On the growth of parameters of approximating neural networks

MARTIN HOLLER AND ERION MORINA

Contact: erion.morina@uni-graz.at
Research Group: Mathematics of Data Science
Department of Mathematics and Scientific Computing
University of Graz

1) Motivation

Empirical risk minimization is a fundamental task in the field of deep learning: For *regular* $f:[0,1]^d \to [0,1]$ determine a network $f_{N,L}$ of width N, depth L and predefined architecture minimizing the empirical risk based on SGD:

$$\mathbb{E}(\|f_{N,L} - f\|_{L^1([0,1]^d,\mathbb{P})}). \qquad (1)$$

Controllability of (1) is provided by **full error analysis** [JW23]:

$$\mathbb{E}(\|f_{N,L} - f\|_{L^1([0,1]^d,\mathbb{P})}) \leq \underbrace{\mathbf{A}(N,L)}_{\text{approximation error}} + \underbrace{\mathbf{G}(N,L,K)c^{L+1}}_{\text{generalization error}} + \underbrace{\mathbf{O}(N,L,M)c}_{\text{optimization error}}$$

with K random initializations of SGD, M i.i.d. training samples and c bounding parameters of $f_{N,L}$.

In practice. Choose K and M large enough to minimize (1).

Issue. c depends on norm of network parameters.

2) Growth of parameters

Realization map for class of neural networks $\mathcal{F}$ and class of parameters Θ

$$\mathcal{R}: \Theta \to \mathcal{F}$$
$$\theta \mapsto \mathcal{N}_\theta.$$

Width and depth. For $\mathcal{N} \in \mathcal{F}$ of width N and depth L denote

$$\mathcal{W}(\mathcal{N}) = N \quad \text{and} \quad \mathcal{D}(\mathcal{N}) = L.$$

Growth of parameters. For $\|\cdot\|_\infty$ the supremum norm on Θ, consider

$$\mathcal{P}: \mathcal{F} \to [0,\infty) \qquad (2)$$
$$\mathcal{N} \mapsto \min_{\theta\in\Theta:\, \mathcal{R}(\theta)=\mathcal{N}} \|\theta\|_\infty.$$

- By standard arguments minimum in (2) is attained and $\mathcal{P}$ well defined.
- For $\tilde{\theta} \in \Theta$ with $\mathcal{R}(\tilde{\theta}) = \mathcal{N}$ it holds $\mathcal{P}(\mathcal{N}) \leq \|\tilde{\theta}\|_\infty$.

3) Objective

Approximation. For $f \in \mathcal{X} \hookrightarrow \mathcal{Y}$ there exist $f_{N,L} \in \mathcal{F}$ of width $\mathcal{W}(f_{N,L}) = w(N)$ and depth $\mathcal{D}(f_{N,L}) = d(L)$ for some increasing $w, d: \mathbb{N} \to \mathbb{N}$ such that

$$\|f_{N,L} - f\|_{\mathcal{Y}} \leq \|f\|_{\mathcal{X}} \alpha_{\mathcal{X}}(N,L) \qquad (3)$$

where $\alpha_{\mathcal{X}}: \mathbb{N}^2 \to [0,\infty)$ decreases in both components - to zero in at least one.

Goal. Determine asymptotical behavior of growth of parameters $\mathcal{P}(f_{N,L})$.

Issue. Networks expandable, describing same network with smaller parameters.

Remedy. Consider network architectures with (nearly) optimal approximation results w.r.t. width/depth/number of nonzero parameters.

Questions.
- Do $f_{N,L}$ as in (3) with (nearly) optimal approximability of f w.r.t. width and depth exist such that $\mathcal{P}(f_{N,L})$ grows polynomially in N, L?
- Difference between shallow/deep approximation results?
- Role of activation function of architecture?

4) The shallow approx. result in [M96] - exponential growth

Assumptions.
- For simplicity $d = 1$ and $f \in C^q((-1,1))$ for $q \geq 2$.
- Existence of $b \in \mathbb{R}$, $\delta > 0$ and $\sigma \in C^\infty((b \pm \delta))$ with $\sigma^{(p)}(b) \neq 0$ for $p \in \mathbb{N}_0$.

