Approximation Error Estimates by Noise-injected Neural Networks

Keito AKIYAMA[1]

[1]Tohoku Daigaku - Aobayama Campus

August 8, 2023

Abstract

One-hidden-layer feedforward neural networks are described as functions having many real-valued parameters. The larger the number of parameters is, neural networks can approximate various functions (universal approximation property). The essential optimal order of approximation bounds is already derived in 1996. We focused on the numerical experiment that indicates the neural networks whose parameters have stochastic perturbations gain better performance than ordinary neural networks, and explored the approimation property of neural networks with stochastic perturbations. In this paper, we derived the quantitative order of variance of stochastic perturbations to achieve the essential approximation order.

Approximation Error Estimates by Noise-injected Neural Networks

Keito AKIYAMA [*] [†] [‡]

Abstract

One-hidden-layer feedforward neural networks are described as functions having many real-valued parameters. The larger the number of parameters is, neural networks can approximate various functions (universal approximation property). The essential optimal order of approximation bounds is already derived in 1996. We focused on the numerical experiment that indicates the neural networks whose parameters have stochastic perturbations gain better performance than ordinary neural networks, and explored the approimation property of neural networks with stochastic perturbations. In this paper, we derived the quantitative order of variance of stochastic perturbations to achieve the essential approximation order.

Keywords: feedforward neural networks, approximation of functions, universal approximation property, approximation order, stochastic perturbations.

Mathematics Subject Classification: 41A30, 62L20, 68T07.

Acknowledgements

This work was partially supported by Grant-in-Aid for Challenging Research (Exploratory), 21K18582, from Japan Society for the Promotion of Science.

ORCID ID

Keito AKIYAMA (`https://orcid.org/0009-0005-3800-8600`)

[*]Mathematical Institute, Tohoku University.

[†]6-3, Aramaki Aza-Aoba, Aoba-ku, Sendai 980-8578, Japan.

[‡]keito.akiyama.p8@dc.tohoku.ac.jp

1 Introduction and Results

Neural networks are mathematical models inspired by the neurons in a biological brain [1], which include many computational units. Each unit has input weights, output weights, biases, and an activation function. Weights are the linear coefficients in units, and an activation function is nonlinear. One-hidden-layer feedforward neural networks, the simplest models, can be described as input-output functions of the form as follows :

$$f_N(x) = \sum_{k_2=1}^{N} w^{(3)}_{1,k_2} \psi \left(\sum_{k_1=1}^{d} w^{(2)}_{k_2,k_1} x_{k_1} - \theta^{(2)}_{k_2} \right), \quad x = (x_{k_1})_{k_1=1}^{d} \in \mathbb{R}^d, \qquad (1)$$

where $N \in \mathbb{N}$ is the number of hidden units, $w^{(2)}_{k_2,k_1} \in \mathbb{R}$ are input weights, $w^{(3)}_{1,k_2} \in \mathbb{R}$, are output weights, and $\theta^{(2)}_{k_2} \in \mathbb{R}$ are biases, for $k_1 = 1, \cdots, d$, $k_2 = 1, \cdots, N$, $\psi : \mathbb{R} \to \mathbb{R}$ indicates an activation function. The most common activation functions are sigmoidals, that is, $\psi : \mathbb{R} \to \mathbb{R}$ is measurable, bounded, and satisfies $\lim_{z\to\infty} \psi(z) = 1$ and $\lim_{z\to-\infty} \psi(z) = 0$.

If N is sufficiently large, it is possible to adjust parameters so that f_N approximates various multivariable functions, referred to as the universal approximation property. Various proofs of the property exist. For example, Cybenko [2, Theorem 1] showed by using the Hahn–Banach theorem that the space of neural networks (1) is dense in the space of continuous functions. Hornik–Stinchcombe–White [3, Theorem 2.4] showed the existence of neural networks approximating a continuous function by the Stone–Wierstrass theorem, while Funahashi [4, Theorem 1] applied the Fourier transform and the piecewise quadrature method.

