Mathematics of Deep Learning
Lecture 3: More on Shallow Networks and Expressive Power of Depth

Prof. Predrag R. Jelenković

Time: Tuesday 4:10-6:40pm

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.ee.columbia.edu/∼predrag

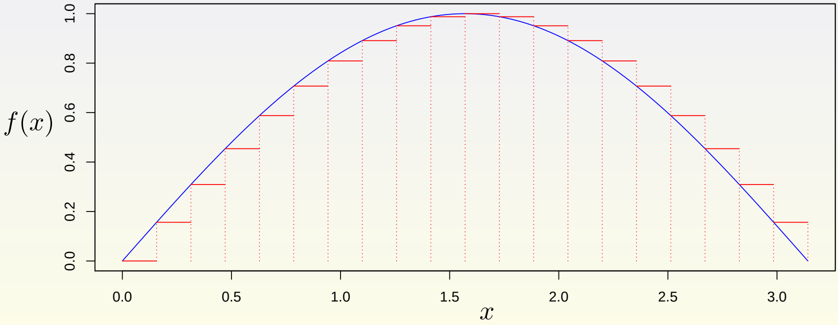
Last Lecture: Expressive Power of Neural Nets

Approximation With Simple Functions

Consider function $f(x) : [0, \pi] \rightarrow [0, 1]$

$$f(x) = \sin(x)$$

Step Functions



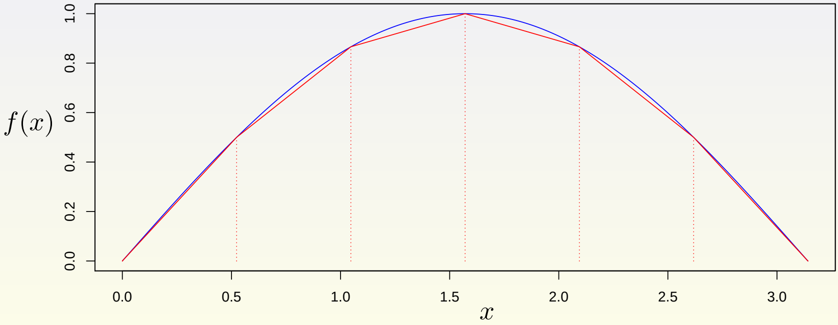
Last Lecture: Expressive Power of Neural Nets

Approximation With Simple Functions

Consider function $f(x) : [0, \pi] \to [0, 1]$

$$f(x) = \sin(x)$$

Piecewise Linear Functions



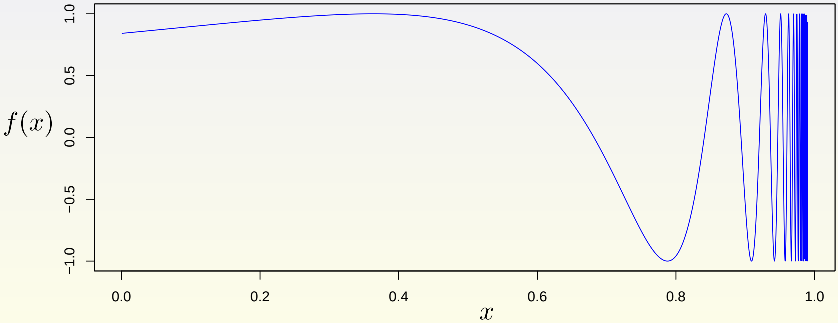
Last Lecture: Need Some Regularity

Consider function $f(x) : [0, 1) \to [-1, 1]$

$$f(x) = \sin\left(\frac{1}{1-x}\right)$$

is continuous and infinitely differentiable.

Difficult to Approximate: Why?



Function Regularity: Uniform Continuity

So, we need to impose some regularity on $f(x)$.
Definition Let $I \subset \mathbb{R}$ be an interval. Function $f(x) : I \to \mathbb{R}$ is uniformly continuous if for any $\epsilon > 0$, there exists $\delta > 0$, such that

$$|x_1 - x_2| < \delta \quad \Rightarrow |f(x_1) - f(x_2)| < \epsilon$$

Remarks

- If I is a closed interval, $[a, b]$, then

$$\text{ordinary continuity} \Leftrightarrow \text{uniform continuity}$$

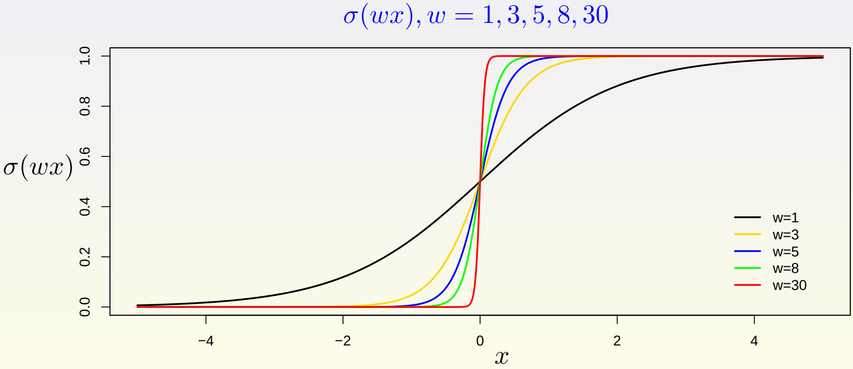
- Hence, to prevent the problem from the preceding example, $f(x)$ needs to be defined on a closed interval $[0, 1]$ instead of $[0, 1)$.
- The preceding definition extends to $\mathbb{R}^k$ spaces with Euclidean norm, and in general metric spaces.
- Finite closed intervals generalize to compact sets.

From Sigmoidal to Step Function

Starting with any sigmoidal functions, say

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

we can approximate arbitrarily close a step function by using $\sigma(wx)$ and w large enough.



Haar Scaling Function: Pulse Function

Haar scaling function (single pulse) is defined

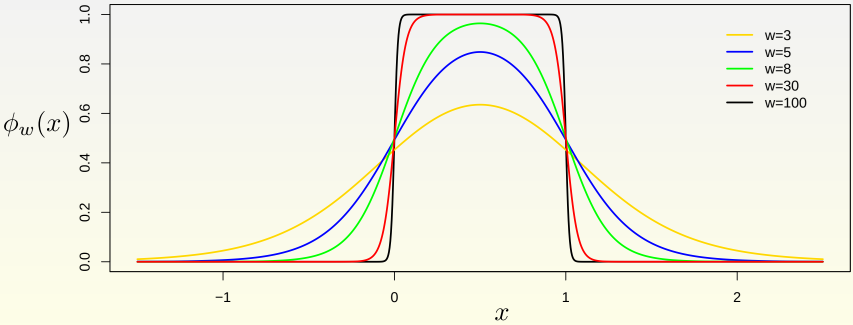
$$\phi(x) = 1_{\{x>0\}} - 1_{\{x-1>0\}}$$

Two sigmoidal functions, we can create a perfect pulse (Haar function)

$$\phi(x) \approx \phi_w(x) = \sigma(wx) - \sigma(w(x-1))$$

for w large enough (which can be used to make Haar wavelet basis)

