

---

# Investigating Decision Boundaries of Trained Neural Networks

---

**Roozbeh Yousefzadeh**

Department of Computer Science  
University of Maryland  
College Park, MD 20742  
[rroozbeh@cs.umd.edu](mailto:rroozbeh@cs.umd.edu)

**Dianne P. O’Leary**

Department of Computer Science and  
Institute for Advanced Computer Studies  
University of Maryland  
College Park, MD 20742  
[oleary@cs.umd.edu](mailto:oleary@cs.umd.edu)

## Abstract

Deep learning models have been the subject of study from various perspectives, for example, their training process, interpretation, generalization error, robustness to adversarial attacks, etc. A trained model is defined by its decision boundaries, and therefore, many of the studies about deep learning models speculate about the decision boundaries, and sometimes make simplifying assumptions about them. So far, finding exact points on the decision boundaries of trained deep models has been considered an intractable problem. Here, we compute exact points on the decision boundaries of these models and provide mathematical tools to investigate the surfaces that define the decision boundaries. Through numerical results, we confirm that some of the speculations about the decision boundaries are accurate, some of the computational methods can be improved, and some of the simplifying assumptions may be unreliable, for models with nonlinear activation functions. We advocate for verification of simplifying assumptions and approximation methods, wherever they are used. Finally, we demonstrate that the computational practices used for finding adversarial examples can be improved and computing the closest point on the decision boundary reveals the weakest vulnerability of a model against adversarial attack.

## 1 Introduction

Interpreting the behavior of trained neural networks, their generalization error, and robustness to adversarial attacks are open research problems that all deal, directly or indirectly with the decision boundaries of these models. The decision boundaries of neural networks have typically been investigated through simplifying assumptions or approximation methods. As we will show in our numerical results, many of these simplifications may lead to unreliable results.

In this work, we show that better information about interpretation, generalization error, and robustness of a model can be obtained by computing *flip points*. Flip points are points on the decision boundaries of the model. For any data point, its closest flip point is the closest point that would flip the decision of the model to indeterminate. The direction between the data point and its closest flip point reveals a great deal about the most influential features in decision making of the model.

Some previous work has highlighted the importance of boundary points. Studies such as Lippmann [1987] investigate the decision boundaries of single-layer perceptrons, while describing the difficulties that arise regarding the complexity of decision boundaries for multi-layer networks. While Spangher et al. [2018] proposed a method to find the least changes in the input that would flip the classification of a model, their method is applicable only to linear classifiers, and they do not actually investigate the decision boundary. Wachter et al. [2018] defined *counterfactuals* as the possible changes in the

input that can produce a different output label. But, for a continuous model, the closest counterfactual is ill-defined, since there are points arbitrarily close to the decision boundaries that can produce different output labels. Furthermore, they solve their problem by enumeration, applicable only to a small number of features.

Some studies on interpretation of deep models have made simplifying assumptions about the decision boundaries. For example, Ribeiro et al. [2016] assumes that the decision boundaries are locally linear. Their approach tries to sample points on two sides of a decision boundary, then perform a linear regression to approximate the decision boundary and explain the behavior of the model. However, as we show in our numerical results, decision boundaries of neural networks can be highly nonlinear, even locally, and a linear regression can lead to unreliable explanations.

Regarding the generalization error of trained models, Elsayed et al. [2018] and Jiang et al. [2019] have shown there is a relationship between the closeness of training points to the decision boundaries and the generalization error of a model. However, they regard computing the distance to the decision boundary as an intractable problem and instead use the derivatives of the output to derive an approximation to the closest distance. In our numerical results, we compare their approximation to our results, and show the advantages of computing the distance directly. Other studies such as Neyshabur et al. [2017] approximate the closest distance to the decision boundary by the closest distance to an input with another label, which can be an overestimate. They use the modified margin between softmax outputs as a measure of distance, which we see later can be misleading.

