Representation Power of Feedforward Neural Networks

Based on work by Barron (1993), Cybenko (1989), Kolmogorov (1957)

Matus Telgarsky <mtelgars@cs.ucsd.edu>

Feedforward Neural Networks

- Two node types:
  - Linear combinations:
$$x \quad \mapsto \quad \sum_i w_i x_i + w_0.$$
  - Sigmoid thresholded linear combinations:
$$x \quad \mapsto \quad \sigma\left(\langle w, x \rangle + w_0\right).$$

Feedforward Neural Networks

- Two node types:
  - Linear combinations:
    $$x \quad \mapsto \quad \sum_i w_i x_i + w_0.$$
  - Sigmoid thresholded linear combinations:
    $$x \quad \mapsto \quad \sigma\left(\langle w, x \rangle + w_0\right).$$
- What can a network of these nodes represent?

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i x_i && \text{one layer,}\\
&\sum_{i=1}^{n} w_i \sigma\left(\sum_{j=1}^{n_i} w_{ji} x_j + w_{j0}\right) && \text{two layers,}\\
&\quad\vdots && \quad\vdots
\end{aligned}
$$

Forget about 1 layer.

- Target set $[0,3]$; target function $\mathbb{1}[x \in [1,2]]$.

Forget about 1 layer.

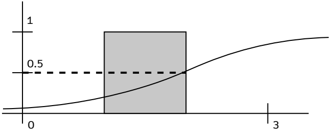
- Target set $[0,3]$; target function $\mathbb{1}[x \in [1,2]]$.
- Standard sigmoid $\sigma_s(x) := 1/(1+e^{-x})$.

Forget about 1 layer.

- Target set $[0, 3]$; target function $\mathbb{1}[x \in [1, 2]]$.
- Standard sigmoid $\sigma_s(x) := 1/(1 + e^{-x})$.
- Consider sigmoid output at $x = 2$.

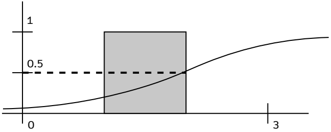
Forget about 1 layer.

- Target set $[0,3]$; target function $\mathbb{1}[x \in [1,2]]$.
- Standard sigmoid $\sigma_s(x) := 1/(1+e^{-x})$.
- Consider sigmoid output at $x = 2$.
  - $w \geq 0$ and $\sigma(2w + w_0) \geq 1/2$: mess up on right side.

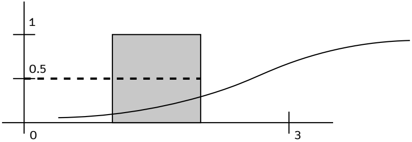


Forget about 1 layer.

- Target set $[0,3]$; target function $\mathbb{1}[x \in [1,2]]$.
- Standard sigmoid $\sigma_s(x) := 1/(1+e^{-x})$.
- Consider sigmoid output at $x = 2$.
  - $w \geq 0$ and $\sigma(2w + w_0) \geq 1/2$: mess up on right side.

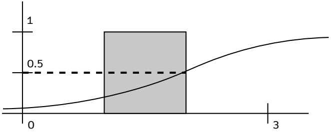


  - $w \geq 0$ and $\sigma(2w + w_0) < 1/2$: mess up on middle bump.

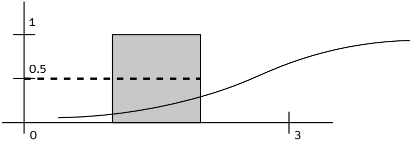


Forget about 1 layer.

- Target set $[0, 3]$; target function $\mathbb{1}[x \in [1, 2]]$.
- Standard sigmoid $\sigma_s(x) := 1/(1 + e^{-x})$.
- Consider sigmoid output at $x = 2$.
  - $w \geq 0$ and $\sigma(2w + w_0) \geq 1/2$: mess up on right side.



  - $w \geq 0$ and $\sigma(2w + w_0) < 1/2$: mess up on middle bump.



- Can symmetrize $(w < 0)$; no matter what, error $\geq 1/2$.

Meaning of “Universal Approximation”

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

Meaning of "Universal Approximation"

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

- For any $\epsilon > 0$, exists NN $\hat{f}$,

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

Meaning of “Universal Approximation”

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

- For any $\epsilon > 0$, exists NN $\hat{f}$,

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

  - This gives NNs $\hat{f}_i \to f$ pointwise.

Meaning of "Universal Approximation"

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

- For any $\epsilon > 0$, exists NN $\hat{f}$,

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

  - This gives NNs $\hat{f}_i \to f$ pointwise.
- For any $\epsilon > 0$, exists NN $\hat{f}$ and $S \subseteq [0,1]^n$, $m(S) \geq 1 - \epsilon$,

$$x \in S \quad \Longrightarrow \quad |f(x) - \hat{f}(x)| < \epsilon.$$

Meaning of "Universal Approximation"

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

- For any $\epsilon > 0$, exists NN $\hat{f}$,

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

  - This gives NNs $\hat{f}_i \to f$ pointwise.
- For any $\epsilon > 0$, exists NN $\hat{f}$ and $S \subseteq [0,1]^n$, $m(S) \geq 1 - \epsilon$,

$$x \in S \quad \Longrightarrow \quad |f(x) - \hat{f}(x)| < \epsilon.$$

  - If (for instance) bounded on S^c, gives NNs $\hat{f}_i \to f$ m-a.e..

Meaning of "Universal Approximation"

Target set $[0,1]^n$; target function $f \in \mathcal{C}([0,1]^n)$.