Maurey(Pisier [5]), Jones [6], and Barron [7, Theorem 1] derived approximation bounds for neural networks, showing that for every H^1 function, a neural network archives approximation error of order $O(N^{-\frac{1}{2}})$. This order is called Maurey–Jones–Barron(MJB) estimation, which does not depend on the input dimension d. Makovoz [8, Theorem 4] proposed that the lower bound of approximation error of order is $N^{-\frac{1}{2}-\frac{1}{d}}$, depending on d. The index $-\frac{1}{2} - \frac{1}{d}$ converges to $-\frac{1}{2}$ as $d \to \infty$, that is, MJB estimation $O(N^{-\frac{1}{2}})$ cannot be essentially improved.

In recent years, glial cells have been revealed in the human brain; these cells are affected to neurons with signal transduction. Ikuta–Uwate–Nishio [9] numerically experimanted with neural networks whose parameters have

stochastic perturbations (as glial cells affecting neurons) gain better performance than ordinary neural networks. Various neural networks contain stochastic perturbations, and Bayesian neural networks(BNNs) are one [10]. Fonng-Burt-Li-Turner [11, Theorem 6] proved the universal approximation property for BNNs whose parameters are distributed by a statistical method, provided the variance is small enough.

This paper derived the quantitative order of variance to achieve an essential approximation order $O(N^{-\frac{1}{2}})$. Let $\widetilde{f}_N$ be

$$\widetilde{f}_N(x) := \sum_{k_2=1}^{N} \widetilde{w}_{1,k_2}^{(3)} \psi \left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)} \right) - \widetilde{\theta}_1^{(3)}, \quad x = (x_{k_1})_{k_1=1}^{d} \in \mathbb{R}^d,$$

where $\widetilde{w}_{k_2,k_1}^{(2)}$, $\widetilde{w}_{1,k_2}^{(3)}$, $\widetilde{\theta}_{k_2}^{(2)}$, $\widetilde{\theta}_1^{(3)}$ are random variables with variance σ^2, and possible values are finite. The following theorem is this paper's main result.

Theorem 1. Let $K_0 \subset \mathbb{R}^d$ be an open bounded set that contains 0, and activation function $\psi : \mathbb{R} \to \mathbb{R}$ be a ReLU, i.e., $\psi(x) = \max\{x, 0\}$. Then, for all $f \in H^1(K_0)$ and $N \in \mathbb{N}$, $C_0 > 0$ exists, such that for all σ^2 with $0 < \sigma^2 \leq C_0 N^{-4}$, the estimate

$$\|E[\widetilde{f}_N] - f\|_{L^2(K_0)} \leq C N^{-\frac{1}{2}}.$$

holds for some $C > 0$.

Several basic tools which are required for the proof in Section 2. The proof of the main theorem is given in Section 3.

2 Preliminaries

We give two definitions for activation functions.

Definition (sigmoidal function)**.** A measurable and bounded function $\psi : \mathbb{R} \to \mathbb{R}$ is called sigmoidal, if ψ satisfies $\lim_{z\to\infty} \psi(z) = 1$ and $\lim_{z\to-\infty} \psi(z) = 0$.

Definition (ReLU function)**.** A function $\psi : \mathbb{R} \to \mathbb{R}$ defined by $\psi(x) = \max\{x, 0\}$ is called ReLU.

Let K be an open set in $\mathbb{R}^d$. We denote by $L^2(K) = L^2(K, \mu)$ as the space of square-integrable functions on K, where μ is the Lebesgue measure. A function f defined on K is said to be in Sobolev space $H^1(K)$, if $f \in L^2(K)$ and if its distributional gradient, ∇f, is a function that is in $L^2(K)$.

We introduce Barron's approximation rates by deterministic neural networks. Denote by $\mathrm{conv}_N G$ the set of all convex combinations of N elements from the set G.

