10707
Deep Learning: Spring 2020

Andrej Risteski

Machine Learning Department

Lecture 2:
Representational power of neural networks

Supervised learning

Empirical risk minimization approach:
minimize a **training** loss l over a class of **predictors** $\mathcal{F}$:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{(x,y):\text{training samples}} l(f(x), y)$$

Three pillars:

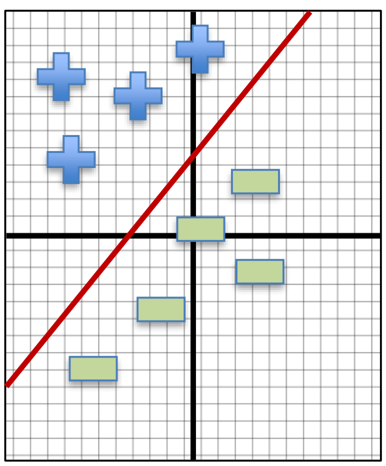
(1) How expressive is the class $\mathcal{F}$? (**Representational power**)

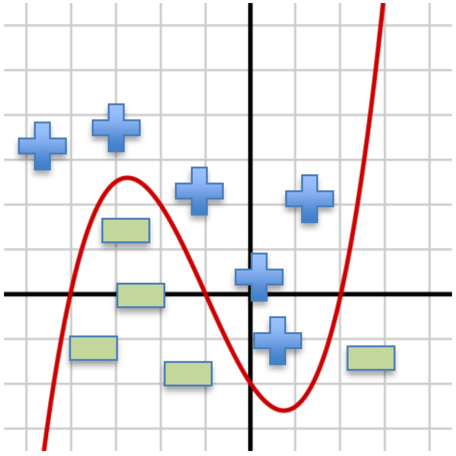
(2) How do we minimize the training loss efficiently? (**Optimization**)

(3) How does $\hat{f}$ perform on unseen samples? (**Generalization**)

Expressivity

What do we mean by expressivity?





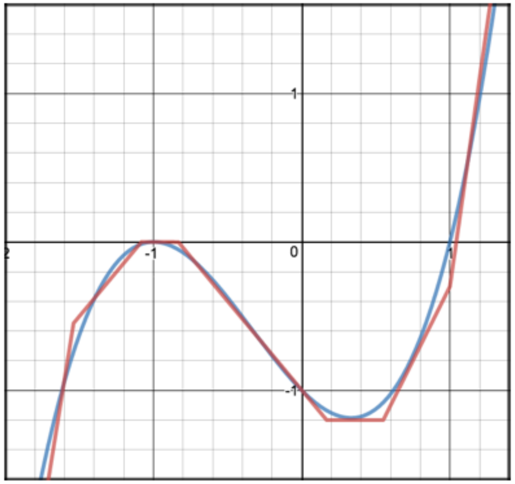
Expressive = functions in class can represent “complicated” functions

"Universal" expressivity of neural networks

(1): Neural networks are **universal approximators**: given any Lipschitz f: $\mathbb{R}^d \to \mathbb{R}$, a **shallow** (3-layer) neural network with $\sim \left(\frac{1}{\epsilon}\right)^{\mathrm{d}}$ neurons can approximate it to within ϵ error.

"curse of dimensionality"

(2): Neural networks can **circumvent** the curse of dimensionality for functions w/ decaying Fourier coefficients:
shallow neural networks with $\sim \left(\frac{1}{\epsilon}\right)$ neurons can approximate them to within ϵ error.



Universal approximation I: Lipschitz function are approximable

Recall, a function f: $[0,1]^d \to \mathbb{R}$ is **L-Lipschitz**, if:
$\forall x, y \in [0,1]^d,\ |f(x) - f(y)| \le L||x - y||_2$

First, we show neural networks are **universal approximators**: given any Lipschitz function f: $[0,1]^d \to \mathbb{R}$, a **shallow** (3-layer) neural network with ~ $\left(\frac{1}{\epsilon}\right)^{\mathrm{d}}$ neurons can approximate it to within ϵ error.

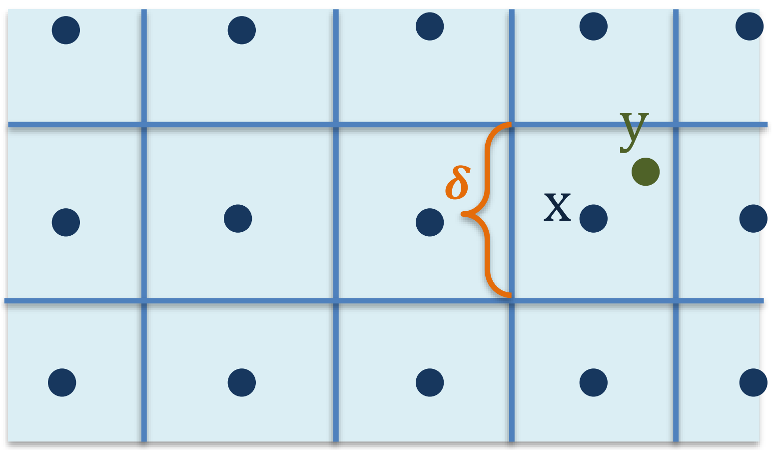
Theorem: For any L-Lipschitz function f: $[0,1]^d \to \mathbb{R}$, there is a 3-layer neural network $\hat{\mathrm{f}}$ with $\mathrm{O}\left(\mathrm{d}\left(\frac{\mathrm{L}}{\epsilon}\right)^{d}\right)$ ReLU neurons, s.t.

$$\int_{[0,1]^d} |f(x) - \hat{f}(x)|\, dx \le \epsilon$$

l_1 *error*

Universal approximation I: Proof intuition

Part 1: using Lipschitzness, we can “query” the values of function f approximately by querying its values on a fine grid.



$$|f(y) - f(x)| \leq L\delta$$

Part 2: we can approximate f as linear combination of “queries”.

