Universal approximation theorem

In the field of machine learning, the **universal approximation theorems** state that neural networks with a certain structure can, in principle, approximate any continuous function to any desired degree of accuracy. These theorems provide a mathematical justification for using neural networks, assuring researchers that a sufficiently large or deep network can model the complex, non-linear relationships often found in real-world data.[1][2]

The best-known version of the theorem applies to feedforward networks with a single hidden layer. It states that if the layer's activation function is non-polynomial (which is true for common choices like the sigmoid function or ReLU), then the network can act as a "universal approximator." Universality is achieved by increasing the number of neurons in the hidden layer, making the network "wider." Other versions of the theorem show that universality can also be achieved by keeping the network's width fixed but increasing its number of layers, making it "deeper."

It is important to note that these are existence theorems. They guarantee that a network with the right structure *exists*, but they do not provide a method for finding the network's parameters (training it), nor do they specify exactly how large the network must be for a given function. Finding a suitable network remains a practical challenge that is typically addressed with optimization algorithms like backpropagation.

Setup

Artificial neural networks are combinations of multiple simple mathematical functions that implement more complicated functions from (typically) real-valued vectors to real-valued vectors. The spaces of multivariate functions that can be implemented by a network are determined by the structure of the network, the set of simple functions, and its multiplicative parameters. A great deal of theoretical work has gone into characterizing these function spaces.

Most universal approximation theorems are in one of two classes. The first quantifies the approximation capabilities of neural networks with an arbitrary number of artificial neurons ("*arbitrary width*" case) and the second focuses on the case with an arbitrary number of hidden layers, each containing a limited number of artificial neurons ("*arbitrary depth*" case). In addition to these two classes, there are also universal approximation theorems for neural networks with bounded number of hidden layers and a limited number of neurons in each layer ("*bounded depth and bounded width*" case).

History

Arbitrary width

The first examples were the *arbitrary width* case. George Cybenko in 1989 proved it for sigmoid activation functions.[3] Kurt Hornik, Maxwell Stinchcombe, and Halbert White showed in 1989 that multilayer feed-forward networks with as few as one hidden layer are universal approximators.[1] Hornik also showed in 1991[4] that it is not the specific choice of the activation function but rather the multilayer feed-forward architecture itself that gives neural networks the potential of being universal approximators. Moshe Leshno *et al* in 1993[5] and later Allan Pinkus in 1999[6] showed that the universal approximation property is equivalent to having a nonpolynomial activation function.

Arbitrary depth

The *arbitrary depth* case was also studied by a number of authors such as Gustaf Gripenberg in 2003,[7] Dmitry Yarotsky,[8] Zhou Lu *et al* in 2017,[9] Boris Hanin and Mark Sellke in 2018[10] who focused on neural networks with ReLU activation function. In 2020, Patrick Kidger and Terry Lyons[11] extended those results to neural networks with *general activation functions* such, e.g. tanh or GeLU.

One special case of arbitrary depth is that each composition component comes from a finite set of mappings. In 2024, Cai [12] constructed a finite set of mappings, named a vocabulary, such that any continuous function can be approximated by compositing a sequence from the vocabulary. This is similar to the concept of compositionality in linguistics, which is the idea that a finite vocabulary of basic elements can be combined via grammar to express an infinite range of meanings.

Bounded depth and bounded width

The bounded depth and bounded width case was first studied by Maiorov and Pinkus in 1999.[13] They showed that there exists an analytic sigmoidal activation function such that two hidden layer neural networks with bounded number of units in hidden layers are universal approximators.

In 2018, Guliyev and Ismailov[14] constructed a smooth sigmoidal activation function providing universal approximation property for two hidden layer feedforward neural networks with less units in hidden layers. In 2018, they also constructed[15] single hidden layer networks with bounded width that are still universal approximators for univariate functions. However, this does not apply for multivariable functions.

In 2022, Shen *et al.*[16] obtained precise quantitative information on the depth and width required to approximate a target function by deep and wide ReLU neural networks.

