

深層ニューラルネットワークの ダイナミクスに対する実装と数学解析

群馬大学・数理データ科学教育研究センター・川澄亮太*,
一橋大学・大学院経済学研究科・米田剛†

Ryota Kawasumi,
Center for Mathematics and Data Science, Gunma University,
4-2 Aramaki-machi, Maebashi City, Gunma 371-8510, Japan.
Tsuyoshi Yoneda,
Graduate School of Economics, Hitotsubashi University,
2-1 Naka, Kunitachi, 186-8601, Tokyo, Japan.

Mathematics Subject Classification 2020 : 41A29, 68T07

Abstract

本研究では、深層ニューラルネットワーク (DNN) の入力層を空間とした場合におけるその空間内の超曲面を研究対象としている。その研究の有意義性を主張するために、まずは、フーリエ級数と DNN の比較数値実験を行った。フーリエ級数の収束問題においてギブス現象、ピンスキー現象、倉坪現象などが発見され、その根底には不確定性原理が横たわっている。それから事実が超曲面をフーリエ級数で近似する際の本質的な難しさとなっている。一方で、機械学習の歴史的進展を勘案すると、DNN においてはそういった本質的な難しさが回避されているように見受けられる。その回避を深堀りするための契機として、本稿では、超曲面、特に凸領域としての超球面を効率的に近似する DNN (合成関数) を具体的に構成した。

1 はじめに

機械学習とは、人間が自然に行える行動を機械でも行えるようにする研究手法であり、機械学習のひとつに深層ニューラルネットワーク (DNN) がある。DNN とは、生物の神経回路網を模した数理モデルのことであり、その起源は 1943 年にさかのぼる。その年に McCulloch 氏および Pitts 氏は人工ニューロンモデルを提案し、神経細胞 (ニューロン) の動作を数学的・論理的に表現できることを示した [8]。その後、半世紀にわたって様々な研究がなされた。例えば、Rosenblatt 氏によってディープラーニングの原型となるパーセプトロンが提案された [9]。その後、年が空き、1984 年に Hopfield 氏はエネルギー関数を用いた連想記憶モデルを提案し、物理学との関係性について取り上げた [4]。さらに、1986 年に Rumelhart 氏および McClelland 氏は学習アルゴリズムのひとつの誤差逆伝播を提示し、DNN に適用させた [10]。そして近年では、Hinton 氏らによって DNN の効率的な学習手法である、逐次的な事前学習 (DNN の各層ごとに学習) が提示され、さらに手書き数字認識 (MNIST) を使った文字の推定実験も行われた [3]。Krizhevsky 氏らは (畳み込み) ニューラルネットワークを用いて画像認識の性能を向上させた [6]。そして、Silver 氏らを含

*E-mail address: r-kawasumi@gunma-u.ac.jp

†E-mail address: t.yoneda@r.hit-u.ac.jp

む Deep Mind（現在は, Google DeepMind）の研究者らによって人工知能システム（AlphaGo）が開発され、人間のプロ囲碁棋士に勝利したことを報告し、当時多くのメディアに取り上げられた。そして、現在では DNN を用いた人工知能システムは実問題に対して高い精度の出力が得られており、生成 AI である Chat GPT をはじめ、日常にも使われることが多くなった（ただし、Chat GPT には最新のアーキテクチャによるトランスフォーマーモデルが使われている）。このように、DNN は目覚ましい発展を遂げている。

目覚ましい発展の一方で、本研究では原点に立ち返り、微積分計算からそういった DNN の数理的構造を深掘りする。より具体的には、二値分類に焦点を置き、入力層を空間とした場合におけるその空間内の超曲面（二値分類によって生成される超曲面）を研究対象とする。なお、本研究では入力層を空間としており、重みとバイアスを空間とした場合における従来の幾多の先行研究（代表的なものとして例えば [1] の第 5 章を参照されたい）とは視座が違う点にご留意頂きたい。

本稿の構成は以下の通りである。2 章では、DNN とフーリエ級数の本質の違いを浮き彫りにするために、数値計算を用いてその比較検討を行う。3 章では、入力層空間内に生成される超曲面、特に凸領域としての超球面を効率的に近似するための DNN（合成関数）を具体的に構成する。4 章では、3 章で述べられている主結果に対する証明の概略を与える。

2 数値実験によるフーリエ級数と DNN の比較

2 次元ユークリッド空間内に閉曲線を用意し、閉曲線の内部を 1、外部を 0 とラベルを付ける。そのとき、ラベルの 1 の集まりと 0 の集まりで領域分割できるか、という問題を考える（いわゆる、2 値分類問題）。以下、我々が行った数値実験の詳細を述べる：2000 個の点 $x \in [-1.5, 1.5]^2$ をランダムに用意し、点 x が原点を中心に半径 1 の円の内部であれば 1 のラベルを、外部であれば 0 のラベルをばり付けた。このとき、以下の二つの手法で分類を試みた。

