(千葉大学審査学位論文)

データ数に基づいたニューラルネットワークの表現能力の指標「表現数」の提案および ReLU ニューラルネットワークにおける表現能力の表現数に基づく評価

2020 年 1 月

千葉大学大学院理学研究科
基盤理学専攻数学・情報数理学コース

井上健太

本論文では, 機械学習で用いられるグラフであるニューラルネットワークの表現能力に関して論じる. 表現能力とは, どれほど複雑な関数を表現できるかを表しており, いくつかの指標が提案されている. しかし, 既存の指標からは, 機械学習を行う際に用いる学習データが与えられたとき, それらのデータが学習を行うニューラルネットワークで表現可能かどうか判別できない. そこで, この問題点を解決するような新たな表現能力の指標として「表現数」を提案する. 表現数はデータ数をどこまで限定すれば任意のデータが表現可能となるかによって定義される指標で, 学習データ数と表現数を比較することでデータの表現可能性を調べることができる. 本論文では, この表現数に関する基本的な性質を証明し, 活性化関数を ReLU 関数に限定したときの表現数の取りうる最大値とニューラルネットワークにおけるニューロン数との関係を調べる. 特に, ニューラルネットワークの中間層が 1 層の場合の最大表現数の上界と下界, 2 層の場合の下界と多層の場合の上界を求める. この結果から, 中間層が 2 層までのニューラルネットワークの最大表現数は, 各層のニューロン数の積に比例することが判明した.

目次

第 1 章

まえがき

1.1 研究背景

ニューラルネットワークは画像認証 [1, 2] や音声認識 [3] 等, 機械学習の様々な分野において有用なグラフとして知られている. ニューラルネットワークを用いる学習では, 関数の入出力の組である有限個の学習データを元に, ニューラルネットワークの各ニューロン, エッジに与えられた変数を調節し, 学習データに近似していく. ニューラルネットワークが広く用いられている理由の一つとして, ニューラルネットワークの持つ表現能力の高さが挙げられる. 表現能力とは, どれほど複雑な関数を表現できるかを表しており, ハイパーパラメータと呼ばれる, 学習の際に変化しない中間層数, 各中間ニューロン数, 活性化関数等のパラメータに依存する. しかしながら, ハイパーパラメータと表現能力の関係について詳しく知られておらず, 実際の学習において, このハイパーパラメータを経験的ないしランダム [4] に決定して使用しているのが現状である. そこで, ニューラルネットワークの表現能力に着目する.

ニューラルネットワークの表現能力, すなわちどれほど複雑な関数を表現できるかを評価するためには, 複雑さの指標を定義する必要がある. 既に, 表現能力の指標として, 線型領域の数に基づいたもの [5] や VC 次元 [6] と呼ばれるものが知られている. しかしながら, 実際の学習において, 既存の指標からは, 与えられた学習データが学習を行うニューラルネットワークで表現可能かどうかを判別できないという問題点がある. そこで本論文では, 表現可能なデータ数に着目した, 新たなニューラルネットワークの表現能力の指標として, 表現数というものを定義する.

表現数は, ニューラルネットワークに対し, データ数 N を固定した任意の N 個のデータが表現可能であるとき, そのニューラルネットワークは表現数 N を持つと定義される. この定義により, 具体的な学習において, 学習データ数が表現数より小さい場合, その学習データが表現可能であることが分かる. 学習データの表現可能性を知ることで, 学習における問題の一つである過学習が起きる可能性を減らせると考えている. それは, 過学習が生じる原因の一つとして, ニューラルネットワークの表現能力が学習データに対して大きすぎる場合が挙げられており [7], 表現数を知ることで, この状態を防ぐことができるためである. また, 表現数はハイパーパラメータの一つである活性化関数の種類を制限しない定義であることも大きな特徴として挙げられる. この性質は表現能力の指標として一般的に持つものではなく, 例えば, 線型領域の数に基づいた指標では, 活性化関数が ReLU 関数ないし区分線型関数のときのみ定義される. したがって, 詳細に述べることはしないが, 表現数によって, 活性化関数の違いによるニューラルネットワークの表現能力の比較を行うことができると期待できる.

本論文では, 表現数に関する数学的な性質について述べる. 特に, ニューラルネットワークにおける入出力の次元と表現数との関係や, 活性化関数を学習において頻繁に用いられる ReLU 関数に固定し

たときの表現数の具体的な値を調べる.

2 章でニューラルネットワークと表現数を定義する. 3 章では, 表現数の持つ一般的な性質を述べる. 4 章で活性化関数を ReLU 関数に固定したときの具体的な表現数の値に関して述べる. 特に, 中間層が 1 層の場合の最大表現数の上界と下界, 多層ニューラルネットワークの表現数の上界と, 中間層が 2 層のときの下界を求める.

1.2 関連研究

ニューラルネットワークにおける表現能力はいくつかの指標によって評価されている. 一つは線型領域の最大数による指標である [5, 8, 9, 10]. 線型領域とは, 活性化関数が ReLU 関数, もしくは区分線型関数のときに定義され, ニューラルネットワークが表す関数の返り値が線型となる極大かつ連結な入力の部分集合である. 複雑な関数を表現するためにはこの線型領域の数が必要になるため, ニューラルネットワークで作ることができる線型領域の個数の最大値を表現能力の指標として用いる. しかし, 各領域は独立ではなく, その傾きに依存関係があるため, 線型領域数の単純な比較によってニューラルネットワークの表現能力を評価できないという問題点がある. また, 線型領域は特定の活性化関数のときのみにしか定義できないため, 異なる活性化関数による表現能力の比較を行うことが難しい. 他の表現能力の指標である結び目 [11] や, 軌跡の長さ [12] によるものも特定の活性化関数でしか定義されないため, 同様のことが言える.

二値分類問題を考えるときには, VC 次元 [6, 13, 14] も表現能力の指標として用いられている. VC 次元は表現数と同じく, データ数に基づいて定義されている. 具体的には, 入力データを固定したとき, 出力が二値であるような任意の指示関数が表現可能となるようなデータ数の最大値で定義される. VC 次元の大きな特徴として, 汎化誤差の上界を求めるのに使用される点が挙げられる. 汎化誤差とは, 学習全体の指標, すなわちニューラルネットワークの表現能力だけでなく学習データや学習アルゴリズムといった学習に関わる要素も含んだ評価を行う指標である. 具体的に, 汎化誤差は学習後のニューラルネットワークに学習データ以外のデータを与えたときの誤差で定義されており, 未知のデータに対し, どれほど元の関数に近い学習ができているかを評価する.

表現数を含む線型領域の数や VC 次元などの指標は, 固定されたハイパーパラメータに対する表現能力の評価である. しかし, ニューラルネットワークにおける表現能力という言葉は他の文脈で使われることがある. それは, ハイパーパラメータは固定せず, むしろ学習を行う関数を固定したとき, どれほどのハイパーパラメータが必要かを評価する場合である. 例えば, 任意の連続関数, 任意の近似誤差に対し, ニューラルネットワークに十分なニューロン数を与えることによって, この関数が近似可能になることが知られている [15, 16]. また, 多変数多項式を表現するのに必要なニューロン数に関する研究もある [17].

本論文では, ニューラルネットワークにおけるハイパーパラメータの一つである活性化関数を ReLU 関数に固定したものを考える. ReLU 関数を用いる理由は, 現在, 学習を行う際に広く用いられいる関数であり, 尚且つ定義がシンプルで扱いやすいためである. 活性化関数は他にも sigmoid 関数や tanh 関数なども用いられるが, 実験的にはこれらの関数を使うより ReLU 関数のほうが学習性能が高いことが知られている [1, 18]. また, 表現能力の観点からも, 同じ関数を近似的に表現する場合, ReLU 関数を用いたものに必要なニューロン数は threshold 関数を用いたものの対数個で済む場合があることが知られている [19]. そのため, ReLU 関数を用いるニューラルネットワークは高い表現能力を有していると考えられる.

第 2 章

準備

この章では, ニューラルネットワークの定義と, 有限個のデータが与えられたときのニューラルネットワークにおける可解性に関する定義を行う.

2.1 記号と関数の表記

$\mathbb{N}$ は 0 を含まない自然数全体の集合を考える. $n, m \in \mathbb{N}$ に対し, $\mathbb{R}^n$ の元を列ベクトル, $\mathbb{R}^{n\times m}$ の元を $n \times m$ 行列で表記する. また, 活性化関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を固定し, $n \in \mathbb{N}$ に対し, $\sigma(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix})$ と書いたとき, $\begin{pmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_n) \end{pmatrix}$ を意味するものだと約束する. また, $\mathbb{R}^n$ における順序関係 $\preceq$ を辞書式順序で定義し, 特に等号を満たさないものを $\prec$ と書くことにする. また, $X \subset \mathbb{R}^n$ に対し, $\dim(X)$ とは X によって生成される線形空間の基底の数によって定義される次元を表すものとする.

証明で用いる定義をいくつか記述しておく.

Definition 1 (ReLU 関数)
$\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}$ を $\mathrm{ReLU}(x) := \max\{0, x\}$ で定義する.

ReLU 関数は活性化関数としてよく用いられる関数である.

Definition 2 (射影関数)
$n \in \mathbb{N}$, $\boldsymbol{w} \in \mathbb{R}^n$ に対し, $\mathrm{proj}_{\boldsymbol{w}} : \mathbb{R}^n \to \mathbb{R}$ を $\mathrm{proj}_{\boldsymbol{w}}(\boldsymbol{x}) := {}^t\boldsymbol{w}\boldsymbol{x}$ で定義する.

Definition 3 (zigzag)
$n, m, k \in \mathbb{N}$, $f : \mathbb{R}^n \to \mathbb{R}^m$, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \in (\mathbb{R}^n)^k$ とする. このとき f が $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ において *zigzag* であるとは, 任意の $1 < i < k$ に対し, $f(\boldsymbol{x}_{i-1}) \prec f(\boldsymbol{x}_i) \succ f(\boldsymbol{x}_{i+1})$ または $f(\boldsymbol{x}_{i-1}) \succ f(\boldsymbol{x}_i) \prec f(\boldsymbol{x}_{i+1})$ を満たすことである. また $X \subset \mathbb{R}^n$ を有限集合としたとき, f が X において *zigzag* とは, X の元を辞書式順序で並べたベクトル $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|X|})$ において f が *zigzag* であることとする.

また, 射影関数に関する性質を示す.

Lemma 1 (辞書式順序を保つ射影関数の存在性)
$n \in \mathbb{N}$ を固定する. このとき任意の有限部分集合 $X \subset \mathbb{R}^n$ に対し, $\mathrm{proj}_{\boldsymbol{w}}$ が X 上の辞書式順序を保つようなベクトル $\boldsymbol{w} \in \mathbb{R}^n$ が存在する.

Proof.
n に関する帰納法を使う.

$n = 1$ のとき, $\boldsymbol{w} := 1$ とおけば明らか.

$n > 1$ のとき, $X' := \{{}^t(x_2, \ldots, x_n) \in \mathbb{R}^{n-1} \mid {}^t(x_1, x_2, \ldots, x_n) \in X\}$ とおくと，帰納法の仮定より $\mathrm{proj}_{\boldsymbol{w}'}$ が X' 上の辞書式順序を保つような $\boldsymbol{w}' \in \mathbb{R}^{n-1}$ が存在する. このとき, $Y := \{x_1 \in \mathbb{R} \mid {}^t(x_1, \ldots, x_n) \in X\}$, $M := \max\{|{}^t\boldsymbol{w}'\boldsymbol{x}'| \mid \boldsymbol{x}' \in X'\}$, $m := \min\{|y_1 - x_1| \mid x_1, y_1 \in Y, x_1 \neq y_1\}$, $w_1 := \begin{cases} \dfrac{2M}{m} + 1 & (|Y| > 1) \\ 0 & (o.w.) \end{cases}$ とおき, $\boldsymbol{w} := \begin{pmatrix} w_1 \\ \boldsymbol{w}' \end{pmatrix}$ と定めれば, $\mathrm{proj}_{\boldsymbol{w}}$ は X 上の辞書式順序を保つ.

2.2 ニューラルネットワーク

まずニューラルネットワークとは何かを定義する.

Definition 4 (ニューラルネットワーク)
$l \in \mathbb{N}$, $(a_0, \ldots, a_l) \in \mathbb{N}^{l+1}$ とする. このとき, 任意の $A \in \prod_{i=1}^{l}(\mathbb{R}^{a_i \times a_{i-1}} \times \mathbb{R}^{a_i})$, つまり $A = ((W_1, \boldsymbol{b}_1), \ldots, (W_l, \boldsymbol{b}_l))$ であり, 各 i に対し $W_i \in \mathbb{R}^{a_i \times a_{i-1}}$, $\boldsymbol{b}_i \in \mathbb{R}^{a_i}$ の形で書けるものを $(a_0, \ldots, a_l)$ ニューラルネットワークと呼び, $(a_0, \ldots, a_l)$ をニューラルネットワーク A の型と呼ぶ.

$(a_0, \ldots, a_l)$ ニューラルネットワークとは, 図 2.1 のように, 入力次元 a_0, 出力次元 a_l で, $1 < l$ のとき, 中間層のニューロン数が入力層から順に $a_1, \ldots, a_{l-1}$ となるものを表しており, A は各階層の係数全てを表している.

Definition 5 (ニューラルネットワーク関数)
$l \in \mathbb{N}$, $A = ((W_1, \boldsymbol{b}_1), \ldots, (W_l, \boldsymbol{b}_l))$ を $(a_0, \ldots, a_l)$ ニューラルネットワークとする. このとき, 任意の $i \in \{1, \ldots, l\}$ に対し, $F_i(\boldsymbol{x}) := W_i\boldsymbol{x} + \boldsymbol{b}_i$ で定義される関数 $F_i : \mathbb{R}^{a_{i-1}} \to \mathbb{R}^{a_i}$ をニューラルネットワーク A における第 i 層関数と呼ぶ. これを用いてニューラルネットワーク関数 $\mathrm{MP}_A^\sigma : \mathbb{R}^{a_0} \to \mathbb{R}^{a_l}$ を以下に定義する.

