A Universal Law of Robustness via Isoperimetry

Sébastien Bubeck & Mark Sellke

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

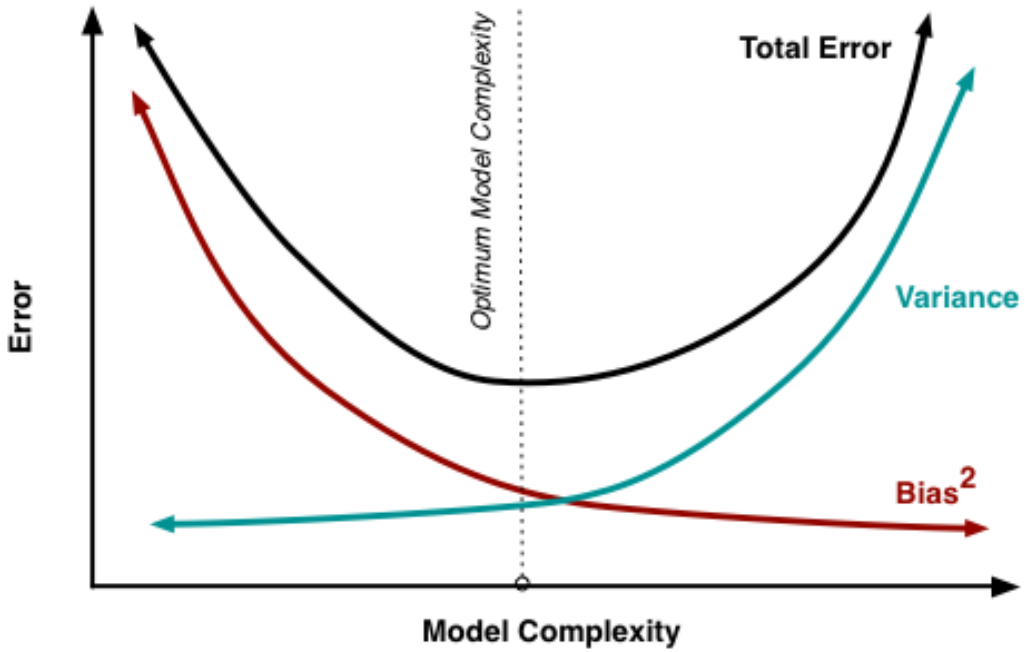
Outline of the Talk

Table of Contents

Classical Statistics vs Modern ML

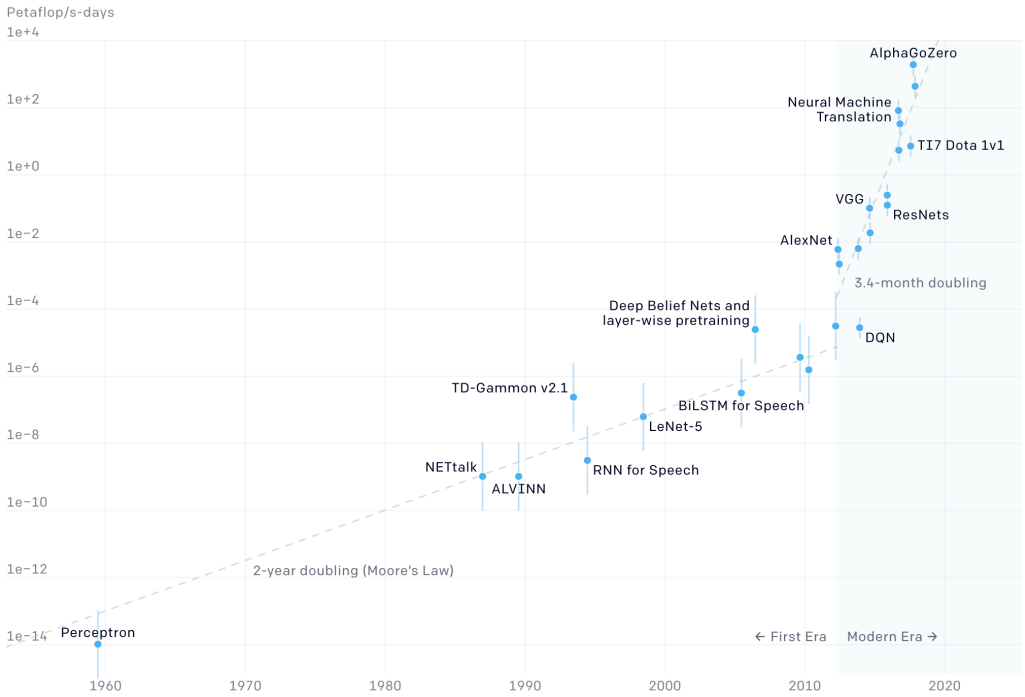
Idea (Classical Statistics)