Regarding adversarial attacks, there are many studies that seek small perturbations in an input that can change the classification of the model. For example, Moosavi-Dezfooli et al. [2016], Jetley et al. [2018], Fawzi et al. [2017] apply small perturbations to the input until its classification changes, but, since they do not attempt to find the closest point on the decision boundaries of the model, they do not reveal its weakest vulnerabilities. Most recent studies on adversarial examples, such as Tsipras et al. [2019] and Ilyas et al. [2019] minimize the loss function of the neural network for the adversarial label, subject to a distance constraint. They impose the distance constraint in order to find an adversarial example similar to the original image. Although this method is an important tool, this form of seeking adversarial examples has certain limitations, regarding the ability to make the models robust, and regarding the measurement of robustness of models, as we explain through numerical examples. We show that finding the closest point on the decision boundary accurately represents the least perturbation needed for adversarial classification, and, therefore, studies on adversarial examples can benefit from direct investigation of decision boundaries.

## 2 Computing the closest flip point

In the results we show later, the output of the neural network has been computed using softmax, which gives us a convenient normalization. For a given data point and its output class, we compute the closest flip point with respect to another output class by solving a numerical optimization problem: minimizing the distance between the given data point and the unknown (flip) point, subject to the constraints that, at the unknown point, the softmax outputs for the two classes are equal, and that no other softmax output is greater [Yousefzadeh and O’Leary, 2019].

Our optimization problem is nonconvex, so we cannot be sure that optimization algorithms will find the global minimizer. One important fact that makes the optimization easier is that we have a good starting point, the data point itself. We have solved our optimization problem using the applicable algorithms in 3 packages, NLOpt [Johnson, 2014], IPOPT [Wächter and Biegler, 2006], and Optimization Toolbox of MATLAB, as well as our own custom-designed homotopy algorithm. The algorithms almost always converge to the same point; in fewer than 5% of images, the interior point algorithms find closer flip points. The variety and abundance of global and local optimization algorithms in the above optimization packages give us confidence that we have indeed usually found the closest flip point. In any case, we demonstrate below that our flip points are closer than those estimated by methods such as linear approximations.

In our numerical results on image data, we measure distance using the  $\ell_2$  norm. In general, the choice of distance measure should be guided by practitioners who understand the nature of the data.

### 3 Numerical experiment setup

To illustrate our ideas, we use a 12-layer feedforward neural network trained on 2 classes of the CIFAR-10 dataset, ships and planes; details are provided in an appendix. To train the network, we have used Tensorflow, with Adam optimizer, learning rate of 0.001, and Dropout with rate 50%.

We use a tunable error function as the activation function. This allows us to introduce nonlinearity into the model while having control over the magnitude of the derivatives. Keep in mind that one can compute flip points for trained models and interpret them regardless of the architecture of the model (number of layers, activation function, etc.), the training set, and the training regime (regularization, etc.).

Inputs to our network are not the pixels, but 200 of the 3D Daubechies-1 wavelet coefficients. We choose the 200 coefficients according to the pivoted QR factorization [Golub and Van Loan, 2012] of the wavelet coefficients for the training set. Using the most significant wavelet coefficients removes redundancies in the features of the image. Figure 1 shows the first ship image in the CIFAR-10 training set along with its reconstructions from subsets of wavelet coefficients. With fewer coefficients, the reconstructed image looks less similar to the original image. Nevertheless, the model is able to correctly classify most of the images by learning from those representations. This result may be in agreement with the arguments of Ilyas et al. [2019] that neural networks learn Gaussian representations of images.

The accuracy we obtain on the testing set is 84.05%. This can be improved to near 95% using the calculated flip points as new training points in order to move decision boundaries, or by using more wavelet coefficients. Notice that using a smaller number of wavelet coefficients leads to a model with smaller feature space and that makes us more likely to find the global solution of our non-convex optimization problem, or a solution close to it.

In our computations, we verify that each computed flip point is a legitimate image, satisfying appropriate upper and lower bounds for each pixel. We never encountered a case in which these bounds were violated, but violations could be handled by extra constraints or by projection.



Figure 1: Reconstruction of an image from a subset of wavelet coefficients leads to different representations. The original image (left), with 4096 wavelet coefficients, is reconstructed using the most significant 2200, 1000, 500, and 200 wavelet coefficients (respectively, from left to right), chosen according to pivoted QR factorization.

