Universal algebra gives universal approximation for neural nets

Charlotte Aten

University of Rochester

2021 March 25

Introduction

- I first tried looking into the formal mathematical treatment of neural nets some time in 2019.
- At the time the only class of results I found were universal approximation theorems.
- Variants of the original results of Cybenko (1989) and Hornik (1991) are still being published.
- Since I have a universal algebra background I naturally asked: ¿Doesn't Murskiĭ's Theorem say something quite similar to these results?
- I will spend the rest of the talk explaining what I mean by this.

Talk outline

- Neural nets and universal approximation
- Discrete neural nets and finite algebras
- Clones and primality
- Statement of Murskiĭ's Theorem
- Sketch of the proof of Murskiĭ's Theorem

Neural nets

Definition (Neural net)

A *neural net* $(V_1, \ldots, V_r, E, \Phi)$ with r layers consists of

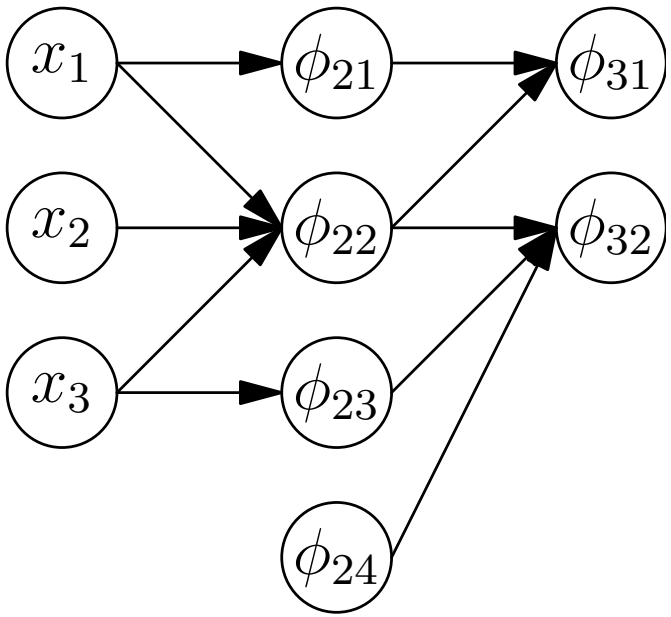
1. a finite digraph (V, E) (the *architecture* of the neural net) and
2. for each $v \in V \setminus V_1$ a function $\Phi(v)\colon \mathbb{R}^{\rho(v)} \to \mathbb{R}$ (the *activation function* of v)

where

1. $V := \bigcup_{i=1}^{r} V_i$,
2. the only edges in E are from vertices in V_i to vertices in V_{i+1} for $i < r$,
3. $\rho(v)$ is the indegree of v in (V, E), and
4. if $i \neq r$ then every vertex $v \in V_i$ has nonzero outdegree.

Neural nets

A typical neural net. A node $v_{ij} \in V_i$ is called a *neuron* in layer i. We will denote $\Phi(v_{ij})$ by ϕ_{ij}.



We think of the labels x_j as variables.

Neural nets

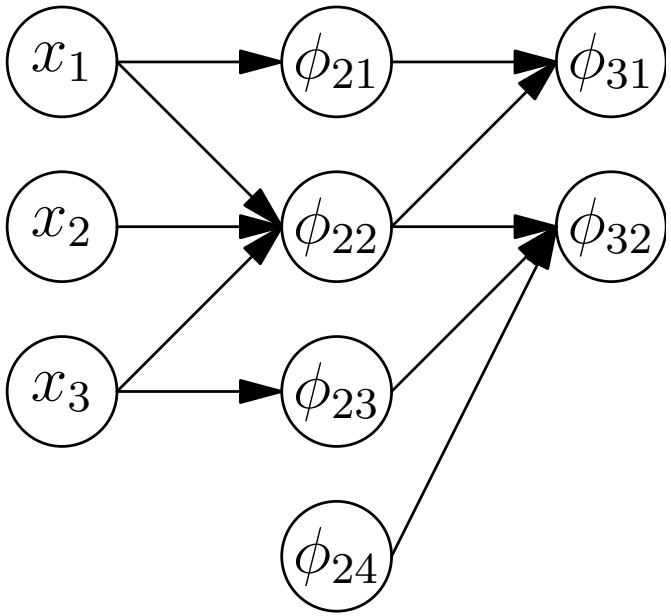
- The activation functions are typically restricted to a nicer class.
- The standard family consists of *sigmoid functions* of the form $\phi_{ij}\colon \mathbb{R}^{\rho(v_{ij})} \to \mathbb{R}$ where

$$\phi_{ij}(z) := \frac{1}{1 + e^{-z \cdot w}}$$

  for some parameter $w \in \mathbb{R}^{\rho(v_{ij})}$, called the *weight*.
- In applications neural nets are initialized with a fixed architecture and some randomly chosen weights, then trained by adjusting these weights.
- In this talk we won't consider how this training is performed, but only what the goal of the training is.