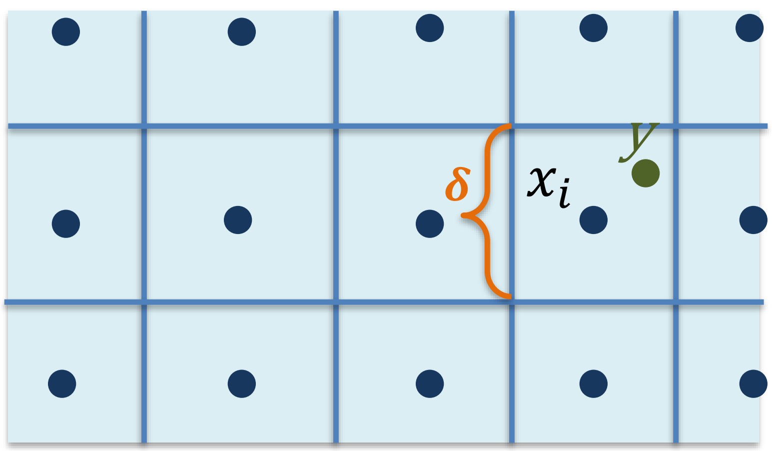
$$f(x) \approx \sum_{\text{cells } C_i} 1_{x \in C_i} f(x_i)$$

Part 3: Approximate the indicators using ReLUs

Universal approximation I: Part 1, formally

Lemma: Let f: $[0,1]^d \to \mathbb{R}$ be L-Lipschitz and $P = (C_1, C_2, \dots, C_N)$ a partition of $[0,1]^d$ into cells of side lengths at most δ. Consider any set $(x_1, x_2, \dots, x_N), x_i \in C_i$. Then:

$$\sup_{i \in N} \sup_{y \in C_i} |f(y) - f(x_i)| \leq L\delta$$



Proof: By Lipschitzness, we have

$$\forall i, y \in C_i\colon\ |f(y) - f(x_i)| \leq L|y - x_i| \leq L\delta \qquad \square$$

Universal approximation I: Part 2, formally

Lemma: Let f: $[0,1]^d \to \mathbb{R}$ be 1-Lipschitz, $P = (C_1, C_2, \ldots, C_N)$ a partition of $[0,1]^d$ into rectangles of side lengths at most δ, and a set $(x_1, x_2, \ldots, x_N), x_i \in C_i$. Then,

$$\mathrm{g}(\mathrm{x}) = \sum_{i=1}^{N} 1_{x \in C_i} f(x_i) \text{ satisfies } \sup_{x \in [0,1]^d} |\mathrm{f}(\mathrm{x}) - \mathrm{g}(\mathrm{x})| \leq L\,\delta$$

Proof: Let x $\in C_i$. Then, $1_{x \in C_i} = 1$, and $1_{x \in C_j} = 0$ for $j \neq i$.

So, $\mathrm{g}(\mathrm{x}) = f(x_i)$.

By Lemma 1, $|\mathrm{f}(\mathrm{x}) - \mathrm{g}(\mathrm{x})| = |\mathrm{f}(\mathrm{x}) - g(\mathrm{x}_i)| \leq L\,\delta$ □

Universal approximation I: approximating indicators of cells

Lemma: Let $C \subseteq \mathbb{R}^d$ be a cell, namely $C = \{x\colon\ x \in [l_i, r_i], i \in d\}$.
Then, there exists a 2-layer network $\tilde{\tilde{h}}(x)$ of size O(d) and ReLU activation, s.t. $\int_{x\in[0,1]^d} \left|\tilde{\tilde{h}}(x) - 1(x \in [l_i, r_i], i \in d)\right| \, dx \to 0$

Proof: First, write indicator for cell as:

$$1(x \in [l_i, r_i], i \in d) = 1\left(\sum_{i=1}^{d}\big(1(x_i \geq l_i) + 1(x_i \leq r_i)\big) \geq 2d - 1\right)$$

Why? x is in cell iff all the indicators $1(x_i \geq l_i) + 1(x_i \leq r_i)$ are on.
All these indicators are on iff they sum to $2d$.
(If at least one is off, they sum to $2d$-1)

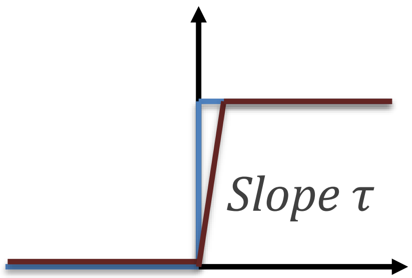
If we can approximate indicators, we're all good!