Theorem 2 (Barron [7, Theorem 1])**.** Let $K_0 \subset \mathbb{R}^d$ be an open bounded set that contains 0. Denote by $G_{\psi,2B}(K_0)$ the set of a sigmoidal function composed with an affine function, i.e.,

$$G_{\psi,2B}(K_0) := \left\{ g : K_0 \to \mathbb{R} : \begin{array}{c} g(x) = w^{(3)}\psi(w^{(2)} \cdot x - \theta^{(2)}), \\ w^{(2)} \in \mathbb{R}^d, w^{(3)}, \theta^{(2)} \in \mathbb{R}, |w^{(3)}| \leq 2B \end{array} \right\}.$$

for a sigmoidal function ψ. Then, for all $f \in H^1(K_0)$ and $N \in \mathbb{N}$,

$$\|f - \mathrm{conv}_N G_{\psi,2B}(K_0)\|_{L^2(K_0)} \leq \sqrt{s^2_{G_{\psi,2B}(K_0)} - \|f\|^2_{L^2(K_0)}} \cdot N^{-\frac{1}{2}}. \tag{2}$$

The element of $\mathrm{conv}_N G_{\psi,2B}(K_0)$ is a linear combination of ψ; therefore, (2) indicates that a neural network exists that approximates f of the order $O(N^{-1/2})$. In case the activation function ψ is ReLU, a similar approximation rate holds since $\psi(\cdot) - \psi(\cdot - 1)$ is sigmoidal.

We denote by $(\Omega, \mathcal{F}, P)$ the probability space; that is, $(\Omega, \mathcal{F})$ is a measurable space and a measure $P : \mathcal{F} \to [0, 1]$ satisfies $P(\emptyset) = 0$ and $P(\Omega) = 1$. For a random variable X on $(\Omega, \mathcal{F}, P)$ and Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$, $E[f(X)]$ indicates the expectation of $f(X)$ concerning P, and $V[f(X)]$ indicates the variance.

We prove the following two lemmas which are used in the proof of the main theorem's proof.

Lemma 3 (Foong-Burt-Li-Turner [11, Lemma 5])**.** Let X be a real-valued random variable, and $\psi : \mathbb{R} \to \mathbb{R}$ be ReLU. Then,

$$V[\psi(X)] \leq V[X].$$

Proof. In the proof, we use the fact that ψ is a contraction mapping. We denote X' by independently and identically distributed random valiables with

X. Then,

$$\begin{aligned}
V[X] &= E[X^2] - E[X]^2 \\
&= \frac{1}{2}(E[X^2] - E[X]^2 + E[X'^2] - E[X']^2) \\
&= \frac{1}{2}(E[X^2] + E[X'^2] - 2E[X]E[X']) \\
&= \frac{1}{2}E[(X - X')^2] \\
&\geq \frac{1}{2}E[(\psi(X) - \psi(X'))^2] \\
&= V[\psi(X)].
\end{aligned}$$

□

Lemma 4. Let $\widetilde{f}_N$ be the function defined by

$$\widetilde{f}_N(x) := \sum_{k_2=1}^{N} \widetilde{w}_{1,k_2}^{(3)} \psi \left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)} \right) - \widetilde{\theta}_1^{(3)}, \quad x = (x_{k_1})_{k_1=1}^{d} \in \mathbb{R}^d,$$

where $\widetilde{w}_{k_2,k_1}^{(2)}$, $\widetilde{w}_{1,k_2}^{(3)}$, $\widetilde{\theta}_{k_2}^{(2)}$, $\widetilde{\theta}_1^{(3)}$ are independently and identically distributed random variables with variance σ^2, and ψ is ReLU. Assume that $L, M > 0$ exists, such that $|\widetilde{w}_{k_2,k_1}^{(2)}|, |\widetilde{w}_{1,k_2}^{(3)}|, |\widetilde{\theta}_{k_2}^{(2)}|, |\widetilde{\theta}_1^{(3)}| \leq L$, and $|x| \leq M$. Then,

$$V\left[\widetilde{f}_N(x)\right] \leq \{L(M^2 d + 1) + (dLM + L)^2 + 1\}\sigma^2 + (M^2 d + 1)\sigma^4.$$

