Adversarial Robustness Through Overparameterization

AJ Barry

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

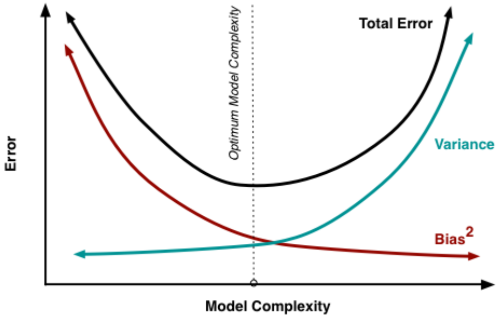
Outline of the Talk

Table of Contents

Classical Statistics vs Modern ML

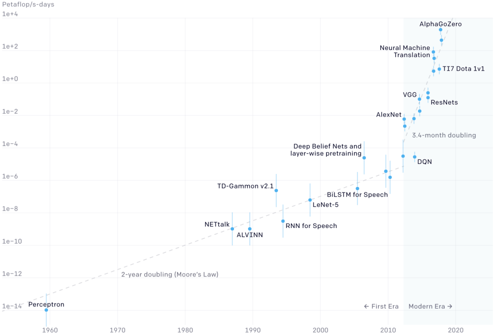
Idea (Classical Statistics)

- Choose number of parameters which minimize MSE (or some approximation thereof).
- There is a clear trade off between model bias and variance.



Classical Statistics vs Modern ML

- Recently there has been an explosion in model size.
- Adding many more parameters $p \gg n$ appears to yield results which outperform traditional methods in empirical work.



[8]

Classical Statistics vs Modern ML

Idea (Modern ML)

- Speculation is that "benign overfitting" is OK when both n, d are large.
- Observed "Double Descent" structure of Test/Training error.[3][2][8]

[8]

Neural Networks in Practice

Theorem

[1]For a number of parameters p and n data points, if $p \geq n$ the model can perfectly memorize the data. Formally, $f(x_i) = y_i \quad \forall\, (x_i, y_i)$ in the training set.

Most cutting edge models have many more parameters than the number of data points.

- MNIST dataset, $n \approx 6 * 10^4$ images, models have $p \approx 10^6$ parameters.
- Imagenet dataset, $n \approx 10^6$ images, models have $p \approx 10^9$ parameters.
- GPT-3, $p \approx 2 * 10^{12}$.

Robustness of Neural Networks

- Neural Networks learn input-output mappings that may be fairly discontinuous. [9]
- While the models generalize well to test data they are highly susceptible to "adversarial attacks".

Adversarial Attacks

These adversarial attacks are small nonrandom perturbations of the data.

Definition (Fast Gradient Sign Adversary)

[6] For x being an input to a NN, y being the target associated with x, θ being the parameters of the model and L being the cost function used to train the NN. The *Fast Gradient Sign Adversary* is:

$$x_{\text{adv}} = x + \epsilon \,\mathrm{sgn}(\nabla_x L(x, \theta, y))$$

What Do We Mean by Robustness?

Consequently in order to ensure robustness against adversarial attacks it is natural to want our output function f to satisfy the following property:

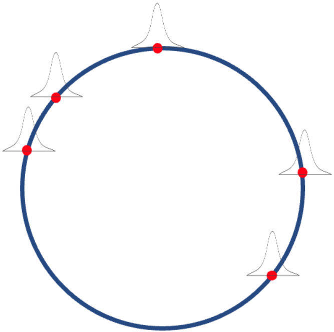
Definition (L-Lipschitz Function)

A function f is $L - Lipschitz$ if for all x, y:

$$\|f(x) - f(y)\| \leq L\|x - y\|_*$$

What about Smoothness?

- While we can memorize data with only n parameters, these constructions will have $\mathrm{Lip}(f) = \Omega(\sqrt{d})$ even for well dispersed data (ex. Uniform on the unit sphere).
- In principle one can memorize data with $\mathrm{Lip}(f) = O(1)$ but will require $p \approx nd$ parameters (sum of bumps construction).



[8]

Theorem

[5][Universal Law of Robustness] Extreme overparametrization (nd parameters) is necessary for robust Neural Networks.

