Approximation Properties of Neural Networks

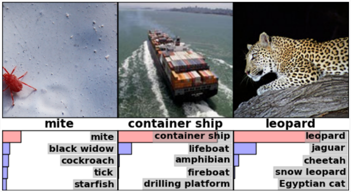
Felix Voigtlaender
`http://voigtlaender.xyz`

Katholische Universität Eichstätt-Ingolstadt
MIDS
Mathematical Institute for Machine Learning and Data Science

Workshop and Summer School on Applied Analysis 2023
Chemnitz, Germany, 18-22 September 2023

Deep learning dramatically changed what computers can do

Image recognition

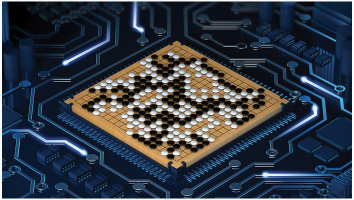


www.infoq.com/presentations/deepmind-q-network

Autonomous driving

www.lindsaysing.com/austin-tech-alliance/

Game intelligence



heise.de

Speech recognition

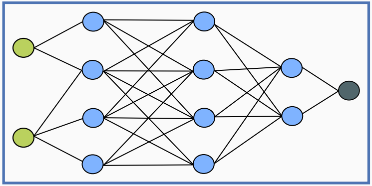
www.quantiphi.com/portfolio-posts/speech-recognition/

“Deep learning” roughly means: Adjust weights of a deep neural network based on training data

Labelled training examples (x_i, y_i)

Loss function, e.g. $\sum_{i=1}^{N} \|y_i - \Phi_W(x_i)\|^2$

Adjust weights



Neural network Φ_W

The performance of a machine learning system is influenced by Expressiveness, Generalization, and Optimization

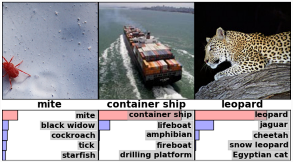
- $\mathcal{X} \times \mathcal{Y}$: set of all possible (input, label) pairs
- $\mathbb{P}$: "ground truth" distribution on $\mathcal{X} \times \mathcal{Y}$ (unknown)

Goal: Minimize the (expected) risk

$$R(f) := \mathbb{P}\big(f(X) \neq Y\big),$$

given only training sample

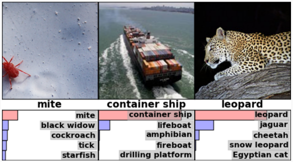
$S = ((X_1, Y_1), \ldots, (X_N, Y_N)) \overset{\text{iid}}{\sim} \mathbb{P}$.



www.infoq.com/presentations/deepmind-q-network

The performance of a machine learning system is influenced by Expressiveness, Generalization, and Optimization

- $\mathcal{X} \times \mathcal{Y}$: set of all possible (input, label) pairs
- $\mathbb{P}$: "ground truth" distribution on $\mathcal{X} \times \mathcal{Y}$ (unknown)



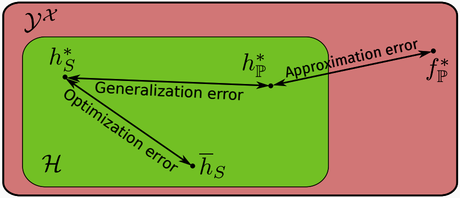
www.infoq.com/presentations/deepmind-q-network

Goal: Minimize the (expected) risk

$$R(f) := \mathbb{P}(f(X) \neq Y),$$

given only training sample

$S = ((X_1, Y_1), \ldots, (X_N, Y_N)) \overset{\text{iid}}{\sim} \mathbb{P}$.

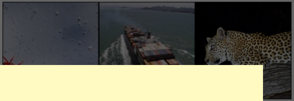


e.g. $h_S^* = \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{N} \mathbb{1}_{h(X_i) \neq Y_i}$

or $h_S^* = \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{N} \|h(X_i) - Y_i\|^2$

The performance of a machine learning system is influenced by Expressiveness, Generalization, and Optimization

- $\mathcal{X} \times \mathcal{Y}$: set of all possible (input, label) pairs
- $\mathbb{P}$: “ground truth” distribution on $\mathcal{X} \times \mathcal{Y}$ (unknown)

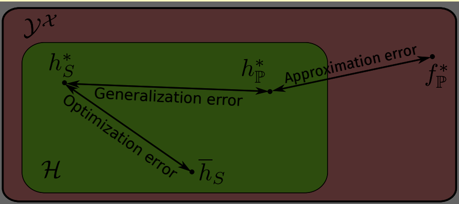


Go

giv

$S =$

In this lecture, we only consider the **approximation** error!

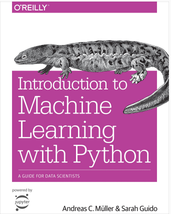


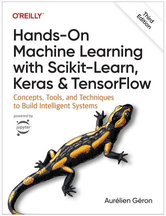
e.g. $h_S^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{1}_{h(X_i) \neq Y_i}$

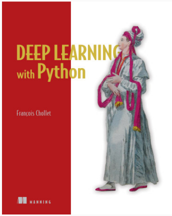
or $h_S^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} \|h(X_i) - Y_i\|^2$

Book recommendations regarding the basics of machine learning

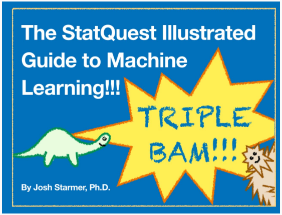
Practice:

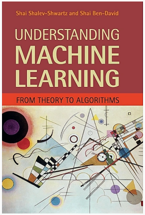






Basic principles and theory:





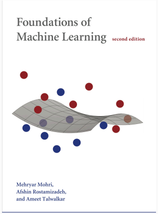
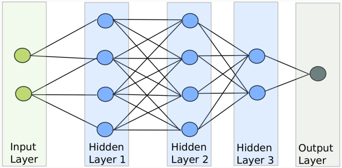
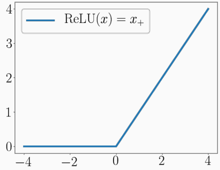


Table of contents

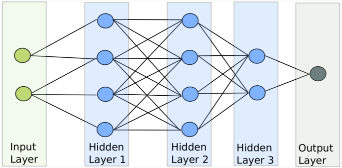
The basics of neural networks

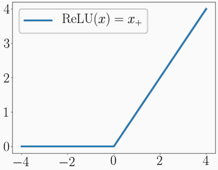
A neural network repeatedly applies affine-linear maps and an activation function





A neural network repeatedly applies affine-linear maps and an activation function





- L: number of (hidden) layers,
- $(N_0, \ldots, N_{L+1})$: neurons per layer
- $T_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_{\ell+1}}, x \mapsto A_\ell x + b_\ell$: connections between neurons (weights),
- $\varrho : \mathbb{R} \to \mathbb{R}$: activation function.

Neural network: $\Phi = (T_0, \ldots, T_L)$

Network function (Realization):

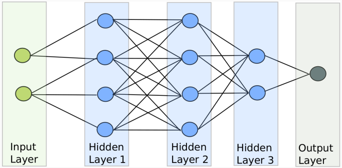
$R_\varrho(\Phi) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$, given by

$$R_\varrho(\Phi) = T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_0)$$

with ϱ applied componentwise, i.e.,

$$\varrho\big((x_1, \ldots, x_K)\big) = \big(\varrho(x_1), \ldots, \varrho(x_K)\big).$$

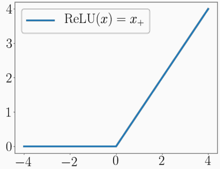
A neural network repeatedly applies affine-linear maps and an activation function



$L(\Phi) = 3$

$N(\Phi) = 13$

$W(\Phi) = \sum_{i=0}^{L} \|A_i\|_{\ell^0} = 34$



- L: number of (hidden) layers,
- $(N_0, \ldots, N_{L+1})$: neurons per layer
- $T_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_{\ell+1}}, x \mapsto A_\ell x + b_\ell$: connections between neurons (weights),
- $\varrho : \mathbb{R} \to \mathbb{R}$: activation function.

Neural network: $\Phi = (T_0, \ldots, T_L)$

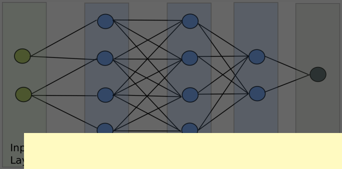
Network function (Realization): $R_\varrho(\Phi) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$, given by

$$R_\varrho(\Phi) = T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_0)$$

with ϱ applied componentwise, i.e.,

$$\varrho((x_1, \ldots, x_K)) = (\varrho(x_1), \ldots, \varrho(x_K)).$$

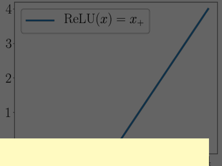
A neural network repeatedly applies affine-linear maps and an activation function



$L(\Phi) = 3$

$N(\Phi) = 13$

$W(\Phi) = \sum_{j=0}^{L} \|A_j\|_{\ell^0} = 34$



These NNs are called **fully connected feed-forward NNs**.

There are other important types of NNs, e.g. CNNs, RNNs, and Transformers.

- $T_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_{\ell+1}}, x \mapsto A_\ell x + b_\ell$: connections between neurons (weights),
- $\varrho : \mathbb{R} \to \mathbb{R}$: activation function.

$$R_\varrho(\Phi) = T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_0)$$

with ϱ applied componentwise, i.e.,

$$\varrho((x_1, \ldots, x_K)) = (\varrho(x_1), \ldots, \varrho(x_K)).$$

The universal approximation theorem

The universal approximation theorem characterizes activation functions for which the associated class of NNs is universal

The universal approximation theorem characterizes activation functions for which the associated class of NNs is universal

A function class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ is called universal if

$$\forall\, g \in C(\mathbb{R}^d), \quad \varepsilon > 0, \quad K \subset \mathbb{R}^d \text{ compact} \quad \exists f \in \mathcal{F} : \quad \sup_{x \in K} |g(x) - f(x)| \leq \varepsilon.$$

The universal approximation theorem characterizes activation functions for which the associated class of NNs is universal

A function class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ is called universal if

$$\forall\, g \in C(\mathbb{R}^d), \quad \varepsilon > 0, \quad K \subset \mathbb{R}^d \text{ compact} \quad \exists f \in \mathcal{F} : \quad \sup_{x \in K} |g(x) - f(x)| \leq \varepsilon.$$

Question: For which activation functions $\varrho \in C(\mathbb{R})$ is the set

$$\mathcal{NN}_\varrho^d := \left\{ x \mapsto \sum_{i=1}^{N} c_i\, \varrho(\langle w_i, x \rangle + b_i) \;:\; N \in \mathbb{N}, w_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R} \right\}$$

of all shallow neural networks with activation function ϱ universal?