- For any $\epsilon > 0$, exists NN $\hat{f}$,

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

  - This gives NNs $\hat{f}_i \to f$ pointwise.
- For any $\epsilon > 0$, exists NN $\hat{f}$ and $S \subseteq [0,1]^n$, $m(S) \geq 1 - \epsilon$,

$$x \in S \quad \Longrightarrow \quad |f(x) - \hat{f}(x)| < \epsilon.$$

  - If (for instance) bounded on S^c, gives NNs $\hat{f}_i \to f$ m-a.e..

Goal: 2-NNs approximate continuous functions over $[0,1]^n$.

Outline

- 2-nn via functional analysis (Cybenko, 1989).
- 2-nn via greedy approx (Barron, 1993).
- 3-nn via histograms (Folklore).
- 3-nn via wizardry (Kolmogorov, 1957).

Overview of Functional Analysis proof (Cybenko, 1989)

- Hidden layer as a basis:

$$B := \{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}.$$

Overview of Functional Analysis proof (Cybenko, 1989)

- Hidden layer as a basis:

$$B := \{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}.$$

- Want to show $\mathrm{cl}(\mathrm{span}(B)) = \mathcal{C}([0,1]^n)$.

Overview of Functional Analysis proof (Cybenko, 1989)

- Hidden layer as a basis:

$$B := \{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}.$$

- Want to show $\mathrm{cl}(\mathrm{span}(B)) = \mathcal{C}([0,1]^n)$.
- Work via contradiction: if $f \in \mathcal{C}([0,1]^n)$ far from $\mathrm{cl}(\mathrm{span}(B))$, can bridge the gap with a sigmoid.

Abstracting σ

- Cybenko needs σ *discriminates*:

$$\mu = 0 \quad \Longleftrightarrow \quad \forall w, w_0 \,.\, \int \sigma(\langle w, x\rangle + w_0) d\mu(x) = 0.$$

Abstracting σ

- Cybenko needs σ *discriminates*:

$$\mu = 0 \quad \iff \quad \forall w, w_0 \,.\, \int \sigma(\langle w, x\rangle + w_0) d\mu(x) = 0.$$

- Satisfied for the standard choices

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

$$\frac{1}{2}(\tanh(x) + 1) = \frac{1}{2}\left(\frac{e^x - e^{-x}}{e^x + e^{-x}} + 1\right) = \sigma_s(2x).$$

Abstracting σ

- Cybenko needs σ *discriminates*:

$$\mu = 0 \qquad \Longleftrightarrow \qquad \forall w, w_0 \,.\, \int \sigma(\langle w, x\rangle + w_0) d\mu(x) = 0.$$

- Satisfied for the standard choices

$$\sigma_s(x) = \frac{1}{1+e^{-x}},$$
$$\frac{1}{2}(\tanh(x)+1) = \frac{1}{2}\left(\frac{e^x - e^{-x}}{e^x + e^{-x}} + 1\right) = \sigma_s(2x).$$

- Most results today only need σ approximates $\mathbb{1}[x \geq 0]$:

$$\sigma(x) \to \begin{cases} 1 & \text{as } x \to +\infty, \\ 0 & \text{as } x \to -\infty. \end{cases}$$

  Combined with σ bounded&measurable gives discriminatory (Cybenko, 1989, Lemma 1).

Proof of Cybenko (1989)

- Consider the subspace

$$S := \operatorname{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right) \quad .$$

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{ \sigma(\langle w, x \rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R} \} \right) \right).$$

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\operatorname{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
    - Can think of projecting f onto $S^{\perp}$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
    - Can think of projecting f onto $S^\perp$.
    - Don't need inner products: Hahn-Banach Theorem.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
    - In $\mathbb{R}^n$, linear functions have form $\langle \cdot, y\rangle$ for some $y \in \mathbb{R}^n$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{ \sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R} \} \right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
    - In $\mathbb{R}^n$, linear functions have form $\langle \cdot, y\rangle$ for some $y \in \mathbb{R}^n$.
    - With infinite dimensions, integrals give inner products.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\} \right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
    - In $\mathbb{R}^n$, linear functions have form $\langle \cdot, y\rangle$ for some $y \in \mathbb{R}^n$.
    - With infinite dimensions, integrals give inner products.
    - A form of the Riesz Representation Theorem.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \operatorname{cl}\left(\operatorname{span}\left(\{ \sigma(\langle w, x \rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R} \} \right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
  - Contradiction: σ is discriminatory.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
  - Contradiction: σ is discriminatory.
    - $L_{|S} = 0$ implies $\forall w, w_0 \,.\, \int \sigma(\langle w, x\rangle + w_0) d\mu(x) = 0$.

Proof of Cybenko (1989)

- Consider the closed subspace

$$S := \operatorname{cl}\left(\operatorname{span}\left(\{ \sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R} \} \right) \right).$$

- Suppose (contradictorily) exists $f \in \mathcal{C}([0,1]^n) \setminus S$.
  - Exists bounded linear $L \neq 0$ with $L_{|S} = 0$.
  - Exists $\mu \neq 0$ so that $L(h) = \int h(x) d\mu(x)$.
  - Contradiction: σ is discriminatory.
    - $L_{|S} = 0$ implies $\forall w, w_0 \,.\, \int \sigma(\langle w, x\rangle + w_0) d\mu(x) = 0$.
    - But "discriminatory" means $\mu = 0 \iff \forall w, w_0 \,.\, \int \ldots$.

What was Cybenko's proof about?

- Set $S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.

What was Cybenko's proof about?

- Set $S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.
- Another way to look at it.

What was Cybenko's proof about?