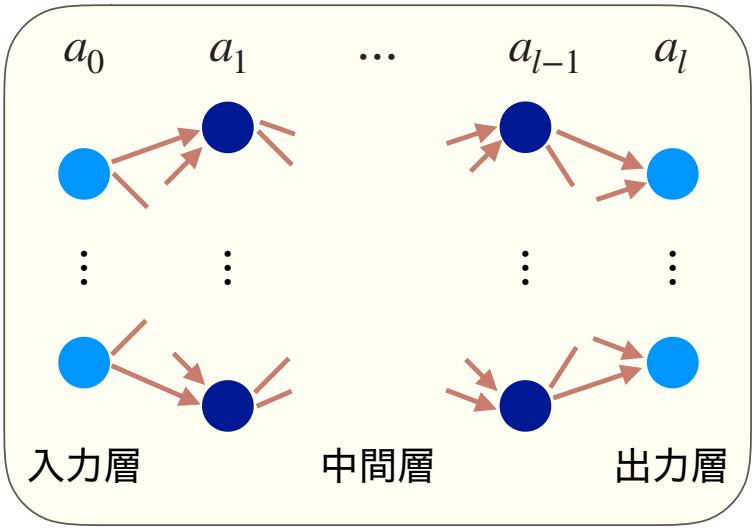
$$\mathrm{MP}_A^\sigma := F_l \circ \sigma \circ F_{l-1} \circ \sigma \circ \cdots \circ \sigma \circ F_2 \circ \sigma \circ F_1$$

特に, 活性化関数 σ が文脈から明らかなときは, MP_A^σ のことを単に MP_A と書く.

ニューラルネットワーク関数 MP_A^σ は, 活性化関数が σ のときのニューラルネットワーク A における入出力を表した関数である.

次章で, ニューラルネットワークにおける表現能力の指標として, 表現数を提案する.

図 2.1　$(a_0, \ldots, a_l)$ ニューラルネットワーク



$(a_0, \ldots, a_l)$ ニューラルネットワークとは, 入力層のニューロン数が a_0 個, 中間層のニューロン数が入力側から順に $a_1, \ldots, a_{l-1}$ 個, 出力層のニューロン数が a_l 個となるニューラルネットワークのグラフの型を表している. $(a_0, \ldots, a_l)$ ニューラルネットワーク A とは, $(a_0, \ldots, a_l)$ ニューラルネットワークが表すグラフの各エッジと入力層以外の各ノードに実数値を割り当てたものであり,

$$\prod_{i=1}^{l} (\mathbb{R}^{a_i \times a_{i-1}} \times \mathbb{R}^{a_i})$$

の要素とみなすことができる.

第 3 章

表現数

本章では, ニューラルネットワークにおける新たな表現能力の指標として, 表現数を定義する. 表現数はデータ数に基づいた指標であり, ニューラルネットワークの持つ表現数以下のデータ数の任意のデータを表現できるという性質を持つ.

まず, 表現数の定義を行う.

3.1 表現数の定義

表現数を定義するためには, データを表現できるとは何かを定義する必要がある. 具体的には, 有限個のデータを与えたときに, それがある型のニューラルネットワークで表現可能かどうかを定義する. 本章では活性化関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を固定する.

Definition 6 (可解性)
$l \in \mathbb{N}$, $(a_0, \ldots, a_l) \in \mathbb{N}^{l+1}$ とし, $n := a_0$, $m := a_l$, $X \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^m$ とする. このとき, **データ** (X, f) **が** $(a_0, \ldots, a_l)$ σ **ニューラルネットワークで可解**とは, $\forall \boldsymbol{x} \in X$, $f(\boldsymbol{x}) = \mathrm{MP}_A^\sigma(\boldsymbol{x})$ となる $(a_0, \ldots, a_l)$ ニューラルネットワーク A が存在することである. 特に活性化関数 σ が文脈上明らかな場合には, 単に $(a_0, \ldots, a_l)$ ニューラルネットワークで可解と書く.

ここで X は任意の部分集合としているが, 本論文では有限集合のみを考えることにする. X とがデータの入力の集合, f が入出力関数を表しており, データ (X, f) が可解であるとは, 図 3.1 のように, 集合 X 上の全ての入力において, その出力と一致するような関数 MP_A^σ がニューラルネットワークで表現できることを意味している. また, 本論文において, 解とは近似解ではなく, 元のデータと出力が一致することを意味することとする.

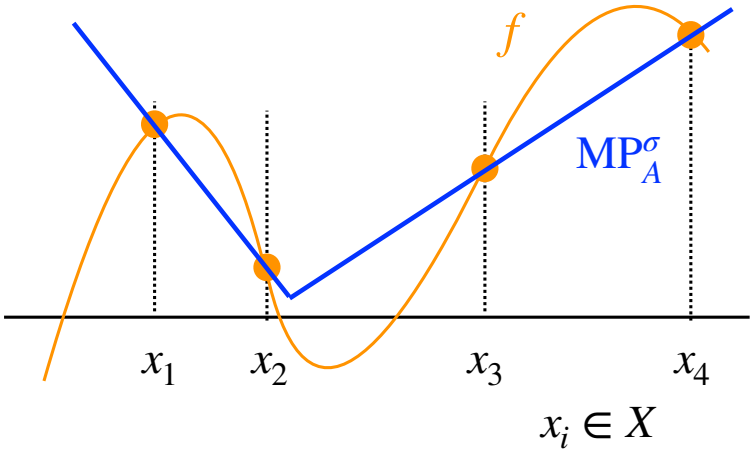
この定義を用いて, ニューラルネットワークの表現能力の指標として提案する表現数を定義する.

Definition 7 (ニューラルネットワークの表現数)
$l, N \in \mathbb{N}$, $(a_0, \ldots, a_l) \in \mathbb{N}^{l+1}$ とし, $n := a_0$, $m := a_l$ とする. このとき, $(a_0, \ldots, a_l)$ σ **ニューラルネットワークが表現数** N **を持つ**とは以下を満たすことである.

- 任意の $|X| = N$ を満たす $X \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^m$ に対し, データ (X, f) は $(a_0, \ldots, a_l)$ σ ニューラルネットワークで可解.

特に, $(a_0, \ldots, a_l)$ σ **ニューラルネットワークの最大表現数**とは, $(a_0, \ldots, a_l)$ σ ニューラルネットワークの表現数の取りうる最大値である. また, 活性化関数 σ が文脈上明らかな場合は, σ ニューラルネットワークの表現数と書く代わりに, 単にニューラルネットワークの表現数と記述する.

図 3.1　データ (X, f) の可解性



データ (X, f) が可解であるとは, 入力の集合 X 上の任意の点 $x_i \in X$ において, その出力 $f(x_i)$ とニューラルネットワーク関数の返り値 $\mathrm{MP}_A^{\sigma}(x_i)$ が一致するようなニューラルネットワークのパラメータ A が存在することを表している.

σ ニューラルネットワークが表現数 N を持つとは, 任意の N 個のデータ全てが可解となるようなものが存在することである. 最大表現数とは, 取りうる表現数の最大数である. この最大表現数をニューラルネットワークの表現能力の指標として提案する. この指標ならば, 具体的なデータが与えられたとき, そのデータ数以上の表現数を持つニューラルネットワークにおいて表現可能だということを保証できる. ただし, データ数が最大表現数を上回る場合でも, 必ずしも表現可能ではないとは言えないことに注意が必要である.

本論文では, データを $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\} \subset \mathbb{R}^n \times \mathbb{R}^m$ のような入力と出力の組ではなく, 入力の集合 X と入出力関数 f で定義している. これは, 前者で定義してしまうと, 表現数を定義する際に, 同じ入力に対し, 複数の出力を返すデータは存在しないことを仮定しないといけないため, 入出力関数で定義する方が記述が簡単であるためである.

表現数は, 任意の活性化関数に対し定義されているため, この指標において, 同じ型の異なる活性化関数の表現能力の比較も行うことができる.

3.2　表現数の持つ一般的な性質

本節では表現数に関する基礎的な性質を証明していく. これらの性質は任意の活性化関数におけるニューラルネットワークで成り立つ.

3.2.1　ニューロン数との間の単調性

まず, 表現数は以下のような直感的に妥当な性質を持つ.

Lemma 2 (中間ニューロン数の大小と表現数)
$n, m, l, N \in \mathbb{N}, (a_1, \ldots, a_l), (b_1, \ldots, b_l) \in \mathbb{N}^l$ とする. このとき任意の $i \in \{1, \ldots, l\}$ に対し $a_i \leq b_i$ かつ, $(n, a_1, \ldots, a_l, m)$ ニューラルネットワークが表現数 N を持つならば, $(n, b_1, \ldots, b_l, m)$ ニューラルネットワークも表現数 N を持つ.

Proof.
任意の $|X| = N$ となる部分集合 $X \subset \mathbb{R}^n$, 関数 $f : \mathbb{R}^n \to \mathbb{R}^m$ に対し, 仮定より, 任意の $\boldsymbol{x} \in X$ に対し, $f(\boldsymbol{x}) = \mathrm{MP}_A(\boldsymbol{x})$ となる $(n, a_1, \ldots, a_{l-1}, m)$ ニューラルネットワーク $A = ((W_1, \boldsymbol{c}_1), \ldots, (W_l, \boldsymbol{c}_l))$ が存在する. このとき $(n, b_1, \ldots, b_{l-1}, m)$ ニューラルネットワーク $B = ((V_1, \boldsymbol{d}_1), \ldots, (V_l, \boldsymbol{d}_l))$ を
$1 < \forall i < l$ に対し, $V_i := \begin{pmatrix} W_i & O \\ O & O \end{pmatrix} \in \mathbb{R}^{b_i \times b_{i-1}}$, $V_1 := \begin{pmatrix} W_1 \\ O \end{pmatrix} \in \mathbb{R}^{b_1 \times n}$, $V_l := \begin{pmatrix} W_l \\ O \end{pmatrix} \in \mathbb{R}^{m \times b_{l-1}}$
$1 \leq \forall i < l$ に対し $\boldsymbol{d}_i := \begin{pmatrix} \boldsymbol{c}_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{b_i}$, $\boldsymbol{d}_l := \boldsymbol{c}_l$ とおけば, 任意の $\boldsymbol{x} \in X$ に対し, $f(\boldsymbol{x}) = \mathrm{MP}_A(\boldsymbol{x}) = \mathrm{MP}_B(\boldsymbol{x})$ を満たすので, $(n, b_1, \ldots, b_{l-1}, m)$ ニューラルネットワークも表現数 N を持つ.

この補題より, 中間層のニューロン数が大きいニューラルネットワークは, 小さいニューラルネットワーク以上の表現数を持つことがわかる. 同様に, 出力のニューロン数に関してはこれとは逆で, 小さいニューラルネットワークは大きいニューラルネットワーク以上の表現数を持つ. すなわち以下の補題が成り立つことが分かる.

Lemma 3 (出力次元の大小と表現数)
$n, m, m', l, N \in \mathbb{N}$, $(a_1, \ldots, a_l) \in \mathbb{N}^l$ とする. このとき $m \leq m'$ かつ $(n, a_1, \ldots, a_l, m')$ ニューラルネットワークが表現数 N を持つならば, $(n, a_1, \ldots, a_l, m)$ ニューラルネットワークも表現数 N を持つ.

Proof.
定義より明らか.

3.2.2 入力次元の独立性

表現数は入力の次元に関して独立である, すなわち表現数は入力次元に依存しないことが以下の定理よりわかる.

Theorem 4 (表現数と入力次元の独立性)
任意の $l, N, n \in \mathbb{N}$, $(a_1, \ldots, a_l) \in \mathbb{N}^l$ に対し, 以下は同値.

- $(1, a_1, \ldots, a_l)$ ニューラルネットワークは表現数 N を持つ.
- $(n, a_1, \ldots, a_l)$ ニューラルネットワークは表現数 N を持つ.

Proof.
(⇐) は定義より明らか. *(⇒)* を示す.
$m := a_l$ とし, 任意の $|X| = N$ となる $X \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^m$ に対し, 補題 *1* より $\mathrm{proj}_{\boldsymbol{w}}$ が X において辞書式順序を保つ, すなわち単射となる $\boldsymbol{w} \in \mathbb{R}^n$ が存在するので, 関数

$$g(y) := \begin{cases} f(\boldsymbol{x}) & (\exists \boldsymbol{x} \in X, y = {}^t\boldsymbol{w}\boldsymbol{x}) \\ (0, \ldots, 0) & (o.w.) \end{cases}$$

が定義できる. 仮定より, データ $({}^t\boldsymbol{w}X, g)$ は $(1, a_1, \ldots, a_l)$ ニューラルネットワークで可解なので, 任意の $y \in {}^t\boldsymbol{w}X$ において $\mathrm{MP}_A(y) = g(y)$ となる $A = ((W_1, \boldsymbol{b}_1), \ldots, (W_l, \boldsymbol{b}_l))$ が存在す

る. そこで $B := ((W_1{}^t\boldsymbol{w}, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2), \ldots, (W_l, \boldsymbol{b}_l))$ とおけば, 任意の $\boldsymbol{x} \in X$ に対し $\mathrm{MP}_B(\boldsymbol{x}) = \mathrm{MP}_A({}^t\boldsymbol{w}\boldsymbol{x}) = g({}^t\boldsymbol{w}\boldsymbol{x}) = f(\boldsymbol{x})$ となるので, データ (X, f) は $(n, a_1, \ldots, a_l)$ ニューラルネットワークで可解. すなわち $(n, a_1, \ldots, a_l)$ ニューラルネットワークは表現数 N を持つ.

この補題により, 学習において, 特に画像認証や音声認識等は入力の次元が非常に大きくなるが, 最大表現数を用いた指標では入力が 1 次元のものに帰着できるので, 証明が簡単になることが予想される.

第 4 章

ReLU ニューラルネットワークにおける表現数

ここから最大表現数の具体的な値を求めていく. しかし, 一般の活性化関数, ニューラルネットワークの型に対して値を求めるのは難しい. そこで本論文では活性化関数を ReLU 関数としたものに限定し, 中間層 1 層のニューラルネットワークに関して具体的な最大表現数を求める.

本章より以降, 活性化関数 $\sigma := \mathrm{ReLU}$ に固定したもののみを考える.

4.1 証明に用いる補題

まず, 最大表現数を求めるのに使用する補題を証明する.

Lemma 5 (データの追加)
$n, k, m \in \mathbb{N}$ を固定する. 任意の有限部分集合 $X \subset \mathbb{R}^n$, 関数 $f : \mathbb{R}^n \to \mathbb{R}^m$ に対し, データ (X, f) が (n, k, m) ニューラルネットワークで可解であるとき, 任意の $\boldsymbol{z} \in \mathbb{R}^n$ に対し, $\forall \boldsymbol{x} \in X, {}^t\boldsymbol{w}\boldsymbol{x} < {}^t\boldsymbol{w}\boldsymbol{z}$ を満たすベクトル $\boldsymbol{w} \in \mathbb{R}^n$ が存在するならば, データ $(X \cup \{\boldsymbol{z}\}, f)$ は $(n, k+1, m)$ ニューラルネットワークで可解.

