Lecture 20: On the Theorem of Maiorov and Pinkus

February 26, 2020

Lecturer: Matthew Hirn

Let us continue our discussion of Theorem 8.13.

Remark 8.16. The proof of Theorem 8.13 is based upon the Kolmogorov Superposition Theorem. In particular it utilizes an improved version, stated below.

Theorem 8.17 (Kolmogorov Superposition Theorem)**.** *There exists d constants $\lambda_j > 0$, $1 \leq j \leq d$, with $\sum_{j=1}^{d} \lambda_j \leq 1$, and $2d+1$ strictly increasing continuous functions $\phi_k : [0,1] \to [0,1]$, $1 \leq k \leq 2d+1$, such that every $F \in \mathbf{C}[0,1]^d$ can be represented as*

$$F(x) = F(x(1), \ldots, x(d)) = \sum_{k=1}^{2d+1} G\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) \right), \qquad (33)$$

for some $G \in \mathbf{C}[0,1]$ depending on F.

Using the Kolmogorov Superposition Theorem we can (crudely) sketch the proof of Theorem 8.13.

Proof sketch of Theorem 8.13. Using the Kolmogorov Superposition Theorem we write $F(x)$ as in (33). Let σ be the same activation function as from Theorem 8.11. We first approximate G using σ. In the proof of Theorem 8.11 (which is Proposition 6.3 and Corollary 6.4 in [13]), it is shown that σ can be constructed in such a way that for each $H \in \mathbf{C}[-1,1]$ and $\eta > 0$, there exist constants $a_1, a_2, a_3 \in \mathbb{R}$ and an integer $m \in \mathbb{Z}$ for which

$$\forall\, z \in [-1,1], \quad |H(z) - (a_1\sigma(z-3) + a_2\sigma(z+1) + a_3\sigma(z+m))| < \eta. \qquad (34)$$

Furthermore, $\sigma(z-3)$ and $\sigma(z+1)$ are linear on $[0,1]$. The construction of σ is accomplished by using the fact that $\mathbf{C}^\infty[-1,1]$ is dense in $\mathbf{C}[-1,1]$, which means there exists a countable collection of functions $\{h_k\}_{k=1}^{\infty} \subset \mathbf{C}^\infty[-1,1]$ so that for each $H \in \mathbf{C}[-1,1]$ and each η there exists $k = k(H, \eta)$ with

$$\sup_{z \in [-1,1]} |H(z) - h_k(z)| < \eta.$$

Pinkus then cleverly constructs σ so that for each $k \geq 1$ there exists constants $a_{1,k}, a_{2,k}, a_{3,k}$ with

$$a_{1,k}\sigma(z-3) + a_{2,k}\sigma(z+1) + a_{3,k}\sigma(z+4k+1) = u_k(z).$$

while also ensuring that $\sigma(z-3)$ and $\sigma(z+1)$ are linear on $[0,1]$ (in fact he places more restrictions on σ, but we will not need them for this discussion).

Anyway, with (34) in hand we can apply it to G with $\eta = \epsilon/2(2d+1)$ and restrict the domain from $[-1,1]$ to $[0,1]$, which gives:

$$\forall\, z \in [0,1]\,, \quad |G(z) - (a_1\sigma(z-3) + a_2\sigma(z+1) + a_3\sigma(z+m))| < \frac{\epsilon}{2(2d+1)}\,.$$

We now use this approximation and the Kolmogorov Superposition Theorem to obtain:

$$\begin{aligned}\forall\, x \in [0,1]^d\,, \quad &\Bigg| F(x) - \sum_{k=1}^{2d+1} \Bigg[a_1\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) - 3\right) \\ &\quad + a_2\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + 1\right) + a_3\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + m\right)\Bigg]\Bigg| < \frac{\epsilon}{2}\,.\end{aligned}$$

Recall that $\sigma(z-3)$ and $\sigma(z+1)$ are linear on $[0,1]$, and that by the Kolmogorov Superposition Theorem $\phi_k : [0,1] \to [0,1]$, so we can combine the first two terms:

$$\begin{aligned}\sum_{k=1}^{2d+1} a_1 \Bigg[\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) - 3\right) + a_2\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + 1\right)\Bigg] \\ = \sum_{k=1}^{2d+2} c_k \sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + \gamma_k\right)\,,\end{aligned}$$

where ϕ_{2d+2} is ϕ_k for some $1 \leq k \leq 2d+1$ and $\gamma_k \in \{-3, 1\}$ for each k. We thus have:

$$\forall\, x \in [0,1]^d\,, \quad \left| F(x) - \sum_{k=1}^{2d+2} c_k \sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + \gamma_k\right) - a_3\sigma\left(\sum_{j=1}^{d} \lambda_j \phi_k(x(j)) + m\right)\right| < \frac{\epsilon}{2}\,.$$

