Deep Learning Generalization and the Convex Hull of Training Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work we study the generalization of deep learning functions in relation to the convex hull of their training sets. A trained image classifier basically partitions its domain via decision boundaries and assigns a class to each of those partitions. The location of decision boundaries inside the convex hull of training set can be investigated in relation to the training samples. However, our analysis shows that in standard image classification datasets, a considerable portion of testing images are outside that convex hull. Therefore, the performance of a trained a model partially depends on how its decision boundaries are extended outside the convex hull of the training data. From this perspective, over-parameterization of deep learning models is a necessity for shaping the extension of decision boundaries. To illustrate this, we investigate the decision boundaries of a neural network, with various degrees of over-parameterization, inside and outside the convex hull of its training set. Moreover, we use a polynomial decision boundary to study the necessity of over-parameterization in shaping its extensions outside the convex hull.

1 Introduction

A deep learning image classifier is a mathematical function that maps images to classes, i.e., a deep learning function [Strang, 2019]. These models/functions have shown to be exceptionally useful in real-world applications. However, generalization of these functions, is considered a mystery by deep learning researchers [Arora et al., 2019]. These models have orders of magnitude more parameters than their training samples [Belkin et al., 2019, Neyshabur et al., 2019], and they can achieve perfect accuracy on their training sets, even when the training images are randomly labeled, or the contents of images are replaced with random noise [Zhang et al., 2017]. The training loss function of these models has infinite number of minimizers, where only a small subset of those minimizers generalize well [Neyshabur et al., 2017a]. If one succeeds in picking a good minimizer of training loss, the model can classify the testing images correctly, nevertheless, for any correctly classified image, there are infinite number of images that look the same, but models will classify them incorrectly (phenomenon known as adversarial vulnerability) [Papernot et al., 2016, Shafahi et al., 2019, Tsipras et al., 2019]. Here, we study some geometric properties of the training and testing sets to provide new insights about what a model can learn from its training data, and how it can generalize.

Specifically, we study the convex hulls of image classification datasets (both in the pixel space and in the wavelet space), and show that a considerable portion of testing images fall outside the convex hull of training sets. Therefore, the performance of a trained model partially depends on how well it can extrapolate. We further investigate this extrapolation in relation to the over-parameterization of the networks, using a polynomial decision boundary as an example.

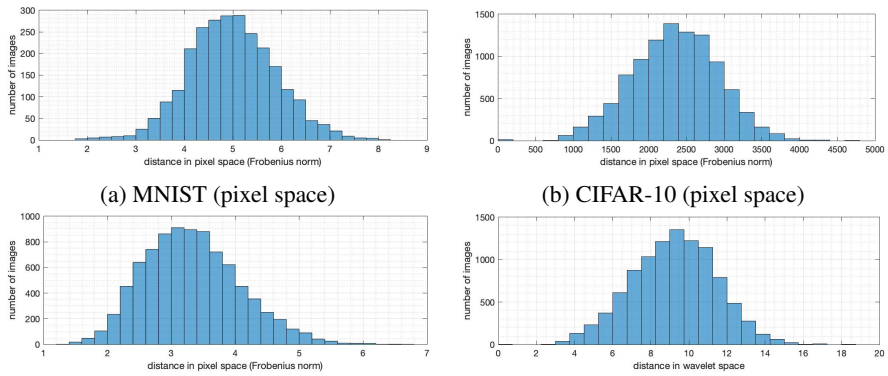
Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

2 Geometry of testing data w.r.t the convex hull of training sets

First, we show that for standard datasets such as MNIST and CIFAR-10, a considerable portion of testing data are outside the convex hull of their training sets. The convex hull of a training set is denoted by $\mathcal{H}^{tr}$.

To verify whether an image/data point is inside its corresponding $\mathcal{H}^{tr}$ or not, we can simply try to fit a hyper-plane separating the point and the training set. If we find such hyper-plane, the point is outside the convex hull, and vice versa. This is basically a linear regression problem and there are many efficient and fast methods to perform it [Goldstein et al., 2015]. For the MNIST dataset, we see that only about 26% of testing data are outside the $\mathcal{H}^{tr}$, in the pixel space. When we transform the images with wavelets (an operation analogous to convolutional neural nets), 87% of testing data fall outside of the $\mathcal{H}^{tr}$. For CIFAR-10, however, more than 99.8% of testing images are outside the $\mathcal{H}^{tr}$, while wavelet transformation increases that percentage to 99.96%.

We can now investigate the testing data outside the $\mathcal{H}^{tr}$, to see how far they are located from it. To approximate the distance of a testing image to the $\mathcal{H}^{tr}$, we measure the distance between that image and the closest image to it in the training set. In high-dimensional space, this approximation is accurate enough for the purpose of our study. Figure 1 shows the histogram of distance to $\mathcal{H}^{tr}$ for the testing images of the above datasets.