The universal approximation theorem characterizes activation functions for which the associated class of NNs is universal

A function class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ is called universal if

$$\forall\, g \in C(\mathbb{R}^d), \quad \varepsilon > 0, \quad K \subset \mathbb{R}^d \text{ compact} \quad \exists f \in \mathcal{F} : \quad \sup_{x \in K} |g(x) - f(x)| \le \varepsilon.$$

Question: For which activation functions $\varrho \in C(\mathbb{R})$ is the set

$$\mathcal{NN}_\varrho^d := \left\{ x \mapsto \sum_{i=1}^{N} c_i\, \varrho(\langle w_i, x\rangle + b_i) \;:\; N \in \mathbb{N}, w_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R} \right\}$$

of all shallow neural networks with activation function ϱ universal?

Quiz: For which activation functions does universality definitely fail?

The universal approximation theorem characterizes activation functions for which the associated class of NNs is universal

A function class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ is called universal if

$$\forall g \in C(\mathbb{R}^d), \quad \varepsilon > 0, \quad K \subset \mathbb{R}^d \text{ compact} \quad \exists f \in \mathcal{F} : \quad \sup_{x \in K} |g(x) - f(x)| \leq \varepsilon.$$

Question: For which activation functions $\varrho \in C(\mathbb{R})$ is the set

$$\mathcal{NN}_\varrho^d := \left\{ x \mapsto \sum_{i=1}^{N} c_i \, \varrho(\langle w_i, x \rangle + b_i) \ : \ N \in \mathbb{N}, w_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R} \right\}$$

of all shallow neural networks with activation function ϱ universal?

Quiz: For which activation functions does universality definitely fail?

Universal approximation theorem (Leshno, Lin, Pinkus, Schocken; 1993).
Let $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous. Then

$$\mathcal{NN}_\varrho^d \text{ is universal} \quad \Longleftrightarrow \quad \varrho \text{ is not a polynomial.}$$

Proof of the universal approximation theorem — Part 0

Stone-Weierstraß theorem. Let X be a compact Hausdorff space. If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$ that separates points, then either $\mathcal{A} = C(X, \mathbb{R})$ or $\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(x_0) = 0\}$ for some $x_0 \in X$.

Stone-Weierstraß theorem. Let X be a compact Hausdorff space. If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$ that separates points, then either $\mathcal{A} = C(X, \mathbb{R})$ or $\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(x_0) = 0\}$ for some $x_0 \in X$.

Remarks:

1. $\mathcal{A}$ being an algebra means it is a vector space and closed under multiplication.
2. $\mathcal{A}$ separates the points if for all $x, y \in X$ with $x \neq y$ there exists $f \in \mathcal{A}$ satisfying $f(x) \neq f(y)$.

Proof.

See Theorem 4.45 in Folland's "Real Analysis" book. □

Proof of the universal approximation theorem — Part 0

Stone-Weierstraß theorem. Let X be a compact Hausdorff space. If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$ that separates points, then either $\mathcal{A} = C(X, \mathbb{R})$ or $\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(x_0) = 0\}$ for some $x_0 \in X$.

Remarks:

1. $\mathcal{A}$ being an algebra means it is a vector space and closed under multiplication.
2. $\mathcal{A}$ separates the points if for all $x, y \in X$ with $x \neq y$ there exists $f \in \mathcal{A}$ satisfying $f(x) \neq f(y)$.

Proof.

See Theorem 4.45 in Folland's "Real Analysis" book. □

Example applications:

1. $\mathbb{R}[X] \subset C([a, b])$ is dense for $a < b$ (why?!).

Proof of the universal approximation theorem — Part 0

Stone-Weierstraß theorem. Let X be a compact Hausdorff space. If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$ that separates points, then either $\mathcal{A} = C(X, \mathbb{R})$ or $\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(x_0) = 0\}$ for some $x_0 \in X$.

Remarks:

1. $\mathcal{A}$ being an algebra means it is a vector space and closed under multiplication.
2. $\mathcal{A}$ separates the points if for all $x, y \in X$ with $x \neq y$ there exists $f \in \mathcal{A}$ satisfying $f(x) \neq f(y)$.

Proof.

See Theorem 4.45 in Folland's "Real Analysis" book. □

Example applications:

1. $\mathbb{R}[X] \subset C([a, b])$ is dense for $a < b$ (why?!).
2. $\mathrm{span}\{e^{\langle a,x \rangle} : a \in \mathbb{R}^d\} \subset C(K)$ is dense for any compact set $\varnothing \neq K \subset \mathbb{R}^d$.

Excursion: Dynkin's multiplicative system theorem is a "measure-theoretic analogue" of the Stone-Weierstraß theorem

Excursion: Dynkin's multiplicative system theorem is a "measure-theoretic analogue" of the Stone-Weierstraß theorem

Let $X \neq \varnothing$ be a set and $\ell^\infty(X) = \{f : X \to \mathbb{R} : \ f \text{ bounded}\}$.

Dynkin's multiplicative system theorem. Let $\mathcal{F} \subset \ell^\infty(X)$ be closed under multiplication and suppose that $\mathcal{A}$ satisfies the following:

1. $\mathcal{A}$ is a subspace of $\ell^\infty(X)$;
2. $\mathcal{F} \subset \mathcal{A}$ and $\mathbb{1}_X \in \mathcal{A}$;
3. $\mathcal{A}$ is closed under bounded pointwise convergence, i.e., whenever $(f_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ satisfies $f_n \to f$ pointwise and $\sup_{n\in\mathbb{N}} \sup_{x\in X} |f_n(x)| < \infty$, then $f \in \mathcal{A}$.

Then $\{f \in \ell^\infty(X) \ : \ f \text{ measurable with respect to } \sigma(\mathcal{F})\} \subset \mathcal{A}$.

Excursion: Dynkin's multiplicative system theorem is a "measure-theoretic analogue" of the Stone-Weierstraß theorem

Let $X \neq \varnothing$ be a set and $\ell^\infty(X) = \{f : X \to \mathbb{R} : f \text{ bounded}\}$.

Dynkin's multiplicative system theorem. Let $\mathcal{F} \subset \ell^\infty(X)$ be closed under multiplication and suppose that $\mathcal{A}$ satisfies the following:

1. $\mathcal{A}$ is a subspace of $\ell^\infty(X)$;
2. $\mathcal{F} \subset \mathcal{A}$ and $\mathbb{1}_X \in \mathcal{A}$;
3. $\mathcal{A}$ is closed under bounded pointwise convergence, i.e., whenever $(f_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ satisfies $f_n \to f$ pointwise and $\sup_{n\in\mathbb{N}} \sup_{x\in X} |f_n(x)| < \infty$, then $f \in \mathcal{A}$.

Then $\{f \in \ell^\infty(X) \ : \ f \text{ measurable with respect to } \sigma(\mathcal{F})\} \subset \mathcal{A}$.

Example application: The set $\operatorname{span}\{e^{-\lambda x} \colon \lambda > 0\} \subset L^2((0,\infty))$ is dense.

Proof: Let $\mathcal{F} = \{e^{-\lambda x} \colon \lambda > 0\} \subset \ell^\infty((0,\infty))$, let $g \in L^2((0,\infty))$ be orthogonal to $\mathcal{F}$, and let $\mathcal{A} = \{f \in \ell^\infty((0,\infty)) : f \text{ measurable and } \langle g \cdot e^{-x}, f\rangle = 0\}$.

Proof of Dynkin’s multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

Proof of Dynkin's multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

1. It is enough to show $\mathbb{1}_M \in \mathcal{A}_0 \subset \mathcal{A}$ for each $M \in \sigma(\mathcal{F})$.

   Reason: Each $\sigma(\mathcal{F})$-measurable $f \in \ell^\infty(X)$ can be approximated by simple functions $\sum_{i=1}^N c_i \, \mathbb{1}_{M_i}$ with $M_i \in \sigma(\mathcal{F})$ (with bounded p.w. convergence).

Proof of Dynkin's multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

1. It is enough to show $\mathbb{1}_M \in \mathcal{A}_0 \subset \mathcal{A}$ for each $M \in \sigma(\mathcal{F})$.

   Reason: Each $\sigma(\mathcal{F})$-measurable $f \in \ell^\infty(X)$ can be approximated by simple functions $\sum_{i=1}^N c_i \, \mathbb{1}_{M_i}$ with $M_i \in \sigma(\mathcal{F})$ (with bounded p.w. convergence).
2. Let $\mathcal{G} := \{M \in \sigma(\mathcal{F}) \colon \mathbb{1}_M \in \mathcal{A}_0\}$. Then $\mathcal{G}$ is a λ-system (closed under complementation and countable disjoint unions).

Proof of Dynkin's multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

1. It is enough to show $\mathbb{1}_M \in \mathcal{A}_0 \subset \mathcal{A}$ for each $M \in \sigma(\mathcal{F})$.

   Reason: Each $\sigma(\mathcal{F})$-measurable $f \in \ell^\infty(X)$ can be approximated by simple functions $\sum_{i=1}^N c_i \, \mathbb{1}_{M_i}$ with $M_i \in \sigma(\mathcal{F})$ (with bounded p.w. convergence).
2. Let $\mathcal{G} := \{M \in \sigma(\mathcal{F}) \colon \mathbb{1}_M \in \mathcal{A}_0\}$. Then $\mathcal{G}$ is a λ-system (closed under complementation and countable disjoint unions).
3. Easy: $\mathcal{A}_0$ is closed under multiplication, since $\mathcal{F}$ is.

   $\Longrightarrow \mathcal{G}$ is a π-system (closed under intersection).

   $\Longrightarrow \mathcal{G}$ is a σ-algebra, by Dynkin's π-λ-theorem.

   Hence, it is enough to show that $\{f^{-1}((a,b)) \colon f \in \mathcal{F}, a < b\} \subset \mathcal{G}$.

Proof of Dynkin's multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

1. It is enough to show $\mathbb{1}_M \in \mathcal{A}_0 \subset \mathcal{A}$ for each $M \in \sigma(\mathcal{F})$.

   Reason: Each $\sigma(\mathcal{F})$-measurable $f \in \ell^\infty(X)$ can be approximated by simple functions $\sum_{i=1}^N c_i \, \mathbb{1}_{M_i}$ with $M_i \in \sigma(\mathcal{F})$ (with bounded p.w. convergence).
2. Let $\mathcal{G} := \{M \in \sigma(\mathcal{F}) \colon \mathbb{1}_M \in \mathcal{A}_0\}$. Then $\mathcal{G}$ is a λ-system (closed under complementation and countable disjoint unions).
3. Easy: $\mathcal{A}_0$ is closed under multiplication, since $\mathcal{F}$ is.
   $\implies \mathcal{G}$ is a π-system (closed under intersection).
   $\implies \mathcal{G}$ is a σ-algebra, by Dynkin's π-λ-theorem.
   Hence, it is enough to show that $\{f^{-1}((a,b)) \colon f \in \mathcal{F}, a < b\} \subset \mathcal{G}$.
4. For each $\varphi \in C(\mathbb{R})$ and $f \in \mathcal{A}_0$, we have $\varphi \circ f \in \mathcal{A}_0$.