Neural nets

In order to understand that goal, we need to see that neural nets represent functions. Our example neural net represents a function $g\colon \mathbb{R}^3 \to \mathbb{R}^2$.



That function is given by

$$g(x_1, x_2, x_3) := (\phi_{31}(\phi_{21}(x_1), \phi_{22}(x_1, x_2, x_3)),$$
$$\phi_{32}(\phi_{22}(x_1, x_2, x_3), \phi_{23}(x_3), \phi_{24}())).$$

Neural nets

Definition (Function represented by a neural net)

Given a neural net $N_r := (V_1, \ldots, V_r, E, \Phi)$ the *function represented* by N_r is $g_r \colon \mathbb{R}^{|V_1|} \to \mathbb{R}^{|V_r|}$ where

1. $g_r = \mathrm{id}_{\mathbb{R}^{|V_1|}}$ when $r = 1$ and
2. when $r > 1$ we set $(g_r(x))_j := \phi_{rj}((g_{r-1}(x))_k)_{(v_{r-1,k}, v_{r,j}) \in E}$ where $x = (x_1, \ldots, x_{|V_1|})$ and g_{r-1} is the function represented by the neural net $N_{r-1} := (V_1, \ldots, V_{r-1}, E', \Phi')$ obtained by deleting the r^{th} layer of N_r.

This definition is using the ordering we placed on the V_i in our example, but this dependence on the ordering can be removed.

Neural nets

- The goal of training is to produce from a neural net with a specified architecture and starting weights a neural net with the same architecture which represents (with an acceptably small error) a target function $h\colon \mathbb{R}^{|V_1|} \to \mathbb{R}^{|V_r|}$.

Universal approximation

- We give a simplified version of Cybenko's result as an example of universal approximation for neural nets.
- Cybenko considers a neural net with $r = 3$ whose activation functions are all either from the sigmoid family described above, are nullary (constant) functions, or are the dot product.
- Each function in the sigmoid family can be written as $\sigma(z \cdot w)$ where
$$\sigma(t) := \frac{1}{1 + e^{-t}}.$$
- Each constant function $\mathbb{R}^0 \to \mathbb{R}$ can be viewed as its image $\theta \in \mathbb{R}$.

Universal approximation

- Such a neural net can be constructed to represent any function of the form

$$g(z) := \sum_{i=1}^{k} \alpha_i \sigma(z \cdot w_i + \theta_i).$$

- Partition $I_n := [0,1]^n$ into s disjoint, measurable subsets $P_1, \ldots, P_s$.
- The *decision function* $h\colon I_n \to \mathbb{R}$ of this partition is given by

$$h(z) := j \text{ when } z \in P_j.$$

- Decision functions are sufficiently general to cover the target functions appearing in applications of neural nets.

Universal approximation

Theorem (Cybenko, 1989)

Given a decision function h for a finite measurable partition of I_n we have for any $\epsilon > 0$ that there is a neural net (of the form described previously) which represents a function

$$g(z) = \sum_{i=1}^{k} \alpha_i \sigma(z \cdot w_i + \theta_i)$$

and a set $D \subset I_n$ with $m(D) \geq 1 - \epsilon$ on which $|g(z) - h(z)| < \epsilon$.

- Cybenko actually proved something more general but this is enough for our purposes.
- This result shows that there is a neural network with $r = 3$ («one hidden layer») which represents with arbitrary precision any target function we could reasonably choose.

Discrete neural nets

Definition (Neural net)

A *discrete neural net* $(V_1, \ldots, V_r, E, \Phi)$ with r layers on a finite set A consists of

1. a finite digraph (V, E) (the *architecture* of the neural net) and
2. for each $v \in V \setminus V_1$ a function $\Phi(v)\colon A^{\rho(v)} \to A$ (the *activation function* of v)

where

1. $V := \bigcup_{i=1}^{r} V_i$,
2. the only edges in E are from vertices in V_i to vertices in V_{i+1} for $i < r$,
3. $\rho(v)$ is the indegree of v in (V, E), and
4. if $i \neq r$ then every vertex $v \in V_i$ has nonzero outdegree.

