Generalisation of some overparametrised models

Stéphane Chrétien

University of Lyon 2
ERIC Laboratory

LOL 2022 – Learning and Optimization in Luminy
CIRM, Luminy, 4 octobre 2022

Introduction

- Modern machine learning and statistics deal with the problem of learning from data:
  - given a training dataset (y_i, x_i) $i \in I$ where
    - $x_i \in \mathbb{R}^d$ is the input
    - $y_i \in \mathbb{R}$ is the output,

    one seeks a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ from a certain function class $\mathcal{F}$ that has good prediction performance on test data $(y_t, x_t))$, $t \in T$, i.e. which has small testing error

$$\sum_{t \in T} \ell(y_t, f(x_t)) \tag{1}$$

- This problem is of fundamental significance and finds applications in numerous scenarios.
    - For instance, in image recognition,
        - the input x corresponds to the raw image
        - the output y is the image category

      and the goal is to find a mapping f that can classify new images with acceptable accuracy.
- Decades of research efforts in statistical machine learning have been devoted to developing methods to find f efficiently with provable guarantees.

- Prominent examples include
  - linear classifiers (e.g., linear / logistic regression, linear discriminant analysis),
  - kernel methods (e.g., support vector machines),
  - tree-based methods (e.g., decision trees, random forests),
  - nonparametric regression (e.g., nearest neighbors, local kernel smoothing), etc.
- Roughly speaking, each aforementioned method corresponds to a different function class $\mathcal{F}$ from which the final classifier f is chosen.

- Deep learning, in its simplest form, consists in looking for functions of the form

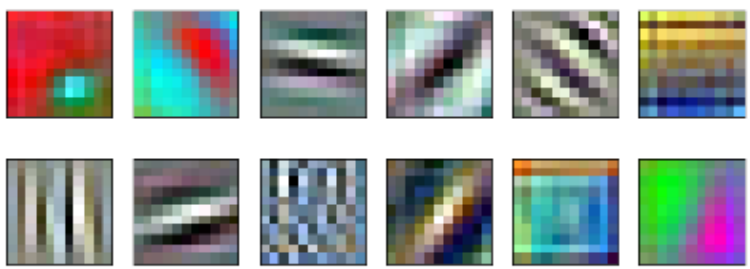
$$\mathcal{F} = \Big\{ f(x,\theta) = W_L(\sigma_L(W_{L-1}(\sigma_{L-1}(\cdots \sigma_2(W_1(x)))))\Big\}.$$

where σ_l is a non-linear function which applies componentwise and W_l is an affine operator, $l = 1, \ldots, L$.

- Deep learning is able to approximate complicated nonlinear maps through composing many simple nonlinear functions.

- The motivation for the multilayer architecture is that there are different levels of features and the layers might be able to properly account for these different levels independently.

- Here, we sample and visualize weights from a pre-trained AlexNet model.



This can be used to generate new images using for instance, Generative Adversarial Networks or Diffusion models.

- Evolution of the performances over the last 7 years . . .

Model	Year	# Layers	# Params	Top-5 error
Shallow	< 2012	—	—	$> 25\%$
AlexNet	2012	8	61M	16.4%
VGG19	2014	19	144M	7.3%
GoogleNet	2014	22	7M	6.7%
ResNet-152	2015	152	60M	3.6%

- It is widely acknowledged that two indispensable factors contribute to the success of deep learning, namely
  - huge datasets that often contain millions of samples and
  - immense computing power resulting from clusters of graphics processing units (GPUs).
- Admittedly, these resources are only recently available.

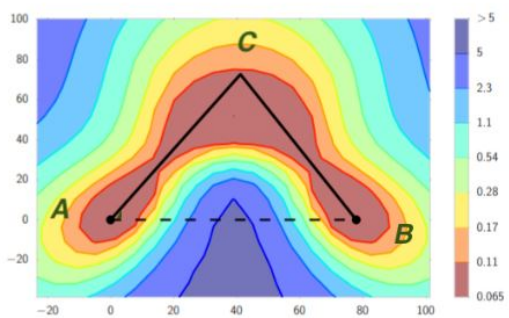
- However, these two alone are not sufficient to explain the mystery of deep learning:
  - Why is over-parametrization not a problem ?
    - overparametrisation should lead to overfitting,
    - BUT ... this is not what we always observe in practice !

- and
  - nonconvexity does not seem to be a problem: even with the help of GPUs, training deep learning models is still NP-hard in the worst case due to the highly nonconvex loss function to minimize.
    - Nevertheless, standard incremental algorithms (Stochastic Gradient Descent, etc) often reach good minimisers of the Empirical Risk
- A lot remains to be understood ! ...

Why overparametrise, to begin with ?

- It is often observed that depth helps efficiently extract features at different scales from the inputs,
- recent studies found that *increasing both depth and width in a shallow model leads to very nice continuous limits, where PDE tools can be put to work...*
- Networks with wide layers (larger than sample size) enjoy connectivity of the minimisers (Nguyen 2019)

> See Figure 1 below from Garipov et al. 2018 for an illustration. Solutions A and B have low cost but the line connecting them goes through solutions with high cost. But we can find C of low cost such that paths AC and CB only pass through low-cost region.