   Reason: For polynomials $\varphi = p$ this is clear, since $\mathcal{A}_0$ is closed under multiplication. Approximate φ uniformly on $\mathrm{range}(f)$ by polynomials p_n.

Proof of Dynkin's multiplicative system theorem

Let $\mathcal{A}_0$ be the minimal set satisfying properties ❶–❸.

1. It is enough to show $\mathbb{1}_M \in \mathcal{A}_0 \subset \mathcal{A}$ for each $M \in \sigma(\mathcal{F})$.

   Reason: Each $\sigma(\mathcal{F})$-measurable $f \in \ell^\infty(X)$ can be approximated by simple functions $\sum_{i=1}^N c_i \, \mathbb{1}_{M_i}$ with $M_i \in \sigma(\mathcal{F})$ (with bounded p.w. convergence).
2. Let $\mathcal{G} := \{M \in \sigma(\mathcal{F}) \colon \mathbb{1}_M \in \mathcal{A}_0\}$. Then $\mathcal{G}$ is a λ-system (closed under complementation and countable disjoint unions).
3. Easy: $\mathcal{A}_0$ is closed under multiplication, since $\mathcal{F}$ is.
   $\Longrightarrow \mathcal{G}$ is a π-system (closed under intersection).
   $\Longrightarrow \mathcal{G}$ is a σ-algebra, by Dynkin's π-λ-theorem.
   Hence, it is enough to show that $\{f^{-1}((a,b)) \colon f \in \mathcal{F}, a < b\} \subset \mathcal{G}$.
4. For each $\varphi \in C(\mathbb{R})$ and $f \in \mathcal{A}_0$, we have $\varphi \circ f \in \mathcal{A}_0$.

   Reason: For polynomials $\varphi = p$ this is clear, since $\mathcal{A}_0$ is closed under multiplication. Approximate φ uniformly on $\mathrm{range}(f)$ by polynomials p_n.
5. Pick $\varphi_n \in C(\mathbb{R})$ with $0 \leq \varphi_n \leq 1$ and $\varphi_n \to \mathbb{1}_{(a,b)}$ pointwise.
   Then $\varphi_n \circ f \to \mathbb{1}_{(a,b)} \circ f = \mathbb{1}_{f^{-1}((a,b))}$ pointwise boundedly. □

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0, \; K \subset \mathbb{R}^d \text{ compact } \; \exists \tilde{f} \in \mathcal{F} : \; \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F}: \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"):

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F}: \ \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F} : \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

$\Longrightarrow$ If $f \in C(\mathbb{R}^d)$ is not a polynomial of degree at most D, it cannot be approximated by elements of $\mathcal{NN}_\varrho^d$ (why?!).

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0, \ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F} : \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

$\Longrightarrow$ If $f \in C(\mathbb{R}^d)$ is not a polynomial of degree at most D, it cannot be approximated by elements of $\mathcal{NN}_\varrho^d$ (why?!).

Step 1 (Reduction to $d = 1$): Claim: If $\mathcal{NN}_\varrho^1$ is universal, then so is $\mathcal{NN}_\varrho^d$.

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F}: \ \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

$\Longrightarrow$ If $f \in C(\mathbb{R}^d)$ is not a polynomial of degree at most D, it cannot be approximated by elements of $\mathcal{NN}_\varrho^d$ (why?!).

Step 1 (Reduction to $d = 1$): Claim: If $\mathcal{NN}_\varrho^1$ is universal, then so is $\mathcal{NN}_\varrho^d$.

Substep ❶: Universality of $\mathcal{NN}_\varrho^1 \quad \Longrightarrow \quad \exp \in \overline{\mathcal{NN}_\varrho^1}$.

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F}: \ \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

$\Longrightarrow$ If $f \in C(\mathbb{R}^d)$ is not a polynomial of degree at most D, it cannot be approximated by elements of $\mathcal{NN}_\varrho^d$ (why?!).

Step 1 (Reduction to $d = 1$): Claim: If $\mathcal{NN}_\varrho^1$ is universal, then so is $\mathcal{NN}_\varrho^d$.

Substep ❶: Universality of $\mathcal{NN}_\varrho^1 \quad \Longrightarrow \quad \exp \in \overline{\mathcal{NN}_\varrho^1}$.

Substep ❷: This implies (how?!) that $(x \mapsto e^{\langle a, x \rangle}) \in \overline{\mathcal{NN}_\varrho^d}$ for all $a \in \mathbb{R}^d$.

Proof of the universal approximation theorem — Part 1

For $\mathcal{F} \subset C(\mathbb{R}^d)$, we write

$$f \in \overline{\mathcal{F}} \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ K \subset \mathbb{R}^d \text{ compact } \ \exists \tilde{f} \in \mathcal{F}: \ \sup_{x \in K} |f(x) - \tilde{f}(x)| \leq \varepsilon.$$

Step 0 (Proving "$\Longrightarrow$"): If ϱ is a polynomial of degree (at most) D, then

$$x \mapsto \varrho(\langle w, x \rangle + b)$$

is a d-variate polynomial of degree at most D.

$\Longrightarrow$ If $f \in C(\mathbb{R}^d)$ is not a polynomial of degree at most D, it cannot be approximated by elements of $\mathcal{NN}_\varrho^d$ (why?!).

Step 1 (Reduction to $d = 1$): Claim: If $\mathcal{NN}_\varrho^1$ is universal, then so is $\mathcal{NN}_\varrho^d$.

Substep ❶: Universality of $\mathcal{NN}_\varrho^1 \quad \Longrightarrow \quad \exp \in \overline{\mathcal{NN}_\varrho^1}$.

Substep ❷: This implies (how?!) that $(x \mapsto e^{\langle a, x \rangle}) \in \overline{\mathcal{NN}_\varrho^d}$ for all $a \in \mathbb{R}^d$.

Substep ❸: Universality of $\mathcal{NN}_\varrho^d$ follows from the Stone-Weierstraß theorem.

Interlude: Computing higher derivatives via divided differences

Let $f : \mathbb{R} \to \mathbb{R}$ and $x_0, \ldots, x_n \in \mathbb{R}$ pairwise distinct. The divided differences of f w.r.t. $x_0, \ldots, x_n$ are defined inductively as

$$f[x_i] := f(x_i)$$
$$f[x_i, \ldots, x_{j+1}] := \frac{f[x_{i+1}, \ldots, x_{j+1}] - f[x_i, \ldots, x_j]}{x_{j+1} - x_i}.$$

Interlude: Computing higher derivatives via divided differences

Let $f : \mathbb{R} \to \mathbb{R}$ and $x_0, \ldots, x_n \in \mathbb{R}$ pairwise distinct. The divided differences of f w.r.t. $x_0, \ldots, x_n$ are defined inductively as

$$f[x_i] := f(x_i)$$

$$f[x_i, \ldots, x_{j+1}] := \frac{f[x_{i+1}, \ldots, x_{j+1}] - f[x_i, \ldots, x_j]}{x_{j+1} - x_i}.$$

Divided differences and interpolation polynomials. Let p be the unique polynomial of degree at most n satisfying $p(x_i) = f(x_i)$. Then $f[x_0, \ldots, x_n]$ is the leading coefficient of p.

Interlude: Computing higher derivatives via divided differences

Let $f : \mathbb{R} \to \mathbb{R}$ and $x_0, \ldots, x_n \in \mathbb{R}$ pairwise distinct. The divided differences of f w.r.t. $x_0, \ldots, x_n$ are defined inductively as

$$f[x_i] := f(x_i)$$
$$f[x_i, \ldots, x_{j+1}] := \frac{f[x_{i+1}, \ldots, x_{j+1}] - f[x_i, \ldots, x_j]}{x_{j+1} - x_i}.$$

Divided differences and interpolation polynomials. Let p be the unique polynomial of degree at most n satisfying $p(x_i) = f(x_i)$. Then $f[x_0, \ldots, x_n]$ is the leading coefficient of p.

Mean-value theorem for divided differences. Let f be n times differentiable and $x_0 < \cdots < x_n$. Then there exists $\xi \in [x_0, x_n]$ such that

$$f[x_0, \ldots, x_n] = \frac{1}{n!} \cdot f^{(n)}(\xi).$$

Reference: Ryaben'kii and Tsynkov: A theoretical introduction to numerical analysis, Section 2.1.2.

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial.

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):
Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta)$$

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta) \quad \Longrightarrow \quad f_x^{(k)}(w) = x^k \cdot \varrho^{(k)}(wx + \theta).$$

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):
Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta) \quad \Longrightarrow \quad f_x^{(k)}(w) = x^k \cdot \varrho^{(k)}(wx + \theta).$$

By the mean-value theorem for divided differences,

$$f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}] = f_x^{(k)}(\xi_{x,n})/k! \qquad \text{with} \qquad 0 \leq \xi_{x,n} \leq \frac{k}{n}.$$

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}^1_\varrho$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta) \quad \Longrightarrow \quad f_x^{(k)}(w) = x^k \cdot \varrho^{(k)}(wx + \theta).$$

By the mean-value theorem for divided differences,

$$f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}] = f_x^{(k)}(\xi_{x,n})/k! \qquad \text{with} \qquad 0 \leq \xi_{x,n} \leq \frac{k}{n}.$$

Substep ❸: For $n \in \mathbb{N}$, define

$$g_n(x) := k! \cdot f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}].$$

Directly from the definitions, we see $g_n \in \mathcal{NN}^1_\varrho$.

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta) \quad \Longrightarrow \quad f_x^{(k)}(w) = x^k \cdot \varrho^{(k)}(wx + \theta).$$

By the mean-value theorem for divided differences,

$$f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}] = f_x^{(k)}(\xi_{x,n})/k! \qquad \text{with} \qquad 0 \le \xi_{x,n} \le \frac{k}{n}.$$

Substep ❸: For $n \in \mathbb{N}$, define

$$g_n(x) := k! \cdot f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}].$$

Directly from the definitions, we see $g_n \in \mathcal{NN}_\varrho^1$. Finally,

$$\left| g_n(x) - \varrho^{(k)}(\theta) \cdot x^k \right| = \left| f_x^{(k)}(\xi_{x,n}) - f_x^{(k)}(0) \right| = x^k \cdot \left| \varrho^{(k)}(\xi_{x,n} x + \theta) - \varrho^{(k)}(\theta) \right| \xrightarrow[n \to \infty]{} 0,$$

with locally uniform convergence (w.r.t. x).