Universal approximation I: approximating indicators of cells

Claim: For $\tau \geq 0,\ x \in \mathbb{R}$:

$$\left|1(x \geq 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right)\right| = \begin{cases} \leq 1, & \text{if } 0 \leq x \leq 1/\tau \\ 0, & \text{otherwise} \end{cases}$$

Proof: Consider several cases:



Case 1, $\boldsymbol{x \leq 0}$: $1(x \geq 0) = 0$ and $\sigma(\tau x) - \sigma(\tau x - 1) = 0$, so

$$1(x \geq 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right) = 0$$

Case 2, $\boldsymbol{x \geq 1/\tau}$: $1(x \geq 0) = 1$ and $\sigma(\tau x) - \sigma(\tau x - 1) = 1$, so

$$1(x \geq 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right) = 0$$

Case 3, $\boldsymbol{0 \leq x \leq 1/\tau}$: $1(x \geq 0) = 1$ and $\sigma(\tau x) - \sigma(\tau x - 1) = \tau x$, so

$$\left|1(x \geq 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right)\right| = \tau x \leq 1$$

Universal approximation I: approximating indicators of cells

$$h(x) := 1\left(\sum_{i=1}^{d}\big(1(x_i \geq l_i) + 1(x_i \leq r_i)\big) \geq 2d - 1\right)$$

Replace all indicators by difference of ReLUs. *What is the error?*

For brevity, let $\tilde{1}(x) = \sigma(\tau x) - \sigma(\tau x - 1)$, for some τ we will choose.

Let $\tilde{h}(x) \coloneqq 1\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$

$\tilde{\tilde{h}}(x) \coloneqq \tilde{1}\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$

(Change the approximations “iteratively”.)

Universal approximation I: approximating indicators of cells

$$h(x) := 1\left(\sum_{i=1}^{d}\left(1(x_i \geq l_i) + 1(x_i \leq r_i)\right) \geq 2d - 1\right)$$
$$\tilde{h}(x) := 1\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$
$$\tilde{\tilde{h}}(x) := \tilde{1}\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$

We have:

$$\int_{x\in[0,1]} \left|\tilde{\tilde{h}}(x) - h(x)\right| dx = \int_{x\in[0,1]^d} \left|\tilde{\tilde{h}}(x) - \tilde{h}(x) + \tilde{h}(x) - h(x)\right| dx$$

$$\leq \int_{x\in[0,1]^d} \left(\left|\tilde{\tilde{h}}(x) - \tilde{h}(x)\right| + \left|\tilde{h}(x) - h(x)\right|\right) dx$$

Triangle inequality

Let's handle two terms one by one.

Universal approximation I: approximating indicators of cells

$$\tilde{h}(x) := 1\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$
$$\tilde{\tilde{h}}(x) := \tilde{1}\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$

Claim: $\left|1(x \geq 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right)\right| = \begin{cases} \leq 1, & \text{if } 0 \leq x \leq 1/\tau \\ 0, & \text{otherwise} \end{cases}$

First: $\tilde{\tilde{h}}(x) - \tilde{h}(x) \neq 0$ only if $\exists i : x_i \in \left(l_i, l_i + \frac{1}{\tau}\right)$ or $x_i \in \left(r_i, r_i - \frac{1}{\tau}\right)$

(If $\tilde{\tilde{h}}(x) - \tilde{h}(x) \neq 0, \sum_i \tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i) - 2d - 1 \in \left(0, \frac{1}{\tau}\right)$, and if condition above isn't satisfied, $\tilde{1} = 1$, above sum is integer)

Measure of such x's is bdd by $\frac{2d}{\tau}$, so $\int_{x\in[0,1]^d} |\tilde{\tilde{h}}(x) - \tilde{h}(x)| \leq \frac{2d}{\tau}$

Universal approximation I: approximating indicators of cells

$$h(x) := 1\left(\sum_{i=1}^{d}\left(1(x_i \ge l_i) + 1(x_i \le r_i)\right) \ge 2d-1\right)$$
$$\tilde{h}(x) := 1\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \ge l_i) + \tilde{1}(x_i \le r_i)\right) \ge 2d-1\right)$$

Claim: $\left|1(x \ge 0) - \left(\sigma(\tau x) - \sigma(\tau x - 1)\right)\right| = \begin{cases} \le 1, & \text{if } 0 \le x \le 1/\tau \\ 0, & \text{otherwise} \end{cases}$

Second: $\displaystyle\int_{x\in[0,1]} \left|\tilde{h}(x) - h(x)\right| dx$

Indicators are equal if inputs are equal

$$\le \int_{x\in[0,1]^d} \left|1\left(\sum_{i=1}^{d}\left(1(x \ge l_i) + 1(x \le r_i)\right) \ne \sum_{i=1}^{d}\left(\tilde{1}(x \ge l_i) + \tilde{1}(x \le r_i)\right)\right)\right| dx$$

$$\le \int_{x\in[0,1]^d} \sum_{i=1}^{d} 1\left(1(x \ge l_i) \ne \tilde{1}(x \ge l_i)\right) + \sum_{i=1}^{d} 1\left(1(x \le r_i) \ne \tilde{1}(x \le r_i)\right) dx$$

By Claim $\le 2d/\tau$

Universal approximation I: approximating indicators of cells

$$h(x) := 1\left(\sum_{i=1}^{d}\left(1(x_i \geq l_i) + 1(x_i \leq r_i)\right) \geq 2d - 1\right)$$
$$\tilde{h}(x) := 1\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$
$$\tilde{\tilde{h}}(x) := \tilde{1}\left(\sum_{i=1}^{d}\left(\tilde{1}(x_i \geq l_i) + \tilde{1}(x_i \leq r_i)\right) \geq 2d - 1\right)$$

Putting together, we have:

$$\int_{x\in[0,1]^d} \left|\tilde{\tilde{h}}(x) - h(x)\right| dx = \int_{x\in[0,1]^d} \left|\tilde{\tilde{h}}(x) - \tilde{h}(x) + \tilde{h}(x) - h(x)\right| dx$$
$$\leq 4d/\tau$$

Also, $\tilde{\tilde{h}}(x)$ is a 2-layer net with ReLU activations and $O(d)$ nodes!

