Universal Approximation

Let's work with the basic model

\[ f(x; w) = \sigma_L (w_L \sigma_{L-1} (\ldots \sigma_1 (w_1 x + b_1) \ldots )) \]

where:

1. each \( w_i \in \mathbb{R}^{n_i \times n_{i-1}} \) and \( b_i \in \mathbb{R}^{n_i} \)

2. each \( \sigma_i \) is a coordinate-wise nonlinear activation function

   e.g. rectified linear unit (ReLU)

   \[ \sigma_i(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases} \]
Main Question: what class of functions can we express with deep networks?

Answer: Everything (that is continuous)

def: we say a class of functions $\mathcal{F}$ is a universal approximator if for any continuous function $g$ and compact domain $D$ and any $\varepsilon > 0$, \( \exists f \in \mathcal{F} \)
\[ |f(x) - g(x)| \leq \varepsilon \quad \forall x \in D \]

From analysis you might remember:

Theorem: Polynomials (with unbounded degree) are universal approximators
Deep networks give us an alternative to polynomials.

**Meta Theorem:** For many activation functions, depth two networks of unbounded width are universal.

e.g. we will prove it for ReLUs, and Hornik et al showed it holds for any continuous \( \sigma \) with

\[
\lim_{z \to -\infty} \sigma(z) = 0 \quad \text{and} \quad \lim_{z \to +\infty} \sigma(z) = 1
\]

which covers sigmoids.

**Univariate Approximations**

**Def:** A function \( g : \mathbb{R} \to \mathbb{R} \) is \( p \)-Lipschitz if

\[
\forall x, y \quad |g(x) - g(y)| \leq p |x - y|
\]

Any Lipschitz function can be approximated arbitrarily well by step functions.
Proof by picture:

\[ \text{Multivariate Approximations} \]

The approach in higher dimensions is also based on creating a "bump".

**Proposition:** Let \( g : \mathbb{R}^d \to \mathbb{R} \) be continuous and suppose \( \varepsilon, \delta > 0 \) s.t.

\[ \| x - y \|_\infty \leq \delta \Rightarrow | g(x) - g(y) | \leq \varepsilon \]

\( \exists \) a depth two network (with cosine act.) \( f \) with \( \frac{C}{\delta^d} \) hidden units and satisfies

\[ \sup_{X \in [0,1]^d} | f(x) - g(x) | \leq \varepsilon \]

A better question is:

What kinds of functions can we **succinctly** express with deep nets?
Barron’s Theorem

Barron’s seminal work identified a rich class of functions that you can approximate by a small depth two network.

For a function \( f: \mathbb{R}^d \rightarrow \mathbb{R} \) we write its Fourier transform as

\[
\hat{f}(\omega) \triangleq \int e^{-2\pi i \omega^T x} f(x) \, dx
\]

Note that \( \hat{f}: \mathbb{R}^d \rightarrow \mathbb{R} \) too. This transform is invertible in the sense that

\[
f(x) = \int e^{2\pi i \omega^T x} \hat{f}(\omega) \, d\omega
\]

under technical conditions on \( f, \hat{f} \) (e.g. both are in \( L_1 \)).

Now we can state Barron’s Theorem:
Theorem: For a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, $f$ is a depth two network with $k$ hidden units that computes $g$

$$\int |f(x) - g(x)|^2 dx \leq \frac{4\, C_f}{k}$$

where $C_f = \| \hat{\nabla} f \|_1^2 = \left( \int \| \hat{f}(w) \| dw \right)^2$

"the square integral of the Fourier transform of the gradient of $f$"

A modern framing of the proof is:

1. rewrite the Fourier inversion formula as a depth two infinite width network

2. sample from this representation to find a good succinct approximation

we will start from 2, which is called Maurey's Lemma
Theorem: Suppose \( X \in \text{conv}(S) \) is a vector of dimension \( d \), possibly infinite, set of vectors of dimension \( d \), possibly infinite.