Proof of the universal approximation theorem — Part 2

Step 2 (Universality of $\mathcal{NN}_\varrho^1$ for $\varrho \in C^\infty$):

Substep ❶: Let $\varrho \in C^\infty$ not a polynomial. Let $k \in \mathbb{N}$ be fixed and $\theta \in \mathbb{R}$ with $\varrho^{(k)}(\theta) \neq 0$.

Substep ❷: For $w, x \in \mathbb{R}$, let

$$f_x(w) := \varrho(wx + \theta) \quad \Longrightarrow \quad f_x^{(k)}(w) = x^k \cdot \varrho^{(k)}(wx + \theta).$$

By the mean-value theorem for divided differences,

$$f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}] = f_x^{(k)}(\xi_{x,n})/k! \qquad \text{with} \qquad 0 \leq \xi_{x,n} \leq \frac{k}{n}.$$

Substep ❸: For $n \in \mathbb{N}$, define

$$g_n(x) := k! \cdot f_x[0, \tfrac{1}{n}, \dots, \tfrac{k}{n}].$$

Directly from the definitions, we see $g_n \in \mathcal{NN}_\varrho^1$. Finally,

$$\left| g_n(x) - \varrho^{(k)}(\theta) \cdot x^k \right| = \left| f_x^{(k)}(\xi_{x,n}) - f_x^{(k)}(0) \right| = x^k \cdot \left| \varrho^{(k)}(\xi_{x,n} x + \theta) - \varrho^{(k)}(\theta) \right| \xrightarrow[n \to \infty]{} 0,$$

with locally uniform convergence (w.r.t. x).

Substep ❹: We have shown $x^k \in \overline{\mathcal{NN}_\varrho^1}$ for all $k \in \mathbb{N}$, and this also holds for $k = 0$ (why?!). Now, the claim follows from the (Stone)-Weierstraß theorem. □

Step 3 (Showing $\varphi * \varrho \in \overline{\mathcal{N}\mathcal{N}_{\varrho}^{1}}$ for $\varphi \in C_c^{\infty}(\mathbb{R})$):

Proof of the universal approximation theorem — Part 3

Step 3 (Showing $\varphi * \varrho \in \overline{\mathcal{NN}_\varrho^1}$ for $\varphi \in C_c^\infty(\mathbb{R})$):

If $\varphi * \varrho \notin \overline{\mathcal{NN}_\varrho^1}$, then there exists $K \subset \mathbb{R}$ compact such that $\varphi * \varrho \notin \overline{\mathcal{NN}_\varrho^1}^{C(K)}$.

Proof of the universal approximation theorem — Part 3

Step 3 (Showing $\varphi * \varrho \in \overline{\mathcal{NN}_\varrho^1}$ for $\varphi \in C_c^\infty(\mathbb{R})$):

If $\varphi * \varrho \notin \overline{\mathcal{NN}_\varrho^1}$, then there exists $K \subset \mathbb{R}$ compact such that $\varphi * \varrho \notin \overline{\mathcal{NN}_\varrho^1}^{C(K)}$.

Thus, there exists a signed Borel measure μ on K satisfying

$$\int_K (\varphi * \varrho)(x)\, d\mu(x) \neq 0 \qquad \text{and} \qquad \int_K \varrho(ax+b)\, d\mu(x) = 0 \quad \forall\, a, b \in \mathbb{R}.$$

Proof of the universal approximation theorem — Part 3

Step 3 (Showing $\varphi * \varrho \in \overline{\mathcal{N}\mathcal{N}_\varrho^1}$ for $\varphi \in C_c^\infty(\mathbb{R})$):

If $\varphi * \varrho \notin \overline{\mathcal{N}\mathcal{N}_\varrho^1}$, then there exists $K \subset \mathbb{R}$ compact such that $\varphi * \varrho \notin \overline{\mathcal{N}\mathcal{N}_\varrho^1}^{C(K)}$.

Thus, there exists a signed Borel measure μ on K satisfying

$$\int_K (\varphi * \varrho)(x)\, d\mu(x) \neq 0 \qquad \text{and} \qquad \int_K \varrho(ax+b)\, d\mu(x) = 0 \quad \forall\, a, b \in \mathbb{R}.$$

But then, Fubini's theorem shows

$$\begin{aligned} 0 \neq \int_K (\varphi * \varrho)(x)\, d\mu(x) &= \int_K \int_{\mathbb{R}} \varphi(y)\varrho(x-y)\, dy\, d\mu(x) \\ &= \int_{\mathbb{R}} \varphi(y) \int_K \varrho(x-y)\, d\mu(x)\, dy \\ &= \int_{\mathbb{R}} \varphi(y) \cdot 0\, dy = 0. \end{aligned}$$

Contradiction.

Step 3 (Showing $\varphi * \varrho \in \overline{\mathcal{NN}^1_\varrho}$ for $\varphi \in C_c^\infty(\mathbb{R})$):

If $\varphi * \varrho \notin \overline{\mathcal{NN}^1_\varrho}$, then there exists $K \subset \mathbb{R}$ compact such that $\varphi * \varrho \notin \overline{\mathcal{NN}^1_\varrho}^{C(K)}$.

Thus, there exists a signed Borel measure μ on K satisfying

$$\int_K (\varphi * \varrho)(x)\, d\mu(x) \neq 0 \qquad \text{and} \qquad \int_K \varrho(ax+b)\, d\mu(x) = 0 \quad \forall\, a, b \in \mathbb{R}.$$

But then, Fubini's theorem shows

$$\begin{aligned} 0 \neq \int_K (\varphi * \varrho)(x)\, d\mu(x) &= \int_K \int_{\mathbb{R}} \varphi(y) \varrho(x-y)\, dy\, d\mu(x) \\ &= \int_{\mathbb{R}} \varphi(y) \int_K \varrho(x-y)\, d\mu(x)\, dy \\ &= \int_{\mathbb{R}} \varphi(y) \cdot 0\, dy = 0. \end{aligned}$$

Contradiction.

Step 4: By the above, we are done if $\varphi * \varrho$ is not a polynomial for some $\varphi \in C_c^\infty(\mathbb{R})$.

Step 5 (Handling the case that $\varphi * \varrho$ is a polynomial for all $\varphi \in C_c^\infty$):

Step 5 (Handling the case that $\varphi * \varrho$ is a polynomial for all $\varphi \in C_c^\infty$):

Substep ❶: $C_c^\infty[-1,1] := \{\varphi \in C_c^\infty(\mathbb{R}) \;:\; \operatorname{supp}\varphi \subset [-1,1]\}$ is a complete metric space with metric

$$d(\varphi,\psi) := \sum_{n=1}^{\infty} 2^{-n} \min\{1, \|\varphi - \psi\|_{C^n}\}.$$

Step 5 (Handling the case that $\varphi * \varrho$ is a polynomial for all $\varphi \in C_c^\infty$):

Substep ❶: $C_c^\infty[-1,1] := \{\varphi \in C_c^\infty(\mathbb{R}) \; : \; \operatorname{supp}\varphi \subset [-1,1]\}$ is a complete metric space with metric

$$d(\varphi, \psi) := \sum_{n=1}^{\infty} 2^{-n} \min\{1, \|\varphi - \psi\|_{C^n}\}.$$

Substep ❷: By assumption,

$$C_c^\infty[-1,1] = \bigcup_{m=1}^{\infty} V_m \qquad \text{for} \qquad V_m := \{\varphi \in C_c^\infty[-1,1] \; : \; \deg(\varphi * \varrho) \leq m\},$$

and each V_m is a closed subspace.

Proof of the universal approximation theorem — Part 4

Step 5 (Handling the case that $\varphi * \varrho$ is a polynomial for all $\varphi \in C_c^\infty$):

Substep ❶: $C_c^\infty[-1,1] := \{\varphi \in C_c^\infty(\mathbb{R}) \;:\; \operatorname{supp}\varphi \subset [-1,1]\}$ is a complete metric space with metric

$$d(\varphi, \psi) := \sum_{n=1}^{\infty} 2^{-n} \min\{1, \|\varphi - \psi\|_{C^n}\}.$$

Substep ❷: By assumption,

$$C_c^\infty[-1,1] = \bigcup_{m=1}^{\infty} V_m \qquad \text{for} \qquad V_m := \{\varphi \in C_c^\infty[-1,1] \;:\; \deg(\varphi * \varrho) \leq m\},$$

and each V_m is a closed subspace.

Substep ❸: By Baire category, some V_m has non-empty interiors, which implies $V_m = C_c^\infty[-1,1]$.

Step 5 (Handling the case that $\varphi * \varrho$ is a polynomial for all $\varphi \in C_c^\infty$):

Substep ❶: $C_c^\infty[-1,1] := \{\varphi \in C_c^\infty(\mathbb{R}) : \operatorname{supp}\varphi \subset [-1,1]\}$ is a complete metric space with metric

$$d(\varphi, \psi) := \sum_{n=1}^{\infty} 2^{-n} \min\{1, \|\varphi - \psi\|_{C^n}\}.$$

Substep ❷: By assumption,

$$C_c^\infty[-1,1] = \bigcup_{m=1}^{\infty} V_m \qquad \text{for} \qquad V_m := \{\varphi \in C_c^\infty[-1,1] : \deg(\varphi * \varrho) \le m\},$$

and each V_m is a closed subspace.

Substep ❸: By Baire category, some V_m has non-empty interiors, which implies $V_m = C_c^\infty[-1,1]$.

Substep ❹: Choose $\varphi_n \in C_c^\infty[-1,1]$ with $\varphi_m \to \delta_0$. Then $\varphi_m * \varrho \to \varrho$, so that ϱ is a polynomial (of degree at most m). Contradiction. □

Quantitative approximation rates for Barron functions

Barron-regular functions can be well approximated by NNs

$f : \mathbb{R}^d \to \mathbb{R}$ is called Barron-regular with constant $C > 0$ (written $f \in B_d(C)$), if

$$f(x) = c + \int_{\mathbb{R}^d} (e^{i\langle x,\xi\rangle} - 1) \cdot F(\xi)\, d\xi \qquad \forall\, x \in \mathbb{R}^d,$$

with $\int_{\mathbb{R}^d} |\xi| \cdot |F(\xi)|\, d\xi \leq C$.

Andrew Barron;
`opc.mfo.de/detail?photo_id=14885`

Barron-regular functions can be well approximated by NNs

$f : \mathbb{R}^d \to \mathbb{R}$ is called Barron-regular with constant $C > 0$ (written $f \in B_d(C)$), if

$$f(x) = c + \int_{\mathbb{R}^d} (e^{i\langle x, \xi \rangle} - 1) \cdot F(\xi)\, d\xi \qquad \forall x \in \mathbb{R}^d,$$

with $\int_{\mathbb{R}^d} |\xi| \cdot |F(\xi)|\, d\xi \leq C$.

