DOI: 10.51790/2712-9942-2023-4-4-04

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ АППРОКСИМАЦИИ И ИНТЕРПОЛЯЦИИ

А. Д. Смородинов[1,2,а], **Т. В. Гавриленко**[1,2,б], **В. А. Галкин**[1,2,в]

[1] *Сургутский филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Сургут, Российская Федерация*

[2] *Сургутский государственный университет, г. Сургут, Российская Федерация*

[а] *ORCID: http://orcid.org/0000-0002-9324-1844,* ✍ *Sachenka_1998@mail.ru*

[б] *ORCID: http://orcid.org/0000-0002-3243-2751, taras.gavrilenko@gmail.com*

[в] *ORCID: http://orcid.org/0000-0002-9721-4026, email: val-gal@yandex.ru*

Аннотация: в статье проведено исследование теоретической базы искусственных нейронных сетей. Изучалось теоретическое обоснование возможности аппроксимации функций многих переменных суперпозицией функций одной переменной. Рассмотрены основные универсальные теоремы аппроксимации, представленные и доказанные к настоящему моменту зарубежными и отечественными авторами. Рассмотрены теоремы аппроксимации, в которых представлено необходимое количество нейронов в слое — ограничение по ширине; теоремы, в которых показано необходимое количество слоев в нейронной сети — ограничение по глубине; теоремы, в которых авторы доказывают минимальные границы одновременно для количества слоев в сети и количества нейронов на слое — ограничения по глубине и ширине.

Ключевые слова: универсальная теорема аппроксимации, теорема Колмогорова—Арнольда, теорема Цыбенко, аппроксимация функций, искусственные нейронные сети.

Для цитирования: Смородинов А. Д., Гавриленко Т. В., Галкин В. А. Теоретические основы искусственных нейронных сетей для решения задачи аппроксимации и интерполяции. *Успехи кибернетики*. 2023;4(4):41–53. DOI: 10.51790/2712-9942-2023-4-4-04.

Поступила в редакцию: 13.12.2023. *В окончательном варианте: 20.12.2023.*

THEORETICAL FOUNDATIONS OF ARTIFICIAL NEURAL NETWORK APPLICATION TO APPROXIMATION AND INTERPOLATION PROBLEMS

A. D. Smorodinov[1,2,a], **T. V. Gavrilenko**[1,2,b], **V. A. Galkin**[1,2,c]

[1] *Surgut Branch of Federal State Institute "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", Surgut, Russian Federation*

[2] *Surgut State University, Surgut, Russian Federation*

[a] *ORCID: http://orcid.org/0000-0002-9324-1844,* ✍ *Sachenka_1998@mail.ru*

[b] *ORCID: http://orcid.org/0000-0002-3243-2751, taras.gavrilenko@gmail.com*

[c] *ORCID: http://orcid.org/0000-0002-9721-4026, val-gal@yandex.ru*

Abstract: we studied the theoretical foundations of artificial neural networks as applied to the possibility of approximating functions of many variables by superposition of functions of one variable. We considered the most important universal approximation theorems. We also studied the approximation theorems with the required number of neurons in a layer (width constraint) or the number of layers in a neural network (depth constraint), and the theorems in which their authors prove the existence of min bounds both for the number of layers and for the number of neurons per layer.

Keywords: universal approximation theorem, Kolmogorov-Arnold theorem, Tsybenko theorem, approximation of functions, artificial neural networks.

Cite this article: Smorodinov A. D., Gavrilenko T. V., Galkin V. A. Theoretical Foundations of Artificial Neural Network Application to Approximation and Interpolation Problems. *Russian Journal of Cybernetics*. 2023;4(4):41–53. DOI: 10.51790/2712-9942-2023-4-4-04.

Original article submitted: 13.12.2023. *Revision submitted: 20.12.2023.*

Одной из основных проблем массового применения систем, основанных на искусственных нейронных сетях (ИНС), является так называемый эффект черного ящика, при котором зачастую даже сами разработчики алгоритмов не имеют четких представлений о том, каким образом достигается результат. Кроме того, на данный момент четкие, теоретически обоснованные и, самое главное, структурированные правила конструирования и обучения ИНС отсутствуют. Существует большое количество эвристических правил, которые помогают разработчикам систем с ИНС, но нет правил, позволяющих гарантировать эффективное решение задачи или четкое достижение цели, что в некоторых случаях ведет к трагичным результатам. В качестве примера можно привести автокатастрофы с участием автомобилей Tesla, когда они «не видят» пешеходов или припаркованные грузовики.

Как отмечает академик В.Б. Бетелин [1], одной из причин слабой применимости систем на основе ИНС является отсутствие каких-либо теоретических обоснований устойчивости и сходимости ИНС, что в свою очередь не гарантирует получение надежного результата. На данный момент вопросы о возможностях ИНС, а также о необходимой размерности, структуре сети, правилах выбора функций активаций и определения весовых коэффициентов активно обсуждается в зарубежном научном сообществе. Целью данной работы является исследование теоретической базы ИНС для решения задачи аппроксимации.

Одной из основных работ, доказывающих возможность применения ИНС для задач аппроксимации функции многих переменных, является работа Колмогорова—Арнольда, в которой представлена одна из центральных теорем искусственного интеллекта — теорема Колмогорова—Арнольда [2]. Формулировка теоремы Колмогорова—Арнольда о представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одной переменной и сложения следующая.

Теорема Колмогорова—Арнольда

При любом целом $n \geq 2$ *существуют такие определенные на единичном отрезке* $E^1 = [0; 1]$ *непрерывные действительные функции* $\psi^{pq}(x)$ *, что каждая определенная на n-мерном единичном кубе* E^n *непрерывная действительная функция* $f(x_1, \ldots, x_n)$ *представима в виде:*

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{q=2n+1} \chi_q \left[\sum_{p=1}^{n} \psi^{pq}(x_p) \right], \tag{1}$$

где функции $\chi_q(y)$ *действительны и непрерывны.*

Результат, показанный в этой работе, намного шире, чем общее решение 13 проблемы Гильберта на случай непрерывных функций (для алгебраических функций данная проблема остается нерешенной). Этой теоремой Колмогоров показал, что любую непрерывную функцию можно представить в виде суперпозиции функций одной переменной. Фактически данная теорема означает, что ИНС может аппроксимировать многомерную функцию.

Ниже представим ИНС, записанную по формуле (1).

Естественно, ИНС, построенная с помощью данной теоремы, не является общепринятой полносвязной ИНС, но именно эта теорема показывает возможность применения ИНС для решения задач аппроксимации. Также данная теорема неконструктивна, т. е. факт существования установлен, но вот алгоритм, который позволил бы найти вид функций χ_q и ψ^{pq}, не показан. Однако благодаря данной теореме уже можно говорить о таких характеристиках нейронной сети, как количество слоев и количество нейронов на слое.