### 4 Investigating the neural network function and the closest flip points

Here, viewing the trained neural network as a function, we investigate the paths between inputs and flip points.

#### 4.1 Lipschitz continuity of the output of trained model

The output of our neural network is a smooth mathematical function. Because it is the composition of a finite set of Lipschitz continuous functions, the output is also Lipschitz continuous. The Lipschitz constant can be bounded using the tuning constants for the error functions and the norms of the matrices applied at each layer.

Why does this matter? As we walk along a direct path connecting one data point to another, the Lipschitz constant can tell us how fine we should discretize that path in order to accurately depict the output of network and identify the locations of decision boundaries. This means that we choose the distance between the discretization points small enough such that the output of network can be considered to change linearly between any consecutive points, with negligible error.

## 4.2 Investigating output of network along direct paths between images

Here, we draw lines between images, discretize those lines, and plot the output of the network along them. Consider two images,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , separated by distance  $d = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$ . The points on the line connecting them may be defined by  $(1 - \alpha)\mathbf{x}_1 + \alpha\mathbf{x}_2$  where  $\alpha$  is a scalar between 0 and 1. This line can be extended beyond  $\mathbf{x}_1$  and  $\mathbf{x}_2$  on either side by choosing  $\alpha < 0$  or  $\alpha > 1$ , respectively.

In Figure 2, zero on the horizontal axis corresponds to image  $\mathbf{x}_1$  and the right boundary corresponds to  $\mathbf{x}_2$ , an image chosen from the same or other class. The lines connecting most pairs of images in the data set resemble the top left plot in Figure 2 in their simplicity; both images are far from the decision boundary, and the line between them crosses the decision boundary once. The other five plots in this figure are hand picked to demonstrate atypical cases. Having multiple boundary crossings is more frequent among the images in the testing set, compared to images in the training set.

Figure 3 shows the output of the model for some lines connecting images to their closest flip points. Notice that the two bottom plots have a much smaller distance scale, and the behavior of the softmax score for correctly and incorrectly classified points is quite similar. These plots clearly show that the decision boundaries in our model are far from linear and a hyperplane would not be able to approximate such boundary surfaces. Fawzi et al. [2018] also have the view that the decision boundaries of deep models are highly curved, but they had not computed exact points on the decision boundaries. Our results confirm their view.

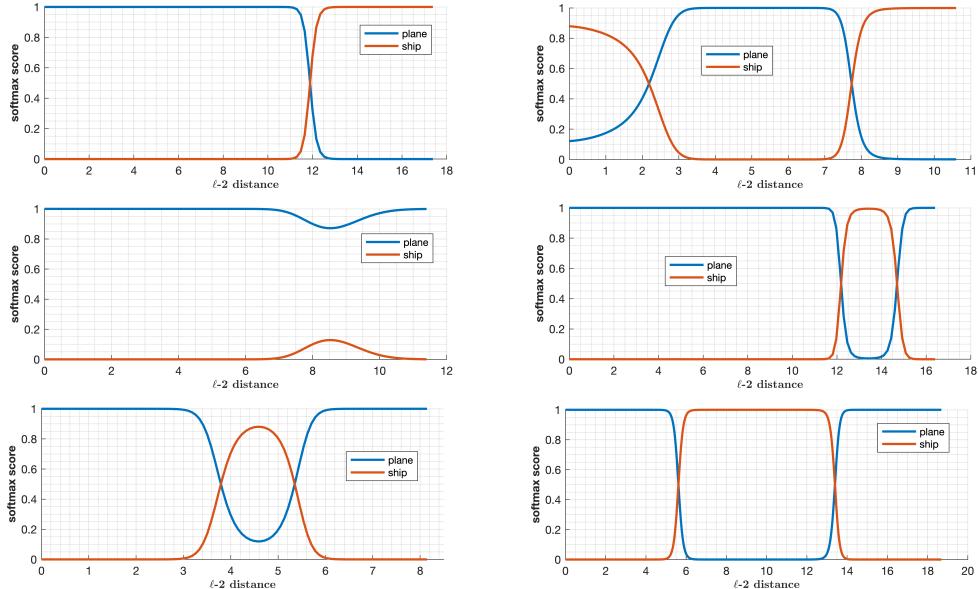


Figure 2: Model output along the line connecting two images.

