions within a network.

2.2 Investigating architectural biases

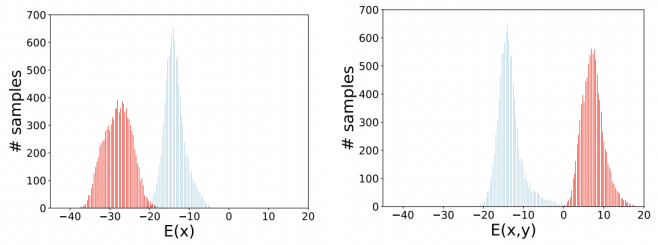
Another line of research highlights the role of architectural inductive biases, independently of learned weights. Ulyanov et al. [17] showed that untrained networks can achieve state-of-the-art results in image restoration, revealing an implicit *data impedance*: convergence on natural images is much faster than on random noise in initialized models, suggesting that network structure, being the only prior present before training, inherently favors data from natural distributions. Similarly, the study of Randomized Neural Networks [3] reveals that architectures where hidden connections remain fixed after initialization can reach representational capabilities on par with fully trained architectures. Lastly, the phenomenon of pruning, where a majority of a model's parameters can be deactivated after training with minimal accuracy loss, further supports the idea that a network's strongest component is its design. While pruning typically occurs post-training, the Lottery Ticket Hypothesis [2] suggests that dense networks contain sparse "winning ticket" sub-networks that are uniquely capable of effective training due to their specific connectivity and fortuitous initialization. Essentially, increasing parameter counts may simply increase the probability of capturing these effective structural priors. To conclude, there is still a great need to further explore the connection between architecture and model performance, with many open questions in deep learning potentially addressed by further findings in this line of inquiry.

2.3 Energy-Based Models and Adversarial Robustness

Joint Energy-based Models (JEM) [5] extend the idea of Energy-Based models [10] to classifiers, providing generative capabilities along with improved robustness and calibration. Several studies [19, 18, 21, 12, 13] have explored this energy-based perspective, showing that it can enhance robustness alongside generative capabilities. Furthermore, [21, 12] demonstrate that robust classifiers tend to exhibit smoother energy landscapes, and that TRADES [20] explicitly encourages this smoothness, which explains its superior robustness compared to standard adversarial training [12]. Building on these insights, we study how architectural design affects the energy landscape of classifiers, with certain configurations inherently producing smoother landscapes that may support higher robustness.

3. Background

Energy Based Models : Energy based models (EBM) define a probability distribution via a scalar function $E_{f_{\boldsymbol{\theta}}}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n \to \mathbb{R}$ called *energy function* . Learning within the energy framework consists in finding an energy function which associates to observed variables configurations lower energy values than to rest of possible configurations. In the context of a discriminant classifier, the logits can be interpreted as the negative



(a) Distribution of marginal energies of normal and adversarial samples

(b) Distribution of joint energies of normal and adversarial samples

FIGURE 2. PGD's impact on joint and marginal energies. ■ indicates adversarial data, while ■ indicates natural data. [12]

of an energy function, since the model aims to assign higher scores to the correct labels. Marginalizing over all labels lets us also define an energy measure for the individual samples without a label:

$$E_{f_{\boldsymbol{\theta}}}(\mathbf{x}, y) = -f_{\boldsymbol{\theta}}(\mathbf{x})[y] \quad E_{f_{\boldsymbol{\theta}}}(\mathbf{x}) = -\log \sum_{k=1}^{K} \exp\left(f_{\boldsymbol{\theta}}(\mathbf{x})[k]\right). \tag{1}$$

Energy and adversarial attacks Using the energy functions, following the work in [12], one can see how adversarial attacks such as projected gradient descent (PGD) [11], create adversaries that increase the joint energy $E(x, y)$ while decreasing the marginal one $E(x)$, as shown in eq. 2. While these adversarial samples cause misclassifications, they are in some sense even more probable than the original samples. Thus being able to significantly perturb a sample's marginal energy directly implies higher chance of changing its predicted label, and smoothness of energy around a point implies then overall resistance to attacks. Rewriting the commonly used cross-entropy loss using these energy functions and taking the gradient we find:

$$\nabla_x \mathcal{L}_{\text{CE}}\big(f_\theta(\mathbf{x}), y_i\big) = \nabla_x E(\mathbf{x}, y_i) - \nabla_x E(\mathbf{x}) \tag{2}$$

Kaiming's initialization We mainly use *Kaiming*'s method [7] to initialize an architecture's weights. Weights for the convolutional layer's filters are sampled from a gaussian distribution with 0 mean and a standard deviation formulated in order to maintain the variance of the final signal bounded and similar to the one of the original input. In particular the variance of a given output layer l, in this case the last layer L is given by

$$\text{Var}[y_L] = \text{Var}[y_1]\Big(\prod_{l=2}^{L} \frac{1}{2} n_l \text{Var}[w_l]\Big) \tag{3}$$

where n_l is the number of connections to a given neuron and the convolution operations are expressed as matrix multiplications. A sufficient condition to maintain the outputs' variance within a proper range is that w_l comes from a gaussian distribution with 0 mean and a standard deviation of $\sqrt{2/n_l}$.

