Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contents

1

FEM and DNN

1.1 Motivation: from finite element to neural network

In this chapter, we will introduce the so-called shallow neural network (deep neural network with one hidden layer) from the viewpoint of finite element method.

Let us first consider the linear finite element functions on the unit interval $\bar{\Omega} = [0, 1]$ in 1D. We then consider a set of equidistant girds Ω_ℓ of level ℓ on the unit interval $\bar{\Omega} = [0, 1]$ and mesh length $h_\ell = 2^{-\ell}$. The grid points $x_{\ell,i}$ are given by

$$x_{\ell,i} := ih_\ell, \quad 0 \le i \le 2^\ell. \tag{1.1}$$

For $\ell = 1$, we denote the special hat function:

$$\varphi(x) = \begin{cases} 2x & x \in [0, \frac{1}{2}] \\ 2(1-x) & x \in [\frac{1}{2}, 1] \\ 0, & \text{others} \end{cases}. \tag{1.2}$$

The next diagram shows this basis function:

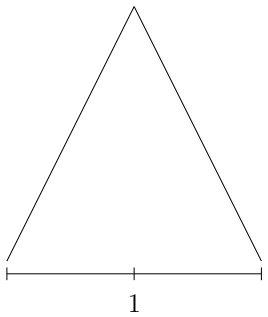

Fig. 1.1. Diagram of $\varphi(x)$

Then, for any nodal basis function $\varphi_{\ell,i}$ as below:

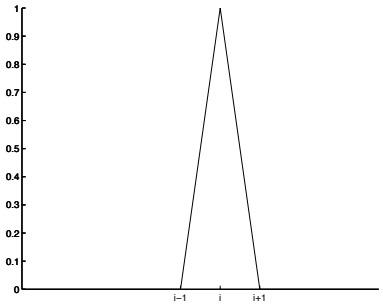

Fig. 1.2. Diagram of $\varphi_{\ell,i}(x)$

in a fine grid $\mathcal{T}_\ell$ can be written as

$$\varphi_{\ell,i} = \varphi(\frac{x - x_{\ell,i-1}}{2h_\ell}) = \varphi(w_\ell x + b_{\ell,i}). \tag{1.3}$$

That is to say, any $\varphi_{\ell,i}(x)$ can be obtained from $\varphi(x)$ ba scaling (dilation) and translation with

$$w_\ell = 2^{\ell-1}, \quad b_{\ell,i} = \frac{-(i-1)}{2}, \tag{1.4}$$

in $\varphi_{\ell,i} = \varphi(w_\ell x + b_{\ell,i})$.

Let recall the finite element interpolation as

$$u(x) \approx u_\ell(x) := \sum_{0 \le i \le 2^\ell} u(x_{\ell,i})\varphi_{\ell,i}(x), \tag{1.5}$$

for any smooth function $u(x)$ on $(0, 1)$. The above interpolation will converge as $\ell \to \infty$, which show that

$$\text{span}\{\varphi(w_\ell x + b_{\ell,i})\} \quad \text{is dense in} \quad H^1(0, 1). \tag{1.6}$$

Thus, we may have the next concise relation:

(1.7)

$$\text{FE space} = \text{span}\left\{\varphi(w_\ell x + b_{\ell,i}) \mid 0 \le i \le 2^\ell, \ell = 1, 2, \cdots\right\} \subset \text{span}\{\varphi(wx + b) \mid w, b \in \mathbb{R}\}.$$

In other words, the finite element space can be understood as the linear combination of $\varphi(wx + b)$ with certain special choice of w and b.

Here, we need to point that this span $\{\varphi(wx + b) \mid w, b \in \mathbb{R}\}$ is exact the deep neural networks with one hidden layer (shallow neural networks) with activation function $\varphi(x)$. More precisely,

$$f \in \text{span}\{\varphi(wx + b) \mid w, b \in \mathbb{R}\}, \tag{1.8}$$

means there exist positive integer N and $w_j, b_j \in \mathbb{R}$ such that

$$f = \sum_{j=1}^{N} a_j \varphi(w_j x + b_j). \tag{1.9}$$

The above function is also called one hidden neural network function with N neurons.

