COS 597G: Toward Theoretical Understanding of Deep Learning **Fall 2018**

Lecture 2: 19 September 2018

Lecturer: Sanjeev Arora *Scribe: Neelesh Kumar, Yuting Wang*

2.1 Expressiveness of Small Neural Networks

The goal of this lecture is to understand what functions can small neural networks represent. More formally, we have a small neural network of s nodes, each of dimension d which take in input $X \in \mathbb{R}^d$ where each $\|X\|_2 = 1$. The network computes a function $f : \mathbb{R}^d \to \mathbb{R}$. The network also has some non linearity like ReLu in between. The problem that we are interested in is that given this network, what functions can be computed using it. Showing that an explicit learning problem *cannot* be computed by a suitably small net is a very difficult problem. Proving lower bounds on circuit size is related to P vs. NP for some discrete version of deep nets. Today we're interested in upper bounds: understanding sufficient conditions under which a function does have small nets. Even this is a difficult problem and we survey the little progress made so far. We discuss an old result called Barron's theorem.

Note that in all our discussion, the size of the network is determined by the number of hidden units in the network.

2.2 Upper bounds on the size of neural network

We turn our attention to upper bounds on the size of the neural network that is sufficient to compute a given function. There are two results that we will be proving in this lecture:

1. In general, a network of size $\exp(d)$ suffices to compute a "reasonably smooth" function f. We will formalize the notion of reasonable later.
2. Under some strong smoothness conditions, a smaller net with 2 layers is enough to compute the function. There are very few results of this type. We will be proving Barron's result [1] in this lecture.

Proving these results will give us insights into concepts like high dimensional geometry, concentration bounds and probabilistic analysis which will be useful later on in the course.

2.3 High Dimensional Geometry

In high dimensionsal spaces there are many points that are pretty far from each other. This observation can be formalized using the fact below.

Theorem 1. *$\exists\ exp(d)$ directions in $\mathbb{R}^d$ that make an angle $> 60^\circ$ with each other.*

Proof. We can prove the above fact using probabilistic method. Pick $2^{d/10}$ random unit vectors in $\mathbb{R}^d$. We are interested in determining the chance that they will make an angle $< 60^\circ$. For this we make use of the following fact: If u is a fixed unit vector and z is a random unit vector, then the inner product $\langle u, z\rangle \sim \mathcal{N}(0,\ 1/d)$
If the angle between fixed u and random z is 60°, then $\langle u, z\rangle = \cos(60^\circ) = 0.5$ which is $\Omega(\sqrt{d})$ standard deviations

away from the mean, making it a very improbable event (with p=$exp(-\frac{(\sqrt{d})^2}{2})$). Union bounding over all pairs of $2^{d/10}$ picked vectors proves the statement. □

The implication of the above statement is that in very high dimensions, there are lot of points pretty far away from each other. Thus even under calculus type conditions like continuity and differentiability the set of "degrees of freedom"is $\exp(d)$. We could construct functions that have any desired distinct values at these $\exp(d)$ points while changing gently between then.

2.4 Multivariate Fourier Transforms

As a build up to proving Barron's result, we discuss here Fourier transforms in higher dimensions. The goal of Fourier transforms is to express any function in terms of Fourier basis functions. We have a space of inputs which we call the X space, where $\|X\|_2 = 1$ and is picked according to some measure, e.g. distribution. The other space is the space of frequencies which we call w space, where $w \in \mathbb{R}^d$and has uniform measure on it.
Using Fourier transforms, we express a function $f : \mathbb{R}^d \to \mathbb{R}$ using basis functions $e^{i\langle w,x\rangle}$ as follows:

$$\widehat{f}(w) = \int f(x)e^{-i\langle w,x\rangle}\, dx \tag{2.1}$$

and the corresponding Fourier inversion is:

$$f(x) = \frac{1}{C}\int \widehat{f}(w)e^{i\langle w,x\rangle}\, dw \tag{2.2}$$

where C is some constant.
The Fourier coefficients satisfy Parseval's identity:

$$\int |\widehat{f}(w)|^2\, dw = 1 \tag{2.3}$$

It is natural to consider Fourier analysis as the branch of classical mathematics that is closest to machine learning. Informally, the objective in machine learning is to fit a model to the data by learning its parameters from the data. Likewise, the Fourier coefficients $\widehat{f}(w)$ can be estimated from the data, which can then be used to compute $f(x)$ using Fourier inversion. Under reasonable conditions, these fourier coefficients can be well-estimated from finite samples.