$\phi_w(x), w = 3, 5, 8, 30, 100$

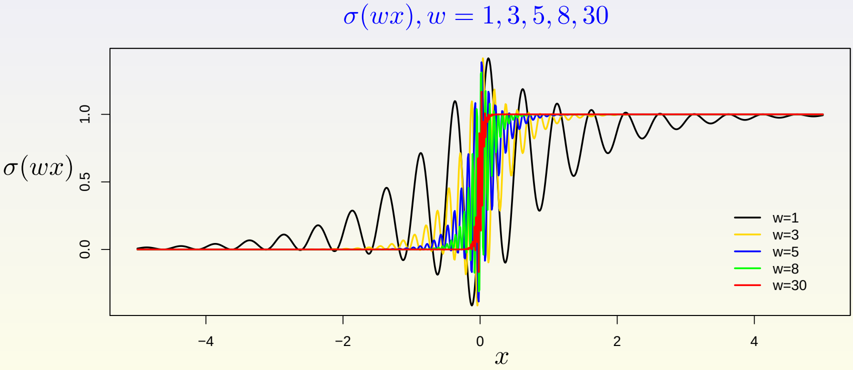


Non-monotonic Sigmoidal to Step Function

Do we need sigmoidal functions to be monotonic? For example

$$\sigma(x) = \frac{1}{1+e^{-x}} + \sin(4\pi x)e^{-|x|}$$

Again, we can approximate arbitrarily close a step function by using $\sigma(wx)$ and w large enough.



We can even relax the continuity by $\sigma(x)$ being bounded.

Universal Approximation Theorem in 1D

Definitions

- A function $\sigma(x) : \mathbb{R} \to \mathbb{R}$ is called sigmoidal if it is bounded and

$$\lim_{x \to -\infty} \sigma(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} \sigma(x) = 1.$$

  (Equivalence to step function: $\sigma(wx) \approx 1_{\{x>0\}}$ for large w.)

- $C([a,b])$: space of continuous functions, $f(x) : [a,b] \to \mathbb{R}$ without loss of generality, we consider $C([0,1])$

Theorem Consider a sigmoidal function, σ and $f \in C([0,1])$. For every $\epsilon > 0$, there exist an integer n and $w > 0$ (depending on n), such that for $x \in [0,1]$

$$\hat{f}(x) \overset{\text{def}}{=} \sum_{k=1}^{n} (f(k/n) - f((k-1)/n))\sigma(w(x - k/n)) + f(0)\sigma(w(x + 1/n)),$$

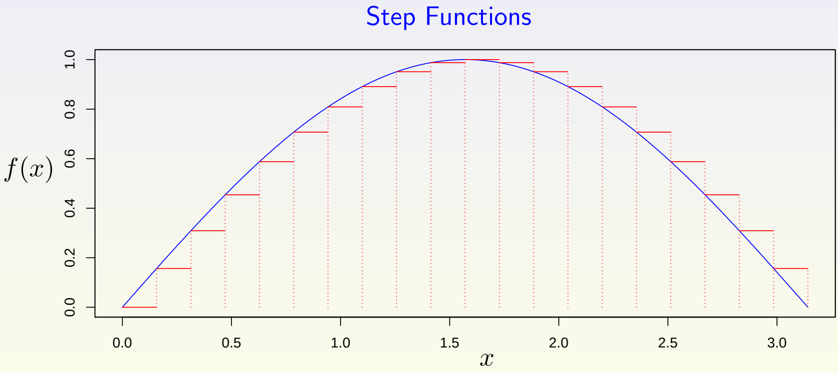
then

$$\sup_{0 \le x \le 1} |\hat{f}(x) - f(x)| < \epsilon.$$

Sigmoidal Approximation

Consider function $f(x) : [0, \pi] \to [0, 1]$

$$f(x) = \sin(x)$$



Expressiveness of ReLU in 1D: Also Universal

- Recall, ReLU activation: $\sigma(x) = \max\{0, x\} =: x^+$
- Approximation is even easier, since we already have linear functions
- The following approximation can be made arbitrarily close to $f \in C([0,1])$, i.e., $|f(x) - \hat{f}(x)| < \epsilon$, for large enough n

$$\hat{f}(x) = \sum_{k=0}^{n-1} w_k \sigma(x - k/n) + f(0),$$

where the slopes w_k are chosen such that $\hat{f}(k/n) = f(k/n)$, i.e.,

  - $w_0 = n(f(1/n) - f(0))$
  - w_1 is such that $f(2/n) = 2w_0/n + w_1/n + f(0)$, yielding

$$w_1 = n(f(2/n) - 2f(1/n) + f(0))$$

  - And so on, we inductively select w_k to satisfy

$$f\left(\frac{k+1}{n}\right) = \hat{f}\left(\frac{k+1}{n}\right) = w_0 \frac{k+1}{n} + w_1 \frac{k}{n} + \cdots + w_k \frac{1}{n} + f(0)$$

Functions Equivalent to ReLU

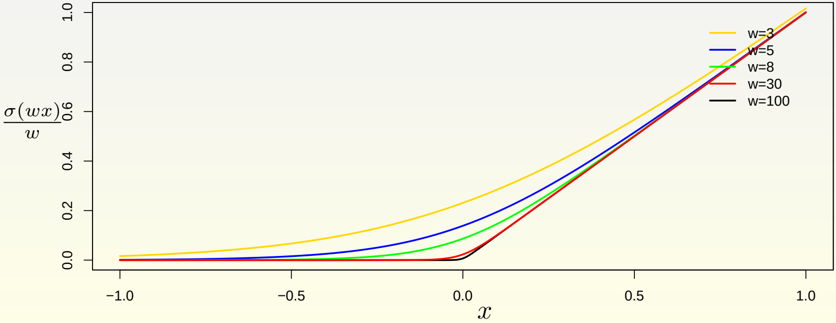
- Similarly to sigmoidal functions, through scaling, many other functions are equivalent to ReLU. For example, one can define ReLU-equivalent functions as: continuous with

$$\lim_{x\to-\infty} \sigma(x) = 0, \qquad \lim_{x\to-\infty} \frac{\sigma(x)}{x} = 1.$$

- Then, as $w \to \infty$,

$$\frac{\sigma(wx)}{w} \to \max(0, x)$$

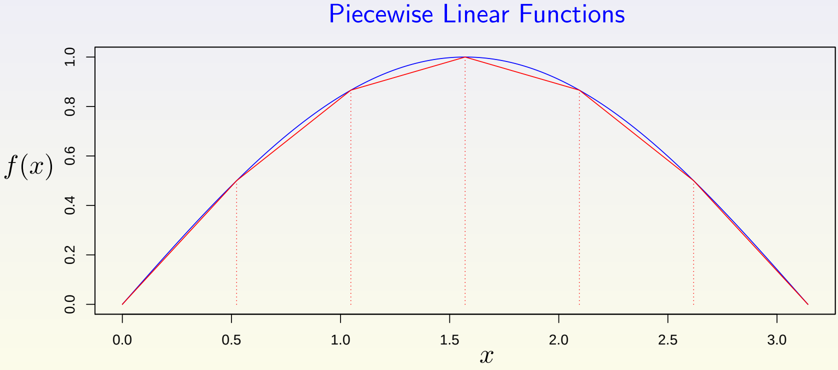
Scaling soft-plus $\sigma(x) = \log(1 + e^{x})$ to ReLU



ReLU Approximation

Consider function $f(x) : [0, \pi] \to [0, 1]$

$$f(x) = \sin(x)$$



Historical Comments on ReLU

- $x^+ = \max(0, x)$ has a long history in statistics
- It is called linear spline basis (or hinge function)
- Bias b: in $\max(0, x + b)$ is called knot
- One hidden layer NN with ReLU is a free knot linear spline
  - Studied for $50+$ years
- In general, one considers polynomial spline bases, for integer $k \geq 0$,

  $$(x^+)^k = (\max(0, x))^k,$$

  e.g., see Chapter 5 of The Elements of Statistical Learning book by Hastie et al.