Remark 1. 1. By making w_ℓ and $b_{\ell,i}$ arbitrary, we get a much larger class of function which is exact a special neural network with activation function $\varphi(x)$.

2. Generalizations:
   a) φ can be different, such as $\mathrm{ReLU}(x) = \max\{0, x\}$.
   b) There is a natural extension for hight dimension d as

$$\{\varphi(w \cdot x + b)\}, \tag{1.10}$$

   where $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $w \cdot x = \sum_{i=1}^{d} w_i x_i$. This is called "deep" neural network with one hidden layer.

1.2 Why we need deep neural networks via composition

1.2.1 FEM ans DNN_1 in 1D

Thanks to the connection between $\varphi(x)$ and $\mathrm{ReLU}(x) = \{0, x\}$

$$\varphi(x) = 2\mathrm{ReLU}(x) - 4\mathrm{ReLu}(x - \frac{1}{2}) + 2\mathrm{ReLU}(x - 1). \tag{1.11}$$

It suffices to show that each basis function $\varphi_{\ell,i}$ can be represented by a ReLU DNN. We first note that the basis function φ_i has the support in $[x_{i-1}, x_{i+1}]$ can be easily written as

$$\varphi_{\ell,i}(x) = \frac{1}{h_\ell}\mathrm{ReLU}(x - x_{\ell,i-1}) - (\frac{2}{h_\ell})\mathrm{ReLU}(x - x_{\ell,i}) + \frac{1}{h_\ell}\mathrm{ReLU}(x - x_{\ell,i+1}). \tag{1.12}$$

More generally, if function φ_i is not on the uniform grid but has support in $[x_{i-1}, x_{i+1}]$ can be easily written as

$$\varphi_i(x) = \frac{1}{h_{i-1}}\mathrm{ReLU}(x - x_{i-1}) - (\frac{1}{h_{i-1}} + \frac{1}{h_i})\mathrm{ReLU}(x - x_i) + \frac{1}{h_i}\mathrm{ReLU}(x - x_{i+1}), \tag{1.13}$$

where $h_i = x_{i+1} - x_i$.

Thus is to say, we have the next theorem.

Theorem 1. *For $d = 1$, and $\Omega \subset \mathbb{R}^d$ is a bounded interval, then* DNN_1 *can be used to cover all linear finite element function in on Ω.*

1.2.2 Linear finite element cannot be recovered by DNN_1 for $d \geq 2$

In view of Theorem 1 and the fact that $\mathrm{DNN}_J \subseteq \mathrm{DNN}_{J+1}$, it is natural to ask that how many layers are needed at least to recover all linear finite element functions in $\mathbb{R}^d$ for $d \geq 2$. In this section, we will show that

$$J_d \geq 2, \quad \text{if} \quad d \geq 2, \tag{1.14}$$

where J_d is the minimal J such that all linear finite element functions in $\mathbb{R}^d$ can be recovered by DNN_J.

In particular, we will show the following theorem.

Theorem 2. *If $\Omega \subset \mathbb{R}^d$ is either a bounded domain or $\Omega = \mathbb{R}^d$, DNN_1 can not be used to recover all linear finite element functions on Ω.*

Proof. We prove it by contradiction. Let us assume that for any continuous piecewise linear function $f : \Omega \to \mathbb{R}$, we can find finite $N \in \mathbb{N}$, $w_i \in \mathbb{R}^{1,d}$ as row vector and $\alpha_i, b_i, \beta \in \mathbb{R}$ such that

$$f = \sum_{i=1}^{N} \alpha_i \mathrm{ReLU}(w_i x + b_i) + \beta,$$

with $f_i = \alpha_i \mathrm{ReLU}(w_i x + b_i)$, $\alpha_i \neq 0$ and $w_i \neq 0$. Consider the finite element functions, if this one hidden layer ReLU DNN can recover any basis function of FEM, then it can recover the finite element space. Thus let us assume f is a locally supported basis function for FEM. Furthermore, if Ω is a bounded domain, we assume that

$$d(\mathrm{supp}(f), \partial\Omega) > 0, \tag{1.15}$$

with

$$d(A, B) = \inf_{x \in A, y \in B} \|x - y\|,$$

as the distance of two closed sets.