Quantitative bounds

The question of minimal possible width for universality was first studied in 2021, Park et al obtained the minimum width required for the universal approximation of L^p functions using feed-forward neural networks with ReLU as activation functions.[17] Similar results that can be directly applied to residual neural networks

were also obtained in the same year by Paulo Tabuada and Bahman Gharesifard using control-theoretic arguments.[18][19] In 2023, Cai obtained the optimal minimum width bound for the universal approximation.[20]

For the arbitrary depth case, Leonie Papon and Anastasis Kratsios derived explicit depth estimates depending on the regularity of the target function and of the activation function.[21]

Kolmogorov network

The Kolmogorov–Arnold representation theorem is similar in spirit. Indeed, certain neural network families can directly apply the Kolmogorov–Arnold theorem to yield a universal approximation theorem. Robert Hecht-Nielsen showed that a three-layer neural network can approximate any continuous multivariate function.[22] This was extended to the discontinuous case by Vugar Ismailov.[23] In 2024, Ziming Liu and co-authors showed a practical application.[24]

Reservoir computing and quantum reservoir computing

In reservoir computing a sparse recurrent neural network with fixed weights equipped of fading memory and echo state property is followed by a trainable output layer. Its universality has been demonstrated separately for what concerns networks of rate neurons [25] and spiking neurons, respectively. [26] In 2024, the framework has been generalized and extended to quantum reservoirs where the reservoir is based on qubits defined over Hilbert spaces. [27]

Variants

Discontinuous activation functions,[5] noncompact domains,[11][28] certifiable networks,[29] random neural networks,[30] and alternative network architectures and topologies.[11][31]

The universal approximation property of width-bounded networks has been studied as a *dual* of classical universal approximation results on depth-bounded networks. For input dimension d_x and output dimension d_y the minimum width required for the universal approximation of the L^p functions is exactly $max\{d_x + 1, d_y\}$ (for a ReLU network). More generally this also holds if *both* ReLU and a threshold activation function are used.[17]

Universal function approximation on graphs (or rather on graph isomorphism classes) by popular graph convolutional neural networks (GCNs or GNNs) can be made as discriminative as the Weisfeiler–Leman graph isomorphism test.[32] In 2020,[33] a universal approximation theorem result was established by Brüel-Gabrielsson, showing that graph representation with certain injective properties is sufficient for universal function approximation on bounded graphs and restricted universal function approximation on unbounded graphs, with an accompanying $\mathcal{O}(|V| \cdot |E|)$-runtime method that performed at state of the art on a collection of benchmarks (where V and E are the sets of nodes and edges of the graph respectively).

There are also a variety of results between non-Euclidean spaces[34] and other commonly used architectures and, more generally, algorithmically generated sets of functions, such as the convolutional neural network (CNN) architecture,[35][36] radial basis functions,[37] or neural networks with specific properties.[38][39]

Arbitrary-width case

A universal approximation theorem formally states that a family of neural network functions is a dense set within a larger space of functions they are intended to approximate. In more direct terms, for any function f from a given function space, there exists a sequence of neural networks $\phi_1, \phi_2, \dots$ from the family, such that $\phi_n \to f$ according to some criterion.[3][1]

A spate of papers in the 1980s—1990s, from George Cybenko and Kurt Hornik etc, established several universal approximation theorems for arbitrary width and bounded depth.[40][1][3][4] See[41][42][6] for reviews. The following is the most often quoted:

> **Universal approximation theorem**—Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.
>
> Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m \times k}$ such that
>
> $$\sup_{x \in K} \|f(x) - g(x)\| < \varepsilon$$
>
> where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$

Also, certain non-continuous activation functions can be used to approximate a sigmoid function, which then allows the above theorem to apply to those functions. For example, the step function works. In particular, this shows that a perceptron network with a single infinitely wide hidden layer can approximate arbitrary functions.

Such an f can also be approximated by a network of greater depth by using the same construction for the first layer and approximating the identity function with later layers.