Научное сообщество широко оценило эту работу, и уже спустя несколько лет появились исследования, дополняющие данную теорему. Так, в 1962 году Джордж Лоренц в математическом ежемесячнике опубликовал статью [3], в которой показал, что число внешних функций χ можно уменьшить до одной. Точнее, основываясь на теореме А.Г. Витушкина [4] о невозможности представления функций нескольких переменных суперпозициями функций меньшего числа переменных, показал, что существуют такие ψ^{pq}, при которых формулу (1) можно записать в следующем виде:

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{q=2n+1} \chi \left[\sum_{p=1}^{n} \psi^{pq}(x_p) \right]. \tag{2}$$

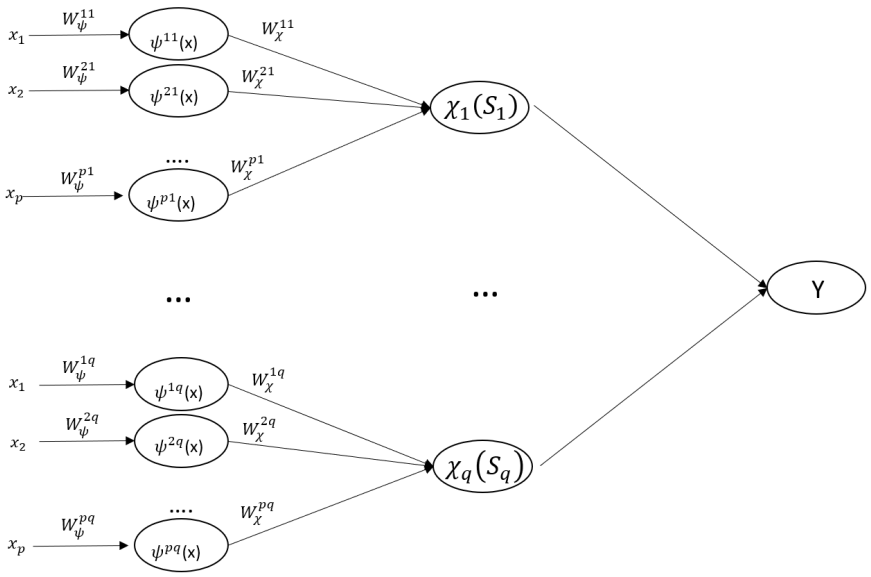

Рис. 1. *Схема ИНС, сконструированной на основе теоремы Колмогорова–Арнольда*

Но так же, как и в исходной работе Колмогорова, в работе Лоренца отсутствует конструктивная часть, т. е. неизвестен алгоритм построения (поиска) функций χ.

Дальнейшие работы показали возможность замены и внутренних функций на одну. Так, в [5] Д. Шпрехер представил теорему, которая содержит результат Колмогорова как частный случай.

Теорема Шпрехера

При любом целом $N \geq 2$ *существует вещественная, монотонно возрастающая функция,* $\psi(x) \in Lip[\frac{ln2}{ln(2N+2)}]$, $\psi\mathcal{E} = \mathcal{E}(\mathcal{E} -$ декартово произведение $\mathcal{E}^n = \prod_{1\leq p\leq n}\mathcal{E}_p)$, *зависящая от N и обладающая следующими свойствами: для каждого предварительно присвоенного номера* $\delta > 0$ *существует рациональное число* ε, $\varepsilon < 0 \leq \delta$, *такое, что для* $2 \leq n \leq N$ *каждая вещественная непрерывная функция от n переменных,* $f(x)$, *определенная на* $\mathcal{E}^n$, *имеет представление в виде:*

$$f(x_1,\ldots,x_n) = \sum_{q=1}^{q=2n+1} \chi\left[\sum_{p=1}^{n} \lambda^p \psi\left(x_p + \varepsilon q\right) + q\right], \tag{3}$$

где функция χ *действительна и непрерывна, а* λ *— константа, не зависящая от* f.

Т. е. фактически Шпрехер показал, что для аппроксимации функции нескольких переменных достаточно двухслойной ИНС с одной функцией активации на слой, n нейронов со смещением на 1 скрытом слое и 2n+1 нейронов со смещением на втором скрытом слое. Схема ИНС, построенной на основе данной теоремы, представлена на рисунке 2.

После был опубликован еще ряд исследований, в которых дорабатывали или критиковали теорему Колмогорова–Арнольда. Так, в работе [6] автор обобщил теорему Колмогорова–Арнольда на компактные метрические пространства. В [7] автор показывает, что в случаях сложных многовариантных функций данная теорема не выполняется. А в работе [8] обсуждают ограничения на практическое использование данной теоремы.

Наиболее интересную и значимую работу по исследованию данной теоремы провел в 1989 году Джордж Цыбенко (Кибенко) [9].

Прежде чем привести формулировку данной теоремы, необходимо дать некоторые определения.

Пусть I_n — n-мерный единичный куб, $C(I_n)$ — пространство непрерывных функций на I_n.

Пространство конечных знаковых регулярных борелевских мер на I_n *обозначается как* $M(I_n)$.

Определение 1. Говорят, что σ является дискриминационной функцией, если для меры $\mu \epsilon M(I_n)$

$$\int_{I_n} \sigma\left(y^T x + \theta\right) d\mu(x) = 0.$$

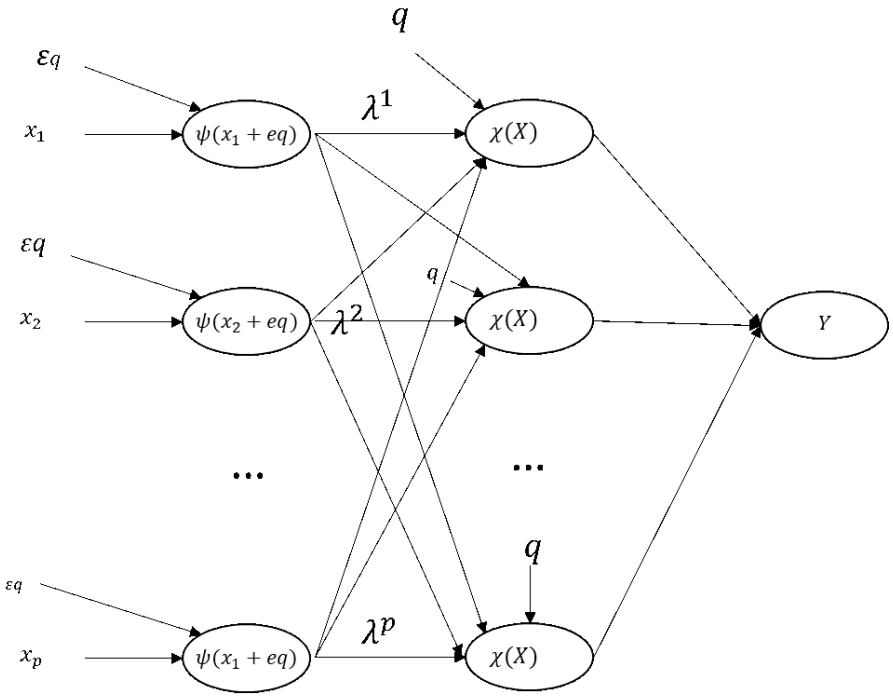

Рис. 2. *Схема ИНС, основанной на теореме Шпрехера*

Для всех $y \in R^n$ и $\theta \epsilon R$ подразумевается, что $\mu = 0$.