Proof.
仮定より, (n, k, m) ニューラルネットワーク $((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ が存在し, 任意の $\boldsymbol{x} \in X$ に対し, $f(\boldsymbol{x}) = W_2\sigma(W_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2$ を満たす. ここで, $M := \max\{{}^t\boldsymbol{w}\boldsymbol{x} \mid \boldsymbol{x} \in X\}$ とおき, $(n, k+1, m)$ ニューラルネットワーク $B := ((V_1, \boldsymbol{c}_1), (V_2, \boldsymbol{c}_2))$ を $V_2 := (W_2, \dfrac{f(\boldsymbol{z}) - (W_2\sigma(W_1\boldsymbol{z} + \boldsymbol{b}_1) + \boldsymbol{b}_2)}{{}^t\boldsymbol{w}\boldsymbol{z} - M})$, $\boldsymbol{c}_2 := \boldsymbol{b}_2$, $V_1 := \begin{pmatrix} W_1 \\ \boldsymbol{w} \end{pmatrix}$, $\boldsymbol{c}_1 := \begin{pmatrix} \boldsymbol{b}_1 \\ -M \end{pmatrix}$ とおくと, 任意の $\boldsymbol{x} \in X \cup \{\boldsymbol{z}\}$ に対し, $\mathrm{MP}_B(\boldsymbol{x}) = f(\boldsymbol{x})$ を満たす. よって, データ $(X \cup \{\boldsymbol{z}\}, f)$ は $(n,\, k+1,\, m)$ ニューラルネットワークで可解.

Lemma 6 (線型領域)
$m, M \in \mathbb{N}$ を固定する. $A = ((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ を $(1, M, m)$ ニューラルネットワークとし, $W_1 = \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}$, $\boldsymbol{b}_1 = \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix}$ とする. ここで, $T := \{-\dfrac{b_i}{w_i} \mid i \in \{1, \ldots, M\}, w_i \neq 0\} \subset \mathbb{R}$ とおき, $\{t_1, \ldots, t_{|T|}\} = T$ を昇順に並べたものとすると, 以下を満たす $(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{|T|}), (\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{|T|}) \in (\mathbb{R}^m)^{|T|+1}$ が存在する.

- $\forall x \leq t_1, \mathrm{MP}_A(x) = x\boldsymbol{u}_0 + \boldsymbol{v}_0$
- $1 \leq \forall i < |T|, t_i \leq \forall x \leq t_{i+1}, \mathrm{MP}_A(x) = x\boldsymbol{u}_i + \boldsymbol{v}_i$
- $t_{|T|} \leq \forall x, \mathrm{MP}_A(x) = x\boldsymbol{u}_{|T|} + \boldsymbol{v}_{|T|}$

Proof.
任意の $x \in \mathbb{R}$ に対し,

$\sigma(W_1(x) + \boldsymbol{b}_1) = \begin{pmatrix} \sigma(xw_1 + b_1) \\ \vdots \\ \sigma(xw_M + b_M) \end{pmatrix}$ となるが, $1 \leq i \leq M$ に対し,

$w_i = 0$ のとき, $\sigma(xw_i + b_i) = \sigma(b_i)$ と書ける.

$w_i \neq 0$ のとき, $t := -\dfrac{b_i}{w_i} \in T$ とおくと $\sigma(xw_i + b_i) = \sigma((x-t)w_i)$ と書け,

$x \leq t$ のとき, $\sigma((x-t)w_i) = -x\sigma(-w_i) + t\sigma(-w_i)$,

$t < x$ のとき, $\sigma((x-t)w_i) = x\sigma(w_i) - t\sigma(w_i)$ となる. そこで $u_i^- := -\sigma(-w_i)$, $v_i^- := t\sigma(-w_i)$, $u_i^+ := \sigma(w_i)$, $v_i^+ := -t\sigma(w_i)$ とおくと, 任意の $x \in \mathbb{R}$ に対し,

$$\sigma(xw_i + b_i) = \begin{cases} \sigma(b_i) & (w_i = 0) \\ xu_i^- + v_i^- & (x \leq t) \\ xu_i^+ + v_i^+ & (t < x) \end{cases}$$

と書け, 特に $x = t$ のとき, $xu_i^- + v_i^- = xu_i^+ + v_i^+$ が成り立つ. ここで $1 \leq j < |T|$ に対し,

$$(u_{i,j}, v_{i,j}) := \begin{cases} (u_i^-, v_i^-) & (w_i = 0 \vee t_j < -\dfrac{b_i}{w_i}) \\ (u_i^+, v_i^+) & (o.w.) \end{cases}$$

とすると, $t_j \leq \forall x \leq t_{j+1}$ に対し, $\sigma(xw_i + b_i) = xu_{i,j} + v_{i,j}$ を満たす.

つまり $\boldsymbol{u}'_j := \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{M,j} \end{pmatrix}$, $\boldsymbol{v}'_j := \begin{pmatrix} v_{1,j} \\ \vdots \\ v_{M,j} \end{pmatrix}$ とおけば任意の $1 \leq j < |T|$, $t_j \leq \forall x \leq t_{j+1}$ に対し, $\sigma(W_1 x + \boldsymbol{b}_1) = x\boldsymbol{u}'_j + \boldsymbol{v}'_j$ を満たす. ここで, $\boldsymbol{u}_j := W_2\boldsymbol{u}'_j$, $\boldsymbol{v}_j := W_2\boldsymbol{v}'_j + \boldsymbol{b}_2$ とおけば, $\mathrm{MP}_A(x) = W_2\sigma(W_1 x + \boldsymbol{b}_1) + \boldsymbol{b}_2 = x\boldsymbol{u}_j + \boldsymbol{v}_j$ と書ける. 同様に $\boldsymbol{u}_0 := \begin{pmatrix} u_1^- \\ \vdots \\ u_M^- \end{pmatrix}$, $\boldsymbol{v}_0 := \begin{pmatrix} v_1^- \\ \vdots \\ v_M^- \end{pmatrix}$,

$\boldsymbol{u}_{|T|} := \begin{pmatrix} u_1^+ \\ \vdots \\ u_M^+ \end{pmatrix}$, $\boldsymbol{v}_{|C|} := \begin{pmatrix} v_1^+ \\ \vdots \\ v_M^+ \end{pmatrix}$ とおくと, 任意の $x \leq t_1$, $t_{|T|} \leq y$ に対し, $\mathrm{MP}_A(x) = x\boldsymbol{u}_0 + \boldsymbol{v}_0$, $\mathrm{MP}_A(y) = y\boldsymbol{u}_{|T|} + \boldsymbol{v}_{|T|}$ を満たす.

Lemma 7 (zigzag 関数の可解性)
$k, m \in \mathbb{N}$ とする. 任意の有限部分集合 $X \subset \mathbb{R}$, 任意の関数 $f : \mathbb{R} \to \mathbb{R}^m$ に対し, f が X において *zigzag* かつ, データ (X, f) が $(1, k, m)$ ニューラルネットワークで可解ならば $|X| \leq k + 2$.

Proof.
仮定より, $\forall x \in X, f(x) = \mathrm{MP}_A(x) = W_2\sigma(xW_1 + \boldsymbol{b}_1) + \boldsymbol{b}_2$ を満たす $(1, k, m)$ ニューラルネットワーク $A = ((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ が存在する. また, この A において補題 *6* より $T := \{-\dfrac{b_i}{w_i} \mid i \in \{1, \ldots, k\}, w_i \neq 0\} \subset \mathbb{R}$ とおき, $\{t_1, \ldots, t_{|T|}\} = T$ を昇順に並べたものとおくと,

- $\forall x \leq t_1, \mathrm{MP}_A(x) = x\boldsymbol{u}_0 + \boldsymbol{v}_0$
- $1 \leq \forall i < |T|, t_i \leq \forall x \leq t_{i+1}, \mathrm{MP}_A(x) = x\boldsymbol{u}_i + \boldsymbol{v}_i$
- $c_{|T|} \leq \forall x, \mathrm{MP}_A(x) = x\boldsymbol{u}_{|T|} + \boldsymbol{v}_{|T|}$

を満たす $(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{|T|}), (\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{|T|}) \in (\mathbb{R}^m)^{|T|+1}$ が存在する.

ここで $x_1, \ldots, x_{|X|}$ を X の元を昇順に並べたものとする. このとき, $f(x_{i-1}) \prec f(x_i) \succ f(x_{i+1})$ のとき $x_{i-1} < t_j < x_{i+1}$ かつ $\boldsymbol{u}_j \prec 0$, $f(x_{i-1}) \succ f(x_i) \prec f(x_{i+1})$ のとき $x_{i-1} < t_j < x_{i+1}$ かつ $0 \prec \boldsymbol{u}_j$ を満たす $j \in \{1, \ldots, |T|\}$ が存在することを示す.

$f(x_{i-1}) \prec f(x_i) \succ f(x_{i+1})$ のとき, $x_{i-1} < t_j < x_{i+1}$ かつ $\boldsymbol{u}_j \prec 0$ となる $j \in \{1, \ldots, |T|\}$ が存在しないと仮定し, $x_{i-1} < t_j \leq x_i$ となる $j \in \{1, \ldots, |T|\}$ が存在するときとしないときの両方で矛盾を示す.

$x_{i-1} < t_j \leq x_i$ となる $j \in \{1, \ldots, |T|\}$ が存在するとき, そのうち最も大きいものを t_j とおくと, $x_{i-1} < t_j \leq x_i < t_l < \cdots < t_h < x_{i+1}$ と書けるが,

$$\begin{aligned} f(x_i) &= x_i\boldsymbol{u}_j + \boldsymbol{v}_j \\ &\preceq t_l\boldsymbol{u}_j + \boldsymbol{v}_j = t_l\boldsymbol{u}_l + \boldsymbol{v}_l \\ &\preceq \cdots \preceq t_h\boldsymbol{u}_{h-1} + \boldsymbol{v}_{h-1} = t_h\boldsymbol{u}_h + \boldsymbol{v}_h \\ &\preceq x_{i+1}\boldsymbol{u}_h + \boldsymbol{v}_h = f(x_{i+1}) \end{aligned}$$

より矛盾.

$x_{i-1} < t_j \leq x_i$ を満たす $j \in \{1, \ldots, |T|\}$ が存在しないとき, $x_i < t_l < \cdots < t_h < a_{i+1}$ と書け, $f(x_{i-1}) = x_{i-1}\boldsymbol{u}_{l-1} + \boldsymbol{v}_{l-1}$, $f(x_i) = x_i\boldsymbol{u}_{l-1} + \boldsymbol{v}_{l-1}$ となり, $f(x_{i-1}) \preceq f(x_i)$, $x_{i-1} < x_i$ より $0 \preceq \boldsymbol{u}_{j-1}$ となるので,

$$\begin{aligned} f(x_i) &= x_i\boldsymbol{u}_{l-1} + \boldsymbol{v}_{l-1} \\ &\prec t_l\boldsymbol{u}_{l-1} + \boldsymbol{v}_{l-1} = t_l\boldsymbol{u}_l + \boldsymbol{v}_l \\ &\preceq \cdots \preceq t_h\boldsymbol{u}_{h-1} + \boldsymbol{v}_{h-1} = t_h\boldsymbol{u}_h + \boldsymbol{v}_h \\ &\preceq x_{i+1}\boldsymbol{u}_h + \boldsymbol{v}_h = f(x_{i+1}) \end{aligned}$$

より矛盾.

以上により $x_{i-1} < t_j < x_{i+1}$ かつ $\boldsymbol{u}_j \prec 0$ となる $j \in \{1, \ldots, |T|\}$ が存在する. $f(x_{i-1}) \succ f(x_i) \prec f(x_{i+1})$ のときも同様.

ここで, $1 < i < |X|$ に対し, $j_i \in \{1, \ldots, |T|\}$ を $f(x_{i-1}) \prec f(x_i) \succ f(x_{i+1})$ のとき $x_{i-1} < t_{j_i} < x_{i+1}$ かつ $0 \succ \boldsymbol{u}_{j_i}$, $f(x_{i-1}) \succ f(x_i) \prec f(x_{i+1})$ のとき $x_{i-1} < t_{j_i} < x_{i+1}$ かつ $0 \prec \boldsymbol{u}_{j_i}$ を満たすものとして 1 つ固定すると, $j_2, \ldots, j_{|X|-1}$ は全て異なる. よって, $|X| - 2 = |\{j_i \in T \mid 1 < i < |X|\}| \leq |T|$ で, T の定義より $|T| \leq k$ なので $|X| \leq k + 2$.

4.2 中間層 1 層の場合

まずは最も簡単なニューラルネットワークである中間層 1 層の場合の最大表現数に関して述べる.

4.2.1 最大表現数の上界

Theorem 8 (中間層 1 層のときの最大表現数の上界)
任意の $n, k, m \in \mathbb{N}$ に対し, (n, k, m) ニューラルネットワークの最大表現数は $k + 2$ 以下である.

Proof.
定理 *4* より $n = 1$ のときを示せばよい. $M \in \mathbb{N}$ に対し, $(1, k, m)$ ニューラルネットワークが表現数 M を持つとする. ここで $X := \{1, \ldots, M\}$, $f : \mathbb{R} \to \mathbb{R}^m$ を $f(x) := (\cos \pi x, 0, \ldots, 0)$ とおくと, f は X において *zigzag* となる. すると補題 *7* より $|X| \leq k + 2$ が成り立つ. 今, $|X| = M$ であるので, (n, k, m) ニューラルネットワークの表現数は $k + 2$ 以下である.

4.2.2 最大表現数の下界

Theorem 9 (中間層 1 層のときの最大表現数の下界)
任意の $n, k, m \in \mathbb{N}$ に対し, (n, k, m) ニューラルネットワークは表現数 $k + 1$ を持つ.

Proof.
定理 *4* より $n = 1$ のときを示せばよい. 任意の $|X| = k + 1$ となる $X \subset \mathbb{R}$ と任意の $f : \mathbb{R} \to \mathbb{R}^m$ に対し, $x_1, \ldots, x_{k+1}$ を X の元を昇順に並べたものとする. ここで k に関する帰納法を使う.

