Mathematical Aspects of Deep Learning

Mathematics of Deep Learning: Lecture 3 – More on depth separation.

Transcribed by Paxton Turner (edited by Asad Lodhia, Elchanan Mossel and Matthew Brennan)

In the Boolean circuits setup, there is a separation result based on depth of the following form:

Theorem. *(Rossman, Servedio, Tan): There exists a function $f : \{0,1\}^n \to \mathbb{R}$ that can be computed by a linear size depth d formula using AND, OR, and NOT gates such that any depth $d-1$ circuit that agrees with f on $1/2 + o(1)$ inputs is of size $\exp(n^{\Omega(1/d)})$*

Question: Can we prove a similar result for deep nets?

Last time we discussed the following separation result.

Theorem. *(Eldan, Shamir): There exists a probability measure on $\mathbb{R}^d$ and $g : \mathbb{R}^d \to [-2,2]$ which is:*

- *supported on $\{x : ||x||^2 \leq Cd\}$ and expressible by a 3-layer network of width $\mathrm{poly}(d)$.*
- *for all f supported by a 2-layer network of width $\exp(\epsilon d)$ it holds that*

$$\mathbb{E}_\mu[|f-g|^2] \geq \epsilon.$$

Today we will discuss a more elegant treatment of the 2 vs. 3 separation result which yields slightly stronger results.

Loading [MathJax]/extensions/MathMenu.js

Danielly's Model

Danielly's Model is formed as follows. Let x, x' chosen uniformly (i.e., with respect to the Haar measure) from S^{d-1} and let $f(x, x') = g(\langle x, x' \rangle)$. The nets of depth ℓ are of the form:

$$\begin{array}{cc} \ell = 2 & w_2^{\mathrm{T}}\sigma(W_1 x + W_2 x' + b_1) + b_2 \\ \ell = 3 & w_3^{\mathrm{T}}\sigma(W_2\sigma(W_1 x + W_2 x' + b_1) + b_2) + b_3 \\ \vdots & \vdots \end{array}$$

on the input (x, x'). Here, we prove the following theorem.

Theorem. *If $g : [-1, 1] \rightarrow [-1, 1]$ is L-Lipschitz, then there exists a 3-layer representation F of f with width $O(d^2 L/\epsilon)$ and weights bounded by $O(1 + L)$ and $||f - F||_\infty \leq \epsilon$.*

This illustrates the "magic of 3." Intuitively, we see the power of quadratic models over linear ones, at least for this specific context.

We begin with the following lemma which is of the flavor of universality of depth 1 nets:

Lemma. *If $f : [-R, R]$ is L-Lipschitz for all $\epsilon > 0$, then f can be approximated with $\beta_i \leq R, \alpha_i \leq 2L$, and $\gamma_i \in \{-1, 1\}$ as follows:*

$$\left\| f(x) - f(0) - \sum_{i=1}^{m} \alpha_i \sigma(\gamma_i x - \beta_i) \right\|_\infty \leq \epsilon, \; m \leq \frac{2RL}{\epsilon}$$

Proof of Lemma (Sketch): Repartition the interval, and note that the Lipschitz condition implies the graph doesn't deviate too far from a straight line.

Proof of the Theorem: We will prove the theorem in the following section and then show a corresponding lower bound.

1. Upper bound

Loading [MathJax]/extensions/MathMenu.js

We first show that the function can be approximated well by a depth 3 net:

$$\langle x, x'\rangle = \left\|\frac{x+x'}{2}\right\|_2^2 - \frac{1}{2} = \sum_{i=1}^{d} \left(\frac{x+x'}{2}\right)^2,$$

which implies that a 2-layer network approximates $\langle x, x'\rangle$ well. We can also compute $\sigma(\langle x, x'\rangle)$ as a linear combination of the previous layers. This in combination with the lemma implies the desired result.

2. Lower bound

We want to show that a depth two net that approximates the function well is very wide. For the analysis it will be useful to answer the following question:

Question: What is the distribution of $\langle x, x'\rangle$?

Approximation: Rotation invariance implies that

$$D(\langle x, x'\rangle) = D'(\langle x, r\rangle) = D(\langle x, e_1\rangle) = D(x_1)$$

Note that this is also true if x' is deterministic and x is random.