- Set $S := \mathrm{cl}\left(\mathrm{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.
- Another way to look at it.
  - Given $f \in \mathcal{C}([0,1]^n)$, current approximant $\hat{f} \in S$.

What was Cybenko's proof about?

- Set $S := \operatorname{cl}\left(\operatorname{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.
- Another way to look at it.
  - Given $f \in \mathcal{C}([0,1]^n)$, current approximant $\hat{f} \in S$.
  - The residue $f - \hat{f}$ must have nonzero projection onto S.

What was Cybenko's proof about?

- Set $S := \operatorname{cl}\left(\operatorname{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.
- Another way to look at it.
  - Given $f \in \mathcal{C}([0,1]^n)$, current approximant $\hat{f} \in S$.
  - The residue $f - \hat{f}$ must have nonzero projection onto S.
  - Add residue projection into $\hat{f}$; does this now match f?

What was Cybenko's proof about?

- Set $S := \operatorname{cl}\left(\operatorname{span}\left(\{\sigma(\langle w, x\rangle + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}\right)\right)$.
- Another way to look at it.
  - Given $f \in \mathcal{C}([0,1]^n)$, current approximant $\hat{f} \in S$.
  - The residue $f - \hat{f}$ must have nonzero projection onto S.
  - Add residue projection into $\hat{f}$; does this now match f?
- Can this be turned into an algorithm?

Outline

- 2-nn via functional analysis (Cybenko, 1989).
- 2-nn via greedy approx (Barron, 1993).
- 3-nn via histograms (Folklore).
- 3-nn via wizardry (Kolmogorov, 1957).

The method of Barron (1993, Section 8)

Input: basis set B,

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \text{cl}(\text{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:

- Can rescale B so $f \in \text{cl}(\text{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \text{cl}(\text{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing

$$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)g)\|_2^2.$$

- Can rescale B so $f \in \text{cl}(\text{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \text{cl}(\text{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)
   $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)g)\|_2^2.$$

- Can rescale B so $f \in \text{cl}(\text{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \text{cl}(\text{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)
   $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g)\|_2^2.$$
   2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \text{cl}(\text{conv}(B))$. (or tweak alg.)

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   - 2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)

     $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g)\|_2^2.$$

   - 2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)
- Is this familiar?

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)

$$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)g)\|_2^2.$$

   2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)
- Is this familiar? Oh let's see..

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)

   $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g)\|_2^2.$$

   2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)
- Is this familiar? Oh let's see.. gradient descent, coordinate descent, Frank-Wolfe, projection pursuit, basis pursuit, boosting.. as usual, cf. Zhang (2003).

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)
   $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g)\|_2^2.$$
   2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)
- Is this familiar? **YES.**

The method of Barron (1993, Section 8)

Input: basis set B, target $f \in \mathrm{cl}(\mathrm{conv}(B))$.

1. Choose arbitrary $\hat{f}_0 \in B$.
2. For $t = 1, 2, \ldots$:
   2.1 Choose (α_t, g_t) apx minimizing (tolerance $\mathcal{O}(1/t^2)$)
   $$\inf_{\alpha \in [0,1], g \in B} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)g)\|_2^2.$$
   2.2 Update $\hat{f}_t := \alpha_t \hat{f}_{t-1} + (1-\alpha_t) g_t$.

- Can rescale B so $f \in \mathrm{cl}(\mathrm{conv}(B))$. (or tweak alg.)
- Is this familiar? **YES.**
- Barron carefully shows rate $\|f - \hat{f}_t\|_2^2 = \mathcal{O}(1/t)$.

Greedy approx as coordinate descent

- Intuition: linear operator A with basis B as "columns":

$$\hat{f}_t = Aw_t, \qquad g = A\mathbf{e}_j.$$

Greedy approx as coordinate descent

- Intuition: linear operator A with basis B as "columns":

$$\hat{f}_t = Aw_t, \qquad g = A\mathbf{e}_j.$$

- Split the update into two steps:

$$\inf_{g \in B} \|f - (\hat{f}_{t-1} + g)\|_2^2 \quad = \quad \inf_i \|f - (Aw_{t-1} + A\mathbf{e}_i)\|_2^2,$$

$$\inf_{\alpha \in [0,1]} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g_t)\|_2^2.$$

Greedy approx as coordinate descent

- Intuition: linear operator A with basis B as "columns":

$$\hat{f}_t = Aw_t, \qquad g = A\mathbf{e}_j.$$

- Split the update into two steps:

$$\inf_{g \in B} \|f - (\hat{f}_{t-1} + g)\|_2^2 \quad = \quad \inf_i \|f - (Aw_{t-1} + A\mathbf{e}_i)\|_2^2,$$

$$\inf_{\alpha \in [0,1]} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g_t)\|_2^2.$$

- Suppose $\|g\|_2 = 1$ for all $g \in B$; first step equiv to

$$\sup_{g \in B} \left\langle g, \hat{f}_{t-1} - f \right\rangle = \sup_i (A\mathbf{e}_i)^\top (Aw_{t-1} - f).$$

Greedy approx as coordinate descent

- Intuition: linear operator A with basis B as "columns":

$$\hat{f}_t = Aw_t, \qquad g = A\mathbf{e}_j.$$

- Split the update into two steps:

$$\inf_{g \in B} \|f - (\hat{f}_{t-1} + g)\|_2^2 \quad = \quad \inf_i \|f - (Aw_{t-1} + A\mathbf{e}_i)\|_2^2,$$

$$\inf_{\alpha \in [0,1]} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g_t)\|_2^2.$$

- Suppose $\|g\|_2 = 1$ for all $g \in B$; first step equiv to

$$\sup_{g \in B} \left\langle g, \hat{f}_{t-1} - f \right\rangle = \sup_i (A\mathbf{e}_i)^\top (Aw_{t-1} - f).$$

- Second version is coordinate descent update!

