

Universal approximation theorem

In the mathematical theory of artificial neural networks, **universal approximation theorems** are results^[1] that establish the density of an algorithmically generated class of functions within a given function space of interest. Typically, these results concern the approximation capabilities of the feedforward architecture on the space of continuous functions between two Euclidean spaces, and the approximation is with respect to the compact convergence topology. However, there are also a variety of results between non-Euclidean spaces^[2] and other commonly used architectures and, more generally, algorithmically generated sets of functions, such as the convolutional neural network (CNN) architecture,^{[3][4]} radial basis-functions,^[5] or neural networks with specific properties.^[6] Most universal approximation theorems can be parsed into 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).

Universal approximation theorems imply that neural networks can *represent* a wide variety of interesting functions when given appropriate weights. On the other hand, they typically do not provide a construction for the weights, but merely state that such a construction is possible.

Contents

History

Arbitrary Width Case

Arbitrary Depth Case

See also

References

History

One of the first versions of the *arbitrary width* case was proved by George Cybenko in 1989 for sigmoid activation functions.^[7] Kurt Hornik showed in 1991^[8] that it is not the specific choice of the activation function, but rather the multilayer feed-forward architecture itself which gives neural networks the potential of being universal approximators. Moshe Leshno *et al* in 1993^[9] and later Allan Pinkus in 1999^[10] showed that the universal approximation property^[11], is equivalent to having a nonpolynomial activation function.

The *arbitrary depth* case was also studied by number of authors, such as Zhou Lu *et al* in 2017,^[12] Boris Hanin and Mark Sellke in 2018,^[13] and Patrick Kidger and Terry Lyons in 2020.^[14] The result minimal width per layer was refined in ^[15].

Several extensions of the theorem exist, such as to discontinuous activation functions^[9], noncompact domains^[14], certifiable networks^[16] and alternative network architectures and topologies^{[14][17]}. A full characterization of the universal approximation property on general function spaces is given by A. Kratsios in ^[11].

Arbitrary Width Case

The classical form of the universal approximation theorem for arbitrary width and bounded depth is as follows.^{[7][8][18][19]} It extends^[10] the classical results of George Cybenko and Kurt Hornik.

Universal Approximation Theorem: Fix a continuous function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ (activation function) and positive integers d, D . The function σ is not a polynomial if and only if, for every continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}^D$ (target function), every compact subset K of \mathbb{R}^d , and every $\epsilon > 0$ there exists a continuous function $f_\epsilon : \mathbb{R}^d \rightarrow \mathbb{R}^D$ (the layer output) with representation

$$f_\epsilon = W_2 \circ \sigma \circ W_1,$$

where W_2, W_1 are composable affine maps and \circ denotes component-wise composition, such that the approximation bound

$$\sup_{x \in K} \|f(x) - f_\epsilon(x)\| < \epsilon$$

holds for any ϵ arbitrarily small (distance from f to f_ϵ can be infinitely small).

The theorem states that the result of first layer f_ϵ can approximate any well-behaved function f . Such a well-behaved function 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.

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.^[12] 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 there was the limited expressive power if the width was less than or equal to n . All Lebesgue integrable functions except for a zero measure set cannot be approximated by ReLU networks of width n . In the same paper^[12] it was shown that ReLU networks with width $n+1$ were sufficient to approximate any continuous function of n -dimensional input variables.^[20] The following refinement, specifies the optimal minimum width for which such an approximation is possible and is due to ^[21]

Universal Approximation Theorem (L1 distance, ReLU activation, arbitrary depth, minimal width). For any Bochner-Lebesgue p-integrable function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and any $\epsilon > 0$, there exists a fully-connected ReLU network F width exactly $d_m = \max\{n + 1, m\}$, satisfying

$$\int_{\mathbb{R}^n} \|f(x) - F(x)\|^p dx < \epsilon.$$

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

Together, the central results of ^[14] and of ^[2] yield the following general universal approximation theorem for networks with bounded width, between general input and output spaces.

Universal Approximation Theorem (non-affine activation, arbitrary depth, Non-Euclidean). Let \mathcal{X} be a compact topological space, $(\mathcal{Y}, d_{\mathcal{Y}})$ be a metric space, $\phi : \mathcal{X} \rightarrow \mathbb{R}^n$ be a continuous and injective feature map and let $\rho : \mathbb{R}^m \rightarrow \mathcal{Y}$ be a continuous readout map, with a section, having dense image $Im(\rho)$ with (possibly empty) collared boundary. Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be any non-affine continuous function which is continuously differentiable at at-least one point, with non-zero derivative at that point. Let $\mathcal{N}_{\phi, \rho}^{\sigma}$ denote the space of feed-forward neural networks with n input neurons, m output neurons, and an arbitrary number of hidden layers each with $n + m + 2$ 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 $f \in C(\mathcal{X}, \mathcal{Y})$, there exists $F \in \mathcal{N}_{\phi, \rho}^{\sigma}$ such that

$$\sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(F(x), f(x)) < \varepsilon.$$

In other words, \mathcal{N} is dense in $C(\mathcal{X}; \mathcal{Y})$ with respect to the uniform distance.