Theorem (Barron; 1993).
Let ϱ be a sigmoidal activation function. Let μ be a probability measure on $\mathbb{R}^d$, let $r > 0$ and $f \in B_d(C)$. For any $N \in \mathbb{N}$, one can achieve

$$\int_{B_r} \left| f(x) - \Phi_N(x) \right|^2 d\mu(x) \leq \left(\frac{2rC}{\sqrt{N}} \right)^2,$$

where Φ_N is a shallow NN with N neurons and activation function ϱ.

Andrew Barron;
opc.mfo.de/detail?photo_id=14885

Barron-regular functions can be well approximated by NNs

$f : \mathbb{R}^d \to \mathbb{R}$ is called Barron-regular with constant $C > 0$ (written $f \in B_d(C)$), if

$$f(x) = c + \int_{\mathbb{R}^d} (e^{i\langle x,\xi\rangle} - 1) \cdot F(\xi)\, d\xi \qquad \forall x \in \mathbb{R}^d,$$

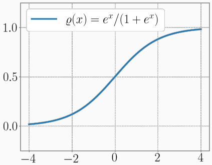
with $\int_{\mathbb{R}^d} |\xi| \cdot |F(\xi)|\, d\xi \leq C$.

Theorem (Barron; 1993).
Let ϱ be a sigmoidal activation function. Let μ be a probability measure on $\mathbb{R}^d$, let $r > 0$ and $f \in B_d(C)$. For any $N \in \mathbb{N}$, one can achieve

$$\int_{B_r} \left|f(x) - \Phi_N(x)\right|^2 d\mu(x) \leq \left(\frac{2rC}{\sqrt{N}}\right)^2,$$

where Φ_N is a shallow NN with N neurons and activation function ϱ.

$\varrho : \mathbb{R} \to \mathbb{R}$ is sigmoidal if it is bounded, measurable, and if $\lim_{x\to\infty} \varrho(x) = 1$ and $\lim_{x\to-\infty} \varrho(x) = 0$.



Andrew Barron;
opc.mfo.de/detail?photo_id=14885

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\mathrm{conv}\, G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \ldots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\operatorname{conv} G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \dots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Proof (Probabilistic method): ❶: Let $\delta > 0$ arbitrary and choose $f^* = \sum_{i=1}^{M} \lambda_i h_i$ with $h_i \in G$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$ satisfying $\|f - f^*\|_{\mathcal{H}} \leq \delta$.

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\mathrm{conv}\, G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \ldots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Proof (Probabilistic method): ➊: Let $\delta > 0$ arbitrary and choose $f^* = \sum_{i=1}^{M} \lambda_i h_i$ with $h_i \in G$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$ satisfying $\|f - f^*\|_{\mathcal{H}} \leq \delta$.

➋: Let $Z \in G$ a random vector with $\mathbb{P}(Z = h_i) = \lambda_i$ for $i \in \{1, \ldots, N\}$, and note $\mathbb{E}Z = f^*$.

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\operatorname{conv} G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \ldots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Proof (Probabilistic method): ❶: Let $\delta > 0$ arbitrary and choose $f^* = \sum_{i=1}^{M} \lambda_i h_i$ with $h_i \in G$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$ satisfying $\|f - f^*\|_{\mathcal{H}} \leq \delta$.

❷: Let $Z \in G$ a random vector with $\mathbb{P}(Z = h_i) = \lambda_i$ for $i \in \{1, \ldots, N\}$, and note $\mathbb{E}Z = f^*$.

❸: Let $Z_1, \ldots, Z_N \overset{iid}{\sim} Z$ and note $\mathbb{E}\langle Z_n - f^*, Z_m - f^* \rangle = 0$ for $n \neq m$ and

$$\mathbb{E}\|Z_n - f^*\|_{\mathcal{H}}^2 = \mathbb{E}\|Z_n\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2 \leq b^2 - \|f^*\|_{\mathcal{H}}^2.$$

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\operatorname{conv} G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \dots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Proof (Probabilistic method): 1: Let $\delta > 0$ arbitrary and choose $f^* = \sum_{i=1}^{M} \lambda_i h_i$ with $h_i \in G$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$ satisfying $\|f - f^*\|_{\mathcal{H}} \leq \delta$.

2: Let $Z \in G$ a random vector with $\mathbb{P}(Z = h_i) = \lambda_i$ for $i \in \{1, \dots, N\}$, and note $\mathbb{E}Z = f^*$.

3: Let $Z_1, \dots, Z_N \overset{iid}{\sim} Z$ and note $\mathbb{E}\langle Z_n - f^*, Z_m - f^* \rangle = 0$ for $n \neq m$ and

$$\mathbb{E}\|Z_n - f^*\|_{\mathcal{H}}^2 = \mathbb{E}\|Z_n\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2 \leq b^2 - \|f^*\|_{\mathcal{H}}^2.$$

4:
$$\mathbb{E}\left\| f^* - \frac{1}{N}\sum_{n=1}^{N} Z_n \right\|_{\mathcal{H}}^2 = \frac{1}{N^2}\, \mathbb{E}\left\| \sum_{n=1}^{N} (Z_i - f^*) \right\|_{\mathcal{H}}^2 = \frac{1}{N^2}\, \mathbb{E} \sum_{n,m=1}^{N} \langle Z_n - f^*, Z_m - f^* \rangle$$

$$= \frac{1}{N^2}\, \mathbb{E} \sum_{n=1}^{N} \|Z_n - f^*\|_{\mathcal{H}}^2 \leq \frac{b^2 - \|f^*\|_{\mathcal{H}}^2}{N}.$$

Main ingredient: Approximability of elements of convex hulls

Lemma (Maurey). Let $\mathcal{H}$ be a Hilbert space, $G \subset \mathcal{H}$ and $b > 0$ with $\|g\|_{\mathcal{H}} \leq b$ for all $g \in G$. Let $f_0 \in \overline{\operatorname{conv} G}$ and $c > b^2 - \|f_0\|_{\mathcal{H}}^2$.
Then for any $N \in \mathbb{N}$ there exist $g_1, \ldots, g_N \in G$ such that

$$f_N = \frac{1}{N} \sum_{n=1}^{N} g_n \quad \text{satisfies} \quad \|f_0 - f_N\|_{\mathcal{H}}^2 \leq \frac{c}{N}.$$

Proof (Probabilistic method): 1: Let $\delta > 0$ arbitrary and choose $f^* = \sum_{i=1}^{M} \lambda_i h_i$ with $h_i \in G$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$ satisfying $\|f - f^*\|_{\mathcal{H}} \leq \delta$.

2: Let $Z \in G$ a random vector with $\mathbb{P}(Z = h_i) = \lambda_i$ for $i \in \{1, \ldots, N\}$, and note $\mathbb{E}Z = f^*$.

3: Let $Z_1, \ldots, Z_N \overset{iid}{\sim} Z$ and note $\mathbb{E}\langle Z_n - f^*, Z_m - f^* \rangle = 0$ for $n \neq m$ and

$$\mathbb{E}\|Z_n - f^*\|_{\mathcal{H}}^2 = \mathbb{E}\|Z_n\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2 \leq b^2 - \|f^*\|_{\mathcal{H}}^2.$$

4:
$$\mathbb{E}\left\| f^* - \frac{1}{N}\sum_{n=1}^{N} Z_n \right\|_{\mathcal{H}}^2 = \frac{1}{N^2}\, \mathbb{E}\left\| \sum_{n=1}^{N} (Z_i - f^*) \right\|_{\mathcal{H}}^2 = \frac{1}{N^2}\, \mathbb{E} \sum_{n,m=1}^{N} \langle Z_n - f^*, Z_m - f^* \rangle$$

$$= \frac{1}{N^2}\, \mathbb{E} \sum_{n=1}^{N} \|Z_n - f^*\|_{\mathcal{H}}^2 \leq \frac{b^2 - \|f^*\|_{\mathcal{H}}^2}{N}.$$

5: For δ small enough, this implies $\mathbb{E}\|f_0 - \frac{1}{N}\sum_{n=1}^{N} Z_n\|_{\mathcal{H}}^2 \leq \frac{c}{N}$, since $\|f_0 - f^*\|_{\mathcal{H}} \leq \delta$. □

Integral formulas imply membership in the closed convex hull

Let (X, μ) be a finite measure space and $G \subset L^2(\mu)$, and let (Ω, ν) be a probability space. Let $g : X \times \Omega \to \mathbb{R}$ be measurable and such that

- $g(\cdot, \omega) \in G$ for all $\omega \in \Omega$;
- $|g(x, \omega)| \leq C$ for all $(x, \omega) \in X \times \Omega$ and some $C < \infty$;
- $f(x) = \int_\Omega g(x, \omega)\, d\nu(\omega)$ for all $x \in X$.

Then $f \in \overline{\mathrm{conv}\, G}$, with the closure taken in $L^2(\mu)$.

Integral formulas imply membership in the closed convex hull

Let (X,μ) be a finite measure space and $G \subset L^2(\mu)$, and let (Ω,ν) be a probability space. Let $g : X \times \Omega \to \mathbb{R}$ be measurable and such that

- $g(\cdot,\omega) \in G$ for all $\omega \in \Omega$;
- $|g(x,\omega)| \leq C$ for all $(x,\omega) \in X \times \Omega$ and some $C < \infty$;
- $f(x) = \int_\Omega g(x,\omega)\, d\nu(\omega)$ for all $x \in X$.

Then $f \in \overline{\mathrm{conv}\, G}$, with the closure taken in $L^2(\mu)$.

Proof: Let $\omega_1, \omega_2, \ldots \overset{iid}{\sim} \mu$. Then

$$\mathbb{E}\int_X \left(f(x) - \frac{1}{N}\sum_{i=1}^N g(x,\omega_i)\right)^2 d\mu(x) = \int_X \mathrm{var}\left(\frac{1}{N}\sum_{i=1}^N g(x,\omega_i)\right) d\mu(x)$$

$$= \frac{1}{N^2}\int_X \sum_{i=1}^N \mathrm{var}[g(x,\omega_i)]\, d\mu(x) \leq \frac{C^2}{N}.$$

Integral formulas imply membership in the closed convex hull

Let (X, μ) be a finite measure space and $G \subset L^2(\mu)$, and let (Ω, ν) be a probability space. Let $g : X \times \Omega \to \mathbb{R}$ be measurable and such that

- $g(\cdot, \omega) \in G$ for all $\omega \in \Omega$;
- $|g(x, \omega)| \leq C$ for all $(x, \omega) \in X \times \Omega$ and some $C < \infty$;
- $f(x) = \int_\Omega g(x, \omega)\, d\nu(\omega)$ for all $x \in X$.

Then $f \in \overline{\mathrm{conv}\, G}$, with the closure taken in $L^2(\mu)$.

