UNIVERSITY OF CAMBRIDGE

Approximation properties of two-layer neural networks with values in a Banach space

Yury Korolev

Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

Analytic and Geometric Approaches to Machine Learning Symposium **29 July 2021**

Layout

Introduction:

- Scalar-valued neural networks and variation norm spaces

Contribution:

- Vector-valued neural networks and variation norm spaces

Disclaimer

New talk – new typos...

M. Thorpe

Two-layer neural networks

Two-layer neural network (NN) $f\colon \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \sum_{i=1}^{n} a_i \sigma(\langle x, b_i \rangle + c_i), \quad x \in \mathbb{R}^d,$$

where

- $\{b_i\}_{i=1}^n \subset \mathbb{R}^d$ are the weights;
- $\{c_i\}_{i=1}^n \subset \mathbb{R}$ are the biases;
- $\sigma\colon \mathbb{R} \to \mathbb{R}$ is the activation function;
- $\{\sigma(\langle x, b_i \rangle + c_i)\}_{i=1}^n$ are the neurons, collectively called the hidden layer of the network;
- $\{a_i\}_{i=1}^n \subset \mathbb{R}$ constitute the second (last) layer of the network;
- $\langle \cdot, \cdot \rangle$ denote the scalar product in $\mathbb{R}^d$.

Approximation by two-layer neural networks

Universal approximation theorems (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993)

> *If σ is not a polynomial then any continuous function on a compact set can be approximated uniformly by two-layer NNs.*

Approximation rates

- in general exponential in dimension d even for Lipschitz functions, error $O(n^{-d})$;
- Monte-Carlo rates $O(1/\sqrt{n})$ for special classes of functions (next slide).

Barron class

Theorem (Barron, 1993)

For any function f on a compact set $B \subset \mathbb{R}^d$ let F be the magnitude of its Fourier transform. For any constant $C > 0$ denote

$$\Gamma_C := \left\{ f : \mathbb{R}^d \to \mathbb{R} \quad \text{s.t.} \int |\omega| \, F(\omega) d\omega < C \right\}.$$

Then for any $n \in \mathbb{N}$ and for any $f \in \Gamma_C$ there exists a two-layer NN f_n with n neurons such that

$$\|f - f_n\|_{L^2(B)} \leqslant \frac{2C}{\sqrt{n}}.$$

The weights of the second layer $\{a_i\}_{i=1}^n$ can be chosen to satisfy

$$\sum_{i=1}^n |a_i| \leqslant 2C.$$

NB: ℓ^1 bound on $\{a_i\}_{i=1}^n$ uniform in n and depends only on C.

Remarks on the Barron class

- Barron's result assumes that σ is sigmoidal (i.e. bounded measurable satisfying $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$), but the result also holds for ReLU;
- all functions in Barron class are $\mathcal{C}^1$, hence piecewise affine functions are not in Barron class, but they can be efficiently approximated by NNs with ReLU activation.

Infinitely wide two-layer neural networks

Infinitely wide two-layer neural network $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \int_{\mathcal{A}} \sigma(\langle x, b \rangle + c)\, da(b, c), \quad x \in \mathbb{R}^d,$$

where $\mathcal{A}$ is a compact topological parameter space and $a \in \mathcal{M}(\mathcal{A})$ is a signed Radon measure. Typically $\mathcal{A} = \mathbb{B}_{\mathbb{R}^{d+1}}$.

Definition (Bach, 2017)

The space of functions that can be represented as above, equipped with the following norm

$$\|f\|_{\mathcal{F}_1} := \inf_a \{\|a\|_{\mathcal{M}} : f(x) = \int_{\mathcal{A}} \sigma(\langle x, b \rangle + c)\, da(b, c),\ x \in \mathbb{R}^d\},$$

is called the variation norm ($\mathcal{F}_1$) space.