Discrete neural nets

- To my knowledge this analogue of neural nets hasn't been discussed before.
- Functions represented by discrete neural nets and target functions for discrete neural nets are defined analogously.
- Training a discrete neural net should be done by varying the activation functions among those in a chosen family.
- The goal of training is to produce from a neural net with a specified architecture and starting activation functions a neural net with the same architecture which represents (with an acceptably small error) a target function $h\colon A^{|V_1|} \to A^{|V_r|}$.
- Universal approximation for discrete neural nets then has to do with which such functions h can be written as composites of some fixed (say finite) family of operations on A.

Finite algebras

Now it's time for a really quick crash course in universal algebra.

Finite algebras

Operations are rules for combining elements of a set together to obtain another element of the same set.

Definition (Operation, arity)

Given a set A and some $n \in \mathbb{W}$ we refer to a function $f\colon A^n \to A$ as an *n-ary operation* on A. When f is an n-ary operation on A we say that f has *arity* n.

Finite algebras

Algebras are sets with an indexed sequence of operations.

Definition (Algebra)

An *algebra* (A, F) consists of a set A and a sequence $F = \{f_i\}_{i \in I}$ of operations on A, indexed by some set I.

An algebra (A, F) is called *finite* when A is a finite set. Whether or not A is finite it is conventional to assume I is finite in most cases, as we will for the rest of this talk.

Finite algebras

- Given an algebra $\mathbf{A} := (A, \{f_i\}_{i \in I})$ we define a map $\rho\colon I \to \mathbb{W}$ where $\rho(i) := n$ when $f_i\colon A^n \to A$ is an n-ary operation on A.
- This map $\rho\colon I \to \mathbb{W}$ is called the *similarity type* of $\mathbf{A}$.
- When two algebras $\mathbf{A} := (A, F)$ and $\mathbf{B} := (B, G)$ have the same similarity type $\rho\colon I \to \mathbb{W}$ we say that $\mathbf{A}$ and $\mathbf{B}$ are *similar algebras*.
- Given a similarity type ρ and $n \in \mathbb{N}$ we define

$$\mathrm{Alg}_{\rho,n} := \{\, \mathbf{A} \mid \mathbf{A} \text{ has type } \rho \text{ and } A = [n] \,\}.$$

Finite algebras

- Given a property P of finite algebras we define

$$\mathrm{Alg}_{\rho,n}[P] \coloneqq \left\{ \mathbf{A} \in \mathrm{Alg}_{\rho,n} \;\middle|\; \mathbf{A} \text{ has } P \right\}.$$

- The probability that a finite algebra of type ρ has property P is then defined to be

$$\mathrm{Pr}_{\rho}(P) \coloneqq \lim_{n\to\infty} \frac{\left|\mathrm{Alg}_{\rho,n}[P]\right|}{\left|\mathrm{Alg}_{\rho,n}\right|}.$$

- This gives a finitely additive probability measure on the set $\mathrm{FinAlg}_{\rho} \coloneqq \bigcup_{n\in\mathbb{N}} \mathrm{Alg}_{\rho,n}$.

Clones

- We will need to keep track of all functions which can be built using the basic operations of an algebra.
- Given $n \in \mathbb{W}$ and a set A we define $\mathrm{Op}_n(A) := A^{(A^n)}$.
- Given $n, k \in \mathbb{W}$, $f \in \mathrm{Op}_n(A)$, and $g_1, \ldots, g_n \in \mathrm{Op}_k(A)$ the *generalized composite*

$$f[g_1, \ldots, g_n] \colon A^k \to A$$

is given by

$$f[g_1, \ldots, g_n](x_1, \ldots, x_k) := f(g_1(x), \ldots, g_n(x)).$$

- Note that $\mathrm{Op}_n(A)$ contains all the coordinate projections p_k^n where

$$p_k^n(x_1, \ldots, x_n) := x_k.$$

Clones

Definition (Clone)

Given a nonempty set A we say that $C \subset \mathrm{Op}(A) := \bigcup_{n \in \mathbb{W}} \mathrm{Op}_n(A)$ is a *clone* when C is closed under generalized composition and contains all the coordinate projection operations.

