Contents

1

Approximation Properties of Neural Network Function Class

1.1 Qualitative convergence results

The first question about DNN_1 is about the approximation properties for any continuous functions. Here we have the next theorem.

The first proof for this lemma above can be found in [**leshno1993multilayer**] and summarized in [**pinkus1999approximation**]. The next theorem plays an important role in the proof of above lemma, which is first proved in [**leshno1993multilayer**] with several steps. Here we can present a more direct and simple version.

Theorem 1 (Universal Approximation Property of Shallow Neural Networks). *Let σ be a Riemann integrable function and $\sigma \in L^\infty_{loc}(\mathbb{R})$. Then $\Sigma_d(\sigma)$ in dense in $C(\Omega)$ for any compact $\Omega \subset \mathbb{R}^n$ if and and only if σ is not a polynomial!*

Namely, if σ is not a polynomial, then, for any $f \in C(\bar{\Omega})$, there exists a sequence $\phi_n \in \mathrm{DNN}_1$ such that

$$\max_{x\in\bar{\Omega}} |\phi_n(x) - f(x)| \to 0, \quad n \to \infty.$$

Proof. Let us first prove the theorem in a special case that $\sigma \in C^\infty(\mathbb{R})$. Since $\sigma \in C^\infty(\mathbb{R})$, it follows that for every ω, b,

$$\frac{\partial}{\partial \omega_j}\sigma(\omega \cdot x + b) = \lim_{n\to\infty} \frac{\sigma((\omega + he_j)\cdot x + b) - \sigma(\omega\cdot x + b)}{h} \in \overline{\Sigma}_d(\sigma) \tag{1.1}$$

for all $j = 1, ..., d$.

By the same argument, for $\alpha = (\alpha_1, ..., \alpha_d)$

$$D^\alpha_\omega \sigma(\omega \cdot x + b) \in \overline{\Sigma}_d(\sigma)$$

for all $k \in \mathbb{N}$, $j = 1, ..., d$, $\omega \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Now

$$D^\alpha_\omega \sigma(\omega \cdot x + b) = x^\alpha \sigma^{(k)}(\omega \cdot x + b)$$

where $k = |\alpha|$ and $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. Since σ is not a polynomial there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. Taking $\omega = 0$ and $b = \theta_k$, we thus see that $x_j^k \in \overline{\Sigma}_d(\sigma)$. Thus, all polynomials of the form $x_1^{k_1} \cdots x_d^{k_d}$ are in $\overline{\Sigma}_d(\sigma)$.

This implies that $\overline{\Sigma}_d(\sigma)$ contains all polynomials. By Weierstrass's Theorem [**stone1948generalized**] it follows that $\overline{\Sigma}_d(\sigma)$ contains $C(K)$ for each compact $K \subset \mathbb{R}^n$. That is $\Sigma_d(\sigma)$ is dense in $C(\mathbb{R}^d)$.

Now we consider the case that σ is only Riemann integrable. Consider the mollifier η

$$\eta(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}.$$

Set $\eta_\epsilon = \frac{1}{\epsilon}\eta(\frac{x}{\epsilon})$. Then consider σ_{η_ϵ}

$$\sigma_{\eta_\epsilon}(x) := \sigma * \eta_\epsilon(x) = \int_{\mathbb{R}} \sigma(x-y)\eta_\epsilon(y)dy \tag{1.2}$$

It can be seen that $\sigma_{\eta_\epsilon} \in C^\infty(\mathbb{R})$. We first notice that $\overline{\Sigma}_1(\sigma_{\eta_\epsilon}) \subset \overline{\Sigma}_1(\sigma)$, which can be done easily by checking the Riemann sum of $\sigma_{\eta_\epsilon}(x) = \int_{\mathbb{R}} \sigma(x-y)\eta_\epsilon(y)dy$ is in $\overline{\Sigma}_1(\sigma)$.