$k = 1$ のとき, $(1, 1, m)$ ニューラルネットワーク $A = ((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ を $W_2 := \dfrac{f(x_2) - f(x_1)}{x_2 - x_1}$, $\boldsymbol{b}_2 := f(x_1)$, $W_1 := 1$, $\boldsymbol{b}_1 := -x_1$, とおけば, $\forall x \in X$, $\mathrm{MP}_A(x) = f(x)$ を満たすので, データ (X, f) は $(1, k, m)$ ニューラルネットワークで可解.

$k > 1$ のとき, 帰納法の仮定より, データ $(\{x_1, \ldots, x_k\}, f)$ は $(1, k - 1, m)$ ニューラルネットワークで可解であり, 任意の $x \in \{x_1, \ldots, x_k\}$ に対し, $x < x_{k+1}$ を満たすので, 補題 *5* より (X, f) は $(1, k, m)$ ニューラルネットワークで可解.

以上により (n, k, m) ニューラルネットワークは表現数 $k + 1$ を持つ.

4.2.3 最大表現数の具体的な値が求まる場合

定理 8 と定理 9 により, (n, k, m) ニューラルネットワークのときの最大表現数は $k + 1$ 以上 $k + 2$ 以下, すなわち $k + 1$ か $k + 2$ のどちらかであることが分かる. 以下の定理のように, 特定の場合においては最大表現数が $k + 1$ と $k + 2$ のどちらであるかを決定することができる.

Theorem 10 (中間層 1 層のときの最大表現数)
任意の $n, k, m \in \mathbb{N}$ に対し, 以下が成り立つ.

(1) $k < m$ のとき (n, k, m) ニューラルネットワークの最大表現数は $k + 1$ である.
(2) $k < 3$ のとき (n, k, m) ニューラルネットワークの最大表現数は $k + 1$ である.
(3) $k \geq 3$ のとき $(n, k, 1)$ ニューラルネットワークの最大表現数は $k + 2$ である.

Proof.
定理 *4* より $n = 1$ のときを示せばよい.

(1) $(1, k, m)$ ニューラルネットワークが表現数 $k + 2$ を持たないことを示せばよい.

$X := \{1, \ldots, k + 2\}$, $f : \mathbb{R} \to \mathbb{R}^m$ を $f(x) = (y_1, \ldots, y_m)$ ただし, $y_i = \begin{cases} 2 & (x \geq k + 2) \\ 1 & (i = x < k + 2) \\ 0 & (o.w.) \end{cases}$ と定義すると, $k < m$ より任意の $\boldsymbol{b} \in \mathbb{R}^m$ に対し $\dim(f(X) - \boldsymbol{b}) \geq k + 1$ が成り立つ. ここで f が X において $(1, k, m)$ ニューラルネットワークで可解とすると, $(1, k, m)$ ニューラルネットワーク $A = ((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ が存在し, $\forall x \in X, f(x) = W_2\sigma(W_1 x + \boldsymbol{b}_1) + \boldsymbol{b}_2$ となるが, $W_2 \in \mathbb{R}^{m \times k}$

表 4.1 $-\dfrac{b_1}{w_1} \leq -\dfrac{b_2}{w_2}$ のときの $w_1, w_2, \boldsymbol{v}_1, \boldsymbol{v}_2$ の正負と関数 MP_A の $\preceq$ における傾きの符号

w_1	w_2	$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\cdots$	$-\dfrac{b_1}{w_1}$	$\cdots$	$-\dfrac{b_2}{w_2}$	$\cdots$
$+$	$+$	$+$	$+$	0		$+$		$+$
$+$	$+$	$+$	$-$	0		$+$		$\star$
$+$	$+$	$-$	$+$	0		$-$		$\star$
$+$	$+$	$-$	$-$	0		$-$		$-$
$+$	$-$	$+$	$+$	$-$		$\star$		$+$
$+$	$-$	$+$	$-$	$+$		$+$		$+$
$+$	$-$	$-$	$+$	$-$		$-$		$-$
$+$	$-$	$-$	$-$	$+$		$\star$		$-$
$-$	$+$	$+$	$+$	$-$		0		$+$
$-$	$+$	$+$	$-$	$-$		0		$-$
$-$	$+$	$-$	$+$	$+$		0		$+$
$-$	$+$	$-$	$-$	$+$		0		$-$
$-$	$-$	$+$	$+$	$-$		$-$		0
$-$	$-$	$+$	$-$	$\star$		$+$		0
$-$	$-$	$-$	$+$	$\star$		$-$		0
$-$	$-$	$-$	$-$	$+$		$+$		0

$\star$ は $+$, $-$, 0 いずれかの傾きを取ることを表している.

より $\dim(f(X) - \boldsymbol{b}_2) = \dim(W_2\sigma(W_1X + \boldsymbol{b}_1)) \leq k$ より矛盾する. よって, $k < m$ のとき (n, k, m) ニューラルネットワークは表現数 $k + 2$ を持たない.

(2) $k = 2$ のとき, $(1, 2, m)$ ニューラルネットワークが表現数 4 を持つと仮定する. そこで $X := \{1, \ldots, 4\}$, $f : \mathbb{R} \to \mathbb{R}^m$ を $f(x) := (\cos \pi x, 0, \ldots, 0)$ とおくと, f は X において *zigzag* となるが, 仮定より $(1, 2, m)$ ニューラルネットワーク $((W_1, \boldsymbol{b}_1), (W_2, \boldsymbol{b}_2))$ が存在し, 任意の $x \in X$ に対し, $f(x) = W_2\sigma(xW_1 + \boldsymbol{b}_1) + \boldsymbol{b}_2$ を満たす. つまり $W_2 = (\boldsymbol{v}_2, \boldsymbol{v}_1) \in \mathbb{R}^{m\times 2}$, $W_1 = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \in \mathbb{R}^{2\times 1}$, $\boldsymbol{b}_1 = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathbb{R}^2$ と書いたとすると, 任意の $x \in X$ に対し, $f(x) = \displaystyle\sum_{i\in\{1,2\}} \sigma(xw_i + b_i)\boldsymbol{v}_i + \boldsymbol{b}_2$ と書ける. ここで $w_1 = 0$ または $w_2 = 0$ とすると, 上式が x に対し単調となり, *zigzag* 関数が表せないため $w_1 \neq 0$ かつ $w_2 \neq 0$ である. すると f はそれぞれ w_1, w_2, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $(\dfrac{b_1}{w_1} - \dfrac{b_2}{w_2})$ の正負[*1]2^5 通りの場合分けで, 表 *4.1* のように変曲点を $\{-\dfrac{b_1}{w_1}, -\dfrac{b_2}{w_2}\}$ とする $\preceq$ に関する増減表を記述できる. しかし, $f(1) \prec f(2) \succ f(3) \prec f(4)$ が成り立つためにはどこかに傾きが $+, -, +$ となる区間が存在するはずであるが, 表よりこのような区間は存在しないため矛盾. すなわち $(1, 2, m)$ ニューラルネットワークは表現数 4 を持たない.

$k = 1$ のとき, $(1, 1, m)$ ニューラルネットワークが表現数 3 を持つと仮定すると補題 5 より $(1, 2, m)$ ニューラルネットワークが表現数 4 を持つので矛盾.

[*1] $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ に関しては, 零ベクトルとの辞書式順序 $\preceq$ における比較で正負を定義する.

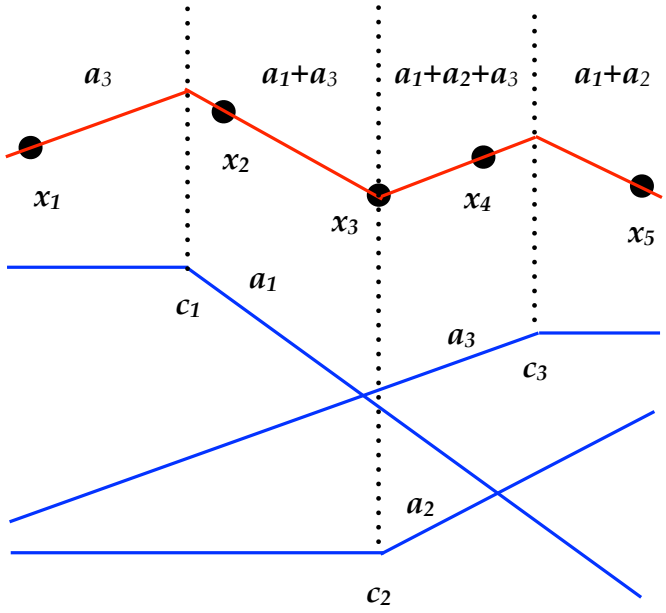


図 4.1 $(1,3,1)$ ニューラルネットワークにおける zigzag 関数の解．下側の 3 本の折れ線は各中間ニューロンの活性化関数を表している．

よって $k<3$ のとき (n,k,m) ニューラルネットワークの最大表現数は $k+1$ である．

(3) $k=3$ のときを示す．$|X|=5$ を満たす任意の $X\subset\mathbb{R}$ と $f:\mathbb{R}\to\mathbb{R}$ に対し，$x_1,\ldots,x_5$ を X を昇順に並べたものとし，$k_1,\ldots,k_4\in\mathbb{R}$ を $k_i:=\dfrac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}$ と定める．

$1<i<5$ で $f(x_{i-1})\le f(x_i)\le f(x_{i+1})$ となるものが存在するとき，$0\le k_{i-1}$, $0\le k_i$ となる．ここで $(1,1,1)$ ニューラルネットワーク $A:=((W_1,b_1),(W_2,b_2))$ を以下のように定める．

$k_{i-1}\le k_i$ のとき，$W_2:=k_i$, $b_2:=f(x_{i-1})$, $W_1:=1$, $b_1:=-(x_i-\dfrac{f(x_i)-f(x_{i-1})}{k_i})$ とする．

$k_{i-1}>k_i$ のとき，$W_2:=-k_{i-1}$, $b_2:=f(x_{i+1})$, $W_1:=-1$, $b_1:=x_i+\dfrac{f(x_{i+1})-f(x_i)}{k_{i-1}}$ とする．

上記のように定めると，いずれの場合も $\mathrm{MP}_A(x_{i-1})=f(x_{i-1})$, $\mathrm{MP}_A(x_i)=f(x_i)$, $\mathrm{MP}_A(x_{i+1})=f(x_{i+1})$ を満たすので，データ $(\{x_{i-1},x_i,x_{i+1}\},f)$ は $(1,1,1)$ ニューラルネットワークで可解となる．したがって，補題 5 よりデータ $(\{x_1,\ldots,x_5\},f)$ は $(1,3,1)$ ニューラルネットワークで可解となる．

$1<i<5$ で $f(x_{i-1})\ge f(x_i)\ge f(x_{i+1})$ となるものが存在するときも同様．

それ以外の場合，f は *zigzag* となる．そこで $f(x_1)<f(x_2)>f(x_3)<f(x_4)>f(x_5)$ のとき，図 *4.1* のように $k_1\le a_3$, $k_2=a_1+a_3$, $k_3=a_1+a_2+a_3$, $k_4\ge a_1+a_2$ を満たすような活性化関数の傾き a_1,a_2,a_3 を見つければよい．[*2] 特に $a_1:=-k_1+k_2-k_3+k_4$, $a_2:=-k_2+k_3$, $a_3:=k_1+k_3-k_4$ と定めると上記の条件を満たす．これらの交点の座標から $c_1:=-\dfrac{f(x_2)-f(x_1)+a_3x_1-k_2x_2}{a_1}$, $c_2:=x_3$, $c_3:=\dfrac{f(x_5)-f(x_4)+k_3x_4+(k_1-k_4)x_5}{a_3}$ と定め，$W_2:=(a_1,a_2,-a_3)$, $b_2:=f(x_1)-a_3x_1+f(x_5)-f(x_4)+k_3x_4+(k_1-k_4)x_5$, $W_1:=\begin{pmatrix}1\\1\\-1\end{pmatrix}$, $\boldsymbol{b}_1:=\begin{pmatrix}-c_1\\-c_2\\c_3\end{pmatrix}$, $A:=((W_1,\boldsymbol{b}_1),(W_2,b_2))$

[*2] 実際には $k_1\le a_3$, $k_2\ge a_1+a_3$, $k_3\le a_1+a_2+a_3$, $k_4\ge a_1+a_2$ かついずれかで等号を満たしていればよい．

と定めると, 任意の $i \in \{1, \ldots, 5\}$ に対し, $\mathrm{MP}_A(x_i) = f(x_i)$ を満たすので, データ (X, f) は $(1, 3, 1)$ ニューラルネットワークで可解.

$f(x_1) > f(x_2) < f(x_3) > f(x_4) < f(x_5)$ のときも同様.

以上より f は X において $(1, 3, 1)$ ニューラルネットワークで可解.

$k > 3$ のとき, $|X| = k + 2$ を満たす任意の $X \subset \mathbb{R}$ と $f : \mathbb{R} \to \mathbb{R}$ に対し, $\{x_1, \ldots, x_{k+2}\}$ を X を辞書式順序で並べたものとすると, データ $(\{x_1, \ldots, x_5\}, f)$ は $(1, 3, 1)$ ニューラルネットワークで可解であるので, 補題 5 より, データ $(\{x_1, \ldots, x_5, x_6, \ldots, x_{k+2}\}, f) = (X, f)$ は $(1, k, 1)$ ニューラルネットワークで可解.

よって $k \geq 3$ のとき $(n, k, 1)$ ニューラルネットワークの最大表現数は $k + 2$ である.

この定理により, (n, k, m) ニューラルネットワークの最大表現数が $k + 1$ であるか $k + 2$ であるかは k と m の値に依存し, そのどちらの値も取りうることがわかる.

4.3 多層ニューラルネットワークの最大表現数の上界

前節では, 中間層 1 層の場合の最大表現数の上界を求めたが, それを多層の場合に一般化する.

Theorem 11 (最大表現数の上界)
$(n, a_1, \ldots, a_l, m)$ ニューラルネットワークの最大表現数は $\left(\prod_{i=1}^{l-1}(a_i + 1)\right)(a_l + 2)$ 以下である.

Proof.
定理 4 より $n = 1$ の場合を示せばよい. $X \subset \mathbb{R}$ を有限集合, $f : \mathbb{R} \to \mathbb{R}$ を X 上 zigzag な関数とし, データ (X, f) が $(1, a_1, \ldots, a_l, m)$ ニューラルネットワークで可解であると仮定する. このとき, $|X| \leq \left(\prod_{i=1}^{l-1}(a_i + 1)\right)(a_l + 2)$ であることを示す.

これを示すために, 以下の命題を l に関する帰納法で示す.