What are the bad consequences of overparametrisation ?

- When some of the layers are not wide, *over-parametrization* usually entails existence of many local minimisers with potentially different statistical performance.
    - Common practice advises to runs stochastic gradient descent with random initialization and converges to parameters with *very good practical prediction accuracy*.
        - Why is this simple approach actually often working ?
- Overfitting should take place in full generality
    - Does the optimisation algorithm help find better networks ?

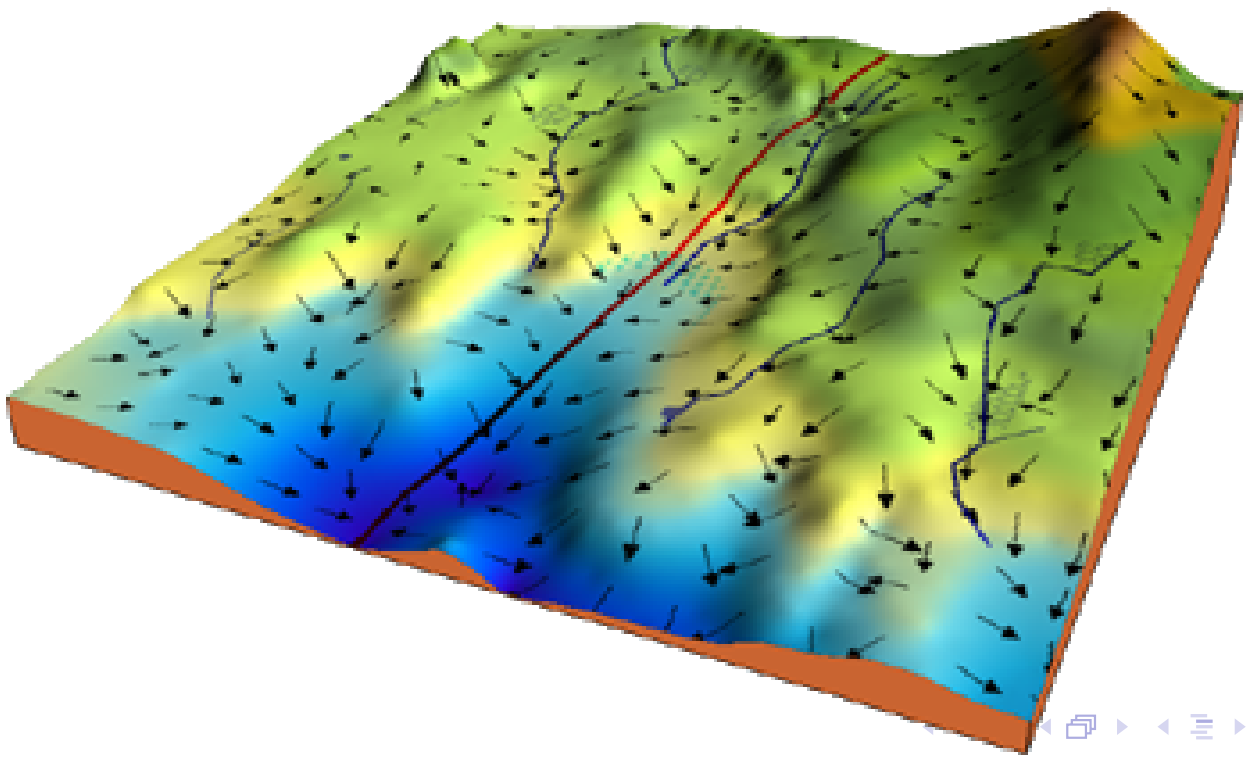
The goal of current research is to resolve these paradoxes !

A striking property of stochastic gradient descent : implicit biais towards least ℓ_2 norm solutions

Implicit biais of gradient descent

- For minimising a function $F(\theta)$, one can use the gradient method :

$$\theta^{(l+1)} = \theta^{(l)} - \eta_l \, \nabla F(\theta^{(l)}) \tag{2}$$



Implicit biais of gradient descent

- if there is a unique global minimizer θ_*, then the goal of optimization algorithms is to find this minimizer,

- when there are multiple minimizers (thus for a function which cannot be strongly convex), one can easily show that

$$F(\theta_t) - \inf_{\theta \in \mathbb{R}^d} F(\theta) \tag{3}$$

is converging to zero.

Implicit biais of gradient descent

- With some extra assumptions, we can show that the algorithm is converging to one of the multiple minimizers of F
    - note that when F is convex, this set is also convex.
- But . . . which one ?

Implicit biais of gradient descent

- This is what is referred to as the implicit regularization property of certain optimization algorithms, and in particular, gradient descent and its variants.
    - This is interesting in overparametrised machine learning because there usually are many minimizers
- In a nutshell, gradient descent usually leads to minimum ℓ_2-norm solutions.
    - This shows that the chosen empirical risk minimizer is not arbitrary !

