Stats 231C - Homework 4

Peter Racioppo (103953689)

May 27, 2020

Preliminary Review for Final Project:

Universal approximation results for shallow neural networks were obtained in the late 1980s and early 1990s ((Cybenko, 1989), (Hornik et al, 1989), (Barron, 1993)). These results showed that any continuous function f on a compact subset of R^d can be approximated to arbitrary accuracy by a shallow, fully-connected neural network (FNN) with a sufficiently large number of neurons and sigmoidal activations. These results were extended in (Leshno, et al, 1993) and (Pinkus, 1999), which show that these networks are universal approximators if and only if they have non-polynomial activations. These results were also extended to deep networks ((Hornik et al, 1989), (Chui et al, 1989), (Mhaskar, 1993)). Results in (Barron, 1993) and (Mhaskar, 1993) gave rates of convergence. Later work proved that deep but narrow neural networks and deep belief networks are also universal approximators, and require no more parameters than shallow networks to achieve universal approximation (resp. (Rojas, 2003), (Sutskever and Hinton, 2008)). Work in the late 2010s, by e.g. (Telgarsky, 2016), (Eldan and Shamir, 2016), (Yarotsky, 2017), (Shaham, et al, 2018) explored the representational power of deep neural networks and the importance of network depth. (Bolcskei, et al, 2018) and (Petersen and Voigtlaender, 2018) showed that, for some function classes, sparsely connected neural networks can achieve approximation error bounds of the same order as those obtained by fully-connected networks. However, these works do not give any conditions on the patterns of sparse connections.

Universal approximation results were extended to CNNs in the late 2010s (Zhou, 2018), (Yarotsky, 2018), (Petersen and Voigtlaender, 2018). In (Zhou, 2018), the author shows that an arbitrary sequence can be factorized into convolutions of a finite sequence of filter masks. Zhou uses this result to construct the sparsely connected weight and bias vectors and, using a result from (Klusowski and Barron, 2018) on approximation by ramp ridge functions, to construct bounds on the approximation error. These results indicate that the bounds on the number of free parameters in a CNN are significantly more favorable than the corresponding bounds for a dense network. As a corollary of the error bound result, Zhou proves a universality result for deep CNNs. In particular, any continuous function on a compact subset of R^d can be ap-

proximated by a sufficiently deep CNN. Formally, given the hypothesis space $H_J^{w,b} = \{\sum_{k=1}^{d_J} c_k h_k^{(J)}(x) : c \in R^{d_J}\}$ of CNNs, for any $\Omega \subseteq R^d$, $f \in C(\Omega)$, $\exists w, b$ and $\exists f_J^{w,b} \in H_J^{w,b}$ s.t. $\lim_{J\to\infty} ||f - f_J^{w,b}||_{C(\Omega)} = 0$.

(Yarotsky, 2018) introduces a "charge-conserving Convnet" model, and shows that it is a universal approximator for continuous equivariant transformations in SE(2), the group of isometries in a two-dimensional Euclidean space. In (Petersen and Voigtlaender, 2018), the authors prove an equivalence result for FNNs and CNNs. Namely, for any FNN approximating a function $f : R^n \mapsto R$, there exists a CNN with the same number of parameters, up to a constant factor, which approximates a translation equivariant function $g : R^n \mapsto R^n$, with f equal to the first coordinate of g. A converse result also holds.

Important recent work on optimization includes (Lee , et al, 2016), in which the Stable Manifold Theorem from Dynamical Systems Theory is used to show that gradient descent always converges to a minimizer, provided some mild conditions are met by the objective function. (Du, et al, 2019) applied this result to show that, again under mild assumptions on the objective function, activation function, and gradient descent step size, gradient descent applied to a sufficiently large FNN always converges to a global minimum during training. An earlier result in (Brutzkus and Globerson, 2017) shows that, for a convolutional neural net with a single hidden layer, with no overlap of filters and a ReLU activation, and with a Gaussian input distribution, gradient descent is guaranteed to converge to a global minimum in polynomial time during training.

It appears that to quantify the efficiency and invariance properties of CNNs, it may be necessary to consider models of data generation which reflect the invariant or hierarchical properties of real-world data. Some recent work in this direction includes (Malach and Shalev-Shwartz, 2018), which describes a method for training deep CNNs layer by layer. The algorithm applies gradient training for a two-layer network followed by a clustering algorithm, and is based on a hierarchical generative model for natural images. In (Bietti and Mairal, 2017), the authors study representations of data by deep CNNs, and their invariance to translations and more general transformations. They also introduce the RKHS norm as a measure of model complexity.

