Generalization in extensive-width neural networks via low-degree polynomials

Guilhem Semerjian

LPENS

14.08.2025 / Cargese

work in progress

Outline

1 Introduction

2 Main results / conjectures

3 Technical details

Outline

1. Introduction
2. Main results / conjectures
3. Technical details

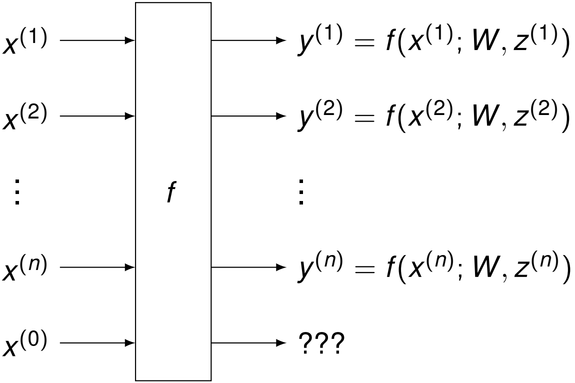
The generalization problem

Black box input-output relation :

$$x \in \mathbb{R}^d \longrightarrow \boxed{f} \longrightarrow y = f(x) \in \mathbb{R}$$

- $d \gg 1$
- $f(x) = f(x; W, z)$
- W : weights, fixed once and for all
- z : noise, i.i.d. at each use of f

The generalization problem



- observations : $\mathcal{O} = \{x^{(1)}, y^{(1)}, \dots, x^{(n)}, y^{(n)}, x^{(0)}\}$
- goal : estimator of $y^{(0)}$ from the observations, $\widehat{y} = \widehat{y}(\mathcal{O})$

The generalization problem

$\widehat{y} = \widehat{y}(\mathcal{O})$ to be built in the Bayesian setting :

- $x^{(i)}$ i.i.d. with a law known to the observer
- $z^{(i)}$ i.i.d. with a law known to the observer
- the law of W is known
- the functional form of $f(x; W, z)$ is known

Quality of the estimator $\widehat{y}$ measured by $\mathrm{MSE}(\widehat{y}) = \mathbb{E}[(y^{(0)} - \widehat{y})^2]$

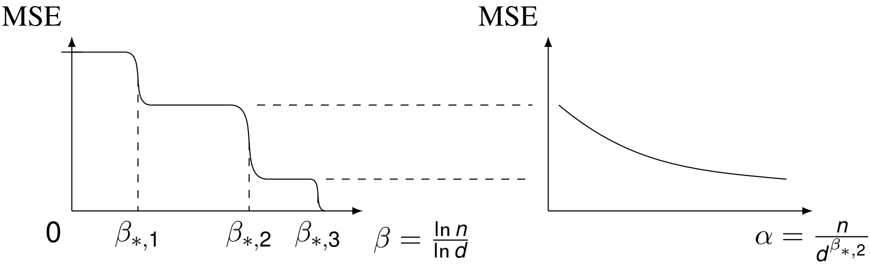
Optimal choice : $\widehat{y} = \mathbb{E}[y^{(0)} | \mathcal{O} = \{x^{(1)}, y^{(1)}, \dots, x^{(n)}, y^{(n)}, x^{(0)}\}]$

difficult to compute in general $\Rightarrow$ systematic approximations

The generalization problem

Depending on the details of f, decrease of MSE with n :

- as a power law
- or step-like behavior when $d \to \infty$, $n = \alpha\, d^\beta$



Goal : these curves for the optimal estimator,

or some efficiently computable approximations

Two-layer architecture

- $y^{(i)} = \frac{1}{\sqrt{m}} \sum_{\mu=1}^{m} \varphi\left(\frac{w_\mu \cdot x^{(i)}}{\sqrt{d}}\right)$
- $\varphi(h) = \sum_{k \geq 1} \widehat{\varphi}_k H_k(h)$ Hermite decomposition
- w_μ i.i.d. with law $\mathcal{N}(0, \mathbb{1}_d)$
- $x^{(i)}$ i.i.d. with law Unif$(\mathbb{S}^{d-1}(\sqrt{d}))$,
  or $\mathcal{N}(0, \mathbb{1}_d)$, or anything orthogonally invariant with norm $\approx \sqrt{d}$
- one could add noise and disorder in the second layer weigths

$d, m, n \to \infty$ with $m = \gamma\, d$, $n = \alpha\, d^\beta$, with α, β and γ are fixed