From ReLU/Polynomial Splines to Sigmoids

- Suppose, for $k = 1, 2, 3, \ldots$, we are given polynomial splines

$$\sigma(x) = \begin{cases} x^k, & \text{if } x \geq 0, \\ ax^k, & \text{if } x < 0, a \neq 1. \end{cases}$$

  Note the case $k = 1, a = 0$ corresponds to ReLU, and $k = 1, 0 < a < 1$ corresponds to leaky ReLU.

- We can obtain sigmoids from these polynomial splines by forming finite differences, e.g., for $k = 1$

$$\Delta_1 \sigma(x) = \sigma(x) - \sigma(x-1) \qquad \text{is a sigmoid.}$$

- Similarly, for $k > 1$, we can obtain a sigmoid by forming the kth finite difference

$$\Delta_k \sigma(x) = \Delta_{k-1}\sigma(x) - \Delta_{k-1}\sigma(x-1).$$

  Check it out for $k = 1$.

From ReLU/Polynomial Splines to Sigmoids

- Suppose, a "soft" version of a polynomial spline, like the soft-plus instead of ReLU, for $k = 1, 2, 3, \ldots$,

$$\lim_{x} \sigma(x) = \begin{cases} x^k, & \text{as } x \to \infty, \\ ax^k, & \text{as } x \to -\infty, a \neq 1. \end{cases}$$

- Then, first scale the activation function using large $w > 0$, such that

$$\sigma^w(x) := \frac{\sigma(wx)}{w} \approx \begin{cases} x^k, & \text{if } x \geq 0, \\ ax^k, & \text{if } x < 0, a \neq 1. \end{cases}$$

- Then, form the finite differences on $\sigma^w(x)$.
- Hence, any result obtained for sigmoids holds also for ReLU, Leaky-ReLU, and polynomial splines.

Results by Cybenko

Theorem (Cybenko (1989)) Let σ be a continuous sigmoidal function (recall $\lim_{x\to-\infty}\sigma(x)=0$ and $\lim_{x\to\infty}\sigma(x)=1$ and pick an $f\in C([0,1]^d)$. Then, the set of approximation functions of the form

$$\hat{f}(x) := \sum_j \alpha_j \sigma(<\boldsymbol{w}_j, \boldsymbol{x}> + b_j)$$

is dense in $C([0,1]^d)$ with the metric $d(f,g)=\sup|f(x)-g(x)|$, $f,g\in C([0,1]^d)$, i.e., for any $f\in C([0,1]^d)$ and $\epsilon>0$, there exists a NN approximation function $\hat{f}(x)$, such that

$$\sup_{x\in[0,1]^d} |f(x)-\hat{f}(x)| < \epsilon.$$

Cybenko's Proof: Outline

Proof by contradiction:

- Consider the subspace M given by $\{\sum \alpha_j \sigma(< \boldsymbol{w}_j, \boldsymbol{x} > + b_j)\}$
- Assume that its closure $\overline{M}$ is not the entire space of functions $C([0,1]^d)$.
- Hence, by Hahn-Banach Theorem, there exists a continuous linear map L on our function space that restricts to 0 on $\overline{M}$ but is not identically zero.
- Then, by Riesz Representation Theorem, this linear functional can be expressed as integral, for any $f \in C([0,1]^d)$

$$L(f) = \int f \, d\mu(x)$$

- Key difficulty (Lemma 1) is to prove that, for all $\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}$,

$$\int_{[0,1]^d} \sigma(< \boldsymbol{w}, \boldsymbol{x} > + b) d\mu(\boldsymbol{x}) = 0 \quad \Rightarrow \quad \mu \equiv 0,$$

which implies $L \equiv 0$ on entire $C([0,1]^d)$, resulting in contradiction.

Hornik's Extensions

Hornik follows the general plan in Cybenko. The key technical result is the generalization of Lemma 1 of Cybenko to bounded and nonconstant activation functions, $\sigma(x)$ (not necessarily sigmoidal).

Theorem 5 (Hornik) Let μ be finite, signed measure on $[0,1]^d$. For any bounded and nonconstant σ, if, for all $\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}$

$$\int_{[0,1]^n} \sigma(<\boldsymbol{w},\boldsymbol{x}>+b)d\mu(x) = 0 \quad \Rightarrow \quad \mu \equiv 0$$

Using this technical result, Hornik obtains:

- **Theorem 1** If σ is unbounded and nonconstant, then NN approximations are dense in $L^p(\mu)$ for all finite measures μ on $\mathbb{R}^d$.
- **Theorem 2** (a generalization of Theorem 2 in Cybenko) If σ is continuous, bounded, and nonconstant, then NN approximations are dense continuous functions, $C(X)$, with compact domain $X \subset \mathbb{R}^k$ and supremum distance $d(f,g) = \sup_{x\in X} |f(x) - g(x)|$.
- **Theorems 3&4** Extend results to Sobolev spaces under the $\ell_p, 1 \leq p < \infty$ and supremum norm.

Another Common Approach to Shallow NN Approximation

- Key assumption: $\sigma(x)$ is not a polynomial
- This method can be found in Section 3 of Pinkus (1999)
- This is also a non-constructive, i.e., existence type, approach.

The following result can be inferred from Section 3 of Pinkus (1999) (see Proposition 3.8):

Theorem Assume that $\sigma : \mathbb{R} \to \mathbb{R}$ is not a polynomial and that it is bounded and Reimann-integrable on any finite interval. Then, for any $f \in C([0,1]^d)$, there exists a shallow (1 hidden layer) NN approximation $\hat{f}$ such that $\sup_{\boldsymbol{x} \in [0,1]^d} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| < \epsilon$.

Outline of the proof: show that

1. Show that this neural network can approximate one-dimensional polynomials of any degree.
2. In particular we can create all polynomial basis: $(1, x, x^2, \ldots, x^k, \ldots)$. This further implies that we can, we can obtain all basis of the form $\mathbb{R}^d$: $(1, \boldsymbol{x} \cdot \boldsymbol{w}, (\boldsymbol{x} \cdot \boldsymbol{w})^2, \ldots, (\boldsymbol{x} \cdot \boldsymbol{w})^k, \ldots)$, implying the Fourier basis $e^{\boldsymbol{x} \cdot \boldsymbol{w}}$.