Proof. In the proof, we calculate $V\left[\widetilde{f}_N(x)\right]$ using the independence of the random variable. It follows from the independence of random variables that

$$\begin{aligned}
V\left[\widetilde{f}_N(x)\right] &= V\left[\sum_{k_2=1}^{N} \widetilde{w}_{1,k_2}^{(3)} \psi \left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)} \right) - \widetilde{\theta}_1^{(3)}\right] \\
&= V\left[\widetilde{w}_{1,k_2}^{(3)} \psi \left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)} \right)\right] + V\left[\widetilde{\theta}_1^{(3)}\right].
\end{aligned}$$

Furthermore, since the equality
$V[XY] = E[X]^2 V[Y] + V[X]E[Y]^2 + V[X]V[Y]$ holds for independent random

variables, X and Y, there holds

$$\begin{aligned}
V\left[\widetilde{f}_N(x)\right] &= E\left[\widetilde{w}^{(3)}_{1,k_2}\right]^2 V\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right] \\
&\quad + V\left[\widetilde{w}^{(3)}_{1,k_2}\right] E\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right]^2 \\
&\quad + V\left[\widetilde{w}^{(3)}_{1,k_2}\right] V\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right] + V\left[\widetilde{\theta}^{(3)}_1\right] \\
&= w^{(3)}_{1,k_2} V\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right] \\
&\quad + \sigma^2 E\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right]^2 \\
&\quad + \sigma^2 V\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right] + \sigma^2.
\end{aligned}$$

Using the bounds of possible values of $\widetilde{w}^{(2)}_{k_2,k_1}$ and $\widetilde{\theta}^{(2)}_{k_2}$, we obtain

$$\begin{aligned}
E\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)\right]^2 &\le E\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right)^2\right] \\
&\le E\left[\left|\sum_{k_1=1}^{d}\widetilde{w}^{(2)}_{k_2,k_1}x_{k_1}-\widetilde{\theta}^{(2)}_{k_2}\right|^2\right] \\
&\le E\left[\left(\sum_{k_1=1}^{d}|\widetilde{w}^{(2)}_{k_2,k_1}||x_{k_1}|+|\widetilde{\theta}^{(2)}_{k_2}|\right)^2\right] \\
&\le (dLM+L)^2.
\end{aligned}$$

By Lemma 3,

$$\begin{aligned} V\left[\psi\left(\sum_{k_1=1}^{d}\widetilde{w}_{k_2,k_1}^{(2)}x_{k_1}-\widetilde{\theta}_{k_2}^{(2)}\right)\right] &\leq V\left[\sum_{k_1=1}^{d}\widetilde{w}_{k_2,k_1}^{(2)}x_{k_1}-\widetilde{\theta}_{k_2}^{(2)}\right] \\ &\leq \sum_{k_1=1}^{d}V\left[\widetilde{w}_{k_2,k_1}^{(2)}\right]x_{k_1}^2+V\left[\widetilde{\theta}_{k_2}^{(2)}\right] \\ &= (M^2d+1)\sigma^2. \end{aligned}$$

Therefore, we have the following inequality

$$\begin{aligned} V\left[\widetilde{f}_N(x)\right] &\leq L(M^2d+1)\sigma^2+\sigma^2(dLM+L)^2+\sigma^2(M^2d+1)\sigma^2+\sigma^2 \\ &= \{L(M^2d+1)+(dLM+L)^2+1\}\sigma^2+(M^2d+1)\sigma^4. \end{aligned}$$

□

3 Proof of Theorem 1

Our proof is completed by revisiting the argument established in Foong–Burt–Li–Turner [11, Theorem 6] to reveal the dependence concerning N of all constants.