The methods in (Zhou, 2018) can perhaps be used to establish approximation bounds for CNNs with pooling. The methods in (Du, et al, 2019) for FNNs using Dynamical Systems Theory might be useful in extending the results in (Brutzkus and Globerson, 2017) to multilayer CNNs. We might also explore alternatives or modifications to the data generation model in (Malach and Shalev-Shwartz, 2018). It would be interesting to study graphical models and other methods of representing spatial hierarchies in data, perhaps along the lines of (Sabour, Frosst, and Hinton, 2017).

References

A. Krizhevsky, I. Sutskever, G. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." ICLR 2015.

Y. LeCun, B. Boser, J. Denker, D. Henderson, E. Howard, W. Hubbard, L. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition." Neural Computation. MIT Press - Journals. 1989.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. "Gradient-based learning applied to document recognition." 1998. Proceedings of the IEEE. 86 (11): 2278–2324.

Y. LeCun and Y. Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks, 3361(10), 1995.

Y. LeCun, Y. Bengio, G. Hinton. Deep learning. Nature, 521(7553):436–444, May 2015.

M. Zeiler, R. Fergus. "Visualizing and Understanding Convolutional Networks." 2013. arXiv:1311.2901.

K. Simonyan, A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." 4 Sept. 2014. arXiv:1409.1556.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. "Going Deeper with Convolutions" 17 Sept. 2014. arXiv:1409.4842.

K. He, X. Zhang, S. Ren, J. Sun. "Deep Residual Learning for Image Recognition." 10 Dec. 2015. arXiv:1512.03385.

G. Larsson, M. Maire, G. Shakhnarovich. "FractalNet: Ultra-Deep Neural Networks without Residuals." arXiv:1605.07648.

S. Sabour, N. Frosst, G. Hinton. "Dynamic Routing Between Capsules." 26 Oct. 2017. arXiv:1710.09829.

G. Cybenko. "Approximations by superpositions of sigmoidal func- tions." 1989. Mathematics of Control, Signals, and Systems 2: 303–314.

K. Hornik, M. Stinchcombe, H. White. "Multilayer feedforward net- works are universal approximators." 1989. Neural networks 2: 359-366.

A. Barron. "Universal approximation bounds for superpositions of a sigmoidal function." 1993. IEEE Transactions on Information Theory 39: 930- 945.

M. Leshno, Y. Lin, A. Pinkus, S. Schocken. "Multilayer feedforward networks with a non-polynomial activation function can approximate any function." 1993. Neural Networks 6: 861–867.

A. Pinkus. "Approximation theory of the MLP model in neural networks." 1999. Acta Numerica 8: 143–195.

H. Mhaskar. "Approximation properties of a multilayered feedfor- ward artificial neural network." 1993. Advances in Computational Mathematics 1: 61–80.

C. Chui, X. Li, H. Mhaskar. "Limitations of the approximation capabilities of neural networks with one hidden layer." 1996. Advances in Computational Mathematics 5: 233–243.

R. Rojas. "Networks of width one are universal classifiers." 2003. In International Joint Conference on Neural Networks, volume 4, pages 3124–3127.

I. Sutskever, G. Hinton. "Deep, narrow sigmoid belief networks are universal approximators." 2008. Neural Computation, 20(11), 2629–2636.