A more important observation is that $\nabla f : \Omega \to \mathbb{R}^d$ is a piecewise constant vector function. The key point is to consider the discontinuous points for $g := \nabla f = \sum_{i=1}^{N} \nabla f_i$.

For more general case, we can define the set of discontinuous points of a function by

$$D_g := \{x \in \Omega \mid x \text{ is a discontinuous point of } g\}.$$

Because of the property that

$$D_{f+g} \supseteq D_f \cup D_g \backslash (D_f \cap D_g), \tag{1.16}$$

we have

$$D_{\sum_{i=1}^{N} g_i} \supseteq \bigcup_{i=1}^{N} D_{g_i} \backslash \bigcup_{i \neq j} \left(D_{g_i} \cap D_{g_j} \right). \tag{1.17}$$

Note that

$$g_i = \nabla f_i(x) = \nabla\left(\alpha_i \mathrm{ReLU}(w_i x + b_i)\right) = \left(\alpha_i H(w_i x + b_i)\right) w_i \in \mathbb{R}^d, \tag{1.18}$$

for $i = 1 : N$ with H be the Heaviside function defined as:

$$H(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{if } x > 0. \end{cases}$$

This means that

$$D_{g_i} = \{x \mid w_i x + b_i = 0\} \tag{1.19}$$

is a $d-1$ dimensional affine space in $\mathbb{R}^d$.

Without loss of generality, we can assume that

$$D_{g_i} \neq D_{g_j}. \tag{1.20}$$

When the other case occurs, i.e. $D_{g_{\ell_1}} = D_{g_{\ell_2}} = \cdots = D_{g_{\ell_k}}$, by the definition of g_i in (1.18) and D_{g_i} in (1.19) , this happens if and only if there is a row vector (w, b) such that

$$c_{\ell_i}\begin{pmatrix} w & b \end{pmatrix} = \begin{pmatrix} w_{\ell_i} & b_{\ell_i} \end{pmatrix}, \tag{1.21}$$

with some $c_{\ell_i} \neq 0$ for $i = 1 : k$. We combine those g_{ℓ_i} as

$$\begin{aligned} \tilde{g}_\ell &= \sum_{i=1}^k g_{\ell_i} = \sum_{i=1}^k \alpha_{\ell_i} H(w_{\ell_i} x + b_{\ell_i}) w_{\ell_i}, \\ &= \sum_{i=1}^k \left(c_{\ell_i} \alpha_{\ell_i} H\left(c_{\ell_i}(wx + b)\right)\right) w, \\ &= \begin{cases} \left(\sum_{i=1}^k c_{\ell_i} \alpha_{\ell_i} H(c_{\ell_i})\right) w & \text{if} \quad wx + b > 0, \\ \left(\sum_{i=1}^k c_{\ell_i} \alpha_{\ell_i} H(-c_{\ell_i})\right) w & \text{if} \quad wx + b \le 0. \end{cases} \end{aligned}$$

Thus, if

$$\left(\sum_{i=1}^k c_{\ell_i} \alpha_{\ell_i} H(c_{\ell_i})\right) = \left(\sum_{i=1}^k c_{\ell_i} \alpha_{\ell_i} H(-c_{\ell_i})\right),$$

$\tilde{g}_\ell$ is a constant vector function, that is to say $D_{\sum_{i=1}^k g_{\ell_i}} = D_{\tilde{g}_\ell} = \emptyset$. Otherwise, $\tilde{g}_\ell$ is a piecewise constant vector function with the property that

$$D_{\sum_{i=1}^k g_{\ell_i}} = D_{\tilde{g}_\ell} = D_{g_{\ell_i}} = \{x \mid wx + b = 0\}.$$

This means that we can use condition (1.21) as an equivalence relation and split $\{g_i\}_{i=1}^N$ into some groups, and we can combine those g_{ℓ_i} in each group as what we do above. After that, we have

$$\sum_{i=1}^{N} g_i = \sum_{\ell=1}^{\tilde{N}} \tilde{g}_\ell,$$

with $D_{\tilde{g}_s} \neq D_{\tilde{g}_t}$. Finally, we can have that $D_{\tilde{g}_s} \cap D_{\tilde{g}_t}$ is an empty set or a $d-2$ dimensional affine space in $\mathbb{R}^d$. Since $\tilde{N} \leq N$ is a finite number,

$$D := \bigcup_{i=1}^{N} D_{\tilde{g}_\ell} \setminus \bigcup_{s \neq t} \left(D_{\tilde{g}_s} \cap D_{\tilde{g}_t} \right)$$

is an unbounded set.