- The largest clone on A is $\mathrm{Op}(A)$ itself.
- The smallest clone on A is $\mathrm{Proj}(A) := \{\, p_k^n \mid 1 \leq k \leq n \in \mathbb{W} \,\}$.
- Clones can be viewed as algebras themselves but that treatment isn't necessary for this discussion.
- For topologists: Clones are examples of operads whose operation spaces are discrete. For what it's worth, clone theory dates back to at least the 1940s.

Clones

Definition (Term)

Given a similarity type $\rho: I \to \mathbb{W}$, a set of variables X, and a set $F := \{f_i\}_{i \in I}$ which we think of as abstract basic operation symbols, a *term* in the language of ρ in the variables X is an element of the set $T_\rho(X) := \bigcup_{n \in \mathbb{W}} T_n$ where

$$T_0 := X \cup \{\, f_i \mid \rho(i) = 0 \,\}$$

and for $n \in \mathbb{W}$ we set

$$T_{n+1} := T_n \cup \{\, f_i[t_1, \ldots, t_k] \mid i \in I,\ k = \rho(i), \text{ and } t_1, \ldots, t_k \in T_n \,\}.$$

Clones

- That is, $T_\rho(X)$ consists of all valid formal composites of the basic operation symbols $\{f_i\}_{i\in I}$ whose arities are given by ρ with variable arguments coming from the set X.
- Given an algebra $\mathbf{A}$ of signature ρ and a term $t(x_1,\ldots,x_n) \in T_\rho(\{x_1,\ldots,x_n\})$ we define the *term operation*

$$t^{\mathbf{A}}\colon A^n \to A$$

  by interpreting all the operation symbols appearing in t as actual basic operations of $\mathbf{A}$ in the obvious way.
- For example, if ρ is the usual signature for groups then $(xy)(x^{-1}y^{-1})$ is a term in the variables $\{x,y\}$ whereas there exists an actual commutator term operation on the symmetric group $\mathbf{S}_3$ which is a binary operation on S_3.

Clones

- Each algebra $\mathbf{A}$ has a corresponding clone of term operations, which is

$$\mathrm{Clo}(\mathbf{A}) := \bigcup_{n \in \mathbb{W}} \mathrm{Clo}_n(\mathbf{A})$$

  where

$$\mathrm{Clo}_n(\mathbf{A}) := \left\{ t^{\mathbf{A}} \;\middle|\; t \in T_\rho(\{x_1, \ldots, x_n\}) \right\}.$$

- This is to say that $\mathrm{Clo}(\mathbf{A})$ consists of all the operations on A which can be built up using the basic operations of A and (implicitly) projections.
- Another way of saying this is that $\mathrm{Clo}(\mathbf{A})$ is the smallest clone in the lattice of clones on A which contains the basic operations of $\mathbf{A}$.

Primal algebras

Definition (Primal algebra)

We say that an algebra $\mathbf{A}$ is *primal* when for each $n > 0$ we have that $\mathrm{Clo}_n(\mathbf{A}) = \mathrm{Op}_n(A)$.

- Primal algebras are those which allow us to express any (nonconstant) operation as a composite of their basic operations.
- Finite fields of the form $\mathbb{F}_p$ for a prime p are primal.
- Finite fields of the form $\mathbb{F}_{p^k}$ for $k > 1$ are not primal.
- The two-element Boolean algebra $\mathbf{B}_2 := (\{0, 1\}, \wedge, \vee, 0, 1, ')$ is primal.
- J. B. Nation has an excellent survey of logic on other planets which describes other primal (or functionally complete) algebras which play the role of $\mathbf{B}_2$ in alien computer systems.
- Some are even based on the game rock-paper-scissors...

Primal algebras

- We introduce one last notion related to primality.
- An operation $f\colon A^n \to A$ is said to be *idempotent* when for all $a \in A$ we have that

$$f(a, a, \ldots, a) = a.$$

- An algebra $\mathbf{A}$ is called *idemprimal* when $\mathrm{Clo}(\mathbf{A})$ contains all idempotent operations on A.

Murskiĭ's Theorem

- For the rest of this talk we take P to denote the property of being primal and I to denote the property of being idemprimal.
- In 1968 R. O. Davies proved that if ρ is a similarity type containing a single k-ary operation symbol with $k > 1$ then $\Pr_\rho(P) = 1/e$.
- In the 1970s V. L. Murskiĭ proved that under the same assumption on ρ we have $\Pr_\rho(I) = 1$.
- He also proved a result about primality for signatures with more basic operations.