Определение 2. Говорят, что σ — сигмоидальная функция, если:

$$\sigma(t) \to \begin{cases} 1 \text{ при } t \to +\infty \\ 0 \text{ при } t \to -\infty \end{cases}.$$

Теорема Цыбенко

Пусть σ — произвольная непрерывная дискриминационная функция. Тогда конечная сумма вида

$$G\left(\overrightarrow{x}\right) = \sum_{j=1}^{N} a_j \sigma\left(y_j^T \overrightarrow{x} + \theta_j\right). \tag{4}$$

Плотная в $C(I_n)$. Другими словами: для $\forall\, f \in C(I_n)$ и $\varepsilon > 0$ $\exists \sum$ вида $G(x)$ (4), для которой

$$|G(x) - f(x)| < \varepsilon \quad \text{для } \forall\, x \in I_n.$$

Эта теорема в математической теории ИНС носит название универсальной теоремы аппроксимации. В настоящее время существуют различные результаты, сформированные в виде теорем, которые показывают возможность использования ИНС в качестве универсальной аппроксимации. Данные теоремы можно условно разделить на три класса:

1. Аппроксимация с произвольным числом нейронов на слое (в некоторых источниках — случай произвольной ширины).
2. Аппроксимация с произвольным числом скрытых слоев (в некоторых источниках — случай произвольной глубины).
3. Аппроксимация с ограниченным числом нейронов и количеством скрытых слоев (в некоторых источниках — случай ограниченной глубины и ширины).

Случай произвольной ширины (и ограниченной глубины)

Теоремы, связанные с произвольной шириной, описывают возможности ИНС аппроксимировать функции с произвольным числом нейронов, но накладывают некоторые ограничения на глубину. К работам, в которых исследуется случай произвольной ширины, можно отнести [9–12]. Самой значимой из этих работ является работа Цыбенко [9] и его теорема, описанная выше. Данную теорему

можно отнести к случаю с произвольной шириной, т. к. авторы показывают, что при ограниченном количестве скрытых слоев (ограничено 1 слоем) и неограниченном количестве нейронов в слое ИНС может аппроксимировать функцию с заранее заданной точностью. В качестве функции активации на скрытом слое следует использовать сигмоидальную функцию.

Причем, как следует из леммы, представленной Цыбенко в работе [9], дискриминационной функцией является любая ограниченная измеримая сигмоидальная функция, в частности, любая непрерывная сигмоидальная функция.

Доказательства приведенных выше и всех последующих теорем и лемм достаточно объемны для приведения в данной работе. Отметим только, что при доказательстве теоремы Цыбенко использовались стандартные методы из функционального анализа, а также теорема Хана—Банаха и теорема представлений Рисса.

Теорема Цыбенко не была первой, но именно ее подразумевают под универсальной теоремой аппроксимации. За 2 года до этого в работе [10] К.-И. Фунохаши доказал теоремы, в которых также в качестве функций активации предлагалось использовать сигмоидальные функции, но количество скрытых слоев, необходимых для аппроксимации, больше одного. Приведем формулировки теорем, которые доказал Фунохаши, используя Фурье-анализ и теорию Пэли—Винера.

Общее замечание при рассмотрении теорем Фунохаши:

Пусть точки n-мерного евклидова пространства R^n обозначаются через $\boldsymbol{x} = (x_1, \ldots, x_n)$ и норма $\boldsymbol{x}$ определяется как $|\boldsymbol{x}| = \left(\sum_{i=0}^{n} x_i^2\right)^{\frac{1}{2}}$.

Теорема Фунохаши I

Пусть функция $\phi(x)$ — непрерывная, монотонно возрастающая функция, не являющаяся константой. Пусть K — компактное подмножество (ограниченное замкнутое подмножество) из R^n и $f(x_1, \ldots, x_n)$ — действительная непрерывная функция на K. Тогда для произвольного $\varepsilon > 0$ существует целое число N и действительные константы c_i, θ_i $(i = 1, \ldots, N)$, $w_{ij} (i = 1, \ldots, N, j = 1, \ldots, n)$, такие, что

$$\overline{f}(x_1, \ldots, x_n) = \sum_{i=1}^{N} c_i \phi\left(\sum_{j=1}^{n} w_{ij} x_j - \theta_i\right) \tag{5}$$

удовлетворяет условию $max_{\boldsymbol{x} \in K} | f(x_1, \ldots, x_n) - C(x_1, \ldots, x_n)| < \varepsilon$.

Фактически данная теорема гласит, что любую непрерывную функцию на компактном подмножестве можно аппроксимировать с заранее заданной точностью ИНС с одним скрытым слоем, в которой функции активации на входных и выходных слоях являются линейными, а на скрытом слое используется функция вида $\phi(x)$. Из этой теоремы вытекает следующая теорема.

Теорема Фунохаши II

Пусть функция $\phi(x)$ — непрерывная, монотонно возрастающая функция, не являющаяся константой. Пусть K — компактное подмножество (ограниченное замкнутое подмножество) из R^n и зафиксируем целое число $k \geq 3$. Тогда любое непрерывное отображение $f : K \to R^m$, определяемое $\boldsymbol{x} = (x_1, \ldots, x_n) \to (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$, может быть аппроксимировано в смысле однородной топологии на K отображениями ИНС из k-уровней ($k - 2$ скрытых слоев), функции активации на скрытых слоях $\phi(x)$, на входном и выходном слое линейные функции активации.

Или, другими словами:

Для любого непрерывного отображения $f : k \to R^m$ и произвольного $\varepsilon > 0$ существует сеть k-го уровня, отображение входа — выхода которого задается как $f : K \to R^m$, таким, что $max_{\boldsymbol{x} \in K} d\left(f(\boldsymbol{x}), \overline{f}(\boldsymbol{x})\right) < \varepsilon$, где $d(,)$ — метрика, которая индуцирует обычную топологию в R^m.