- If $\Omega = \mathbb{R}^d$,
$$\text{supp(f)} \supseteq D_g = D_{\sum_{i=1}^{N} g_i} = D_{\sum_{\ell=1}^{\tilde{N}} \tilde{g}_\ell} \supseteq D,$$
is contradictory to the assumption that f is locally supported.
- If Ω is a bounded domain,
$$d(D, \partial\Omega) = \begin{cases} s > 0 & \text{if} \quad D_{\tilde{g}_i} \cap \Omega = \emptyset, \forall i \\ 0 & \text{otherwise.} \end{cases}$$
Note again that all $D_{\tilde{g}_i}$'s are $d-1$ dimensional affine spaces, while $D_{\tilde{g}_i} \cap D_{\tilde{g}_j}$ is either an empty set or a d-2 dimensional affine space. If $d(D, \partial\Omega) > 0$, this implies that ∇f is continuous in Ω, which contradicts the assumption that f is a basis function in FEM. If $d(D, \partial\Omega) = 0$, this contradicts the previous assumption in (1.15).

Hence DNN_1 cannot recover any piecewise linear function in Ω for $d \geq 2$. □

Following the proof above, we have the following theorem.

Theorem 3. $\{\text{ReLU}(w_i x + b_i)\}_{i=1}^{m}$ *are linearly independent if* (w_i, b_i) *and* (w_j, b_j) *are linearly independent in* $\mathbb{R}^{1\times(d+1)}$ *for any* $i \neq j$.

1.3 Definition of neural network space

1. Primary variables $n_0 = d$
$$x^0 = x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$
2. n_1 hyperplanes
$$W^1 x + b^1 = \begin{pmatrix} w_1^1 x + b_1^1 \\ w_2^1 x + b_2^1 \\ \vdots \\ w_n^1 x + b_n^1 \end{pmatrix}$$
3. n_1-neurons:
$$x^1 = \sigma(W^1 x + b^1) = \begin{pmatrix} \sigma(w_1^1 x + b_1^1) \\ \sigma(w_2^1 x + b_2^1) \\ \vdots \\ \sigma(w_n^1 x + b_n^1) \end{pmatrix}$$
4. n_2-hyperplanes
$$W^2 x^1 + b^2 = \begin{pmatrix} w_1^2 x^1 + b_1^2 \\ w_2^2 x^1 + b_2^2 \\ \vdots \\ w_n^2 x^1 + b_n^2 \end{pmatrix}$$

Shallow neural network functions:

(1.22)
$${}_n\mathrm{N}(n_1, n_2) = {}_n\mathrm{N}(\sigma; n_1, n_2) = \left\{ W^2 x^1 + b^2, x^1 = \sigma(W^1 x + b^1) \text{ with } W^\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, b^\ell \in \mathbb{R}^{n_\ell}, \ell = 1, 2, n_0 = d \right\}$$

(1.23)
$${}_n\mathrm{N}(\sigma; n_1, n_2, \ldots, n_L) = \left\{ W^L x^{L-1} + b^L, x^\ell = \sigma(W^\ell x^{\ell-1} + b^\ell) \text{ with } W^\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, b^\ell \in \mathbb{R}^{n_\ell}, \ell = 1 : L, n_0 = d, x^0 = x \right\}$$

First, let us first define the so-called 1-hidden layer (shallow) neural network.

The 1-hidden layer (shallow) neural network is defined as:

(1.24)
$${}_n\mathrm{N} = {}_n\mathrm{N}(\sigma) = {}_n\mathrm{N}^1(\sigma) = \bigcup_{n_1 \geq 1} {}_n\mathrm{N}(\sigma; n_1, 1)$$

The 2-hidden layer (shallow) neural network is defined as:

(1.25)
$${}_n\mathrm{N}^2(\sigma) = \bigcup_{n_1, n_2 \geq 1} {}_n\mathrm{N}(\sigma; n_1, n_2, 1)$$