FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

NATURWISSENSCHAFTLICHE
FAKULTÄT

Universal Approximation Theorem

Lecture “Mathematics of Learning” 2021/2022

Frauke Liers
Friedrich-Alexander-Universität Erlangen-Nürnberg

WIRTSCHAFTS
MATHEMATIK

Brief Summary

What we have learned on neural networks so far:

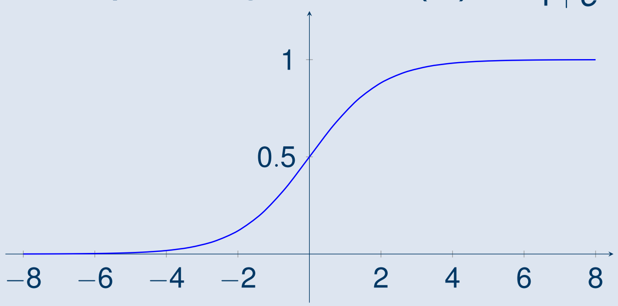
- They perform great in task like classification, image and language processing, gaming, etc.
- They are typically composed of elementary artificial neurons of the form $\Phi(x) = \psi(w \cdot x + b)$ with weights w, bias b, and activation ψ.

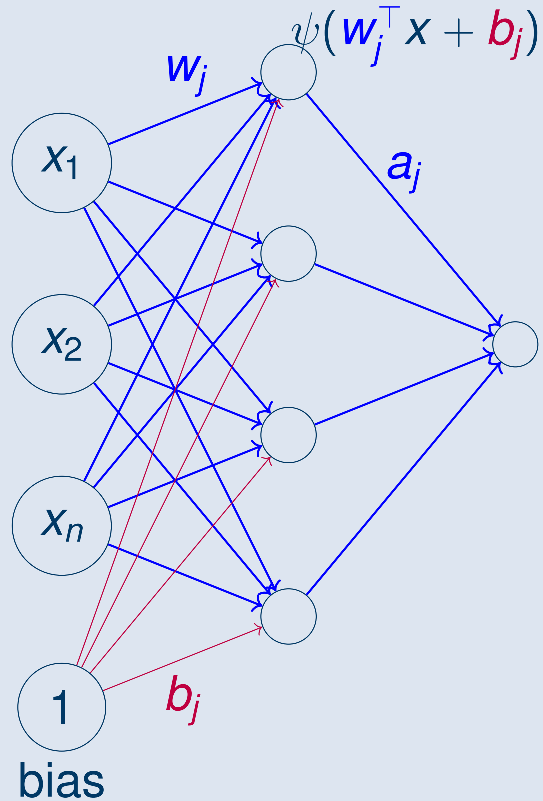
Feedforward Neural Networks with Activation Function ψ

Function $\psi : \mathbb{R} \to \mathbb{R}$ is called *sigmoidal*, if

$$\psi(z) \to \begin{cases} 1 & \text{for } \lim\limits_{z \to \infty} \\ 0 & \text{for } \lim\limits_{z \to -\infty} \end{cases}$$

example: sigmoid $\psi(z) = \frac{1}{1+e^{-z}}$





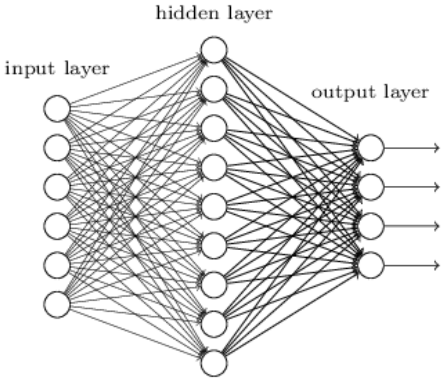
$\sum_{j=1}^{N} a_j \psi(w_j^\top x + b_j)$

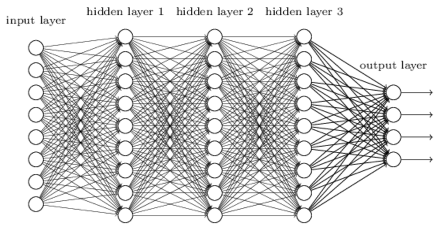
Which class of functions can be modelled (approximated) with feedforward neural networks? Does this depend on the size of the network?

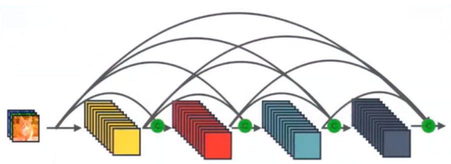
Universal Approximation Theorems

There are plenty of different Universal Approximation Theorems, which coarsely can be structured as follows:

- Arbitrary-width [1980s-1990s]
- Arbitrary-depth [2010s-2020s]
- Specific situations (discontinuous activations, architectures like *residual* or *convolutional* networks, *stable* networks) [1990s-2020s]







Topic of the lecture

- Let $X \subset \mathbb{R}^n . C(X)$: space of continuous functions in X.
- $I_n := [0,1]^n$: n-dimensional hyper cube
- supremum norm $|f| := \sup_{x \in I_n} |f(x)|$
- $M(I_n)$: space of signed regular Borel measures on I_n.