Т. е. фактически данная теорема показывает, что с помощью ИНС можно аппроксимировать любое отображение с заранее заданной точностью. Доказательство данных теорем представлено в [10]. Глубину сетей автор ограничивает минимум двумя скрытыми слоями, а количество нейронов зависит от требуемой точности аппроксимации отображения. В качестве функции активации предлагается использовать сигмоидальные функции, т. к. они удовлетворяют ограничениям, наложенным на функции активации, кроме того, еще в модели Маккалока—Питса было показано, что они сходятся к пороговой функции, и с помощью разработанной ими модели можно спроектировать любую логическую схему.

Исходя из этого и на основании теоремы Фунохаши II можно сделать вывод, что любое непрерывное отображение можно представить многослойными ИНС с сигмоидальными функциями активаций.

Стоит также рассмотреть работу К. Хорника с соавторами [11], которая примечательна следующим. В ней исследуются сети под названием П-сигма, отличительной особенностью которых является то, что в них входы нейрона можно не только складывать, но и перемножать. Такой подход нужен для решения проблемы представления многочленов. Схема частного случая ИНС типа П-сигма представлена на рисунке 3.

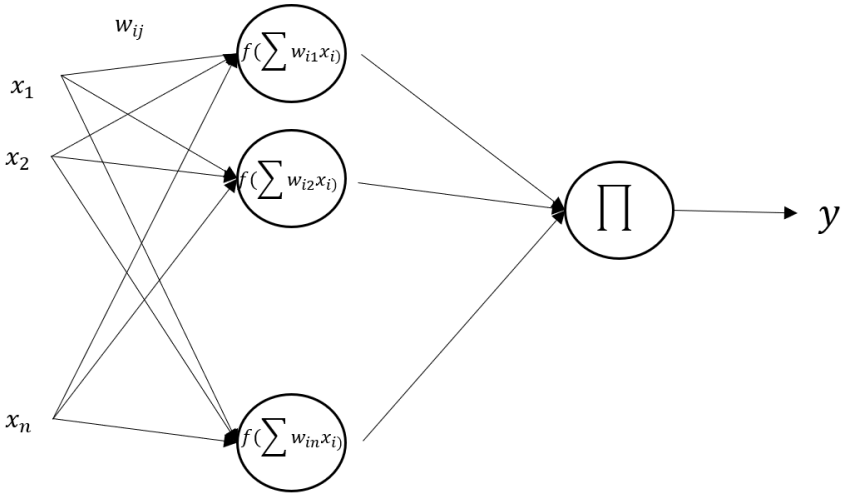

Рис. 3. *Схема ИНС типа П-сигма*

В математическом представлении такую ИНС можно записать в следующем виде [13]:

$$y = \sigma(\prod(\sum_j w_{ij}x_i + \theta_{ij})). \tag{6}$$

Хорником вместе с соавторами [11] были сформулированы и доказаны универсальные теоремы аппроксимации для сетей П-сигма. Прежде чем привести данные теоремы, необходимо ввести некоторые определения.

Определение 3. *Для любого* $r \in N \equiv \{1,2,\ \ldots\}$ A^r *— множество всех аффинных функций от* R^r до R, *т. е. множество всех функций вида* $A(x) = \boldsymbol{w} \cdot \boldsymbol{x} + b$, *где* $\boldsymbol{w}$ и $\boldsymbol{x}$ *векторы* в R^r. «·» *обозначает обычное скалярное произведение векторов,* $b \in R$ *— скаляр.*

Фактически в данном определении $\boldsymbol{w}$ — весовые коэффициенты, $\boldsymbol{x}$ — входные данные ИНС, b — смещение.

Следующие определение — это фактически математическая запись классической ИНС.

Определение 4. *Для любой (борелевской) измеримой функции* $G(\cdot)$, *отображающей* R в R, *и* $r \in N$ *пусть* $\sum_r(G)$ *класс функций:*

$$\{f : R^r \to R \ : f(x) = \sum_{i=1}^{q} B_i G(A_i(x)),\ \ x \in R^r,\ B_i \in R,\ A_i \in A^r,\ \ q = 1,2,\ldots\}.$$

Следующее определение необходимо для понимания того, какой класс функций можно использовать в качестве функции активации.

Определение 5. *Функция* $\psi : R \to [0,1]$ *— сжимающая функция, если она не убывающая, т. е.* $lim_{x\to+\infty}\psi(x) = 1$ *и* $lim_{x\to-\infty}\psi(x) = 0$.

Фактически данное определение повторяет **определение 2,** которое было введено для изложения теоремы Цыбенко. В качестве примера функции, удовлетворяющей **определению 5**, авторы используют пороговые функции, в то время как Цыбенко используются сигмоидальные функции, и, хотя данные функции на бесконечности ведут себя одинаково, стоит отметить, что графики данных

функций различны. Пороговые функции имеют точку разрыва первого рода и фактически принимают только два значения — 0 и 1.

Далее дадим формальное определение П-сигмы, которое используют авторы статьи.

Определение 6. *Для любой измеримой функции $G(\cdot)$, отображающей R в R, и $r \in N$ пусть $\sum\prod_r G$* класс функций:

$$\{f : R^r \to R \ : f(x) = \sum_{i=1}^{q} B_i \prod_{k=1}^{l_i} G(A_{ik}(x)),\ \ x \in R^r,\ B_i \in R,\ A_{ik} \in A^r,\ l_i \in N,\ \ q = 1,2,\dots\}.$$

Определение 7. *Пусть C^r — множество непрерывных функций от R^r* до *R и пусть M^r — множество борелевских измеримых функций от R^r* в *R. Обозначим борелевскую σ-алгебру на R^r как B^r.*

Определение 8. *Подмножество S-метрического пространства (X,p) является p-плотным в подмножестве T, если для каждого $\varepsilon > 0$ и для каждого $t \in T$ существует $s \in S$, такое, что $p(s,t) < \varepsilon$.*

Другими словами, элемент S может аппроксимировать элемент T с любой наперед заданной точностью.

В теоремах Хорника T и X соответствуют C^r и M^r, а S — это класс функций $\sum\prod_r G$ или $\sum_r(G)$ для конкретных G, p выбирается соответсвующим образом, т. е. фактически это метрика для оценки качества обучения ИНС.

Определение 9. *Подмножество $S \in C^r$ называется равномерно плотным на компактах в C^r, если для каждого компактного подмножества $K \subset R^r$ S является p_k-плотным в C^r, где для $f,g \in C^r$ $p_k(f,g) \equiv sup_{x\in K}|f(x) - g(x)|$. Последовательность функций $\{f_n\}$ сходится к функции f равномерно на компактах, если для всех компактов $K \subset R^r$ $p_k(f_n,f) \to 0$ как $n \to \infty$.*

Далее приведем 1 теорему Хорника для сетей П-сигма.