A simple analysis of the linear case with iid Gaussian design *(from the lecture notes by Francis Bach)*

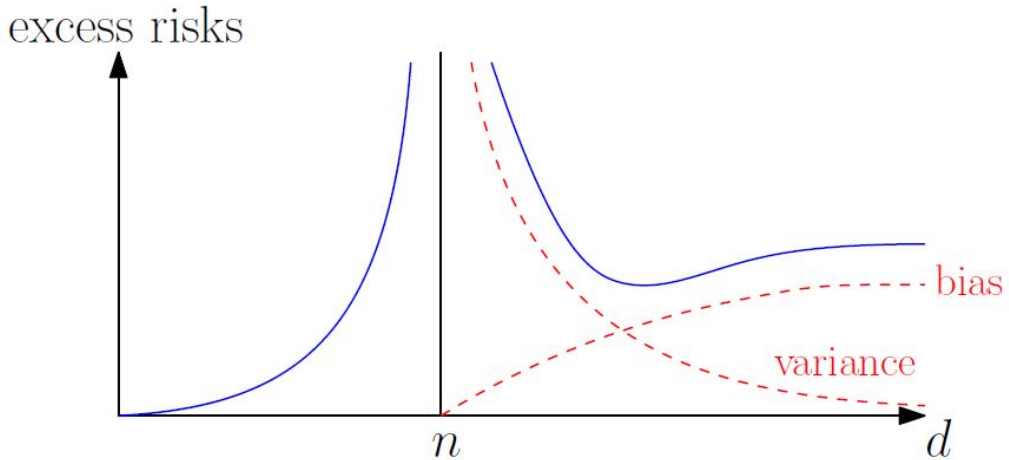
Linear regression with iid Gaussian design : the smallest ℓ_2-norm estimator

- We have

$$\text{if } d \leqslant n-2, \quad \mathbb{E}[R(\hat{\theta})] = \sigma^2 \frac{d}{n-d-1}$$

$$\text{if } d \geqslant n+2, \quad \mathbb{E}[R(\hat{\theta})] = \frac{\sigma^2 n}{d-n-1} + \|\theta_*\|_2^2 \frac{d-n}{d}.$$

This leads to the following picture.



A slightly more general nonlinear regression setup: ridge functions

(work with Emmanuel Caron, Univ. Avignon, France)

Mathematical Model

Let $Z_i = (X_i, Y_i)$ in $\mathbb{R}^{d+1} \times \mathbb{R}$, $i = 1, \ldots, n$ be observations drawn from the following model

$$Y_i = f^*(X_i) + \varepsilon_i \tag{4}$$

$i = 1, \ldots, n$, where we assume that

- the vectors X_i, $i = 1, \ldots, n$ are random and i.i.d., taking values in $\mathbb{R}^d$
- the values ε_i, $i = 1, \ldots, n$ are the random observation errors.

The goal is to estimate f^* based on the observation $Z_1, \ldots, Z_n$.

The estimation of f^* will be based on restricting the search to a subset $\mathcal{F}$ of functions of a Banach space $\mathcal{B}$.

In order to generalise, the estimator should be chosen in the set of stationary points of the empirical version of the risk $\mathcal{R} : \mathcal{F} \to \mathbb{R}$ defined by

$$\mathcal{R}(f) = \mathbb{E}\left[\ell(Y, f(X))\right],$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies

- $\ell(y, y) = 0$ for all $y \in \mathbb{R}$ and
- $\ell(y, \cdot)$: $\mathbb{R} \mapsto \mathbb{R}$ is a strictly convex twice continuously differentiable nonnegative function

Let $\hat{R}_n(f)$ denote the empirical risk defined by

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)). \tag{5}$$

Then, the Empirical Risk Minimizer $\hat{f}^{ERM}$ will be a solution to

$$\hat{f}^{ERM} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_n(f). \tag{6}$$

Let us start with ridge type functions

Ridge type functions

We consider a statistical model of the form

$$\mathbb{E}[Y_i \mid X_i] = f(X_i^t \theta^*), \qquad i = 1, \ldots, n, \tag{7}$$

where

- $\theta^* \in \mathbb{R}^p$
- the function f: $\mathbb{R} \mapsto \mathbb{R}$ is assumed increasing

- A random variable ξ is called sub-Gaussian if there exists a number $a \in [0, \infty)$ such that

$$\mathbb{E} \exp\{\lambda \xi\} \leq \exp\left\{\frac{a^2 \lambda^2}{2}\right\}$$

for all $\lambda \in \mathbb{R}$.

- The number

$$\|\xi\|_{\psi_2} = \inf\left\{a \geq 0 : \mathbb{E} \exp\{\lambda \xi\} \leq \exp\left\{\frac{a^2 \lambda^2}{2}\right\}, \lambda \in \mathbf{R}\right\}$$

is called the sub-Gaussian norm of the random variable ξ.

- A random variable ξ is sub-Gaussian if and only if

$$\|\xi\|_{\psi_2} < \infty$$

.

- A random vector ξ with values in $\mathbb{R}^p$ is subGaussian with subGaussian constant K_ξ if

$$\|\langle w, \xi \rangle\|_{\psi_2} \leq K_X \tag{8}$$

for all $w \in \mathbb{R}^p$ with $\|w\|_2 = 1$.