Question

Let us consider finite sums of the form

$$G(x) = \sum_{j=1}^{N} a_j \psi(w_j^\top x + b_j).$$

$G(x)$ consists of all functions that are generated by a feedforward neural network with a **single** hidden (inner) layer.
Which space do the functions $G(x)$ span?

Main Result

Universal Approximation Theorem

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function. Then finite sums of form

$$G(x) = \sum_{j=1}^{N} a_j \psi(w_j^\top x + b_j)$$

are dense in $C(I_n)$. This means: For an arbitrary continuous function $f \in C(I_n)$ and an arbitrary $\epsilon > 0$ there exists an N and a function G as above such that:

$$|G(x) - f(x)| < \epsilon \quad \forall\, x \in I_n.$$

This means:
neural feedforward networks with a discriminatory function and *a single* hidden layer can approximate continuous functions *arbitrarily well*.

Discussion

- The domain $[0, 1]^d$ is no severe restriction.
- There is not estimate on how large $N \in \mathbb{N}$ has to be.
- As we will see, the theorem holds true for a larger class of activations ψ, namely *discriminatory* ones.
- The proof uses abstract functional analysis and does not provide an algorithm how to compute G.
- For students without background in functional analysis, the proof may be difficult to understand. Even without this background, please make sure that you understand the *meaning* of the theorem!

Discriminatory Activations

Discriminatory Activations

Let $\Omega \subset \mathbb{R}^d$ be a compact set. A continuous function $\psi : \mathbb{R} \to \mathbb{R}$ is called discriminatory if for any signed measure $\mu \in M(\Omega)$:

$$\int_\Omega \psi(w \cdot x + b)\, \mathrm{d}\mu(x) = 0,\ \forall w \in \mathbb{R}^d,\ b \in \mathbb{R} \implies \mu = 0.$$

Counterexamples: $\psi(t) = 1$, $\psi(t) = t$, general polynomials [exercise]
Examples:

- sigmoidal functions, e.g., $\psi(t) = \frac{1}{1+e^{-t}}$ [later]
- rectified linear units, $\psi(t) = \mathrm{ReLU}(t) = \max(t, 0)$ [exercise]
- hyperbolic tangent, $\psi(t) = \tanh(t)$
- **any function which is no polynomial**

Preparations for the Proof

The Hahn–Banach Theorem

Let U be a normed vector space over $\mathbb{R}$ and $V \subset U$ a subspace such that $\overline{V} \neq U$. Then there exists a continuous linear map $\ell : U \to \mathbb{R}$ such that $\ell(x) = 0$ for all $x \in V$ but $\ell \not\equiv 0$.

For us: U = continuous functions on and V = single layer perceptrons

The Riesz Representation Theorem

Let $\Omega \subset \mathbb{R}^d$ be compact and $C(\Omega)$ denote the space of continuous functions on Ω. For any continuous linear map $\ell : C(\Omega) \to \mathbb{R}$ there exists a signed measure $\mu \in \mathfrak{M}(\Omega)$ such that

$$\ell(f) = \int_\Omega f(x)\,\mathrm{d}\mu(x).$$

For us: f = single layer perceptron with discriminatory activation

Main Result

Universal Approximation Theorem

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function. Then finite sums of form

$$G(x) = \sum_{j=1}^{N} a_j \psi(w_j^\top x + b_j)$$

are dense in $C(I_n)$. This means: For an arbitrary continuous function $f \in C(I_n)$ and an arbitrary $\epsilon > 0$ there exists an N and a function G as above such that:

$$|G(x) - f(x)| < \epsilon \quad \forall\, x \in I_n.$$

This means:
neural feedforward networks with a discriminatory function and *a single* hidden layer can approximate continuous functions *arbitrarily well*.

Proof Universal Approximation (Cybenko 1988)

Let $S \subset C(I_n)$ set of functions of form $G(x) \Rightarrow S$ is linear subspace from $C(I_n)$.
We show: $\overline{S} = C(I_n)$.
Assumption: This is not true, i.e., $R := \overline{S} \neq C(I_n)$.
Following the Hahn-Banach Theorem there exists a continuous linear functional L such that $L \neq 0$, but $L(R) = L(S) = 0$.
After Riesz representation theorem exists a $\mu \in M(I_n)$, such that L can be represented as

$$L(h) = \int_{I_n} h(x) d\mu(x)$$

for all $h \in C(I_n)$.
(to show: $L = 0$).
For all $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ we have in particular $\psi(w^\top x + b) \in R$. Thus it necessarily is true that for all $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ we have:

$$\int_{I_n} \psi(w^\top x + b) d\mu(x) = 0.$$

We have assumed that ψ is discrimatory, thus it is $\mu = 0$. $\rightarrow L(h) = 0$ for arbitrary $h \in C(I_n)$. This is a contradiction to Hahn-Banach Theorem, and so S is dense in $C(I_n)$. $\square$

Which functions are discrimatory?

