Lecture 2: constructive 3-layer, non-constructive 2-layer apx

1. Basic approximation results (continued)

Here’s our three layer approximation theorem.

[We stated this last class, sketched a picture proof.]

**Theorem.** Let \( \text{cont} f \) and \( \delta > 0 \) and ReLU \( \sigma \) be given, There exists \( g(x) := W_3 \sigma(W_2 \sigma(W_1 x + b_1)) \) of width \( \mathcal{O}(d/\delta^s) \) with \( \| f - g \|_1 = \int_{[0,1]^d} |f(x) - g(x)| \, dx \leq 2\omega_f(\delta) \).

**Remarks.**

- The curse of dimension (exponential in \( d \)) is there again.
- The metric has been relaxed to \( L_1 \); the same proof can do \( \| \cdot \|_u \), but it takes a lot of care, and we’ll have better constructions anyway.
- The construction will have an unbounded Lipschitz constant. It seems unlikely gradient descent can learn such representations.

**Proof (continued).** Let’s do what we did in the univariate case, putting nodes where the function value changes. For each \( R_i := \times_{j=1}^d [a_j, b_j] \), pick \( \gamma > 0 \)

\[
g_{\gamma,j}(x) = \sigma \left( \frac{z - (a_j - \gamma)}{\gamma} \right) - \sigma \left( \frac{z - a_j}{\gamma} \right) - \sigma \left( \frac{z - b_j}{\gamma} \right) + \sigma \left( \frac{z - (b_j + \gamma)}{\gamma} \right)
\]

and \( g_\gamma(x) = \sigma(\sum_j g_{\gamma,j}(x_j) - (d - 1)) \) (adding the additional ReLU layer is the key step!), whereby

\[
g_\gamma(x) = \begin{cases} 1 & x \in R_i, \\ 0 & x \not\in \times_j[a_j - \gamma, b_j + \gamma], \\ [0,1] & \text{otherwise}. \end{cases}
\]

Since \( g_\gamma \to \mathbb{1}_{R_i} \) pointwise, there exists \( \gamma \) with \( \| g_\gamma - \mathbb{1}_{R_i} \|_1 \leq \frac{\omega_f(\delta)}{\sum_j |a_j|} \).
2. Non-constructive approximation with 2 layers
The previous section developed $g_\gamma$ such that

$$g_\gamma(x) \approx \mathbb{1}_{x \in \times_i [a_i, b_i]}.$$ If deep networks could multiply, we could do

$$x \mapsto \prod_i \mathbb{1}_{x_i \in [a_i, b_i]}.$$ Who thinks deep networks can multiply?

**Multiplication with shallow networks**

Let's introduce some notation:

$$H_{\sigma,d} := H_{\sigma} := \left\{ x \mapsto \sigma(a^T x + b) : (a,b) \in \mathbb{R}^{d+1} \right\},$$

$$\text{span}(\mathcal{F}) := \left\{ x \mapsto \sum_{i=1}^N \alpha_i f_i(x) : N \in \mathbb{Z}_{\geq 0}, \alpha_i \in \mathbb{R}, f_i \in \mathcal{F} \right\}.$$ Let's consider $H_{\cos}$. Since

$$2 \cos(y) \cos(z) = \cos(x + y) + \cos(x - y),$$

$$\cos(a^T x + b) \cos(r^T x + s) = \frac{1}{2} \left( \cos((a + r)^T x + (b + s)) + \cos((a - r)^T x + (b - s)) \right),$$

Therefore $\text{span}(H_{\cos})$ is closed under products!

**Remark.** $\text{span}(H_{\sigma})$ denotes single hidden layer networks; second bias unneeded since $H_{\sigma}$ includes constant functions.

$H_{\cos}$ is closed under multiplication

Note that this gives us "bumps" via

$$x \mapsto \prod_{i=1}^d \cos(x_i),$$

and we can linearly combine bumps to get continuous functions.

Where does this leave us?

- **Polynomials** are also closed under addition and multiplication, and they are universal approximators (Weierstrass 1885).

- An extended version, the "Stone-Weierstrass theorem", says "polynomial-like" classes of functions are also universal approximators.

- These "polynomial-like" properties are satisfied by $\text{span}(H_{\cos}).$

- Since $\cos$ can be approximated by $\text{span}(H_{\sigma,1})$, we also have $\text{span}(H_{\sigma}) \approx \text{span}(H_{\cos}) \approx \text{cont}.$

**Stone-Weierstrass theorem**

Any class of functions that has multiplication behaves like polynomials, and has nice interpolation properties.

**Theorem** (Stone-Weierstrass; (Folland 1999, Theorem 4.45)). Let functions $\mathcal{F}$ be given as follows.

1. Each $f \in \mathcal{F}$ is continuous.
2. For every $x$, there exists $f \in \mathcal{F}$ with $f(x) \neq 0$.
3. $\mathcal{F}$ separates points, meaning for every $x \neq x'$ there exists $f \in \mathcal{F}$ with $f(x) \neq f(x').$
4. $\mathcal{F}$ is closed under multiplication and vector space operations ($\mathcal{F}$ is an algebra).