We construct $\widetilde{f}_N$ using the deterministic optimal neural network. By Theorem 2, for all $N \in \mathbb{N}$ and nonnegative $f \in H^1(K_0)$, there exists $C' > 0$ and a neural network

$$f_N(x) = \sum_{k_2=1}^{N} w_{1,k_2}^{(3)}\psi\left(\sum_{k_1=1}^{d} w_{k_2,k_1}^{(2)}x_{k_1}-\theta_{k_2}^{(2)}\right),$$

such that

$$\|f_N - f\|_{L^2(K_0)} \leq C'N^{-\frac{1}{2}}. \tag{3}$$

For $k_1 = 1, \cdots, d$ and $k_2 = 1, \cdots, N$, let $\xi_{k_2,k_1}^{(2)}$, $\xi_{1,k_2}^{(3)}$, $\xi_{k_2}^{(2)}$, $\xi_1^{(3)}$ be random variables on probability space $(\Omega, \mathcal{F}, P)$, which are independently and identically distributed. Assume that the expectation of $\xi_{k_2,k_1}^{(2)}$, $\xi_{1,k_2}^{(3)}$, $\xi_{k_2}^{(2)}$, $\xi_1^{(3)}$ is

zero, the variance is $\sigma^2 > 0$, and possible values are finite. We define the random variables $\widetilde{w}^{(2)}_{k_2,k_1}$, $\widetilde{w}^{(3)}_{1,k_2}$, $\widetilde{\theta}^{(2)}_{k_2}$, $\widetilde{\theta}^{(3)}_1$ by

$$\begin{aligned} \widetilde{w}^{(2)}_{k_2,k_1} &:= w^{(2)}_{k_2,k_1} + \xi^{(2)}_{k_2,k_1}, & \widetilde{w}^{(3)}_{1,k_2} &:= w^{(3)}_{1,k_2} + \xi^{(3)}_{1,k_2}, \\ \widetilde{\theta}^{(2)}_{k_2} &:= \theta^{(2)}_{k_2} + \xi^{(2)}_{k_2}, & \widetilde{\theta}^{(3)}_1 &:= \xi^{(3)}_1. \end{aligned}$$

We remark that there exists $L > 0$ such that $|\widetilde{w}^{(2)}_{k_2,k_1}|, |\widetilde{w}^{(3)}_{1,k_2}|, |\widetilde{\theta}^{(2)}_{k_2}|, |\widetilde{\theta}^{(3)}_1| \leq L$ since possible values of $\xi^{(2)}_{k_2,k_1}$, $\xi^{(3)}_{1,k_2}$, $\xi^{(2)}_{k_2}$, $\xi^{(3)}_1$ are finite. Then, we define $\widetilde{f}_N$ by

$$\widetilde{f}_N(x) := \sum_{k_2=1}^{N} \widetilde{w}^{(3)}_{1,k_2} \psi \left(\sum_{k_1=1}^{d} \widetilde{w}^{(2)}_{k_2,k_1} x_{k_1} - \widetilde{\theta}^{(2)}_{k_2} \right) - \widetilde{\theta}^{(3)}_1.$$

We prove the probability $P\left(\left|E[(\widetilde{f}_N(x))] - f_N(x)\right| \geq N^{-\frac{1}{2}}\right)$ vanishes if σ^2 satisfies an appropriate bound. By applying the triangle inequality, we obtain

$$\begin{aligned} &P\left(\left|E[(\widetilde{f}_N(x))] - f_N(x)\right| \geq N^{-\frac{1}{2}}\right) \\ &\leq P\left(\left|E[(\widetilde{f}_N(x))] - f_N(\omega, x)\right| \geq \frac{N^{-\frac{1}{2}}}{2}\right) \\ &\quad + P\left(|f_N(\omega, x) - f_N(x)| \geq \frac{N^{-\frac{1}{2}}}{2}\right), \end{aligned}$$

for a.e. $x \in K_0$ and fixed $\omega \in \Omega$. We denote the first and second term as I_1 and I_2, respectively.

Let $M > 0$ satisfy $|x| \leq M$ for all $x \in K_0$. It follows from Chebyshev's inequality and Lemma 4, such that

$$I_1 \leq 4NV\left[\widetilde{f}_N(x)\right] = C_1 N\sigma^2 + C_2 N\sigma^4, \tag{4}$$

where $C_1 := 4\left\{L(M^2 d + 1) + (dLM + L)^2 + 1\right\}$, $C_2 := 4(M^2 d + 1)$.