Теорема Хорника I для сетей П-сигма

Пусть G — любая непрерывная непостоянная функция от R^r в *R. Тогда $\sum\prod_r(G)$ равномерно плотная на компактах C^r.*

Или, другими словами, *сети П-сигма способны сколь угодно точно аппроксимировать любую вещественнозначную непрерывную функцию на компактном множестве*.

Требование компактности множества выполняется всякий раз, когда возможные значения входных данных x ограничены. Кроме того, функции активации в таких сетях могут быть любой непрерывной непостоянной функцией.

Для того чтобы представить еще один важный результат из статьи, необходимо ввести следующие два определения.

Определение 10. *Пусть μ–мера вероятности на (R^r,B^r). Если $f \in M^r$* и *$g \in M^r$, говорят, что они μ* – эквиваленты, *если $\mu\{x \in R^r : f(x) = g(x)\} = 1$.*

Меру вероятности авторы вводят для удобства, и, строго говоря, их результаты справедливы для любых произвольных конечных мер.

Метрику для классов μ – эквивалентных функций авторы задают следующим образом.

Определение 11. *Учитывая меру вероятности μ на (R^r,B^r), определим метрику p_μ от $M^r \times M^r$ до R с помощью $p_\mu(f,g) = inf$*

$$\{\varepsilon > 0 \ : \ \mu\{x : |f(x) - g(x)| > \varepsilon\} < \varepsilon\}.$$

Т. е. две функции близки в этой метрике тогда и только тогда, когда существует лишь небольшая вероятность того, что они существенно различаются.

Задав метрику, можно приводить следующие теоремы, которые являются главным результатом работы Хорника.

Теорема Хорника II для сетей П-сигма

Для каждой непрерывной непостоянной функции G любой r и любой меры вероятности μ на (R', B'), $\sum\prod_r(G)$ p_μ–плотно в M^r.

Другими словами, данную теорему можно пояснить следующим образом.

Сети прямого доступа с одним скрытым слоем $\sum\prod$ могут сколь угодно хорошо аппроксимировать любую измеримую функцию, независимо от используемой непрерывной непостоянной функции G, независимо от размерности входного пространства r и независимо от среды входного пространства μ.

Именно в этом точном смысле авторы считают, что сети П-сигма являются универсальными аппроксиматорами.

Теорема Хорника III для сетей П-сигма

Для каждой сжимающей функции ψ, каждого r и каждой меры вероятности μ на (R^r, B^r), $\sum\prod_r(\psi)$ равномерно плотен на компактах в C^r и p_μ плотен в M'.

Фактически в данной теореме говорится, что в качестве функции активации должна использоваться сжимающаяся функция или, что одно и то же, сигмоидальная функция, определенная Цыбенко (см. определение 5).

Аналогичную теорему авторы приводят для классических ИНС с сумматором.

Теорема Хорника IV для сетей П-сигма

Для каждой сжимающей функции ψ, каждого r и каждой меры вероятности μ на (R^r, B^r) $\sum_r(\psi)$ равномерно плотен на компактах в C^r и p_μ плотен в M'.

Т. е. фактически классические ИНС прямого распространения, имея только один скрытый слой, могут равномерно аппроксимировать любую непрерывную функцию на любом компактном множестве и любую измеримую функцию сколь угодно хорошо в метрике p_μ, независимо от сжимающей функции ψ (непрерывной или нет), независимо от размерности входного пространства r и независимо от среды входного пространства μ.

Доказательство своих теорем Хорник с соавторами основывает на теореме Стоуна—Вейерштрасса, в которой говорится о возможности представления любой непрерывной функции на хаусдорфовом пространстве в виде предела равномерно сходящейся последовательности функций особого класса — алгебры Стоуна.

Данные работы являются основными для случая произвольной ширины.

Случай произвольной глубины (и ограниченной ширины)

Предыдущие теоремы показывают возможность применения ИНС с произвольной шириной и ограниченной глубиной, т. е. данные теоремы налагали ограничения на количество слоев, но не делали ограничений для количества нейронов. И фактически количество нейронов могло расти экспоненциально по отношению к размерности входных данных. А следовательно, наличие только одной возможности аппроксимировать функции с помощью ИНС не достаточно для их применения, т. к. сети с большим количеством нейронов в слое требуют дополнительных вычислительных затрат как при обучении, так и для их последующего применения. Поэтому начиная с недавнего времени исследователи ИНС обратили внимание на то, возможно ли заменить сети произвольной ширины с ограниченным числом слоев на сети с произвольной глубиной и ограниченной шириной. Но перед этим ряд ученых задался вопросом о том, существует ли сеть глубины n, которую нельзя представить сетью меньшей глубины. Так, в серии статей [15, 16] этот вопрос исследовался и на него был дан положительный ответ, но, строго говоря, сформулированные в данных работах теоремы неконструктивны.

Рассмотрим работу, в которой авторы приводят универсальную теорему аппроксимации для сетей ReLU с ограниченной шириной [14].

Теорема универсальной аппроксимации сетей с ограниченной шириной и функцией активации ReLU

Для любой интегрируемой по Лебегу функции $f : R^n \to R$ и любого $\varepsilon > 0$ существуют полносвязные сети $\mathcal{A}$ с функцией активации ReLU с шириной (количество нейронов на слое) $d_m \leq n + 4$, такие, что функция $F_{\mathcal{A}}$,представленная этой сетью, удовлетворяет:

$$\int_{R^n} |f(x) - F_{\mathcal{A}}(x)|\, dx < \varepsilon.$$

Авторы не приводят доказательств данной теоремы. Вместо этого они явно строят ИНС, описанную в теореме, так, чтобы она могла аппроксимировать функцию с заданной точностью. Конструируемая ими ИНС состоит из многих блоков, каждый из которых удовлетворяет следующим условиям:

1) ИНС с глубиной (4n+1) и шириной (n+4) с функциями активации ReLU;
2) он может с высокой точностью аппроксимировать любую интегрируемую по Лебегу функцию, которая равномерно равна нулю вне куба с длиной δ;
3) он может сохранять выходные данные предыдущего блока, т. е. аппроксимацию других интегрируемых функций Лебега на разных кубах;
4) он может суммировать свое текущее приближение и память предыдущих приближений.

Один из таких блоков представлен на рисунке 4.