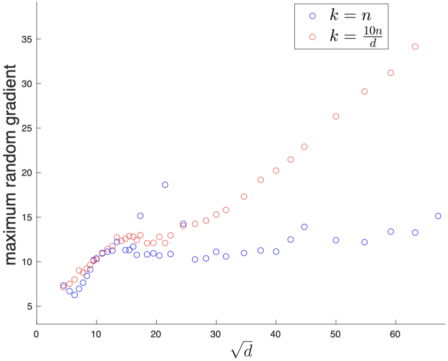
Table of Contents

Setup (Random Data)

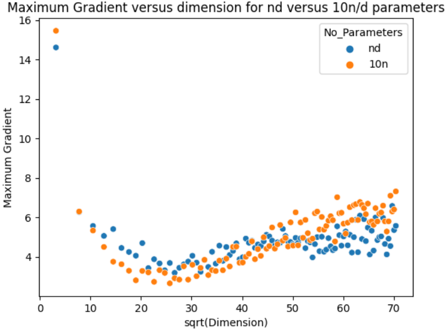
We aim to replicate the findings of [4][5.2] in investigating the case of $p = nd$ and $p = 10n$. We fix $n = 10^4$ and generate random data from an $x_i \overset{iid}{\sim} N(0, \frac{1}{d} I_d)$, and labels $y_i \overset{iid}{\sim} U(\{\pm 1\}$. We will sweep values of $d \in [10, 5000]$ by 50.

- Train a neural network using nd parameters $(f_{nd}(x))$, and one that has n parameters $f_n(x)$. (using the adam optimizer, and least squares loss, $\epsilon = 0.1$ for thresholding)
- Compute the maximum random gradient by generating 1000 random samples $z_j \overset{iid}{\sim} N(0, \frac{1}{d} I_d)$ and computing $\max_j \|\nabla f(z_j)\|_2$ for each NN.

Results (Random Data)



Results from [4]



My results

Table of Contents

Mnist Experiment

Goal: Examine the sufficiency of overparameterization for robustness.

The Data Set

MNIST is a data set containing black and white images of hand written digits. There are $n = 60,000$ training images and $10,000$ test images each with a corresponding label. The images are 28×28 and we will normalize pixel values to be between $[0, 1]$.

Our Models

We train two models, one with $p = 120,000$ parameters, and the other with $p = 3 * 10^6$ parameters.

- For MNIST, $d = 748$. However, effective dimension is estimated to be on the order of $[5, 20]$.
- If Universal Law of Robustness held only to true dimension we would need $47 * 10^6$ parameters for a Lipschitz model.

We will then test the two models against both a white noise attack, and FGSM (1.4).

How Models Were Trained

- Models were simple 3-layer networks with the hidden layer having ReLU activations and the output layer having a Softmax activation function.
- Models were trained until loss was 0 and were using the Adam optimizer, and categorical cross entropy loss function.

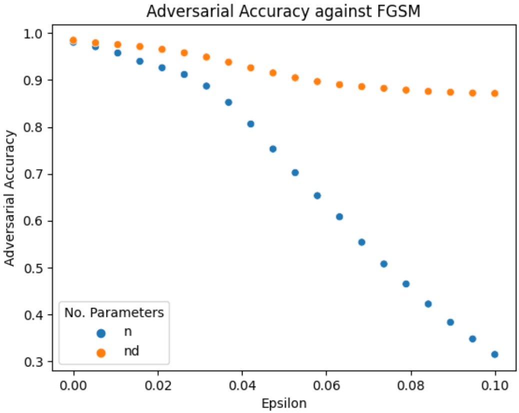
Clean Data Results

Let f_{nd} denote the model with $3 * 10^6$ parameters, and f_n denote the model with $120,000$ parameters.

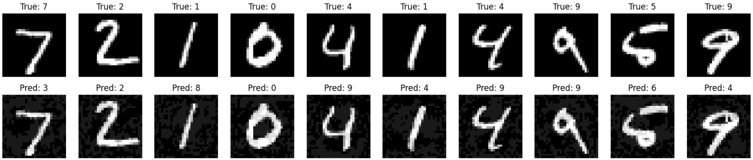
Model	**Test Accuracy**	**Test Loss**
f_{nd}	0.985	0.587
f_n	0.980	0.211

FGSM

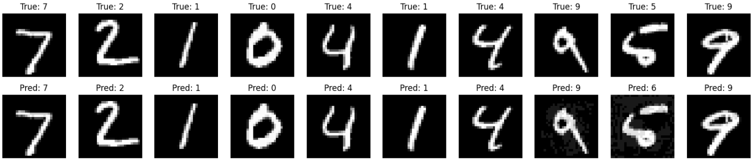
Tested both models against FGSM adversary for twenty equally spaced values of $\epsilon \in [0, 0.1]$.