Proof: Let $\omega_1, \omega_2, \ldots \overset{iid}{\sim} \mu$. Then

$$\mathbb{E} \int_X \left(f(x) - \frac{1}{N} \sum_{i=1}^N g(x, \omega_i) \right)^2 d\mu(x) = \int_X \mathrm{var}\left(\frac{1}{N} \sum_{i=1}^N g(x, \omega_i) \right) d\mu(x)$$

$$= \frac{1}{N^2} \int_X \sum_{i=1}^N \mathrm{var}[g(x, \omega_i)] d\mu(x) \leq \frac{C^2}{N}.$$

By Fatou's lemma, this implies

$$\mathbb{E}\left[\liminf_{N \to \infty} \left\| f - \frac{1}{N} \sum_{i=1}^N g(\cdot, \omega_i) \right\|_{L^2(\mu)}^2 \right] \xrightarrow[N \to \infty]{} 0. \qquad \square$$

Proof of Barron's result

For $f : \mathbb{R}^d \to \mathbb{R}$, write $f \in B_d^*(C)$ if

$$f(x) = \int_{\mathbb{R}^d} (e^{i\langle x,\omega\rangle} - 1) \cdot F(\omega)\, d\omega \qquad \forall x \in \mathbb{R}^d, \qquad (*)$$

for some F with $C_F := \int_{\mathbb{R}^d} |\omega| \cdot |F(\omega)|\, d\omega \leq C$. Thus, $B_d(C) = \mathbb{R} + B_d^*(C)$.

Proof of Barron's result

For $f : \mathbb{R}^d \to \mathbb{R}$, write $f \in B_d^*(C)$ if

$$f(x) = \int_{\mathbb{R}^d} (e^{i\langle x,\omega\rangle} - 1) \cdot F(\omega)\, d\omega \qquad \forall x \in \mathbb{R}^d, \qquad (*)$$

for some F with $C_F := \int_{\mathbb{R}^d} |\omega| \cdot |F(\omega)|\, d\omega \leq C$. Thus, $B_d(C) = \mathbb{R} + B_d^*(C)$.

Lemma: Let $c \in \mathbb{R}$ be arbitrary, and let $H(x) := \mathbb{1}_{(0,\infty)} + c \cdot \mathbb{1}_{\{0\}}$. We have $B_d^*(C) \subset \overline{\operatorname{conv} G_H}$, where the closure is taken in $L^2(B_r; \mu)$ for any finite measure μ, and where

$$G_H := \left\{ \gamma \cdot H(\langle w, \bullet \rangle + b) \ : \ |\gamma| \leq 2rC,\, w \in \mathbb{R}^d,\, b \in \mathbb{R} \right\}.$$

Proof of Barron's result

For $f : \mathbb{R}^d \to \mathbb{R}$, write $f \in B_d^*(C)$ if

$$f(x) = \int_{\mathbb{R}^d} (e^{i\langle x,\omega\rangle} - 1) \cdot F(\omega)\, d\omega \qquad \forall x \in \mathbb{R}^d, \qquad (*)$$

for some F with $C_F := \int_{\mathbb{R}^d} |\omega| \cdot |F(\omega)|\, d\omega \leq C$. Thus, $B_d(C) = \mathbb{R} + B_d^*(C)$.

Lemma: Let $c \in \mathbb{R}$ be arbitrary, and let $H(x) := \mathbb{1}_{(0,\infty)} + c \cdot \mathbb{1}_{\{0\}}$. We have $B_d^*(C) \subset \overline{\operatorname{conv} G_H}$, where the closure is taken in $L^2(B_r; \mu)$ for any finite measure μ, and where

$$G_H := \Big\{ \gamma \cdot H(\langle w, \bullet \rangle + b) \; : \; |\gamma| \leq 2rC,\, w \in \mathbb{R}^d,\, b \in \mathbb{R} \Big\}.$$

Proof: ❶: A direct computation shows for $c > 0$ and $|t| \leq c$ that

$$e^{it} - 1 = i \int_0^c \mathbb{1}_{u<t} \cdot e^{iu} - \mathbb{1}_{u<-t} \cdot e^{-iu}\, du = i \int_0^c H(t-u)\, e^{iu} - H(-u-t)\, e^{-iu}\, du.$$

Proof of Barron's result

For $f : \mathbb{R}^d \to \mathbb{R}$, write $f \in B_d^*(C)$ if

$$f(x) = \int_{\mathbb{R}^d} (e^{i\langle x,\omega\rangle} - 1) \cdot F(\omega)\, d\omega \qquad \forall x \in \mathbb{R}^d, \qquad (*)$$

for some F with $C_F := \int_{\mathbb{R}^d} |\omega| \cdot |F(\omega)|\, d\omega \leq C$. Thus, $B_d(C) = \mathbb{R} + B_d^*(C)$.

Lemma: Let $c \in \mathbb{R}$ be arbitrary, and let $H(x) := \mathbb{1}_{(0,\infty)} + c \cdot \mathbb{1}_{\{0\}}$. We have $B_d^*(C) \subset \overline{\mathrm{conv}\, G_H}$, where the closure is taken in $L^2(B_r;\mu)$ for any finite measure μ, and where

$$G_H := \Big\{ \gamma \cdot H(\langle w, \bullet\rangle + b) \ : \ |\gamma| \leq 2rC,\ w \in \mathbb{R}^d,\ b \in \mathbb{R} \Big\}.$$

Proof: ❶: A direct computation shows for $c > 0$ and $|t| \leq c$ that

$$e^{it} - 1 = i \int_0^c \mathbb{1}_{u<t} \cdot e^{iu} - \mathbb{1}_{u<-t} \cdot e^{-iu}\, du = i \int_0^c H(t-u)\, e^{iu} - H(-u-t)\, e^{-iu}\, du.$$

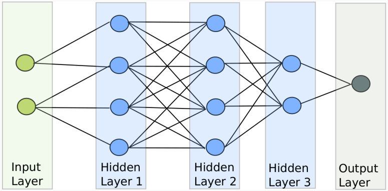
❷: Using $(*)$ and the formula from ❶ with $t = \langle \omega, x\rangle$ and $c = r \cdot |\omega|$, and writing $F(\omega) = e^{i\theta(\omega)}|F(\omega)|$, we finally see

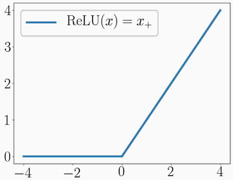
$$f(x) = \mathrm{Re}\Big(i \int_{\mathbb{R}^d} \int_0^{r\cdot|\omega|} F(\omega) \cdot \Big(H(\langle \omega, x\rangle - u)\, e^{iu} - H(\langle -\omega, x\rangle - u)\, e^{-iu} \Big)\, du\, d\omega \Big)$$

$$= \sum_{j=0}^{1} \int_{\mathbb{R}^d} \int_0^1 \frac{|\omega| \cdot |F(\omega)|}{2C_F} \cdot (-1)^{j+1} 2rC_F \cdot \sin\big(\theta(\omega) + (-1)^j r|\omega|t\big) \cdot H\big(\langle (-1)^j \omega, x\rangle - r|\omega|t\big)\, dt d\omega. \square$$

Universal approximation for complex-valued neural networks

The definition of complex-valued neural networks (CVNNs)





- L: number of (hidden) layers,
- N_ℓ: number of neurons in layer ℓ,
- $T_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_{\ell+1}}, x \mapsto A_\ell x + b_\ell$: connections between neurons (weights).

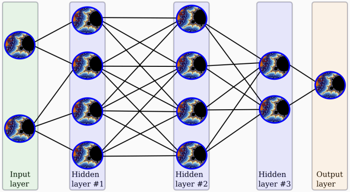
$\varrho : \mathbb{R} \to \mathbb{R}$: activation function

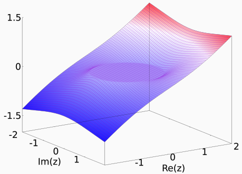
Network function $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ given by

$$\Phi = T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_0)$$

with ϱ applied componentwise.

The definition of complex-valued neural networks (CVNNs)





- L: number of (hidden) layers,
- N_ℓ: number of neurons in layer ℓ,
- $T_\ell : \mathbb{C}^{N_\ell} \to \mathbb{C}^{N_{\ell+1}}, x \mapsto A_\ell x + b_\ell$: connections between neurons (weights).

$\sigma : \mathbb{C} \to \mathbb{C}$: activation function

Network function $\Phi : \mathbb{C}^{N_0} \to \mathbb{C}^{N_{L+1}}$ given by

$$\Phi = T_L \circ (\sigma \circ T_{L-1}) \circ \cdots \circ (\sigma \circ T_0)$$

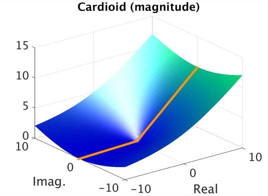
with σ applied componentwise.

CVNNs have advantages for tasks with naturally ℂ-valued inputs

Virtue, Yu, Lustig: *Better than real: Complex-valued Neural Nets for MRI fingerprinting*, ICIP, 2017:

Goal: From ℂ-valued MRI measurements, determine if tissue is benign or malignant.

> *CVNNs outperform 2-channel real-valued networks for almost all of our experiments, and this advantage cannot be explained away by the twice large model capacity.*



Differentiability is always understood in the sense of real variables

[unless mentioned otherwise]

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not ???.

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

Here, $g : \mathbb{C} \to \mathbb{C}$ is polyharmonic if $g \in C^\infty$ and $\Delta^m g \equiv 0$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ denotes the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$.

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

Here, $g : \mathbb{C} \to \mathbb{C}$ is polyharmonic if $g \in C^\infty$ and $\Delta^m g \equiv 0$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ denotes the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$.

Remark: g polyharm. $\Longleftrightarrow$ $\mathrm{Re}\, g$ and $\mathrm{Im}\, g$ of the form $\mathrm{Re}\big(\sum_{k=0}^m \overline{z}^k \cdot f_k(z)\big)$ with all f_k entire.

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

Here, $g : \mathbb{C} \to \mathbb{C}$ is polyharmonic if $g \in C^\infty$ and $\Delta^m g \equiv 0$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ denotes the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$.

Remark: g polyharm. $\iff \mathrm{Re}\, g$ and $\mathrm{Im}\, g$ of the form $\mathrm{Re}\big(\sum_{k=0}^m \overline{z}^k \cdot f_k(z)\big)$ with all f_k entire.

Theorem (deep case; FV; 2020)

Let $L \in \mathbb{N}_{\geq 2}$. The set $\mathcal{NN}_{\sigma,L}^d$ of deep CVNNs with L hidden layers is universal if and only if none(!) of the following hold:

- *σ is holomorphic or $\overline{\sigma}$ is holomorphic,*
- *$\sigma(z) = p(z, \overline{z})$ with a polynomial p.*

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

Here, $g : \mathbb{C} \to \mathbb{C}$ is polyharmonic if $g \in C^\infty$ and $\Delta^m g \equiv 0$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ denotes the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$.