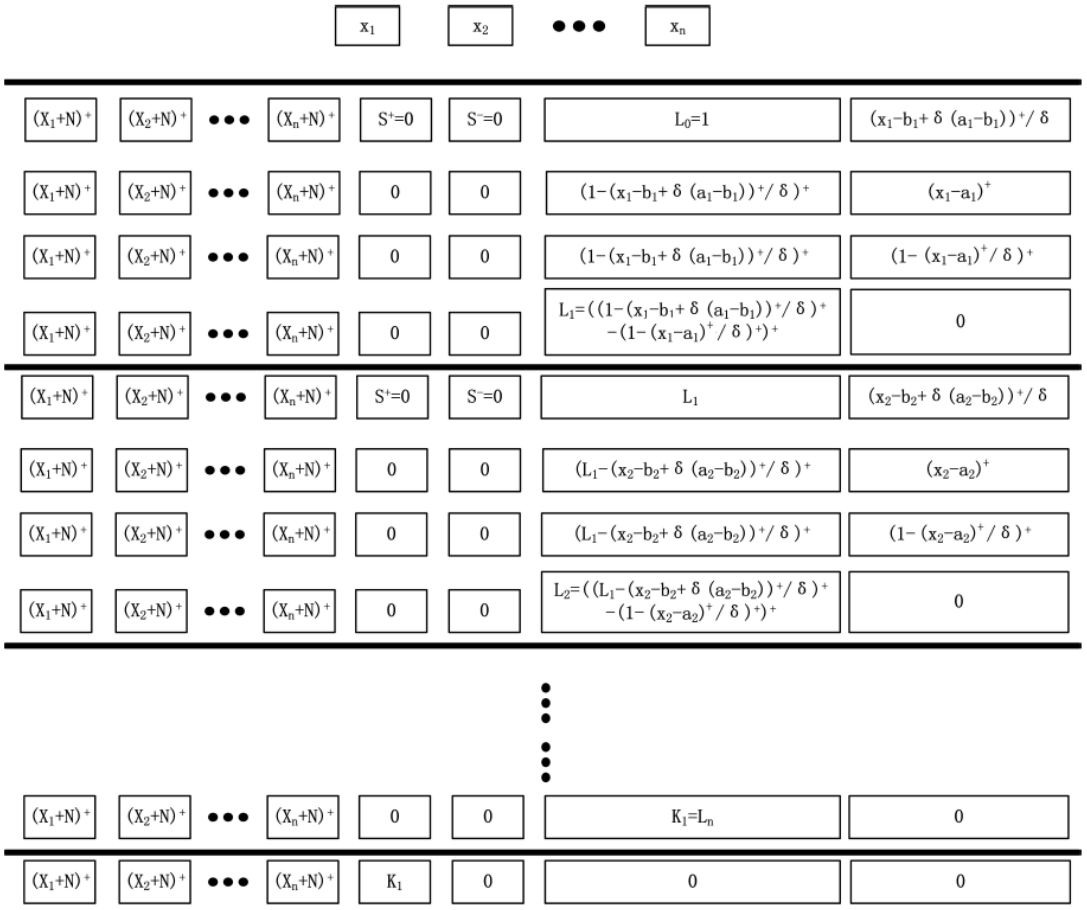

Рис. 4. *Один блок ИНС, аппроксимирующей любую интегрируемую функцию по Лебегу [14]*

В этом блоке каждый слой содержит n + 4 нейрона. Каждый прямоугольник на рисунке 4 представляет собой нейрон, а символы в прямоугольнике описывают выходные данные этого нейрона как функцию блока. Среди n + 4 нейронов n нейронов просто передают входные координаты. Для остальных 4 нейронов 2 нейрона сохраняют приближение, выполненное предыдущими блоками. Остальные 2 нейрона помогают выполнить аппроксимацию на текущем кубе. Топология блока довольно проста. Он очень разрежен, каждый нейрон соединяется максимум с 2 нейронами в следующем слое.

В этой же работе авторы приводят теорему, в которой показывают, что, вообще говоря, при уменьшении числа нейронов аппроксимирующая способность ИНС аппроксимировать интегрируемую по Лебегу функцию теряется. Формулировка теоремы представлена ниже.

Теорема о невозможности аппроксимации ИНС с шириной n функции, интегрируемой по Лебегу

Для любой интегрируемой по Лебегу функции $f : R^n \to R$ *, удовлетворяющей тому, что* $\{x : f(x) \neq 0\}$ *является положительной мерой, заданной в мере Лебега, и любая функция* $F_{\mathcal{A}}$*, представляющаяся полносвязной сетью* $\mathcal{A}$ *с функцией активации ReLU с шириной (количество нейронов на слое)* $d_m \leq n$*, выполняется следующее условие:*

$$\int_{R^n} |f(x) - F_{\mathcal{A}}(x)|\, dx = +\infty.$$

Если ограничить функцию на ограниченном множестве, то, как показывают авторы статьи, справедлива следующая теорема.

Теорема об аппроксимации функции ИНС с шириной n

Для любой непрерывной функции $f : [-1,1]^n \to R$, *которая не является постоянной ни в одном направлении, существует универсальное значение* $\epsilon^* > 0$, *такое, что для любой функции* F_A, *являющейся полносвязной ИНС с функцией активации ReLU с шириной (количество нейронов на слое)* $d_m \leq n-1$, *расстояние* L^1 *между* f и F_A *равно по меньшей мере:*

$$\int_{[-1,1]^n} |f(x) - F_A(x)|\, dx \geq \epsilon^*.$$

Основным результатом работы [14] является следующая теорема.

Теорема. *Пусть* n — *входное измерение. Для любого целого числа* $k \geq n+4$ *существует* $F_{\mathcal{A}} : R^n \to R$, *являющаяся нейронной сетью* $\mathcal{A}$ *с функцией активации ReLU с шириной (количество нейронов в слое)* $d_m = 2k^2$ *и глубиной (количество скрытых слоев)* $h = 3$, *такой, что для любой константы* $b > 0$ *существует* $\epsilon > 0$ *и для любой функции* $F_{\mathfrak{B}} : R^n \to R$, *являющейся нейронной сетью* $\mathfrak{B}$, *параметры которой ограничены в* $[-b,b]$, *с шириной (количество нейронов в слое)* $d_m = k^{3/2}$ *и глубиной* $h \leq k+2$ *(количество скрытых слоев), выполняется следующее неравенство:*

$$\int_{R^n} (F_{\mathcal{A}} - F_{\mathfrak{B}})^2 dx \geq \epsilon.$$

Фактически в данной теореме говорится о том, что существуют сети, в которых уменьшение ширины требует увеличения глубины.

В другой работе [17] был представлен первый результат, полностью характеризующий минимальную ширину сетей ReLU для универсальной аппроксимации, в частности, авторы доказывают теорему, в которой верхняя граница минимальной ширины нейронной сети существенно меньше, чем в других работах похожей тематики, к примеру, другой известный результат представлен в работе [19].