> **Proof sketch**
>
> It suffices to prove the case where $m = 1$, since uniform convergence in $\mathbb{R}^m$ is just uniform convergence in each coordinate.
>
> Let F_σ be the set of all one-hidden-layer neural networks constructed with σ. Let $C_0(\mathbb{R}^d, \mathbb{R})$ be the set of all $C(\mathbb{R}^d, \mathbb{R})$ with compact support.
>
> If the function is a polynomial of degree d, then F_σ is contained in the closed subspace of all polynomials of degree d, so its closure is also contained in it, which is not all of $C_0(\mathbb{R}^d, \mathbb{R})$.
>
> Otherwise, we show that F_σ's closure is all of $C_0(\mathbb{R}^d, \mathbb{R})$. Suppose we can construct arbitrarily good approximations of the ramp function
>
> $$r(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } |x| \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

then it can be combined to construct arbitrary compactly-supported continuous function to arbitrary precision. It remains to approximate the ramp function.

Any of the commonly used activation functions used in machine learning can obviously be used to approximate the ramp function, or first approximate the ReLU, then the ramp function.

if σ is "squashing", that is, it has limits $\sigma(-\infty) < \sigma(+\infty)$, then one can first affinely scale down its x-axis so that its graph looks like a step-function with two sharp "overshoots", then make a linear sum of enough of them to make a "staircase" approximation of the ramp function. With more steps of the staircase, the overshoots smooth out and we get arbitrarily good approximation of the ramp function.

The case where σ is a generic non-polynomial function is harder, and the reader is directed to.[6]

The above proof has not specified how one might use a ramp function to approximate arbitrary functions in $C_0(\mathbb{R}^n, \mathbb{R})$. A sketch of the proof is that one can first construct flat bump functions, intersect them to obtain spherical bump functions that approximate the Dirac delta function, then use those to approximate arbitrary functions in $C_0(\mathbb{R}^n, \mathbb{R})$.[43] The original proofs, such as the one by Cybenko, use methods from functional analysis, including the Hahn-Banach and Riesz–Markov–Kakutani representation theorems. Cybenko first published the theorem in a technical report in 1988,[44] then as a paper in 1989.[3]

Notice also that the neural network is only required to approximate within a compact set K. The proof does not describe how the function would be extrapolated outside of the region.

The problem with polynomials may be removed by allowing the outputs of the hidden layers to be multiplied together (the "pi-sigma networks"), yielding the generalization:[1]

Universal approximation theorem for pi-sigma networks—With any nonconstant activation function, a one-hidden-layer pi-sigma network is a universal approximator.

Arbitrary-depth case

The "dual" versions of the theorem consider networks of bounded width and arbitrary depth. A variant of the universal approximation theorem was proved for the arbitrary depth case by Zhou Lu et al. in 2017.[9] They showed that networks of width $n + 4$ with ReLU activation functions can approximate any Lebesgue-integrable function on n-dimensional input space with respect to L^1 distance if network depth is allowed to grow. It was also shown that if the width was less than or equal to n, this general expressive power to approximate any Lebesgue integrable function was lost. In the same paper[9] it was shown that ReLU networks with width $n + 1$ were sufficient to approximate any continuous function of n-dimensional input variables.[10] The following refinement, specifies the optimal minimum width for which such an approximation is possible and is due to.[45]

Universal approximation theorem ***(L1 distance, ReLU activation, arbitrary depth, minimal width)***—For any Bochner–Lebesgue p-integrable function $f : \mathbb{R}^n \to \mathbb{R}^m$ and any $\varepsilon > 0$, there exists a fully connected ReLU network F of width exactly $d_m = \max\{n+1, m\}$, satisfying

$$\int_{\mathbb{R}^n} \|f(x) - F(x)\|^p \, \mathrm{d}x < \varepsilon.$$

Moreover, there exists a function $f \in L^p(\mathbb{R}^n, \mathbb{R}^m)$ and some $\varepsilon > 0$, for which there is no fully connected ReLU network of width less than $d_m = \max\{n+1, m\}$ satisfying the above approximation bound.

Remark: If the activation is replaced by leaky-ReLU, and the input is restricted in a compact domain, then the exact minimum width is[20] $d_m = \max\{n, m, 2\}$.