Then \( \exists \) a convex combination

\[
\| X - \frac{1}{k} \sum_{i=1}^{k} \alpha_i V_i \| \leq \frac{\sup_{V \in S} \| V \|}{k}
\]

where \( V_i \in S \)

Actually, this theorem holds even in infinite dimensions (i.e., \( x \) and \( V_i \)'s are functions).

Proof: If \( x \in \text{conv}(S) \) then there is a r.v. \( V \) supported in \( S \) with

\[
X = \mathbb{E}_{\mu}[V]
\]

Let \( V_1, \ldots, V_k \) be iid draws of \( V \). Then

\[
\mathbb{E} \left[ \| X - \frac{1}{k} \sum_{i=1}^{k} V_i \|^2 \right] = \mathbb{E} \left[ \| \frac{1}{k} \sum_{i=1}^{k} X - V_i \|^2 \right]
\]

\[
\leq \mathbb{E} \left[ \| V \|^2 \right] \leq \frac{\sup_{V \in S} \| V \|^2}{k}
\]
is based on massaging the Fourier inversion formula, we get:

\[ f(x) - f(b) = - \frac{\|x\|}{\|w\|} \int_{0}^{\|w\|} \int_{0}^{\|w\|} \frac{\sin(2\pi b + 2\pi \Theta(w))}{\|w\|} \|\hat{f}(w)\| \, db \, dw \]

\[ + \frac{\|x\|}{\|w\|} \int_{-\|w\|}^{0} \int_{0}^{\|w\|} \frac{\sin(2\pi b + 2\pi \Theta(w))}{\|w\|} \|\hat{f}(w)\| \, db \, dw \]

where \( \hat{f}(w) = |\hat{f}(w)| e^{2\pi i \Theta(w)} \), i.e. \( \Theta(w) \) is the angle.

Why is this useful?

It is an infinite width, depth two rep. that we can sample from to get a good approx.
Depth Separations

so far, we’ve identified interesting classes of functions that can be well-approximated by shallow (depth=2) networks

Main Question: Does increasing the depth buy you expressivity?

Ideally, we want an algorithmic answer:

“I can solve problem A with SGD on a depth $d$ network, but not on a depth $d-1$ network”

We’ll settle for the more modest goal of finding explicit functions that can be that can be represented w/ depth $L$, but not $L-1$

These are called depth separations
Main Theorem [Telgarsky] There is a function $f$ that can be computed by a ReLU network with:

\[ 0 \leq s(x) \leq 1 \]
\[ \text{depth} = O(L^2) \]
\[ \# \text{units} = O(L^2) \]

but for any function $g$ computed with:

\[ \text{depth} \leq L \]
\[ \# \text{units} \leq 2^L \]

we must have:

\[ \int_{[0,1]} |f(x) - g(x)| \, dx \geq \frac{1}{32} \]

Recall, you can approximate any continuous function arbitrarily well by a depth two network, but: you can and do need exponentially many more units.
Intuition: A ReLU network computes a function made up of many flat regions and:

- flat regions grow polynomially with width, but exponentially in depth

Let’s see a simple construction where the flat regions is exponential:

\[
m(x) = \begin{cases} 
2x & \text{if } 0 \leq x \leq \frac{1}{2} \\
2-2x & \text{if } \frac{1}{2} < x \leq 1 \\
0 & \text{else}
\end{cases}
\]

What happens if we iterate \( m(x) \)?

i.e.

\[
m(m(x)) \Delta m^{[2]}(x)
\]
Similarly $m^{(n)}(x)$ has $2^n$ "teeth."

The key point is $m(x)$ can be computed by a simple depth two ReLU network:

claim: $m(x) = \sigma(2\sigma(x) - 4\sigma(x-\frac{1}{2}))$

**Why can’t much shallower networks get low approx. error?**

**Proof by Picture:** If you don’t have enough pieces in your piece-wise linear approximation:

![Graph showing piece-wise linear approximation with an error approximation](image-url)