M. Telgarsky. "Benefits of depth in neural networks." 2016. 29th Annual Con-

ference on Learning Theory PMLR 49:1517–1539.
R. Eldan, O. Shamir. "The power of depth for feedforward neural networks." 2016. COLT: 907-940.
D. Yarotsky. "Error bounds for approximations with deep ReLU networks." 2017. Neural Networks 94: 103–114.
U. Shaham, A. Cloninger, R. Coifman. "Provable approximation properties for deep neural networks." 2018. Applied and Computational Har- monic Analysis 44: 537–557.
H. Bolcskei, P. Grohs, G. Kutyniok, P. Petersen. "Optimal approximation with sparsely connected deep neural networks." 2018. arXiv:1705.01714v4.
P. Petersen, V. Voigtlaender. "Optimal approximation of piecewise smooth functions using deep ReLU neural networks." 2018. arXiv:1709.05289v4.
D.X. Zhou. "Universality of Deep Convolutional Neural Networks." arXiv:1805.10769. 28 May 2018.
D. Yarotsky. "Universal approximations of invariant maps by neural networks." arXiv:1804.10306. 26 April 2018.
P. Petersen and F. Voigtlaender. "Equivalence of approximation by convolutional neural networks and fully-connected networks." arXiv:1809.00973. 4 Sept. 2018.
J. Klusowski , A. Barron. "Uniform approximation by neural networks activated by first and second order ridge splines." 2018. arXiv:1607.07819v2.
J. Lee , M. Simchowitz , M. Jordan, B. Recht. "Gradient Descent Only Converges to Minimizers." 2016. JMLR: Workshop and Conference Proceedings vol 49:1–12, 2016.
S. Du, J. Lee, H. Li, L. Wang, X. Zhai. "Gradient Descent Finds Global Minima of Deep Neural Networks." 2019. arXiv:1811.03804.
A. Brutzkus, A. Globerson. "Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs." arXiv:1702.07966. 26 Feb 2017.
Q. Nguyen, M. Hein. "Optimization Landscape and Expressivity of Deep CNNs." 30 Oct 2017. arXiv:1710.10928.
E. Malach, S. Shalev-Shwartz. "A Provably Correct Algorithm for Deep Learning that Actually Works." School of Computer Science, The Hebrew University, Israel. arXiv:1803.09522v2 [cs.LG]. 24 June 2018.
A. Bietti, J. Mairal. "Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations." arXiv:1706.03078. 9 Jun 2017.

Stats 231C - Homework 4

Peter Racioppo (103953689)

May 26, 2020

Exercise 2. *(Averaged SGD with variable step size, 10 points). Prove an analog version of following theorem that we had in class, but using a step size* $\eta = \frac{B}{\rho\sqrt{t}}$ *at iteration* $t = 1, ..., T$.

Lemma 1. *Let* $v_1, ..., v_T$ *be a sequence of vectors. Any algorithm with initialization* $w^{(1)} = 0$ *and an update rule* $w^{(t+1)} = w^{(t)} - \eta v_t$ *satisfies*
$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} ||v_t||^2$.

Theorem. *Let* $B, \rho > 0$. *Let* f *be a convex function and let* $w^* \in argmin_{w:||w|| \leq B} f(w)$. *Assume that SGD is run for* T *iterations with step size* $\eta = \frac{B}{\rho\sqrt{T}}$. *Assume also that for all* t, $||v_t|| \leq \rho$ *with probability 1. Then* $E[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}$, *where* $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$.

Write $v_{1:t}$ for the sequence $v_1, ..., v_t$. We have that:
$f(\bar{w}) - f(w^*) = f(\frac{1}{T} \sum_{t=1}^{T} w^{(t)}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) - f(w^*)$ (Jensen's inequality) $\leq \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, f(w^{(t)}) \rangle$ (by the convexity of f).
Taking the expectation of each side,
$E_{v_{1:t}}(f(\bar{w}) - f(w^*)) \leq E_{v_{1:t}}(\frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) - f(w^*))$.

Let $z_t = \sqrt{\frac{T}{t}} v_t$. Thus, $w^{(t+1)} = w^{(t)} - \eta v_t = w^{(t)} - \frac{B}{\rho\sqrt{t}} v_t = w^{(t)} - \frac{B}{\rho\sqrt{T}} z_t$
$= w^{(t)} - H z_t = w^{(t)}$, where $H = \frac{B}{\rho\sqrt{T}}$.
By Lemma 1, $\sum_{t=1}^{T} \langle w^{(t)} - w^*, z_t \rangle \leq \frac{||w^*||^2}{2H} + \frac{H}{2} \sum_{t=1}^{T} ||z_t||^2$, so,
$\sum_{t=1}^{T} \sqrt{\frac{T}{t}} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{||w^*||^2}{2H} + \frac{H}{2} \sum_{t=1}^{T} \frac{T}{t} . ||v_t||^2$.
But, since $\langle w^{(t)} - w^*, v_t \rangle$ is positive for all t,
$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \sum_{t=1}^{T} \sqrt{\frac{T}{T}} \langle w^{(t)} - w^*, v_t \rangle \leq \sum_{t=1}^{T} \sqrt{\frac{T}{t}} \langle w^{(t)} - w^*, v_t \rangle$.
Thus, $\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{||w^*||^2}{2H} + \frac{H}{2} \sum_{t=1}^{T} \frac{T}{t} ||v_t||^2$
$= \frac{\rho\sqrt{T}||w^*||^2}{2B} + \frac{BT}{2\rho\sqrt{T}} \sum_{t=1}^{T} \frac{1}{t} ||v_t||^2$
$\leq \frac{\rho B^2\sqrt{T}}{2B} + \frac{\rho^2 BT}{2\rho\sqrt{T}} \sum_{t=1}^{T} \frac{1}{t}$ ($\forall$w* with $||w^*|| \leq B$)
$= \frac{\rho B\sqrt{T}}{2}(1 + \sum_{t=1}^{T} \frac{1}{t})$.
Thus, $\frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\rho B}{2\sqrt{T}}(1 + \sum_{t=1}^{T} \frac{1}{t})$.