Approximation. For some $C > 0$

$$\|f_N - f\|_{L^\infty((-1,1))} \leq C N^{-q} \|f\|_{C^q((-1,1))}.$$

Network.

$$f_N(x) = \sum_{0\leq r\leq p\leq k\leq 2N} C_{r,p,k}\, \sigma(h(2r-p)\cdot x + b)$$

with $C_{r,p,k}$ trigonometric coefficients depending on f, b, δ, N and h certain step size decreasing to zero for increasing N. Note that $\mathcal{W}(f_N) = \mathcal{O}(N)$.

Theorem (Exponential growth of parameters). For
- $f^{(q)}$ absolutely continuous and $f^{(q+1)}$ discontinuous
- activation $\sigma(x) = \exp(-x^2)$ or $\sigma(x) = (1+\exp(-x))^{-1}$

Exponential growth of parameters follows, i.e., there exists $(N_l)_l \subset \mathbb{N}$, $c > 0$:

$$\mathcal{P}(f_{N_l}) \gtrsim c^{N_l}.$$

5) Modified deep approx. result in [LSYZ21] - polynomial growth

Approximation. For $f \in C^q([0,1]^d)$ with Lipschitz constant $\tilde{L}$ there exist ReLU feed forward neural networks $f_{N,L}$ of width $\mathcal{W}(f_{N,L}) = \mathcal{O}(N \log N)$ and depth $\mathcal{D}(f_{N,L}) = \mathcal{O}(L^2 \log L)$ such that for some $C > 0$

$$\|f_{N,L} - f\|_{L^\infty([0,1]^d)} \leq C\|f\|_{C^q([0,1]^d)} N^{-2q/d} L^{-2q/d}.$$

Network. For e_i the i-th canonical unit vector, $f_{N,L} = \psi_d^{N,L}$ is constructed by

$$\psi_{i+1}^{N,L}(x) = \operatorname{median}(\psi_i^{N,L}(x - \delta e_{i+1}), \psi_i^{N,L}(x), \psi_i^{N,L}(x + \delta e_{i+1}))$$

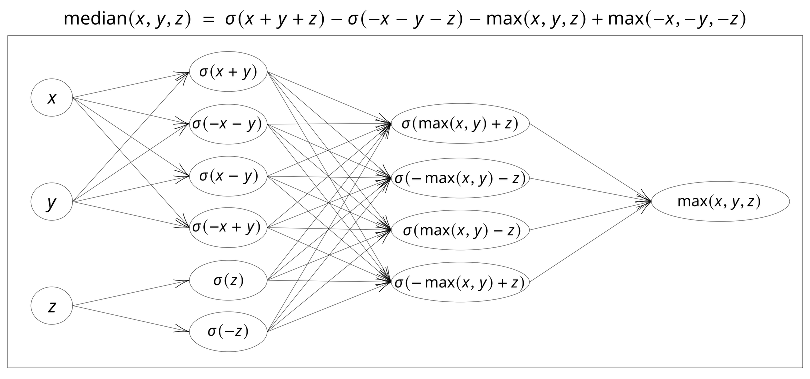
with $0 < \delta \leq d^{-1}\tilde{L}^{-1} N^{-2q/d} L^{-2q/d}$ such that

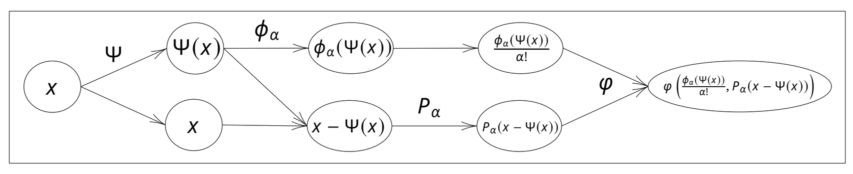
$$\psi_0^{N,L}(x) = \sum_{\|\alpha\|_1 \leq q-1} \varphi^{N,L}\left(\frac{1}{\alpha!}\phi_\alpha^{N,L}(\Psi^{N,L}(x)), P_\alpha^{N,L}(x - \Psi^{N,L}(x))\right)$$