$0 < \gamma < \infty$, extensive width : this is not

- a multi-index model (would be m finite, $\gamma = 0$)
- a Gaussian process (would be $m \to \infty$ first, $\gamma = \infty$)

Two-layer architecture

- $y^{(i)} = \frac{1}{\sqrt{m}} \sum_{\mu=1}^{m} \varphi\left(\frac{w_\mu \cdot x^{(i)}}{\sqrt{d}} \right)$
- $\varphi(h) = \sum_{k \geq 1} \widehat{\varphi}_k H_k(h)$
- $d, m, n \to \infty$ with $m = \gamma\, d$, $n = \alpha\, d^\beta$

Some recent studies in this regime :

- $\varphi(h)$ arbitrary, $\beta = 1$ [Cui, Zdeborová, Krzakala 23]
- $\varphi(h) = H_2(h)$, dynamics on population risk [Martin, Bach, Biroli 24]
- $\varphi(h) = H_2(h)$, $\beta = 2$ [Maillard, Troiani, Martin, Krzakala, Zdeborová 24]
- $\varphi(h)$ arbitrary, $\beta = 2$ [Barbier, Camilli, Nguyen, Pastore, Skerk 25]

Approximate Bayesian estimation

Quality of an estimator $\widehat{y}(\mathcal{O})$ of $y^{(0)}$: $\mathrm{MSE}(\widehat{y}) = \mathbb{E}[(y^{(0)} - \widehat{y})^2]$

Optimal choice : $\widehat{y} = \mathbb{E}[y^{(0)}|\mathcal{O}]$, too complicated in general

Low-degree polynomial method :

- for hypothesis testing [Hopkins, Steurer 17] [Kunisky, Wein, Bandeira 22]
- for estimation [Schramm, Wein 22] [Montanari, Wein 22]
- for constraint satisfaction problems [Bresler, Huang 22]

proofs of hardness results,

thought to emulate polynomial-time algorithms

Approximate Bayesian estimation

introduce a variational space with basic functions (e.g. polynomials)

$$\widehat{y}(\mathcal{O}) = \sum_{\omega \in \mathcal{A}} c_\omega \, b_\omega(\mathcal{O}) \ , \ \mathcal{A} : \text{finite set}, \ \ c : \text{variational parameters}$$

reduces to a quadratic optimization problem in a smaller space:

$$\begin{aligned} \mathrm{MSE}(\widehat{y}) &= \mathbb{E}[(y^{(0)})^2] + \sum_{\omega,\omega' \in \mathcal{A}} c_\omega \mathcal{M}_{\omega,\omega'} c_{\omega'} - 2 \sum_{\omega \in \mathcal{A}} c_\omega \mathcal{R}_\omega \\ &= \mathbb{E}[(y^{(0)})^2] + c^T \mathcal{M} c - 2 c^T \mathcal{R} \ , \end{aligned}$$

where $\mathcal{M}$ is a square matrix and $\mathcal{R}$ a vector, both of size $|\mathcal{A}|$:

$$\mathcal{M}_{\omega,\omega'} = \mathbb{E}[b_\omega(\mathcal{O}) b_{\omega'}(\mathcal{O})] \qquad \mathcal{R}_\omega = \mathbb{E}[y^{(0)} b_\omega(\mathcal{O})]$$

Approximate Bayesian estimation

$$\mathcal{M}_{\omega,\omega'} = \mathbb{E}[b_\omega(\mathcal{O}) b_{\omega'}(\mathcal{O})] \qquad \mathcal{R}_\omega = \mathbb{E}[y^{(0)} b_\omega(\mathcal{O})]$$

optimal MSE in this subspace:

$$\mathrm{MMSE}_{\mathcal{A}} = \mathbb{E}[(y^{(0)})^2] + \inf_{c \in \mathbb{R}^{|\mathcal{A}|}} [c^T \mathcal{M} c - 2 c^T \mathcal{R}]$$

reached for $\mathcal{M} c = \mathcal{R}$, yields

$$\mathrm{MMSE}_{\mathcal{A}} = \mathbb{E}[(y^{(0)})^2] - \mathcal{R}^T \mathcal{M}^{-1} \mathcal{R}$$

one should aim for "$\mathcal{R}$ big, $\mathcal{M}$ small"

low-degree polynomial method for estimation :

$$\{b_\omega(\mathcal{O})\}_{\omega \in \mathcal{A}} = \text{polynomials in } \mathcal{O} \text{ of small degree}$$

Approximate Bayesian estimation

here $b_\omega(\mathcal{O}) = b_\omega(x^{(1)}, y^{(1)}, \dots, x^{(n)}, y^{(n)}, x^{(0)})$,

quite a lot of polynomials from $\mathbb{R}^{n(d+1)+d}$ to $\mathbb{R}$...