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.^{[12][13][22]}

See also

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

References

1. Balázs Csanád Csáji (2001) Approximation with Artificial Neural Networks; Faculty of Sciences; Eötvös Loránd University, Hungary
2. Kratsios, Anastasis; Bilokopytov, Eugene (2020). *Non-Euclidean Universal Approximation* (<http://papers.nips.cc/paper/2020/file/786ab8c4d7ee758f80d57e65582e609d-Paper.pdf>) (PDF). *Advances in Neural Information Processing Systems* 33. Curran Associates, Inc.
3. Zhou, Ding-Xuan (2020) Universality of deep convolutional neural networks; Applied and computational harmonic analysis 48.2 (2020): 787-794.
4. A. Heinecke, J. Ho and W. Hwang (2020); Refinement and Universal Approximation via Sparsely Connected ReLU Convolution Nets; IEEE Signal Processing Letters, vol. 27, pp. 1175-1179.
5. Park, Jooyoung, and Irwin W. Sandberg (1991); Universal approximation using radial-basis-function networks; Neural computation 3.2, 246-257.
6. Yarotsky, Dmitry (2018); Universal approximations of invariant maps by neural networks.
7. Cybenko, G. (1989) "Approximation by superpositions of a sigmoidal function" (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.441.7873&rep=rep1&type=pdf>), *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. doi:10.1007/BF02551274 (<https://doi.org/10.1007/9782FBF02551274>)

8. Kurt Hornik (1991) "[1] (<https://www.sciencedirect.com/science/article/abs/pii/089360809190009T>)", *Neural Networks*, 4(2), 251–257. doi:10.1016/0893-6080(91)90009-T (<https://doi.org/10.1016%2F0893-6080%2891%2990009-T>)
9. Leshno, Moshe; Lin, Vladimir Ya.; Pinkus, Allan; Schocken, Shimon (January 1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". *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>).
10. 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>).
11. Kratsios, Anastasis (August 7, 2020). "The Universal Approximation Property". *Annals of Mathematics and Artificial Intelligence*. doi:10.1007/s10472-020-09723-1 (<https://doi.org/10.1007%2FS10472-020-09723-1>) (inactive 2020-12-13).
12. Lu, Zhou; Pu, Hongming; Wang, Feicheng; Hu, Zhiqiang; Wang, Liwei (2017). "The Expressive Power of Neural Networks: A View from the Width" (<http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width>). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.: 6231–6239. arXiv:1709.02540 (<https://arxiv.org/abs/1709.02540>).
13. Hanin, Boris; Sellke, Mark (March 2019). "Approximating Continuous Functions by ReLU Nets of Minimal Width" (<https://www.mdpi.com/2227-7390/7/10/992>). *Mathematics*. MDPI. arXiv:1710.11278 (<https://arxiv.org/abs/1710.11278>).
14. Kidger, Patrick; Lyons, Terry (July 2020). *Universal Approximation with Deep Narrow Networks*. Conference on Learning Theory. arXiv:1905.08539 (<https://arxiv.org/abs/1905.08539>).
15. Park, Sejun; Yun, Chulhee; Lee, Jaeho; Shin, Jinwoo (October 2020). *Minimum Width for Universal Approximation*. Conference on Learning Theory. arXiv:1905.08539 (<https://arxiv.org/abs/1905.08539>).
16. Baader, Maximilian; Mirman, Matthew; Vechev, Martin (2020). *Universal Approximation with Certified Networks* (<https://openreview.net/forum?id=B1gX8kBtPr>). ICLR.
17. Lin, Hongzhou; Jegelka, Stefanie (2018). *ResNet with one-neuron hidden layers is a Universal Approximator* (<https://papers.nips.cc/paper/7855-resnet-with-one-neuron-hidden-layers-is-a-universal-approximator>). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. pp. 6169–6178.
18. Haykin, Simon (1998). *Neural Networks: A Comprehensive Foundation*, Volume 2, Prentice Hall. ISBN 0-13-273350-1.
19. Hassoun, M. (1995) *Fundamentals of Artificial Neural Networks* MIT Press, p. 48
20. Hanin, B. (2018). *Approximating Continuous Functions by ReLU Nets of Minimal Width*. arXiv preprint arXiv:1710.11278.
21. Park, Yun, Lee, Shin, Sejun, Chulhee, Jaeho, Jinwoo (2020-09-28). "Minimum Width for Universal Approximation" (<https://openreview.net/forum?id=O-XJwyoIF-k>). ICLR. arXiv:2006.08859 (<https://arxiv.org/abs/2006.08859>).
22. Johnson, Jesse (2019). *Deep, Skinny Neural Networks are not Universal Approximators* (<https://openreview.net/forum?id=ryGgSsAcFQ>). International Conference on Learning Representations.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Universal_approximation_theorem&oldid=999802138"

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.