Heaviside Networks Approximation Theorem

Choeun Kim
IDEA

June 4, 2026

Table of Contents

Section 1

Notations

Notations: Functional Norms & Spaces

Basic Notation & Multi-index

Let $m = (m_1, \ldots, m_d) \in \mathbb{N}_0^d$ be a multi-index, where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $|m|_1 := \sum_{j=1}^d |m_j|$.
For a function $f : [0,1]^d \to \mathbb{R}$:

- **Sup-norm:** $\|f\|_\infty := \sup_{x \in [0,1]^d} |f(x)|$
- **L_p-norm:** $\|f\|_{p,\mu} := \left(\int_X |f(x)|^p \mu(dx)\right)^{1/p}$ for $1 \leq p < \infty$.
- **Mixed partial derivative of order m:** $\partial^m f := \frac{\partial^{|m|_1} f}{\partial x_1^{m_1} \ldots \partial x_d^{m_d}}$

Definition: The Hölder Space $\mathcal{H}_d^\beta(R)$

For a smoothness parameter $\beta > 0$, let $\lfloor \beta \rfloor$ be the greatest integer strictly less than β. The Hölder space $\mathcal{H}_d^\beta(R)$ is defined as $\{f \in \mathcal{C}_d^{\lfloor \beta \rfloor} : \|f\|_{\mathcal{H}_d^\beta} \leq R\}$, where:

$$\|f\|_{\mathcal{H}_d^\beta} := \sum_{m \in \mathbb{N}_0^d : |m|_1 \leq \lfloor \beta \rfloor} \|\partial^m f\|_\infty + \sum_{m \in \mathbb{N}_0^d : |m|_1 = \lfloor \beta \rfloor} \sup_{x_1, x_2 \in [0,1]^d, x_1 \neq x_2} \frac{|\partial^m f(x_1) - \partial^m f(x_2)|}{|x_1 - x_2|_\infty^{\beta - \lfloor \beta \rfloor}}$$

Notations: Network Architectures

Deep Neural Networks with HTAF

Let $HTAF_{\alpha_0,\alpha_1}(x) := \sigma(\alpha_1\psi(\alpha_0 x))$. A DNN with HTAF, having L layers, width p_l at the l-th layer, input dimension $p_0 = d$, and output dimension $p_{L+1} = 1$, is parameterized as:

$$f^{HTAF}_{\theta,\alpha_0,\alpha_1}(x) := A_{L+1} \circ HTAF_{\alpha_0,\alpha_1} \circ A_L \circ \cdots \circ HTAF_{\alpha_0,\alpha_1} \circ A_1(x)$$

where $A_l(x) = W_l x + b_l$ is an affine map with weight matrix $W_l \in \mathbb{R}^{p_l \times p_{l-1}}$ and bias vector $b_l \in \mathbb{R}^{p_l}$. The entire parameter set is denoted by $\theta := (W_1, b_1, \ldots, W_{L+1}, b_{L+1})$.

Deep Heaviside Networks (DHNs)

Furthermore, if we replace the activation function $HTAF_{\alpha_0,\alpha_1}$ with the Heaviside step function $\sigma_H(x) = \mathbb{I}(x > 0)$, we define the corresponding Deep Heaviside Network as $f^{DHN}_\theta(x)$ with the same parameter set θ.

Notations: Function Classes

HTAF Function Classes: $\mathcal{F}_{\alpha_0,\alpha_1}(L,N)$ and $\mathcal{F}_{\alpha_0,\alpha_1}(L,N,S,B)$

For a given parameter θ, let $L(\theta)$ be the number of hidden layers (depth), and let $p_{\max}(\theta) := \max_{1\le l\le L} p_l$ be the maximum number of hidden units per layer (width).

Let $|\theta|_0 = \sum_{l=1}^{L+1}(|vec(W_l)|_0 + |b_l|_0)$ and $|\theta|_\infty = \max\{\max_l |vec(W_l)|_\infty, \max_l |b_l|_\infty\}$. We define:

- $\mathcal{F}_{\alpha_0,\alpha_1}(L,N) := \{f^{HTAF}_{\theta,\alpha_0,\alpha_1} \mid L(\theta) \le L,\ p_{\max}(\theta) \le N\}$
- $\mathcal{F}_{\alpha_0,\alpha_1}(L,N,S,B) := \{f^{HTAF}_{\theta,\alpha_0,\alpha_1} \in \mathcal{F}_{\alpha_0,\alpha_1}(L,N) \mid |\theta|_0 \le S,\ |\theta|_\infty \le B\}$

DHN Function Class: $\mathcal{F}_{DHN}(L,\mathbf{p})$

Similarly, let $\mathbf{p} = (d, p_1, p_2, \ldots, p_L, p_{L+1})$ be a hidden width vector. We define $\mathcal{F}_{DHN}(L,\mathbf{p})$ as the class of DHNs (f^{DHN}_θ) consisting of L hidden layers, where the width of the l-th layer is strictly specified by p_l.