The individual components $\langle x, e_k\rangle$ of a uniform random vector in $x \in S^{d-1}$ approaches a Gaussian as $d \to \infty$. Indeed,

$$\begin{aligned} d\mu_d(x) \quad &= \mathrm{Vol}\left((1-x^2)^{1/2} S^{d-2}\right) \\ &= \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)} \left(1-x^2\right)^{\frac{d-3}{2}} \end{aligned}$$

which is a t-statistic distribution (and as $d \to \infty$ converges to a gaussian). The same reasoning applies also if one of the vectors x, x’ is fixed. Therefore, if $g = \psi(\langle v, x\rangle, \langle v', x'\rangle)$, then

$$||g||_2^2 = ||\psi||_{L^2(\mu_d \times \mu_d)}^2,$$

Loading [MathJax]/extensions/MathMenu.js

and if $f = \phi(\langle x, x' \rangle)$, then

$$||f||_2^2 = ||\phi||_{L^2(\mu_d)}^2 = \int_{S^{d-1} \times S^{d-1}} f^2(x, x')\, d(m \times m),$$

where $d(m \times m)$ indicates the Haar measure.

Orthogonal Polynomials: Define $P_0 = 1, P_1 = x$, and

$$P_n = \frac{2n + d - 4}{n + d - 3} \cdot P_{n-1}(x) - \frac{n - 1}{n + d - 3} \cdot P_{n-2}(x)$$

We observe that

$$||P_n||_2^2 = 1/N_{n,d} \quad \text{where} \quad N_{n,d} = \frac{(2n + d - 2)(n + d - 3)!}{n!(d - 2)!}$$

and that $||P_n||_\infty = 1$. Now define

$$h_n(x, x') = \frac{P_n(\langle x, x' \rangle)}{\sqrt{N_{n,d}}} \quad \text{and} \quad L_n^x(x) = h_n(x, x).$$

The key fact we use is that

$$\langle L_i^v, L_j^v \rangle = \mathbb{E}[L_i^v(x) L_j^{v'}(x)] = \mathbb{E}[h_i(\langle v, x \rangle) h_j(\langle v', x \rangle)] = \delta_{i,j} P_i(\langle v, v' \rangle),$$

which implies

$$\mathbb{E}[h_n(x, x') L_i^v(x) L_j^{v'}(x)] = 1(i = j = n) N_{n_1}^{-1/2} P_n(\langle v, v' \rangle).$$

To flesh out the above more carefully,

$$\begin{aligned} \mathbb{E}[h_n(x, x') L_i^v(x) L_j^{v'}(x')] &= \mathbb{E}_x \left[L_i^v(x) \mathbb{E}_{x'}[h_n(x, x') L_j^{v'}(x')] \right] \\ &= \mathbb{E}_x [L_i^v \delta_{j=n} P_n(v', x)] \\ &= \mathbb{E}_x \left[L_i^v(x) \frac{L_j^{v'}(x)}{\sqrt{N_{n,d}}} \right] \end{aligned}$$

Loading [MathJax]/extensions/MathMenu.js

Lower Bound Idea: We now outline how to obtain the desired lower bound. Expand $g(x, x') = \sum \alpha_i h_i(x, x')$ and choose f that is "noise sensitive", i.e., has a lot of mass at higher Fourier levels. For the rest of the argument assume that $f = h_n$ for a large n (though h_n is not a Lip function, but we can find a Lip function that has a lot of weight on high levels).

We want $||g - \sum_{j=1}^{m} g_j||_2^2$ to be large when $g_j = b_j \sigma(\langle v_j, x\rangle) \geq c$ unless either b_j is large. Otherwise, there exists j such that $\langle g, g_j\rangle \geq \frac{c}{m}$. Note that if h_n is close to $\sum_{j=1}^{m} g_j$, then we need to have at least one j where $\langle g_j, f\rangle \geq \frac{||f||_2^2}{4m}$.

We now compute $\langle g_j, f\rangle$. Observe that

$$g_j = \sum \beta_{k,\ell} L_k^{v}(x) L_\ell^{v'}(x')$$

$$g_j(x, x') = \psi(\langle v, x\rangle, \langle v', x'\rangle)$$

$$\sum \beta_{k,\ell}^2 = ||g_j||_2^2.$$

Therefore,

$$|\langle g_j, h_n\rangle| = N_{n,d}^{-1/2} |\beta_{n,n} P_n(\langle v, v'\rangle)| \leq N_{n,d}^{-1/2} |\beta_{n,n}|$$

so m is very large or $||g_j^2||_2^2$ is very large. Informally, we get a bad approximation of $h_n(\langle x, x'\rangle)$ unless

- Width $\geq N_{n,d}^{\Omega(1)}$ or
- Max-weight $\geq N_{n,d}^{\Omega(1)}$

Open Questions: 1.Can it really be that large weights solve the approximation problem? 2. Instead of working with the Haar measure, can this computation be tly? We lose rotation invariance but it is a bit

Loading [MathJax]/extensions/MathMenu.js

strange we work with measure that are in some sense getting close to Gaussian measure and cannot prove it directly in Gaussian space.