Architecture configurations : Finally we describe the ways models' architectures are modified to study a relationship with final robustness: at the time of this article the most commonly used model for robust image classification is the *Wide Residual Network* (WRN) and its architecture is mainly configurable by two parameters called *depth* and *width*. We focus on the former which describes the number of basic blocks contained within each of the 3 stages of a common WRN configuration. Each block consists of two residual convolutional layers with batch normalization and activation functions and a final fully connected layer. From the grid search on architectures conducted in [8] we already posses the final robustness percentages of many possible depth configurations, which we later compare with their untrained versions.

4. Methodology and Evaluation

4.1 Method

To isolate the effects of architectural design on adversarial robustness, we build on the insights from the previous sections and leverage PGD attacks to systematically analyze the energy landscape of already robust and of *untrained models*. Specifically, we construct various WRN models with different architectural configurations while keeping the number of parameters *constant*. After initialization with the Kaiming's method, we feed both natural inputs and their corresponding adversarial counterparts generated by PGD into the network and compute the difference in marginal energies between them. This difference provides a direct measure of the local smoothness of the energy landscape around each input point and more in general it represents a measure of how much a network's logit can be perturbed within a neighborhood of a sample. All the following experiments used PGD with step size $\alpha = 2/255$, perturbation bound of $\epsilon = 8/255$, which will represent the size of the neighborhood we're exploring relative to the original sampe, using ℓ_∞ norm, random initialization, and an increasing sequence of attack iterations of $[1, 2, 5, 10, 20, 50]$. Lastly, all reported results are averaged over multiple runs with different Pytorch seeds.

Our central hypothesis is that aspects of robustness of an untrained model might be predictable from architecture alone, *prior to undergoing a resource intensive adversarial training*.

In particular, we argue that when PGD produces perturbed samples whose energy differs only minimally from that of the original inputs, the architecture itself is enforcing local smoothness around the data, which translates into higher robustness. Prior works have shown that such local energy smoothness is associated with higher robustness in *adversarially trained models* [21], [12], and our analysis suggests that this desirable property is, to some extent, already encoded in the architecture before any adversarial training.

4.2 The Energy Landscape of Robust Models

To first highlight how local energy smoothness is a property of robust models when compared to standard *non-robust* model, we tested a variety of robust models available from the `RobustBench` [1] leaderboard by attacking them with increasingly stronger PGD (as before, increasing the number of iterations following the sequence $[1, 2, 5, 10, 20, 50]$ and collecting the resulting energies of the original and their adversarial counterparts. This analysis focuses on WRN-28-10 architectures, which are among the most widely adopted WRN variants due to their favorable balance between performance and model size.

It should be noted that the absolute numerical values of energies are generally arbitrary. To evaluate local smoothness, we therefore focus on the *energy differences* between clean samples and their adversarial counterparts. Comparing robust models with standard non-robust models trained on the same CIFAR-10 dataset, we find that robust models maintain difference in marginal energies $\Delta E(x)$ largely invariant to the strength of the attack, while differences in the joint energies $\Delta E(x, y)$, associated with class predictions, are perturbed but remain close to their original values, as shown in Fig. 1. In contrast, the non-robust model exhibits substantial deviations in both marginal and joint energies under attack.

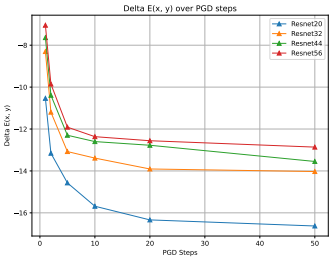


FIGURE 3. $\Delta E(x, y)$ of different ResNets over increasing attack strength

To extend the use of energy as a tool for analysing a network's properties, we also conducted the same studies on the non-robust baseline models for residual networks [6]. Specifically, four different sizes of ResNet architecture standardly trained on the CIFAR-10 dataset underwent the same test suite as the robust models: ResNet-20, ResNet-32, ResNet-44, and ResNet-56. The two digits refer to the number of convolutional layers within the network, including residual connections and with the addition of the last fully connected layer producing the logits of the model.

