Towards Explaining Adversarial Examples Phenomenon in Artificial Neural Networks

Barati R., Safabakhsh R., Rahmati M.

ramin.barati@aut.ac.ir safa@aut.ac.ir rahmati@aut.ac.ir

Computer engineering Department
University of Amirkabir

December 7, 2020

Overview

Introduction

- ANNs has become one of the more popular tools in many industries. However, it was that deep networks are sensitive to adversarial examples[1].
- Research has shown that adversarial samples can be misused in physical world as well[2].
- Providing an explanation for the existence of adversarial examples is crucial in minimizing their effect.
- We describe a new perspective on the reasons behind the existence of adversarial examples and discuss the consequences of our approach.

Related work

Low Probability Pockets[1]

- Viewed the adversarial examples as pockets in the input space which have a low probability of being observed and correctly classified.
- Further described with an analogy between the real numbers as the natural samples and the rational numbers as the adversarial examples.
- Does not justify why would a classifier show such a behavior and it is only considered as a side effect of nonlinearity of ANNs.[3]
- It was shown in [4] that adversarial examples are not isolated points and they form dense regions in the input space.

Related work (cont.)

Linearity Hypothesis[5]

- Relates the adversarial phenomenon to a side effect of linear classifiers in high dimensions.
- Predicts that the classifier would be fooled if we choose a perturbation to be the sign of the gradient of the loss function with respect to the input.
- Using this theory, Goodfellow et al. developed the *Fast Gradient Search Method* (FGSM) which was able to easily produce adversarial examples that would also transfer to other networks.
- It was shown that there are interpretations of the hypothesis that does not produce correct predictions[3].

Related work (cont.)

Nonrobust Features[6]

- Explain the existence of adversarial examples by attributing them to features that are predictive, but nonrobust.
- A useful but nonrobust feature is a feature that is highly predictive of the true label on the empirical distribution of samples and labels, but if we add adversarial perturbations to samples, it would not be as useful anymore.
- The authors show that these features consistently exist in standard datasets and tie the phenomenon they observed to a misalignment between the human-specified notion of robustness and the inherent geometry of data.

Proposed Notion

We propose that the phenomenon occurs when a universally **consistent learning rule** on a **nonuniform learnable** hypothesis class does not guarantee uniform convergence.

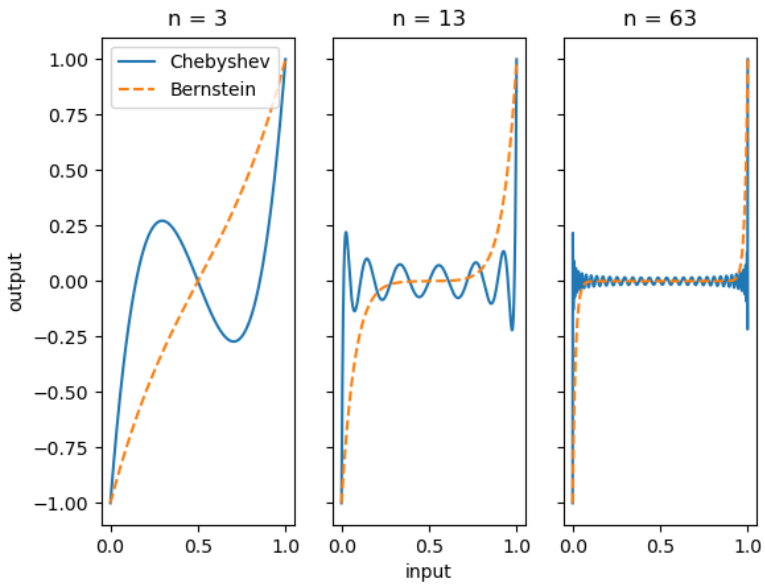


Figure 2: Pointwise convergence of Chebyshev polynomials

Proposed Notion (cont.)

Sample complexity of a consistent learning rule depends on the generating distribution of the data as well as the hypothesis.