We use the explicit representation of f_N and $\widetilde{f}_N$ to estimate the second

term I_2. Applying the triangle inequality, we obtain

$$
\begin{aligned}
I_2 = P\Bigg(\Bigg|\Bigg\{\sum_{k_2=1}^{N} \widetilde{w}_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right) - \widetilde{\theta}_1^{(3)}\Bigg\} \\
- \Bigg\{\sum_{k_2=1}^{N} w_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} w_{k_2,k_1}^{(2)} x_{k_1} - \theta_{k_2}^{(2)}\right)\Bigg\}\Bigg| \geq \frac{N^{-\frac{1}{2}}}{2}\Bigg) \\
= P\Bigg(\Bigg|\Bigg\{\sum_{k_2=1}^{N} \widetilde{w}_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right) - \widetilde{\theta}_1^{(3)}\Bigg\} \\
- \Bigg\{\sum_{k_2=1}^{N} w_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right)\Bigg\}\Bigg| \geq \frac{N^{-\frac{1}{2}}}{4}\Bigg) \\
+ P\Bigg(\Bigg|\Bigg\{\sum_{k_2=1}^{N} w_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \theta_{k_2}^{(2)}\right) - \widetilde{\theta}_1^{(3)}\Bigg\} \\
- \Bigg\{\sum_{k_2=1}^{N} w_{1,k_2}^{(3)} \psi\left(\sum_{k_1=1}^{d} w_{k_2,k_1}^{(2)} x_{k_1} - \theta_{k_2}^{(2)}\right)\Bigg\}\Bigg| \geq \frac{N^{-\frac{1}{2}}}{4}\Bigg),
\end{aligned}
$$

for all $w \in \Omega$ and a.e. $x \in K_0$. Again, we denote the first and second term as $I_{2,1}$ and $I_{2,2}$, respectively.

It is easy to see that

$$
\begin{aligned}
I_{2,1} &\leq P\left(\sum_{k_2=1}^{N} \left|\widetilde{w}_{1,k_2}^{(3)} - w_{1,k_2}^{(3)}\right| \left|\psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right)\right| + \left|\widetilde{\theta}_1^{(3)}\right| \geq \frac{N^{-\frac{1}{2}}}{4}\right) \\
&\leq \sum_{k_2=1}^{N} P\left(\left|\widetilde{w}_{1,k_2}^{(3)} - w_{1,k_2}^{(3)}\right| \left|\psi\left(\sum_{k_1=1}^{d} \widetilde{w}_{k_2,k_1}^{(2)} x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right)\right| \geq \frac{N^{-\frac{1}{2}}}{4(N+1)}\right) \\
&\quad + P\left(\left|\widetilde{\theta}_1^{(3)}\right| \geq \frac{N^{-\frac{1}{2}}}{4(N+1)}\right).
\end{aligned}
$$

Since ψ is a contraction mapping, by Chebyshev's inequality, there holds

$$\begin{aligned}
I_{2,1} &\leq \sum_{k_2=1}^{N} P\left(\left|\widetilde{w}_{1,k_2}^{(3)} - w_{1,k_2}^{(3)}\right| (dLM+L) \geq \frac{N^{-\frac{1}{2}}}{4(N+1)}\right) \\
&\quad + P\left(\left|\widetilde{\theta}_1^{(3)}\right| \geq \frac{N^{-\frac{1}{2}}}{4(N+1)}\right) \\
&\leq \sum_{k_2=1}^{N} \frac{16(N+1)^2(dLM+L)^2}{N^{-1}}\sigma^2 + \frac{16(N+1)^2}{N^{-1}}\sigma^2 \\
&= \frac{16(N+1)^2}{N^{-1}}\left\{1+(dLM+L)^2\right\}\sigma^2 \\
&\leq 64N^3\left\{1+(dLM+L)^2\right\}\sigma^2. \qquad (5)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&I_{2,2} \\
&\leq P\left(\sum_{k_2=1}^{N}\left|w_{1,k_2}^{(3)}\right| \left|\psi\left(\sum_{k_1=1}^{d}\widetilde{w}_{k_2,k_1}^{(2)}x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right)\right.\right. \\
&\qquad\qquad \left.\left. -\psi\left(\sum_{k_1=1}^{d} w_{k_2,k_1}^{(2)}x_{k_1} - \theta_{k_2}^{(2)}\right)\right| \geq \frac{N^{-\frac{1}{2}}}{4}\right) \\
&\leq P\left(\sum_{k_2=1}^{N} L\left|\left(\sum_{k_1=1}^{d}\widetilde{w}_{k_2,k_1}^{(2)}x_{k_1} - \widetilde{\theta}_{k_2}^{(2)}\right) - \left(\sum_{k_1=1}^{d} w_{k_2,k_1}^{(2)}x_{k_1} - \theta_{k_2}^{(2)}\right)\right| \geq \frac{N^{-\frac{1}{2}}}{4}\right),
\end{aligned}$$