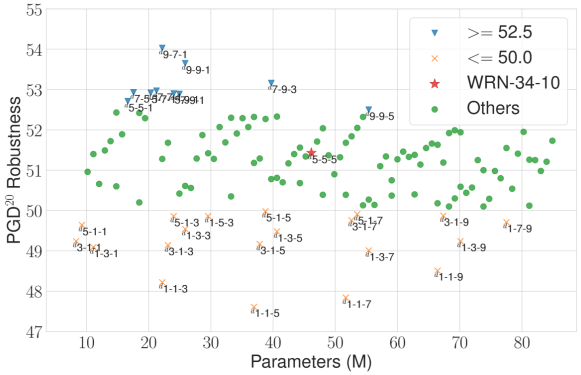


FIGURE 4. Points represent models with different depth configurations, y-axis contains robustness on PGD-20 after adversarial training, x-axis the number of parameters in the models. Image courtesy of [8]

Interestingly, deeper models are more resistant to perturbations as can be seen in Table 1. We also note that for deeper models, $\Delta E(x, y)$ appears more smoother than the shallow models as shown in Fig. 3. It then seems that a greater number of parameters contribute here to a stronger inherited robustness and we conjecture this might be due to higher redundancy in learned features in neurons while a smaller model requires fewer activations to be changed and thus a smaller difference with respect to the original sample to change its predictions. However, as will be discussed in the following section, a direct correspondence between model scale and adversarial robustness is not consistently observed.

ResNet	PGD^1	PGD^2	PGD^5
20	15.28 %	1.10 %	0.00 %
32	30.25 %	5.29 %	0.00 %
44	34.15 %	7.99 %	0.20 %
56	**37.45** %	**11.68** %	0.10%

TABLE 1. Different Residual Network sizes and their adversarial accuracy under different PGD step sizes

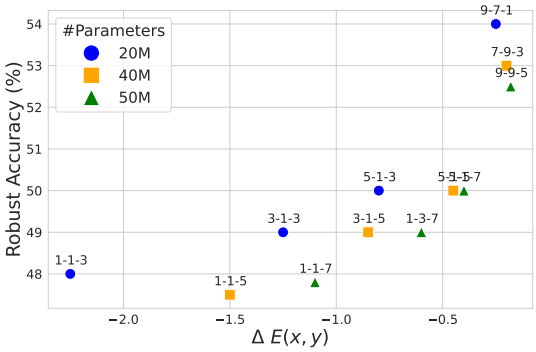


FIGURE 5. Scatterplot of $\Delta E(x, y)$ computed on **untrained** classifiers and the corresponding PGD robust accuracy computed from the same exact model architecture yet **adversarially trained**. Robust accuracies from [8]

4.3 The Energy Landscape of Untrained Models

As noted earlier model size isn't necessarily connected to final robustness as shown in [14] and [8], and to further investigate this fact we study 4 different architecture configurations across three different model sizes ($\approx 20, \approx 40, \approx 50$ Million parameters) over the CIFAR10 dataset. The grid search results reported in [8] (Fig. 4) provide the established final robustness percentages for these configurations.

When analyzing untrained architectures, we note that since the models do not go through any kind of learning, the marginal energies $E(x)$ give little information on the properties of the network as they are always a byproduct of training. On the contrary, joint energies $E(x, y)$, although agnostic to the actual label distribution, give an insight on the *sensitivity* of the network to adversarial attacks. Key to this study is the initialization method, as a network in an untrained state tends to output numerically unstable logits and gradients, rendering both attacks and energy metrics unfeasible unless methods such as Kaiming's [7] which are intentionally developed to contain variances of layers in both forward and backward passes are used.

After initializing the models, we proceed to attack samples and collect their energy metrics, progressively increasing the number of iterations of PGD and thus of its strength following this sequence $[1, 2, 5, 10, 20, 50]$. All reported results are averaged over multiple runs with different Pytorch seeds as the overall nature of this kind of analysis is inherently stochastic and requires a statistical analysis rather than a qualitative one.

As can be seen in 6a and 6b which contain the average values for $\Delta E(x, y) = E(x, y) - E(x^*, y)$, where $E(x, y)$ corresponds to the joint energy distribution defined in Eq. 1 and x^* is the adversarial sample found by PGD, the more a model will be robust after training the less its joint energy can be bent when initialized. Figure 6c shows the results for the bigger models with approximately ~ 50 M parameters.

We further analyze these results using scatter plots to investigate the relationship between energy differences, number of trainable parameters and robustness, shown in Fig. 5. Where models with smaller $|\Delta E(x, y)|$ achieve higher robust accuracy, regardless of the number of parameters.
We report both the Pearson Correlation coefficient r and the r_s Spearman Rank Correlation coefficient with 10 degrees of freedom in table Tab. 2 of the values shown in Fig. 5, with both metrics along with their p-values suggesting strong positive correlations, though we stress the limited number of trials that were conducted due to the speculative nature of this paper.