Computational Complexity of a NN Approximation

Recall the Bochner-Riesz approximation on d-torus $(C(T^d))$:

$$\hat{f}_R(\boldsymbol{x}) := \sum_{\boldsymbol{m}:\|\boldsymbol{m}\|_2<R} \left(1-\frac{\|\boldsymbol{m}\|_2^2}{R^2}\right)^{\alpha} \left(a_{\boldsymbol{m}R}\cos(2\pi i\boldsymbol{m}\cdot\boldsymbol{x}) - a_{\boldsymbol{m}I}\sin(2\pi i\boldsymbol{m}\cdot\boldsymbol{x})\right).$$

Hence, we can obtain a NN approximation with the 2-step procedure:

1. Approximate uniformly with NNS 1D functions $\cos(u)$ and $\sin(u)$ for $|u| \leq 2\pi\sqrt{d}R$. (Note that $\boldsymbol{x} \in [0,1]^d$ and $\|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|_2$ imply $|\boldsymbol{m}\cdot\boldsymbol{x}| \leq \sum|m_i| = \|\boldsymbol{m}\|_1 \leq \sqrt{d}\|\boldsymbol{m}\|_2 \leq \sqrt{d}R$.)
2. Replace the NN approximation of $\sin / \cos$ into the Fourier Bochner-Riesz approximation.

Complexity:

- Under some conditions: $|f(\boldsymbol{x}) - \hat{f}_R(\boldsymbol{x})| = O(1/R) < \epsilon$
- The number of summands in the Bochner-Riesz approximation is $\sum_{\|\boldsymbol{m}\|_2<R} 1 \approx \text{Vol}(n\text{-Ball}) = O(R^d) = O(e^{d\log R}) \approx e^{d\log(1/\epsilon)}$
- $\Rightarrow$ Curse of dimensionality

Lipschitz Continuity

Commonly used notion of continuity, which is stronger than uniform continuity.

Definition Function $f(x) : \mathbb{R}^k \to \mathbb{R}^m$ is Lipschitz continuous if there exists a finite constant $K > 0$, such that

$$\|f(x_1) - f(x_2)\|_p < K\|x_1 - x_2\|_p,$$

where $\|x_1 - x_2\|_p = \sqrt[p]{\sum_i |x_i|^p}$ is the L^p norm.

Remarks

- $\|x_1 - x_2\|_p = \sqrt[p]{\sum_i |x_i|^p}$ can be replaced with any metric, and thus, Lipschitz continuity can be defined on any metric space.
- Implies U.C.: $\|x_1 - x_2\|_p < \epsilon \Rightarrow \|f(x_1) - f(x_2)\|_p < K\epsilon$
- On a closed finite (compact) set

Continuous different. $\Rightarrow$ Lipschitz cont. $\Rightarrow$ uniform cont. $\Leftrightarrow$ ordinary cont.

- Strictly stronger than U.C.: $f(x) = \sqrt{x}$ is U.C. on $[0, 1]$ but not L.C.
- Can extend U.C. to infinite (non compact) sets.

Simple Function Approximation in $\mathbb{R}^d$

How about multivariate functions $f(x) : [0,1]^d \to \mathbb{R}$?

- Build piecewise constant (or linear) approximations over small hypercubes of size δ^d using sigmoids (or ReLUs)

Example:

- Let $\delta = 1/n$, d - dimension of $\boldsymbol{x}$, Euclidean norm $\|\boldsymbol{x}\|_2 = \sqrt{\sum x_i^2}$
- If f is Lipschitz continuous: $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2 \leq K\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$
- Then, piecewise constant approximation, $\hat{f}$, of f on hypercubes c_i of size $[0, 1/n]^d$ will be ϵ-accurate if $K\sqrt{d}/n \leq \epsilon$ since

$$
\begin{aligned}
\|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\|_2 &\leq \sup_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in c_i} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2 \\
&\leq K \sup_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in c_i} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 = K\frac{\sqrt{d}}{n} \leq \epsilon
\end{aligned}
$$

Where is the problem?

Curse of Dimensionality

- To achieve this accuracy, we need the number of grid point

$$\# \text{ of grid points} = n^d \geq \left(\frac{K\sqrt{d}}{\epsilon} \right)^d$$

- Example: let $K = 1, d = 10,000, \epsilon = 1/100$

$$\# \text{ of grid points} \geq 10^{40,000}$$

  Not feasible.

- Can we do better than this?
  - Fourier doesn't work, simple functions don't work. How about some other basis?
  - Is it possible to find an accurate approximation in high dimensions with a fixed set of basis?

Barron Functions (1993)

Let $f : \mathbb{R}^d \to \mathbb{R}$. Barron considers a class of functions Γ_C, $C > 0$, satisfying the following integrability condition

$$C_f = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2 |\tilde{f}(\boldsymbol{\omega})| d\boldsymbol{\omega} \leq C, \tag{1}$$

where $\|\boldsymbol{\omega}\|_2 = \sqrt{\boldsymbol{\omega} \cdot \boldsymbol{\omega}}$ and $f(\boldsymbol{x}), \tilde{f}(\boldsymbol{\omega})$ is the Fourier transform pair

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

Remark: Functions in Γ_C are continuously differentiable and finite with the gradient having representation

$$\nabla f(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} \widetilde{\nabla f}(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} i\boldsymbol{\omega} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega},$$

i.e., integrability in Equation (3) is equivalent integrability of the Fourier transform of the gradient.

Lower Bound on Approximation With n Fixed Basis

We'll show that picking in advance n bases can never achieve a uniformly accurate approximation.
Consider functions, $f : [0,1]^d \to \mathbb{R} \subset \Gamma_C$ with

$$d(f,g)^2 = \|f-g\|_2^2 = \int_{[0,1]^d} (f(\boldsymbol{x}) - g(\boldsymbol{x}))^2 d\boldsymbol{x}.$$

The distance between a function f and a set of functions G is defined as $\inf_{g \in G} d(f,g)$.

- Consider a set of fixed basis $h_1, \ldots, h_n$ and a set of all linear combinations of these basis be denoted as span$(h_1, \ldots, h_n)$.
- Kolmogorov n-width of the class Γ_C

$$W_n := \inf_{h_1,\ldots,h_n} \sup_{f \in \Gamma_C} d(f, \mathsf{span}(h_1, \ldots, h_n)).$$

If W_n is small, then Γ_C can be approximated well with n basis.

Lower Bound on Approximation With n Fixed Basis

Theorem 6 (Barron) For any choice of fixed basis functions $h_1, h_2, \ldots, h_n$,

$$\sup_{f \in \Gamma_C} d(f, \mathsf{span}(h_1, \ldots, h_n)) \geq \kappa \frac{C}{d} \left(\frac{1}{n}\right)^{1/d},$$

where $\kappa = 1/(8\pi e^{\pi-1})$.