The proof proceeds by applying (34) to each $H = \phi_k$ for η small enough, and again using the fact that $\sigma(z-3)$ and $\sigma(z+1)$ are linear on $[0,1]$. After combining terms, the result is obtained. For more details on the proof, see [13, Theorem 7.2]. □

Remark 8.18. The proof of Theorem 8.13 gives the structure of this two-layer network, and in fact the number of connections between the first hidden layer and the second hidden layer is quite small. In particular, $f(x;\theta)$ can be written as:

$$f(x;\theta) = \sum_{\ell=1}^{4d+3} \alpha(\ell)\sigma\left(\sum_{k=1}^{2d+1} w_{2,\ell}(k)\sigma(\langle x, w_{1,k,\ell}\rangle + b_1(k,\ell)) + b_2(\ell)\right)\,,$$

where $w_{1,k,\ell} \in \mathbb{R}^d$ for $1 \leq k \leq 2d+1$ and $1 \leq \ell \leq 4d+3$, $b_1 \in \mathbb{R}^{(2d+1)\times(4d+3)}$, $w_{2,\ell} \in \mathbb{R}^{2d+1}$, and $b_2 \in \mathbb{R}^{4d+3}$. The network is illustrated in Figure 27. In [13], Pinkus says the network has $2d+1$ units in the first layer and $4d+3$ units in the second layer, but by most definitions of neurons, as well as our own, it has $(2d+1)(4d+3)$ neurons in the first layer and $4d+3$ neurons in the second layer.

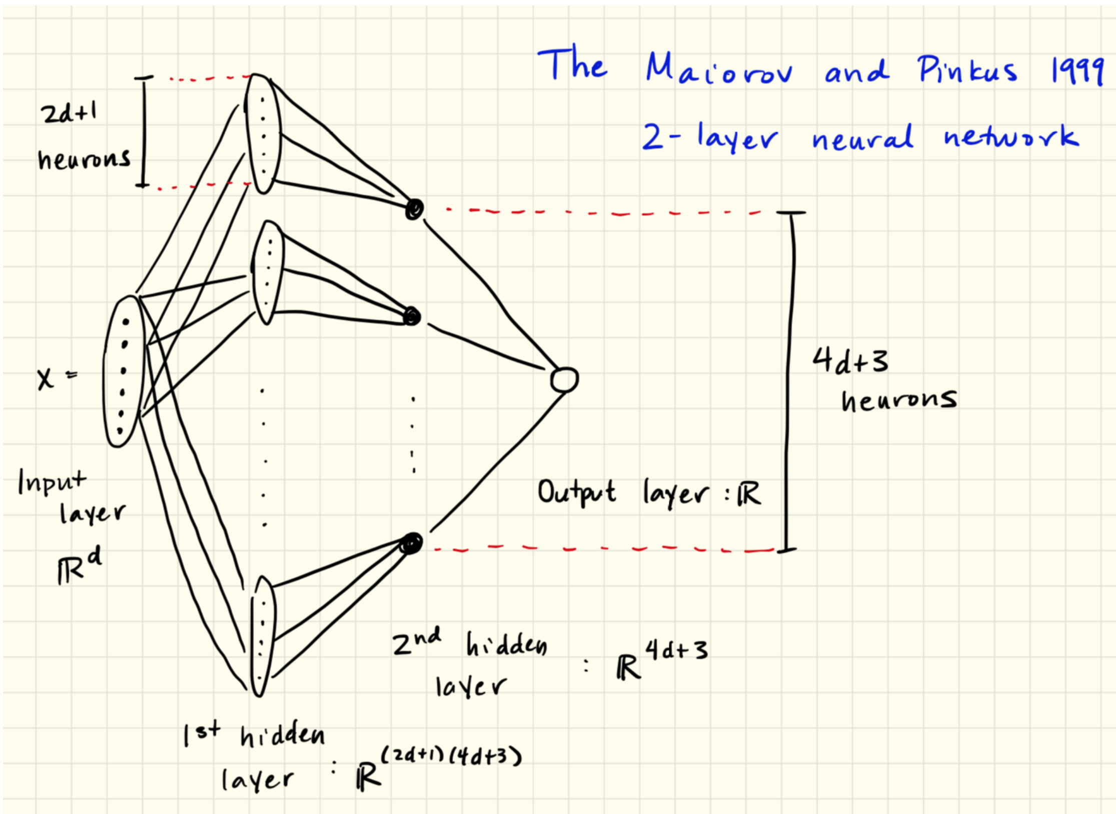

Figure 27: *Drawing of the Maiorov and Pinkus (1999) two-layer neural network that achieves the result of Theorem 8.13.*

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] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

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

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

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

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

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

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

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

[16] 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.