Following the argument in the beginning of the proof proposition, we want to show that $\overline{\Sigma}_1(\sigma_{\eta_\epsilon}))$ contains all polynomials. For this purpose, it suffices to show that there exists θ_k and σ_{η_ϵ} such that $\sigma_{\eta_\epsilon}^{(k)}(\theta_k) \neq 0$ for each k. If not, then there must be k_0 such that $\sigma_{\eta_\epsilon}^{(k_0)}(\theta) = 0$ for all $\theta \in \mathbb{R}$ and all $\epsilon > 0$. Thus σ_{η_ϵ}'s are all polynomials with degree at most $k_0 - 1$. In particular, It is known that $\eta_\epsilon \in C_0^\infty(\mathbb{R})$ and $\sigma * \eta_\epsilon$ uniformly converges to σ on compact sets in $\mathbb{R}$ and $\sigma * \eta_\epsilon$'s are all polynomials of degree at most $k_0 - 1$. Polynomials of a fixed degree form a closed linear subspace, therefore σ is also a polynomial of degree at most $k_0 - 1$, which leads to contradiction. □

Properties of polynomials using Fourier transform

We make use of the theory of tempered distributions (see [**strichartz2003guide**, **stein2016introduction**] for an introduction) and we begin by collecting some results of independent interest, which will also be important later. We begin by noting that an activation function σ which satisfies a polynomial growth condition $|\sigma(x)| \leq C(1+|x|)^n$ for some constants C and n is a tempered distribution. As a result, we make this assumption on our activation functions in the following theorems. We briefly note that this condition is sufficient, but not necessary (for instance an integrable function need not satisfy a pointwise polynomial growth bound) for σ to be represent a tempered distribution.

We begin by studying the convolution of σ with a Gaussian mollifier. Let η be a Gaussian mollifier

$$\eta(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}. \tag{1.3}$$

Set $\eta_\epsilon = \frac{1}{\epsilon}\eta(\frac{x}{\epsilon})$. Then consider σ_ϵ

$$\sigma_\epsilon(x) := \sigma * \eta_\epsilon(x) = \int_{\mathbb{R}} \sigma(x-y)\eta_\epsilon(y)dy \tag{1.4}$$

for a given activation function σ.

It is clear that $\sigma_\epsilon \in C^\infty(\mathbb{R})$. Moreover, by considering the Fourier transform (as a tempered distribution) we see that

$$\hat{\sigma}_\epsilon = \hat{\sigma}\hat{\eta}_\epsilon = \hat{\sigma}\eta_{\epsilon^{-1}}. \tag{1.5}$$

We begin by stating a lemma which characterizes the set of polynomials in terms of their Fourier transform.

Lemma 1. *Given a tempered distribution σ, the following statements are equivalent:*

1. *σ is a polynomial*
2. *σ_ϵ given by* (1.4) *is a polynomial for any $\epsilon > 0$.*
3. $\mathrm{supp}(\hat{\sigma}) \subset \{0\}$.

Proof. We begin by proving that (3) and (1) are equivalent. This follows from a characterization of distributions supported at a single point (see [**strichartz2003guide**], section 6.3). In particular, a distribution supported at 0 must be a finite linear combination of Dirac masses and their derivatives. In particular, if $\hat{\sigma}$ is supported at 0, then

$$\hat{\sigma} = \sum_{i=1}^{n} a_i \delta^{(i)}. \tag{1.6}$$

Taking the inverse Fourier transform and noting that the inverse Fourier transform of $\delta^{(i)}$ is $c_i x^i$, we see that σ is a polynomial. This shows that (3) implies (1), for the converse we simply take the Fourier transform of a polynomial and note that it is a finite linear combination of Dirac masses and their derivatives.

Finally, we prove the equivalence of (2) and (3). For this it suffices to show that $\hat{\sigma}$ is supported at 0 iff $\hat{\sigma}_\epsilon$ is supported at 0. This follows from equation 1.5 and the fact that $\eta_{\epsilon^{-1}}$ is nowhere vanishing. □

As an application of Lemma 1, let us give a simple proof of the following result. The first proof of this result can be found in [**leshno1993multilayer**] and is summarized in [**pinkus1999approximation**]. Extending this result to the case of non-smooth activation is first done in several steps in [**leshno1993multilayer**]. Our contribution is to provide a much simpler argument based on Fourier analysis.