Remark: g polyharm. $\iff \mathrm{Re}\, g$ and $\mathrm{Im}\, g$ of the form $\mathrm{Re}\big(\sum_{k=0}^m \overline{z}^k \cdot f_k(z)\big)$ with all f_k entire.

Theorem (deep case; FV; 2020)

Let $L \in \mathbb{N}_{\geq 2}$. The set $\mathcal{NN}_{\sigma,L}^d$ of deep CVNNs with L hidden layers is universal if and only if none(!) of the following hold:

- *σ is holomorphic or $\overline{\sigma}$ is holomorphic,*
- *$\sigma(z) = p(z, \overline{z})$ with a polynomial p.*

Example: $\sigma(z) = \overline{z} \cdot e^z$ is polyharmonic, but $\mathcal{NN}_{\sigma,L}^d$ is universal if $L \geq 2$.

The universal approximation theorem for CVNNs

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be continuous.

Theorem (shallow case; FV; 2020)

The set $\mathcal{NN}_\sigma^d$ of shallow CVNNs is universal if and only if σ is not polyharmonic.

Here, $g : \mathbb{C} \to \mathbb{C}$ is polyharmonic if $g \in C^\infty$ and $\Delta^m g \equiv 0$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ denotes the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$.

Remark: g polyharm. $\Longleftrightarrow$ $\mathrm{Re}\, g$ and $\mathrm{Im}\, g$ of the form $\mathrm{Re}\big(\sum_{k=0}^m \bar{z}^k \cdot f_k(z) \big)$ with all f_k entire.

Theorem (deep case; FV; 2020)

Let $L \in \mathbb{N}_{\geq 2}$. The set $\mathcal{NN}_{\sigma,L}^d$ of deep CVNNs with L hidden layers is universal if and only if none(!) of the following hold:

- *σ is holomorphic or $\overline{\sigma}$ is holomorphic,*
- *$\sigma(z) = p(z, \bar{z})$ with a polynomial p.*

Example: $\sigma(z) = \bar{z} \cdot e^z$ is polyharmonic, but $\mathcal{NN}_{\sigma,L}^d$ is universal if $L \geq 2$.

Remark: Some (very) partial results were already known [Arena, Fortuna, Re, Xibilia; 1995].

Proof ingredients

Ingredient 1: Wirtinger calculus

Identifying $f : U \subset \mathbb{C} \to \mathbb{C}$ with $(x, y) \mapsto f(x + iy)$, define

$$\partial f := \frac{1}{2}\left(\partial_1 f - i\,\partial_2 f\right) \qquad \text{and} \qquad \overline{\partial} f := \frac{1}{2}\left(\partial_1 f + i\,\partial_2 f\right).$$

Ingredient 1: Wirtinger calculus

Identifying $f : U \subset \mathbb{C} \to \mathbb{C}$ with $(x, y) \mapsto f(x + iy)$, define

$$\partial f := \tfrac{1}{2}\big(\partial_1 f - i\,\partial_2 f\big) \qquad \text{and} \qquad \overline{\partial} f := \tfrac{1}{2}\big(\partial_1 f + i\,\partial_2 f\big).$$

Properties:

- $f \in C^1(U; \mathbb{C})$ is holomorphic $\iff \overline{\partial} f \equiv 0$.

  In this case, ∂f is the usual complex derivative of f.
- $\Delta f = 4 \cdot \partial \overline{\partial} f$ for $f \in C^2(U; \mathbb{C})$.

Ingredient 1: Wirtinger calculus

Identifying $f : U \subset \mathbb{C} \to \mathbb{C}$ with $(x, y) \mapsto f(x + iy)$, define

$$\partial f := \tfrac{1}{2}\big(\partial_1 f - i\,\partial_2 f\big) \qquad \text{and} \qquad \overline{\partial} f := \tfrac{1}{2}\big(\partial_1 f + i\,\partial_2 f\big).$$

Properties:

- $f \in C^1(U; \mathbb{C})$ is holomorphic $\iff \overline{\partial} f \equiv 0$.

  In this case, ∂f is the usual complex derivative of f.

- $\Delta f = 4 \cdot \partial \overline{\partial} f$ for $f \in C^2(U; \mathbb{C})$.

- Product rule:
  $$\partial(f \cdot g) = (\partial f) \cdot g + f \cdot \partial g \quad \text{and} \quad \overline{\partial}(f \cdot g) = (\overline{\partial} f) \cdot g + f \cdot (\overline{\partial} g).$$

- Chain rule:
  $$\partial(f \circ g) = [(\partial f) \circ g] \cdot \partial g + [(\overline{\partial} f) \circ g] \cdot \overline{\partial} g$$
  $$\text{and} \quad \overline{\partial}(f \circ g) = [(\partial f) \circ g] \cdot \overline{\partial} g + [(\overline{\partial} f) \circ g] \cdot \overline{\partial}\, \overline{g}.$$

Weyl's lemma

Let $U \subset \mathbb{R}^d$ be open and suppose that $\gamma \in \mathcal{D}'(U)$ [i.e., γ is a distribution] satisfies $\Delta\gamma = g$ for some $g \in C^\infty(U)$. Then $\gamma \in C^\infty(U)$.

Ingredient 2: Weyl's lemma

Weyl's lemma

Let $U \subset \mathbb{R}^d$ be open and suppose that $\gamma \in \mathcal{D}'(U)$ [i.e., γ is a distribution] satisfies $\Delta\gamma = g$ for some $g \in C^\infty(U)$. Then $\gamma \in C^\infty(U)$.

Corollary

Suppose that $f \in L^1_{\mathrm{loc}}(U)$ satisfies $\int_U f \cdot \Delta^m \theta \, dx = 0$ for all $\theta \in C_c^\infty(U)$. Then $f \in C^\infty(U)$ and $\Delta^m f \equiv 0$.

Ingredient 2: Weyl's lemma

Weyl's lemma

Let $U \subset \mathbb{R}^d$ be open and suppose that $\gamma \in \mathcal{D}'(U)$ [i.e., γ is a distribution] satisfies $\Delta\gamma = g$ for some $g \in C^\infty(U)$. Then $\gamma \in C^\infty(U)$.

Corollary

Suppose that $f \in L^1_{\mathrm{loc}}(U)$ satisfies $\int_U f \cdot \Delta^m \theta \, dx = 0$ for all $\theta \in C_c^\infty(U)$. Then $f \in C^\infty(U)$ and $\Delta^m f \equiv 0$.

Corollary

If $(f_n)_{n\in\mathbb{N}} \subset C^\infty(\mathbb{C};\mathbb{C})$ with $\Delta^m f_n \equiv 0$ for all $n \in \mathbb{N}$ and $f_n \to f$ with locally uniform convergence, then $f \in C^\infty(\mathbb{C};\mathbb{C})$ and $\Delta^m f \equiv 0$.

Necessity

(Universality $\implies \sigma$ is not polyharmonic / ...)

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Recall: Each shallow network $\Psi \in \mathcal{NN}_\sigma^1$ is of the form

$$\Psi(z) = c + \sum c_j \, \sigma(a_j z + b_j).$$

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Recall: Each shallow network $\Psi \in \mathcal{NN}_\sigma^1$ is of the form

$$\Psi(z) = c + \sum c_j \, \sigma(a_j z + b_j).$$

Step ❶: Using $\Delta = 4 \, \partial \overline{\partial}$ and Wirtinger calculus shows $\Delta^m \Psi \equiv 0$ for $\Psi \in \mathcal{NN}_\sigma^1$.

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Recall: Each shallow network $\Psi \in \mathcal{NN}_\sigma^1$ is of the form

$$\Psi(z) = c + \sum c_j \, \sigma(a_j z + b_j).$$

Step ❶: Using $\Delta = 4\,\partial\overline{\partial}$ and Wirtinger calculus shows $\Delta^m \Psi \equiv 0$ for $\Psi \in \mathcal{NN}_\sigma^1$.

Step ❷: By Weyl's lemma: If $(\Psi_n)_{n \in \mathbb{N}} \subset \mathcal{NN}_\sigma^1$ satisfies $\Psi_n \to F$ locally uniformly, then $F \in C^\infty$.

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Recall: Each shallow network $\Psi \in \mathcal{NN}_\sigma^1$ is of the form

$$\Psi(z) = c + \sum c_j \, \sigma(a_j \, z + b_j).$$

Step ❶: Using $\Delta = 4\,\partial\overline{\partial}$ and Wirtinger calculus shows $\Delta^m \Psi \equiv 0$ for $\Psi \in \mathcal{NN}_\sigma^1$.

Step ❷: By Weyl's lemma: If $(\Psi_n)_{n\in\mathbb{N}} \subset \mathcal{NN}_\sigma^1$ satisfies $\Psi_n \to F$ locally uniformly, then $F \in C^\infty$.

$\Longrightarrow$ Universality fails if $\Delta^m \sigma \equiv 0$.

Necessity for shallow networks

Suppose that $\Delta^m \sigma \equiv 0$ for some $m \in \mathbb{N}$.

To prove: Universality fails.

Recall: Each shallow network $\Psi \in \mathcal{NN}_\sigma^1$ is of the form

$$\Psi(z) = c + \sum c_j \, \sigma(a_j z + b_j).$$

Step ❶: Using $\Delta = 4\,\partial\overline{\partial}$ and Wirtinger calculus shows $\Delta^m \Psi \equiv 0$ for $\Psi \in \mathcal{NN}_\sigma^1$.

Step ❷: By Weyl's lemma: If $(\Psi_n)_{n \in \mathbb{N}} \subset \mathcal{NN}_\sigma^1$ satisfies $\Psi_n \to F$ locally uniformly, then $F \in C^\infty$.

$\Longrightarrow$ Universality fails if $\Delta^m \sigma \equiv 0$.

Necessity for deep networks

Necessity for deep networks

Case ❶: σ holomorphic.

Then Ψ is holomorphic for any $\Psi \in \mathcal{NN}_{\sigma,L}$.

$\rightsquigarrow$ Locally uniform limits also holomorphic

$\rightsquigarrow$ Universality fails!

Necessity for deep networks

Case ❶: σ holomorphic.

Then Ψ is holomorphic for any $\Psi \in \mathcal{NN}^1_{\sigma,L}$.

$\rightsquigarrow$ Locally uniform limits also holomorphic

$\rightsquigarrow$ Universality fails!

Case ❷: σ is anti-holomorphic (i.e., $\overline{\sigma}$ is holomorphic).

Then Ψ is holomorphic or anti-holomorphic for any $\Psi \in \mathcal{NN}^1_{\sigma,L}$.

$\rightsquigarrow$ As above: Universality fails!

Necessity for deep networks

Case ❶: σ holomorphic.