Figure 4 considers lines connecting various pairs of flip points. If decision boundaries were linear, we would expect the red and blue curves to have softmax scores of 0.5 all along these lines, and that is certainly not what the plots show. If decision boundaries were convex/concave, then we would expect behavior such as that in the upper right plot, but the other three plots show that the true behavior of the decision boundaries is much more complicated.

## 5 Comparing with approximation methods

Here, we compare our calculated minimum distance to the decision boundaries with approximation methods in the literature. We also compare the direction to the closest point on decision boundary with that predicted by first order derivatives. In both comparisons we observe that relying on approximation methods may be misleading.

Regarding the minimum distance to the decision boundaries, Elsayed et al. [2018] suggested estimating the distance using a approximation method based on first order Taylor expansion, building

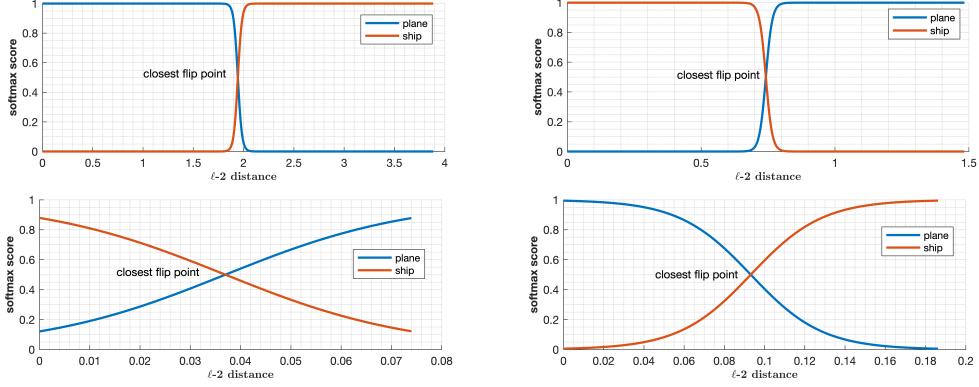


Figure 3: Model output along the line connecting an image with its closest flip point. Images for the top row are correctly classified, while images for the bottom row are misclassified.

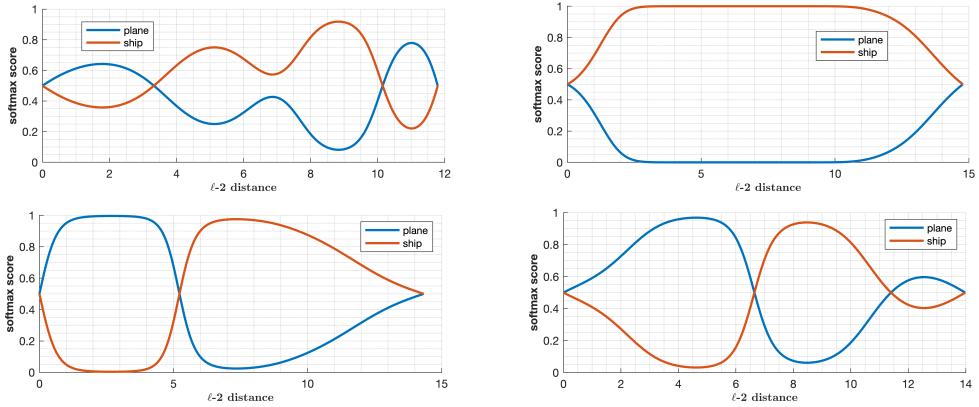


Figure 4: Model output along the line connecting two flip points.

on other suggestions for linear approximation of the distance, e.g., Matyasko and Chau [2017] and Hein and Andriushchenko [2017]. The approximation method of Elsayed et al. [2018] has also been used by Jiang et al. [2019] to study the generalization error of models. Figure 5 shows the distances computed using their approximation method versus the actual distances we have computed using flip points. For distances less than 0.01, the Taylor approximation underestimates the distance by about a factor of 2. For larger distances, the Taylor approximation underestimates by as much as a factor of 20 or more, as shown in Figure 6. We note that their approximation method estimates distance to decision boundaries without finding actual points on the decision boundaries. We find that the estimated distances are generally underestimates of both the true distance in that direction and the true distance to a flip point.