where the ReLU FFNN
- $\Psi^{N,L}$ realize projections of subcubes of $[0,1]^d$ to one corner of subcube
- $P_\alpha^{N,L}$ approximate multinomials of order at most $q-1$
- $\phi_\alpha^{N,L}$ achieve fitting partial derivatives of f of order at most $q-1$ at the corners of the subcubes to which $\Psi^{N,L}$ projects
- $\varphi^{N,L}$ approximate binomials

Architectures.

$\operatorname{median}(x,y,z) = \sigma(x+y+z) - \sigma(-x-y-z) - \max(x,y,z) + \max(-x,-y,-z)$





Theorem [★] (Polynomial growth of parameters). It holds true that

$$\mathcal{P}(f_{N,L}) = \mathcal{O}(\max(N^{(6q-3)/d} L^{(6q-2)/d}, NL(N + L^2))).$$

6) Comparison to existing literature

For $f \in C^q([0,1]^d)$ under normalized width and $\epsilon > 0$:

Result	Width	Depth	Approximation	Growth of parameters	Activation
Th. [★]	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}(N^{\frac{-2q}{d(1+\epsilon)}} L^{\frac{-q}{d(1+\epsilon)}})$	$\mathcal{O}(N^{\frac{6q-3}{d}} L^{\frac{3q-1}{d}} \vee N^2 L^{3/2})$	ReLU
[BNPS23]	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$\mathcal{O}(N^{-q/d})$	$\mathcal{O}(1)$	ReQU
[DLM21]	$\mathcal{O}(N)$	3	$\mathcal{O}(N^{-q/d})$	$\mathcal{O}(N^{(d+q^2)/2})$	tanh
[L21]	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$\mathcal{O}(N^{-2q/d})$	$\mathcal{O}(N^{(16q+2d+9)/d})$	$\frac{1}{1+\exp(-x)}$

- Except for [BNPS23] the growth of parameters of Theorem [★] is slower in most cases (in particular $d \geq 3$).

Result	Nonzero weights	Approximation	Growth of parameters	Activation
Th. [★]	$\mathcal{O}(W)$	$\mathcal{O}(W^{-q/d})$	$\mathcal{O}(W^{\frac{9q-4}{2d} \vee \frac{7}{4}})$	ReLU
[GR21]	$\mathcal{O}(W)$	$\mathcal{O}(W^{-q/d})$	$\mathcal{O}(W^{4+2q/d})$	RePU, soft+, ...

- Growth of parameters of Theorem [★] is slower if $18q \leq 7d + 8$. For $18q > 7d + 8$ only if $5q \leq 8d + 4$.

7) Significance

- **Polynomial rates** achievable for growth of parameters for (nearly) optimal feed forward neural network approximation
- Direct consequences for **full error analysis** and neural network training
- Obtained growth for analyzed deep approximation result slower compared to literature for **high dimensional input** (except for [BNPS23] with ReQU)

8) References

[BNPS23] D. Belomestny, A. Naumov, N. Puchkin, and S. Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. Neural Networks, 2023.

[DLM21] T. De Ryck, S. Lanthaler, and S. Mishra. On the approximation of functions by tanh neural networks. Neural Networks, 2021.

[GKP20] I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep relu neural networks in $W^{s,p}$ norms. Analysis and Applications, 2020.

[GR21] I. Gühring and M. Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. Neural Networks, 2021.

[JW23] A. Jentzen and T. Welti. Overall error analysis for the training of deep neural networks via stochastic gradient descent with random initialisation. Applied Mathematics and Computation, 2023.

[L21] S. Langer. Approximating smooth functions by deep neural networks with sigmoid activation function. Journal of Multivariate Analysis, 2021.

[LSYZ21] J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. SIAM Journal on Mathematical Analysis, 2021.

[M96] H. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. Neural Computation, 1996.