CMSE 890-002: Mathematics of Deep Learning, MSU, Spring 2020

Lecture 24: Deep Approximation of Compositional Functions II

March 13, 2020

Lecturer: Matthew Hirn

We prove the following:

Theorem (Poggio, et al., [18]). *Let $\sigma \in \mathbf{C}^{\infty}(\mathbb{R})$ not be a polynomial. Let $F \in \mathbf{C}_2^s[-1,1]^d$ and let $\{H_\lambda \in \mathbf{C}^s[-1,1]^2\}_\lambda$ be the constituent functions of F, each satisfying $\|H_\lambda\|_{\mathbf{C}^s[-1,1]^2} \leq 1$. Then,*

$$\inf_{f \in \mathcal{D}_{N,2}(\sigma)} \|F - f\|_{\mathbf{L}^\infty[-1,1]^d} \leq C(d,s) N^{-s/2} \, .$$

Stated another way, in order to guarantee

$$\inf_{f \in \mathcal{D}_{N,2}(\sigma)} \|F - f\|_{\mathbf{L}^\infty[-1,1]^d} \leq \epsilon$$

for an arbitrary $F \in \mathbf{C}_2^s[-1,1]^d$ with $\|H_\lambda\|_{\mathbf{C}^s[-1,1]^2} \leq 1$, one must take

$$N = C'(d,s) \epsilon^{-2/s} \, .$$

Proof. Let $d = 2^J$. Recall that each node of the network $f \in \mathcal{D}_{N,2}$ has $m = N/|V| = N/(d-1)$ neurons inside the node. Let $H_\lambda \in \mathbf{C}^s[-1,1]^2$ be one of the constituent functions of F. We can apply Theorem 9.6 to conclude that there is a node $\overline{\eta}_\lambda \in \mathcal{M}_m(\sigma)$ such that

$$\|H_\lambda - \overline{\eta}_\lambda\|_{\mathbf{L}^\infty[-1,1]^2} \leq C m^{-s/2} \, . \tag{40}$$

That works for the individual functions making up F, but we have to check that when we compose different H_λ functions, the error does not get too large. So let us consider $d = 4$, which means the label function is of the form

$$F = H_1(H_{11}, H_{12}) \, ,$$

and let $\overline{\eta}_1$, $\overline{\eta}_{11}$, and $\overline{\eta}_{12}$ be the nodes that approximate H_1, H_{11}, and H_{12}, respectively. Then:

$$\begin{aligned}
&\|H_1(H_{11}, H_{12}) - \overline{\eta}_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4} \\
&\quad = \|H_1(H_{11}, H_{12}) - H_1(\overline{\eta}_{11}, \overline{\eta}_{12}) + H_1(\overline{\eta}_{11}, \overline{\eta}_{12}) - \overline{\eta}_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4} \\
&\quad \leq \|H_1(H_{11}, H_{12}) - H_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4} + \|H_1(\overline{\eta}_{11}, \overline{\eta}_{12}) - \overline{\eta}_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4}
\end{aligned} \tag{41}$$

For the second term we can apply (40) nearly directly. Write

$$z = (z(1), z(2), z(3), z(4)) \in [-1,1]^4$$

as $z = (z_1, z_2)$ with $z_1 = (z(1), z(2))$ and $z_2 = (z(3), z(4))$. We have:

$$\begin{aligned}
&\|H_1(\overline{\eta}_{11}, \overline{\eta}_{12}) - \overline{\eta}_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4} \\
&\quad = \sup_{z\in[-1,1]^4} |H_1(\overline{\eta}_{11}(z_1), \overline{\eta}_{12}(z_2)) - \overline{\eta}_1(\overline{\eta}_{11}(z_1), \overline{\eta}_{12}(z_2))| \\
&\quad = \sup_{u\in[-1,1]^2} |H_1(u(1), u(2)) - \overline{\eta}_1(u(1), u(2))| \,, \quad [u(1) = \overline{\eta}_{11}(z_1)\,,\, u(2) = \overline{\eta}_{12}(z_2)] \\
&\quad = \|H_1 - \overline{\eta}_1\|_{\mathbf{L}^\infty[-1,1]^2} \\
&\quad \leq Cm^{-s/2} \,.
\end{aligned} \tag{42}$$

For the first term, let $u, \bar{u} \in \mathbb{R}^2$. We first observe:

$$|H_1(u) - H_1(\bar{u})| \leq \sup_{v\in\mathbb{R}^2} \|\nabla H_1(v)\|_2 \|u - \bar{u}\|_2 \,. \tag{43}$$