Universal approximation I: Putting everything together

By Part 1+2, $\sup_{x\in[0,1]^d} \left|\mathrm{f(x)} - \sum_{i=1}^{N} 1_{x\in C_i}\, f(x_i)\right| \leq L\,\delta$

Moreover, the number of cells N can be bounded by $\left(\frac{1}{\delta}\right)^d$

By indicator approximation: can approximate arbitrarily well by taking $\tau \to \infty$ with a 2-layer ReLU net.

Combining the above two points, we get a $\left(\frac{1}{\delta}\right)^d$ −sized 3-layer net s.t.

$$\int_{[0,1]^d} |\, f(x) - \hat{f}(x)|\, dx \leq L\,\delta$$

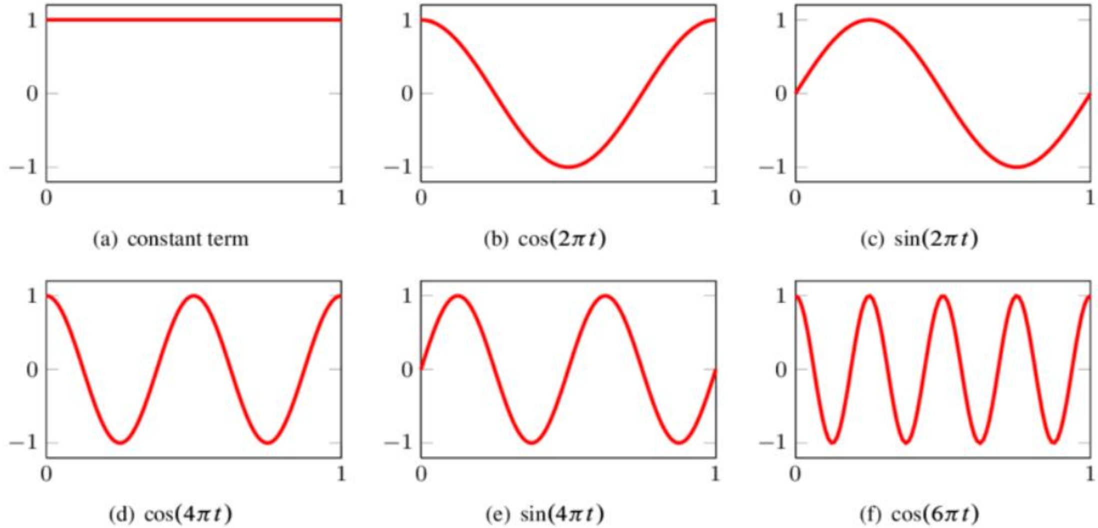
Taking $\delta = \frac{\epsilon}{L}$, the theorem follows.

Escaping the curse of dimensionality: Nuggets of Fourier analysis

Can the $\left(\frac{1}{\epsilon}\right)^d$ dependence be avoided for "nice" functions?

Yes! Relevant property is decay of the Fourier coefficients.

Recall: The Fourier basis for "nice" functions from $\mathbb{R}^d \to \mathbb{R}$ consists of basis functions $\{e_w(x) = e^{i\langle w,x \rangle} = \cos(\langle w,x \rangle) + i\,\sin(\langle w,x \rangle) | w \in \mathbb{R}^d\}$.



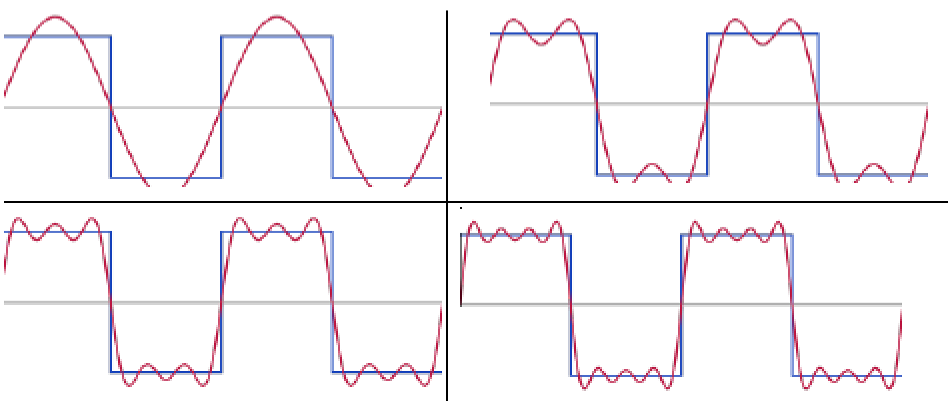
||w|| is larger: function oscillates more

Escaping the curse of dimensionality: Nuggets of Fourier analysis

Can the $\left(\frac{1}{\epsilon}\right)^d$ dependence be avoided for "nice" functions?

Yes! Relevant property is decay of the Fourier coefficients.

Recall: The Fourier basis for "nice" functions from $\mathbb{R}^d \to \mathbb{R}$ consists of basis functions $\{e_w(x) = e^{i\langle w,x \rangle} = \cos(\langle w, x \rangle) + i\sin(\langle w, x \rangle) | w \in \mathbb{R}^d\}$.