Then Ψ is holomorphic for any $\Psi \in \mathcal{NN}^1_{\sigma,L}$.

$\rightsquigarrow$ Locally uniform limits also holomorphic

$\rightsquigarrow$ Universality fails!

Case ❷: σ is anti-holomorphic (i.e., $\overline{\sigma}$ is holomorphic).

Then Ψ is holomorphic or anti-holomorphic for any $\Psi \in \mathcal{NN}^1_{\sigma,L}$.

$\rightsquigarrow$ As above: Universality fails!

Case ❸: $\sigma(z) = p(z, \overline{z})$ for a polynomial p.

Then Ψ is a polynomial of degree $N = N(L, p)$ for any $\Psi \in \mathcal{NN}^1_{\sigma,L}$.

$\rightsquigarrow$ Universality fails!

Sufficiency

Sufficiency: It is enough to consider networks with 1D input

Lemma

If $\mathcal{NN}^1_{\sigma,L}$ is universal, then so is $\mathcal{NN}^d_{\sigma,L}$ for any $d \in \mathbb{N}$.

Sufficiency: It is enough to consider networks with 1D input

Lemma

If $\mathcal{NN}^{1}_{\sigma,L}$ is universal, then so is $\mathcal{NN}^{d}_{\sigma,L}$ for any $d \in \mathbb{N}$.

Proof.

Step 1: Assumption ensures:

$$\left(z \mapsto e^{\operatorname{Re} z}\right) \in \overline{\mathcal{NN}^{1}_{\sigma,L}}.$$

Step 2: This implies

$$\left(\mathbf{z} \mapsto e^{\operatorname{Re}\langle \mathbf{a},\mathbf{z}\rangle}\right) \in \overline{\mathcal{NN}^{d}_{\sigma,L}} \qquad \forall\, \mathbf{a} \in \mathbb{C}^{d}.$$

Step 3: By Stone-Weierstraß: The functions from Step 2 span a dense subspace of $C(K)$ for $K \subset \mathbb{C}^{d}$ compact. □

Proof of sufficiency for shallow complex-valued networks

For simplicity: Assume $\sigma \in C^\infty$ is smooth

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{N}\mathcal{N}^1_{\sigma,1}}$.

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{N}\mathcal{N}^1_{\sigma,1}}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \big[\sigma(wz + \theta)\big] = z^m \, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz + \theta)$$

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{N}\mathcal{N}^1_{\sigma,1}}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \big[\sigma(wz + \theta)\big] = z^m \, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz + \theta)$$

and hence

$$\partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \big[\sigma(wz + \theta)\big] = z^m \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(\theta).$$

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{NN}^1_{\sigma,1}}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \big[\sigma(wz+\theta)\big] = z^m\, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz+\theta)$$

and hence

$$\partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \big[\sigma(wz+\theta)\big] = z^m \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(\theta).$$

❷: We have $\big[z \;\mapsto\; \partial_w^m \overline{\partial}_w^\ell \big|_{w=0}\; \sigma(wz+\theta)\big] \in \overline{\mathcal{NN}^1_{\sigma,1}}$.

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{NN}_{\sigma,1}^1}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \left[\sigma(wz+\theta)\right] = z^m \, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz+\theta)$$

and hence

$$\boxed{\partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \left[\sigma(wz+\theta)\right] = z^m \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(\theta).}$$

❷: We have $\left[z \;\mapsto\; \partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \; \sigma(wz+\theta)\right] \in \overline{\mathcal{NN}_{\sigma,1}^1}$.

Proof idea: approximate derivative via difference quotient:

$$\frac{\partial}{\partial a} \sigma\big((a+ib)z+\theta\big) = \lim_{h\to 0} \frac{1}{h} \Big[\underbrace{\sigma\big((a+h+ib)z+\theta\big) - \sigma\big((a+ib)z+\theta\big)}_{\in \mathcal{NN}_{\sigma,1}^1 \text{ as a function of } z}\Big],$$

with locally uniform convergence. □

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{NN}_{\sigma,1}^1}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \left[\sigma(wz + \theta)\right] = z^m \, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz + \theta)$$

and hence

$$\partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \left[\sigma(wz + \theta)\right] = z^m \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(\theta).$$

❷: We have $\left[z \mapsto \partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \sigma(wz + \theta)\right] \in \overline{\mathcal{NN}_{\sigma,1}^1}$.

Proof idea: approximate derivative via difference quotient:

$$\frac{\partial}{\partial a} \sigma\big((a + ib)z + \theta\big) = \lim_{h \to 0} \frac{1}{h} \Big[\underbrace{\sigma\big((a + h + ib)z + \theta\big) - \sigma\big((a + ib)z + \theta\big)}_{\in \mathcal{NN}_{\sigma,1}^1 \text{ as a function of } z}\Big],$$

with locally uniform convergence. □

Corollary. If σ is not polyharmonic, then $\overline{\mathcal{NN}_{\sigma,1}^1} = C(\mathbb{C}; \mathbb{C})$.

Proof of sufficiency for shallow complex-valued networks

Proposition. If $m, \ell \in \mathbb{N}_0$ such that $\partial^m \overline{\partial}^\ell \sigma \not\equiv 0$, then $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{NN}^1_{\sigma,1}}$.

Proof sketch: ❶: Wirtinger calculus shows

$$\partial_w^m \overline{\partial}_w^\ell \big[\sigma(wz+\theta)\big] = z^m\, \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(wz+\theta)$$

and hence

$$\partial_w^m \overline{\partial}_w^\ell \big|_{w=0} \big[\sigma(wz+\theta)\big] = z^m \overline{z}^\ell \cdot (\partial^m \overline{\partial}^\ell \sigma)(\theta).$$

❷: We have $\big[z \;\mapsto\; \partial_w^m \overline{\partial}_w^\ell \big|_{w=0}\; \sigma(wz+\theta)\big] \in \overline{\mathcal{NN}^1_{\sigma,1}}$.

Proof idea: approximate derivative via difference quotient:

$$\frac{\partial}{\partial a} \sigma\big((a+ib)z+\theta\big) = \lim_{h\to 0} \frac{1}{h} \Big[\underbrace{\sigma\big((a+h+ib)z+\theta\big) - \sigma\big((a+ib)z+\theta\big)}_{\in \mathcal{NN}^1_{\sigma,1} \text{ as a function of } z}\Big],$$

with locally uniform convergence. □

Corollary. If σ is not polyharmonic, then $\overline{\mathcal{NN}^1_{\sigma,1}} = C(\mathbb{C};\mathbb{C})$.

Proof: ❶: We have $0 \not\equiv \Delta^k \sigma = 4^k \cdot \partial^k \overline{\partial}^k \sigma$ for all $k \in \mathbb{N}$.

❷: By the proposition, $(z \mapsto z^m \overline{z}^\ell) \in \overline{\mathcal{NN}^1_{\sigma,1}}$ for all m, ℓ. □

Sufficiency for deep networks

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be smooth, but not holomorphic, anti-holomorphic, or a polynomial.

Sufficiency for deep networks

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be smooth, but not holomorphic, anti-holomorphic, or a polynomial.

1

$$\left.\begin{array}{l} \sigma \text{ not holom.} \implies \overline{\partial}\sigma \not\equiv 0 \xrightarrow{\text{as before}} (z \mapsto \overline{z}) \in \overline{\mathcal{NN}^1_{\sigma,1}} \\ \sigma \text{ not anti-holom.} \implies \partial\sigma \not\equiv 0 \xrightarrow{\text{as before}} \boxed{(z \mapsto z) \in \overline{\mathcal{NN}^1_{\sigma,1}}} \end{array}\right\} \implies \boxed{(z \mapsto \operatorname{Re} z) \in \overline{\mathcal{NN}^1_{\sigma,1}}.}$$

Sufficiency for deep networks

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be smooth, but not holomorphic, anti-holomorphic, or a polynomial.

1

$$\left.\begin{array}{l} \sigma \text{ not holom.} \implies \overline{\partial}\sigma \not\equiv 0 \xRightarrow{\text{as before}} (z \mapsto \overline{z}) \in \overline{\mathcal{NN}^1_{\sigma,1}} \\ \sigma \text{ not anti-holom.} \implies \partial\sigma \not\equiv 0 \xRightarrow{\text{as before}} \boxed{(z \mapsto z) \in \overline{\mathcal{NN}^1_{\sigma,1}}} \end{array}\right\} \implies \boxed{(z \mapsto \operatorname{Re} z) \in \overline{\mathcal{NN}^1_{\sigma,1}}.}$$

2 Since σ is not a polynomial, we have

$$\forall m \in \mathbb{N}_0 : \quad \partial^m \sigma \not\equiv 0 \quad \text{or} \quad \overline{\partial}^m \sigma \not\equiv 0$$

$$\xRightarrow{\text{as before}} \forall m \in \mathbb{N}_0 : \quad (z \mapsto z^m) \in \overline{\mathcal{NN}^1_{\sigma,1}} \quad \text{or} \quad (z \mapsto \overline{z}^m) \in \overline{\mathcal{NN}^1_{\sigma,1}}$$

Sufficiency for deep networks

Let $\sigma : \mathbb{C} \to \mathbb{C}$ be smooth, but not holomorphic, anti-holomorphic, or a polynomial.

1

$$\left.\begin{array}{l} \sigma \text{ not holom.} \implies \overline{\partial}\sigma \not\equiv 0 \xrightarrow{\text{as before}} (z \mapsto \overline{z}) \in \overline{\mathcal{NN}^1_{\sigma,1}} \\ \sigma \text{ not anti-holom.} \implies \partial\sigma \not\equiv 0 \xrightarrow{\text{as before}} (z \mapsto z) \in \overline{\mathcal{NN}^1_{\sigma,1}} \end{array}\right\} \implies (z \mapsto \operatorname{Re} z) \in \overline{\mathcal{NN}^1_{\sigma,1}}.$$

2 Since σ is not a polynomial, we have

$$\forall\, m \in \mathbb{N}_0 : \quad \partial^m \sigma \not\equiv 0 \quad \text{or} \quad \overline{\partial}^m \sigma \not\equiv 0$$

$$\xrightarrow{\text{as before}} \quad \forall\, m \in \mathbb{N}_0 : \quad (z \mapsto z^m) \in \overline{\mathcal{NN}^1_{\sigma,1}} \quad \text{or} \quad (z \mapsto \overline{z}^m) \in \overline{\mathcal{NN}^1_{\sigma,1}}$$

3 Since we consider deep networks ($L \geq 2$), 1 and 2 imply

$$\forall\, m \in \mathbb{N}_0 : \left[z \mapsto (\operatorname{Re} z)^m\right] \in \overline{\mathcal{NN}^1_{\sigma,L}}.$$

This easily implies universality. □

Thanks for your attention ☺

Questions, comments, counterexamples?