Ridge type functions

- the data $X_1, \ldots, X_n$ will be assumed isotropic and subGaussian
- the matrix

$$X^\top = [X_1, \ldots, X_n] \tag{9}$$

  is full rank with probability one.
- for all $i = 1, \ldots, n$, the random vectors X_i are assumed
  - to have a second moment matrix $\mathbb{E}[X_i X_i^t] = I_p$,
  - to have ℓ_2-norm equal[1] to $\sqrt{p}$.
- the errors $\epsilon_i = Y_i - \mathbb{E}[Y_i]$ are independent subGaussian centered random variables with ψ_2-norm upper bounded by K_ϵ.

[1] notice that this is different from the usual regression model, where the columns are assumed to be normalised

Ridge type functions

In order to estimate θ^*, the Empirical Risk Minimizer $\hat{\theta}$ is defined as a solution to the following optimisation problem

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \hat{R}_n(\theta) \tag{10}$$

with

$$\hat{R}_n(\theta) = \operatorname{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i - f(X_i^t \theta)). \tag{11}$$

Moreover, we assume that $\ell'(0) = 0$ and ℓ'' is upper bounded by a constant $C_{\ell''} > 0$.

Ridge type functions

Theorem

(Overparametrised setting) Let $\mu > 0$, $\nu > 0$ *and let* $\beta \in (0, 1)$. *Assume that* p *and* n *are such that*

$$(\alpha + C_{K_X})^2 \; n < p. \tag{12}$$

Let

$$r = \frac{12 C' \sqrt{C}\, C_{\ell''} K_\epsilon \sqrt{p}}{(\sqrt{p} - (\alpha + C_{K_X})\sqrt{n})\; \delta}. \tag{13}$$

Assume that $f'(z) \leq C_{f'}$ *and* ℓ *and* f *are such that*

$$\ell''(w)\; f'(z)^2 - \ell'(w)\; f''(z) \geq \delta$$

for all z *in* $X\mathcal{B}_2(\theta^*, r)$ *(Trivial in the linear case).*

Ridge type functions

Theorem

(Overparametrised setting) Then, there exists a first order stationary point $\hat{\theta}$ to the ERM problem such that, with probability larger than or equal to

$$1 - \left(2\exp\left(-c_{K_X}\alpha^2 n\right) + \exp\left(-\frac{n}{2}\right) + 2n\left(\exp\left(-\frac{\nu^2 \log(n)}{C_{\ell''}^2 K_\epsilon^2}\right)\right)\right)$$

we have

$$\|\hat{\theta} - \theta^*\|_2 \leq r. \tag{14}$$

The DNN case

An handy result from Neuberger about the distance of the solution of a zero finding problem, i.e. consisting in solving

$$F(\hat{f}) = 0,$$

to the initial guess f^*.

The Continuous Newton's Method, Inverse Functions, and Nash-Moser

J. W. Neuberger

1. INTRODUCTION. The conventional Newton's method for finding a zero of a function $F : R^n \to R^n$, assuming that $(F'(y))^{-1}$ exists for at least some y in R^n, is the familiar iteration: pick z_0 in R^n and define

$$z_{k+1} = z_k - (F'(z_k))^{-1} F(z_k) \quad (k = 0, 1, 2, \ldots),$$

hoping that $z_1, z_2, \ldots$ converges to a zero of F. What can stop this process from finding a zero of F? For one thing, there might not *be* a zero of F. For another, the process

The DNN case

Theorem (Neuberger's theorem)

Suppose that $r > 0$, that $\theta^* \in \mathbb{R}^p$ and that the map F is continuous on $\overline{B_r}(\theta^*)$, with the property that for each θ in $B_r(\theta^*)$ there exists a vector d in $\overline{B_r}(0)$ such that,

$$\lim_{t \downarrow 0} \frac{F(\theta + td) - F(\theta)}{t} = -F(\theta^*). \tag{15}$$

Then there exists u in $\overline{B_r}(\theta^*)$ such that $F(u) = 0$.

Ridge type functions

Theorem (Neuberger's theorem for ERM)

Suppose that $r > 0$, that $\theta^* \in \mathbb{R}^p$ and that the Jacobian $D\hat{R}_n(\cdot)$ is a continuous map on $\mathcal{B}(\theta^*, r)$ with the property that for each θ in $\mathcal{B}(\theta^*, r)$ there exists a vector d in $\overline{\mathcal{B}(0, r)}$ such that,

$$\lim_{t \downarrow 0} \frac{D\hat{R}_n(\theta + td) - D\hat{R}_n(\theta)}{t} = -D\hat{R}_n(\theta^*). \tag{16}$$

Then there exists u in $\overline{\mathcal{B}(\theta^*, r)}$ such that $D\hat{R}_n(u) = 0$.

