ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Adversarial Robustness of Linear Models: Regularization and Dimensionality

István Megyeri[1], István Hegedűs[1] and Márk Jelasity[1,2] *

1- University of Szeged, Hungary
2- MTA-SZTE Research Group on Artificial Intelligence, Hungary

Abstract. Many machine learning models are sensitive to adversarial input, meaning that very small but carefully designed noise added to correctly classified examples may lead to misclassification. The reasons for this are still poorly understood, even in the simple case of linear models. Here, we study linear models and offer a number of novel insights. We focus on the effect of regularization and dimensionality. We show that in very high dimensions adversarial robustness is inherently very low due to some mathematical properties of high-dimensional spaces that have received little attention so far. We also demonstrate that—although regularization may help—adversarial robustness is harder to achieve than high accuracy during the learning process. This is typically overlooked when researchers set optimization meta-parameters.

1 Introduction

The high sensitivity of most machine learning models to adversarial examples was pointed out not long ago [1, 2]. A number of methods have been proposed to create better adversarial examples [3, 4] as well as to provide defense mechanisms against these [5, 6].

Here, we focus on the adversarial robustness of linear machine learning models. The theoretical basis of the problem is still lacking. Some results are known e.g. Fawzi et al. [7] offer bounds on robustness for the linear case based on the distance of classes, but their study is orthogonal to ours. Goodfellow at al. [1] suggested that higher-dimensional linear models are more sensitive because the same amount of noise in each dimension can result in a larger Euclidean distance from the point simply due to the larger number of dimensions, provided the sign of the noise is the same as the sign of the value in the given dimension. However, we argue that the Euclidean distance is of limited interest simply because classes and data points in general will also have larger Euclidean distances from each other in higher dimensions.

In this paper, we propose novel insights, which provide an alternative explanation to the adversarial sensitivity of linear models. We focus on the effect of regularization and dimensionality. We will show that in very high dimensions adversarial robustness is inherently very low due to the fact that a random hyperplane is very close to any data point. This property which has received little attention so far, is highly counter-intuitive.

*This study was supported by the National Research, Development and Innovation Office of Hungary through the Artificial Intelligence National Excellence Program (grant 2018-1.2.1-NKP-2018-00008) and by the Hungarian Ministry of Human Capacities (grant 20391-3/2018/FEKUSTRAT).

We also point out that regularization has a profound effect on adversarial robustness. From the point of view of prediction accuracy and adversarial robustness the amount of regularization required will be different. We should add that the current practice of setting meta-parameters based only on prediction accuracy might result in very high sensitivity to adversarial examples. This is because the convergence of robustness is much slower than that of accuracy and because robustness requires stronger regularization.

We shall also provide a thorough experimental evaluation of our claims where we study the effect of dimensionality, regularization, and the interaction of these two factors. In this evaluation, we will use artificial datasets as well as a subset of the MNIST dataset.

2 Linear Models in High Dimensional Spaces

We are given a set of training instances of the form (x, y), $x \in \mathbb{R}^n$, $y \in \{0, 1\}$, and we are looking for a hyperplane $Pl(w) = \{z|\langle w, z\rangle = 0\}$ defined by $w \in \mathbb{R}^n$ such that $Pl(w)$ separates the data points with different labels. This plane is typically found via optimizing a loss function based on the examples and w. Model optimization typically starts with a random initial model, or, equivalently, an initial model that is independent of the optimal model. The following result implies that such a random model will be extremely close to *any* point in expectation, hence, it should also be very close to each instance. This highly unintuitive property implies that a random plane has a very high sensitivity to adversarial examples.

