Uniform Convergence of Interpolators: Gaussian width, Norm Bounds and Benign Overfitting

Frederic Koehler, *Simons Institute*
Lijia Zhou, *University of Chicago*
Danica J. Sutherland, *UBC + Amii*
Nathan Srebro, *TTI-Chicago*

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

- The phenomenon of **Interpolation learning** - achieving low population error while training error is exactly zero in a noisy, non-realizable setting, is one of the core mysteries in deep learning
- **Uniform convergence** is the fundamental technique used in learning theory:

$$L(\hat{w}) \leq \hat{L}(\hat{w}) + \sup_{w \in \mathscr{K}} |L(w) - \hat{L}(\hat{w})|$$

But it does not seem to be tight enough for benign overfitting.

- We consider a different, yet standard notion: the **uniform convergence of interpolators**

$$L(\hat{w}) \leq \sup_{w \in \mathscr{K}, \hat{L}(w)=0} L(w)$$

In this work, we analyze the above gap for high dimensional linear regression with Gaussian data and arbitrary data covariance

Summary of results

- We prove a generic bound in terms of an arbitrary class $\mathscr{K}$'s Gaussian width
- Taking $\mathscr{K}$ to be the Euclidean norm ball and combining with an analysis of the minimal norm required to perfectly fit the data, we recover the consistency result of Bartlett et al. (2020) for the minimal ℓ_2 norm interpolator
- We extend this to any norm and prove novel result for the minimal ℓ_1 norm (basis pursuit)

Setting: Linear Regression

We assume the data (X, Y) is generated as

$$X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\Sigma), \quad \xi \sim \mathcal{N}(0,\sigma^2 I_n), \quad Y = Xw^* + \xi$$

where X_i are the rows of $X \in \mathbb{R}^{n \times d}$ and ξ, X are independent. When $d > n$, there exists w such that

$$Xw = Y \implies \hat{L}(w) = \frac{1}{n}\|Y - Xw\|_2^2 = 0$$

and we can consider the minimal norm interpolator

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d : \hat{L}(w)=0} \|w\|$$

Benign overfitting: in many cases, it holds that $L(\hat{w}) \to \sigma^2$. Bartlett et al. (2020) covers the case when $\|\cdot\|$ is the Euclidean norm.

Main result

Gaussian width: natural measure of "complexity" of a set, long used in generalization theory (e.g. [Bartlett-Mendelson, 2002])

$$W(\mathscr{K}) = \mathbb{E}_{H \sim N(0,I_d)} \left[\sup_{w \in \mathscr{K}} |\langle H, w \rangle| \right]$$

Theorem (Informal): for any covariance matrix Σ, for any splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that $\mathrm{rank}(\Sigma_1) = o(n)$, it holds with high probability that

$$\sup_{w \in \mathscr{K}, \hat{L}(w)=0} L(w) \leq (1 + o(1)) \cdot \frac{W(\Sigma_2^{1/2}\mathscr{K})^2}{n}$$

ℓ_2 norm ball $\{w : \|w\|_2 \leq B\}$

Since it holds that

$$W(\Sigma_2^{1/2}\mathscr{K}) = B \cdot \mathbb{E}_{H \sim N(0,I_d)} \|\Sigma_2^{1/2} H\|_2 \leq \sqrt{B^2 \mathbb{E}\|x\|_2^2}$$

our main bound shows

$$\sup_{\|w\| \leq B, \hat{L}(w)=0} L(w) \leq (1 + o(1)) \frac{B^2 \mathbb{E}\|x\|_2^2}{n}$$

Norm bound:

$$\|\hat{w}\|_2^2 \leq (1 + o(1)) \frac{\sigma^2 n}{\mathbb{E}\|x\|_2^2}$$

Plugging in, we establish consistency $L(\hat{w}) \to \sigma^2$ under the benign overfitting conditions

$$\frac{\mathrm{rank}(\Sigma_1)}{n} \to 0, \quad \|w^*\|_2 \sqrt{\frac{tr(\Sigma_2)}{n}} \to 0, \quad \frac{n}{R(\Sigma_2)} \to 0$$

where

$$r(\Sigma) = \frac{tr(\Sigma)}{\|\Sigma\|} \quad \text{and} \quad R(\Sigma) = \frac{tr(\Sigma)^2}{tr(\Sigma^2)}.$$

ℓ_1 norm ball (Basis Pursuit)

The effective rank becomes

$$R_1(\Sigma) = \frac{\left(\mathbb{E}\|\Sigma^{1/2} H\|_\infty\right)^2}{\max_i \Sigma_{ii}}$$

and the benign overfitting condition is

$$\frac{\mathrm{rank}(\Sigma_1)}{n} \to 0, \quad \|w^*\|_1 \frac{\mathbb{E}\|\Sigma^{1/2} H\|_\infty}{\sqrt{n}} \to 0, \quad \frac{n}{R_1(\Sigma_2)} \to 0$$

This is satisfied, for example, by the "junk feature" model. We can also extend this to arbitrary norm.