> 任意の $l \in N$, $a_1, \ldots, a_l \in \mathbb{N}$, $Y \subset \mathbb{R}$ に対し, f が Y 上 zigzag であり, データ (Y, f) が $(1, a_1, \ldots, a_l, m)$ ニューラルネットワークで可解ならば, $|Y| \leq \left(\prod_{i=1}^{l-1}(a_i + 1)\right)(a_l + 2)$ が成り立つ.

$l = 1$ のときは補題 7 より明らか.

$l > 1$ のとき, 仮定より, $\forall x \in X, \mathrm{MP}_A(x) = f(x)$ となる $(1, a_1, \ldots, a_l, m)$ ニューラルネットワーク A が存在する. そこで, $A = ((W_1, \boldsymbol{b}_1), \ldots, (W_l, \boldsymbol{b}_l))$ とおき, $B := ((W_2, \boldsymbol{b}_2), \ldots, (W_l, \boldsymbol{b}_l))$ としたとき, $\mathrm{MP}_A(x) = \mathrm{MP}_B(\sigma(W_1 x + \boldsymbol{b}_1))$ と書ける. そこで, $W_1 = \begin{pmatrix} w_0 \\ \vdots \\ w_{a_1 - 1} \end{pmatrix}$, $\boldsymbol{b}_1 = \begin{pmatrix} b_0 \\ \vdots \\ b_{a_1 - 1} \end{pmatrix}$ と書いたとき, $K := \{-\frac{b_i}{w_i} \mid w_i \neq 0, 0 \leq i < a_1\}$ と置き, $k_0, \ldots, k_{|K|-1} \in K$ を K の全ての要素を昇順に並べたものとする. このとき $k_0, \ldots, k_{|K|-1}$ は, ニューラルネットワーク A における線型領域の境界になっている. ここで X の分割 $X_0, \ldots, X_{|K|}$ を

$$X_j := \begin{cases} \{x \in X \mid x \leq k_0\} & (j = 0) \\ \{x \in X \mid k_{j-1} < x \leq k_j\} & (0 < j < |K|) \\ \{x \in X \mid k_{|K|-1} < x\} & (j = |K|) \end{cases}$$

と定義する ($|K| = 0$ のとき, $X_0 := X$ と約束する). 今, $j_0 \in \underset{0\leq j\leq |K|}{\arg\max} |X_j|$, $Y := X_{j_0}$ とおくと, 鳩の巣原理より $|Y| \geq \dfrac{|X|}{|K|+1} \geq \dfrac{|X|}{a_1+1}$ が成り立つ. そこで, $k_{-1} := k_0 - 1$, $k_{|K|} := k_{|K|-1} + 1$ と約束し, $I_{j_0} := \{i \mid w_i k_{j_0-1} + b_i \geq 0 \wedge w_i k_{j_0} + b_i \geq 0\}$ とおくと, 任意の $x \in Y$ に対し, $i \in I_{j_0}$ であることと $w_i x + b_i \geq 0$ であることが同値になる. そこで, $0 \leq i < a_1$ に対し, $w'_i := \begin{cases} w_i & (i \in I_{j_0}) \\ 0 & (o.w.) \end{cases}$, $b'_i := \begin{cases} b_i & (i \in I_{j_0}) \\ 0 & (o.w.) \end{cases}$ と定義すると, $W' := \begin{pmatrix} w'_0 \\ \vdots \\ w'_{a_1-1} \end{pmatrix}$, $\boldsymbol{b}' := \begin{pmatrix} b'_0 \\ \vdots \\ b'_{a_1-1} \end{pmatrix}$ とおいたとき, 任意の $x \in Y$ に対し, $\sigma(W_1 x + \boldsymbol{b}_1) = W'x + \boldsymbol{b}'$ と書ける. したがって $(1, a_2, \ldots, a_l, m)$ ニューラルネットワーク C を $C := ((W_2W', W_2\boldsymbol{b}' + \boldsymbol{b}_2), (W_3, \boldsymbol{b}_3), \ldots, (W_l, \boldsymbol{b}_l))$ と定義すると, $f(x) = \mathrm{MP}_A(x) = \mathrm{MP}_B(\sigma(W_1 x + \boldsymbol{b}_1)) = \mathrm{MP}_B(W'x + \boldsymbol{b}') = \mathrm{MP}_C(x)$ が成り立つ. すなわち, データ (Y, f) は $(1, a_2, \ldots, a_l, m)$ ニューラルネットワークで可解となる. 今, f は Y において zigzag であるので, 帰納法の仮定より $|Y| \leq \left(\prod_{i=2}^{l-1}(a_i+1)\right)(a_l+2)$ が成り立つ. よって, $|X| \leq (a_1+1)|Y| \leq \left(\prod_{i=1}^{l-1}(a_i+1)\right)(a_l+2)$ である.

多層ニューラルネットワークの最大表現数は中間層のうち最終層のニューロン数だけ 2 大きい値で, それ以外の各層のニューロン数に 1 を加えた値の積で上から押さえられる. 特に, $(n, a_1, \ldots, a_l, m)$ReLU ニューラルネットワークの最大表現数は $o(\prod_{i=1}^{l} a_i)$ である.

4.4 中間層 2 層のときの最大表現数の下界

最大表現数の下界も多層ニューラルネットワークへ一般化を行いたいが, 本論文では中間層 2 層までの証明を行い, 3 層以上の場合は将来の課題とする.

まずは出力を 1 次元に限定した場合の最大表現数の下界を示す.

Theorem 12 (中間層 2 層, 出力 1 次元のときの最大表現数の下界)
$(n, a_1, a_2, 1)$ ニューラルネットワークは表現数 $a_1 a_2$ を持つ.

Proof.
定理 4 より $n = 1$ のときを示せばよい.

証明を以下の 3 つのステップに分ける.

> **Step 1**: $a_1 a_2$ 個のデータを昇順に並べ, 小さい順に a_2 個ずつ a_1 個のグループに分ける. これにより, 1 層目のパラメータを定義する.
> **Step 2**: ある連立不等式を解き, 解となる 2 層目, 3 層目のパラメータを見つける.
> **Step 3**: Step 1, Step 2 によって定義されたパラメータが与えられたデータに対する解となっていることを示す.

Step 1 任意の $|X| = a_1 a_2$ を満たす部分集合 $X \subset \mathbb{R}$ と任意の $f : \mathbb{R} \to \mathbb{R}$ に対し, 任意の $x \in X$ に対し, $\mathrm{MP}_A(x) = f(x)$ となる $(1, a_1, a_2, 1)$ ニューラルネットワーク A が存在することを示す.

$0 \leq i < a_1$, $0 \leq j < a_2$ に対し, X の要素 $x_{i,j}$ を $x_{0,0} < x_{0,1} < \cdots < x_{0,a_2-1} < x_{1,0} < x_{1,1} < \cdots < x_{1,a_2-1} < \cdots < x_{a_1-1,0} < x_{a_1-1,1} < \cdots < x_{a_1-1,a_2-1}$ を満たすように並べる. ここで

$b_i := \begin{cases} x_{0,0} - 1 & (i = 0) \\ (x_{i-1,a_2-1} + x_{i,0})/2 & (0 < i < a_1) \\ x_{a_1-1,a_2-1} + 1 & (i = a_1) \end{cases}$ と置くと, $b_0 < x_{0,0} < \cdots < x_{0,a_2-1} < b_1 < x_{1,0} < \cdots < x_{1,a_2-1} < b_2 < \cdots < b_{a_1-1} < x_{a_1-1,0} < \cdots < x_{a_1-1,a_2-1} < b_{a_1}$ を満たし, $b_1, \ldots, b_{a_1-1}$ によって X を a_1 個の部分集合 $\{x_{i,0}, \ldots, x_{i,a_2-1}\}_{0 \le i < a_1}$ に分割する. そこで, 1 層目のパラメータとして, $W_1 := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{a_1 \times 1}$, $\boldsymbol{b}_1 := -\begin{pmatrix} b_0 \\ \vdots \\ b_{a_1-1} \end{pmatrix} \in \mathbb{R}^{a_1}$ と置く. すると, 1 層目の関数 $F_1 : \mathbb{R} \to \mathbb{R}^{a_1}$ は $F_1(x) = \sigma(W_1 x + \boldsymbol{b}_1)$ と書ける. このとき任意の $0 \le i < a_1$, $0 \le j < a_2$ に対し, $F_1(x_{i,j}) = \begin{pmatrix} x_{i,j} - b_0 \\ \vdots \\ x_{i,j} - b_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ と書ける.

Step 2 2 層目, 3 層目のパラメータは, 以下の等式, 不等式を満たす変数 $k_{i,j} \in \mathbb{R}$, $\boldsymbol{w}_j \in \mathbb{R}^{a_1}$, $c_j \in \mathbb{R}$, $C \in \mathbb{R}$ $(0 \le i < a_1, 0 \le j < a_2)$ によって構成する.

(1) $(-1)^i x'_{i,j-1} < (-1)^i k_{i,j} < (-1)^i x'_{i,j}$
(2) ${}^t\boldsymbol{w}_j F_1(k_{i,j}) = c_j$
(3) $\sum_{u=0}^{j} ({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) + C = f(x'_{i,j})$

ただし, $x'_{i,j} := \begin{cases} x_{i,j} & (i \bmod 2 = 0) \\ x_{i,a_2-1-j} & (i \bmod 2 = 1) \end{cases}$,

$(x'_{i,-1}, x'_{i,a_2}) := \begin{cases} (\dfrac{b_i + x'_{i,0}}{2}, \dfrac{x'_{i,a_2-1} + b_{i+1}}{2}) & (i \bmod 2 = 0) \\ (\dfrac{b_{i+1} + x'_{i,0}}{2}, \dfrac{x'_{i,a_2-1} + b_i}{2}) & (i \bmod 2 = 1) \end{cases}$ と約束する.

ここで, (1), (2) は, 1 層目のパラメータで a_1 個に分割した線型領域を, それぞれ図 4.2 のように, 超平面 $\{\boldsymbol{x} \in \mathbb{R}^{a_1} \mid {}^t\boldsymbol{w}_j \boldsymbol{x} = c_j\}$ によって分割を行うために必要な不等式を表している.

(1) – (3) を満たす 1 つの解は以下で与えられる.

- $C := \min_{0 \le i < a_1} f(x'_{i,0}) - 1$.
- 変数 c_j と $k_{i,j}$ は以下のように相互再帰的に定義する. このとき変数を定義する順番は $c_0, k_{0,0}, \ldots, k_{a_1-1,0}, c_1, k_{0,1}, \ldots, k_{a_1-1,1}, c_2, \ldots$ である.
  $0 \le j < a_2$ を固定し, 任意の $0 \le t < j$, $0 \le i < a_1$ に対し, c_t, $k_{i,t}$ が定義されていたとする. このとき, 以下のように $M_{i,j} \in \mathbb{R}$, $B_{i,j}, D_{i,j} : \mathbb{R} \to \mathbb{R}$, $E_{i,j}, E'_{i,j} \in \mathbb{R}$ が定義できる.

$$M_{i,j} := \prod_{s=0}^{i-1} \frac{b_{s+1} - k_{s,j}}{k_{s,j} - b_s}$$

$$B_{i,j}(x) := \frac{(-1)^i (x - k_{i,j})}{k_{i,j} - b_i}$$

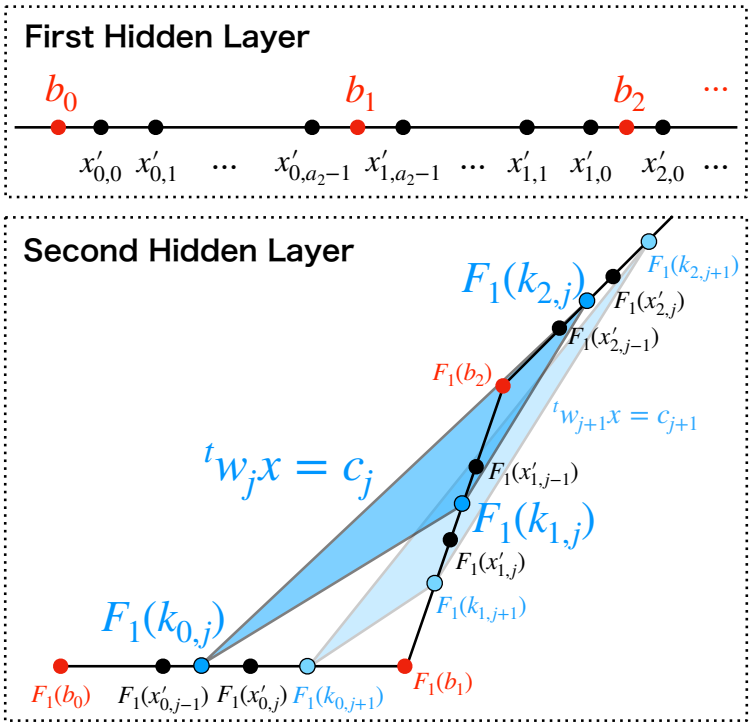


図 4.2 入力データ $x'_{i,j}$ と 1 層目と 2 層目のパラメータとの関係を表した図．1 層目では，$b_0, \dots, b_{a_1}$ によって，$a_1 a_2$ 個の入力データを a_2 個ずつの a_1 個のグループに分割する．2 層目では，1 層目で分割した入力データ $F_1(x'_{i,j})$ を，超平面 $\{\boldsymbol{x} \in \mathbb{R}^{a_1} \mid {}^t\boldsymbol{w}_j \boldsymbol{x} = c_j\}_{0 \le j < a_2}$ によって，それぞれ 1 個ずつに分割する．

$$D_{i,j}(x) := f(x) - C - \sum_{t=0}^{j-1} B_{i,t}(x) M_{i,t} c_t$$

$$E_{i,j} := \frac{\max\{x'_{i,j-1}, x'_{i,j}\} - b_i}{(-1)^i (x'_{i,j} - x'_{i,j-1})} |D_{i,j}(x'_{i,j})|$$

$$E'_{i,j} := \frac{\max\{x'_{i,j-1}, x'_{i,j}\} - b_i}{(-1)^i (x'_{i,j+1} - x'_{i,j})} |D_{i,j}(x'_{i,j+1})|$$

ここで，$K_{i,j} := \max\{|D_{i,j}(x'_{i,j})|, E_{i,j}, E'_{i,j}\} + 1$ とおくと，$K_{i,j} > |D_{i,j}(x'_{i,j})|$, $K_{i,j} > E_{i,j}$, $K_{i,j} > E'_{i,j}$ を満たす．そこで c_j を

$$c_j := (-1)^j \max_{0 \le i < a_1} \left(\prod_{s=0}^{i-1} \frac{\max\{x'_{s,j-1}, x'_{s,j}\} - b_s}{b_{s+1} - \max\{x'_{s,j-1}, x'_{s,j}\}} \right) K_{i,j}$$

と定義し，$k_{i,j}$ を以下のように i に関して再帰的に定義する．

$$k_{i,j} := \frac{(-1)^i x'_{i,j} M_{i,j} c_j + b_i D_{i,j}(x'_{i,j})}{(-1)^i M_{i,j} c_j + D_{i,j}(x'_{i,j})}$$

これを式変形したものは以下のようになり，後の証明で使う．

$$(k_{i,j} - b_i) D_{i,j}(x'_{i,j}) = (-1)^i (x'_{i,j} - k_{i,j}) M_{i,j} c_j \tag{e1}$$

この式で各 c_j, $k_{i,j}$ が c_0, $k_{0,0}$, $\dots$, $k_{a_1-1,0}$, c_1, $k_{0,1}$, $\dots$, $k_{a_1-1,1}$, c_2, $\dots$ の順番で定義できることが図 4.2 より分かる．

表 4.2　各変数に含まれる変数 c_s, $k_{s,t}$ の添字 s, t が満たす条件

	c_t	$k_{s,t}$
$M_{i,j}$	$-$	$s < i \wedge t = j$
$B_{i,j}(x)$	$-$	$s = i \wedge t = j$
$D_{i,j}(x)$	$t < j$	$s = i \wedge t < j$
$E_{i,j}$	$t < j$	$s = i \wedge t < j$
$E'_{i,j}$	$t < j$	$s = i \wedge t < j$
$K_{i,j}$	$t < j$	$s = i \wedge t < j$
c_j	$t < j$	$s < a_1 \wedge t < j$
$k_{i,j}$	$t \leq j$	$(s = i \wedge t < j) \vee (s < i \wedge t = j)$

変数 c_j, $k_{i,j}$ の依存関係を表している.