Sums of form $G(x) = \sum_{j=1}^{N} a_j \psi(w_j^\top)x + b_j$ are dense in $C(I_n)$, if ψ is continuous and discriminatory. We show that each sigmoidal

$$\psi(z) \to \begin{cases} 1 & \text{for } \lim\limits_{z \to \infty} \\ 0 & \text{for } \lim\limits_{z \to -\infty} \end{cases}$$

is discriminatory.

Theorem

Each bounded measurable sigmoidal $\psi : \mathbb{R} \to \mathbb{R}$ is discriminatory.

Proof

Let ψ be such a sigmoidal We assume that we have for a measure $\mu \in M(I_n)$:

$$\int \psi(w^\top x + b) d\mu(x) = 0 \text{ for all } w \in \mathbb{R}^n, b \in \mathbb{R}.$$

We show that $\mu = 0$ is true.
For each x, w, b, ϕ it is

$$\psi(\lambda(w^\top x + b) + \phi)) \to \begin{cases} 1 & \text{for } w^\top x + b > 0 \text{ if } \lim_{\lambda\to\infty} \\ 0 & \text{for } w^\top x + b < 0 \text{ if } \lim_{\lambda\to\infty} \\ = \psi(\phi) & \text{for } w^\top x + b = 0 \text{ for all } \lambda. \end{cases}$$

This means that $\psi_\lambda(x) = \psi(\lambda(w^\top x + b) + \phi))$ for $\lambda \to \infty$ converges pointwise and bounded to

$$\gamma(x) = \begin{cases} 1 & \text{for } w^\top x + b > 0 \\ 0 & \text{for } w^\top x + b < 0 \\ \psi(\phi) & \text{for } w^\top x + b = 0. \end{cases}$$

Proof

$$\gamma(x) = \begin{cases} 1 & \text{for } w^\top x + b > 0 \\ 0 & \text{for } w^\top x + b < 0 \\ \psi(\phi) & \text{for } w^\top x + b = 0. \end{cases}$$

After theorem from Lebesgue

$$\int_{I_n} \gamma(x) d\mu(x) = \lim_{\lambda \to \infty} \int_{I_n} \psi_\lambda(\lambda(w^\top x + b) + \phi) d\mu(x) = 0.$$

We denote:

$$\begin{aligned} \text{hyperspace } \Pi_{w,b} &:= \{x \mid w^\top x + b = 0\} \\ \text{open halfspace } H^+_{w,b} &:= \{x \mid w^\top x + b > 0\} \\ \text{open halfspace } H^-_{w,b} &:= \{x \mid w^\top x + b < 0\} \end{aligned}$$

We then have for all w, b

$$\begin{aligned} \int_{I_n} \gamma(x) d\mu(x) &= \int_{H^+_{w,b}} 1 d\mu(x) + \int_{\Pi_{w,b}} \psi(\phi) d\mu(x) + \int_{H^-_{w,b}} 0 d\mu(x) \\ &= \mu(H^+_{w,b}) + \psi(\phi)\mu(\Pi_{w,b}) \\ &= 0. \end{aligned}$$

Proof

$\mu(H^+_{w,b}) + \psi(\phi)\mu(\Pi_{w,b}) = 0$ is true for arbitrary ϕ.
For $\phi \to \infty$ it is $\psi(\phi) = 1$ and thus we have

$$\mu(H^+_{w,b}) + \psi(\phi)\mu(\Pi_{w,b}) = \mu(H^+_{w,b}) + \mu(\Pi_{w,b}) = 0$$

Similarly: For $\phi \to -\infty$ ist $\psi(\phi) = 0 \Rightarrow$

$$\mu(H^+_{w,b}) = 0,$$

i.e., the measure of all halfspaces is zero. We need to show:

$$\mu(\Pi_{w,b}) = 0 \Rightarrow \mu = 0$$

This is not obvious, as due to the Theorem by Riesz we have μ is not necessarily zero.