Ridge type functions

Since the loss is twice differentiable, the empirical risk $\hat{R}_n$ is itself twice differentiable. The Gradient of the empirical risk is given by

$$\begin{aligned}\nabla \hat{R}_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i - f(X_i^t\theta))\ f'(X_i^t\theta)X_i \\ &= -\frac{1}{n}X^t D(\nu)\ l'(\epsilon)\end{aligned}$$

where $\ell'(\epsilon)$ is to be understood componentwise, and

$$\nu_i = f'(X_i^t\theta) \tag{17}$$

and the Hessian is given by

$$\begin{aligned}\nabla^2 \hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \Big(&\ell''(Y_i - f(X_i^t\theta))\ f'(X_i^t\theta)^2 \\ &- \ell'(Y_i - f(X_i^t\theta))\ f''(X_i^t\theta)\Big)X_iX_i^t.\end{aligned} \tag{18}$$

Ridge type functions

The condition we have to satisfy in order to use Neuberger's theorem is

$$\nabla^2 \hat{R}_n(\theta) d = -\nabla \hat{R}_n(\theta^*) \tag{19}$$

for all $\theta \in \mathcal{B}(\theta^*, r)$. The Hessian matrix can be rewritten as

$$\nabla^2 \hat{R}_n(\theta) = \frac{1}{n} X^t D(\mu) X \tag{20}$$

where $D_{Y,X}$ is a diagonal matrix given by

$$\mu_i = \left(\ell''(Y_i - f(X_i^t \theta))\ f'(X_i^t \theta)^2 - \ell'(Y_i - f(X_i^t \theta))\ f''(X_i^t \theta) \right)$$

Ridge type functions

We have to solve Neuberger's equation

$$\frac{1}{n} X^t D(\mu) X d = \frac{1}{n} X^t D(\nu) \ell'(\epsilon) \tag{21}$$

which can be solved by finding the least norm solution of the interpolation problem

$$D(\mu) X d = D(\nu) \ell'(\epsilon). \tag{22}$$

i.e.

$$d = X^\dagger D(\mu)^{-1} D(\nu) \ell'(\epsilon). \tag{23}$$

Ridge type functions

Given the compact SVD of $X = U\Sigma V^t$, where $U \in O(n)$ and $V \in \mathbb{R}^{p\times n}$ with orthonormal columns, i.e. V belongs to the Stiefel manifold, we get

$$d = V\Sigma^{-1}U^t D(\mu^{-1})D(\nu)\ell'(\epsilon). \tag{24}$$

We then have

$$\|d\|_2 = \|V\Sigma^{-1}U^t D(\mu^{-1})D(\nu)\ell'(\epsilon)\|_2 \tag{25}$$

i.e.

$$\|d\|_2 = \|\Sigma^{-1}U^t D(\mu^{-1})D(\nu)\ell'(\epsilon)\|_2 \leq \frac{\|U^t D(\mu^{-1})D(\nu)\ell'(\epsilon)\|_2}{s_{\min}(X^t)}.$$

We will need the following maximal inequality.

Theorem

Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. Then

$$\mathbb{E}\left[\max_{\theta \in \mathcal{B}_2} \theta^\top X\right] = \mathbb{E}\left[\max_{\theta \in \mathcal{B}_2} \left|\theta^\top X\right|\right] \leq 4\sigma\sqrt{d}$$

Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\max_{\theta \in \mathcal{B}_2} \theta^\top X = \max_{\theta \in \mathcal{B}_2} \left|\theta^\top X\right| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$$

Ridge functions

After computing the sub-Gaussian constant of the numerator, we get

$$\|U^t\ D(\mu)^{-1}D(\nu)\ell'(\epsilon)\|_2 \leq 2C'\sqrt{C}\,C_{\ell''}\,\frac{\max_{i'=1}^n \nu_i'}{\min_{i'=1}^n \mu_{i'}}\,K_\epsilon(2\sqrt{p}+u)$$
$$\leq \frac{4C'\sqrt{C}\,C_{\ell''}\,C_{f'}\,K_\epsilon(2\sqrt{p}+u)}{\delta}, \tag{26}$$

with probability

$$1-\left(\exp\left(-\frac{u^2}{2}\right)+2n\left(\exp\left(-\frac{t^2}{C_{\ell''}^2K_\epsilon^2}\right)\right)\right).$$

Ridge functions

Taking $u = \sqrt{p}$, equation (26) yields

$$\|d\|_2 = \left\| \nabla^2 \hat{R}_n(\theta)^{-1} \, \nabla \hat{R}_n(\theta^*) \right\|_2 \leq \frac{12 C' \sqrt{C} \, C_{\ell''} \, C_{f'} \, K_\epsilon \sqrt{p}}{s_{\min}(X) \delta},$$

with the same probability.

Ridge type functions

We also have with probability $1 - 2\exp\left(-c_{K_X}\alpha^2 n\right)$

$$s_{\min}(X^t) \geq \left(\sqrt{p} - (\alpha + C_{K_X})\sqrt{n}\right). \tag{27}$$

Therefore, with probability larger than or equal to

$$1 - \left(2\exp\left(-c_{K_X}\alpha^2 n\right) + \exp\left(-\frac{n}{2}\right) + 2n\left(\exp\left(-\frac{t^2}{C_{\ell''}^2 K_\epsilon^2}\right)\right)\right),$$

we have

$$\|d\|_2 \leq \frac{12C'\sqrt{C}\,C_{\ell''}C_{f'}K_\epsilon\sqrt{p}}{\left(\sqrt{p} - (\alpha + C_{K_X})\sqrt{n}\right)\,\delta}.$$

Finally replace η with $\nu\sqrt{\log(n)}$ and t with $\upsilon\sqrt{\log(n)}$ and the proof is completed.

What about the smallest ℓ_2-norm estimator ?