Section 2

Expressivity Limits of DHNs

Expressivity Limits of DHNs

Theorem 1

Let $f_0 : [0,1]^d \to \mathbb{R}$ be any target function. For a Deep Heaviside Network $f : [0,1]^d \to \mathbb{R}$ belonging to the function class $\mathcal{F}_{DHN}(L, \mathbf{p})$, the worst-case approximation error is bounded below by:

$$\inf_{f \in \mathcal{F}_{DHN}(L,\mathbf{p})} \|f - f_0\|_\infty \geq \frac{\sup f_0 - \inf f_0}{2(p_1 + 1)} \tag{1}$$

where p_1 is the width of the first hidden layer. Achieving an error $\leq \epsilon$ requires $p_1 \geq \mathcal{O}(1/\epsilon)$, **regardless of the total depth** L.

Implication: In a Plain DHN, **depth is entirely useless** for improving the worst-case approximation error if the first layer is narrow. Unlike ReLU networks where depth brings exponential expressivity, Plain DHNs suffer from a severe structural bottleneck dictated solely by p_1.

Proof Sketch

Why does depth fail? (The 1-Bit Bottleneck)

1. **Initial Partitioning:** The first hidden layer consists of p_1 Heaviside neurons. In a 1D input space, these p_1 hyperplanes partition the input line into at most $p_1 + 1$ distinct intervals.

Proof Sketch

Why does depth fail? (The 1-Bit Bottleneck)

1. **Initial Partitioning:** The first hidden layer consists of p_1 Heaviside neurons. In a 1D input space, these p_1 hyperplanes partition the input line into at most $p_1 + 1$ distinct intervals.
2. **Information Collapse:** Because the Heaviside activation outputs strictly $\{0, 1\}$, the continuous input x is collapsed into a single discrete vector $h_1 \in \{0, 1\}^{p_1}$. For any x within the same interval, the output h_1 is identical.

Proof Sketch

Why does depth fail? (The 1-Bit Bottleneck)

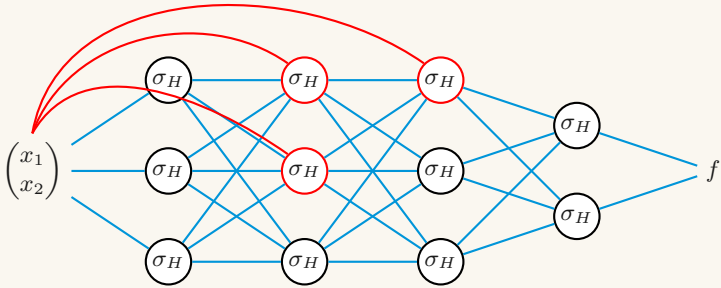
1. **Initial Partitioning:** The first hidden layer consists of p_1 Heaviside neurons. In a 1D input space, these p_1 hyperplanes partition the input line into at most $p_1 + 1$ distinct intervals.
2. **Information Collapse:** Because the Heaviside activation outputs strictly $\{0, 1\}$, the continuous input x is collapsed into a single discrete vector $h_1 \in \{0, 1\}^{p_1}$. For any x within the same interval, the output h_1 is identical.
3. **Propagation of Constants:** The second layer receives this discrete vector and applies an affine transformation followed by another Heaviside step. Since the input to the second layer is constant for each of the $p_1 + 1$ intervals, the output of the second layer must also be constant on those intervals.

Proof Sketch

Why does depth fail? (The 1-Bit Bottleneck)

1. **Initial Partitioning:** The first hidden layer consists of p_1 Heaviside neurons. In a 1D input space, these p_1 hyperplanes partition the input line into at most $p_1 + 1$ distinct intervals.
2. **Information Collapse:** Because the Heaviside activation outputs strictly $\{0, 1\}$, the continuous input x is collapsed into a single discrete vector $h_1 \in \{0, 1\}^{p_1}$. For any x within the same interval, the output h_1 is identical.
3. **Propagation of Constants:** The second layer receives this discrete vector and applies an affine transformation followed by another Heaviside step. Since the input to the second layer is constant for each of the $p_1 + 1$ intervals, the output of the second layer must also be constant on those intervals.
4. **Conclusion:** By induction, the entire network's output is **piecewise constant with at most $p_1 + 1$ distinct values**. No amount of depth can subdivide these initial intervals.

Skip connection augmented DHN



Example of a skip-connection augmented DHN. The graph represents the network class $\mathcal{F}_{DHN_{skip}}(L, \mathbf{p}, \mathbf{s})$ with $L = 4, \mathbf{p} = (2, 3, 3, 3, 2, 1)$, and $\mathbf{s} = (2, 1, 0)$.