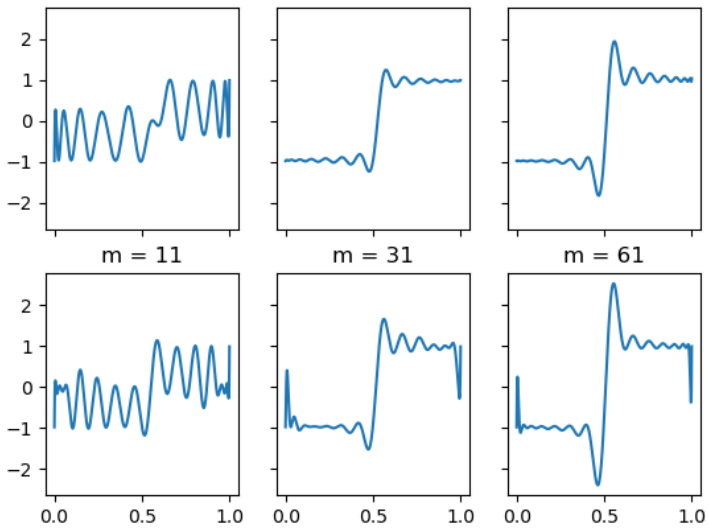


Figure 3: Chebyshev grid minimizes the sample complexity for Chebyshev polynomials

Proposed Notion (cont.)

Optimal training points: Let $\mathcal{X}$ be a domain set, let $\mathcal{H}$ be a hypothesis class and let A be a universally consistent learning rule with respect to $\mathcal{H}$. For every $X \subset \mathcal{X}$ and every $h \in \mathcal{H}$, let

$$\hat{h}_X = A(\{(x, h(x)) \,|\, x \in X\}).$$

The optimal training points of A is a solution to the following problem plus the boundary $\partial\mathcal{X}$ of $\mathcal{X}$,

$$\begin{aligned} \underset{X}{\arg\min} \quad & \int_{\mathcal{H}} \sum_{x \in X} \|\nabla(\hat{h}_X(x) - h(x))\| \, dh \\ \text{subject to} \quad & X \text{ is feasible} \end{aligned} \tag{1}$$

Proposed Notion (cont.)

Computing the objective of (1) is not tractable for any practical purpose. However, steps could be taken to calculate an approximation.

Hard training points of H: The hard training points of A with respect to a distribution H on $\mathcal{H}$ is a solution to the following problem,

$$\begin{aligned} \arg\max_{X} \quad & \mathbb{E}_H[\sum_{x \in X} L(h(x), h(x))] \\ \text{subject to} \quad & X \text{ is feasible} \end{aligned} \tag{2}$$

Adversarial training points of h: The adversarial training points of A with respect to $h \in \mathcal{H}$ is a solution to the following problem,

$$\begin{aligned} \arg\max_{X} \quad & \sum_{x \in X} L(h(x), h(x)) \\ \text{subject to} \quad & X \text{ is feasible} \end{aligned} \tag{3}$$

Experiments

The definition of hard training points is consistent with the observations made about Chebyshev polynomials.

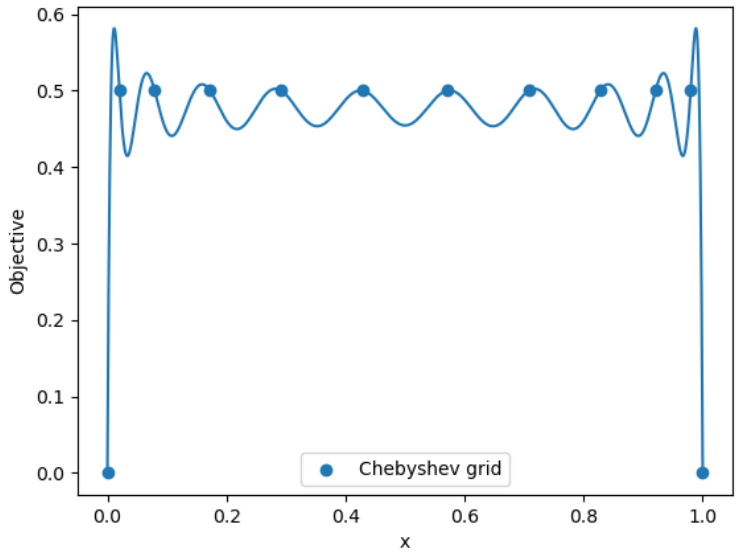


Figure 4: Comparing the position of a Chebyshev grid with the Hard training points objective of Chebyshev basis polynomials.

Experiments (cont.)

Even though the proposed notion can explain the existence and abundance of adversarial examples in MLPs, it cannot further explain their transfer between different architectures of MLPs.