The proof is based on the following lemma
Lemma No linear subspace of dimension n can have a squared distance less than $1/2$ from every basis function in an orthonormal basis of a $2n$-dimensional space.
Proof: Let $G_n = \mathsf{span}(g_1, \ldots, g_n)$ be a linear subspace of dimension n. Then

- $d^2(e_j, G_n) = \|e_j\|^2 - \|Pe_j\|^2 = 1 - \|Pe_j\|^2$, where P is the projection onto G_n
- Hence, it is enough to show $\|Pe_j\|^2 \leq 1/2$ for some j

Lower Bound on Approximation With n Fixed Basis

Proof of the lemma:

- Without loss of generality, assume that $g_1, \ldots, g_n$ are orthonormal basis of G_n
- Then, $Pe_j = \sum_{i=1}^{n} \langle e_j, g_i \rangle g_i$, implying

$$
\begin{aligned}
\frac{1}{2n} \sum_{j=1}^{2n} \|Pe_j\|^2 &= \frac{1}{2n} \sum_{j=1}^{2n} \sum_{i=1}^{n} \langle e_j, g_i \rangle^2 \\
&= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2n} \langle e_j, g_i \rangle^2 \\
&= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2n} \|g_i\|^2 = \frac{n}{2n} = \frac{1}{2}
\end{aligned}
$$

- Hence, $\|Pe_j\|^2 \leq 1/2$ for some j, which completes the proof.

Lower Bound on Approximation With n Fixed Basis

Proof of the theorem: Let $h_1, h_2, \ldots,$ be function $\cos(2\pi \boldsymbol{k}\cdot\boldsymbol{x}), \boldsymbol{k}\in\mathbb{Z}^d$ ordered in terms of the ℓ_1 norm $\|\boldsymbol{k}\|_1 = \sum_{i=1}^d |k_i|$.

- Let H_{2n} be the span of $h_1, \ldots h_{2n}$
- Let G_n be the n-dimensional linear subspace of H_{2n}. Then

$$\begin{aligned}
W_n &\geq \inf_{G_n} \sup_{f\in H_{2n}\cap\Gamma_C} d(f, G_n)\\
&\geq \inf_{G_n} \sup_{f\in\{C\cos(2\pi\boldsymbol{k}_j\cdot\boldsymbol{x})/(2\pi\|\boldsymbol{k}_j\|_1), j=1,\ldots,2n\}} d(f, G_n)\\
&\geq \min_{j=1,\ldots,2n} \frac{C}{2\pi\|\boldsymbol{k}_j\|_1} \inf_{G_n} \sup_{f\in\{\cos(2\pi\boldsymbol{k}_j\cdot\boldsymbol{x}), j=1,\ldots,2n\}} d(f, G_n)\\
&\geq \min_{j=1,\ldots,2n} \frac{C}{2\pi\|\boldsymbol{k}_j\|_1}\frac{1}{2} \quad \text{(lemma)}\\
&\geq \frac{C}{4\pi m}
\end{aligned}$$

where $\binom{m+d}{d} \geq 2n$, implying (Stirling's) $m < e^{\pi-1}dn^{1/d}$, and completing the proof.

Lower Bound on Approximation With n Fixed Basis

Simple geometric intuition:

Claim For any set of n points $\boldsymbol{x}_i \in [0,1]^d, i = 1, \dots, n$, there exists a point $\boldsymbol{x}_0 \in [0,1]^d$ such that

$$\min_{i=1,\dots,n} \|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_2 \geq \frac{1}{4} \frac{1}{n^{1/d}}$$

(prove it)

Hence

- Fixed, predetermined, set of n basis cannot do a good job in high dimensions
- Can we do better if we search for bases, like in NN approach?
- Let us look at more general results on uniform bounds.

Smooth Function Classes

Recall the usual notation:

- For $\boldsymbol{s} \in \mathbb{Z}_+^d, \boldsymbol{s} = (s_1, \dots, s_d) \geq \mathbf{0}, |\boldsymbol{s}| = s_1 + \cdots + s_d$, we can define a differential operator

$$D^{\boldsymbol{s}} := \frac{\partial^{|\boldsymbol{s}|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$$

  and a class of functions with up to $s \geq 1$ partial derivatives

$$C^s(\mathbb{R}^d) := \left\{ f : D^{\boldsymbol{k}} f \in C(\mathbb{R}^d), \text{ for all } |\boldsymbol{k}| \leq s \right\}$$

- For these functions, it make sense to measure the distance between the functions and all of their derivatives, motivating the following norm

$$\|f\|_{s,p} := \begin{cases} \left(\sum_{0 \leq |\boldsymbol{k}| \leq s} \|D^{\boldsymbol{k}} f\|_p^p\right)^{1/p}, & 1 \leq p < \infty \\ \max_{0 \leq |\boldsymbol{k}| \leq s} \|D^{\boldsymbol{k}} f\|_\infty, & p = \infty \end{cases}$$

  where

$$\|g\|_p := \begin{cases} \left(\int_K |g(\boldsymbol{x})|^p d\boldsymbol{x}\right)^{1/p}, & 1 \leq p < \infty \\ \operatorname{ess\,sup}_{\boldsymbol{x} \in K} |g(\boldsymbol{x})|, & p = \infty \end{cases}$$

  over some domain $K \in \mathbb{R}^d$.

Sobolev Spaces

- We will pick K to be a d-ball (or d-cube)

$$B^d = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2^2 = x_1^2 + \cdots + x_d^2 \leq 1\}$$

- Sobolev space $\mathcal{W}_p^s = \mathcal{W}_p^s(B^d)$ is a completion of $C^s(B^d)$ with the respect to the p norm. In addition we assume that all $f \in \mathcal{W}_p^s$ have a bounded norm $\|f\|_{s,p}$.
- Also, we define a space of all NN approximations with n neurons

$$\mathcal{N}_n := \left\{ \sum_{i=1}^{n} a_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i), \boldsymbol{w}_i \in \mathbb{R}^d, a_i, b_i \in \mathbb{R} \right\}$$

- and d-width (d same as n before)

$$E_{s,p}(\mathcal{W}_p^s, \mathcal{N}_n) := \sup_{f \in \mathcal{W}_p^s} \inf_{\hat{f} \in \mathcal{N}_n} \|f - \hat{f}\|_{s,p}$$

Uniform Upper and Lower Bounds

For one layer NNs this is well understood:

Theorem Assume that σ is infinitely differentiable and not a polynomial. Also, assume that the NN approximation $\hat{f}_n$, as a mapping $\mathcal{W}_p^s \to \mathcal{N}_n$, is continuous in parameters $(\boldsymbol{w}, a, b)$, than, for each $1 \leq p \leq \infty, m \geq 1$, there is a finite constant C, such that

$$C^{-1} n^{-s/d} \leq E_{s,p}(\mathcal{W}_p^s, \mathcal{N}_n) \leq C n^{-s/d}$$

Remarks:

- Proving this would take too much time away from deep networks. Instead, see Theorem 1 and the notes after it in the recent survey by Poggio et al. (2017). For more rigorous presentation see Section 6 in Pinkus (1999), Theorems 6.6 & 6.8.
- An immediate consequence of the preceding theorem is that if we want a uniform ϵ approximation, then the number of neurons n needed is

  $$n^{-s/d} = \epsilon \;\Rightarrow\; n = O(\epsilon^{-d/s}) \quad \text{(curse of dimensionality)}$$

  We will use this for comparison with deep learning results.
- Can we do better than this?