by the contraction property of ψ. Applying the triangle inequality and

Chebyshev's inequality, we obtain

$$
\begin{aligned}
I_{2,2} &\le \sum_{k_2=1}^{N} P\left(\left|\sum_{k_1=1}^{d}\left(\widetilde{w}_{k_2,k_1}^{(2)}-w_{k_2,k_1}^{(2)}\right)x_{k_1}-\left(\widetilde{\theta}_{k_2}^{(2)}-\theta_{k_2}^{(2)}\right)\right|\ge\frac{N^{-\frac{1}{2}}}{4NL}\right)\\
&\le \sum_{k_2=1}^{N} P\left(\sum_{k_1=1}^{d}\left|\widetilde{w}_{k_2,k_1}^{(2)}-w_{k_2,k_1}^{(2)}\right||x_{k_1}|+\left|\widetilde{\theta}_{k_2}^{(2)}-\theta_{k_2}^{(2)}\right|\ge\frac{N^{-\frac{1}{2}}}{4NL}\right)\\
&\le \sum_{k_2=1}^{N}\sum_{k_1=1}^{d}\left\{P\left(\left|\widetilde{w}_{k_2,k_1}^{(2)}-w_{k_2,k_1}^{(2)}\right|M\ge\frac{N^{-\frac{1}{2}}}{4N(d+1)L}\right)\right.\\
&\qquad\left.+\,P\left(\left|\widetilde{\theta}_{k_2}^{(2)}-\theta_{k_2}^{(2)}\right|\ge\frac{N^{-\frac{1}{2}}}{4N(d+1)L}\right)\right\}\\
&\le \sum_{k_2=1}^{N}\left(\sum_{k_1=1}^{d}\frac{16N^2(d+1)^2L^2M^2}{N^{-1}}\sigma^2+\frac{16N^2(d+1)^2L^2}{N^{-1}}\sigma^2\right)\\
&= 16N^4(dM^2+1)(d+1)^2L^2\sigma^2. \qquad (6)
\end{aligned}
$$

By (5) and (6), we obtain the estimate

$$
I_2 \le C_3N^3\sigma^2+C_4N^4\sigma^2, \qquad (7)
$$

where $C_3 := 64\left\{1+(dLM+L)^2\right\}$, $C_4 := 16(dM^2+1)(d+1)^2L^2$.

Combining (4) and (7), we get

$$
\begin{aligned}
&P\left(\left|E[(\widetilde{f}_N(x))]-f_N(x)\right|\ge N^{-\frac{1}{2}}\right)\\
&\le C_1N\sigma^2+C_2N\sigma^4+C_3N^3\sigma^2+C_4N^4\sigma^2\\
&\le C_5\left(N\sigma^2+N\sigma^4+N^3\sigma^2+N^4\sigma^2\right),
\end{aligned}
$$

where $C_5 := \max\{C_1,C_2,C_3,C_4\}$. Setting $C_0=(2C_5)^{-2}$, if $0<\sigma^2\le C_0N^{-4}$,

$$
\begin{aligned}
&P\left(\left|E[(\widetilde{f}_N(x))]-f_N(x)\right|\ge N^{-\frac{1}{2}}\right)\\
&\le \frac{1}{4N^3}+\frac{1}{16N^7}+\frac{1}{4N}+\frac{1}{4}\\
&\le \frac{13}{16}.
\end{aligned}
$$