- 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:

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

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)

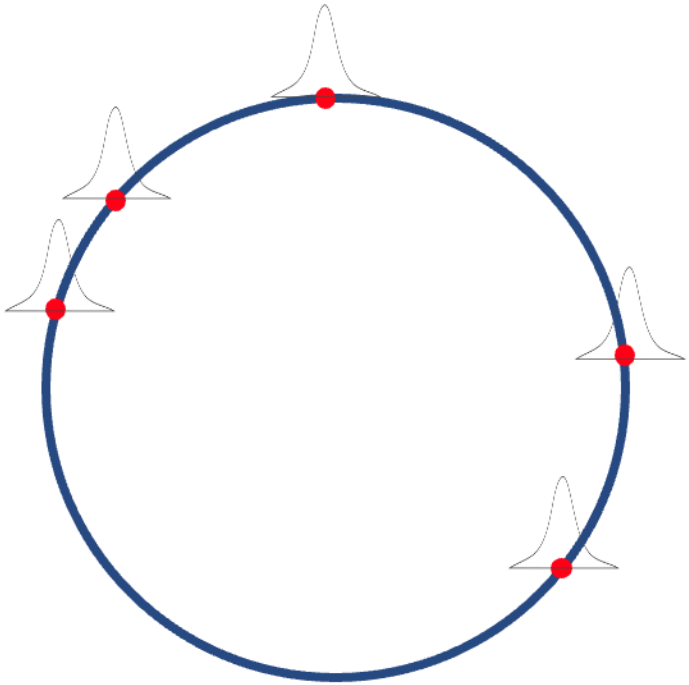
A function f is $L - Lipschitz$ if for all x, y:

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

This is in fact how Bubeck & Sellke define robustness for the paper.

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]

Overparametrization and Robustness

Claim

[5] Extreme overparametrization (*nd* parameters) is *necessary* for robust Neural Networks.

Table of Contents

Neural Networks

Definition

A neural Network of depth D can be written as:

$$f(x) = W_D \circ \alpha \circ W_{D-1} \circ \cdots \circ \alpha \circ W_1(x)$$

Where $\alpha(t)$ is an activation function applied entry wise.

- For our purposes, let w be a parameter vector of length p which encodes all entries in all of the W_i.
- Lipschitz constants in w and x are upper bounded by $\prod_{i=1}^{D} \|W_i\|_{op}$.
  - At worst our Lipschitz constant L is logarithmic in everything except D.
  - This product is often explicitly controlled to ensure training stability.

Isoperimetry

Definition

[5] A probability measure μ on $\mathbb{R}^d$ satisfies *c-Isoperimetry* if for any bounded L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, and any $t \geq 0$:

$$\mathbb{P}[|f(x) - \mathbb{E}[f]| \geq t] \leq 2e^{-\frac{dt^2}{2cL^2}}$$

Examples include:

- Uniform on the sphere.
- Gaussian (plus small independent noise).
- Mixtures of isoperimetric distributions.

Isoperimetry

Recall: If a scalar random variable X satisfies $\mathbb{P}[|X| \geq t] \leq 2e^{-t^2/c}$, then it is c-Subgaussian. If we have isoperimetry, this implies that the output of any Lipschitz function is subgaussian under suitable rescaling.