Ridge type functions

Theorem

Let $\hat{\theta}^\circ$ denote the minimum norm stationary point of the empirical risk $\hat{R}_n$ under the constraint that $X\theta = X\hat{\theta}$, i.e.

$$\text{argmin}_\theta \ \|\theta\|_2 \quad \text{subject to } X\theta = X\hat{\theta}. \tag{28}$$

Under the same assumptions as before, we have

$$|f(X_{n+1}^\top \hat{\theta}^\sharp) - f(X_{n+1}\hat{\theta}^*)| \leq t \ \frac{C \ K_X K_\epsilon(\sqrt{n}+1)}{((1+\alpha)\sqrt{p} - C_{K_X}\sqrt{n})}$$
$$+ tK_\epsilon + t \ \frac{6\sqrt{C}\, C_{\ell''} C_{f'} K_\epsilon \sqrt{n}}{\delta(r)((1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n})}$$

Ridge type functions

Theorem

with probability at least

$$1 - 2\exp\left(-c_{K_X}\alpha^2 p\right) - \exp\left(-\frac{n}{2}\right) - \exp(-c_{K_X} p) - 3\exp(-t^2/2).$$

where r is a solution of

$$r = \frac{C_{\ell'',f',\varepsilon}\sqrt{n}}{\delta(r)\left((1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n}\right)}. \tag{29}$$

Ridge type functions: proof

Recall that $\hat{\theta}^{\circ}$ denote the minimum norm solution to the ERM, i.e.

$$\text{argmin}_{\theta} \ \|\theta\|_2 \quad \text{subject to } X\theta = X\hat{\theta}.$$

Let $\hat{\theta}^{\sharp}$ solves

$$\text{argmin}_{\theta} \ \|\theta\|_2 \quad \text{subject to } \begin{bmatrix} X \\ X_{n+1}^{\top} \end{bmatrix} \theta = \begin{bmatrix} X \\ X_{n+1}^{\top} \end{bmatrix} \hat{\theta},$$

where $\hat{\theta}$ is the solution to the ERM problem which is close to θ^*.

Ridge type functions: proof

Then,

$$|X_{n+1}^\top(\hat{\theta}^\circ - \theta^*)| \leq |X_{n+1}^\top(\hat{\theta}^\circ - \hat{\theta}^\sharp)| + |\underbrace{X_{n+1}^\top(\hat{\theta}^\sharp - \hat{\theta})}_{=0 \text{ by definition}}| + |X_{n+1}^\top(\hat{\theta} - \theta^*)|,$$

$$\leq |X_{n+1}^\top(\hat{\theta}^\circ - \hat{\theta}^\sharp)| + |X_{n+1}^\top(\hat{\theta} - \theta^*)|.$$

Ridge type functions: proof

Then,

$$|X_{n+1}^\top(\hat{\theta}^\circ - \theta^*)| \leq |X_{n+1}^\top(\hat{\theta}^\circ - \hat{\theta}^\sharp)| + |\underbrace{X_{n+1}^\top(\hat{\theta}^\sharp - \hat{\theta})}_{=0 \text{ by definition}}| + |X_{n+1}^\top(\hat{\theta} - \theta^*)|,$$

$$\leq |X_{n+1}^\top(\underbrace{\hat{\theta}^\circ - \hat{\theta}^\sharp}_{\text{we have formulas}})| + |X_{n+1}^\top(\underbrace{\hat{\theta} - \theta^*}_{\text{we know a bound}})|.$$

The remainder of the proof is a sequence of standard computations.

The DNN case

The Deep Neural Network case

The DNN case

Assumption

The sample satisfies the following separation

$$\min_{i,i'=1}^{n} \|X_i - X_{i'}\|_2 \geq cn^{-1/\nu} \tag{30}$$

with probability larger than or equal to $1-\delta$*, for some positive constants* c*,* ν *and for* $\delta \in (0,1)$*.*

The Holder exponent ν is usually interpreted as a surrogate for the intrinsic dimension of the data manifold. E.g., this intrinsic dimension was estimated to be less than 20 for the MNIST dataset .

Intrinsic Dimensionality Estimation of Submanifolds in $\mathbb{R}^d$

Matthias Hein MH@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

The DNN case

Here is a Banach space version of the Neuberger theorem.

Theorem (Neuberger's theorem)

Suppose that $\mathcal{B}$, $\mathcal{J}$, and $\mathcal{K}$ are three Banach spaces and that $\mathcal{B}$ is compactly embedded in $\mathcal{J}$.
Suppose that $F : \mathcal{B} \to \mathcal{K}$ is continuous with respect to the topologies of $\mathcal{J}$ and $\mathcal{K}$.
Suppose that $f \in \mathcal{B}$, that $r > 0$, and that for each g in $B_r(f)$, there is an h in $\bar{B}_r(0)$ such that

$$\lim_{t \to 0^+} \frac{1}{t}(F(g + th) - F(g)) = -F(f).$$

Then there is $\hat{f}$ in $\bar{B}_r(f)$ such that $F(\hat{f}) = 0$.

For $r > 0$ and u in $\mathcal{B}$, $B_r(u)$ and $\bar{B}_r(u)$ will denote the open and closed balls in $\mathcal{B}$, respectively, with center u and radius r.