Greedy approx as coordinate descent

- Intuition: linear operator A with basis B as "columns":

$$\hat{f}_t = Aw_t, \qquad g = A\mathbf{e}_j.$$

- Split the update into two steps:

$$\inf_{g \in B} \|f - (\hat{f}_{t-1} + g)\|_2^2 \quad = \quad \inf_i \|f - (Aw_{t-1} + A\mathbf{e}_i)\|_2^2,$$

$$\inf_{\alpha \in [0,1]} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha) g_t)\|_2^2.$$

- Suppose $\|g\|_2 = 1$ for all $g \in B$; first step equiv to

$$\sup_{g \in B} \left\langle g, \hat{f}_{t-1} - f \right\rangle = \sup_i (A\mathbf{e}_i)^\top (Aw_{t-1} - f).$$

- Second version is coordinate descent update!
- (Standard rate proofs use curvature of $\|\cdot\|_2^2$.)

Story for 2-layer NNs so far

- Can L^2 apx any $f \in \mathcal{C}([0,1]^n)$ at rate $\mathcal{O}(1/\sqrt{t})$..

Story for 2-layer NNs so far

- Can L^2 apx any $f \in \mathcal{C}([0,1]^n)$ at rate $\mathcal{O}(1/\sqrt{t})$..
- ..so why aren't we using this algorithm for neural nets?

Story for 2-layer NNs so far

- Can L^2 apx any $f \in \mathcal{C}([0,1]^n)$ at rate $\mathcal{O}(1/\sqrt{t})$..
- ..so why aren't we using this algorithm for neural nets?
  - This is not an easy optimization problem:

$$\inf_{\alpha \in [0,1], w \in \mathbb{R}^n, w_0 \in \mathbb{R}} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)\sigma(\langle w, \cdot \rangle + w_0)\|_2^2.$$

Story for 2-layer NNs so far

- Can L^2 apx any $f \in \mathcal{C}([0,1]^n)$ at rate $\mathcal{O}(1/\sqrt{t})$..
- ..so why aren't we using this algorithm for neural nets?
  - This is not an easy optimization problem:

$$\inf_{\alpha \in [0,1], w \in \mathbb{R}^n, w_0 \in \mathbb{R}} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)\sigma(\langle w, \cdot \rangle + w_0)\|_2^2.$$

  - For a general basis B, must check each basis element..

Story for 2-layer NNs so far

- Can L^2 apx any $f \in \mathcal{C}([0,1]^n)$ at rate $\mathcal{O}(1/\sqrt{t})$..
- ..so why aren't we using this algorithm for neural nets?
  - This is not an easy optimization problem:

$$\inf_{\alpha \in [0,1], w \in \mathbb{R}^n, w_0 \in \mathbb{R}} \|f - (\alpha \hat{f}_{t-1} + (1-\alpha)\sigma(\langle w, \cdot \rangle + w_0)\|_2^2.$$

  - For a general basis B, must check each basis element..
- From an algorithmic perspective, much remains to be done.

Lower bound from Barron (1993, Section 10)

- The greedy algorithm gets to search whole basis.

Lower bound from Barron (1993, Section 10)

- The greedy algorithm gets to search whole basis.
- If an n-element basis B_n is fixed, then

$$\sup_{f \in \mathrm{cl}(\mathrm{span}_C(B))} \inf_{\hat{f} \in \mathrm{span}(B_n)} \|f - \hat{f}\|_2 = \Omega\left(\frac{1}{n^{1/d}}\right),$$

  the curse of dimension!

Lower bound from Barron (1993, Section 10)

- The greedy algorithm gets to search whole basis.
- If an n-element basis B_n is fixed, then

$$\sup_{f \in \mathrm{cl}(\mathrm{span}_C(B))} \inf_{\hat{f} \in \mathrm{span}(B_n)} \|f - \hat{f}\|_2 = \Omega\left(\frac{1}{n^{1/d}}\right),$$

  the curse of dimension!
- Can defeat this with more layers (cf. Kolmogorov (1957)).

Outline

- 2-nn via functional analysis (Cybenko, 1989).
- 2-nn via greedy approx (Barron, 1993).
- 3-nn via histograms (Folklore).
- 3-nn via wizardry (Kolmogorov, 1957).

Folklore proof of NN power

- What's an easy way to write function as sum of others?

Folklore proof of NN power

- What's an easy way to write function as sum of others?
  - Project onto favorite basis (Fourier, Chebyshev, ...)!

Folklore proof of NN power

- What's an easy way to write function as sum of others?
  - Project onto favorite basis (Fourier, Chebyshev, ...)!
- Let's use this to build NNs.

Folklore proof of NN power

- What's an easy way to write function as sum of others?
  - Project onto favorite basis (Fourier, Chebyshev, ...)!
- Let's use this to build NNs.
  - Easiest basis: histograms!

Histogram basis for $\mathcal{C}([0,1]^n)$

- Grid $[0,1]^n$ into $\{R_i\}_{i=1}^m$; consider

$$\hat{f}(x) = \sum_{i=1}^{m} a_i \mathbb{1}[x \in R_i],$$

where target $f(x) = a_i$ for some $x \in R_i$.

Histogram basis for $\mathcal{C}([0,1]^n)$

- Grid $[0,1]^n$ into $\{R_i\}_{i=1}^m$; consider

$$\hat{f}(x) = \sum_{i=1}^{m} a_i \mathbb{1}[x \in R_i],$$

where target $f(x) = a_i$ for some $x \in R_i$.