Теорема об аппроксимации L^p функций с помощью ИНС с функцией активации ReLU

Для любого $p \in (1,\infty)$ *нейронные сети с функцией активации ReLU шириной* w *можно аппроксимировать функцию из пространства* $L^p(R^{dx},R^{dy})$ *с любой требуемой точностью тогда и только тогда, когда* $w \geq max(dx+1,dy)$, *где* dx — *входная размерность нейронной сети,* dy — *выходная размерность.*

В той же работе показывается, что, строго говоря, для $C(\mathcal{K},\mathbb{R}^{d_y})$ на компакте $\mathcal{K} \subset \mathbb{R}^{d_x}$ данная теорема не выполняется для $dx = 1$, $dy = 2$, и в таком случае справедлива следующая теорема.

Теорема. *ИНС с функцией активации ReLU и шириной* w *может аппроксимировать функцию в пространстве* $C\left([0,1],\mathbb{R}^2\right)$ *с любой требуемой точностью тогда и только тогда, когда* $w \geq 3$.

Предыдущие теоремы показывают, что для сетей с функцией активации ReLU, чтобы аппроксимировать функции из $C(\mathcal{K},\mathbb{R}^{d_y})$, требуется большее количество нейронов в слое, чем для аппроксимации функции $L^p(R^{dx},R^{dy})$. Но авторы в той же работе показывают, что, если добавить ступенчатую функцию активации (функция Хевисайда), это позволит уменьшить ширину.

Теорема. *ИНС с функцией активации ReLU и функцией активации Хевисайда и шириной* w *может аппроксимировать функцию из* $C(\mathcal{K},\mathbb{R}^{d_y})$ *тогда и только тогда, когда* $w \geq max(d_x+1,d_y)$.

Предыдущие теоремы показывают, что минимальная ширина универсальных аппроксимаций зависит от выбора функций активации, что противоречит классическим результатам [20], где показывают, что сети с функцией активации ReLU и двумя слоями являются универсальным аппроксиматорами.

Кроме того, авторы доказывают теорему, в которой они значительно улучшают результат, представленный в работе [19], и показывают верхнюю границу.

Теорема. *Пусть* $p:\mathbb{R} \to \mathbb{R}$ — *любая непрерывная не полиноминальная функция, непрерывно дифференцируемая по крайне мере в одной точке, с ненулевой производной в этой точке. Тогда ИНС с функцией активации* p *с шириной* w *аппроксимирует с любой требуемой точностью функцию из* $L^p(\mathcal{K},\mathbb{R}^{dy})$ *для всех* $p \in [1,\infty)$, *если* $w \geq max(d_x+2,d_y+1)$.

Представленные выше теоремы на данный момент являются лучшим, но не единственным результатом, в котором показаны границы на ширину нейронных сетей. Часть результатов, которые могут

быть интересны при исследовании универсальных теорем аппроксимации, представлена в таблице ниже. Указанные там работы рассматривать отдельно не будем, представим только основные результаты, которые сформированы в работах как теоремы и доказаны авторами соответствующих работ.

Таблица

Верхние и нижние границы на ширину нейронной сети для аппроксимации различных классов функций

Ссылка на работу	Класс функций	Функция активации	Верхняя/нижняя граница
14	$L^1(\mathbb{R}^{d_x},\mathbb{R})$	ReLU	$d_x+1 \le w_{min} \le d_x+4$
14	$L^1(\mathcal{K},\mathbb{R})$	ReLU	$w_{min} \ge d_x$
21	$C(\mathcal{K},\mathbb{R}^{d_y})$	ReLU	$d_x+1 \le w_{min} \le d_x+d_y$
22	$C(\mathcal{K},\mathbb{R})$	Равномерная, непрерывная функция*	$w_{min} \ge d_x+1$
23	$C(\mathcal{K},\mathbb{R}^{d_y})$	Непрерывная не полиноминальная функция**	$w_{min} \le d_x+d_y+1$
23	$C(\mathcal{K},\mathbb{R}^{d_y})$	Не аффинный полином	$w_{min} \le d_x+d_y+1$
23	$L^p(\mathbb{R}^{dx},\mathbb{R}^{dy})$	ReLU	$w_{min} \le d_x+d_y+1$
18	$L^p(\mathbb{R}^{dx},\mathbb{R}^{dy})$	ReLU	$w_{min} = \max(d_x+1,d_y)$
18	$C\left([0,1],\mathbb{R}^2\right)$	ReLU	$w_{min} = 3 > \max(d_x+1,d_y)$
18	$C\left([0,1],\mathbb{R}^2\right)$	ReLU + функция Хевисайда	$w_{min} = \max(d_x+1,d_y)$
18	$L^p(\mathcal{K},\mathbb{R}^{dy})$	непрерывная не полиноминальная функция**	$w_{min} \le \max(d_x+2,d_y+1)$

* Требует, чтобы ρ равномерно аппроксимировалось последовательностью взаимно однозначных функций.

** Требует, чтобы ρ было непрерывно дифференцируемо по крайней мере в одной точке (скажем, z), причем $p'(z) \neq 0$.

Случай ограниченной глубины и ограниченной ширины

Далее рассмотрим работы, в которых авторы исследовали возможность определения точных размеров нейронной сети не только на глубину или на ширину по отдельности, а универсальную теорему аппроксимации, в которых установлены ограничения одновременно и на ширину, и на глубину.

Первый значимый результат был получен в 1999 году в работе [24], в которой показано существование функции активации, при которой существует нейронная сеть с ограничениями на ширину и глубину.

Теорема универсальной аппроксимации с ограничением по ширине и глубине

Существует функция активации p, которая является аналитической, строго возрастающей, сигмоидальной и обладает следующим свойством: для любого $f \in C[0,1]^d$ *и* $\varepsilon > 0$ *существуют константы* d_i , c_{ij} , θ_{ij}, γ_i *и векторы* $\boldsymbol{w}^{ij} \in \mathbb{R}^d$, *для которых:*

$$\left| f(\boldsymbol{x}) - \sum_{i=1}^{6d+3} d_i p\left(\sum_{j=1}^{3d} c_{ij} p\left(\boldsymbol{w}^{ij} \cdot \boldsymbol{x} - \theta_{ij}\right) - \gamma_i \right) \right| < \varepsilon.$$

Для всех $\boldsymbol{x} = (x_1,\ldots,x_d) \in [0,1]^d$.

Хоть данный результат и является теоретическим, спустя почти 20 лет в работах [25, 26] был представлен алгоритм, который позволяет строить такие функции активации. В данных работах авторы формулируют теорему, и в качестве ее доказательства приводят построения алгоритма, который позволяет построить функцию активации. Кроме того, авторы дают ссылку на исходный код программы, который позволяет строить такие функции активации. Приведем формулировку теоретического результата.