Variation norm spaces: also known as

Variation norm ($\mathcal{F}_1$) spaces

- Bach (2017). Breaking the curse of dimensionality with convex neural networks;

Barron spaces (not to be confused with Barron class)

- E, Ma, Wu (2019). Barron spaces and the compositional function spaces for neural network models;
- E, Wojtowytsch (2020). Representation formulas and pointwise properties for Barron functions;

Radon-BV^2 spaces

- Ongie, Willett, Soudry, Srebro (2020). A function space view of bounded norm infinite width ReLU nets: The multivariate case;
- Parhi, Nowak (2021). Banach space representer theorems for neural networks and ridge splines;

Mean field approach

- Rotskoff, Vanden-Eijnden (2018). Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks;
- Mei, Montanari, Nguyen (2018). A mean field view of the landscape of two-layer neural networks;
- Chizat, Bach (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport;
- Sirignano, Spiliopoulos (2020). Mean field analysis of neural networks: A law of large numbers.

Linear-nonlinear decomposition

Linear-nonlinear decomposition of a two-layer NN $f\colon \mathbb{R}^d \to \mathbb{R}$

$$f(x) = A\sigma(Bx + c), \quad x \in \mathbb{R}^d,$$

where

$$B\colon \mathbb{R}^d \to \mathbb{R}^n, \quad c \in \mathbb{R}^n \quad \text{and} \quad A\colon \mathbb{R}^n \to \mathbb{R}$$

for a NN with $n < \infty$ neurons

and

$$B\colon \mathbb{R}^d \to \mathcal{C}(\mathbb{R}^{d+1}), \quad c \in \mathcal{C}(\mathbb{R}^{d+1}) \quad \text{and} \quad A\colon \mathcal{C}(\mathbb{R}^{d+1}) \to \mathbb{R}$$

for an infinitely wide NN.

If σ is positively one-homogeneous, parameters can be chosen on the unit ball $\mathbb{B}_{\mathbb{R}^{d+1}}$. In this case $\mathcal{C}(\mathbb{R}^{d+1})$ is replaced by $\mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}})$.

(E and Wojtowytsch, 2020)

Linear-nonlinear decomposition and variation norm

Linear-nonlinear decomposition of an infinitely wide two-layer NN $f\colon \mathbb{R}^d \to \mathbb{R}$

$$f(x) = A\sigma(Bx + c), \quad x \in \mathbb{R}^d,$$

where

$$B\colon \mathbb{R}^d \to \mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}}), \quad c \in \mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}}) \quad \text{and} \quad A\colon \mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}}) \to \mathbb{R}.$$

That is, A is a linear functional on $\mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}})$ and can be identified with a Radon measure $a \in \mathcal{M}(\mathbb{B}_{\mathbb{R}^{d+1}})$. Then

$$\|f\|_{\mathcal{F}_1} = \inf_a \{\|a\|_{\mathcal{M}} : f(x) = \langle \sigma(Bx + c),\, a\rangle\,,\ x \in \mathbb{R}^d\},$$

where $\langle \cdot, \cdot \rangle$ is the dual pairing between $\mathcal{C}(\mathbb{B}_{\mathbb{R}^{d+1}})$ and $\mathcal{M}(\mathbb{B}_{\mathbb{R}^{d+1}})$.

(more details to follow)

Monte-Carlo rates in variation norm spaces

Theorem (direct approximation; E, Ma and Wu, 2019)

Let $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ be a probability measure with $p \geqslant 1$ finite moments and let $f \in \mathcal{F}_1(\mathbb{R}^d)$. Then for any $n \in \mathbb{N}$ there exists a two-layer NN f_n with n neurons such that

$$\|f - f_n\|_{L^2_\mu(\mathbb{R}^d)} \leqslant \frac{2\|f\|_{\mathcal{F}_1}}{\sqrt{n}}$$

and

$$\sum_{i=1}^n |a_i| \|b_i\| \leqslant 2\|f\|_{\mathcal{F}_1} .$$

Cf. Barron's theorem: $\|f\|_{\mathcal{F}_1}$ plays the role of C in the Barron class Γ_C.

Inverse approximation also holds (E, Ma and Wu, 2019).