- Fact: for any $\epsilon > 0$, exists histogram $\hat{f}$ with

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

Histogram basis for $\mathcal{C}([0,1]^n)$

- Grid $[0,1]^n$ into $\{R_i\}_{i=1}^m$; consider

$$\hat{f}(x) = \sum_{i=1}^{m} a_i \mathbb{1}[x \in R_i],$$

  where target $f(x) = a_i$ for some $x \in R_i$.
- Fact: for any $\epsilon > 0$, exists histogram $\hat{f}$ with

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

  - Pf. Continuous function over compact set is uniformly continuous.

Histogram basis for $\mathcal{C}([0,1]^n)$

- Grid $[0,1]^n$ into $\{R_i\}_{i=1}^m$; consider

  $$\hat{f}(x) = \sum_{i=1}^{m} a_i \mathbb{1}[x \in R_i],$$

  where target $f(x) = a_i$ for some $x \in R_i$.
- Fact: for any $\epsilon > 0$, exists histogram $\hat{f}$ with

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

  - Pf. Continuous function over compact set is uniformly continuous.
- Now just write individual histogram bars as a NNs.

Histograms via 0/1 NNs

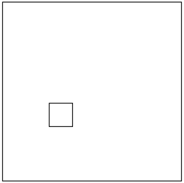
- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x \rangle \geq w_0]$.

Histograms via 0/1 NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x \rangle \geq w_0]$.
- 2-D example. First carve out axes.

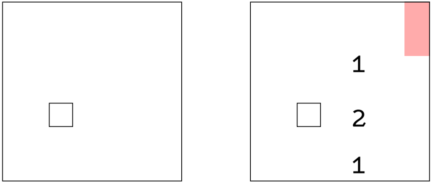
Histograms via 0/1 NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x\rangle \geq w_0]$.
- 2-D example. First carve out axes.



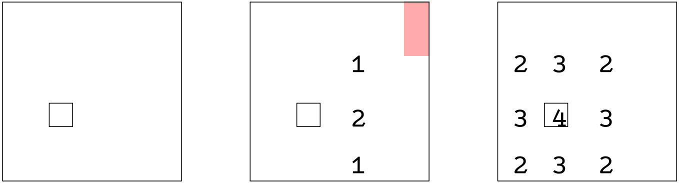
Histograms via 0/1 NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x \rangle \geq w_0]$.
- 2-D example. First carve out axes.



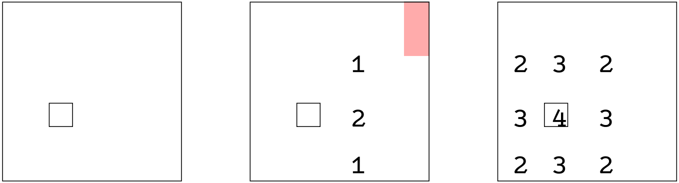
Histograms via 0/1 NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x\rangle \geq w_0]$.
- 2-D example. First carve out axes.

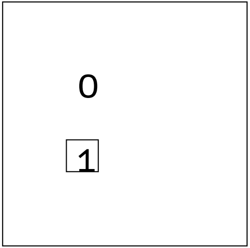


Histograms via 0/1 NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\mathbb{1}[\langle w, x\rangle \geq w_0]$.
- 2-D example. First carve out axes.



- Now pass through a second layer $\mathbb{1}[\text{input} \geq 3.5]$.



Histograms via sigmoidal NNs

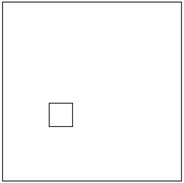
- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x\rangle + w_0)$.

Histograms via sigmoidal NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x\rangle + w_0)$.
- 2-D example. First carve out axes, with fuzz region.

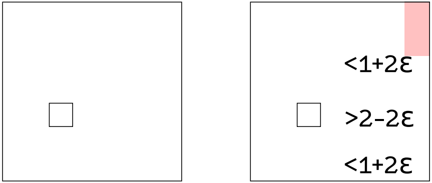
Histograms via sigmoidal NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x\rangle + w_0)$.
- 2-D example. First carve out axes, with fuzz region.



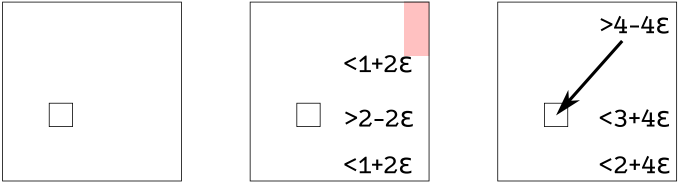
Histograms via sigmoidal NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x \rangle + w_0)$.
- 2-D example. First carve out axes, with fuzz region.



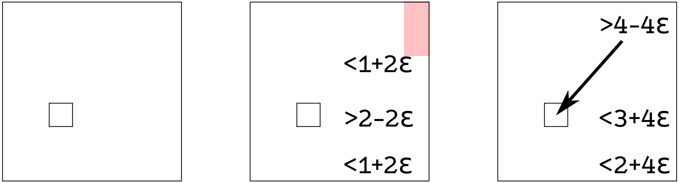
Histograms via sigmoidal NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x\rangle + w_0)$.
- 2-D example. First carve out axes, with fuzz region.