まず, $t < 0$ を満たす自然数 t が存在しないことから, c_0 が定義できる. c_i が定義できると, $k_{i,0}$ が定義でき, $k_{i,1}$, $k_{i,2}$, …, k_{i,a_2-1} と定義できる. k_{i,a_2-1} まで定義できると, c_{i+1} が定義でき, これを繰り返していけば, 任意の i, j で c_j, $k_{i,j}$ が定義できる.

- 変数 $\boldsymbol{w}_j$ は

$$w_{i,j} := \begin{cases} \dfrac{c_j}{k_{0,j} - b_0} & (i = 0) \\ (-1)^i \dfrac{(k_{i,j} - k_{i-1,j}) M_{i-1,j} c_j}{(k_{i,j} - b_i)(k_{i-1,j} - b_{i-1})} & (i > 0) \end{cases}$$

と置いたとき, $\boldsymbol{w}_j := \begin{pmatrix} w_{0,j} \\ \vdots \\ w_{a_1-1,j} \end{pmatrix}$ と定義する.

ここで定義した変数は以下の命題を満たす (付録参照).

(4) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $\displaystyle\sum_{s=0}^{i} w_{s,j} = (-1)^i \frac{M_{i,j} c_j}{k_{i,j} - b_i}$ を満たす.

(5) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $(-1)^s x'_{s,j-1} < (-1)^s k_{s,j} < (-1)^s x'_{s,j}$ が任意の $0 \leq s < i$ で成り立っているとき, $M_{i,j}(-1)^j c_j \geq K_{i,j}$ を満たす.

(6) 任意の $0 \leq j < a_2$ に対し, もし任意の $0 \leq i < a_1$ に対して $(-1)^j D_{i,j}(x'_{i,j}) > 0$ が成り立っていたとすると, 任意の $0 \leq i < a_1$ に対し, $(-1)^i x'_{i,j-1} < (-1)^i k_{i,j} < (-1)^i x'_{i,j}$ が成り立つ.

(7) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $(-1)^j D_{i,j}(x'_{i,j}) > 0$ を満たす.

先ほど定義した $k_{i,j} \in \mathbb{R}$, $\boldsymbol{w}_j \in \mathbb{R}^{a_1}$, $c_j \in \mathbb{R}$, $C \in \mathbb{R}$ $(0 \leq i < a_1, 0 \leq j < a_2)$ が (1) – (3) を満たすことを示す.

(1) は (6), (7) より直ちに従う.

(2) は i に関する帰納法で示す. ここで $F_1(k_{i,j}) = \begin{pmatrix} k_{i,j} - b_0 \\ \vdots \\ k_{i,j} - b_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ と書けることに注意すると,

$i=0$ のとき, ${}^t\boldsymbol{w}_j F_1(k_{0,j}) = w_{0,j}(k_{0,j}-b_0) = c_j$ より明らか.

$i>0$ のとき, 帰納法の仮定より ${}^t\boldsymbol{w}_j F_1(k_{i-1,j}) = c_j$ が成り立つので,

$$\begin{aligned}
{}^t\boldsymbol{w}_j F_1(k_{i,j}) - c_j &= {}^t\boldsymbol{w}_j F_1(k_{i,j}) - {}^t\boldsymbol{w}_j F_1(k_{i-1,j}) \\
&= \sum_{s=0}^{i} w_{s,j}(k_{i,j}-b_s) - \sum_{s=0}^{i-1} w_{s,j}(k_{i-1,j}-b_s) \\
&= w_{i,j}(k_{i,j}-b_i) + \sum_{s=0}^{i-1} w_{s,j}(k_{i,j}-k_{i-1,j}) \\
&= (-1)^i \frac{k_{i,j}-k_{i-1,j}}{k_{i-1,j}-b_{i-1}} M_{i-1,j}c_j + (k_{i,j}-k_{i-1,j}) \sum_{s=0}^{i-1} w_{s,j} \\
&= (k_{i,j}-k_{i-1,j})\Big(\sum_{s=0}^{i-1} w_{s,j} - (-1)^{i-1}\frac{M_{i-1,j}c_j}{k_{i-1,j}-b_{i-1}}\Big)
\end{aligned}$$

(4) より $\displaystyle\sum_{s=0}^{i-1} w_{s,j} = (-1)^{i-1}\frac{M_{i-1,j}c_j}{k_{i-1,j}-b_{i-1}}$ を満たすので, ${}^t\boldsymbol{w}_j F_1(k_{i,j}) - c_j = 0$ が言える. 以上より, 任意の $0 \le i < a_1$ に対し, ${}^t\boldsymbol{w}_j F_1(k_{i,j}) = c_j$ が成り立つ.

(3) を示す.

$$\begin{aligned}
\sum_{u=0}^{j}({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) + C &= \sum_{u=0}^{j}({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - {}^t\boldsymbol{w}_u F_1(k_{i,u})) + C \\
&= \sum_{u=0}^{j}\sum_{s=0}^{i} w_{s,u}(x'_{i,j}-k_{i,u}) + C \\
&= \sum_{u=0}^{j}(x'_{i,j}-k_{i,u})\sum_{s=0}^{i} w_{s,u} + C \\
&= \sum_{u=0}^{j} B_{i,u}(x'_{i,j})M_{i,u}c_u + C \\
&= B_{i,j}(x'_{i,j})M_{i,j}c_j + \sum_{u=0}^{j-1} B_{i,u}(x'_{i,j})M_{i,u}c_u + C \\
&= \frac{(-1)^i(x'_{i,j}-k_{i,j})}{k_{i,j}-b_i}M_{i,j}c_j - D_{i,j}(x'_{i,j}) + f(x'_{i,j})
\end{aligned}$$

等式 (e1) より $(k_{i,j}-b_i)D_{i,j}(x'_{i,j}) = (-1)^i(x'_{i,j}-k_{i,j})M_{i,j}c_j$ が成り立つので, $\displaystyle\sum_{u=0}^{j}({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) + C = f(x'_{i,j})$ が得られる.

Step 3 Step 2 で定義した変数を用いて, $(1, a_1, a_2, 1)$ ニューラルネットワーク A を定義する.

$$W_2 := {}^t(\boldsymbol{w}_0, -\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, (-1)^{a_2-1}\boldsymbol{w}_{a_2-1}),\ \boldsymbol{c} := \begin{pmatrix} c_0 \\ -c_1 \\ c_2 \\ \vdots \\ (-1)^{a_2-1}c_{a_2-1} \end{pmatrix},\ W_3 := (1, -1, 1, \ldots, (-1)^{a_2-1}),$$

$d = C$ と置き, ニューラルネットワーク $A := ((W_1, \boldsymbol{b}), (W_2, \boldsymbol{c}), (W_3, d))$ と定義する. このとき, 任

意の $x'_{i,j} \in X$ に対し, $\mathrm{MP}_A(x'_{i,j}) = f(x'_{i,j})$ となることを示す.

$$\begin{aligned}\mathrm{MP}_A(x'_{i,j}) &= \sum_{u=0}^{a_2-1}(-1)^u\sigma((-1)^u \cdot {}^t\boldsymbol{w}_u F_1(x'_{i,j}) - (-1)^u c_u) + C \\ &= \sum_{u=0}^{a_2-1}(-1)^u\sigma((-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u)) + C\end{aligned}$$

このとき $(-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) \geq 0$ が $u \leq j$ と同値であることを示す.

$$\begin{aligned}(-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) &= (-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - {}^t\boldsymbol{w}_u F_1(k_{i,u})) \\ &= (-1)^u(\sum_{s=0}^{i} w_{s,u}(x'_{i,j} - k_{i,u})) \\ &= (-1)^u(x'_{i,j} - k_{i,u})\sum_{s=0}^{i} w_{s,u} \\ &= (-1)^i(x'_{i,j} - k_{i,u})\frac{M_{i,u}(-1)^u c_u}{k_{i,u} - b_i}\end{aligned}$$

ここで, $M_{i,u} > 0$, $(-1)^u c_u > 0$, $k_{i,u} - b_i > 0$ であることが分かる. また, (1) より, $(-1)^i x'_{i,-1} < \cdots < (-1)^i x'_{i,u-1} < (-1)^i k_{i,u} < (-1)^i x'_{i,u} < \cdots < (-1)^i x'_{i,a_2-1}$ が成り立つ. よって $u \leq j$ と $(-1)^i k_{i,u} < (-1)^i x_{i,j}$ が同値になる. したがって, $(-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) \geq 0$ と $u \leq j$ は同値になる.

すなわち

$$\begin{aligned}\mathrm{MP}_A(x'_{i,j}) &= \sum_{u=0}^{a_2-1}(-1)^u\sigma((-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u)) + C \\ &= \sum_{u=0}^{j}(-1)^u(-1)^u({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) + C \\ &= \sum_{u=0}^{j}({}^t\boldsymbol{w}_u F_1(x'_{i,j}) - c_u) + C = f(x'_{i,j}) \quad (\because (3))\end{aligned}$$

が言える.

したがって, $(1, a_1, a_2, 1)$ ニューラルネットワークは表現数 $a_1 a_2$ を持つ.

この定理は出力を 1 次元に固定した結果だが, これを任意の出力次元に一般化する.

Theorem 13 (中間層 2 層のときの最大表現数の下界)
(n, a_1, a_2, m) ニューラルネットワークは表現数 $\max\{a_1(a_2 \operatorname{div} m) + a_2 \bmod m, a_2 + 1\}$ を持つ.

Proof.
$(1, a_1, a_2, m)$ ニューラルネットワークが表現数 $a_1(a_2 \operatorname{div} m) + a_2 \bmod m$ と $a_2 + 1$ の両方を持つことを示せばよい.

先に表現数 a_2+1 を持つことを示す. $|X| = a_2+1$ を満たす部分集合 $X \subset \mathbb{R}$ と任意の $f : \mathbb{R} \to \mathbb{R}^m$ を固定する. このとき, $W_1 := \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \in \mathbb{R}^{a_1 \times 1}$, $\boldsymbol{b}_1 := \begin{pmatrix}-\min X\\0\\\vdots\\0\end{pmatrix} \in \mathbb{R}^{a_1}$ and $g(x) := \sigma(W_1 x + \boldsymbol{b}_1)$

とおく．このとき, X において g は単射であるので, 任意の $x \in X$ に対し, $h(g(x)) = x$ となるような $h : \mathbb{R}^{a_1} \to \mathbb{R}$ が存在する．また, 定理 9 より (a_1, a_2, m) ニューラルネットワークは表現数 $a_2 + 1$ を持ち, $|g(X)| = a_2 + 1$ であるので, 任意の $y \in g(X)$ に対し, $\mathrm{MP}_A(y) = (f \circ h)(y)$ となる (a_1, a_2, m) ニューラルネットワーク A が存在する．そこで, $(1, a_1, a_2, m)$ ニューラルネットワーク $B := ((W_1, \boldsymbol{b}_1), A)$ と置くと, 任意の $x \in X$ に対し,

$$\mathrm{MP}_B(x) = \mathrm{MP}_A(\sigma(W_1 x + \boldsymbol{b}_1)) = \mathrm{MP}_A(g(x)) = (f \circ h)(g(x)) = f(x)$$