To keep $|\mathcal{A}|$ finite as $d \to \infty$, use symmetries (Hunt-Stein lemma) :

- permutation symmetry between the n samples
- orthogonal invariance, $x^{(i)} \to Ox^{(i)}$, $w_\mu \to Ow_\mu$ for all $O \in O_d$

Weyl's First Fundamental Theorem :

$$f(Ox^{(0)}, \dots, Ox^{(n)}) = f(x^{(0)}, \dots, x^{(n)}) \qquad \forall O \in O_d$$

$$\Rightarrow \qquad f(x^{(0)}, \dots, x^{(n)}) = \widetilde{f}(\{x^{(i)} \cdot x^{(j)}\})$$

Approximate Bayesian estimation

one can thus restrict to

$$b_\omega(\mathcal{O}) = \sum_{\substack{i_1,\dots,i_p=1 \\ \text{all} \neq}}^{n} \mathcal{W}_\omega(y^{(i_1)},\dots,y^{(i_p)})\, \mathcal{Y}_\omega(x^{(0)},x^{(i_1)},\dots,x^{(i_p)})$$

- $p = p_\omega$
- $\mathcal{W}_\omega$ polynomial on $\mathbb{R}^p$
- $\mathcal{Y}_\omega(x^{(0)},x^{(1)},\dots,x^{(p)})$ polynomial in the $\{x^{(i)} \cdot x^{(j)}\}$

$\Rightarrow$ number of polynomials of a given degree independent of d

convenient choices of $\mathcal{W}$ and $\mathcal{Y}$ to be discussed later

Outline

1 Introduction

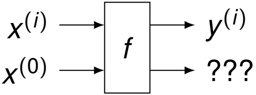
2 Main results / conjectures

3 Technical details

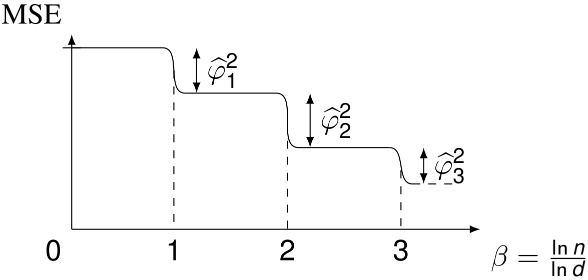
Main results / conjectures

- $p = 1$, variational set : $\left\{ b_\ell(\mathcal{O}) = \sum_{i=1}^{n} y^{(i)} \mathcal{Y}_\ell(x^{(0)} \cdot x^{(i)}) \right\}_{\ell \geq 1}$

$\mathcal{Y}_\ell$: Gegenbauer polynomial



achieves



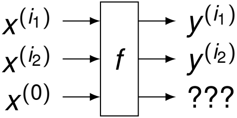
consistent with [Mei, Misiakiewicz, Montanari 21]

can one do better ?

Main results / conjectures

- no if $\gamma = \infty$, yes if $0 < \gamma < \infty$:
- with $p = 2$

$$b_\omega(\mathcal{O}) = \sum_{i_1 \neq i_2 = 1}^{n} \mathcal{W}(y^{(i_1)}, y^{(i_2)})\, \mathcal{Y}_\omega(x^{(0)}, x^{(i_1)}, x^{(i_2)})$$



  - for $\varphi(h) = H_k(h)$, k even,
    step at $\beta = \frac{2+3k}{4}$ instead of k
  - there is “cooperation” between Hermite modes
    reminiscent of staircase effect
    [Abbe, Boix-Adsera, Brennan, Bresler, Nagaraj 21]

    for instance if $\widehat{\varphi}_5 \neq 0$ and $\widehat{\varphi}_6 \neq 0$,
    step of height $\widehat{\varphi}_6^2$ in $\beta = \frac{9}{2} < 5 < 6$

Main results / conjectures

- what is the best one can hope for with finite degree polynomials ?

  i.e. p arbitrary but independent on d

  for $\varphi(h) = H_k(h)$, β has to be larger than $\frac{2+k}{2}$

  naive IT guess is $\beta = 2$, would imply
  a statistical-computational gap for all $k > 2$

Outline

1 Introduction

2 Main results / conjectures

3 Technical details

Technical details

how to choose $\mathcal{W}(y^{(1)},\dots,y^{(p)})$ and $\mathcal{Y}(x^{(0)},\dots,x^{(p)})$?