Kane-Williams Model

Threshold gates are given by

$$\mathbf{1}\left\{\sum w_i x_i \geq t\right\}.$$

Theorem. *There exists a function* $A_n : \{0,1\}^n \to \{0,1\}$ *that can be computed*

- *Using a 3-layer threshold circuit with* $O(n)$ *gates,*
- *but requires at least* $\Omega(n^{3/2})$ *gates to be computed correctly on* $.5 + \epsilon$ *of inputs by a 2-layer circuit.*

The basic unit threshold is

$$f(x) = \operatorname{sgn}(w_0 + \sum_{i=1}^{n} w_i x_i).$$

We assume a distribution that is uniform over $\{0,1\}^n$.

Meta-Question: Can we use reduction between models (eg, the previous result) to get better separation in the binary case?

Partial answer: Due to complexity-theoretic issues, better than $\operatorname{poly}(n)$ isn't possible.

Question: Can we replace new activations into the previous result (or this one)?

Answer: Isn't written down, but should be straightforward.

Proof. First define the function $A_n = A_{2^{k+1}}$.

$$M(x_1, \ldots, x_{2^k}, a_1, \ldots, a_k) = x_{a_1} \ldots x_{a_k}$$

Loading [MathJax]/extensions/MathMenu.js

is a Mux function generalization. Next,

$$A_n(x_1, \dots, x_{2^k}, a_1, \dots, a_{2^k}) = M(x_1, \dots, x_{2^k}, \oplus_{i=1}^{2^k/k} a_i, \oplus_{i=2^k/k+1}^{2 \cdot 2^k/k}, \dots),$$

where $\oplus_{i=1}^{2^k/k} a_i$ indicates the parity of the slice of the string.

Claim 1: M can be computed depth 2 using $O(n \log n)$ gates.

Claim 2: Parity can be computed using threshold gates.

Proof of Claim 2: Let

$$\sum x_i \geq 1 \to b_1$$

$$\vdots$$

$$\sum x_i \geq n \to b_n.$$

It can be shown that there exist weights $w_1, \dots w_n$ such that

$$\text{sgn}(\sum w_i b_i) = \oplus x_i.$$

In fact, this can be done with $1 \leq w \leq \text{poly}(k)$, though we don't show this here.

Lower bound ingredients: We now move onto the lower bound.

Lemma: For all but ϵ of n-bit Boolean functions f, there is no depth 2 network of size less than $o(\epsilon^2 2^n/n)$ that agrees with f on more than $.5 + \epsilon$ of the inputs.

Proof sketch of lemma: The number of threshold functions is less than $2^{O(n^2)}$, while the total number of functions is 2^{2^n}. That is, there aren't too many threshold functions in general. Also, for most Boolean functions f on $\{0, 1\}^n$, a 2-layer network of size $O(\frac{\xi 2^n}{n^2})$ does not approximate f well.

Loading [MathJax]/extensions/MathMenu.js

Returning to the proof of the lemma, the idea is to generate a random function, allowing us to use the lemma's statement that most functions do not approximate f well. The construction is:

- Pick x randomly.
- In each block of a_i, fix all coordinates but one randomly.
- This generates a random function on k variables.

Question: How does a 2-layer look like after fixing randomly almost all bits?

Intuitive answer: Once enough bits are fixed, most gates become constant and die. Then we're left with a smaller network which can't approximate f well.

Main Lemma: A random restriction of a threshold gate will result in a constant function with probability at least

$$1 - O(\log n / n^{1/2}).$$

Thus, the number of gates that remain is $O(n/\log^2(n))$, which is not enough to represent a random function.

■

PAC Learning

We close this lecture with the definition of **PAC-learning** . We have a space X_n and are interested in functions $f : X_n \to \{0, 1\}$. Let C be the class of such functions (eg, 2-layer networks).

Definition. The class C is PAC-learnable if for all ϵ, δ and for all distributions D on X_n, there exists an algorithm A such that given $\text{poly}(n, 1/\epsilon, 1/\delta)$ samples

$$(x_i, y_i);\ x \sim D;\ y_i = f(x_i);\ f \in C,$$

with probability larger than $1 - \delta$, A returns a function $h : X_n \to \{0, 1\}$ such

Loading [MathJax]/extensions/MathMenu.js

$$\mathbb{P}[h(x) \neq f(x)] \leq \epsilon$$

and A runs in poly$(n, 1/\epsilon, 1/\delta)$.

Here, δ is interpreted as a *global parameter* , and ϵ is the *probabilistic error* .

elmos / April 14, 2017

Mathematical Aspects of Deep Learning / Proudly powered by WordPress

Loading [MathJax]/extensions/MathMenu.js