Proposition. *Let $w \in \mathbb{R}^n$ define a random plane $Pl(w) = \{z|\langle w, z\rangle = 0\}$. Let w_i $(i = 1, \ldots, n)$ be i.i.d. random variables with $P(w_i = -1) = P(w_i = 1) = 0.5$. Let $d(\mathbf{1}, Pl(w))$ denote the distance between $Pl(w)$ and the point $\mathbf{1} = (1, \ldots, 1)$. Then we have $\lim_{n\to\infty} \mathbb{E}(d(\mathbf{1}, Pl(w))) = O(1)$.*

Proof. We have $d(\mathbf{1}, Pl(w)) = |\langle \mathbf{1}, w\rangle| / \|w\|_2 = \frac{1}{\sqrt{n}} |\sum_{i=1}^n w_i|$. Also, we have $\sum_{i=1}^n w_i \to \sqrt{n}\mathcal{N}(0, \sigma^2)$ due to the central limit theorem, where $\sigma^2 = 0.25$ is the variance of w_i. The mean of $|\mathcal{N}(0, \sigma^2)|$ is finite and it does not depend on n, so it is $O(1)$; thus $\mathbb{E}(\frac{1}{\sqrt{n}} |\sum_{i=1}^n w_i|) \to \frac{1}{\sqrt{n}} \sqrt{n} O(1) = O(1)$, which completes the proof. □

Note that there exists a plane for which the distance from $\mathbf{1}$ is $\sqrt{n} = O(\sqrt{n})$, namely when $w = \mathbf{1}$. However, according to the result above, a random plane is of distance $O(1)$ in expectation. The result is not specific to $\mathbf{1}$ because it is invariant to rotation. The striking consequence is that a random plane will result in a high sensitivity to adversarial examples, *no matter how the classes are positioned*. This means that the optimal plane in terms of distance is very special, regardless of the difficulty of the classification problem, so we suspect that this very special plane is hard to find during optimization. Our experimental results are consistent with this view.

3 Linear Models and Regularization

Here, we argue that regularization is closely related to the geometric properties outlined in Section 2. Assuming n examples (x_i, y_i), $x_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$, $i = 1, \ldots, n$, let us now consider logistic regression where the goal is to approximate the data using the logistic function $y \approx \sigma(w^T x + b) = 1/(1 + e^{-w^T x + b})$. This will lead to a linear separator defined by w and b and a logistic probability approximation as a function of the distance from the separating hyperplane.

The loss function typically used to find the best model (that is, w and b) is the negative log likelihood function $L(w, b) = -\sum_{i=1}^{n} y_i \cdot log(\sigma(x_i; w, b)) + (1 - y_i) \cdot log(1 - \sigma(x_i; w, b))$. To handle noisy data, it is customary to add a regularization term to the loss function. Here we focus on the so-called L2 regularization: $L(w, b) + \alpha\|w\|_2^2$, where α is the regularization coefficient.

We would like to study the effect of regularization from the point of view of adversarial robustness. L2 regularization results in preventing the length of w from growing indefinitely. This in turn results in preventing the derivative of the model from growing indefinitely. To see this, consider the derivative $\sigma(a \cdot x)' = a\sigma(a \cdot x)\sigma(1 - a \cdot x)$. Clearly, increasing the length of w will make the logistic curve steeper. Without regularization, the model in practice becomes a step function so the loss function will simply attempt to minimize the number of misclassified examples. With regularization, all the examples will affect the orientation of the separating hyperplane.

This means that if regularization is not strong enough then noisy examples will have too much influence, forcing the hyperplane out of optimal position, which in turn will result in very high adversarial sensitivity according to the proposition provided in Section 2. Accordingly, we expect that for optimal robustness one will have to use quite strong regularization.

4 Experimental Results

In order to evaluate the effect of dimensionality and regularization, we carried out a systematic experimental study. Now let us describe the experimental setup and the methodology in detail.

4.1 Binary Classification Problems

We will use two binary classification problems that are described below. The first dataset is a subset of the MNIST dataset [8] that includes two classes: 3 and 7 (also used by the authors of [1]). We will refer to this dataset as MNIST-73. It contains about 6000 samples per class. The raw pixel values were normalized to the range $[0, 1]$.

We will also use an artificial dataset called 2-Gauss. The two classes are defined by the distributions $\mathcal{N}(\mathbf{1}, \Sigma)$ and $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ is the origin and Σ is the diagonal matrix $4\mathbf{I}$, hence the variance is $\sigma^2 = 4$, which is the same for each dimension. Note that the Euclidean distance of the class centers is $\sqrt{n}$, where n is the dimension. Here, we sampled 6000 instances per class.