Figure 7 illustrates the distance to the decision boundary along the direction defined by the Taylor series approximation, compared to the distance to the closest flip point. Finding the Taylor direction and then finding the intersection with the decision boundary (a one-dimensional optimization problem) are both very inexpensive operations, and our results indicate that this approach usually gives a good approximation to the closest flip point distance (average of 1.06 times the true distance), but for 7% of the data, a step in that direction goes outside the feasible set of images before passing through a decision boundary. This indicates that it might not be wise to limit the search to the direction of first-order derivatives. And if we limit the search to the direction of first-order derivatives (or any other direction), it would be most reliable to search along that direction for a flip point rather than just estimating the distance. Based on the results in Figures 6 and 7, we can conclude that the distance obtained by searching the direction of first-order derivatives can be considered a much better approximation method compared to the distance obtained by the first-order Taylor series approximation method.

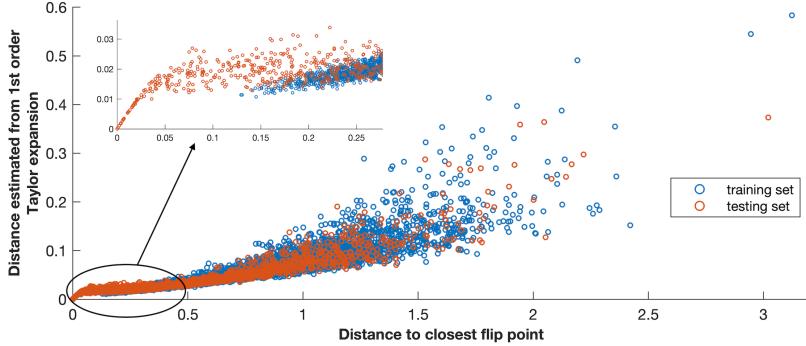


Figure 5: Using the first-order Taylor expansion for estimating the minimum distance to decision boundaries significantly underestimates the distance, except when points are very close to the decision boundaries (closer than 0.01).

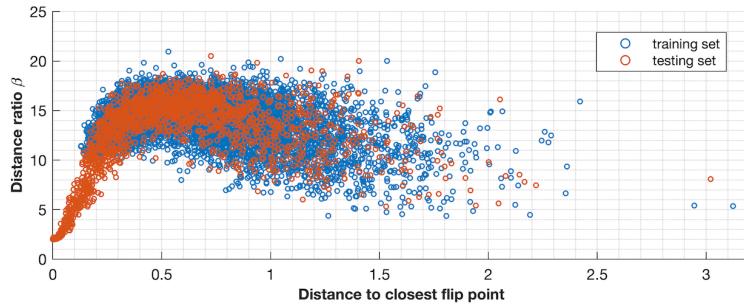


Figure 6:  $\beta$  is the distance to the closest flip point divided by the distance predicted by the Taylor approximation. Since these ratios are far from 1, the approximation is not a reliable measure.

Unfortunately, the Taylor direction itself is not so reliable. We look at the angle between the direction defined by the Taylor approximation and that defined by the calculated closest flip point. Large angle between the two directions means the derivative does not point near the closest point on decision boundary. Figure 8 shows the distribution of the angles (in degrees) vs the distance to the closest flip points. This clearly shows that the farther an image is from the decision boundary, the larger the angle tends to be. In Figure 8, we observe that the lower bound for the angles linearly increases with the distance.

All these observations show that the simplifying assumptions used by Elsayed et al. [2018] and Jiang et al. [2019] can be unreliable, and signify the importance of verification for such simplifying assumptions, whenever used for models with nonlinear activation functions.