Higher and higher frequencies => better approximation

Escaping the curse of dimensionality: Nuggets of Fourier analysis

Can the $\left(\frac{1}{\epsilon}\right)^d$ dependence be avoided for "nice" functions?

Yes! Relevant property is decay of the Fourier coefficients.

Recall: The Fourier basis for "nice" functions from $\mathbb{R}^d \to \mathbb{R}$ consists of basis functions $\{e_w(x) = e^{i\langle w,x \rangle} = \cos(\langle w, x \rangle) + i \sin(\langle w, x \rangle) | w \in \mathbb{R}^d\}$.

Recall: The Fourier integral theorem gives coefficients for this basis:

Defining $\hat{f}(w) = \int_{\mathbb{R}^d} f(x) e^{-i\langle w,x\rangle} dx$, we have:

$$f(x) = \int_{\mathbb{R}^d} \underbrace{\hat{f}(w)}_{\substack{\textit{Coefficient for} \\ \textit{basis fn } e^{i\langle w,x \rangle}}} e^{i\langle w,x\rangle} dw$$

Escaping the curse of dimensionality: Nuggets of Fourier analysis

Can the $\left(\frac{1}{\epsilon}\right)^d$ dependence be avoided for "nice" functions?

Yes! Relevant property is decay of the Fourier coefficients.

Def.: The **Barron constant** of a function f is the quantity

$$C = \int_{\mathbb{R}^d} ||w|| \, \left|\hat{f}(w)\right| dw$$

Interpretation: the higher-order Fourier coefficients (i.e. high-oscillation parts of f) are small.

We will look for $O_C\left(\frac{1}{\epsilon}\right)$ dependence of the size of the network.

Escaping the curse of dimensionality: Barron's Theorem

Def.: The **Barron constant** of a function f is the quantity

$$C = \int_{\mathbb{R}^d} ||w|| \; |\hat{f}(w)| \, dw$$

$=\{x \in \mathbb{R}^d : ||x|| \leq 1\}$

Theorem (Barron '93): For any f: $\mathbb{B} \to \mathbb{R}$, there is a 3-layer neural network $\hat{f}$ with $O\left(\frac{C^2}{\epsilon}\right)$ ReLU neurons, s.t.

$$\int_{\mathbb{B}} \left(f(x) - \hat{f}(x)\right)^2 dx \leq \epsilon$$

$$= \mathbb{E}_x\left[\left(f(x) - \hat{f}(x)\right)^2\right]$$

l_2 *error*

Barron's theorem: proof idea

Step 1: Show that any continuous function f can be written as an "infinite" convex combination of cosine-like activations .
(**Main tool**: Fourier integral theorem)

Step 2: Show that a function f with small Barron constant can in fact be approximately written as a convex combination of a **small** number of cosine-like activations.
(**Main tool**: subsampling the above infinite combination and concentration bounds.)

Step 3: Show that the cosine non-linearities can be approximated by sigmoid non-linearities.
(**Main tool**: classical approximation theory.)

Step 1: infinite convex combination of cosine-like activations

By Fourier integral theorem, we have:

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(w) e^{i\langle w,x \rangle} dw$$

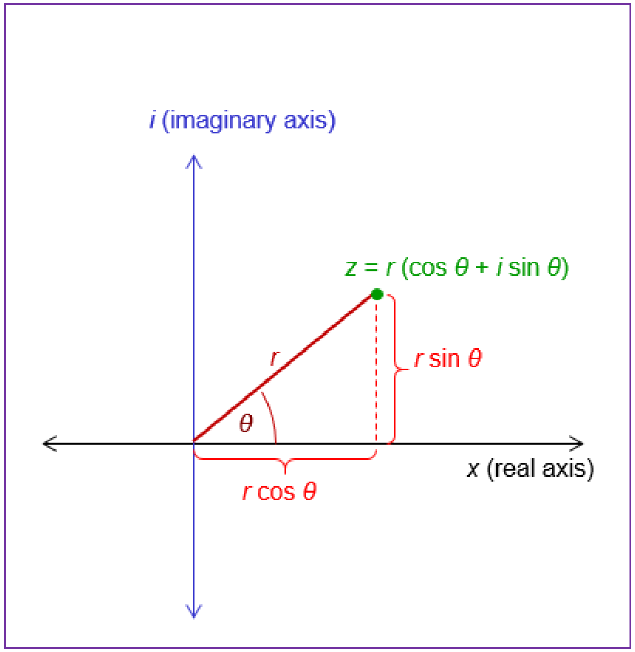
$$= f(0) + \int_{\mathbb{R}^d} \hat{f}(w) \left(e^{i\langle w,x \rangle} - 1 \right) dw$$

$$= \int_{\mathbb{R}^d} \hat{f}(w) dw$$

Step 1: infinite convex combination of cosine-like activations

By Fourier integral theorem, $f(x) = f(0) + \int_{\mathbb{R}^d} \hat{f}(w)\left(e^{i\langle w,x\rangle} - 1\right)dw$

Recall the **polar form** of a complex number:



$z = |z|\, e^{i\,\phi_z}$

$= |z|\,(\cos\phi_z + i\sin\phi\_z)$

Step 1: infinite convex combination of cosine-like activations

By Fourier integral theorem, $f(x) = f(0) + \int_{\mathbb{R}^d} \hat{f}(w)\left(e^{i\langle w,x\rangle} - 1\right)dw$