2.5 Barron's smoothness condition

Barron used the following notion of smoothness of functions in his result:

$$\int |w||\widehat{f}(w)|\, dw < C \tag{2.4}$$

where C is some constant.
We note that this is a fairly strong assumption. Informally, if we have $x_1, x_2, ...x_n$ where $\Sigma x_i^2 = 1$, then the sum of these numbers can be very large, with the maximum value being $\Sigma|x_i| \leq \sqrt{n}$. In fact, this quantity is expected to diverge, contrary to the assumption. Note that $|w|$ doesn't matter that much as it is very tightly concentrated. The smoothness condition, therefore is essentially a bound on $|\widehat{f}(w)|$. Somebody noted during lecture that Barron's condition says that the gradient of the fourier transform is absolutely integrable.

2.6 Barron's Theorem

Functions satisfying Barron's condition can be approximated by small depth 2 nets. Recall that deep nets have a nonlinear at the nodes, like sigmoid, ReLU etc. The result holds for any reasonable nonlinearity at the nodes, as explained later.

Theorem 2. *Any continuous function f that satisfies the smoothness condition stated in (2.5) can be approximated by a neural network g with single hidden layer and containing $\mathcal{O}(\frac{C^2}{\epsilon})$ hidden units such that $\forall x$ in the domain of f*

$$\mathbb{E}_{\mathbb{x}\sim\mu}[\|f(x) - g(x)\|^2] \leq \epsilon$$

where C is a constant and μ is the measure (distribution) from which x is picked.

Proof. For a class of functions F, the convex hull of F is : $conv(F) = \{g : g = \sum_{f\in F} \alpha_f f\}$, where $\alpha_f \geq 0$, and $\sum_{f\in F} \alpha_f = 1$.

Step 1: We first show that any continuous function that satisfies Barron's smoothness continuous function can be represented in the convex hull of:

$$conv\big(\frac{C}{\|w\|_2}\cos(\langle w, x\rangle + b)\big)$$

Using the inverse Fourier transform, we can write f as (up to some constant):

$$f(x) = \int \widehat{f}(w)e^{i\langle w,x\rangle}\, dw \tag{2.5}$$

Representing the Fourier coefficient $\widehat{f}(w)$ in polar form, Equation (2.5) can be rewritten as :

$$f(x) = \int |\widehat{f}(w)|e^{ib_w}e^{i\langle w,x\rangle}\, dw \tag{2.6}$$

Since f is real valued, we consider the real part of Equation (2.6). The imaginary part integrates out to 0. Consequently,

$$f(x) = \int |\widehat{f}(w)|\cos(\langle w, x\rangle + b_w)\, dw \tag{2.7}$$

Multiplying and dividing the expression by $\|w\|_2$ and C:

$$f(x) = \int \frac{|\widehat{f}(w)|\|w\|_2}{C} C\frac{\cos(\langle w, x\rangle + b_w)}{\|w\|_2}\, dw \tag{2.8}$$

The term $\int \frac{|\widehat{f}(w)|\|w\|_2}{C}\, dw$ evaluates to ≤ 1 as per the smoothness condition. Hence Equation (2.8) can be written as :

$$f(x) = \int C\, \frac{\cos(\langle w, x\rangle + b_w)}{\|w\|_2}\, dw \tag{2.9}$$

where $f(x)$ is represented in the convex hull.

Equation (2.9) implies that if we allow for only cosine non-linearities, then f can be computed by an infinite network as below:

$$f = \Sigma_w \alpha_w \cos(\langle w, x\rangle + b_w)\frac{C}{\|w\|_2} \tag{2.10}$$

Step 2: Next we show that f can be well-approximated by a finite net g. This involves a probabilistic sampling technique:

- Pick one w out of this infinite set of w's, where w is picked with probability α_w.
- Repeat the above steps r times.

Let g be approximation of f, computed by the sum of these finite samples. By definition of f in Equation (2.10), the expected value of picked gates, $\mathbb{E}[g] = f$.
The variance of g is:

$$\frac{1}{r}[\mathbb{E}(g^2) - \mathbb{E}(g)^2] \leq \frac{1}{r}(1 - \|f\|^2)$$

As a random variable, g approximates f within $1/r$ in expectation. This is possible only if there is an outcome of the random variable that achieves the bound. So, there must be a fixed $g_1, ..., g_r$ for which $\|g - f\|^2 \leq \frac{1}{r}(1 - \|f\|^2)$.