を満たす．したがって, $(1, a_1, a_2, m)$ ニューラルネットワークは表現数 $a_2 + 1$ を持つ．

$(1, a_1, a_2, m)$ ニューラルネットワークが表現数 $a_1(a_2 \operatorname{div} m) + a_2 \bmod m$ を持つことを示す．$p := a_1(a_2 \operatorname{div} m)$, $q := a_2 \bmod m$ とおき, $|X| = p + q$ を満たす $X \subset \mathbb{R}$ と $f : \mathbb{R} \to \mathbb{R}^m$ を固定する．まず, $x_1, \ldots, x_{p+q}$ を X の要素を昇順に並べたものとし, $X' := \{x_1, \ldots, x_p\}$, $X'' := \{x_{p+1}, \ldots, x_{p+q}\}$ とおく．また, $0 \le i < m$ に対し, $f(x) = {}^t(y_0(x), \ldots, y_{m-1}(x))$ と書いたとき, 関数 $f_i : \mathbb{R} \to \mathbb{R}$ を $f_i(x) := y_i(x)$ と定義する．そこで, 定理 12 の証明において, データ (X, f) に対してそれが可解となるようなニューラルネットワーク A を与えたが, 各 $0 \le i < m$ において, データ (X', f_i) に対しても同じように $(1, a_1, a_2 \operatorname{div} m, 1)$ ニューラルネットワーク A_i を与える．このとき, 任意の $x \in X'$ に対し, $\mathrm{MP}_{A_i}(x) = f_i(x)$ となる．そこで, $A_i = ((W_{1,i}, \boldsymbol{b}_i), (W_{2,i}, \boldsymbol{c}_i), (W_{3,i}, d_i))$ とおくと, 定理 12 の証明の Step1 におけるパラメータの定義より, 1 層目のパラメータ $(W_{1,i}, \boldsymbol{b}_i)$ は f_i に依存しない, すなわち i に依存しない値となることが分かる．よって, 任意の $0 \le i, j < m$ に対し, $(W_{1,i}, \boldsymbol{b}_i) = (W_{1,j}, \boldsymbol{b}_j)$ となる．そこで, $W_1 := W_{1,0}$, $\boldsymbol{b} := \boldsymbol{b}_0$, $W_2 := \begin{pmatrix} W_{2,0} \\ \vdots \\ W_{2,m-1} \end{pmatrix}$, $\boldsymbol{c} := \begin{pmatrix} \boldsymbol{c}_0 \\ \vdots \\ \boldsymbol{c}_{m-1} \end{pmatrix}$, $W_3 := \begin{pmatrix} W_{3,0} & & O \\ & \ddots & \\ O & & W_{3,m-1} \end{pmatrix}$, $\boldsymbol{d} := \begin{pmatrix} d_0 \\ \vdots \\ d_{m-1} \end{pmatrix}$ とおき, $(1, a_1, m(a_2 \operatorname{div} m), m)$ ニューラルネットワーク $B := ((W_1, \boldsymbol{b}), (W_2, \boldsymbol{c}), (W_3, \boldsymbol{d}))$ とおくと, 任意の $x \in X'$ に対し, $\mathrm{MP}_B(x) = \begin{pmatrix} y_0(x) \\ \vdots \\ y_{m-1}(x) \end{pmatrix} = f(x)$ を満たす．

また, $\boldsymbol{b} = -\begin{pmatrix} b_0 \\ \vdots \\ b_{a_1-1} \end{pmatrix}$ とおいたとき, $W_2' := \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{q \times a_1}$, $\boldsymbol{c}' := -\begin{pmatrix} x_p - b_{a_1-1} \\ \vdots \\ x_{p+q-1} - b_{a_1-1} \end{pmatrix} \in \mathbb{R}^q$ とおく．さらに, 任意の $0 \le i < q$ に対し, $\boldsymbol{w}'_{3,i} \in \mathbb{R}^m$ を

$$\boldsymbol{w}'_{3,i} := \frac{f(x_{p+i+1}) - \mathrm{MP}_B(x_{p+i+1}) - \sum_{j=0}^{i-1} (x_{p+i+1} - x_{p+j}) \boldsymbol{w}'_{3,j}}{x_{p+i+1} - x_{p+i}}$$

と i に関して再帰的に定義し, $W_3' := (\boldsymbol{w}'_{3,0}, \ldots, \boldsymbol{w}'_{3,q-1})$ とおく．そこで, $W_2'' := \begin{pmatrix} W_2 \\ W_2' \end{pmatrix}$, $\boldsymbol{c}'' := \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{c}' \end{pmatrix}$, $W_3'' := (W_3\, W_3')$ とおき, $(1, a_1, a_1(a_2 \operatorname{div} m) + a_2 \bmod m, m)$ ニューラルネットワーク

$B' := ((W_1, \boldsymbol{b}), (W_2'', \boldsymbol{c}''), (W_3'', \boldsymbol{d}))$ とおく．ここで $W_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ であることに注意すると

$$\begin{aligned}
\mathrm{MP}_{B'}(x) &= W_3''\sigma(W_2''\sigma(W_1x + \boldsymbol{b}) + \boldsymbol{c}'') + \boldsymbol{d} \\
&= (W_3\, W_3')\sigma\begin{pmatrix} W_2\sigma(W_1x + \boldsymbol{b}) + \boldsymbol{c} \\ W_2'\sigma(W_1x + \boldsymbol{b}) + \boldsymbol{c}' \end{pmatrix} + \boldsymbol{d} \\
&= \mathrm{MP}_B(x) + W_3'\sigma(W_2'\sigma(W_1x + \boldsymbol{b}) + \boldsymbol{c}') \\
&= \mathrm{MP}_B(x) + W_3'\sigma\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\sigma(x - b_{a_1-1}) + \boldsymbol{c}'\right) \\
&= \mathrm{MP}_B(x) + W_3'\sigma\begin{pmatrix} \sigma(x - b_{a_1-1}) - x_p + b_{a_1-1} \\ \vdots \\ \sigma(x - b_{a_1-1}) - x_{p+q-1} + b_{a_1-1} \end{pmatrix}
\end{aligned}$$

と書ける．このとき，$b_{a_1-1} < x_p < \cdots < x_{p+q-1}$ であるので，任意の $0 \le i < q$，$x \in \mathbb{R}$ に対し，$\sigma(x - b_{a_1-1}) - x_{p+i} + b_{a_1-1} > 0$ となることと $x_{p+i} < x$ が同値となる．このとき，任意の $x \in X = X' \cup X''$ において，$\mathrm{MP}'_B(x) = f(x)$ となることを示す．

$x \in X'$ のとき，$x \le x_p$ より $\mathrm{MP}_{B'}(x) = \mathrm{MP}_B(x) = f(x)$ より明らか．

$x \in X''$ のとき，$x = x_{p+i}$ と書けるので，

$$\begin{aligned}
\mathrm{MP}_{B'}(x_{p+i}) &= \mathrm{MP}_B(x_{p+i}) + W_3'\sigma\begin{pmatrix} \sigma(x_{p+i} - b_{a_1-1}) - x_p + b_{a_1-1} \\ \vdots \\ \sigma(x_{p+i} - b_{a_1-1}) - x_{p+q-1} + b_{a_1-1} \end{pmatrix} \\
&= \mathrm{MP}_B(x_{p+i}) + W_3'\begin{pmatrix} x_{p+i} - x_p \\ \vdots \\ x_{p+i} - x_{p+i-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \mathrm{MP}_B(x_{p+i}) + \sum_{j=0}^{i-1}(x_{p+i} - x_{p+j})\boldsymbol{w}'_{3,j}
\end{aligned}$$

となる．このとき，$i > 0$ なので，

$$\begin{aligned}
&\mathrm{MP}_B(x_{p+i}) + \sum_{j=0}^{i-1}(x_{p+i} - x_{p+j})\boldsymbol{w}'_{3,j} \\
&= \mathrm{MP}_B(x_{p+i}) + (x_{p+i} - x_{p+j})\boldsymbol{w}'_{i-1} + \sum_{j=0}^{i-2}(x_{p+i} - x_{p+j})\boldsymbol{w}'_{3,j} \\
&= \mathrm{MP}_B(x_{p+i}) + f(x_{p+i}) - \mathrm{MP}_B(x_{p+i}) - \sum_{j=0}^{i-2}(x_{p+i} - x_{p+j})\boldsymbol{w}'_{3,j} + \sum_{j=0}^{i-2}(x_{p+i} - x_{p+j})\boldsymbol{w}'_{3,j} \\
&= f(x_{p+i})
\end{aligned}$$

よって，$\mathrm{MP}_{B'}(x_{p+i}) = f(x_{p+i})$ となる．したがって，$(1, a_1, a_2, m)$ ニューラルネットワークは表現数 $a_1(a_2 \operatorname{div} m) + a_2 \bmod m$ を持つ．

この定理より, (n, a_1, a_2, m) ニューラルネットワークの最大表現数は $\max\{a_1(a_2 \operatorname{div} m) + a_2 \bmod m, a_2 + 1\}$ 以上であることが分かる. 特に, $a_1 \leq m$ または $a_2 < m$ のとき, $a_1(a_2 \operatorname{div} m) + a_2 \bmod m \leq a_2 + 1$ となる. しかし, 定理 9 より中間層 1 層の (n, a_2, m) ニューラルネットワークも表現数 $a_2 + 1$ を持つため, 中間層 2 層のニューラルネットワークを使うときは, 中間層数は出力次元よりも大きい値が好ましいと予想できる.

また, (n, a_1, a_2, m)ReLU ニューラルネットワークの最大表現数は $O(a_1a_2/m)$ であることが言える. ここで機械学習において, 出力次元 m は定数であることを考えると, 中間層 2 層の最大表現数は $O(a_1a_2)$ であると言える. 今, 最大表現数が $o(a_1a_2)$ であることから, $\Theta(a_1a_2)$ となることが分かる. 特に, $m = 1$ のとき, a_1a_2 と漸近的に等しい.

4.5 まとめ

まとめると表 4.3 のようになる. 上界に関しては, 中間層の最終層以外の各層のニューロン数の積

表 4.3 ReLU ニューラルネットワークの最大表現数の下界, 上界

中間層数	下界	上界
1	$a_1 + 1$	$a_1 + 2$
2	$a_1(a_2 \operatorname{div} m) + a_2 \bmod m$ $a_2 + 1$	$(a_1 + 1)(a_2 + 2)$
$l\,(> 2)$	$\star$	$\left(\prod_{i=1}^{l-1}(a_i + 1)\right)(a_l + 2)$

表は $(n, a_1, \ldots, a_l, m)$ReLU ニューラルネットワークの最大表現数である. 中間層 2 層の下界は, この 2 つのどちらも下界になる.

$\star$: 中間層が 3 層以上の下界は判明していない.

に 1 を足した値の積に最終層のニューロン数に 2 を足した値を掛け合わせた値で押さえられる. すなわち, $(n, a_1, \ldots, a_l, m)$ReLU ニューラルネットワークの最大表現数は $o(\prod_{i=1}^{l} a_i)$ である. 下界に関しては, 中間層 1 層の場合は中間ニューロン数 $+1$ であり, 上界の結果と合わせると, 中間ニューロン数 $+1$ か $+2$ のどちらかとなる. これがどちらであるかというのは表 4.4 のように 1 つに決定できず, 中間ニューロン数, 出力の次元によってどちらの場合もありうる.

表 4.4 (n, a_1, m)ReLU ニューラルネットワークの最大表現数

	$a_1 < \max\{3, m\}$	$\max\{3, m\} \leq a_1$
$m = 1$	$a_1 + 1$	$a_1 + 2$
$1 < m$	$a_1 + 1$	$a_1 + 1$ または $a_1 + 2$

表現数は入力の次元 n に依存しない.

中間層が 2 層の場合の最大表現数は $O(a_1a_2/m)$ となる. ここで, 機械学習を行う際の出力の次元が定数であることを考えると, a_1 と a_2 が m に対して十分に大きい場合, $O(a_1a_2)$ となる. 特に, 出力の次元が 1 のとき, 表 4.5 のようになり, 最大表現数 N は $a_1a_2 \leq N \leq (a_1 + 1)(a_2 + 2)$ を満たすので, 特に $N \sim a_1a_2$ である.

表 4.5 $(n, a_1, a_2, 1)$ReLU ニューラルネットワークの最大表現数の下界, 上界

中間層数	下界	上界
2	$a_1 a_2$	$(a_1+1)(a_2+2)$

出力が 1 次元の場合, 最大表現数は $a_1 a_2$ に漸近的に等しい.

第 5 章

むすび

ニューラルネットワークの型に対する表現能力に関する指標を, 最大いくつの任意のデータが表現可能かによって表現数という名前で定義した. この指標は, 導入部の関連研究で述べた既存の指標と違い, 具体的なデータが与えられた際に, それらが表現可能だという性質を保証することができ, また, 異なる活性化関数による表現能力の比較を行うことができる.

本論文では表現可能性, つまり可解性の定義において, 出力が元のデータと一致するものを可解と考えているが, 実際の学習においては近似解が存在すれば十分であることも多いため, 任意の近似誤差で近似解が存在することを可解の定義とすることも考えられる. しかしながら, 近似解による定義は, 誤差を評価する距離関数ないし二乗誤差関数等の学習における損失関数に依存してしまうという問題点が挙げられる. 損失関数は学習手法やデータの種類によって様々な関数を用いることがあり, 本論文では, 学習手法に依存しないニューラルネットワークの型による表現能力の指標を述べるため, 出力と一致する解が存在することを可解と定義した. なお, この 2 つの定義は同値ではなく, 本論文における可解性を仮定すれば, 任意の近似誤差に対して近似解が存在することが言えるが, その逆は一般に成り立たない[*1].

また, 最大表現数の定義は二値分類問題における学習器の複雑さの指標である VC 次元 [6] の定義に似ている. VC 次元は主に二値分類問題で使用される指標で, ニューラルネットワークにおいては, 入力の部分集合 $X \subset \mathbb{R}^n$ に対し, 任意の $F : \mathbb{R}^n \to \{1, -1\}$, 任意の $\boldsymbol{x} \in X$ で $\mathrm{MP}_A(\boldsymbol{x})$ と $F(\boldsymbol{x})$ の符号が一致するようなニューラルネットワーク A が存在するときの X の濃度の取りうる最大値と定義される [14]. この定義では, 学習における解の存在性を出力の符号と元の関数 F との等しさで考えてはいるが, 最大表現数と同様に入力の部分集合に限定したときに解となる濃度を指標としている. しかし, VC 次元は, 解が存在する部分集合が存在するような濃度の最大値であり, 最大表現数は, 濃度を固定した任意の部分集合が可解となるような濃度の最大値, つまり指標と一致する濃度の部分集合の存在性か任意性かに大きな違いがある. この違いは, VC 次元は主に学習における汎化誤差の上限の評価に用いられ, その証明において最もデータを表現しやすい入力で評価すること, 最大表現数は主に具体的なデータの表現可能性の判定に用いることに由来している. また, 最大表現数と VC 次元の値に関して, 中間層数, 各層のニューロン数の等しいニューラルネットワークにおいて, 最大表現数は VC 次元以下であることがいえる.

また, 本論文では活性化関数が ReLU 関数であるニューラルネットワークにおいて, 最大表現数の上界, 下界を求めた. 特に, 中間層 1 層の場合の最大表現数は, そのニューラルネットワークの中間

[*1] 例えば, 活性化関数を tanh とする (1, 1, 1) ニューラルネットワークにおいて, 誤差関数 $d(x, y) := |x - y|$ としたとき, データ $(\{-1, 1, 2\}, \mathrm{sgn})$ において任意の近似誤差 $\epsilon > 0$ に対し近似解が存在するが, tanh の単射性により可解ではない.