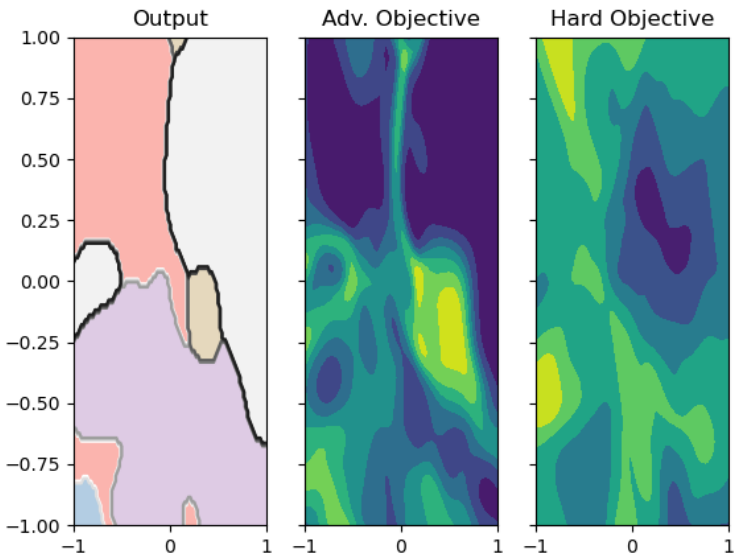
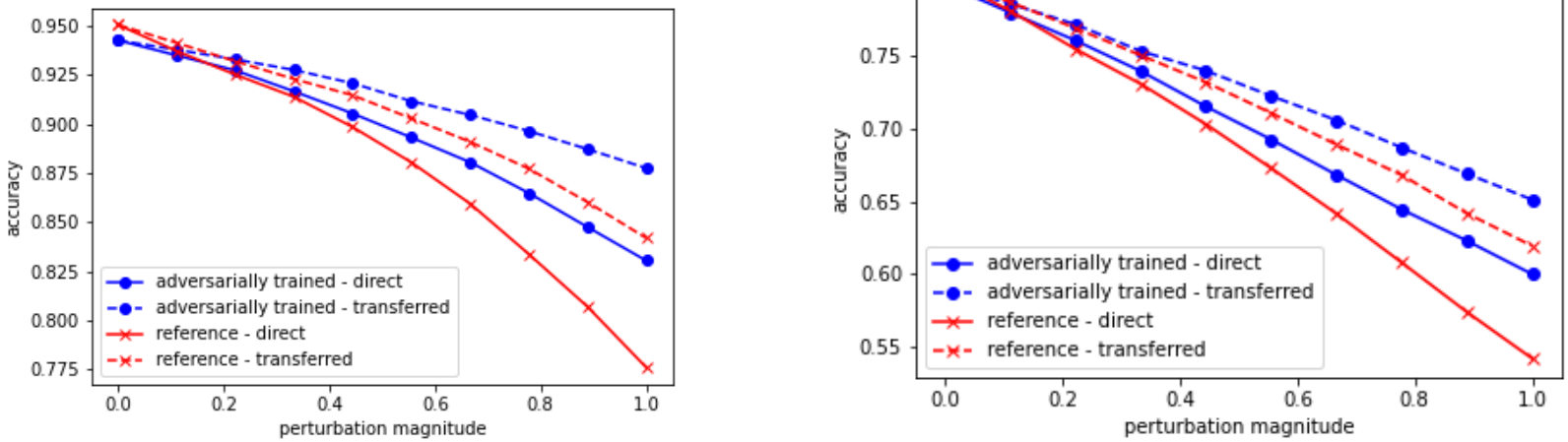


Figure 5: A comparison between the output, the adversarial objective and hard objective of a random network.

Experiments (cont.)

Adversarial training using (2) would make the trained MLP more robust against adversarial attacks by attacking random hypothesises that use the same feature layer instead of attacking the hypothesis itself.



(a) MNIST (b) Fashion-MNIST

Figure 6: Adversarial performance of adversarially trained networks

Conclusion and future work

- Defined the adversarial examples as the critical points of the error function.
- Constructed the first instance of a classifier with adversarial examples that are dense in the input domain.
- MLPs do follow the principle in case of existence and training, but it is insufficient for explaining transferable adversarial examples.
- If approximation theory is any indicator of the path forward, our next step should be to analyse the classifier independently from the training points.

Bibliography

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. arXiv: 1607.02533. [Online]. Available: http://arxiv.org/abs/1607.02533.

[3] T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," *arXiv preprint arXiv:1608.07690*, 2016.

[4] P. Tabacof and E. Valle, "Exploring the space of adversarial images," *2016 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2016. DOI: 10.1109/ijcnn.2016.7727230. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2016.7727230.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

Bibliography (cont.)

[6] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *arXiv preprint arXiv:1905.02175*, 2019.