Further if we allow $r > \frac{4C^2}{\epsilon}$, then we can bound the error using Chebyshev's inequality resulting in:

$$\mathbb{E}_{\mathbb{x}\sim\mu}[\|f(x) - g(x)\|^2] \leq \epsilon$$

Step 3: So far we have constructed a finite net g where the nonlinearity is cosine. Barron's theorem says that any continuous function can be approximated by a linear combination of $\phi(w.x + b)$ where $\phi(.)$ is any non-linearity. Now we show how to allow other reasonable non-linearities. We show here the proof sketch of the above statement for sigmoid non-linearity. This consists of two steps:

1. We approximate the non-linearity that we used in steps 1 and 2 of proof, namely cosine, as step functions: $step(z) = 1_{\{z\geq 0\}}$. Since the cosine function is uniformly continuous in the domain of interest, it follows that it is uniformly well approximated by piecewise constant functions $h(z) = 1_{\{0\leq a1\leq z\leq a2\}}$ for any sequence of partitions of the domain into intervals of maximum width tending to zero. Such piecewise constant functions can be represented as linear combinations of unit step functions. The cosine function can then be written as the sum of unit step functions (Figure 2.1).

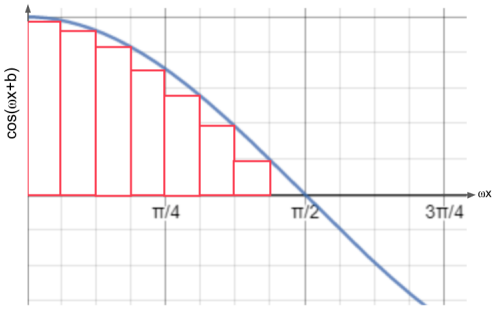

Figure 2.1: Approximation of cosine function using step function.

2. Next, we rewrite each step function using a standard non-linearity. Let's consider the sigmoid function, $\sigma(z) = 1/(1 + e^{-az})$, where a is a constant. By suitably modifying a, $\sigma(z)$ can be approximated as a step function. Figure 2.2 shows how this can be done.

□

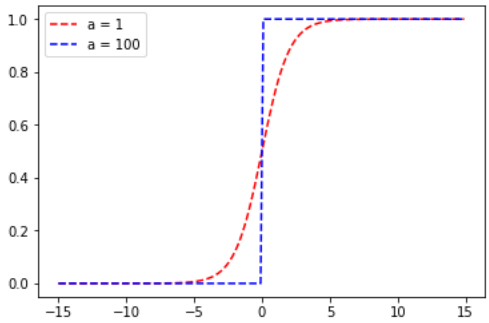

Figure 2.2: Approximation of step function using sigmoid function. The red curve shows the sigmoid function for $a = 1$. The blue curve is the sigmoid function for $a = 100$, which is an approximation of step function.

2.7 Universal approximation theorem

Barron's assumption of smoothness is pretty strong. In general, if f is continuous, a network with single hidden layer and $exp(d)$ number of units and containing non-linear activation function is sufficient to compute f. More formally, the theorem can be stated as:

Theorem 3. *Let $\phi(.)$ be a non-constant, bounded and continuous function. Let I_m denote the m-dimensional unit hypercube $[0, 1]^m$. The space of continuous functions on I_m is denoted by $C(I_m)$. Then, given any $\epsilon > 0$ and any function $f \in C(I_m)$, $\exists\ \alpha_i, b_i \in \mathbb{R}$ and $w_i \in \mathbb{R}^m$ such that :*

$$F(x) = \Sigma \alpha_i \phi(\langle w_i, x \rangle + b_i)$$

and

$$\mathbb{E}_{\mathbb{x}}[|F(x) - f(x)]^2] \leq \epsilon$$

$\forall x \in I_m$

The theorem was originally proved by Cybenko (1989) for sigmoidal activation functions[2]. Later, Hornik (1991) showed it to be true for any non-linear activation functions[3]. The proof is very similar to Barron's.
The universal approximation theorem demonstrates the capability of a shallow neural network to approximate any continuous function. However the theorem doesn't help us much in practice, because such a neural network could have very large number of hidden units.

References

1. Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory, 39(3), 930-945.

2. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

3. Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2), 251-257.