Quantitative refinement: In the case where $f : [0,1]^n \to \mathbb{R}$, (i.e. $m = 1$) and σ is the ReLU activation function, the exact depth and width for a ReLU network to achieve ε error is also known.[16] If, moreover, the target function f is smooth, then the required number of layer and their width can be exponentially smaller.[46] Even if f is not smooth, the curse of dimensionality can be broken if f admits additional "compositional structure".[47][48]

Together, the central result of[11] yields the following universal approximation theorem for networks with bounded width (see also[7] for the first result of this kind).

Universal approximation theorem (Uniform non-affine activation, arbitrary depth, constrained width).—Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any non-affine continuous function which is continuously differentiable at at least one point, with nonzero derivative at that point. Let $\mathcal{N}^{\sigma}_{d,D:d+D+2}$ denote the space of feed-forward neural networks with d input neurons, D output neurons, and an arbitrary number of hidden layers each with $d + D + 2$ neurons, such that every hidden neuron has activation function σ and every output neuron has the identity as its activation function, with input layer ϕ and output layer ρ. Then given any $\varepsilon > 0$ and any $f \in C(\mathcal{X}, \mathbb{R}^D)$, there exists $\hat{f} \in \mathcal{N}^{\sigma}_{d,D:d+D+2}$ such that

$$\sup_{x \in \mathcal{X}} \left\| \hat{f}(x) - f(x) \right\| < \varepsilon.$$

In other words, $\mathcal{N}$ is dense in $C(\mathcal{X}; \mathbb{R}^D)$ with respect to the topology of uniform convergence.

Quantitative refinement: The number of layers and the width of each layer required to approximate f to ε precision known;[21] moreover, the result hold true when $\mathcal{X}$ and $\mathbb{R}^D$ are replaced with any non-positively curved Riemannian manifold.

Certain necessary conditions for the bounded width, arbitrary depth case have been established, but there is still a gap between the known sufficient and necessary conditions.[9][10][49]

Bounded depth and bounded width case

The first result on approximation capabilities of neural networks with bounded number of layers, each containing a limited number of artificial neurons was obtained by Maiorov and Pinkus.[13] Their remarkable result revealed that such networks can be universal approximators and for achieving this property two hidden layers are enough.

Universal approximation theorem:[13]—There exists an activation function σ which is analytic, strictly increasing and sigmoidal and has the following property: For any $f \in C[0,1]^d$ and $\varepsilon > 0$ there exist constants $d_i, c_{ij}, \theta_{ij}, \gamma_i$, and vectors $\mathbf{w}^{ij} \in \mathbb{R}^d$ for which

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

for all $\mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d$.

This is an existence result. It says that activation functions providing universal approximation property for bounded depth bounded width networks exist. Using certain algorithmic and computer programming techniques, Guliyev and Ismailov efficiently constructed such activation functions depending on a numerical parameter. The developed algorithm allows one to compute the activation functions at any point of the real axis instantly. For the algorithm and the corresponding computer code see.[14] The theoretical result can be formulated as follows.

Universal approximation theorem:[14][15]—Let $[a,b]$ be a finite segment of the real line, $s = b - a$ and λ be any positive number. Then one can algorithmically construct a computable sigmoidal activation function $\sigma \colon \mathbb{R} \to \mathbb{R}$, which is infinitely differentiable, strictly increasing on $(-\infty, s)$, λ -strictly increasing on $[s, +\infty)$, and satisfies the following properties:

1. For any $f \in C[a,b]$ and $\varepsilon > 0$ there exist numbers c_1, c_2, θ_1 and θ_2 such that for all $x \in [a,b]$

   $$|f(x) - c_1 \sigma(x - \theta_1) - c_2 \sigma(x - \theta_2)| < \varepsilon$$

2. For any continuous function F on the d-dimensional box $[a,b]^d$ and $\varepsilon > 0$, there exist constants e_p, c_{pq}, θ_{pq} and ζ_p such that the inequality

   $$\left| F(\mathbf{x}) - \sum_{p=1}^{2d+2} e_p \sigma \left(\sum_{q=1}^{d} c_{pq} \sigma(\mathbf{w}^q \cdot \mathbf{x} - \theta_{pq}) - \zeta_p \right) \right| < \varepsilon$$

   holds for all $\mathbf{x} = (x_1, \ldots, x_d) \in [a,b]^d$. Here the weights $\mathbf{w}^q$, $q = 1, \ldots, d$, are fixed as follows:

   $$\mathbf{w}^1 = (1, 0, \ldots, 0), \quad \mathbf{w}^2 = (0, 1, \ldots, 0), \quad \ldots, \quad \mathbf{w}^d = (0, 0, \ldots, 1).$$

   In addition, all the coefficients e_p, except one, are equal.