The DNN case

We recall that $f \in \mathcal{F}$, and $d' \in \mathcal{B}$ such that $\mathcal{F} \subset \mathcal{B}$. Let us compute the directional derivative of $\hat{R}_n$

$$D\hat{R}_n(f) \cdot h' = \lim_{t \to 0} \frac{\hat{R}_n(f + th') - \hat{R}_n(f)}{t}$$

$$= \lim_{t \to 0} \frac{\frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i) + t\ h'(X_i)) - \ell(Y_i, f(X_i))}{t}$$

$$= \lim_{t \to 0} \frac{\frac{1}{n} \sum_{i=1}^n \partial_2 \ell(Y_i, f(X_i))\ t\ h'(X_i) + c\ \partial_2^2 \ell(Y_i, f(X_i))\ t^2\ h'^2(X_i)}{t}$$

with $c \in [0, 1]$, and thus

$$D\hat{R}_n(f) \cdot h' = \frac{1}{n} \sum_{i=1}^n \partial_2 \ell(Y_i, f(X_i)) h'(X_i).$$

The DNN case

In the same spirit, we get

$$D^2 \hat{R}_n(f) \cdot (h', h) = \frac{1}{n} \sum_{i=1}^{n} \partial_2^2 \, \ell(Y_i, f(X_i)) \; h'(X_i) h(X_i).$$

The DNN case

Based on these computations, Neuberger's theorem resorts to obtaining a bound on the norm of an appropriate solution h to the following linear system

$$\frac{1}{n}\sum_{i=1}^{n} \partial_2^2\, \ell(Y_i, f(X_i))\; h'(X_i)h(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \partial_2 \ell(Y_i, f^*(X_i))h'(X_i)$$

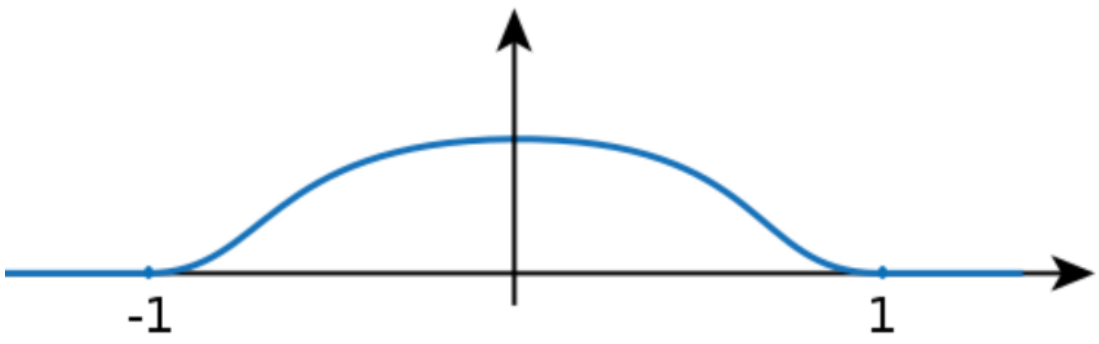
for all $f \in B_r(f^*)$ and for all $h' \in \mathcal{B}$.

The DNN case

Let ψ denote the bump function

$$\psi(x) = \begin{cases} \exp\left(1 - \frac{1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2^2 \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

and let $\psi_\sigma = \psi(\cdot/\sigma)$.



Let $\psi_\sigma = \psi(\cdot/\sigma)$.

The DNN case

Theorem

Suppose that $f^* \in C^\kappa\left([0,1]^d\right)$ *with* $\kappa \in \mathbb{N}^+$ *satisfies* $\|\partial^\alpha f\|_{L^\infty\left([0,1]^d\right)} < 1$ *for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| \leq \kappa$.

The DNN case

Theorem

Suppose that $f^* \in C^{\kappa}\left([0,1]^d\right)$ *with* $\kappa \in \mathbb{N}^+$ *satisfies* $\|\partial^{\alpha} f\|_{L^{\infty}\left([0,1]^d\right)} < 1$ *for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| \leq \kappa$.

Let $\hat{f}$ *denote any estimator of* f^*.

The DNN case

Theorem

Suppose that $f^* \in C^{\kappa}\left([0,1]^d\right)$ *with* $\kappa \in \mathbb{N}^+$ *satisfies* $\|\partial^{\alpha} f\|_{L^{\infty}\left([0,1]^d\right)} < 1$ *for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| \leq \kappa$.

Let $\hat{f}$ *denote any estimator of* f^*.