Finally, since $E_{v_{1:t}}(\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^*)) \leq E_{v_{1:t}}(\frac{1}{T}\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t\rangle)$,

$$E[f(\bar{w}) - f(w^*)] \leq \frac{\rho B}{2\sqrt{T}}(1 + \sum_{t=1}^{T}\frac{1}{t}) \tag{1}$$

This bound converges, though much more slowly than the previous bound. Hence, we have at least shown that SGD with this variable step size converges. (We can bound this function with simpler functions using the fact that $\sum_{t=1}^{T}\frac{1}{t} \leq \log_2(T+1) \leq \sqrt{T}$, but this is not very useful.)

Another bound can be obtained by using the fact that:
$\langle w^{(t)} - w^*, v_t\rangle = \frac{1}{\eta_t}\langle w^{(t)} - w^*, \eta_t v_t\rangle$
$= \frac{1}{2\eta_t}(-||w^{(t)} - w^* - \eta_t v_t||^2 + ||w^{(t)} - w^*||^2 + \eta_t^2||v_t||^2)$
$= \frac{1}{2\eta_t}(-||w^{(t+1)} - w^*||^2 + ||w^{(t)} - w^*||^2) + \frac{\eta_t}{2}||v_t||^2$.
Thus, $\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t\rangle = \sum_{t=1}^{T}\frac{1}{2\eta_t}(||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2) + \sum_{t=1}^{T}\frac{\eta_t}{2}||v_t||^2$
$= \sum_{t=1}^{T}\frac{\rho\sqrt{t}}{2B}(||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2) + \sum_{t=1}^{T}\frac{B}{2\rho\sqrt{t}}||v_t||^2$
$\leq \sum_{t=1}^{T}\frac{\rho\sqrt{T}}{2B}(||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2) + \sum_{t=1}^{T}\frac{B}{2\rho\sqrt{t}}||v_t||^2$
(since every term in the first sum is positive)
$= \frac{\rho\sqrt{T}}{2B}(||w^{(1)} - w^*||^2 - ||w^{(T+1)} - w^*||^2) + \frac{B}{2\rho}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}||v_t||^2$
(using a telescopic identity for the sum)
$\leq \frac{\rho\sqrt{T}}{2B}(||w^{(1)} - w^*||^2) + \frac{B}{2\rho}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}||v_t||^2$
$= \frac{\rho\sqrt{T}}{2B}(||w^*||^2) + \frac{B}{2\rho}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}||v_t||^2$ (since $w^{(1)} = 0$)
$\leq \frac{\rho\sqrt{T}}{2B}B^2 + \frac{B}{2\rho}\rho^2\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$
$= \frac{\rho B\sqrt{T}}{2} + \frac{\rho B}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$.
Thus, $\frac{1}{T}\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t\rangle \leq \frac{\rho B}{2\sqrt{T}} + \frac{\rho B}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$.
$= \frac{\rho B}{2\sqrt{T}}(1 + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}})$. It follows that:

$$E[f(\bar{w}) - f(w^*)] \leq \frac{\rho B}{2\sqrt{T}}(1 + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}) \tag{2}$$

Furthermore, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \sum_{t=1}^{T}\frac{1}{t}$ for all $T \geq 1$, so this is a better bound. In fact, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$ is a monotonically increasing function for positive T and $\lim_{t\to\infty}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} = 2$. It follows that:

$$E[f(\bar{w}) - f(w^*)] \leq \frac{3\rho B}{2\sqrt{T}} \tag{3}$$

This is the same bound, up to a constant (3/2), as our bound for the case of constant step size.