Lecture 11: Depth separations

Announcements

* Hwk1 due tomorrow.
* Opt starting on Thursday!...
* Typed notes Sunday.

Last lecture: only lecture so far with depth ≥ 2

Remark/reminder: little emphasis on depth in course material since all results (except for these two lectures) worsen, whereas in practice they improve.

Last lecture:

* "Triangle mapping" $\Delta = \begin{cases} 2x & x \in [0, 1/2) \\ 2(1-x) & x \in [1/2, 1] \\ 0 & \text{o.w.} \end{cases}$

* "viral fractal property"

  $\Delta^L = 2^{L-1}$ shrunken copies of Δ;

  $f \circ \Delta^L$ where $f(1-z) = f(z)$ for $z \in [0,1]$
  $\Rightarrow 2^L$ shrunken copies of f

* x^2 can be efficiently approximated via "a few" triangles:
  affine interpolation h_i of x^2 on $[0,1]$ using $2^i + 1$ interpolation points can be written with $O(i)$ ReLU layers, and $\sup_{x \in [0,1]} |x^2 - h_i(x)| \leq \frac{1}{4^{i+1}}$.

Didn't do last time (topic for today):

necessity of depth.

Two theorems on necessity of depth:

Theorem. $\forall L \geq 2$,
$\forall g: \mathbb{R} \to \mathbb{R}$ ReLU networks
with $\leq 2^L$ nodes, $\leq L$ layers

$$\int_0^1 \left| g(x) - \Delta^{L^2+2}(x) \right| dx \geq \frac{1}{32},$$

where Δ^{L^2+2} has $\leq 3(L^2+2)$ nodes
$\leq 2(L^2+2)$ layers.

Words. For any depth L width $\leq 2^L$
$\exists$ $O(1)$ width L^2 depth function
you can't approximate.

Theorem. $\forall L \geq 1$, $\forall N \geq 1$
$\forall g: \mathbb{R} \to \mathbb{R}$ ReLU networks of
width $\leq N$, depth $\leq L$,

$$\int_0^1 \left(x^2 - g(x)\right)^2 dx \geq \frac{1}{5760\left(\frac{2N}{L}\right)^{4L}}.$$

Words: If we fix depth L and increase N, error can't go down faster than "polynomially": $\frac{1}{(N)^{O(1)}}$,

whereas if we choose $N = L = O(\ln(1/\varepsilon))$, get error ε.

Remark (L_1 norm.) For upper bounds, uniform norm makes sense $\Rightarrow$ can do well on any data distribution.

For lower bounds, L_1 makes sense $\Rightarrow$ do poorly on any "spread out" distribution.

Also, as in lectures 1-2, these choices affect tractability.

Proof scheme for both.

① Prove ReLU networks of small depth have few affine pieces.

② Use region counting argument to show poor approximation

* For approximating Δ^{L^2+L}, must exist many regions where approximant is a single affine function, & Δ^{L^2+L} oscillates a lot.
* For approximating x^2, must " " " " " " " " " " , & x^2 is highly curved.

<u>Remark.</u> <u>All</u> proofs use region counting. This limitation may be related to the lack of stronger theorems using dim > 1.

We'll establish "fixed depth has few affine pieces" via 2 lemmas.

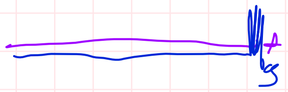
Definition. $N_A(f)$ denotes the cardinality of the smallest partition of $\mathbb{R}$ into intervals such that f is affine within each interval, or ∞ if no such partition exists.

Examples: $N_A(x \mapsto \max\{0, x\}) = 2.$ [a single relu]

$N_A(x \mapsto \max\{0, x\} - \max\{0, -x\}) = 1$ [identity]

Remark. Can abstractly view $N_A(f)$ as a "complexity measure"

This alone does not prove depth separation in L_1:



$N_A(f) = 1$, $N_A(g)$ = arbitrarily large

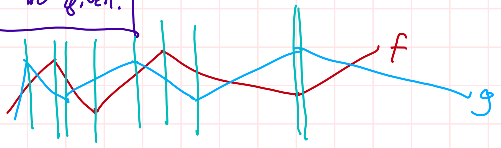
$\int_0^1 |f(x) - g(x)|\,dx$ = arbitrarily small.

Lemma. Let $f, g, (g_1, \dots, g_k)$, $(a_1, \dots, a_k, b)$ be given.

1. $N_A(f+g) \le N_A(f) + N_A(g)$. not tight
2. $N_A(\sum_{i=1}^k a_i g_i + b) \le \sum_{i=1}^k N_A(g_i)$.
3. $\boxed{N_A(f \circ g) \le N_A(f)\, N_A(g).}$
4. $N_A(x \mapsto f(\sum_{i=1}^k a_i g_i + b)) \le N_A(f) \sum_{i=1}^k N_A(g_i)$.

Remark: purest (?) form of power of depth in these lectures (?). Rare for this inequality to be almost tight; captures part of why Δ is special.

Proof: (1)



$f+g$ is affine between adjacent changepoints, & there are $\le N_A(f) + N_A(g) - 1$ changepoints.

(2) Proof by induction, noting $N_A(ag) \le N_A(g)$ and $N_A(g+b) = N_A(g)$.
can be loose if $a=0$ & $N_A(g) > 1$.

(3) Define $P_A(g)$ = pieces of g (in a smallest partition), and consider a single piece U:

U interval $\Rightarrow$ $g(U)$ interval

$\Rightarrow (f \circ g)_{|U} = f_{|g(U)} \Rightarrow (f \circ g)_{|U}$ has $\le N_A(f)$ pieces

$$N_A(f \circ g) \le \sum_{U \in P_A(g)} N_A((f \circ g)_{|U}) \le \sum_{U \in P_A(g)} N_A(f_{|g(U)})$$

$$\le \sum_{U \in P_A(g)} N_A(f) \le N_A(g) \cdot N_A(f).$$

(4) Combination of (2) & (3).

Via induction, that lemma implies the following lemma.

Lemma. Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network of widths $(m_1, \dots, m_L)$, $m_L = 1$, & $B = \sum_i m_i$.

Then $N_A(f) \le \left(\frac{2B}{L}\right)^L$.

Proof idea. Proceed inductively over nodes of the network, using previous lemma.

Reminder. Δ^{L^2+2} has $N_A(\Delta^{L^2+2}) = 2 \cdot (2^{L^2+2-1}) + 2$ affine pieces

and here we're saying depth $\le L$ nodes $\le B$ $\Rightarrow N_A(\cdot) \le \left(\frac{2B}{L}\right)^L$.

Open problems

1. Prove or disprove a near initialization embedding for Δ^k.
2. Low norm approximation (in what norm? Both for network & for target).
3. Other architectures (anything modern: attention, ...)
4. Characterize multivariate multilayer approximation (e.g., like Δ)
5. Depth L vs $L+1$ in depth separation.