Then there exists a neural network $f_{\hat{W}}$ *which (nearly) minimizes the empirical risk such that*

$$\begin{aligned}\|f_{\hat{W}} - \hat{f}\|_{W^{k,p}(\mathcal{D})} \leq 3(\kappa+1)^d \; 8^{\kappa-k} \beta_{width}^{-2(\kappa-k)/d} \beta_{depth}^{-2(\kappa-k)/d} \\ \cdot \left(1 + n^{\frac{k}{\nu}} \max_{|\alpha|\leq K} \|\partial^{\alpha}\psi\|_{L^{\infty}\left([0,1]^d\right)}\right) \\ + 6\left(\frac{c}{2}\right)^{d/p-k} \; K_{\epsilon} \; n^{(1+\frac{k-d/p}{\nu})} \|\psi\|_{W^{k,p}(\mathbb{R}^d)} \\ + \|\hat{f} - f^*\|_{W^{k,p}(\mathcal{D})}.\end{aligned}$$

The DNN case

Theorem

Suppose that $f^* \in C^{\kappa}\left([0,1]^d\right)$ *with* $\kappa \in \mathbb{N}^+$ *satisfies* $\|\partial^{\alpha} f\|_{L^{\infty}([0,1]^d)} < 1$ *for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| \leq \kappa$.

Let $\hat{f}$ *denote any estimator of* f^*.

The neural network $f_{\hat{W}}$ *can be chosen with*

width

$$16\kappa^{d+1} d(\beta_{width} + 2)\log_2(8\beta_{width})$$

and depth

$$27\kappa^2(\beta_{depth} + 2)\log_2(4\beta_{width}).$$

The DNN case

Sketch of the proof

The DNN case

We can decouple the problem and

- first solve it in a Sobolev space, and then

- approximate the solution by a deep neural network

Simultaneous Neural Network Approximation for Smooth Functions

Sean Hon

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR

Haizhao Yang

Department of Mathematics, Purdue University, IN 47907, USA

Abstract

We establish in this work approximation results of deep neural networks for smooth func-

The DNN case

Notice that for all $f \in B_s(f_{W^*})$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell}{\partial_2}(Y_i, f(X_i))\ h'(X_i) = -\frac{1}{n}\sum_{i=1}^{n} (Y_i - f(X_i))\ \ h'(X_i),$$

and that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell}{\partial_2^2}(Y_i, f(X_i))\ h'(X_i)h(X_i) = \frac{1}{n}\sum_{i=1}^{n} h'(X_i)h(X_i).$$

Then, using the fact that ℓ is the ℓ_2^2 loss, Neuberger's condition reads

$$\frac{1}{n}\sum_{i=1}^{n} h'(X_i)h(X_i) = \frac{1}{n}\sum_{i=1}^{n} h'(X_i)\,(Y_i - f_{W^*}(X_i)).$$

The DNN case

One possible solution can be obtained by setting

$$h(X_i) = Y_i - f_{W^*}(X_i) = \varepsilon_i$$

$i = 1, \ldots, n$, i.e. using a noise interpolating solution.

One simple option is to take

$$h(x) = \sum_{i=1}^{n} \epsilon_i \, \psi\left(\frac{x - X_i}{\sigma}\right)$$

where $\psi : \mathbb{R}^p \to \mathbb{R}$ is a kernel function and $\sigma > 0$ is a bandwidth.

The DNN case

- Let

$$\psi_\sigma = \psi\left(\cdot/\sigma\right).$$

Now, observe that, based on Assumption 1, the functions $\psi((x - X_i)/\sigma)$, and their successive derivatives up to k, $i = 1, \ldots, n$, have disjoint supports for with probability larger than or equal to $1 - \delta$ as long as $\sigma \leq cn^{-1/\nu}$.

- We thus obtain that

$$\|h\|_{\mathcal{B}} \leq \|\epsilon\|_1 \ \|\psi_\sigma\|_{\mathcal{B}}$$

- Moreover, as is well known for subGaussian vectors, the norm is controlled by

$$\|\epsilon\|_2 \leq 6K_\epsilon\sqrt{n}.$$

with probability at least $1 - \exp(-n)$.

The DNN case

The proof for the deep neural network case is completed by using the approximation result of Hon and Wang.

Simultaneous Neural Network Approximation for Smooth Functions

Sean Hon

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR

Haizhao Yang

Department of Mathematics, Purdue University, IN 47907, USA

Abstract

We establish in this work approximation results of deep neural networks for smooth functions measured in Sobolev norms, motivated by recent development of numerical solvers for partial differential equations using deep neural networks. Our approximation results are nonasymptotic in the sense that the error bounds are explicitly characterized in terms of both the width and depth of the networks simultaneously with all involved constants explicitly determined. Namely, for $f \in C^s([0,1]^d)$, we show that deep ReLU networks of

- The number of layers may have to increase logarithmically with the number of samples

- The total number of parameters blows up polynomially in the number of samples and exponentially in the dimension of the problem

Conclusion and perspectives

Conclusion and perspectives

- This simple exercice in using quantitative zero finding theorems such as Neuberger's theorem shows that we can easily prove results that do not blow up with the number of layers with interpolating networks
- We can easily study local minimisers as well using the same technique

- We would need to explore approximation theory in unusual/non standard directions:
  - improve the Hon and Wang theorem by introducing the constraint that the network be a flat minimiser
    - This would explain that Stochastic Gradient methods can find the correct approximation with large probability (?)

Biblio

Some papers:

- A finite sample analysis of the double descent phenomenon for ridge function estimation, Emmanuel Caron and Stephane Chretien: arXiv preprint arXiv:2007.12882

- On the problem of estimating a Sobolev function using deep neural networks, Hadrien Bigot-Balland, Emmanuel Caron and Stephane Chretien (soon on Arxiv)