Recall the **polar form** of a complex number: $z = |z|\, e^{i\,\phi_z}$

Hence, we can rewrite the Fourier integral formula as:

$$f(x) = f(0) + \int_{\mathbb{R}^d} \left|\hat{f}(w)\right|\left(e^{i(b_w + \langle w,x\rangle)} - e^{ib_w}\right) dw$$

Step 1: infinite convex combination of cosine-like activations

$$f(x) = f(0) + \int_{\mathbb{R}^d} \left|\hat{f}(w)\right| (e^{i(b_w + \langle w,x \rangle)} - e^{ib_w})\, dw$$

Recall the expansion of complex exponentials: $e^{iy} = \cos(y) + i\,\sin(y)$

As f is a real-valued function, only the real part of the above expression will survive. Hence,

$$f(x) = \int_{\mathbb{R}^d} \left|\hat{f}(w)\right| (\cos(b_w + \langle w, x \rangle) - \cos(b_w))dw$$

Linear combination of cosine functions, but not *convex!*

(As $\int_{\mathbb{R}^d} \left|\hat{f}(w)\right|$ integrates potentially to > 1)

Step 1: infinite convex combination of cosine-like activations

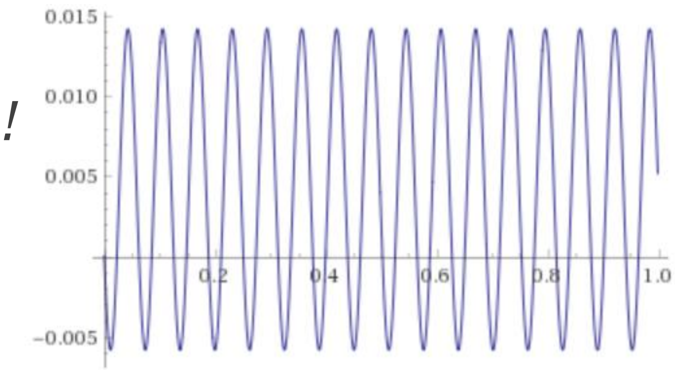
We will rewrite: $f(x) = \int_{\mathbb{R}^d} |\hat{f}(w)| \left(\cos(b_w + \langle w, x \rangle) - \cos(b_w)\right) dw$

$$f(x) = \int_{\mathbb{R}^d} |\hat{f}(w)| \left(\cos(b_w + \langle w, x \rangle) - \cos(b_w)\right) dw$$

$$= \int_{\mathbb{R}^d} \frac{|\hat{f}(w)| \, ||w||}{C} \left(\frac{C}{||w||} \left(\cos(b_w + \langle w, x \rangle) - \cos(b_w)\right) \right) dw$$

Convex combination of cosine-like activations*!*

(As $\int_{\mathbb{R}^d} \frac{|\hat{f}(w)| \, ||w||}{C} = 1$)



Step 2: convex combination of small number of cosine-like activations

Recall: $f(x) = \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|\,||w||}{C} \left(\frac{C}{||w||} (\cos(b_w + \langle w, x \rangle) \; - \cos(b_w)) \right)$

We will prove that there is a set S of w's, s.t.

$$f(x) \approx \frac{1}{|S|} \sum_{w \in S} \frac{C}{||w||} (\cos(b_w + \langle w, x \rangle) \; - \cos(b_w))$$

Natural idea: **subsampling**!

Remember, these *integrate* to 1, so form a **distribution** over w's.

Repeat **r** times:

Choose a new $w \in \mathbb{R}$ to add to S with probability $\frac{|\hat{f}(w)|\,||w||}{C}$

Step 2: convex combination of small number of cosine-like activations

Recall: $f(x) = \int_{\mathbb{R}^d} \frac{|\hat{f}(w)| ||w||}{C} \left(\frac{C}{||w||} (\cos(b_w + \langle w, x \rangle) - \cos(b_w)) \right)$

Repeat **r** times:

Choose a new $w \in \mathbb{R}$ to add to S with probability $\frac{|\hat{f}(w)| ||w||}{C}$

Let g_i be a random variable, denoting the i-th selected w.

Let $g = \frac{1}{r} \sum_{i=1}^{r} g_i$. Then, we have:

$$\mathbb{E}_x \mathbb{E}_g [(g(x) - f(x))^2] = \mathbb{E}_x \mathbb{E}_{g_i} \left[\left(\sum_i \left(\frac{1}{r} g_i - \frac{1}{r} f \right) \right)^2 \right] = \frac{1}{r^2} \mathbb{E}_x \mathbb{E}_{g_i} [(\sum_i (g_i - f))^2]$$

Direct substitution

All $g_i - f$ are mean-0, (since $\mathbb{E}[g_i] = f$), and independent.