Layout

Introduction:

Scalar-valued neural networks and variation norm spaces

Contribution:

Vector-valued neural networks and variation norm spaces

Learning in infinite-dimensional spaces

Reproducing kernel Hilbert/Banach spaces a.k.a. random feature models

- Micchelli, Pontil (2005). On learning vector-valued functions;
- Zhang, Zhang (2013). Vector-valued reproducing kernel Banach spaces with applications to multi-task learning;
- Álvarez, Rosasco, Lawrence (2012). Kernels for vector-valued functions: A review;
- Nelsen, Stuart (2020). The random feature model for input-output maps between Banach spaces.

Apparently, no such results for variation norm spaces.

Vector-valued two-layer neural networks – 1

Vector-valued two-layer NN $f\colon \mathcal{X} \to \mathcal{Y}$

$$f(x) = A\sigma(Bx + c), \quad x \in \mathcal{X},$$

where

- $\mathcal{X}$, $\mathcal{Y}$ are Banach spaces with separable preduals,
- σ is the generalised ReLU function that we will define using partially ordered spaces (vector lattices),
- A, B and c are yet to be defined.

We will slightly abuse notation and write

$$f(x) = A\sigma(Bx), \quad x \in \mathcal{X},$$

where we have identified $\mathcal{X}$ with $\mathcal{X} \times \mathbb{R}$ and B with an operator (B, c) acting on $\mathcal{X} \times \mathbb{R}$ as $(x, \alpha) \mapsto Bx + \alpha c$.

For inputs of the form $(x, 1)$ the two formulas are the same.

Vector lattices, a.k.a. Riesz spaces

Partial order " $\leqslant$ " on a set S is a reflexive, antisymmetric and transitive binary relation " $\leqslant$ " $\subset S \times S$.

Vector lattices, a.k.a. Riesz spaces

Partial order " $\leqslant$ " on a set S is a reflexive, antisymmetric and transitive binary relation " $\leqslant$ " $\subset S \times S$.

Vector space $\mathcal{X}$ with partial order " $\leqslant$ " called an *ordered vector space* if

$$x \leqslant y \implies x + z \leqslant y + z \qquad \forall\ x, y, z \in \mathcal{X},$$
$$x \leqslant y \implies \lambda x \leqslant \lambda y \qquad \forall\ x, y \in \mathcal{X} \text{ and } \lambda \in \mathbb{R}_+.$$

Vector lattices, a.k.a. Riesz spaces

Partial order " $\leqslant$ " on a set S is a reflexive, antisymmetric and transitive binary relation " $\leqslant$ " $\subset S \times S$.

Vector space $\mathcal{X}$ with partial order " $\leqslant$ " called an *ordered vector space* if

$$x \leqslant y \implies x + z \leqslant y + z \qquad \forall\ x, y, z \in \mathcal{X},$$
$$x \leqslant y \implies \lambda x \leqslant \lambda y \qquad \forall\ x, y \in \mathcal{X} \text{ and } \lambda \in \mathbb{R}_+.$$

A *vector lattice* (or a *Riesz space*) is an ordered vector space $\mathcal{X}$ with well defined suprema and infima

$$\forall x, y \in \mathcal{X} \quad \exists\, x \vee y \in \mathcal{X},\ x \wedge y \in \mathcal{X};$$
$$x \vee 0 = x_+, \quad (-x)_+ = x_-, \quad x = x_+ - x_-, \quad |x| = x_+ + x_-.$$

Examples of vector lattices

- Sequence spaces ℓ^p, $1 \leqslant p \leqslant \infty$

$$x \geqslant y \iff x^i \geqslant y^i \quad i \in \mathbb{N};$$

- Space of signed Radon measures $\mathcal{M}(\Omega)$

$$\mu \geqslant \nu \iff \mu(A) \geqslant \nu(A) \quad \forall A \subset \Omega;$$