We also have:

$$\begin{aligned}
\sup_{v\in[-1,1]^2} \|\nabla H_1(v)\|_2 &\leq \sup_{v\in[-1,1]^2} \|\nabla H_1(v)\|_1 \\
&= \sup_{v\in[-1,1]^2} [|\partial_1 H_1(v)| + |\partial_2 H_1(v)|] \\
&= \||\partial_1 H_1| + |\partial_2 H_1|\|_{\mathbf{L}^\infty[-1,1]^2} \\
&\leq \|\partial_1 H_1\|_{\mathbf{L}^\infty[-1,1]^2} + \|\partial_2 H_1\|_{\mathbf{L}^\infty[-1,1]^2} \\
&\leq \|H_1\|_{\mathbf{C}^s[-1,1]^2} \\
&\leq 1 \,.
\end{aligned}$$

Therefore, combining with (43) we have

$$|H_1(u) - H_1(\bar{u})| \leq \|u - \bar{u}\|_2 \,.$$

Now plug in

$$\begin{aligned}
u &= (H_{11}(z_1), H_{12}(z_2)) \\
\bar{u} &= (\overline{\eta}_{11}(z_1), \overline{\eta}_{12}(z_2)) \,.
\end{aligned}$$

We get:

$$\begin{aligned}
&|H_1(H_{11}(z_1), H_{12}(z_2)) - H_1(\overline{\eta}_{11}(z_1), \overline{\eta}_{12}(z_2))|^2 \\
&\qquad \leq |H_{11}(z_1) - \overline{\eta}_{11}(z_1)|^2 + |H_{12}(z_2) - \overline{\eta}_{12}(z_2)|^2 \\
&\qquad \leq \|H_{11} - \overline{\eta}_{11}\|^2_{\mathbf{L}^\infty[-1,1]^2} + \|H_{12} - \overline{\eta}_{12}\|^2_{\mathbf{L}^\infty[-1,1]^2} \\
&\qquad \leq 2Cm^{-s} \,.
\end{aligned} \tag{44}$$

Therefore:

$$\|H_1(H_{11}, H_{12}) - H_1(\overline{\eta}_{11}, \overline{\eta}_{12})\|_{\mathbf{L}^\infty[-1,1]^4} \leq \sqrt{2}Cm^{-s/2} \,.$$

Combining (41), (42), and (44) we get:

$$\| \underbrace{H_1(H_{11}, H_{12})}_{F} - \underbrace{\overline{\eta}_1(\overline{\eta}_{11}, \overline{\eta}_{12})}_{f} \|_{\mathbf{L}^\infty[-1,1]^4} \leq (1+\sqrt{2})Cm^{-s/2}\,.$$

If the $d > 4$ we can recursively apply the bounds computed above. The number of times we will apply these bound will depend only on d. On the right hand side, they are all of the form $Cm^{-s/2}$, and so we will get (for general $d = 2^J$)

$$\|F - f\|_{\mathbf{L}^\infty[-1,1]^d} \leq C(d)m^{-s/2}\,.$$

Now recall that $m = N/(d-1)$. Thus in fact we have:

$$\|F - f\|_{\mathbf{L}^\infty[-1,1]^d} \leq C(d)\left(\frac{N}{d-1}\right)^{-s/2} = C(d,s)N^{-s/2}\,.$$

□

Remark 9.8. Remember, N is a proxy for the complexity of the network in terms of the number of trainable parameters, which is $O(N)$ for both the shallow and deep networks. Thus the deep networks are much more efficient learners than shallow networks for compositional functions.

Remark 9.9. Suppose F 2-compositional but we guess a network in $\mathcal{D}_{N,q}$, that is to say, it aggregates information q variables at a time as opposed to 2 variables at a time. Another scenario is if F is q-compositional and we use a network from $\mathcal{D}_{N,q}$ (which matches the compositional structure of F). In either case the learning rate is:

$$\inf_{f \in \mathcal{D}_{N,q}(\sigma)} \|F - f\|_{\mathbf{L}^\infty[-1,1]^d} \leq C(d,s,q)N^{-s/q}\,.$$

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550:354—359, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[5] Larry Greenemeier. AI versus AI: Self-taught AlphaGo Zero vanquishes its predecessor. *Scientific American*, October 18, 2017.

[6] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. arXiv:1803.08823, 2018.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[8] J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic - unbiased estimation of the inverse covariance matrix. *Astronomy and Astrophysics*, 464(1):399–404, 2007.

[9] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, San Diego, CA, USA, 2015.

[12] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

[13] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[14] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[15] Vitaly E. Maiorov. On best approximation by ridge functions. *Journal of Approximation Theory*, 99:68–94, 1999.

[16] Vitaly E. Maiorov and Allan Pinkus. Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25:81–91, 1999.

[17] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940. PMLR, 2016.

[18] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[19] Hrushikesh Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8:164–177, 1996.