ニューロン数 $+1$ か $+2$ の値であることを解明した. 中間層 2 層の場合は, 各中間層のニューロン数の積に比例して大きくなることが分かり, 特に出力が 1 次元の場合はこの比例定数が 1 であることが解明した. また, 多層の場合の最大表現数は各中間ニューロン数の積のオーダーで押さえられることが判明した. この事実から, 多層ニューラルネットワークの最大表現数も, 各中間ニューロン数の積に比例するのではないかと予想している. この予想は, 線型領域の数などの他のニューラルネットワークの表現能力の指標において, 層数に対して指数的に指標が増加すること [9, 12, 20] からも妥当であると考えることができる.

今後の研究として, 3 つの課題が挙げられる.

1 つ目は, 中間層数が 3 以上の ReLU ニューラルネットワークにおける最大表現数の下界を求めることである. 3 層以上の最大表現数の上界は各中間層のニューロン数の積のオーダーで抑えられることを証明したが, このオーダーのデータ数が表現可能であるか, 下界も同様のオーダーであるかどうかは判明していない. 3 層以上のニューラルネットワークにおける最大表現数のオーダーを求めるためには, 最大表現数の下界を求める必要がある.

2 つ目の課題として, 活性化関数が ReLU 関数以外の場合において, 表現数がどのように変化するのかが挙げられる. 本論文では, 機械学習において広く用いられる活性化関数である ReLU 関数に限定した場合において検証を行なったが, 他の活性化関数の場合において, ReLU 関数の結果と同様な結果が得られるかどうかは未知数である. 同じ型のニューラルネットワークにおける異なる活性化関数を用いた場合の表現数を比較することによって, 活性化関数による表現能力の違いを検証することができると考えられる.

3 つ目は, 表現数と学習との関係について明らかにすることである. 表現数によって, あるデータがあるニューラルネットワークで表現可能であることが判明していたとしても, そのデータを用いた学習が収束するとは限らない. すなわち, 学習によって得られるニューラルネットワークのパラメータの集合と, パラメータ全体の集合とのギャップがあるはずである. このギャップ, すなわち学習によって得られるニューラルネットワークのパラメータの集合がどのようなものであるか, また, このギャップと表現数との関係がどのようになっているかを調べたいと考えている.

謝辞

本論文の作成にあたり, ご指導いただきました山本光晴先生に感謝の意を表します. また, 助言いただきました萩原学先生, 桜井貴文先生, 久我健一先生に感謝の意を表します.

参考文献

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

[2] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 20, No. 1, pp. 23–38, 1998.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97, 2012.

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, Vol. 13, pp. 281–305, 2012.

[5] Razvan Pascanu, Guido Montúfar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations, 2013.

[6] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[7] Igor V. Tetko, David J. Livingstone, and Alexander I. Luik. Neural network studies, 1. comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, Vol. 35, pp. 826–833, 1995.

[8] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2*, NIPS'14, pp. 2924–2932, 2014.

[9] Thiago Serra and Christian Tjandraatmadja. Bounding and counting linear regions of deep neural networks. *International Conference on Learning Representations*, 2018.

[10] Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 2596–2604, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[11] Kevin K. Chen. The upper bound on knots in neural networks, 2016.

[12] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 2847–2854, 2017.

[13] Wolfgang Maass. Neural nets with superlinear VC-dimension. *Neural Computation*, Vol. 6,

No. 5, pp. 877–884, 1994.

[14] Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory*, Vol. 65, pp. 1064–1068, 2017.

[15] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, Vol. 2, No. 5, pp. 359–366, 1989.

[16] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems 30*, pp. 6231–6239, 2017.

[17] David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*, 2018.

[18] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[19] Xingyuan Pan and Vivek Srikumar. Expressiveness of rectifier networks. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, pp. 2427–2435, 2016.

[20] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pp. 3360–3368. Curran Associates, Inc., 2016.

[21] Yoshua Bengio and Olivier Delalleau. On the expressive power of deep architectures. In *International Conference on Algorithmic Learning Theory*, pp. 18–36, 2011.

[22] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958, 2014.

[23] A. Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT '88, pp. 139–154, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[24] Kenta Inoue. Expressive power of neural networks by the number of data that can be expressed. *IEICE*, Vol. J102-D, No. 6, 2019. In Japanese.

[25] Kenta Inoue. Expressive numbers of two or more hidden layer relu neural networks. In *2019 Seventh International Symposium on Computing and Networking Workshops*, pp. 129–135, 2019.

付録 A

本文で省略した証明

本付録では, 4.4 節の定理 12 で省略した証明を記述する.

A.1 定理 12 における命題 (4) – (7) の証明

(4) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $\sum_{s=0}^{i} w_{s,j} = (-1)^i \frac{M_{i,j}c_j}{k_{i,j} - b_i}$ を満たす.

i に関する帰納法で示す.

$i = 0$ のときは明らか.

$i > 0$ のとき, 帰納法の仮定より $\sum_{s=0}^{i-1} w_{s,j} = (-1)^{i-1} \frac{M_{i-1,j}c_j}{k_{i-1,j} - b_{i-1}}$ を満たすので,

$$\begin{aligned}
\sum_{s=0}^{i} w_{s,j} &= w_{i,j} + \sum_{s=0}^{i-1} w_{s,j} \\
&= (-1)^i \left(\frac{(k_{i,j} - k_{i-1,j})M_{i-1,j}c_j}{(k_{i,j} - b_i)(k_{i-1,j} - b_{i-1})} - \frac{M_{i-1,j}c_j}{k_{i-1,j} - b_{i-1}}\right) \\
&= (-1)^i \left(\frac{k_{i,j} - k_{i-1,j}}{k_{i,j} - b_i} - 1\right) \frac{M_{i-1,j}c_j}{k_{i-1,j} - b_{i-1}} \\
&= (-1)^i \frac{M_{i,j}c_j}{k_{i,j} - b_i}
\end{aligned}$$

より, 任意の i に対し, $\sum_{s=0}^{i} w_{s,j} = (-1)^i \frac{M_{i,j}c_j}{k_{i,j} - b_i}$ が言える.

(5) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $(-1)^s x'_{s,j-1} < (-1)^s k_{s,j} < (-1)^s x'_{s,j}$ が任意の $0 \leq s < i$ で成り立っているとき, $M_{i,j}(-1)^j c_j \geq K_{i,j}$ を満たす.

仮定より, $\min\{x'_{s,j-1}, x'_{s,j}\} < k_{s,j} < \max\{x'_{s,j-1}, x'_{s,j}\}$ が任意の $s < i$ で成り立つので, $b_s < k_{s,j} < b_{s+1}$ が言え, かつ $M_{i,j} > 0$ が成り立っている. よって,

$$\begin{aligned}
M_{i,j}(-1)^j c_j &\geq M_{i,j} \left(\prod_{s=0}^{i-1} \frac{\max\{x'_{s,j-1},\, x'_{s,j}\} - b_s}{b_{s+1} - \max\{x'_{s,j-1},\, x'_{s,j}\}} \right) K_{i,j} \\
&\geq M_{i,j} \left(\prod_{s=0}^{i-1} \frac{k_{s,j} - b_s}{b_{s+1} - k_{s,j}} \right) K_{i,j} = K_{i,j}
\end{aligned}$$

が成り立つ.

(6) 任意の $0 \leq j < a_2$ に対し, $(-1)^j D_{i,j}(x'_{i,j}) > 0$ が任意の $0 \leq i < a_1$ で成り立っているとき, 任意の $0 \leq i < a_1$ に対し, $(-1)^i x'_{i,j-1} < (-1)^i k_{i,j} < (-1)^i x'_{i,j}$ が成り立つ.

i に関する完全帰納法を用いる.

まず, $(-1)^i k_{i,j} < (-1)^i x'_{i,j}$ を示す.

$$\begin{aligned}(-1)^i(x'_{i,j} - k_{i,j}) &= (-1)^i(x'_{i,j} - \frac{(-1)^i x'_{i,j} M_{i,j} c_j + b_i D_{i,j}(x'_{i,j})}{(-1)^i M_{i,j} c_j + D_{i,j}(x'_{i,j})}) \\ &= \frac{(-1)^i(x'_{i,j} - b_i) D_{i,j}(x'_{i,j})}{(-1)^i M_{i,j} c_j + D_{i,j}(x'_{i,j})} \\ &= \frac{(x'_{i,j} - b_i)(-1)^j D_{i,j}(x'_{i,j})}{M_{i,j}(-1)^j c_j + (-1)^i (-1)^j D_{i,j}(x'_{i,j})} \\ &= \frac{(x'_{i,j} - b_i)|D_{i,j}(x'_{i,j})|}{M_{i,j}(-1)^j c_j + (-1)^i |D_{i,j}(x'_{i,j})|} > 0\end{aligned}$$

最後の不等号は $x'_{i,j} - b_i > 0$ であることと, (5) と帰納法の仮定より $M_{i,j}(-1)^j c_j \geq K_{i,j}$ が言え, $K_{i,j} > |D_{i,j}(x'_{i,j})|$ を満たすことから示せる.

次に $(-1)^i x'_{i,j-1} < (-1)^i k_{i,j}$ を示す.

(e1) より

$$(k_{i,j} - b_i) D_{i,j}(x'_{i,j}) = (-1)^i (x'_{i,j} - k_{i,j}) M_{i,j} c_j$$

が言えるので, これを変形すると

$$\frac{k_{i,j} - b_i}{(-1)^i(x'_{i,j} - k_{i,j})} = \frac{M_{i,j} c_j}{D_{i,j}(x'_{i,j})}$$

となる. よって

$$\begin{aligned}\frac{k_{i,j} - b_i}{(-1)^i(x'_{i,j} - k_{i,j})} &= \frac{M_{i,j} c_j}{D_{i,j}(x'_{i,j})} \\ &= \frac{M_{i,j}(-1)^j c_j}{(-1)^j D_{i,j}(x'_{i,j})} \\ &\geq \frac{K_{i,j}}{|D_{i,j}(x'_{i,j})|} \\ &> \frac{E_{i,j}}{|D_{i,j}(x'_{i,j})|} = \frac{\max\{x'_{i,j-1}, x'_{i,j}\} - b_i}{(-1)^i(x'_{i,j} - x'_{i,j-1})} \\ &\geq \frac{x'_{i,j-1} - b_i}{(-1)^i(x'_{i,j} - x'_{i,j-1})} \quad (\because (-1)^i(x'_{i,j} - x'_{i,j-1}) > 0)\end{aligned}$$

となる. このとき両辺の分母はどちらも正なので,

$$(-1)^i(x'_{i,j} - x'_{i,j-1})(k_{i,j} - b_i) > (-1)^i(x'_{i,j} - k_{i,j})(x'_{i,j-1} - b_i)$$

が成り立つ. これを変形すると

$$\begin{gathered}(-1)^i k_{i,j}(x'_{i,j} - x'_{i,j-1} + x'_{i,j-1} - b_i) > (-1)^i(x'_{i,j}(x'_{i,j-1} - b_i) + b_i(x'_{i,j} - x'_{i,j-1})) \\ (-1)^i k_{i,j}(x'_{i,j} - b_i) > (-1)^i x'_{i,j-1}(x'_{i,j} - b_i)\end{gathered}$$

となり, $x'_{i,j} - b_i > 0$ より $(-1)^i k_{i,j} > (-1)^i x'_{i,j-1}$ が成り立つ.

(7) 任意の $0 \leq i < a_1$, $0 \leq j < a_2$ に対し, $(-1)^j D_{i,j}(x'_{i,j}) > 0$ を満たす.

j に関する完全帰納法で示す.

$j = 0$ のとき, $(-1)^j D_{i,j}(x'_{i,j}) = f(x'_{i,0}) - C > 0$ より明らか.

$j > 0$ のとき,

$$\begin{aligned}(-1)^j D_{i,j}(x'_{i,j}) &= (-1)^j (f(x'_{i,j}) - C - \sum_{t=0}^{j-1} B_{i,t}(x'_{i,j}) M_{i,t} c_t) \\ &= (-1)^j (D_{i,j-1}(x'_{i,j}) - B_{i,j-1}(x'_{i,j}) M_{i,j-1} c_{j-1}) \\ &= (-1)^j D_{i,j-1}(x'_{i,j}) + B_{i,j-1}(x'_{i,j}) M_{i,j-1} (-1)^{j-1} c_{j-1}\end{aligned}$$

帰納法の仮定と (6) より, $(-1)^i x'_{i,j-2} < (-1)^i k_{i,j-1} < (-1)^i x'_{i,j-1} < (-1)^i x'_{i,j}$ が成り立つので

$$B_{i,j-1}(x'_{i,j}) = \frac{(-1)^i (x'_{i,j} - k_{i,j-1})}{k_{i,j-1} - b_i} > 0$$

が言える. 同様に, 帰納法の仮定と (6), (5) より, $M_{i,j-1}(-1)^{j-1} c_{j-1} \geq K_{i,j-1} > E'_{i,j-1}$ が成り立つので,

$$\begin{aligned}B_{i,j-1}(x'_{i,j}) M_{i,j-1} (-1)^{j-1} c_{j-1} &> B_{i,j-1}(x'_{i,j}) E'_{i,j-1} \\ &= \frac{(-1)^i (x'_{i,j} - k_{i,j-1})}{k_{i,j-1} - b_i} \cdot \frac{\max\{x'_{i,j-2}, x'_{i,j-1}\} - b_i}{(-1)^i (x'_{i,j} - x'_{i,j-1})} |D_{i,j-1}(x'_{i,j})| \\ &= \frac{(-1)^i (x'_{i,j} - k_{i,j-1})}{(-1)^i (x'_{i,j} - x'_{i,j-1})} \cdot \frac{\max\{x'_{i,j-2}, x'_{i,j-1}\} - b_i}{k_{i,j-1} - b_i} |D_{i,j-1}(x'_{i,j})| \\ &> |D_{i,j-1}(x'_{i,j})|\end{aligned}$$

最後の不等式は $(-1)^i k_{i,j-1} < (-1)^i x'_{i,j-1}$ と $k_{i,j-1} < \max\{x'_{i,j-2}, x'_{i,j-1}\}$ から従う. 以上により

$$(-1)^j D_{i,j}(x'_{i,j}) > (-1)^j D_{i,j-1}(x'_{i,j}) + |D_{i,j-1}(x'_{i,j})| \geq 0$$

が言える. したがって, 任意の i, j に対し, $(-1)^j D_{i,j}(x'_{i,j}) > 0$ が成り立つ.