(a) MNIST (pixel space) (b) CIFAR-10 (pixel space)

(c) MNIST (wavelet space) (d) CIFAR-10 (wavelet space)

Figure 1: Variations of distance of testing images to $\mathcal{H}^{tr}$.

2.1 Related work about geometry of data and deep learning

Convex hulls of training sets are rarely considered in deep learning studies. Recently, Yousefzadeh and Huang [2020] reported that in the wavelet space, distance of the testing images to the convex hull for each training class can predict the label for more than 97% of MNIST testing data. Previously, Haffner [2002] considered the convex hull of MNIST data for Support Vector Machines. Similarly, Vincent and Bengio [2002] considered the convex hulls for K-Nearest Neighbor (KNN) algorithms. However, those methods do not generalize to deep learning functions.

At the same time, some researchers have studied other geometrical aspects of deep learning models, e.g., [Cohen et al., 2020, Fawzi et al., 2018, Cooper, 2018, Kanbak et al., 2018, Neyshabur et al., 2017b]. To our knowledge, those studies do not investigate the generalization of deep neural networks in relation to the convex hull of training sets.

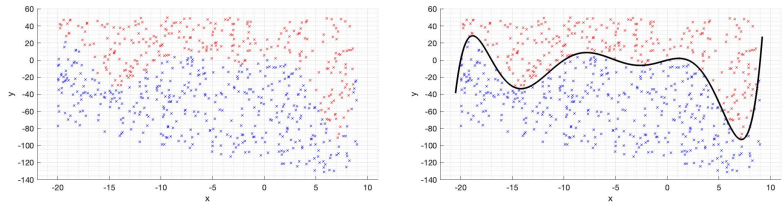
3 Learning outside the convex hull: A polynomial decision boundary

In the previous section, we showed that a considerable portion of testing data are outside the convex hull of training set, while the distance to the $\mathcal{H}^{tr}$ resembles a normal distribution. Hence, a trained deep learning model somehow manages to define its decision boundaries accurately enough outside the boundaries of what it has observed during training. But how does a model achieve that, or more precisely, how do we manage to train a model such that its decision boundaries have our desirable form outside the $\mathcal{H}^{tr}$?

Let's use a polynomial fitting problem as an example to gain some intuitive insights.[1] Figure 2a shows two point sets colored in blue and red, each set belonging to a class. These sets are non-linearly separable, because they have no overlap. If we use the polynomial

$$y = 10^{-5}(x+20)(x+17)(x+10)(x+5)(x)(x-2)(x-9),$$

as our decision boundary, we will achieve perfect accuracy in separating these two sets, as shown in Figure 2b.



(a) (b)

Figure 2: **(a)** Training data with 2 classes, colored with blue and red. **(b)** Non-linear separation of 2 classes with a polynomial of degree 7.

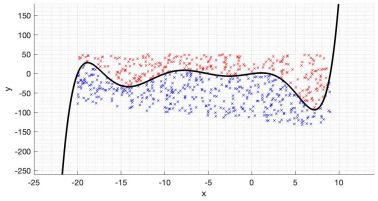


Figure 3: Extension of the polynomial in Figure 2, outside the convex hull of its training data.

Now that we have obtained this polynomial, i.e., decision boundary, how does it generalize to unseen data? Figure 3 shows the extension of this polynomial outside its $\mathcal{H}^{tr}$. If our polynomial can correctly separate and label our testing data, we would say that our polynomial is generalizing well, and vice versa. But, what is reasonable to expect from the testing data? In what regions of the domain should we be hopeful that our polynomial can generalize? What if the domain is much larger than the $\mathcal{H}^{tr}$? Is the extension of our polynomial on both sides reasonable enough?

Clearly, the answer to the above questions can be different inside and outside the $\mathcal{H}^{tr}$. Inside the $\mathcal{H}^{tr}$, if the unseen data has a similar label distribution as the training set, we can be hopeful that our decision boundary will generalize well. However, outside the $\mathcal{H}^{tr}$ is uncharted territory and hence, there will be less hope/confidence about the generalization of our decision boundary, especially when we go far outside the $\mathcal{H}^{tr}$.

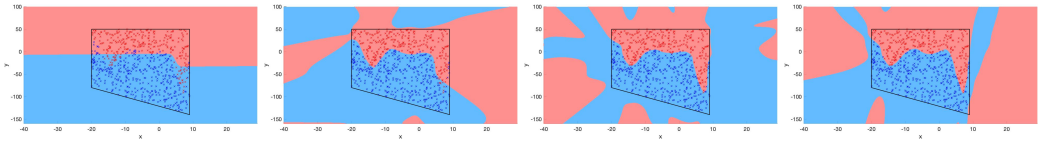
Now in Figure 3, imagine that from some prior knowledge, we expect the classes to be separable outside the convex hull, according to a horizontal line approximable by $y = 0$. Then, our current decision boundary will generalize poorly outside the $\mathcal{H}^{tr}$, because it almost has a vertical shape.