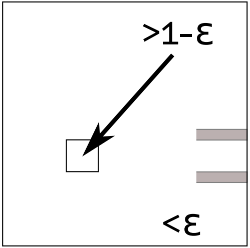


Histograms via sigmoidal NNs

- Write $\mathbb{1}[x \in R]$ as sum/composition of $\sigma(\langle w, x\rangle + w_0)$.
- 2-D example. First carve out axes, with fuzz region.



- Now pass through a second layer $\sigma(\text{huge} \cdot (\text{input} - 3.5))$.



Histograms apx via sigmoidal NNs, some math

- Histogram region is clearly 2-layer 0/1 NN:

$$
\begin{aligned}
&\mathbb{1}[\forall i \,.\, x_i \geq l_i \wedge x_i \leq u_i] \\
&= \mathbb{1}\left[\sum_{i=1}^{n} (\mathbb{1}[x_i \geq l_i] + \mathbb{1}[x_i \leq u_i]) \geq (2n-1)/2n\right].
\end{aligned}
$$

Histograms apx via sigmoidal NNs, some math

- Histogram region is clearly 2-layer 0/1 NN:

$$\begin{aligned}&\mathbb{1}[\forall i \,\textbf{.}\, x_i \geq l_i \wedge x_i \leq u_i] \\ &= \mathbb{1}\left[\sum_{i=1}^{n}(\mathbb{1}[x_i \geq l_i] + \mathbb{1}[x_i \leq u_i]) \geq (2n-1)/2n\right].\end{aligned}$$

- Choose B huge; apx $\mathbb{1}[\langle w, x\rangle \geq w_0]$ with $\sigma(B(\langle w, x\rangle - w_0))$.

Histograms apx via sigmoidal NNs, some math

- Histogram region is clearly 2-layer 0/1 NN:

$$
\begin{aligned}
&\mathbb{1}[\forall i \centerdot x_i \geq l_i \wedge x_i \leq u_i] \\
&= \mathbb{1}\left[\sum_{i=1}^{n}(\mathbb{1}[x_i \geq l_i] + \mathbb{1}[x_i \leq u_i]) \geq (2n-1)/2n\right].
\end{aligned}
$$

- Choose B huge; apx $\mathbb{1}[\langle w, x\rangle \geq w_0]$ with $\sigma(B(\langle w, x\rangle - w_0))$.
- $\epsilon/2$-apx f with histogram $\{a_i, R_i\}_{i=1}^{m}$;
  $\epsilon/(2m)$-apx each bar with sigmoids.

Histograms apx via sigmoidal NNs, some math

- Histogram region is clearly 2-layer 0/1 NN:

$$
\begin{aligned}
&\mathbb{1}[\forall i \,.\, x_i \geq l_i \wedge x_i \leq u_i] \\
&= \mathbb{1}\left[\sum_{i=1}^{n} (\mathbb{1}[x_i \geq l_i] + \mathbb{1}[x_i \leq u_i]) \geq (2n-1)/2n\right].
\end{aligned}
$$

- Choose B huge; apx $\mathbb{1}[\langle w, x\rangle \geq w_0]$ with $\sigma(B(\langle w, x\rangle - w_0))$.
- $\epsilon/2$-apx f with histogram $\{a_i, R_i\}_{i=1}^{m}$;
  $\epsilon/(2m)$-apx each bar with sigmoids.
- Final NN $\hat{f}(x) := \sum_{i=1}^{m} a_i(\text{sigmoidal apx to } \mathbb{1}[x \in R_i])$.

Histograms apx via sigmoidal NNs, some math

- Histogram region is clearly 2-layer 0/1 NN:

$$
\begin{aligned}
&\mathbb{1}[\forall i \centerdot x_i \geq l_i \wedge x_i \leq u_i] \\
&= \mathbb{1}\left[\sum_{i=1}^{n}(\mathbb{1}[x_i \geq l_i] + \mathbb{1}[x_i \leq u_i]) \geq (2n-1)/2n\right].
\end{aligned}
$$

- Choose B huge; apx $\mathbb{1}[\langle w, x\rangle \geq w_0]$ with $\sigma(B(\langle w, x\rangle - w_0))$.
- $\epsilon/2$-apx f with histogram $\{a_i, R_i\}_{i=1}^m$;
  $\epsilon/(2m)$-apx each bar with sigmoids.
- Final NN $\hat{f}(x) := \sum_{i=1}^m a_i(\text{sigmoidal apx to } \mathbb{1}[x \in R_i])$.
- Have fuzz $S \subseteq [0,1]^n$ with $m(S) \geq 1 - \epsilon$,

$$
x \in S \quad \Longrightarrow \quad |f(x) - \hat{f}(x)| < \epsilon.
$$

Folklore proof discussion

- Curse of dimension!
- Still, high level features useful.

Outline

- 2-nn via functional analysis (Cybenko, 1989).
- 2-nn via greedy approx (Barron, 1993).
- 3-nn via histograms (Folklore).
- 3-nn via wizardry (Kolmogorov, 1957).

Preface to the Kolmogorov Superposition Theorem

- In 1957, at 19, Kolmogorov's student Vladimir Arnold solved Hilbert's 13^{th} problem:

Preface to the Kolmogorov Superposition Theorem

- In 1957, at 19, Kolmogorov's student Vladimir Arnold solved Hilbert's 13^{th} problem:

  Can the solution to

  $$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

  be written as a finite number of 2-variable functions?

Preface to the Kolmogorov Superposition Theorem

- In 1957, at 19, Kolmogorov's student Vladimir Arnold solved Hilbert's 13^{th} problem:

  Can the solution to

  $$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

  be written as a finite number of 2-variable functions?

- Kolmogorov's generalization is a NN apx theorem.