Step 2: convex combination of small number of cosine-like activations

Repeat **r** times:

Choose a new $w \in \mathbb{R}$ to add to S with probability $\frac{|\hat{f}(w)||w|}{C}$

$\mathbb{E}_x \mathbb{E}_{g_i,g_j}[(g_i - f)(g_j - f)]=$
$\mathbb{E}_x[\mathbb{E}_{g_i}[(g_i - f)]\mathbb{E}_{g_j}[(g_j - f)]] = 0$

Then, we have:

$$\mathbb{E}_x \mathbb{E}_g[(g-f)^2] = \frac{1}{r^2} \mathbb{E}_x \mathbb{E}_{g_i}[(\textstyle\sum_i (g_i - f))^2]$$

$\mathbb{E}[g_i] = f$

$$= \frac{1}{r^2}\left(\textstyle\sum_i \mathbb{E}_x \mathbb{E}_g[(g_i - f)^2] + \sum_{i \neq j} \mathbb{E}_x \mathbb{E}_{g_i,g_j}[(g_i - f)(g_j - f)]\right)$$

$$= \frac{1}{r^2}\left(\textstyle\sum_i \mathbb{E}_x \mathbb{E}_g[(g - \mathbb{E}[g])^2]\right) = \frac{1}{r} \mathbb{E}_x \mathbb{E}_g[(g - \mathbb{E}_g[g])^2]$$

$$= \frac{1}{r}(\mathbb{E}_x \mathbb{E}_g[g^2] - \mathbb{E}_x \mathbb{E}_g[g]^2) \leq \frac{1}{r} \mathbb{E}_g \mathbb{E}_x[g^2]$$

Change order of expectations

$$\leq \frac{1}{r} \max_w \mathbb{E}_x[g_w^2]$$

Slight abuse of notation

Step 2: convex combination of small number of cosine-like activations

Repeat **r** times:

Choose a new $w \in \mathbb{R}$ to add to S with probability $\frac{|\hat{f}(w)||w|}{C}$

Let g_i denote the i-th selected w. Let $g = \frac{1}{r}\sum_{i=1}^{r} g_i$

Then, we have: $\mathbb{E}_x \mathbb{E}_g[(g-f)^2] \le \frac{1}{r} \max_w \mathbb{E}_x[\, g_w^2]$

Writing out $\mathbb{E}_x[\, g_w^2]$ explicitly, we will show that:

$$\forall w\colon \int_{x \in \mathbb{B}} \left(\frac{C}{||w||} (\cos(b_w + \langle w, x \rangle) - \cos(b_w)) \right)^2 dx \le C^2$$

Step 2: convex combination of small number of cosine-like activations

Claim: $\forall w$: $\int_{x\in\mathbb{B}} \left(\frac{C}{||w||}(\cos(b_w + \langle w, x\rangle) - \cos(b_w))\right)^2 dx \le C^2$

Note, cos is 1-Lipschitz (show this if you don't see it!). Hence:

$$|(\cos(b_w + \langle w, x\rangle) - \cos(b_w))| \le |\langle w, x\rangle| \le ||w|| \; ||x||$$

So, $\left(\frac{C}{||w||}(\cos(b_w + \langle w, x\rangle) - \cos(b_w))\right)^2 \le C^2 ||x||^2 \le C^2$

Integrating, the claim follows.

Step 2: convex combination of small number of cosine-like activations

Repeat **r** times:

Choose a new $w \in \mathbb{R}$ to add to S with probability $\frac{|\hat{f}(w)||w|}{C}$

Let g_i denote the i-th selected w. Let $g = \frac{1}{r}\sum_{i=1}^{r} g_i$

Plugging in previous bound: $\mathbb{E}_g \mathbb{E}_x[(g-f)^2] \leq \frac{C^2}{r}$

If the expectation of a random variable is $\leq \frac{C^2}{r}$, there must be some realization of it w/ value $\leq \frac{C^2}{r}$. Hence:

There exist some g, s.t. $\mathbb{E}_x[(g(x)-f(x))^2] \leq \frac{C^2}{r}$

Almost there! g is a width r network, with cosine-like activation.

Step 3: approximating the cosines

Finally, we approximate the cosine-like activations using sigmoids.

Let us denote $g_w(x) = \frac{C}{||w||}(\cos(b_w + \langle w, x\rangle) - \cos(b_w))$

Namely, we show that: there exists a 2-layer neural net G_0 of size $O\left(\frac{1}{\epsilon}\right)$ with sigmoid activations, s.t. $\sup_{x \in \mathbb{B}} | G_0(x) - g_w(x)| \leq \epsilon$

Step 3: approximating the cosines

$$g_w(x) = \frac{C}{||w||}(\cos(b_w + \langle w, x\rangle) - \cos(b_w))$$

Exists 2-layer neural net G_0 of size $O\left(\frac{1}{\epsilon}\right)$ with sigmoid activations, s.t. $\sup_{x\in\mathbb{B}} | G_0(x) - g_w(x)| \leq \epsilon$

First, we rewrite $g_w(x)$ slightly:

$$g_w(x) = \frac{C}{||w||}\left(\cos\left(b_w + ||w|| \underbrace{\left\langle \frac{w}{||w||}, x \right\rangle}_{y}\right) - \cos(b_w)\right)$$

$$:= h_w(y)$$

Hence, $g_w(x) = h_w\left(\left\langle \frac{w}{||w||}, x\right\rangle\right)$, i.e. a composition of a **linear function** and $\boldsymbol{h_w}$, and the domain of h_w is [-1,1] (univariate!). Suffices to approx. h_w using sigmoids.