How can we incorporate that prior knowledge into our decision boundary so that it can generalize well both inside and outside the $\mathcal{H}^{tr}$? How can we change its shape outside the $\mathcal{H}^{tr}$, while maintaining its current shape inside the $\mathcal{H}^{tr}$? Clearly, we should add to the degree of our polynomial, or in other words, we should over-parameterize it. The necessity of over-parameterization for achieving that goal can be rigorously shown using the orthogonal system of Legendre polynomials [Ascher and Greif, 2011].

[1] This choice seem appropriate since many recent studies on generalization of deep learning consider the regression models that interpolate, e.g. [Belkin et al., 2018b,a, 2019, Liang et al., 2020, Verma et al., 2019], but those studies do not consider the convex hull of training sets.

4 Output of deep learning functions outside their $\mathcal{H}^{tr}$

In this section, we investigate a 2-layer neural network with ReLU activation. We train the model with various number of neurons on the data we investigate in previous section, depicted in Figure 2b. We then investigate the output of the trained models inside and outside of the $\mathcal{H}^{tr}$, as shown in Figure 4.



(a) model with 2 neurons (b) model with 10 neurons (c) model with 20 neurons (d) model with 20 neurons

Figure 4: Variations of model output inside and outside the $\mathcal{H}^{tr}$, as we change the number of neurons in the model. The black trapezoid depicts the $\mathcal{H}^{tr}$. The colors red and blue correspond to our 2 classes, described earlier. The models (c) and (d) both have 20 neurons and achieve perfect accuracy on the training set, but they are trained using different hyper-parameters and as a result their decision boundaries outside the convex hull are completely different. Models with 2 and 10 neurons do not achieve perfect accuracy inside the convex hull.

We observe that when the model is under-parameterized (e.g., model with 2 neurons), the training regime does not have a significant effect on the resulting model. The model output also follows a relatively simple pattern inside and outside the $\mathcal{H}^{tr}$. For an under-parameterized model, the training loss function has limited number of minimizers, none of which lead to zero loss. Finding the same minimizer of training loss is equivalent to obtaining the same trained model, hence unlike the over-parameterized setting, the training regime is focused on finding the best shape for the decision boundary inside the convex hull.

When the model is over-parameterized, however, we have infinite number of parameter configurations that minimize the training loss to zero, which is equivalent to developing a decision boundary that perfectly separates our two sets of data points. So, forming the decision boundary inside the convex hull is easily achievable. What is different about those infinite number of models is the extension of their decision boundaries outside the convex hull of the data.

This seems to explain why we need over-parameterized models for deep learning and also explain why the generalization of deep learning models are so susceptible to different training regimes: 1- generalization of these models relies on how they extrapolate, 2- to desirably shape the extension of decision boundaries, we over-parameterize the models and then minimize their loss in a particular way.[2]

5 Conclusion and future work

We show that a considerable portion of testing data for standard image classification models lie outside the convex hull of training sets, both in pixel space and in wavelet space. Therefore, the generalization of a deep network partially relies on its capability to extrapolate outside the boundaries of the data it has seen during training. Base on this observation, the significant number of studies that focus on interpolation regimes in order to explain the generalization of deep networks seem to be inadequate to explain the extrapolation requirement.

From this perspective, over-parameterization of models is a necessity to desirably form the extension of decision boundaries outside the convex hull of data. This can be proven for polynomial regression models using the orthogonal system of Legendre polynomials. Moreover, we showed that the training regime can significantly affect the shape of decision boundaries outside the convex hulls, affecting the accuracy of a model in its extrapolation. We investigated a polynomial decision boundary to demonstrate these phenomena.

[2]This particular way of minimizing trining loss implies the massive literature that aim to find the best training regime for achieving the best testing accuracy. Much of that literature has relied on the knowledge of testing sets in order to develop those training methods, shaping the extension of decision boundaries desirably.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

Uri M Ascher and Chen Greif. *A first course on numerical methods*. SIAM, 2011.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pages 2300–2311, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018b.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13, 2020.

Yaim Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

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.

Tom Goldstein, Min Li, and Xiaoming Yuan. Adaptive primal-dual splitting methods for statistical learning and image processing. In *Advances in Neural Information Processing Systems*, pages 2089–2097, 2015.

Patrick Haffner. Escaping the convex hull with extrapolated vector machines. In *Advances in Neural Information Processing Systems*, pages 753–760, 2002.

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017a.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017b.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019.

Gilbert Strang. *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

Pascal Vincent and Yoshua Bengio. K-local hyperplane and convex distance nearest neighbor algorithms. In *Advances in neural information processing systems*, pages 985–992, 2002.

Roozbeh Yousefzadeh and Furong Huang. Using wavelets and spectral methods to study patterns in image-classification datasets. *arXiv preprint arXiv:2006.09879*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.