Results by Barron (1993)

Recall Barron class of functions from a few slides ago:
Let $f : \mathbb{R}^d \to \mathbb{R}$. Barron considers a class of functions Γ_C, $C > 0$, satisfying the following integrability condition

$$C_f = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2 |\tilde{f}(\boldsymbol{\omega})| d\boldsymbol{\omega} \leq C, \tag{2}$$

where $\|\boldsymbol{\omega}\|_2 = \sqrt{\boldsymbol{\omega} \cdot \boldsymbol{\omega}}$ and $f(\boldsymbol{x}), \tilde{f}(\boldsymbol{\omega})$ is the Fourier transform pair

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

Remark: Functions in Γ_C are continuously differentiable and finite with the gradient having representation

$$\nabla f(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} \widetilde{\nabla f}(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} i\boldsymbol{\omega} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega},$$

i.e., integrability in Equation (3) is equivalent integrability of the Fourier transform of the gradient.

Uniform Bounds of Barron

- Let $\Gamma_{C,B}$ be functions from Γ_C with a bounded domain $B \in \mathbb{R}^d$

Typical result in Barron:

Theorem 1 (Barron) For any $f \in \Gamma_{C,B}$, probability measure μ, there exists a NN approximation $f_n(\boldsymbol{x})$ with n neurons, such that

$$\int_B (f(\boldsymbol{x}) - f_n(\boldsymbol{x}))^2 \mu(dx) \leq \frac{4C^2}{n}.$$

The linear coefficients, c_k, in f_n can be chosen to satisfy $\sum |c_k| \leq 2C$.

- Barron's bound is $O(1/\sqrt{n})$, no curse of dimensionality!: Where is his magic?

Comments on Barron

Simple answer: Barron defined a much smaller class of functions, which is much more suitable for approximation with NNs

$$\Gamma_C \subset \mathcal{W}^1_\infty \subset \mathcal{W}^1_2$$

Let $f : \mathbb{R}^d \to \mathbb{R}$. Barron considers a class of functions Γ_C, $C > 0$, satisfying the following integrability condition

$$C_f = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2 |\tilde{f}(\boldsymbol{\omega})| d\boldsymbol{\omega} \leq C, \tag{3}$$

where $\|\boldsymbol{\omega}\|_2 = \sqrt{\boldsymbol{\omega} \cdot \boldsymbol{\omega}}$ and $f(\boldsymbol{x}), \tilde{f}(\boldsymbol{\omega})$ is the Fourier transform pair

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}\cdot\boldsymbol{x}} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

Remark: Informally, one could think of condition (3) as f being "band limited", i.e., no significant amount high frequency components.

Comments of Barron

Theorem 1 (Barron) For any $f \in \Gamma_{C,B}$, probability measure μ, there exists a NN approximation $f_n(\boldsymbol{x})$ with n neurons, such that

$$\int_B (f(\boldsymbol{x}) - f_n(\boldsymbol{x}))^2 \mu(dx) \leq \frac{4C^2}{n}. \tag{4}$$

The linear coefficients, c_k, in f_n can be chosen to satisfy $\sum |c_k| \leq 2C$.

Proof in Barron is very long. The following is much shorter argument of Makovoz (1996), see also Theorem 6.13 in Pinkus (1996), based on the following observation

$$\Gamma_{C,B} \subset \left\{ f(x) : f(x) = \int_{\|\boldsymbol{w}\|_2 = 1, |b| \leq 1} a(\boldsymbol{w}, b) \sigma(\boldsymbol{w} \cdot \boldsymbol{x} - b) d\pi(\boldsymbol{w}, b) \right\}$$

where π is the probability measure and $\int |a(\boldsymbol{w}, b)| d\pi(\boldsymbol{w}, b) \leq C$.
Proof of (4): Make n random samples $(\boldsymbol{w}_i, b_i)$ from distribution π and make $\hat{f}(\boldsymbol{x}) = 1/n \sum_{i=1}^n a(\boldsymbol{w}_i, b_i) \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i)$. Then, (4) is a Var($\hat{f}$).

General Takeaways

- In terms of approximating a general class of smooth functions on bounded domain:
  - The worst case bounds for NNs with n parameters are equivalent to any other model with n-parameters, e.g., Fourier, wavelets, polynomials, etc.
- We can do better if we assume more structure, i.e., smaller classes of functions, like in Barron.
- How does depth help?

Impact of Depth

- With two hidden layers it is easy to create a pulse in $\mathbb{R}^d$:
  - Hence, we can obtain an elementary proof of Cybenko, and others, using only calculus. Haven't seen this idea in the literature, so please refer to this lecture notes if you decide to use it. I might write a short note about it☺.
- Namely, let $\sigma(x) = 1_{\{x \geq 0\}}$ and $\boldsymbol{e}_i$ be the standard ith base in $\mathbb{R}^d$, then for $d \geq 2, \boldsymbol{x} \in \mathbb{R}^d$,

$$\Pi_\delta(\boldsymbol{x}) := \sigma\left(\sum_{i=1}^{d}(\sigma(\boldsymbol{x}\cdot\boldsymbol{e}_i) - \sigma(\boldsymbol{x}\cdot\boldsymbol{e}_i - \delta)) - d + 1/2\right) = \begin{cases} 1, & \text{if } x \in [0,\delta]^d, \\ 0, & \text{otherwise.} \end{cases}$$

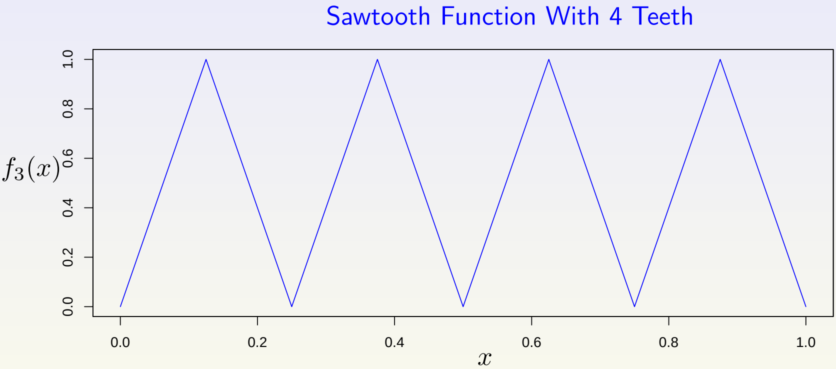
- Now, choose large n, set $\delta = 1/n$, and consider a grid $S = [0, 1/n, 2/n, \ldots, (n-1)/n]^d$. With this grid we can obtain an arbitrarily accurate NN approximation for any $f \in C([0,1]^d)$ using

$$\hat{f}(\boldsymbol{x}) = \sum_{s \in S} f(s)\Pi_{1/n}(\boldsymbol{x} - \boldsymbol{s}).$$

- If instead of a step function $\sigma(x) = 1_{\{x \geq 0\}}$, we have a general sigmoid, ReLU, generalized polynomial spline, we can combine scaling and finite differences to obtain an arbitrarily close approximation to a prefect pulse.

