Towards Explaining Adversarial Examples Phenomenon in ANNs

Ramin Barati, Reza Safabakhsh, Mohammad Rahmati

University of Amirkabir

Introduction

With the recent advancements in processing power and algorithms, machine learning has become one of the more popular tools in many industries. However, it was shown that deep networks are sensitive to adversarial examples[3]. Adversarial examples are points close to a natural example in which the classifier output changes drastically. As a result, many concerns has been raised in regards to the use of ANNs in safety-critical applications[6]. Providing an explanation for the existence of adversarial examples is crucial in minimizing their effect. Here, we study the adversarial examples existence and adversarial training from the standpoint of convergence and provide evidence that pointwise convergence in ANNs can explain these observations. The main contribution of our proposal is that it relates the objective of the evasion attacks and adversarial training with concepts already defined in learning theory.

Related Work

Low probability pockets: Szegedy et al. viewed the adversarial examples as pockets in the input space which have a low probability of being observed and correctly classified[3]. This view has been further described with an analogy between the real numbers as the natural samples and the rational numbers as the adversarial examples. Nevertheless, the argument does not justify why would a classifier show such a behavior[5]. This view is further undermined when it was shown in [4] that adversarial examples are not isolated points and they form dense regions in the input space.

Linearity hypothesis: According to the linearity perspective, nonlinear classifiers like ANNs are showing the phenomenon because they are trained with algorithms that prefer linear models. As stated in [1], since the sensory precision is fairly limited, the input domain is discrete. But, on a computer, we have to represent these quantities using floating-point precision numbers. If the input dimension is large enough, the accumulated error of floating-point arithmetic in the output could be considerable even if the perturbation was actually small. Linearity means that the trained classifiers show a similar behavior and would be fooled if we choose the direction of the perturbation to be the sign of the gradient of the loss function with respect to the input. Linearity perspective explains adversarial transfer as the side effect of converging to the optimal linear classifier.

Nonrobust features: The proposal of Ilyas et al. is also of interest in our discussion[2]. They 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

Pointwise convergence: Consider the training set $S = \{(0, -), (1, +)\}$. This training set consists of two samples from $[0, 1]$ interval. Suppose that we want to find the maximum margin classifier of S. To do so, we need to choose a set of features first. First, we use the Bernstein basis polynomials as features and find the maximum margin classifier f_n for a Bernstein polynomial of degree n. It could be checked that f_n is robust and does not show any adversarial regions.
Similarly, the same classification task could be posed for shifted Chebyshev basis polynomials as features instead. Unlike Bernstein basis polynomials, Chebyshev basis polynomials do not guarantee uniform convergence and reveal the error caused by pointwise convergence. Let g_n be the maximum margin classifier for the Chebyshev basis.

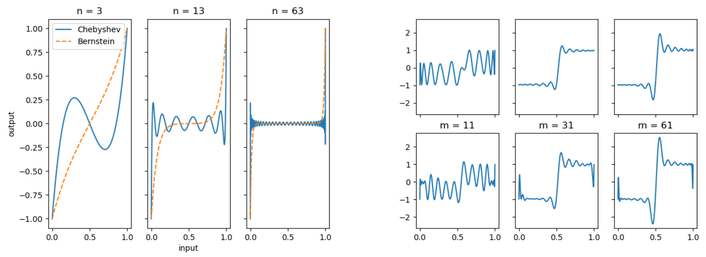

Fig. 1: Comparison of the Bernstein and Chebyshev maximum margin classifiers.