Theorem 3

Let $f_0 : [0,1]^d \to \mathbb{R}$ be any target function. For a skip-connection augmented DHN $f : [0,1]^d \to \mathbb{R}$ belonging to the function class $\mathcal{F}_{DHN_{skip}}(L, \mathbf{p}, \mathbf{s})$, the worst-case approximation error is bounded below by:

$$\inf_{f \in \mathcal{F}_{DHN_{skip}}(L,\mathbf{p},\mathbf{s})} \|f - f_0\|_\infty \geq \frac{\sup f_0 - \inf f_0}{2(p_1+1)\prod_{l=2}^{L}(s_l+1)} \tag{2}$$

Notice that compared to (1), the term $\prod_{l=2}^{L}(s_l+1)$ has been added to the denominator, effectively driving down the lower bound. Consequently, this yields a much stronger approximation performance for the given function class.

Minimax Optimality

- Nonparametric Regression setting : n i.i.d training samples $(\mathbf{X}_i, y_i)$ drawn from the joint distribution of $(\mathbf{X}, Y)$ where

$$\mathbf{X} \sim P_X, \quad Y = f_0(\mathbf{X}) + \epsilon, \tag{3}$$

  $P_{\mathbf{X}}$ denotes the marginal distribution of $\mathbf{X}$ over $\mathcal{X}$, and ϵ an independent noise term following a standard normal distribution.

- The authors proved that skip-DHNs with appropriately chosen architectures achieve the minimax convergence rate over $\mathcal{H}_d^\beta$ up to $\log(n)$-factors.

Section 3

HTAF Theoretical Theorems

HTAF expressivity

Theorem B.1

For any deep ReLU network f^{ReLU} with L hidden layers and N hidden units per layer, and any $\alpha_0, \alpha_1, \epsilon > 0$, there exists a DNN with HTAF satisfying: [1]

$$f_{\theta,\alpha_0,\alpha_1}^{HTAF} \in \mathcal{F}_{\alpha_0,\alpha_1}(2L, 3N)$$

and

$$\|f_{\theta,\alpha_0,\alpha_1}^{HTAF} - f^{ReLU}\|_\infty \leq \epsilon$$

Theoretical Significance: This proves that the expressivity properties of deep ReLU networks carry over directly to DNNs with HTAF, provided the network is larger by only a constant factor. It establishes that **HTAF networks do not suffer from the expression bottleneck of plain DHNs.**

[1] Proof is done by applying the Theorem 1 of Zhang 2024.

Theorem B.2 (Hölder Function Approximation)

There exist positive constants L_0, N_0, S_0 and B_0 depending only on (d, β, R), such that for any $f_0 \in \mathcal{H}_d^\beta(R)$ and $\epsilon > 0$, there exists a DNN with HTAF satisfying:

$$f_{\theta,\alpha_0,\alpha_1}^{HTAF} \in \mathcal{F}_{\alpha_0,\alpha_1}\left(L_0 \log(1/\epsilon),\ N_0\epsilon^{-d/\beta},\ S_0\epsilon^{-d/\beta},\ B_0\epsilon^{-4(d/\beta+1)}\right)$$

which satisfies the uniform approximation bound:

$$\|f_{\theta,\alpha_0,\alpha_1}^{HTAF} - f_0\|_\infty \leq \epsilon$$

Core Insight: Importantly, the scaling parameters α_0 and α_1 affect neither the depth and width required to bound the approximation error nor the explicit scale of the internal network parameters. Refer to Ohn and Kim 2019.

Theorem B.3 (Statistical Convergence Bound)

For any $\kappa, R, \alpha_0, \alpha_1 > 0$, there exist positive constants L_0, N_0, S_0, and B_0 such that the HTAF DNN estimator $\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$ with

$$\mathcal{F}_n := \{ f_{\xi,\alpha_0,\alpha_1}^{HTAF} \in \mathcal{F}(L_0 \log n, N_0 n^{\frac{d}{2\beta+d}}, S_0 n^{\frac{d}{2\beta+d}} \log n, B_0 n^{\kappa}) : \| f_{\xi,\alpha_0,\alpha_1}^{HTAF} \|_\infty \leq 2R \}$$

satisfies

$$\sup_{f_0 \in \mathcal{H}_d^\beta(R)} \mathbb{E}_n \left[\| \hat{f}_n - f_0 \|_{2,P_X}^2 \right] \leq C n^{-\frac{2\beta}{2\beta+d}} \left(\log^3 n + \log^2 n \log(\alpha_0 \alpha_1) \right)$$

for some constant $C > 0$.

Conclusion on Statistical Optimality: DNN estimators with HTAF achieve the near-minimax optimal convergence rate over the Hölder class under mild assumptions on α_0 and α_1. The explicit parameter dependence on α_0 and α_1 appears minimally through a logarithmic factor.

Section 4

References

References


Ohn, Ilsang and Yongdai Kim (2019). “Smooth function approximation by deep neural networks with general activation functions”. In: *Entropy* 21.7, p. 627.


Zhang Lu, Zhao (2024). “Deep network approximation: Beyond relu to diverse activation functions”. In: *Journal of Machine Learning Research* 25.35, pp. 1–39.