## 6 Shape and connectedness of decision regions

We consider all images of correctly-classified ships in the testing set, and investigate the lines (in image space) connecting each pair of images. 89% of those lines stay within the “ship” class for the model, while 11% do not. The least-connected ship is connected to 220 other ships by lines that do not exit the “ship” region, and there are paths (some using multiple lines) that connect every pair of ships without exiting the “ship” region. This indicates that the “ship” region is star-shaped, providing another reason why linear approximations to decision boundaries are inadequate. These observations also hold for images in the training set. Therefore, the trained network has formed a connected sub-region (in the domain) that defines the “ship” region. This result aligns with the observations reported by Fawzi et al. [2018] that classification regions created by a deep neural network can be connected and a single large region may contain all points of the same label. Fawzi et al. [2018], however, did not investigate the output of network along direct paths between images of the same class.

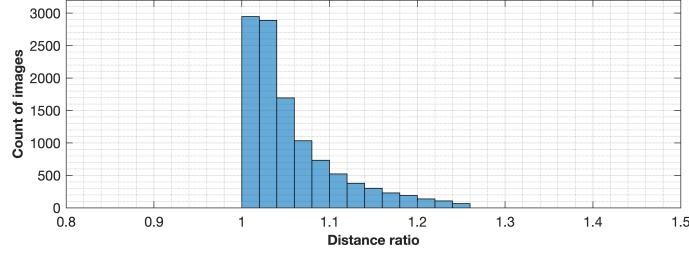


Figure 7: Finding the flip point along the direction indicated by the first-order Taylor expansion often gives an accurate estimate of distance to the decision boundary. The horizontal axis is the ratio of the distance to the decision boundary along the Taylor direction to the distance to the closest flip point.

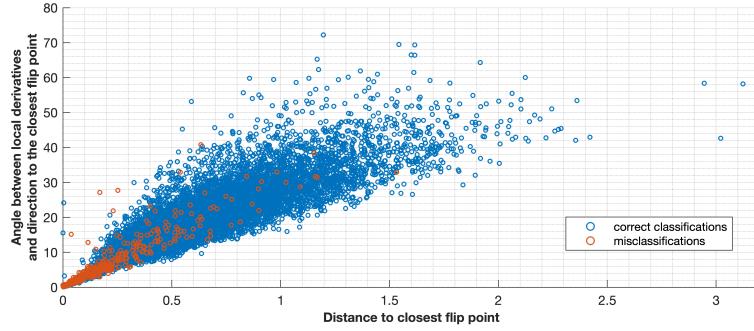


Figure 8: Angle between direction of first-order Taylor approximation and direction to closest flip points. These angles are far from 0, indicating that Taylor approximation gives misleading results.

We performed our analysis by building the adjacency matrix of directly connected images. Performing spectral clustering [Von Luxburg, 2007] on the graph and the Laplacian derived from the adjacency matrix that incorporates the distance between images, may provide additional insights.

## 7 Adversarial examples and decision boundaries

In most recent studies about adversarial examples, the inputs with adversarial label are obtained by minimizing the loss function of the neural network for that label, subject to a distance constraint [Tsipras et al., 2019, Ilyas et al., 2019]. The distance constraint keeps the adversarial image close to the original image. There are limitations to this approach, as we explain via examples.

Consider the image on the left in Figure 9. Figure 9 (right) shows the value of the softmax score on the line from this image to its closest flip point and beyond. We compare this with the result of minimizing the loss function of the model for the adversarial label “plane”, subject to  $\ell_2$  distance constraint of 0.5, as suggested by Ilyas et al. [2019]. The adversarial image obtained by this method is much further away, a distance of 0.494 instead of a distance of 0.178 for the closest flip point that we found. So their calculation underestimates the vulnerability of the model. It is also interesting that the line between the image and the adversarial image found by their method crosses a flip point at a distance much less than 0.494, as shown in Figure 10, yielding a much better assessment of the vulnerability of the model.

There is another difficulty associated with minimizing the loss for an adversarial label subject to a distance constraint. As an example, consider the image in Figure 11, which is at a distance of 2.14 from its closest flip point. Seeking an adversarial image for this image with distance constraint 0.5 will be unsuccessful, as the optimization problem has no feasible solution. Finding the closest flip point yields much better information about robustness.