Теорема о существовании алгоритма построения ИНС с ограничением по ширине и глубине

Пусть $[a,b] \in \mathbb{R}$, $s = b - a$ *и* $\lambda-$ *любое достаточно малое положительное действительное число. Тогда можно алгоритмически построить вычислимую, бесконечно дифференцируемую, сигмоидальную функцию активации* $p:\mathbb{R} \to \mathbb{R}$, *строго возрастающую на* $(-\infty,s)$, λ *строго возрастает на* $[s, +\infty)$ *и удовлетворяет следующим свойствам:*

для любой непрерывной функции f на d-мерной квадратной области $[a,b]^d$ *и* $\varepsilon > 0$ *существуют константы* e_i , c_{ij} , θ_{ij}, γ_i , *такие, что неравенство:*

$$\left| f(\boldsymbol{x}) - \sum_{i=1}^{2d+2} e_i p\left(\sum_{j=1}^{d} c_{ij} p\left(\boldsymbol{w}^{ij} \cdot \boldsymbol{x} - \theta_{ij}\right) - \gamma_i \right) \right| < \varepsilon$$

справедливо для всех $\boldsymbol{x} = (x_1, \ldots, x_d) \in [a,b]^d$. *Веса* w^j, $j = 1, \ldots, d$ *фиксированы следующим образом:*

$$w^1 = (1,0,\ldots 0)\,,\ w^2 = (0,1,\ldots,0)\,,\ \ldots,\ w^d = (0,0,\ldots,1).$$

Кроме того, все коэффициенты e_i, *кроме одного, равны.*

Т. е. фактически данная теорема говорит о том, что нейронная сеть с глубиной 2 и шириной 2 является универсальным аппроксиматором для одномерных функций, а также нейронная сеть с глубиной 3 и шириной (2d+2) является универсальным аппроксиматором для функций d-переменных при условии правильного выбора функции активации. Алгоритм построения функций активации, при которых указанные нейронные сети являются универсальными аппроксиматорами, описан в [25, 26].

В настоящий момент для различных классов функций создана большая теоретическая база для их аппроксимации с помощью ИНС. Из существующих теорий ясно, что сконструировать одну единственную нейронную сеть, которая будет аппроксимировать любую функцию только лишь путем переопределения весовых коэффициентов, не получится. С другой стороны, уже существует ряд теоретических обоснований, которые позволяют конструировать ИНС не на основе эмпирических правил, а на основе теорем, в которых доказана максимальная размерность сети.

ЛИТЕРАТУРА

1. Бетелин В. Б. О проблеме доверия к технологиям искусственного интеллекта. *Успехи кибернетики*. 2021;2(3):6—7. DOI: 10.51790/2712-9942-2021-2-3-1.
2. Колмогоров А. Н. О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения. *Докл. АН СССР.* 1957;114(5):953—956.
3. Lorentz G. G. Metric Entropy, Widths, and Superpositions of Functions. *American Mathematical Monthly.* 1962;69(6):469—485. DOI: 10.1080/00029890.1962.11989915.
4. Колмогоров А. Н., Тихомиров В. М. ε-энтропия и ε-емкость множеств в функциональных пространствах. *Успехи мат. наук.* 1959;14(2):3—86.
5. Sprecher D. On the Structure of Continuous Functions of Several Variables. *Transactions of the American Mathematical Society*. 1965;115(3):340—355. DOI: 10.2307/1994273.
6. Ostrand P. A. Dimension of Metric Spaces and Hilbert's Problem 13. *Bulletin of the American Mathematical Society*. 1965;71(4):619—623. DOI: 10.1090/s0002-9904-1965-11363-5.
7. Akashi S. Application of ϵ-entropy Theory to Kolmogorov—Arnold Representation Theorem. *Reports on Mathematical Physics*. 2001;48(1—2):19—26. DOI: 10.1016/s0034-4877(01)80060-4.
8. Girosi F., Poggio T. Representation Properties of Networks: Kolmogorov's Theorem is Irrelevant. *Neural Computation*. 1989;1(4):465—469. DOI: 10.1162/neco.1989.1.4.465.
9. Cybenko G. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*. 1989;2(4):303—314. CiteSeerX: 10.1.1.441.7873. DOI: 10.1007/BF02551274.
10. Funahashi K.-I. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks*. 1989;2(3):183—192. DOI: 10.1016/0893-6080(89)90003-8.
11. Hornik K., Stinchcombe M., White H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*. 1989;2(5):359—366. DOI: 10.1016/0893-6080(89)90020-8.

12. Hornik K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*. 1991;4(2):251–257. DOI: 10.1016/0893-6080(91)90009-T.
13. Husaini N. A., Ghazali R., Nazri M. N., Lokman H. I., Mustafa M. D., Tutut H. Pi-Sigma Neural Network for a One-Step-Ahead Temperature Forecasting. *International Journal of Computational Intelligence and Applications.* 2014;13(4):1450023. DOI: 10.1142/S1469026814500230.
14. Lu Z., Pu H., Wang F., Hu Z., Wang L. The Expressive Power of Neural Networks: A View from the Width. Режим доступа: https://doi.org/10.48550/arXiv.1709.02540.
15. Eldan R., Shamir O. The Power of Depth for Feedforward Neural Networks. *Proceedings of Machine Learning Research.* 2016;49:907–940.
16. Cohen N., Sharir O., Shashua A. On the Expressive Power of Deep Learning: A Tensor Analysis. *Proceedings of Machine Learning Research.* 2016;49:698–728.
17. Telgarsky M. Benefits of Depth in Neural Networks. *Proceedings of Machine Learning Research.* 2016;49:1517–1539.
18. Park S., Yun C., Lee J., Shin J. Minimum Width for Universal Approximation. Режим доступа: https://arxiv.org/abs/2006.08859.
19. Kidger P., Lyons T. Universal Approximation with Deep Narrow Networks. *Proceedings of Machine Learning Research.* 2020;125:2306–2327.
20. Leshno M., Ya Lin V., Pinkus A., Schocken S. Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function. *Neural Networks*. 1993;6(6):861–867.
21. Hanin B., Sellke M. *Approximating Continuous Functions by ReLU Nets of Minimal Width*. Режим доступа: https://arxiv.org/abs/1710.11278.
22. Johnson J. *Deep, Skinny Neural Networks are not Universal Approximators*. Режим доступа: https://arxiv.org/abs/1810.00393.
23. Kidger P., Lyons T. *Universal Approximation with Deep Narrow Networks*. Режим доступа: https://arxiv.org/abs/1905.08539.
24. Maiorov V., Pinkus A. Lower Bounds for Approximation by MLP Neural Networks. *Neurocomputing*. 1999;25(1–3):81–91. DOI: 10.1016/S0925-2312(98)00111-8.
25. Guliyev N., Ismailov V. Approximation Capability of Two Hidden Layer Feedforward Neural Networks with Fixed Weights. *Neurocomputing*. 2018;316:262–269.
26. Guliyev N., Ismailov V. On the Approximation by Single Hidden Layer Feedforward Neural Networks with Fixed Weights. *Neural Networks.* 2018;98:296–304.