- Lebesgue spaces $\mathcal{L}^p$, $1 \leqslant p \leqslant \infty$

$$f \geqslant g \iff f(x) \geqslant g(x) \quad \text{a.e. in } \Omega;$$

- Space of continuous functions $\mathcal{C}(\Omega)$, space of Lipschitz functions $\mathrm{Lip}(\Omega)$

$$f \geqslant g \iff f(x) \geqslant g(x) \quad \forall x \in \Omega;$$

- Space of functions of bounded variation on an interval $\mathrm{BV}([0,1])$

$$f \geqslant g \iff f(\cdot) - g(\cdot) \quad \text{is non-decreasing;}$$

- Space of linear operators between two partially ordered spaces $\mathcal{L}^r(\mathcal{X};\mathcal{Y})$

$$A \geqslant B \iff \forall x \geqslant 0 \text{ it holds that } Ax \geqslant Bx.$$

Vector-valued two-layer neural networks – 2

Vector-valued two-layer NN $f\colon \mathcal{X} \to \mathcal{Y}$

$$f(x) = A\sigma(Bx), \quad x \in \mathcal{X},$$

where

$\mathcal{X}$, $\mathcal{Y}$ have separable preduals and $\mathcal{Y}$ is also a vector lattice,

$\sigma\colon \mathcal{Y} \to \mathcal{Y}$ is the generalised ReLU function,

$$\sigma(y) := y_+ = y \vee 0 \quad \text{in the lattice sense,}$$

$B\colon \mathcal{X} \to \mathcal{C}(\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}; \mathcal{Y})$ maps

$$x \mapsto \mathcal{L}_x(\cdot) \quad \text{such that} \quad \mathcal{L}_x(K) = Kx,$$

$A\colon \mathcal{C}(\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}; \mathcal{Y}) \to \mathcal{Y}$ maps

$$\varphi(\cdot) \mapsto \int_{\mathbb{B}_{\mathcal{L}}} \varphi(K)\, da(K), \quad \text{where } a \in \mathcal{M}(\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}).$$

Caveats – 1

The parameter space is $\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}$. To make sure it is compact, we need to

- make sure that $\mathcal{L}(\mathcal{X};\mathcal{Y})$ is a dual space and
- use the weak* topology.

Theorem (Ryan. Introduction to tensor products of Banach spaces. 2002)

Suppose that $\mathcal{X}$ and $\mathcal{Y}$ have separable preduals $\mathcal{X}^\diamond$ and $\mathcal{Y}^\diamond$ and that either $\mathcal{X}$ or $\mathcal{Y}^\diamond$ has the approximation property. Then the dual of the space of nuclear operators $\mathcal{N}(\mathcal{Y}^\diamond;\mathcal{X}^\diamond)$ can be identified with the space of bounded operators $\mathcal{L}(\mathcal{X};\mathcal{Y})$

$$(\mathcal{N}(\mathcal{Y}^\diamond;\mathcal{X}^\diamond))^* \simeq \mathcal{L}(\mathcal{X};\mathcal{Y}).$$

Consequently, the unit ball $\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}$ is weakly compact and metrisable.*

Caveats – 2

Since $\mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})}$ is equipped with the weak* topology, we need to make sure that

- the function $\mathcal{L}_x : \mathbb{B}_{\mathcal{L}(\mathcal{X};\mathcal{Y})} \to \mathcal{Y}$ such that $\mathcal{L}_x(K) = Kx$ is weakly-* continuous $\quad \to$ true if $\mathcal{Y}$ is equipped with the weak* topology;
- the nonlinearity σ is weakly-* continuous
  $\to$ turns out to be quite restrictive for the ReLU!

Examples:

✔ Sequence spaces ℓ^p, $p > 1$; Lipschitz space $\mathrm{Lip}(\Omega)$;

✗ Lebesgue spaces L^p_μ (unless μ is atomic); space of linear operators $\mathcal{L}^r(\mathcal{X};\mathcal{Y})$ (except in special cases); space of Radon measures $\mathcal{M}(\Omega)$ (unless Ω is discrete).

Caveats – 3

In order to obtain convergence rates in Bochner spaces L^p, we need to metrise the weak* topology on the unit ball in $\mathcal{Y}$. This is typically done using the following metric

$$d(y,z) = \sum_{i=1}^{\infty} 2^{-i} \frac{|\langle \eta_i,\, y-z\rangle|}{1+|\langle \eta_i,\, y-z\rangle|},$$

where $\{\eta_i\}_{i\in\mathbb{N}}$ is a countable dense system in the predual.