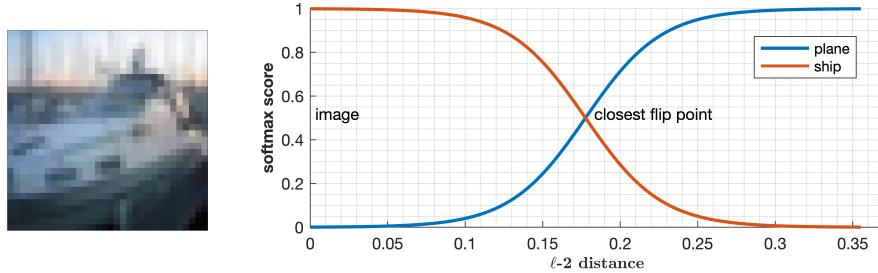


Figure 9: Finding the closest flip point reveals the least changes that would lead to an adversarial label for the image.

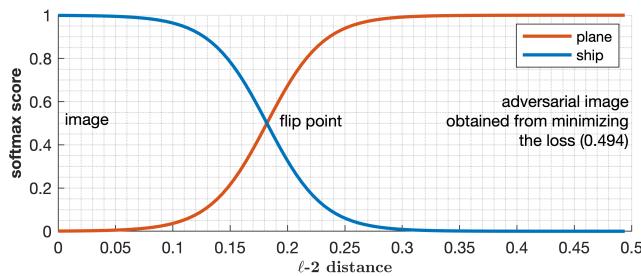


Figure 10: Minimizing the loss function subject to a distance constraint may find adversarial examples far from the original image.

We also measure the angle between the direction to the closest flip point, and the direction to the adversarial example found by minimizing the loss function. For the image in Figure 10, the angle is 12.7 degrees.

The cost of finding a flip point is comparable to the cost of minimizing the loss function, and it provides much better information.

The distance constraint used by Ilyas et al. [2019] can be viewed as a ball around the input. We showed that choosing the size of that ball can be challenging. If the size of ball is too small, their optimization problem becomes infeasible. If the size of ball is large, they do not find the adversarial example closest to the input. Their problem is non-convex, like ours. The examples above demonstrated that our approach finds a closer adversarial example compared to their approach. The computation times for our method and theirs are quite similar.

Moreover, Figure 12 shows that the closest distance to the decision boundary can have a large variation among the images in a dataset. Therefore, tuning the distance constraint for one image may not be insightful for most of the other images in a dataset.

These observations would still hold for networks trained on the pixels rather than wavelet coefficients.

Regarding the reason for excessive vulnerability of trained neural networks towards adversarial examples [Goodfellow et al., 2014], there are studies that speculate about decision boundaries. For example, Tanay and Griffin [2016] argue that “adversarial examples exist when the classification boundary lies close to the submanifold of sampled data”, but their analysis is limited to linear classifiers. Shamir et al. [2019] also explain the adversarial examples via geometric structure of partitions defined by decision boundaries; however, they do not consider the actual distance to the decision boundaries, nor the feasibility of changes in the input space, and their analysis is focused on linear decision boundaries.

The analysis provided in this paper shows that studies focused on adversarial examples can benefit from using the closest flip points and from direct investigation of decision boundaries, for measuring and understanding the vulnerabilities, and for making the models more robust.

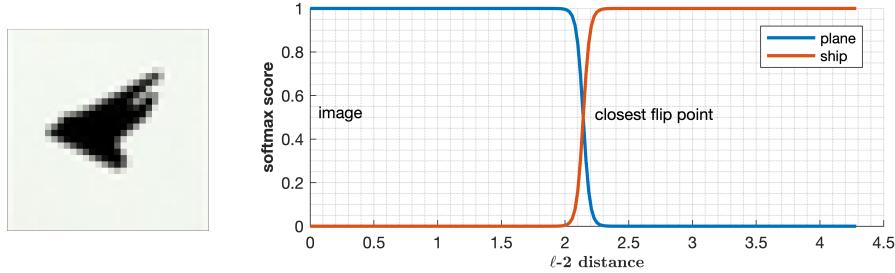


Figure 11: Minimizing the loss function subject to a tight distance constraint may not have a feasible solution and would not reveal how robust the model actually is.