Impact of Depth

Example: "High frequency" sawtooth function
Consider sawtooth $f_n(x) : [0,1] \to [0,1]$ with 2^{n-1} equally spaced teeth



How many ReLU neurons are needed to represent f_n with 1 hidden layer?

2^n neurons

Depth To The Rescue: Hat Function

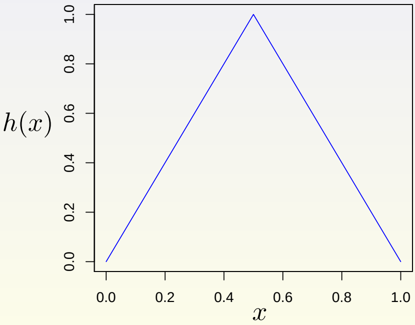
Use 3 ReLUs, $\sigma(x) : \max(0, x)$, to define "hat" function

$$h(x) = 2\sigma(x) - 4\sigma(x - 1/2) + 2\sigma(x - 1)$$

or, equivalently

$$h(x) = \sigma\left(2\sigma(x) - 4\sigma(x - 1/2)\right)$$

Hat Function

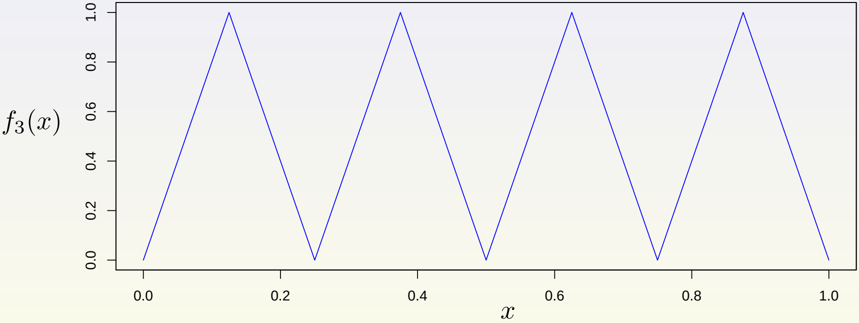


Depth To The Rescue: Hat Function Composition

Define composition of functions $h^{(n)}$ as

$$h^{(n)}(x) = h \circ h \cdots h \circ h(x) = h(h(\cdots h(x) \cdots))$$

$$f_3(x) = h^{(3)}(x) = h(h(h(x)))$$



Claim Function f_n with 2^{n-1} teeth can be represented by a deep network of depth n and $3n$ ReLU neurons.

Depth To The Rescue: Hat Function Composition

Key idea: Hat function composition - exploited in:

- Classification: Telgarsky (2015)
- Function approximation: Yarotsky (2017), Liang and Srikant (2017), Montanelli and Du (2018)

We will cover parts of these papers today.

Hat function composition allows

more efficient representation = smaller # of neurons

Intuition:

- Depth increase the number of oscillations (frequency) exponentially
- While width increases it additively

Hat Function Composition - Classification

Telgarsky (2015) - Classification setup:

- Use ReLU: $\sigma(x) = \max(0, x)$ (for the remainder of today's lecture)
- Classification problem: 2 classes - $\{0, 1\}$
- Classifier: for any function $f : \mathbb{R} \to \mathbb{R}$, define classifier

$$\tilde{f}(x) = 1_{\{f(x) \geq 1/2\}}$$

- Empirical error: for a sequence of points $\{(x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$, define the empirical error

$$\bar{R}(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(x_i) \neq y_i\}}$$

- $\mathcal{H}(\sigma; m, l)$: Hypothesis class of functions that can be constructed using NN with l layers, each with width of at most m nodes.

Telgarsky's (2015) Theorem

Theorem Let positive integer k, number of layers l, and number of nodes per layer m be given with $m \leq 2^{(k-3)/l-1}$. Then there exists a collection of $n := 2^k$ points $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in [0,1]$ and $y_i \in \{0,1\}$ such that

$$\min_{f \in \mathcal{H}(\sigma;2,2k)} \bar{R}(f) = 0 \quad \text{and} \quad \min_{f \in \mathcal{H}(\sigma;m,l)} \bar{R}(f) \geq \frac{1}{6}$$

Proof: Choose $x_i = i2^{-k}$ and set $y_{2i} = 0, y_{2i+1} = 1, 0 \leq i \leq 2^k - 1$

Upper bound: Showing that $2 \times (2k)$ network can produce a perfect classifier: $\bar{R}(f) = 0$.

- Construct k-fold composition of "hat" function, $h^{(k)}(x)$, using $2 \times (2k)$ neural net, i.e., $h^{(k)} \in \mathcal{H}(\sigma; 2, 2k)$
- Note that $h^{(k)}(x)$ perfectly reproduces data

$$h^{(k)}(x_i) = y_i$$

Hence, $\bar{R}(h^{(k)}) = 0$

Telgarsky's (2015) Theorem

Proof: *Lower bound:* Showing that $m \times l$ network cannot produce a perfect classifier, i.e., $\bar{R}(f) \geq 1/6$. This is a bit more involved.

- $f : \mathbb{R} \to \mathbb{R}$ is t-*sawtooth* if it is piecewise linear (affine) with t linear pieces.

Lemma 1 Let $\{(x_i, y_i)\}_{i=1}^n, n = 2^k$, be the set of points defined earlier. Then, every t-sawtooth function $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$\bar{R}(f) \geq \frac{n - 4t}{3n}.$$

Proof lemma. Recall the classifier $\tilde{f}(x) = 1_{\{f(x) \geq 1/2\}}$.

- f crosses $1/2$ at most once in every interval of its linear growth.
- $\tilde{f}(x)$ is piecewise constant (0 or 1) over $2t$ intervals. Meaning, n points must fall in $2t$ buckets (intervals).
- At least $n - 4t$ intervals (buckets) must have at least 3 points.
- At least 1/3 of points in each of the $n - 4t$ are labeled incorrectly (since points alternate). This completes the proof of the lemma.

Telgarsky's (2015) Theorem

Proof: *Lower bound:* (continued)

Lemma 2 If f is p-sawtooth and g is q-sawtooth, then $f+g$ is at most $p+q$-sawtooth and $f \circ g$ is at most pq-sawtooth.

Proof Lemma 2.

- Addition: "joints" of f can fall within linear intervals of g, and adding a linear function does not change the sawtooth degree.
- Composition: each interval on which f is linear has as its domain at most the range of g. Hence, on such intervals, $f(g(x))$ is at most q-sawtooth. Since there are p such intervals, $f \circ g$ is at most pq-sawtooth.