Correlation Metric	Value	p-value
Pearson Correlation (r)	0.7624	0.0039
Spearman Rank Correlation (r_s)	0.8937	< 0.0001

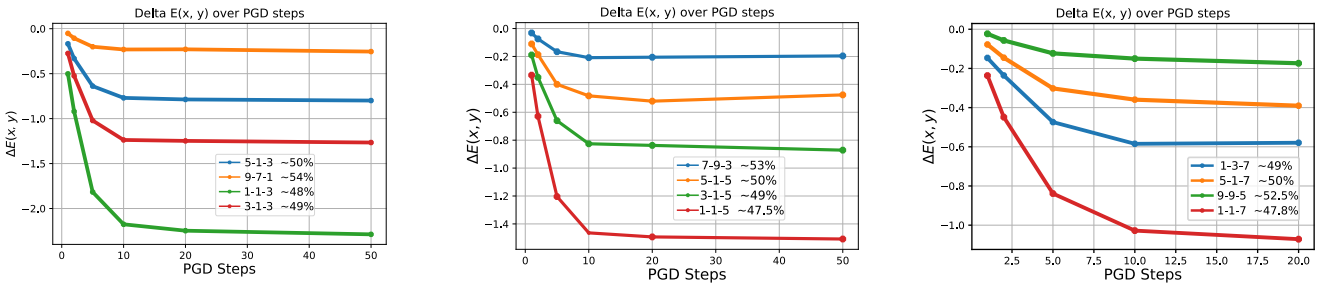
TABLE 2. Correlation metrics of values plotted in Fig 5.

Although the final accuracies of different models often differ by only a few percentage points, some smaller models with 20M parameters can achieve a robustness which is even greater than some of the 50M counterparts highlighting that the connection between robustness and model size is much nuanced.

Finally, we want to highlight that if architecture truly imposes inherit robustness, adversarial training isn't then able to compensate for these intrinsic differences as the margin between models will persist in the end, reinforcing the requirement for optimality of architecture during the model design stage.

5 Conclusions and Future Work

In this study, we investigate a methodology aimed at connecting network architecture with final robustness, namely by studying the impact of adversarial examples on untrained models with different architecture configurations. Our findings suggest a potential correlation between the sensitivity to perturbations when untrained and the resulting adversarial accuracy after training. Future work aims to expand analysis on both the amount of models tested and on the different types of architectures.



(a) $\sim$ 20 M parameters **(b)** $\sim$ 40 M parameters **(c)** $\sim$ 50 M parameters

FIGURE 6. Average $\Delta E(x, y)$ between adversarial sample and original of different model configurations with $\sim$ 20 M, $\sim$ 40 M, and $\sim$ 50 M parameters. Labels correspond to depth configurations.

References

[1] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. RobustBench: A standardized adversarial robustness benchmark, Oct. 2021.

[2] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019.

[3] C. Gallicchio and S. Scardapane. Deep Randomized Neural Networks, Feb. 2021.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples, Mar. 2015.

[5] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One, Sept. 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition, Dec. 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Feb. 2015.

[8] H. Huang, Y. Wang, S. M. Erfani, Q. Gu, J. Bailey, and X. Ma. Exploring Architectural Ingredients of Adversarially Robust Deep Neural Networks, Jan. 2022.

[9] T. Huster, C.-Y. J. Chiang, and R. Chadha. Limitations of the Lipschitz constant as a defense against adversarial examples, July 2018.

[10] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, Sept. 2019.

[12] M. H. Mirza, M. R. Briglia, S. Beadini, and I. Masi. Shedding more light on robust classifiers under the lens of energy-based models. In *ECCV*, 2024.

[13] M. H. Mirza, A. D'Orazio, O. Melamed, and I. Masi. A provable energy-guided test-time defense boosting adversarial robustness of large vision-language models. In *CVPR*, 2026.

[14] S. Peng, W. Xu, C. Cornelius, M. Hull, K. Li, R. Duggal, M. Phute, J. Martin, and D. H. Chau. Robust Principles: Architectural Design Principles for Adversarially Robust CNNs, Sept. 2023.

[15] C. Sitawarin, S. Chakraborty, and D. Wagner. SAT: Improving Adversarial Training via Curriculum-Based Loss Smoothing. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pages 25–36, Nov. 2021.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, Feb. 2014.

[17] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep Image Prior. *International Journal of Computer Vision*, 128(7):1867–1888, July 2020.

[18] X. Yin, S. Li, and G. K. Rohde. Learning energy-based models with adversarial training. Springer, 2022.

[19] X. Yin, C. Zhang, J. Steele, N. N. Shavit, and T. T. Wang. Joint discriminative-generative modeling via dual adversarial training. 2026.

[20] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy, June 2019.

[21] Y. Zhu, J. Ma, J. Sun, Z. Chen, R. Jiang, and Z. Li. Towards Understanding the Generative Capability of Adversarially Robust Classifiers, Sept. 2021.