Figure 1 (left) compares a few members of f_n and g_n. We propose that the phenomenon occurs when a universally consistent learning rule on a nonuniform learnable hypothesis class does not guarantee uniform convergence. According to the definition of consistency, sample complexity of a consistent learning rule depends on the generating distribution of the data as well as the hypothesis. As a result, there could be distributions of data that maximize or minimize the sample complexity of the learning rule for any hypothesis.
To showcase the proposal, we keep the degree of the Chebyshev polynomial constant and instead increase the number of training points. The training points get labeled according to the nearest neighbour classifier of S. We construct two training sets, E_m and C_m. E_m is constructed using an equispaced grid of m points, C_m on the other hand use a Chebyshev grid as samples. The Chebyshev maximum margin classifier of degree 30 of E_m and C_m is depicted in Figure 1 (right) for a few choices of m. It could be seen in the figure that in case of E_m (bottom), uniform convergence does not occur even for relatively large m. In contrast, classifiers trained on C_m (top) converge in the expected number of samples. As a matter of fact, Chebyshev polynomials guarantee uniform convergence on a Chebyshev grid, and hence, do not suffer from the phenomenon.

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 $\mathcal{X}$,

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

The intuition behind the definition of the optimal points is that by minimizing the empirical risk on the critical points of the error function, we are effectively minimizing the maximum norm of the true error. Computing the objective of (1) is not tractable for any practical purpose. However, steps could be taken to calculate an approximation. First, we can utilize the loss function L as a surrogate for the objective of (1). Second, we approximate the integral in (1) by restricting the hypothesis class.

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,

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

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,

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

Experiments

To exhibit that the definition of hard training points is consistent with the observations made about Chebyshev polynomials, we find the hard training points of these polynomials in a maximum margin classification setting and show that a Chebyshev grid is very close to the optimal solution.
Next, we analyze MLPs through our framework. To this end, we sample a random MLP classifier and visualize the adversarial and hard training points objectives with respect to that network. By comparing the output of the network with the adversarial objective of the network, we can see that the adversarial objective is sensitive to the position of the decision boundary. Contrasting the adversarial objective and the hard objective reveals that a hypothesis agnostic set of points would not be as helpful as for the case of polynomials. In other words, while we could reach a definitive set of optimal points for polynomials, the optimal points of MLPs are not unique even up to the layers of a single network. As a result, 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.
Finally, we show that sampling according to the hard objective described in our proposal does indeed optimize the sample complexity of the robust hypothesis for a MLP on real-world data.

Conclusion and future work

Present paper introduces a framework to explain the adversarial examples phenomenon. We defined the adversarial examples as the critical points of the error function. We showed that this principle can fully explain adversarial examples existence, training and transfer for pointwise converging polynomials.
Unfortunately, this approach does not enjoy the same success in case of MLPs. We could demonstrate that MLPs do follow the principle in case of existence and training, but the same principle proved to be insufficient in explaining transferable adversarial examples. Consequently, for future work, we will focus on extending the framework to account for transfer of adversarial examples between ANNs as well.
With regards to other proposals in literature, our definition is more aligned with the low probability pockets perspective. We constructed the first instance of a classifier with adversarial examples that are dense in the input domain.
With respect to the linearity hypothesis, our analysis does not confirm the view that the phenomenon is a result of representing the input by floating-point numbers. Nevertheless, we share the idea that the transfer is rooted in properties of the optimal classifier.
The proposed notion is similar to the nonrobust features perspective in that we also relate the phenomenon to a form of optimal training points. Our proposal can accommodate a notion of pointwise converging features similar to nonrobust features as well.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[2] Andrew Ilyas et al. “Adversarial examples are not bugs, they are features”. In: *arXiv preprint arXiv:1905.02175* (2019).

[3] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).

[4] Pedro Tabacof and Eduardo Valle. “Exploring the space of adversarial images”. In: *2016 International Joint Conference on Neural Networks (IJCNN)* (July 2016). DOI: 10.1109/ijcnn.2016.7727230. URL: http://dx.doi.org/10.1109/IJCNN.2016.7727230.

[5] Thomas Tanay and Lewis Griffin. “A boundary tilting persepective on the phenomenon of adversarial examples”. In: *arXiv preprint arXiv:1608.07690* (2016).

[6] Xiaoyong Yuan et al. “Adversarial Examples: Attacks and Defenses for Deep Learning”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2019), pp. 1–20. ISSN: 2162-2388. DOI: 10.1109/tnnls.2018.2886017. URL: http://dx.doi.org/10.1109/TNNLS.