If $\{\eta_i\}_{i\in\mathbb{N}}$ are normalised, the following equivalent metric can be used

$$d_*(y,z) = \sum_{i=1}^{\infty} 2^{-i} \left|\langle \eta_i,\, y-z\rangle\right|.$$

It can be used to define a norm in the weak* completion of $\mathcal{Y}$.

Vector-valued variation norm space

Definition (Vector-valued $\mathcal{F}_1$ functions)

Let $\mathcal{X}, \mathcal{Y}$ have separable preduals and let $\mathcal{Y}$ be such that lattice operations are 1-Lipschitz with respect to the d_* metric. We define the space of $\mathcal{Y}$-valued $\mathcal{F}_1$ functions as follows

$$\mathcal{F}_1(\mathcal{X}; \mathcal{Y}) := \{f \in \mathrm{Lip}_0 : \ \|f\|_{\mathcal{F}_1} < \infty\},$$

where Lip_0 is the space of Lipschitz functions with respect to the d_* metric in $\mathcal{Y}$ that vanish at zero and

$$\|f\|_{\mathcal{F}_1} := \inf_{a \in \mathcal{M}(\mathbb{B}_{\mathcal{L}})} \{\|a\|_{\mathcal{M}} : f(x) = \int_{\mathbb{B}_{\mathcal{L}}} \sigma(\mathcal{L}_x(K))\, da(K) \ \forall x \in \mathcal{X}\}.$$

Monte-Carlo rates in vector-valued $\mathcal{F}_1$ spaces

Theorem (direct approximation; YK 2021)

Let above assumptions be satisfied and let $f \in \mathcal{F}_1(\mathcal{X};\mathcal{Y})$. Then for any $n \in \mathbb{N}$ there exists a two-layer neural network with n neurons

$$f_n(x) := \sum_{i=1}^{n} \alpha_i (K_i x)_+, \quad x \in \mathcal{X},$$

where K_i have finite rank and $\|K_i\|_{\mathcal{L}(\mathcal{X};\mathcal{Y})} \leqslant 1$, such that

- *for any $x \in \mathcal{X}$*

$$d_* \left(f(x), f_n(x)\right) \leqslant \frac{2\sqrt{2}\,\|f\|_{\mathcal{F}_1}\,\|x\|}{\sqrt{n}};$$

- *if $\mu \in \mathcal{P}_p(\mathcal{X})$ and $m_p(\mu) < \infty$ is its p-th moment, $p \geqslant 1$, then*

$$\|f - f_n\|_{L^p_\mu} \leqslant \frac{2\sqrt{2}\,\|f\|_{\mathcal{F}_1}\,(m_p(\mu))^{\frac{1}{p}}}{\sqrt{n}}.$$

Conclusions

- ✔ Generalised variation norm spaces with ReLU activation to networks with values in a Banach space;
- ✔ Proved inverse and direct approximation theorems, obtained Monte-Carlo rates;
- ✔ Obtained results for generalised ReLU, but they hold for any weakly-* continuous activation;
- ✗ Saw a limitation – weak* continuity of σ often not fulfilled by ReLU $\rightarrow$ Is the use of weak* topologies a technicality?
- ✗ More interesting architectures.

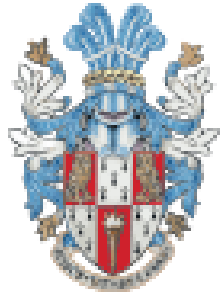
YK (2021). Two-layer neural networks with values in a Banach space. arXiv:2105.02095

So long, and thanks for all the ~~fish~~ funding

Engineering and
Physical Sciences
Research Council

CCIMI
CANTAB CAPITAL INSTITUTE FOR THE
MATHEMATICS OF INFORMATION

NPL
National Physical Laboratory



HUGHES HALL
UNIVERSITY OF CAMBRIDGE