For each dataset, $100/6 \approx 16.7\%$ of the data was separated to form a test set. In the preprocessing step, the training data values were translated so as to have a zero mean. The mean was estimated over the training set, and the test set was translated as well using this value.

To examine the effect of the dimensionality on adversarial robustness, we will use a range of input dimensions. The dataset 2-Gauss can naturally be generated in any dimensions. The MNIST-73 examples were scaled using image processing algorithms. The original dimension of the images was 28×28. We performed preliminary tests with different interpolation methods (cubic, linear, nearest-neighbor) that gave similar results. Here, we applied the nearest-neighbor method.

4.2 Methodology

Our two main measures of interest are *accuracy* (i.e. the proportion of correctly classified examples) and the distance of the examples from the hyperplane normalized by the dimension $\sqrt{n}$. The latter measure characterizes the sensitivity to adversarial examples; namely the smaller the distance, the higher the sensitivity. Here, we normalize the distance by $\sqrt{n}$ for two reasons. First, it is more meaningful to measure sensitivity *relative* to the distance of the two classes, and the distance of the two classes grows with $\sqrt{n}$. Second, in the case of image data, this also means that we characterize the sensitivity of each pixel, which is a more natural measure. We will call this measure the *normalized distance*.

We used ADAM [9] as our optimizer with a minibatch size of 32. Since we were interested in the actual optimal model (to avoid artifacts due to early stopping) we ran the algorithm with an extremely small stopping threshold of 10^{-10}. We will also include results with a 10^{-4} stopping threshold that is often used as a default. We can still study the effect of early stopping, since we record the convergence history as well. In our plots, we will indicate the regularization coefficient used in the case of $n = 28 \times 28$, however, for different dimensionalities, the regularization value was scaled proportional to n to make the strength of regularization in different dimensions comparable.

4.3 Results

Figure 1 shows some of the results of our experiments. The MNIST-73 results indicate that normalized distance and accuracy behave very differently in terms of regularization. Most importantly, one is normally interested in prediction performance, and the meta-parameters optimal for that purpose perform rather badly for adversarial robustness. To optimize the distance, it is good to have a regularization coefficient that is as large as possible, whereas accuracy displays a degrading trend with increased regularization. These observations hold true regardless of the problem dimension. In other words, in each dimension we see that they have almost the same values.

The 2-Gauss problem behaves slightly differently because no noisy examples are added and because in high dimensions there is a wide linear separation margin between the classes and this grows with n. The optimal values for distance and accuracy are found in almost every case. However, we noticed that for low regularization values the optimizer struggles to find the optimum in high dimensions. For no regularization, even the smaller stopping threshold is insufficient

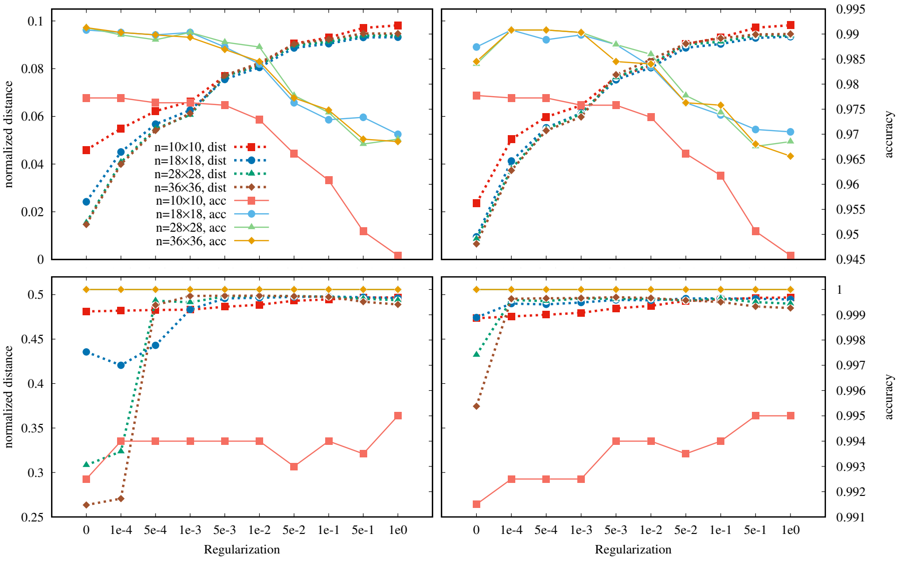

Fig. 1: Normalized distance and accuracy as a function of regularization coefficient and dimension for the MNIST-73 dataset (top) and the 2-Gauss dataset (bottom), and stopping threshold 10^{-4} (left) and 10^{-10} (right).

to find the theoretically optimal model. This is because then the loss function is extremely flat. This effect is closely related to the dimensionality n, and the problem is more severe with larger values of n.