Preface to the Kolmogorov Superposition Theorem

- In 1957, at 19, Kolmogorov's student Vladimir Arnold solved Hilbert's 13^{th} problem:

  Can the solution to

  $$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

  be written as a finite number of 2-variable functions?

- Kolmogorov's generalization is a NN apx theorem.
- The "transfer functions" (i.e., sigmoids), are not fixed across nodes.

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

for

any $f \in \mathcal{C}([0,1]^n)$,

$$f(x) = \sum_{q=1} \chi_q \left(\sum_{p=1} \psi^{p,q}(x_p) \right)$$

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}$,

$$f(x) = \sum_{q=1} \chi_q \left(\sum_{p=1} \psi^{p,q}(x_p) \right)$$

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

for

any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}$,

$$f(x) = \sum_{q=1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

for

any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

There exist continuous $\{\psi^{p,q} : p \in [n], q \in [2n+1]\}$, so that for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

There exist continuous $\{\psi^{p,q} : p \in [n], q \in [2n+1]\}$, so that for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- The proof is *also* histogram based.

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

There exist continuous $\{\psi^{p,q} : p \in [n], q \in [2n+1]\}$, so that for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- The proof is *also* histogram based.
  - It must remove the "fuzz" gaps.

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

There exist continuous $\{\psi^{p,q} : p \in [n], q \in [2n+1]\}$, so that for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- The proof is *also* histogram based.
  - It must remove the “fuzz” gaps.
  - It must remove dependence on a particular $\epsilon > 0$.

The Kolmogorov Superposition Theorem

Theorem. (Kolmogorov, 1957)

There exist continuous $\{\psi^{p,q} : p \in [n], q \in [2n+1]\}$, so that for any $f \in \mathcal{C}([0,1]^n)$, there exist continuous $\{\chi_q\}_{q=1}^{2n+1}$,

$$f(x) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

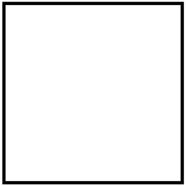
- The proof is *also* histogram based.
  - It must remove the "fuzz" gaps.
  - It must remove dependence on a particular $\epsilon > 0$.
- The magic is within $\psi^{p,q}$!!

Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.

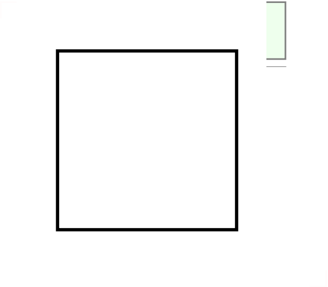
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



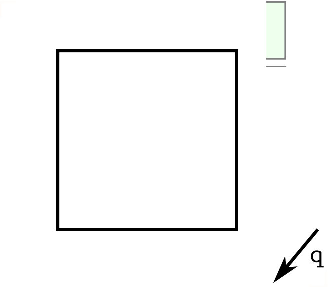
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



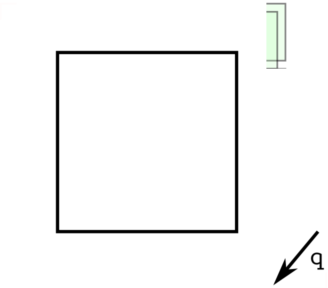
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



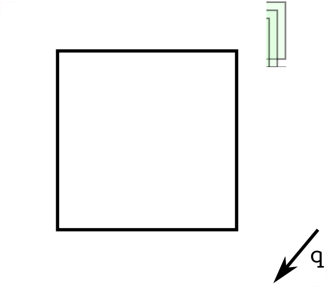
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



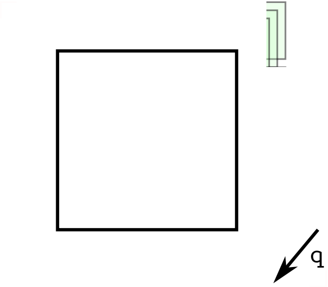
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



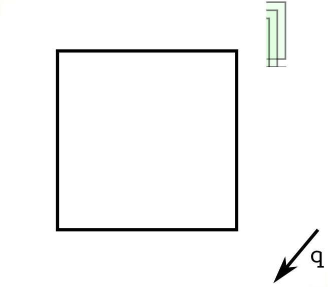
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



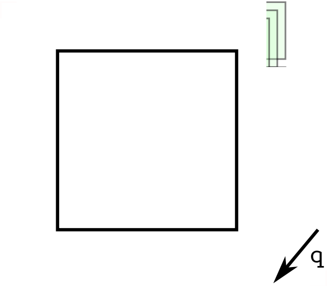
Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



Constructing $\psi^{p,q}$

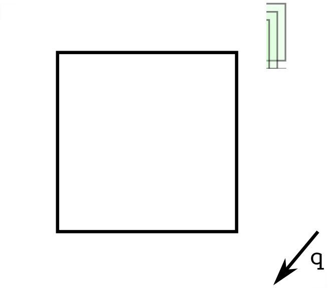
- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



- Let $S^q_{k,i_1,\ldots,i_m}$ range over the cells.

Constructing $\psi^{p,q}$

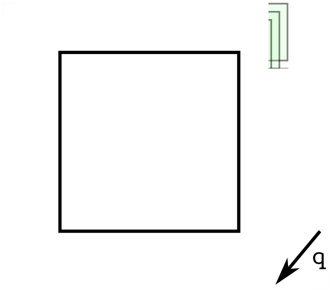
- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



- Let $S^q_{k,i_1,\ldots,i_m}$ range over the cells.
- Construct $\psi^{p,q}$ so that for each k and q, $\sum_{p=1}^{n} \psi^{p,q}$ maps each $S^q_{k,i_1,\ldots,i_m}$ to its own specific interval of $\mathbb{R}$.