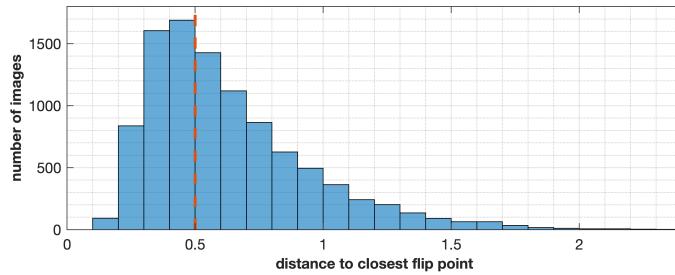


Figure 12: Distance to the closest flip point has large variation among images in the training set, which shows that a single distance constraint would not be able to reveal the vulnerabilities of a model for all images. For example, a distance constraint of 0.5 cannot yield an adversarial example for the large fraction of images that are farther than 0.5 from the decision boundaries. It also would not reveal the weakest vulnerabilities for images which are much closer than 0.5 to the decision boundaries.

## 8 Conclusions

We showed the complexities of decision regions of a model can make linear approximation methods quite unreliable, when nonlinear activation functions are used for the neurons. Instead, we used flip points to provide improved estimates of distance and direction of data points to decision boundaries. These estimates can provide measures of confidence in classifications, explain the smallest change in features that change the decision, and generate adversarial examples. Closest flip points are computed by solving a non-convex optimization problem, but the cost of this is comparable to methods used to compute an adversarial point that may be much further away. The closest flip point along a particular direction can be easily computed by a bisection algorithm. Our example involved only two classes and continuous input data, but we have also implemented our method for problems with multiple classes and discrete features.

## References

- Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, pages 842–852, 2018.
- Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. The robustness of deep networks: A geometrical perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, 2017.
- Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018.
- Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU Press, Baltimore, 4th edition, 2012.

- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pages 2266–2276, 2017.
- Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.
- Saumya Jetley, Nicholas Lord, and Philip Torr. With friends like these, who needs adversaries? In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, pages 10749–10759, 2018.
- Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations (ICLR 2019)*, 2019.
- Steven G. Johnson. The NLopt nonlinear-optimization package, <http://ab-initio.mit.edu/nlopt>, 2014.
- Richard P Lippmann. An introduction to computing with neural nets. *Artificial Neural Networks: Theoretical Concepts*, 4(2):4–22, 1987.
- Alexander Matyasko and Lap-Pui Chau. Margin maximization for robust classification using deep learning. In *International Joint Conference on Neural Networks*, pages 300–307. IEEE, 2017.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.
- Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pages 5947–5956, 2017.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
- Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small Hamming distance. *arXiv preprint arXiv:1901.10861*, 2019.
- Alexander Spangher, Berk Ustun, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning*, 2018.
- Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.
- Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR 2019)*, 2019.
- Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2018.
- Roozbeh Yousefzadeh and Dianne P O’Leary. Interpreting neural networks using flip points. *arXiv preprint arXiv:1903.08789*, 2019.

## Appendix A: Information about neural network used in our numerical examples

Here, we provide more information about the model we have trained and used in the paper. Our model is a fully connected feed-forward neural network with 12 hidden layers. The inputs to the model are 200 wavelet coefficients for any image, as explained earlier. The number of neurons for each layer are shown in Table A1.

Table A1: Number of nodes in neural network used for each data set.

DATA SET	RESTRICTED CIFAR-10
INPUT LAYER	200
LAYER 1	700
LAYER 2	600
LAYER 3	510
LAYER 4	440
LAYER 5	375
LAYER 6	325
LAYER 7	285
LAYER 8	250
LAYER 9	215
LAYER 10	160
LAYER 11	100
LAYER 12	40
OUTPUT LAYER	2

The activation function we have used in the nodes is the error function

$$activation(y|\sigma) = \text{erf}\left(\frac{y}{\sigma}\right) = \frac{1}{\sqrt{\pi}} \int_{-\frac{y}{\sigma}}^{+\frac{y}{\sigma}} e^{-t^2} dt,$$

where  $y$  is the result of applying the weights and bias to the node's inputs. The tuning parameter  $\sigma$  is constant among the nodes on each layer and is optimized during the training process.

We have used softmax on the output layer, and cross entropy for the loss function.