Brief Repetition

Repetition: Theorem majorised convergence (Lebesgue)

Let X be measureable space, μ Borel measure on X, $f : X \to \mathbb{R}$ measurable and $\{f_n\}$ series of measurable functions such that

- $\lim\limits_{n\to\infty} f_n(x) = f(x)$ for μ-almost all $x \in X$
- There exists $g \in \mathcal{L}^p(X)$ with $|f_n(x)| \leq g(x)$ for μ-almost all $x \in X$

Then f is μ-integrable and it is $\lim\limits_{n\to\infty} \int_X f_n(x) d\mu(x) = \int_X f(x) d\mu(x)$.

Proof, continued

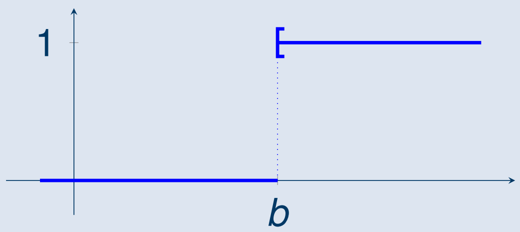
Let w arbitrary but fixed. For arbitrary measurable function h we define linear functional F:

$$F(h) := \int_{I_n} h(w^\top x) d\mu(x)$$

F is a bounded functional in Lebesgue space $L^\infty(\mathbb{R})$ because μ is a bounded signed measure.

Let h indicator function of $[b, \infty]$, i.e., $h(u) = 1$ for $u \geq b$, $h(u) = 0$ for $u < b$.

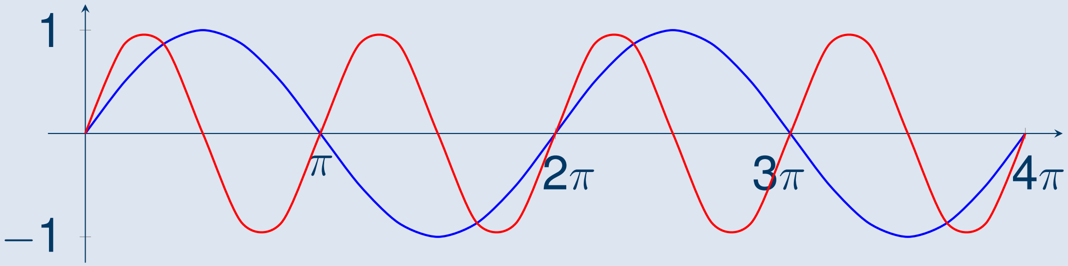
$$\Pi_{w,b} := \{x \mid w^\top x + b = 0\}$$
$$H^+_{w,b} := \{x \mid w^\top x + b > 0\}$$



$$\Rightarrow F(h) = \int_{I_n} h(w^\top x) d\mu(x) = \mu(\Pi_{w,-b}) + \mu(H^+_{w,-b}) = 0$$

Analogously: $F(h) = 0$ on (b, ∞). Due to the linearity we thus have $F(h) = 0$ for all indicator functions of an arbitrary interval and thus also for each simple function (sum of indicator functions and intervals.)
we know: simple functions are dense in $L^\infty(\mathbb{R}) \Rightarrow F = 0$.

Proof



concrete selection: for bounded measurable function
$s(u) = \sin(mu), c(u) = \cos(mu)$ we have for all $m \in \mathbb{R}$

$$F(s + ic) = \int_{I_n} \sin(mx) + i\cos(mx) d\mu(x) = \int_{I_n} \exp(imx) d\mu(x) = 0$$

$\Rightarrow$ Fourier transform of μ is zero. As this is dense in the space of continuous functions $\Rightarrow \mu = 0 \Rightarrow \psi$ discriminatory.

Remarks

- One prove analogous statements using other dualities, e.g., $L^1(\Omega)$ and $L^\infty(\Omega)$, adapting the Riesz representation and the definition of discriminatory activations.
- Any ReLU-activated single layer perceptron with N neurons can be rewritten as deep N-layer perceptron with width $d + 2$.
- This yields a universal approximation theorem for deep ReLU networks with bounded width.

Observations

Summary

- **Remark**: Cybenko's statement follows from choosing $\Omega = [0, 1]^d$ and using that sigmoidal functions are discriminatory.
- As each continuous sigmoidal is discriminatory, the functions $G(x)$ are dense in $C(I_n)$.
- feedforward neural networks with a single hidden layer and sigmoidal functions ($\frac{1}{1+e^{-x}}, \mathtt{tanh}(x)$...) as activation function *approximate* continuous functions arbitrarily well, (not necessarily exact)
- This existence result is not constructive, and it does not tell us how to construct the network!
- This is typically not possible for non-continuous functions (or only as good as a noncontinuous function can be approximated by a continuous function.

References

- Cybenko. Approximation by superpositions of a sigmoidal function. Math Cont Sig Syst, 1989.
- Hornik, Stinchcombe, White. Multi-layer feedforward networks are universal approximators, 1988.
- Hornik. Approximation capabilities of multilayer feedforward networks, Neural Networks, 1991.