Constructing $\psi^{p,q}$

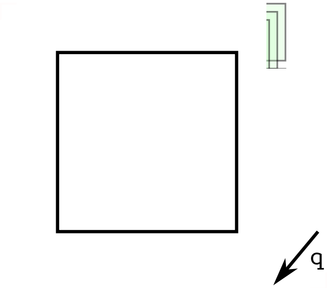
- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



- Let $S^q_{k,i_1,\ldots,i_m}$ range over the cells.
- Construct $\psi^{p,q}$ so that for each k and q, $\sum_{p=1}^{n} \psi^{p,q}$ maps each $S^q_{k,i_1,\ldots,i_m}$ to its own specific interval of $\mathbb{R}$.
- Holds for all q: *gracefully handles the fuzz regions*.

Constructing $\psi^{p,q}$

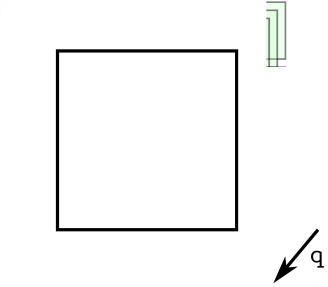
- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



- Let $S^q_{k,i_1,\ldots,i_m}$ range over the cells.
- Construct $\psi^{p,q}$ so that for each k and q, $\sum_{p=1}^{n} \psi^{p,q}$ maps each $S^q_{k,i_1,\ldots,i_m}$ to its own specific interval of $\mathbb{R}$.
- Holds for all q: *gracefully handles the fuzz regions.*
- Holds for all k, *simultaneously handles all resolutions.*

Constructing $\psi^{p,q}$

- Given resolution $k \in \mathbb{Z}_{++}$, build a staggered gridding.



- Let $S^q_{k,i_1,\ldots,i_m}$ range over the cells.
- Construct $\psi^{p,q}$ so that for each k and q, $\sum_{p=1}^n \psi^{p,q}$ maps each $S^q_{k,i_1,\ldots,i_m}$ to its own specific interval of $\mathbb{R}$.
- Holds for all q: *gracefully handles the fuzz regions.*
- Holds for all k, *simultaneously handles all resolutions.*
- *The functions $\psi^{p,q}$ are fractals.*

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r \to \infty} \chi_r^q$.

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r \to \infty} \chi_r^q$.
- To build χ_{r+1}^q from χ_r^q:

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r\to\infty} \chi_r^q$.
- To build χ_{r+1}^q from χ_r^q:
  - Choose k_{r+1} so gridding fluctuations sufficiently small.

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r\to\infty} \chi_r^q$.
- To build χ_{r+1}^q from χ_r^q:
  - Choose k_{r+1} so gridding fluctuations sufficiently small.
  - Make χ_q some value of f in each cell, smooth interpolation in fuzz.

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r\to\infty} \chi_r^q$.
- To build χ_{r+1}^q from χ_r^q:
  - Choose k_{r+1} so gridding fluctuations sufficiently small.
  - Make χ_q some value of f in each cell, smooth interpolation in fuzz.
  - Every $x \in [0,1]^n$ hits at least $n+1$ cells $S^q_{k,i_1,\ldots,i_n}$; i.e., since $q \in [2n+1]$, more than half the approximations are good.

Constructing χ^q

- Recall goal:

$$f(x) = \sum_{q=1}^{2n+1} \chi^q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right)$$

- Start $\chi_0^q := 0$; eventually $\chi^q := \lim_{r\to\infty} \chi_r^q$.
- To build χ_{r+1}^q from χ_r^q:
  - Choose k_{r+1} so gridding fluctuations sufficiently small.
  - Make χ_q some value of f in each cell, smooth interpolation in fuzz.
  - Every $x \in [0,1]^n$ hits at least $n+1$ cells $S^q_{k,i_1,\ldots,i_n}$; i.e., since $q \in [2n+1]$, more than half the approximations are good.
- (I'm leaving a lot out =( )

Discussion 1

NOTE — **Communicated by Halbert White**

Representation Properties of Networks: Kolmogorov's Theorem Is Irrelevant

Federico Girosi
Tomaso Poggio
Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA 02142 USA
and
Center for Biological Information Processing, Whitaker College, Cambridge, MA 02142 USA

Many neural networks can be regarded as attempting to approximate a multivariate function in terms of one-input one-output units. This note considers the problem of an exact representation of nonlinear mappings in terms of simpler functions of fewer variables. We review Kolmogorov's theorem on the representation of functions of several variables in terms of functions of one variable and show that it is irrelevant in the context of networks for learning.

1 Kolmogorov's Theorem: An Exact Representation Is Hopeless

A crucial point in approximation theory is the choice of the representation

Discussion 2

- It needs different transfer functions..

Discussion 2

- It needs different transfer functions..
- ..but these can be approximated by multiple layers!

Discussion 2

- It needs different transfer functions..
- ..but these can be approximated by multiple layers!
- It shows the value of powerful nodes, whereas the standard apx results just suggest a very wide hidden layer.

Conclusion

- Some fancy mechanisms give 2-NNs $\hat{f}_i \to f \in \mathcal{C}([0,1]^n)$.
- Histogram constructions hint to the power of deeper networks.

Thanks!

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

Andrei N. Kolmogorov. On the representation of continuous functions of several veriables as superpositions of continuous functions of one variable and addition. *Dokl. Acad. Nauk SSSR*, 114(5):953–956, 1957. Translation to English: V. M. Volosov.

Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.