Example FGSM Adversaries for f_n using $\epsilon = 0.1$.

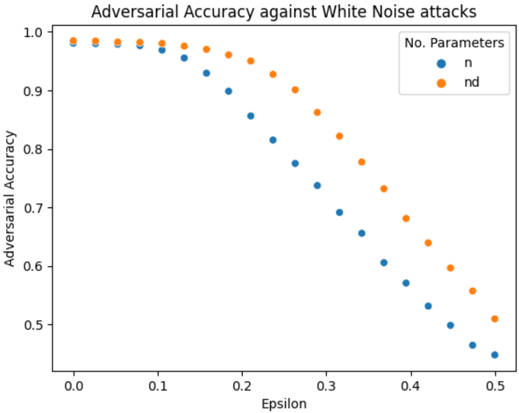


Example FGSM Adversaries for f_{nd} using $\epsilon = 0.1$.



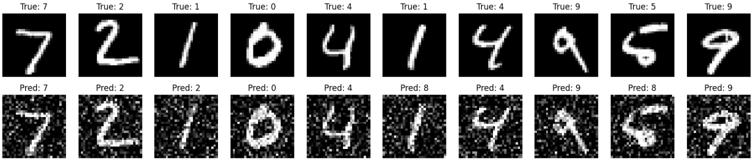
White Noise

Tested both models against simple Gaussian white noise attacks for twenty equally spaced values of $\epsilon \in [0, 0.5]$ (where ϵ is the standard deviation of the noise).



WN-Visualization, n

Example FGSM Adversaries for f_n using $\epsilon = 0.3$.



WN-Visualization, nd

Example FGSM Adversaries for f_{nd} using $\epsilon = 0.3$.

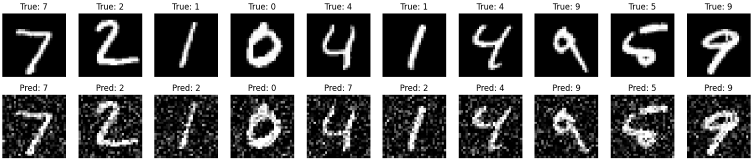


Table of Contents

[1] Eric B Baum. "On the capabilities of multilayer perceptrons". en. In: *Journal of Complexity* 4.3 (Sept. 1988), pp. 193–215. DOI: 10.1016/0885-064X(88)90020-9.

[2] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. *To understand deep learning we need to understand kernel learning*. Tech. rep. arXiv:1802.01396. arXiv:1802.01396 [cs, stat] type: article. arXiv, June 2018.

[3] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". en. In: *Proceedings of the National Academy of Sciences* 116.32 (Aug. 2019), pp. 15849–15854. DOI: 10.1073/pnas.1903070116.

[4] Sébastien Bubeck, Yuanzhi Li, and Dheeraj Nagaraj. *A law of robustness for two-layers neural networks*. Tech. rep. arXiv:2009.14444. arXiv:2009.14444 [cs, stat] type: article. arXiv, Nov. 2020.

[5] Sébastien Bubeck and Mark Sellke. *A Universal Law of Robustness via Isoperimetry*. Tech. rep. arXiv:2105.12806. arXiv:2105.12806 [cs, stat] type: article. arXiv, Dec. 2022.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. Tech. rep. arXiv:1412.6572. arXiv:1412.6572 [cs, stat] type: article. arXiv, Mar. 2015.

[7] Song Mei and Andrea Montanari. *The generalization error of random features regression: Precise asymptotics and double descent curve*. Tech. rep. arXiv:1908.05355. arXiv:1908.05355 [math, stat] type: article. arXiv, Dec. 2020.

[8] Preetum Nakkiran et al. *Deep Double Descent: Where Bigger Models and More Data Hurt*. Tech. rep. arXiv:1912.02292. arXiv:1912.02292 [cs, stat] type: article. arXiv, Dec. 2019.

[9] Christian Szegedy et al. *Intriguing properties of neural networks*. Tech. rep. arXiv:1312.6199. arXiv:1312.6199 [cs] type: article. arXiv, Feb. 2014.