Murskiĭ's Theorem

Theorem (Murskiĭ, 1970s)

If ρ is a similarity type which contains at least two basic operations, at least one of which is not unary, then $\Pr_\rho(P) = 1$.

Murskiĭ's Theorem (neural net interpretation)

Theorem (Murskiĭ, 1970s)

If ρ is a similarity type which contains at least two basic operations, at least one of which is not unary, then $\Pr_\rho(P) = 1$.

Theorem (Murskiĭ, 1970s, interpreted in our context)

If ρ is a similarity type which contains at least two basic operations, at least one of which is not unary, then a randomly-selected finite algebra $\mathbf{A}$ *of signature ρ has (with probability 1) the property that given any target function $h\colon A^n \to A^m$ there exists a discrete neural net $(V_1, \ldots, V_r, E, \Phi)$ whose activation functions are all basic operations of* $\mathbf{A}$ *or projections which represents h.*

Proof sketch

- The proof that $\mathrm{Pr}_\rho(P) = 1$ given that ρ has at least two basic operations, at least one of which is nonunary, is relatively direct once we have that $\mathrm{Pr}_\rho(I) = 1$ in the case that ρ has a single binary operation.
- One can show that a finite algebra is primal if and only if it is idemprimal and has no trivial subalgebras.
- Let E denote the property of having no trivial subalgebras.
- We have that $P = E \cap I$ and we denote by $\bar{I}$ the complement of I in FinAlg_ρ.
- It follows from the fact that an algebra $(A, f\colon A^k \to A)$ with $k > 2$ has a binary operation in its clone that $\mathrm{Pr}_\rho(I) = 1$ for any ρ with a single k-ary operation.

Proof sketch

- We have that $\Pr_\rho(E) = 1$ in the case that ρ has at least two operations, at least one of which is nonunary.
- Since
$$E = (I \cap E) \coprod (\bar{I} \cap E) = P \coprod (\bar{I} \cap E)$$
we have that
$$1 = \Pr_\rho(E) = \Pr_\rho(P) + \Pr_\rho(\bar{I} \cap E).$$
- Assuming we can show $\Pr_\rho(I) = 1$ in this case we have that $\Pr_\rho(\bar{I} \cap E) \leq \Pr_\rho(\bar{I}) = 0$ and hence $\Pr_\rho(P) = 1$.

Proof sketch

- Let's now look at the proof that a random finite magma is idemprimal with probability 1.
- This is done by splitting the class of nonidemprimal magmas into 10 (overlapping) classes and showing that each of these occurs with probability 0.

Proof sketch

Let $\mathbf{A} := (A, \cdot)$ be a finite magma of order n. If $\mathbf{A}$ is not idemprimal then at least one of the following holds where $X, Y \subset A$, $a, b, c \in A$, and $\alpha, \beta \in \mathrm{Perm}(A)$.

1. $(\exists X)(2 \leq |X| \leq n-1 \text{ and } X \cdot X \subset X)$
2. $(\exists X)(3 \leq |X| \leq n-1 \text{ and } |X \cdot X| \leq |X|)$
3. $(\exists X)(|X| = 2 \text{ and } |X \cdot X| = 1)$
4. $(\exists X, Y)(|X| = |Y| = 2,\ X \cdot X = Y, \text{ and } |Y \cdot Y| = 2)$
5. $A \cdot A \neq A$
6. $(\exists a, b)(a \neq b,\ a \cdot a = a \cdot b = b \cdot a = a)$
7. $(\exists X)(1 \leq |X \cdot A| \leq |X| \leq n-1)$
8. $(\exists X)(1 \leq |A \cdot X| \leq |X| \leq n-1)$
9. $(\exists a, b)(a \neq b \text{ and } (\forall c)(a \cdot c = b \cdot c))$
10. $(\exists \alpha, \beta)(\alpha \neq \mathrm{id}_A \text{ and } (\forall a, b)(\alpha(a) \cdot \alpha(b) = \beta(a \cdot b)))$