Now, return to the *proof of the theorem.*

- Note: $\sigma(x)$ is 2-sawtooth. Hence, function f, made by $(m \times l)$-NN is at most $(2m)^l$-sawtooth. Thus, by assumption $m \leq 2^{(k-3)/l-1}$,

$$(2m)^l \leq 2^{k-3}$$

- Finally, by Lemma 1,

$$\bar{R}(f) \geq \frac{2^k - 4 \cdot 2^{k-3}}{3 \cdot 2^k} = \frac{1}{6},$$

which completes the proof.

General Function Approximation

Use "hat" function composition to efficiently represent general functions, $f : [0,1]^d \to \mathbb{R}$

efficient representation = smaller # of neurons

Papers: Yarotsky (2017), Liang & Srikant (2017), Montanelli & Du (2018)

General approximation plan:

1. Use "hat" function composition to efficiently represent some nice functions, e.g. polynomials, f_P, by neural nets, f_N, so that, in some norm
$$\|f_P - f_N\| \leq \frac{\epsilon}{2}$$
2. Show that polynomials f_P approximate well more general functions, f, (in some larger functional space, e.g., Sobolev space)
$$\|f - f_P\| \leq \frac{\epsilon}{2}$$
3. Put the preceding steps together via triangular inequality
$$\|f - f_N\| = \|(f - f_P) + (f_P - f_N)\| \leq \|f - f_P\| + \|f_P - f_N\| \leq \epsilon$$

Approximating Polynomials: $f(x) = x^2$

First step in approximating multivariate polynomials on $[0,1]^d$ is to construct a quadratic on $[0,1]$:

$$f(x) = x^2$$

Proposition (2, Yarotski, 2017) For any $\epsilon > 0$, the function $f(x) = x^2$ on the $[0,1]$ can be approximated by a ReLU network function, $f_N(x)$, having depth, number of weights and computation units $O(\ln(1/\epsilon))$, such that

$$|f(x) - f_N(x)| < \epsilon$$

Proof: The key observation is that $f(x) = x^2$ can be approximated by a linear combination of "hat" composition functions $h^{(s)}$.

- Let $f_m(x)$ be a piecewise linear interpolation of $f(x)$ with $2^m + 1$ uniformly distributed break points $k2^{-m}, k = 0, 1, \ldots, 2^m$

$$f\left(\frac{k}{2^m}\right) = f_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2$$

Approximating Polynomials: $f(x) = x^2$

Proof: (continued)

- Next,

$$|f_m(x) - f(x)| \leq \frac{1}{2^{2m+2}}$$

- Also, note that

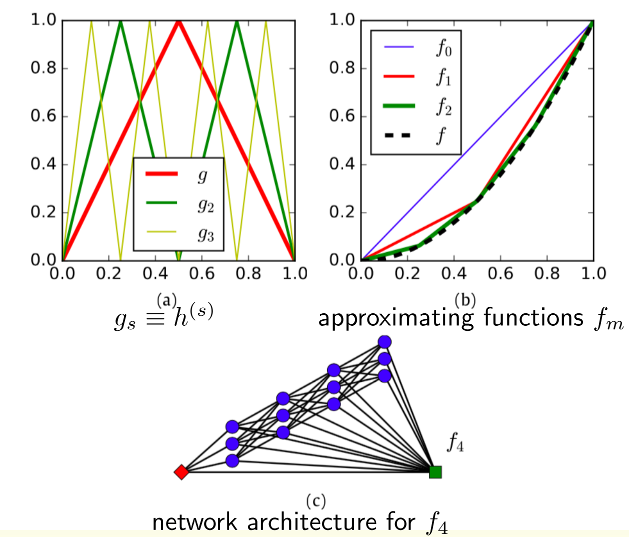
$$f_{m-1}(x) - f_m(x) = \frac{h^{(m)}(x)}{2^{2m}}$$

- From which it follows

$$f_m(x) = x - \sum_{s=1}^{m} \frac{h^{(s)}(x)}{2^{2s}}$$

Approximating Polynomials: $f(x) = x^2$

Efficient approximation of $f(x) = x^2$



(a)

$g_s \equiv h^{(s)}$

(b)

approximating functions f_m

(c)

network architecture for f_4

Approximating Multivariate Polynomials

Next step in approximating multivariate polynomials on $[0,1]^d$ is to construct a product function on $[0,1]^2$:

$$f(x,y) = xy$$

Proposition (3, Yarotski, 2017) For any $\epsilon > 0$, the function $f(x,y) = xy$ on the $[0,1]^2$ can be approximated by a ReLU network function, $f_N(x,y)$, having depth, number of weights and computation units $O(\ln(1/\epsilon))$, such that

$$|f(x,y) - f_N(x,y)| < \epsilon$$

Proof: Follows immediately from Proposition 2 and

$$xy = \frac{1}{2}((x+y)^2 - x^2 - y^2)$$

Approximating General Smooth Functions

1. Use Propositions 1&2 to approximate polynomials of any degree on $[0,1]^d$
2. Assume that functions, f, are smooth enough (have enough derivatives), to bound the error between f and its polynomial - Taylor expansion

Putting 1. and 2. together yields
Theorem (1, Yarotski, 2017) For any $\epsilon > 0$, a function $f(\boldsymbol{x})$ on the $[0,1]^d$ with up to n partial derivatives can be approximated by a ReLU network function, $f_N(\boldsymbol{x})$, having depth $O(\ln(1/\epsilon))$, and number of weights and computation units $O(e^{-d/n}\ln(1/\epsilon))$, such that

$$|f(\boldsymbol{x}) - f_N(\boldsymbol{x})| < \epsilon$$

Today's Class Reading

- Universal approximation bounds for superpositions of a sigmoidal function, by Barron 1993.
- Approximation theory of the mlp model in neural networks, by Pinkus, 1999.

Expressive Power Depth We tried to answer the question: Can deep neural networks of depth $(d+1)$ express functions much more efficiently in terms of the number of neurons compared to networks of depth d?

- Representation Benefits of Deep Feedforward Networks, by Telgarsky, 2015.
- Error bounds for approximations with deep ReLU networks, by Yarotsky, 2017.
- WHY DEEP NEURAL NETWORKS FOR FUNCTION APPROXIMATION?, Liang and Srikant, 2017.
- NEW ERROR BOUNDS FOR DEEP RELU NETWORKS USING SPARSE GRIDS, by Montanelli and Du, 2018.
- Check references and citations of these papers for additional reading.

Next Class

- Cover some more results from today's papers.
- **Optional reading:** If interested in how to approximate functions that don't always have derivatives, check: weak derivatives, distributions and Sobolev spaces. (E.g., see Functional Anal. by Rudin, or Real Anal. by Folland.) In general, don't worry about these, I might say a few words next week.
- Will cover some ideas from: The Power of Depth for Feedforward Neural Networks, by Eldan and Shamir, 2016.
  The techniques in this paper are different from today's papers.
- Start a new topic (hopefully): maybe connection between Neural Nets and Kernels. Some good books on Kernels: (available online through CU Library)
  1. Chapters 1, 2, 13-16 in: B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.
  2. Chapters 1, 5.3, 6, 7 in (this book is mathematically advanced) A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer, 2004.

Have Fun!