Since the above estimate indicates that the event $\left\{\left|E[(\widetilde{f}_N(x))] - f_N(x)\right| \geq N^{-\frac{1}{2}}\right\}$ does not occur, we have the following:

$$P\left(\left|E[(\widetilde{f}_N(x))] - f_N(x)\right| \geq N^{-\frac{1}{2}}\right) = 0.$$

This implies that

$$\left|E[(\widetilde{f}_N(x))] - f_N(x)\right| \leq N^{-\frac{1}{2}}. \tag{8}$$

Finally, combining the inequality (3) and (8), we obtain

$$\begin{aligned}\|E[\widetilde{f}_N] - f\|_{L^2(K_0)} &\leq \left(\int_{K_0} \left|E[(\widetilde{f}_N(x))] - f_N(x)\right|^2 d\mu(x)\right)^{\frac{1}{2}} + \|f_N - f\|_{L^2(K_0)} \\ &\leq \mu(K_0)^{\frac{1}{2}} N^{-\frac{1}{2}} + C' N^{-\frac{1}{2}} \\ &= C N^{-\frac{1}{2}},\end{aligned}$$

where $C := \mu(K_0)^{\frac{1}{2}} + C'$. This completes the proof of Theorem 1.

References

[1] W. S. McCulloch, W. Pitts, *A logical calculus of the ideas immanent in nervous activity* , The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943, doi: 10.1007/BF02478259, Available: `https://doi.org/10.1007/BF02478259`.

[2] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Systems, vol. 2, no. 4, pp. 303–314, 1989, issn: 0932-4194, doi: 10.1007/BF02551274, Available: `https://doi.org/10.1007/BF02551274`.

[3] K. Hornik, M. Stinchcombe, H. White, *Multilayer feedforward networks are universal approximators*, Neural Networks, vol. 2, no. 5, pp. 359–366, 1989, issn: 0893-6080. doi: 10.1016/0893-6080(89)90020-8. Available: `https://www.sciencedirect.com/science/article/pii/0893608089900208`.

[4] K. Funahashi, *On the approximate realization of continuous mappings by neural networks*, Neural Networks, vol. 2, no. 3, pp. 183–192, 1989, issn: 0893-6080, doi: 10.1016/0893-6080(89)90003-8, Available: https://www.sciencedirect.com/science/article/pii/0893608089900038.

[5] G. Pisier, *Remarques sur un résultat non publié de b.maurey*, fre, Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz"), pp. 1–12, 1980-1981. Available: http://eudml.org/doc/109255.

[6] L. K. Jones, *A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training*, The Annals of Statistics, vol. 20, no. 1, pp. 608–613, 1992, doi: 10.1214/aos/1176348546, Available: https://doi.org/10.1214/aos/1176348546.

[7] A. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory, vol. 39, no. 3, pp. 930–945, 1993, doi: 10.1109/18.256500.

[8] Y. Makovoz, *Random approximants and neural networks*, Journal of Approximation Theory, vol. 85, no. 1, pp. 98–109, 1996, issn: 0021-9045, doi: 10.1006/jath.1996.0031, Available: https://www.sciencedirect.com/science/article/pii/S0021904596900313.

[9] C. Ikuta, Y. Uwate, Y. Nishio, *Chaos glial network connected to multilayer perceptron for solving two-spiral problem*, in Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp. 1360–1363. doi: 10.1109/ISCAS.2010.5537060.

[10] D. J. C. MacKay, *A Practical Bayesian Framework for Backpropagation Networks*, Neural Computation, vol. 4, no. 3, pp. 448–472, May 1992, issn: 0899-7667, doi: 10.1162/neco.1992.4.3.448, eprint: https://direct.mit.edu/neco/article-pdf/4/3/448/812348/neco.1992.4.3.448.pdf. Available: https://doi.org/10.1162/neco.1992.4.3.448.

[11] A. Y. K. Foong, D. R. Burt, Y. Li, R. E. Turner, *On the expressiveness of approximate inference in bayesian neural networks*, 2020, arXiv: 1909.00719.