Let us also examine the dynamics of convergence during optimization, which is shown in Figure 2. Clearly, the convergence of distance is significantly slower than that of accuracy in each case. For the 2-Gauss problem, this effect is more marked. With $\alpha = 10^{-4}$, due to the wide separation margin and relatively large weights, the loss function practically vanishes and gives only a very weak signal to the optimizer, while the accuracy attains its optimum quite quickly.

With the MNIST-73 dataset we see there is a local optimum for distance when no regularization is applied. This is due to the length of the parameter vector w gradually increasing. With the 2-Gauss dataset we have no noisy examples that could make the model go in the wrong direction as w grows due to the lack of regularization, so this effect is not so marked.

5 Conclusions

In this study, we demonstrated that even in the case of simple binary classification problems with linear models, the adversarial problem is real and it strongly depends on regularization and the less obvious properties of high-dimensional spaces. We presented an experimental evaluation where we showed that the optimal regularization strength is very different for adversarial robustness and prediction accuracy, and that the convergence of adversarial robustness is much slower than that of the accuracy metric. Also, in higher dimensions an overly

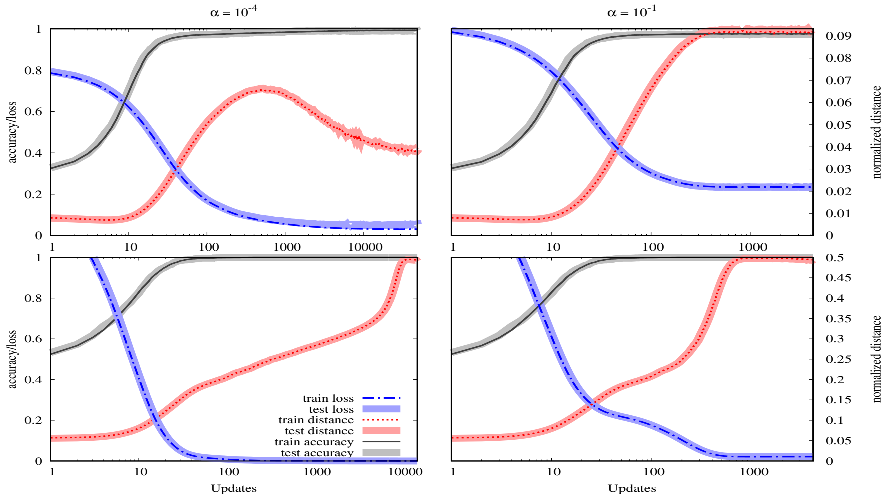

Fig. 2: Convergence of normalized distance and accuracy in $n = 28 \times 28$ dimensions for the MNIST-73 dataset (top) and the 2-Gauss dataset (bottom), with regularization coefficient $\alpha = 10^{-4}$ (left) and $\alpha = 10^{-1}$ (right).

weak regularization setting might result in a significantly harder optimization problem in some cases.

References

[1] Ian J. Goodfellow and Jonathon Shlens Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd Intl. Conf. on Learning Representations (ICLR)*, 2015.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd Intl. Conf. on Learning Representations (ICLR)*, 2014.

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, June 2016.

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th Intl. Conf. on Learning Representations (ICLR)*, 2018.

[6] Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th Intl. Conf. on Learning Representations (ICLR)*, 2018.

[7] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, Mar 2018.

[8] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, November 1998.

[9] Jimmy Ba and Diederik Kingma. Adam: A method for stochastic optimization. In *3rd Intl. Conf. on Learning Representations (ICLR)*, 2015.