- $\mathcal{W}(y^{(1)},\dots,y^{(p)})$: Wick polynomial (for the law conditional on the x)
  - they are to monomials what cumulants are to moments
    generalized centering
    called “normal order” in Quantum Field Theory
  - $\mathcal{W}(X_1,\dots,X_N) = [t_1\dots t_N]\frac{e^{t_1X_1+\dots+t_NX_N}}{\mathbb{E}[e^{t_1X_1+\dots+t_NX_N}]}$
  - $\mathcal{W}(X) = X - \mathbb{E}[X]$
    $\mathcal{W}(X_1,X_2) = X_1X_2 - X_1\mathbb{E}[X_2] - \mathbb{E}[X_1]X_2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] - \mathbb{E}[X_1X_2]$
    $\mathcal{W}(X_1,X_2,X_3) = X_1X_2X_3 - \dots$
  - $\mathbb{E}[X_0\mathcal{W}(X_1,\dots,X_N)] = \kappa[X_0,X_1,\dots,X_N]$
    whereas $\mathbb{E}[X_0X_1\dots X_N]$ involves a sum
    on the partitions of $\{0,1,\dots,N\}$

Technical details

- $\mathcal{Y}(x^{(1)}, \dots, x^{(p)})$: "Multi-spherical harmonics"

  [Jones, Potechin 21]

  reminder on spherical harmonics :

  - $\mathcal{P}$: polynomials of $\mathbb{R}^d \to \mathbb{R}$
  - $\mathcal{P}_\ell$: polynomials of $\mathbb{R}^d \to \mathbb{R}$ homogeneous of degree ℓ
  - $\mathcal{P} = \bigoplus_{\ell \geq 0} \mathcal{P}_\ell$
  - $\mathcal{H}_\ell \subset \mathcal{P}_\ell$: harmonic ($\Delta P = 0$) polynomials homogeneous, degree ℓ
  - $\mathcal{P}_\ell = \mathcal{H}_\ell \oplus (x \cdot x)\mathcal{H}_{\ell-2} \oplus (x \cdot x)^2 \mathcal{H}_{\ell-4} \oplus \dots$
  - $\mathfrak{P}_{\ell,x}$: projector from $\mathcal{P}_\ell$ onto $\mathcal{H}_\ell$

Technical details

multi-spherical harmonics generalization :

- $P(x^{(1)}, \dots, x^{(p)})$ polynomial of $(\mathbb{R}^d)^p \to \mathbb{R}$
- any P is a linear combination of $(x^{(1)} \cdot x^{(1)})^{j_1} \dots (x^{(p)} \cdot x^{(p)})^{j_p} Q(x^{(1)}, \dots, x^{(p)})$
- with $Q \in \mathcal{H}_{\ell_1,\dots,\ell_p}$, homogeneous and harmonic in each variable
- $\mathcal{H}^{(\mathrm{inv})}_{\ell_1,\dots,\ell_p}$: those that are invariant by simultaneous orthogonal transformations $\overset{\text{Weyl FFT}}{\Rightarrow}$ functions of $x^{(i)} \cdot x^{(j)}$
- $\mathcal{H}^{(\mathrm{inv})}_{\ell_1,\dots,\ell_p}$ spanned by $\mathfrak{P}_{\ell_1,x^{(1)}} \dots \mathfrak{P}_{\ell_p,x^{(p)}} M_G$

  $G = \{m_{i,j} \geq 0\}_{i<j}$ $\qquad M_G(x^{(1)}, \dots, x^{(p)}) = \prod_{1 \leq i < j \leq p} (x^{(i)} \cdot x^{(j)})^{m_{i,j}}$

  G : multi-graph on vertices $\{1, \dots, p\}$, with no self-loops, $m_{i,j}$ edges between vertices i and j, vertex i has degree ℓ_i

Perspectives

- connection with the multi-index regime (m finite, $\gamma \to 0$),
  focus was here on $y^{(0)}$ and not on W
- other (deeper) architectures, all you need are the cumulants $\kappa[y^{(1)}, \dots, y^{(p)} | x^{(1)}, \dots, x^{(p)}]$
- to reach a given accuracy,
  sample vs compute time complexity tradeoff
- for $\varphi(h) = H_2(h)$, order by order identification of the terms of the AMP+RIE algorithm of
  [Maillard, Troiani, Martin, Krzakala, Zdeborová 24]
- universality in the weights distribution
  [Barbier, Camilli, Nguyen, Pastore, Skerk 25]