Here "$\sigma: \mathbb{R} \to \mathbb{R}$ is λ-strictly increasing on some set X" means that there exists a strictly increasing function $u: X \to \mathbb{R}$ such that $|\sigma(x) - u(x)| \leq \lambda$ for all $x \in X$. Clearly, a λ-increasing function behaves like a usual increasing function as λ gets small. In the "*depth-width*" terminology, the above theorem says that for certain activation functions depth-2 width-2 networks are universal approximators for univariate functions and depth-3 width-$(2d+2)$ networks are universal approximators for d-variable functions ($d > 1$).

See also

- Kolmogorov–Arnold representation theorem
- Representer theorem
- No free lunch theorem
- Stone–Weierstrass theorem
- Fourier series

References

1. Hornik, Kurt; Stinchcombe, Maxwell; White, Halbert (January 1989). "Multilayer feedforward networks are universal approximators". *Neural Networks*. **2** (5): 359–366. doi:10.1016/0893-6080(89)90020-8 (https://doi.org/10.1016%2F0893-6080%2889%2990020-8).
2. Balázs Csanád Csáji (2001) Approximation with Artificial Neural Networks; Faculty of Sciences; Eötvös Loránd University, Hungary
3. Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems*. **2** (4): 303–314. Bibcode:1989MCSS....2..303C (https://ui.adsabs.harvard.edu/abs/1989MCSS....2..303C). CiteSeerX 10.1.1.441.7873 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.441.7873). doi:10.1007/BF02551274 (https://doi.org/10.1007%2FBF02551274). S2CID 3958369 (https://api.semanticscholar.org/CorpusID:3958369).
4. Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". *Neural Networks*. **4** (2): 251–257. doi:10.1016/0893-6080(91)90009-T (https://doi.org/10.1016%2F0893-6080%2891%2990009-T). S2CID 7343126 (https://api.semanticscholar.org/CorpusID:7343126).
5. Leshno, Moshe; Lin, Vladimir Ya.; Pinkus, Allan; Schocken, Shimon (January 1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function" (http://archive.nyu.edu/handle/2451/14329). *Neural Networks*. **6** (6): 861–867. doi:10.1016/S0893-6080(05)80131-5 (https://doi.org/10.1016%2FS0893-6080%2805%2980131-5). S2CID 206089312 (https://api.semanticscholar.org/CorpusID:206089312).
6. Pinkus, Allan (January 1999). "Approximation theory of the MLP model in neural networks". *Acta Numerica*. **8**: 143–195. Bibcode:1999AcNum...8..143P (https://ui.adsabs.harvard.edu/abs/1999AcNum...8..143P). doi:10.1017/S0962492900002919 (https://doi.org/10.1017%2FS0962492900002919). S2CID 16800260 (https://api.semanticscholar.org/CorpusID:16800260).
7. Gripenberg, Gustaf (June 2003). "Approximation by neural networks with a bounded number of nodes at each level". *Journal of Approximation Theory*. **122** (2): 260–266. doi:10.1016/S0021-9045(03)00078-9 (https://doi.org/10.1016%2FS0021-9045%2803%2900078-9).
8. Yarotsky, Dmitry (October 2017). "Error bounds for approximations with deep ReLU networks". *Neural Networks*. **94**: 103–114. arXiv:1610.01145 (https://arxiv.org/abs/1610.01145). doi:10.1016/j.neunet.2017.07.002 (https://doi.org/10.1016%2Fj.neunet.2017.07.002). PMID 28756334 (https://pubmed.ncbi.nlm.nih.gov/28756334). S2CID 426133 (https://api.semanticscholar.org/CorpusID:426133).