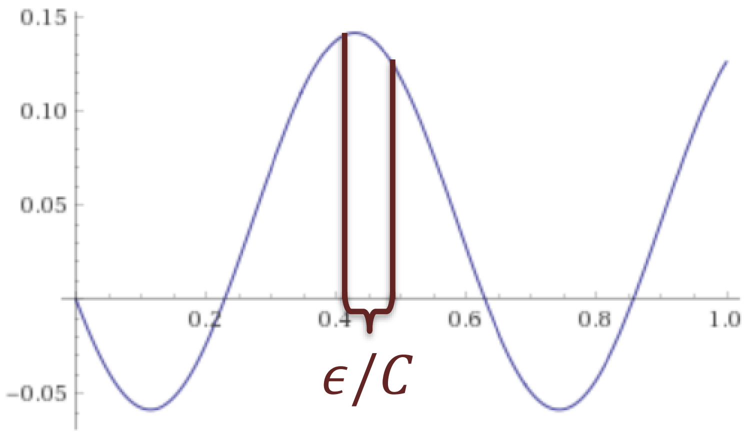
Step 3: approximating the cosines

$$h_w(y) = \frac{C}{||w||}(\cos(b_w + ||w||y) - \cos(b_w))$$

Exists 2-layer neural net G_0 of size $O\left(\frac{1}{\epsilon}\right)$ with sigmoid activations, s.t.

$$\sup_{x\in[-1,1]} |G_0(y) - h_w(y)| \le \epsilon$$

Check derivative bd gives Lipschitzness!



1. h_w is C-Lipschitz:

$$h'_w(y) = C\sin(b_w + ||w||y)$$

2. Grid the interval [-1,1] into intervals $[l_i, r_i]$ of size ϵ/C. Pick arbitrary $y_i \in [l_i, r_i]$

Same as in the first theorem, we have

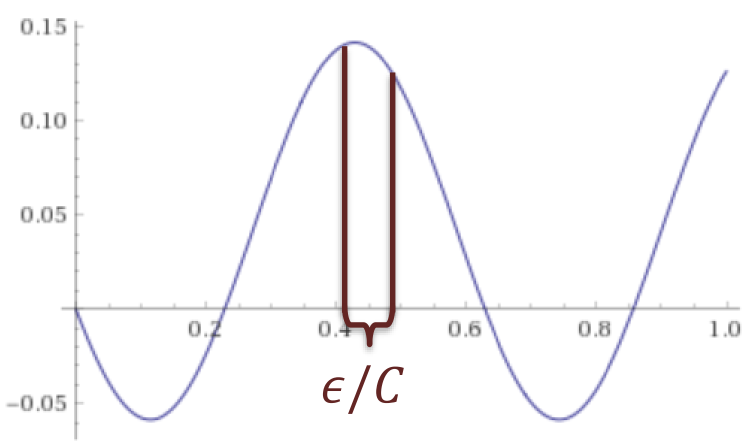
$$\sup_{x\in[-1,1]} \left| \sum_i 1(y \in [l_i, r_i]) h_w(y_i) - h_w(y) \right| \le \epsilon$$

Step 3: approximating the cosines

$$h_w(y) = \frac{C}{||w||}(\cos(b_w + ||w||y) \ - \cos(b_w))$$

Exists 2-layer neural net G_0 of size $O\left(\frac{1}{\epsilon}\right)$ with sigmoid activations, s.t.

$$\sup_{x\in[-1,1]} |\, G_0(y) - h_w(y)| \leq \epsilon$$



$$\sup_{x\in[-1,1]} \left| \sum_i 1(y \in [l_i, r_i]) h_w(y_i) - h_w(y) \right| \leq \epsilon$$

3. We can write the indicators as differences of step functions:

$$1(y \in [l_i, r_i]) = 1(y \geq l_i) - 1(y \geq r_i)$$

Hence, it suffices to approximate a step function using a sigmoid.

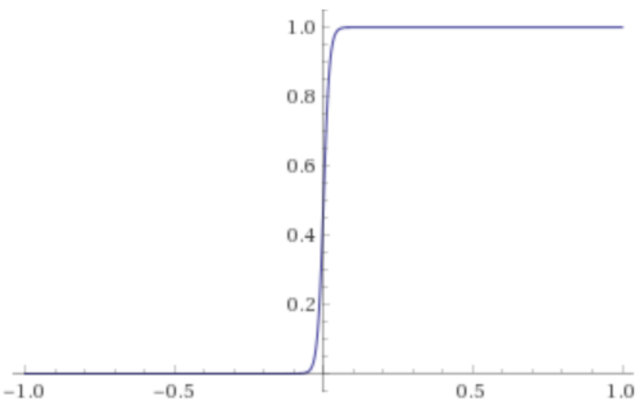
Step 3: approximating the cosines

$$h_w(y) = \frac{C}{||w||}\left(\cos(b_w + ||w||y) \; - \cos(b_w)\right)$$

Exists 2-layer neural net G_0 of size $O\left(\frac{1}{\epsilon}\right)$ with sigmoid activations, s.t.

$$\sup_{x \in [0,1]^d} |\, G_0(x) - h_w(x)| \leq \epsilon$$

Approximating a step function using a sigmoid:



$$\lim_{\tau \to \infty} \sup_{x \in [0,1]^d} \left| 1(x \geq a) - \frac{1}{1 + e^{-\tau(x-a)}} \right| = 0$$

Hence, taking τ large enough, we can drive the error as small as possible (in the l_∞ sense.)

Putting everything together, the claim follows.

Parting thoughts

All results we proved are **existential**: they prove that a good approximator exists. Finding one efficiently (much less so using gradient descent) is a different matter.

The choices of non-linearities are usually very **flexible**: most results of the type we saw can be re-proven using different non-linearities. (Examples in homework.)

Many other results of similar flavor. For instance, there are also results that deep, but narrow networks are universal approximators.