A Universal Law of Robustness via Isoperimetry

Sébastien Bubeck (Microsoft Research Redmond), Mark Sellke (Stanford University). NeurIPS 2021. https://arxiv.org/abs/2105.12806

① Motivation

- "Fact" 1: neural networks memorize training data to zero error.
- "Fact" 2: overparametrized models are better for robustness

What's going on? Are these related?

② A Model for Memorization

- Input: $n = d^{O(1)}$ random points $x_1, \dots x_n$ on unit sphere $\mathbb{S}^d$.
  Labels $y_i = g(x_i) + Z_i$: signal + centered noise.
  Noise level $\mathbb{E}[Var[y_i \mid x_i]] = \sigma^2$
- Partial memorization: fit data much better than the signal:

$$\sum_i (f(x_i) - y_i)^2 \leq \frac{1}{2} \sum_i Z_i^2.$$

- Robust classifier: $f: \mathbb{R}^d \to \mathbb{R}$ is $O(1)$-Lipschitz.

③ Memorizing with Parametrized Function Classes

- If some f ∈ $\mathcal{F}$ (robustly) memorizes, how large must the function class $\mathcal{F}$ be?
- Measure size by # parameters P. Formally, $w \to f_w \in \mathcal{F}$ for $w \in \mathbb{R}^P$ with

$$|w| \leq poly(d), \quad |f_w(x) - f_v(x)| \leq poly(d) \cdot |w - v|.$$

- Captures "true" parameter count for convolutional nets, weight sharing, …
- $P = n$ parameters suffice to memorize
  - [Baum 1988]: use a 2-layer neural network with n/d neurons. Not robust.
  - Intuition: fitting n data-points ≈ solving n equations, requires n unknowns.
- $P = nd$ parameters suffice to robustly memorize.
  - Use 1 radial basis function for each of n inputs ⇒ nd parameters.

④ The Law of Robustness

- Conjecture [Bubeck-Li-Nagaraj 20]: $L \geq \sqrt{\frac{nd}{P}}$ for 2-layer neural networks.
- Theorem [Bubeck-S. 21]: for P -parameter function classes $\mathcal{F}$, if there exists f ∈ $\mathcal{F}$ partially memorizing the noisy data, then (w.h.p.):

$$Lip(f) \gg \sigma^2 \sqrt{\frac{nd}{P}}.$$

  - Input distribution can be mixture of $n^{0.99}$ isoperimetric components.
- Tight for any P: project to dimension $\tilde{d} = P/n$, use RBF construction in $\mathbb{R}^{\tilde{d}}$.
- Definition: μ is isoperimetric if Lipschitz functions have sub-Gaussian tail on μ.
  - Typical when μ is "genuinely high-dimensional". Spheres, Gaussians, …

⑤ Proof for Perfect Memorization with 1 Component + Pure Noise

- Claim: if labels y_i are IID ± 1, then robust memorization needs $P \geq nd$.
- Assume *balanced labels*: $\# y_i = 1$ in $\left[\frac{n}{3}, \frac{2n}{3}\right]$. $\mathbb{P}[false] \leq \exp(-n)$.
- Fix an $f \in \mathcal{F}$. Isoperimetry implies:

$$\min(\mathbb{P}^\mu[f(x) = 1], \mathbb{P}^\mu[f(x) = -1]) \leq \exp(-\Omega(d)).$$

$$\Rightarrow \mathbb{P}[f \text{ outputs unlikely label on} \geq \tfrac{n}{3} \text{ of } x_1, \dots, x_n] \leq \exp(-nd).$$

$$\Rightarrow \mathbb{P}[f \text{ fits all (or even most) labels}] \leq \exp(-nd).$$

- Union bound over f ∈ $\mathcal{F}$ ⇒ $|\mathcal{F}| \geq \exp(nd)$.
- P parameters ⇒ discretization of $\mathcal{F}$ has size ≈ $\exp(P) \geq \exp(nd)$. ∎

- Mixtures: assume balanced labels in each component.
- Some further results: generalization perspective, construction showing polynomially bounded parameters necessary even for depth 3 networks.