Table of Contents

The Theorem

Theorem

[5] Let $\mathcal{F}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i \in [n]}$ be i.i.d. input output pairs in $\mathbb{R}^n \times [-1, 1]$.

The Theorem

Theorem

[5] Let $\mathcal{F}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i\in[n]}$ be i.i.d. input output pairs in $\mathbb{R}^n \times [-1, 1]$. If:

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most poly(n, d).*

The Theorem

Theorem

[5] Let $\mathcal{F}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i\in[n]}$ be i.i.d. input output pairs in $\mathbb{R}^n \times [-1, 1]$. If:

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most poly(n, d).*
2. *The distribution μ of the covariates x_i satisfies Isoperimetry (or some mixture thereof).*

The Theorem

Theorem

[5] Let $\mathcal{F}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i\in[n]}$ be i.i.d. input output pairs in $\mathbb{R}^n \times [-1, 1]$. If:

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most poly(n, d).*
2. *The distribution μ of the covariates x_i satisfies Isoperimetry (or some mixture thereof).*
3. *The expected conditional variance of the output is strictly positive. Denoted $\sigma^2 \equiv \mathbb{E}_\mu[Var(y|x)] > 0$.*

Then w.h.p. over the sampling data, for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \leq \sigma^2 - \epsilon \Rightarrow Lip(f) \geq \tilde{\Omega}\left(\frac{\epsilon}{\sigma}\sqrt{\frac{nd}{p}}\right)$$

Notes

- If f is a perfect memorizer, then for $\text{Lip}(f) = O(1)$ (doesn't scale w.r.t. n, d) we require at least nd parameters.
- Distribution of x_i's can be a mixture of $o(\frac{n}{\log(n)})$ isoperimetric distributions.
- Heteroscedastic label noise is acceptable.

Simplified Proof

- Fix some function class $\mathcal{F}$ described by p parameters of size at most poly(n, d).
- Consider $f \in \mathcal{F}$. Let $y_i \sim$ i.i.d. ± 1.
- Via Isoperimetry for any arbitrary $f \in F$:

$$\min(\mathbb{P}_\mu[f(x) = 1], \mathbb{P}_\mu[f(x) = -1]) \leq \exp\left(\frac{-d}{\text{Lip}(f)^2}\right)$$

- This implies that the probability that f fits all n data points is $\leq \exp\left(\frac{-nd}{\text{Lip}(f)^2}\right)$.

Simplified Proof - Continued

- From here, because $\mathcal{F}$ is a family of L-Lipschitz functions we can use Boole's inequality over the function class to show that:

$$\mathbb{P}[\cup_{f\in\mathcal{F}}(f(x_i)=y_i\ \forall i)] \leq \exp\left(\log(|\mathcal{F}|) - \frac{-nd}{L^2}\right)$$

- If $L \ll \sqrt{\frac{nd}{\log(|\mathcal{F}|)}}$ the probability of finding a function $f \in \mathcal{F}$ is vanishingly small.

Simplified Proof - Continued

- If $\mathcal{F}$ admits a parametrization by p (polynomially bounded) parameters, then we can approximate $\mathcal{F}$ by $\exp(p)$ size discretization. So $|\mathcal{F}|$ is exponential in p.
- So for Lipschitz memorization to be possible w.h.p, we need $\text{Lip}(f) \gtrsim \sqrt{\frac{nd}{p}}$.

Generalization

- Generalizes to mixtures of isoperimetric distributions.
- Partial Memorization:

$$\sum_{i=1}^{n}(f(x_i)-y_i)^2 \leq \sum_{i=1}^{n} Z_i^2$$

Implications

- Begins to provide theoretical justification for the scale of modern Neural Networks.
- Optimistically holds with respect to lower effective dimension (Ex: MNIST).
- Implies that ImageNet models may need more parameters to be adversarially robust.
- Lipschitz is essentially the strongest notion of robustness against adversarial attacks. Perhaps better bounds with some weaker definition.
- Model distillation would not work to ensure robustness.

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.