Proof sketch

- We'll only consider case (2), which says that

$$(\exists X)(3 \leq |X| \leq n-1 \text{ and } |X \cdot X| \leq |X|).$$

- One can show that for any $c \in \{1, \ldots, 10\}$ a random magma satisfies condition (c) and not condition (2) with probability 0.
- We will now show that the probability that a random magma of order n satisfies condition (2) is at most $\sum_{k=3}^{n-1} \psi_n(k)$ where

$$\psi_n(k) := \binom{n}{k}^2 \left(\frac{k}{n}\right)^{k^2}.$$

Proof sketch

- We will now show that the probability that a random magma of order n satisfies condition (2) is at most $\sum_{k=3}^{n-1} \psi_n(k)$ where

$$\psi_n(k) := \binom{n}{k}^2 \left(\frac{k}{n}\right)^{k^2}.$$

- Condition (2) is equivalent to having that

$$(\exists X, Y \subset A)(3 \leq |X| = |Y| \leq n-1 \text{ and } X \cdot X \subset Y).$$

- Consider $X, Y \in \binom{A}{k}$ where $3 \leq k \leq n-1$.
- The probability that $X \cdot X \subset Y$ is the proportion of Cayley tables for a binary operation so that the k^2-many positions corresponding to $X \times X$ all take values in Y.
- A particular spot on the Cayley table is in Y with probability k/n.
- Since there are k^2 spots and $\binom{n}{k}$ ways to choose each of X and Y the claim follows.

Proof sketch

- We will be done if we can show that

$$\lim_{n\to\infty} \sum_{k=3}^{n-1} \psi_n(k) = 0.$$

- We do this by showing that there are $c, d \in (0,1)$ such that each of the sums

$$\sum_{3\le k\le cn} \psi_n(k) \qquad \sum_{cn\le k\le dn} \psi_n(k) \qquad \sum_{dn\le k\le n-1} \psi_n(k)$$

converges to 0 as $n \to \infty$.

Proof sketch

- We'll just look at one of the three cases we need to cover, namely that there exists some $d \in \left(\frac{1}{2}, 1\right)$ such that

$$\sum_{dn \leq k \leq n-1} \psi_n(k) \to 0.$$

- For this we need a sharp version of Stirling's approximation for $x!$ which was given by Robbins in 1955.
- For all $x \geq 1$ we have

$$\sqrt{2\pi} x^{x+\frac{1}{2}} e^{-x} e^{(12x+1)^{-1}} \leq x! \leq \sqrt{2\pi} x^{x+\frac{1}{2}} e^{-x} e^{(12x)^{-1}}.$$

- It follows for integers a, b where $a > b \geq 1$ that

$$\binom{a}{b} < \sqrt{a} \left(\left(\frac{b}{a}\right)^{\frac{b}{a}} \left(\frac{a-b}{a}\right)^{\frac{a-b}{a}} \right)^{-a}.$$

Proof sketch

- Taking $\frac{1}{2} < d < 1$ and setting $u = \lfloor dn \rfloor$ we have that

$$\binom{n}{u}^2 < n\left(\left(\frac{u}{n}\right)^{\frac{u}{n}}\left(\frac{n-u}{n}\right)^{\frac{n-u}{n}}\right)^{-2n}.$$

- This bound can be used to show that

$$\binom{n}{u}^2 < n\left(\left(d-\frac{1}{n}\right)^{d}(1-d)^{1-d}(1-d)^{\frac{1}{n}}\right)^{-2n}.$$

- We also have that

$$\left(\frac{k}{n}\right)^{k^2} \leq \left(\frac{n-1}{n}\right)^{(dn)^2}.$$

Proof sketch

- Combining these estimates yields

$$\psi_n(k) \leq n \left(\left(\frac{1}{\left(d - \frac{1}{n}\right)^d (1-d)^{1-d}} \right)^2 \left(\frac{1}{e} \right)^{d^2} \right)^n (1-d)^{-2}.$$

- For $n > 4$ we have that

$$\lim_{d \to 1^-} de^{d^2} \left(\left(d - \frac{1}{n} \right)^d (1-d)^{1-d} \right)^2 \geq 1.$$

- Choosing $d < 1$ large enough so that the inequality is realized we have that

$$\psi_n(k) \leq n d^n (1-d)^{-2}$$

and hence

$$\sum_{dn \leq k \leq n-1} \psi_n(k) \leq n^2 d^n (1-d)^2 \to 0.$$

References

- G. Cybenko. "Approximation by Superpositions of a Sigmoidal Function". In: *Math. Control Signals Systems* 2 (1989), pp. 303–314
- Toshinori Munakata. *Fundamentals of the New Artificial Intelligence*. Springer, 2008. ISBN: 978-1-84628-838-8
- Clifford Bergman. *Universal Algebra: Fundamentals and Selected Topics*. Chapman and Hall/CRC, 2011. ISBN: 978-1-4398-5129-6