Then for every continuous $g : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon > 0$, there exists $f \in \mathcal{F}$ with $\|f - g\|_u \leq \epsilon$. ($\mathcal{F}$ is universal.)
Remarks on Stone-Weierstrass.

- It is heavyweight, but a good tool to have.
- Proofs are not constructive, but seem to require size $O(1/\epsilon^u)$.
- Proofs are interesting:
  - We will revisit the standard one due to Bernstein, which picks a fine grid and interpolating polynomials that are well-behaved off the grid.
  - Weierstrass's original proof convolved the target with a Gaussian, which makes it analytic, and also leads to good polynomial approximation.
- As a technical point, we could also approximately satisfy the properties, and apply the theorem to the closure of $\mathcal{F}$.
- The second and third conditions are necessary; if there exists $x$ so that $f(x) \neq 0 \forall f \in \mathcal{F}$, then we can't approximate $g$ with $g(x) = 0$; if we can't separate points $x \neq x'$, then we can't approximate functions with $g(x) \neq g(x')$.

Universal approximation via Stone-Weierstrass.

**Lemma (Hornik, Stinchcombe, and White 1989).** $\text{span}(\mathcal{H}_{\cos})$ is universal.

**Proof.** Let's check the Stone-Weierstrass conditions:

1. Each $f \in \text{span}(\mathcal{H}_{\cos})$ is continuous.
2. For each $x$, $\cos(0^T x) = 1 \neq 0$.
3. For each $x \neq x'$, $f(z) := ((z - x')^T (x - x')/\|x - x'\|^2) \in \mathcal{H}_{\cos}$ satisfies $f(x) = \cos(1) \neq \cos(0) = f(x')$.
4. $\text{span}(\mathcal{H}_{\cos})$ is closed under products and VS ops as before.

Arbitrary activations

**Theorem (Hornik, Stinchcombe, and White 1989).** Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous, and

$$\lim_{z \to -\infty} \sigma(z) = 0, \quad \lim_{z \to +\infty} \sigma(z) = 1.$$  

Then $\text{span}(\mathcal{H}_{\sigma})$ is universal.

**Proof.** Given a continuous function $g$, pick $f \in \text{span}(\mathcal{H}_{\cos})$ (or $\text{span}(\mathcal{H}_{\exp})$) with $\|f - g\|_u \leq \epsilon/2$. Then use the earlier univariate approximation results to replace each cos (or exp) with elements of $\text{span}(\mathcal{H}_{\sigma})$. (Details will be in homework.)

**Remarks.**

- ReLU is fine: use $z \mapsto \sigma(z) - \sigma(z - 1)$ and split nodes.
- exp didn't need bias, but ReLU apx of exp needs bias.
- Weakest conditions on $\sigma$ (Leshno et al. 1993): universal apx iff not a polynomial
- Never forget: curse of dimension (size $O(1/\epsilon^n)$).

If you don’t like $\cos$...

**Lemma (Hornik, Stinchcombe, and White 1989).** $\text{span}(\mathcal{H}_{\exp})$ is universal.

**Proof.** Let’s check the Stone-Weierstrass conditions:

1. Each $f \in \text{span}(\mathcal{H}_{\exp})$ is continuous.
2. For each $x$, $\exp(0^T x) = 1 \neq 0$.
3. For each $x \neq x'$,

$$f(z) := \exp((z - x')^T (x - x')/\|x - x'\|^2) \in \mathcal{H}_{\exp} \text{ satisfies } f(x) = \exp(1) \neq \exp(0) = f(x').$$
4. $\text{span}(\mathcal{H}_{\cos})$ is closed under VS ops by construction; for products,

$$\left(\sum_{i=1}^n r_i \exp(a_i^T x)\right) \left(\sum_{j=1}^m s_j \exp(b_j^T x)\right) = \sum_{i=1}^n \sum_{j=1}^m r_is_j \exp((a + b)^T x).$$

**Remark.** Biases? $x \mapsto \exp(a^T x + b) = e^b \cdot \exp(a^T x)$. 

$$\text{span}(\mathcal{H}_{\exp})$$
Other universal approximation proofs

▶ (Cybenko 1989) Assume contradictorily you miss some functions. By duality, \( 0 = \int \sigma(a^T x - b) \, d\mu(x) \) for some signed measure \( \mu \), all \((a, b)\). Using Fourier, can show this implies \( \mu = 0 \ldots \)

▶ (Leshno et al. 1993) If \( \sigma \) a polynomial, \ldots; else can (roughly) get derivatives and polynomials of all orders (we'll have homework problems on this).

▶ (Barron 1993) Use inverse Fourier representation to construct an infinite-width network; we'll cover this in class. It can beat the worst-case curse of dimension!

▶ (Funahashi 1989) I'm sorry, I haven't read it. Also uses Fourier.

References I


References II


http://dblp.uni-trier.de/db/journals/nn/nn6.html#LeshnoLPS93.