- フーリエ級数におけるフーリエ係数を勾配降下法で最適化（最小二乗法）
- DNN における重みとバイアスを勾配降下法で最適化（最小二乗法）

下の図がその数値結果となる。

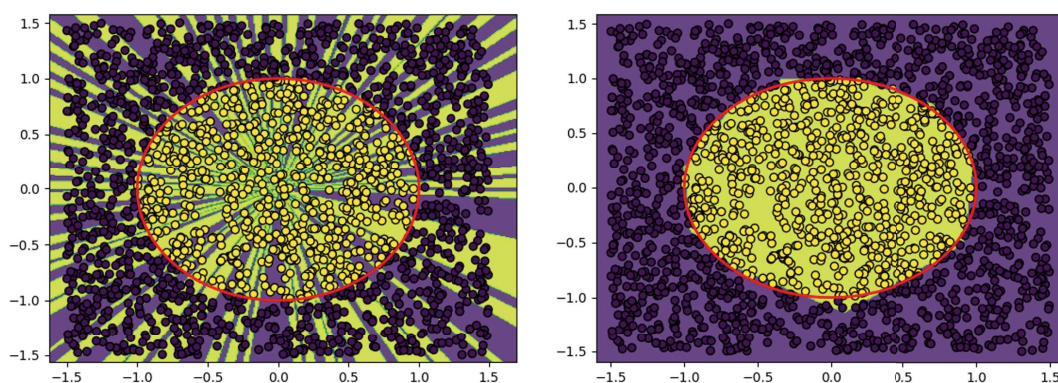


Figure 1: 実行結果（左がフーリエ級数の場合、右が DNN を用いた場合）

右の図、すなわち DNN における勾配降下ではうまく分類を推定しているのに対して、左の図、すなわちフーリエ級数における勾配降下では、分類が放射状に広がってしまい、推定がうまくできてい

ない。しかしながら、フーリエ解析学に立脚すると、この左図の数値結果は至極当然の成り行きであろう。その理由を以下に述べる。上述の数値計算を勘案して、超球 $\Omega := \{x \in [-1, 1]^d : |x| < 1/2\}$ に対して正解関数 f° を

$$f^\circ(x) = \begin{cases} 1 & \text{if } x \in \Omega, \\ 0 & \text{if } x \notin \Omega \end{cases}$$

と定義する（便宜上、数値計算の設定から少しだけスケール変換している）。 f_N を以下のフーリエ部分和とし、 W^t ($t = 0, 1, 2, \dots$) をそのフーリエ係数の集合とする（厳密には、この W^t は N に依存する）：

$$f_N(W^t, x) := \sum_{k \in \mathbb{Z}^d, |k| < N} (a_k^t \cos(k\pi \cdot x) + b_k^t \sin(k\pi \cdot x)),$$

$$W^t := \{a_k^t\}_{|k| < N} \cup \{b_k^t\}_{|k| < N} \subset \mathbb{R}.$$

$W^{t=0}$ は勾配降下に対するフーリエ係数の初期値となる。今、この $f_N(W^t, \cdot)$ と正解関数 $f^\circ(\cdot)$ の誤差関数 E を考える（厳密には、この E 自体も N に依存する）。パーセバルの等式を適用することで以下の等式が得られる：

$$\begin{aligned} E(W^t) &:= \frac{1}{2} \int_{[-1, 1]^d} |f_N(W^t, x) - f^\circ(x)|^2 dx \\ &= \frac{1}{2} \sum_{k \in \mathbb{Z}^d, |k| < N} ((a_k^t - a_k^\infty)^2 + (b_k^t - b_k^\infty)^2) + \frac{1}{2} \sum_{k \in \mathbb{Z}^d, |k| \geq N} ((a_k^\infty)^2 + (b_k^\infty)^2). \end{aligned}$$

ここで a_k^∞, b_k^∞ は正解関数 f° のフーリエ係数である：

$$a_k^\infty := \int_{[-1, 1]^d} f^\circ(x) \cos(k\pi \cdot x) dx \quad \text{and} \quad b_k^\infty := \int_{[-1, 1]^d} f^\circ(x) \sin(k\pi \cdot x) dx.$$

そして、勾配降下法によって以下の収束を得ることができる：任意の $W^{t=0}$ に対して

$$\lim_{t \rightarrow \infty} E(W^t) = \sum_{k \in \mathbb{Z}^d, |k| \geq N} ((a_k^\infty)^2 + (b_k^\infty)^2).$$

実際のところ、勾配降下は

$$a_k^{t+1} = a_k^t - \eta \partial_{a_k^t} E(W^t)$$

と表現される（ b_k^t も同様）。 $\eta \in (0, 1)$ は学習係数と呼ばれる。 $\partial_{a_k^t} E(W^t) = a_k^t - a_k^\infty$ なので

$$(a_k^{t+1} - a_k^\infty) = (1 - \eta)(a_k^t - a_k^\infty)$$

と計算され、このことにより

$$a_k^t - a_k^\infty = (1 - \eta)^t (a_k^0 - a_k^\infty) \rightarrow 0 \quad (t \rightarrow \infty)$$

が得られる。よって $x \in [-1, 1]^d$ に対して、

$$\lim_{t \rightarrow \infty} f_N(W^t, x) = \sum_{k \in \mathbb{Z}^d, |k| < N} (a_k^\infty \cos(k\pi \cdot x) + b_k^\infty \sin(k\pi \cdot x))$$

となることが理解できる。しかしながら、これはフーリエ級数の収束問題に抵触する。例えば近年では以下のことが知られている（倉坪現象と呼ばれている）。

Proposition 2.1 ([7]). $d \geq 5$ とする。そのとき、任意の $x \in \mathbb{Q}^d \cap [-1, 1]^d$ に対して。

$$f_N(W^\infty, x) - f^\circ(x)$$

は $N \rightarrow \infty$ として発散する。

この先行結果を大雑把に翻訳すると、実空間情報（すなわち、入力層空間内の超曲面）とフーリエ空間情報（すなわちフーリエ係数）は、本質的に“**明確な対応関係が持てない**”、ということになる。実のところ、上述の数値計算は2次元なので、倉坪現象の結果を直接適用することができない。しかしながら不確定性原理（[2]の第5章4節を参照されたい）を勘案することによっても、同様の見解が得られることだろう。よって、そういった入力層空間内に生成される超曲面を近似する際、「足し合わせ」ではないやり方を考え出さないといけない、と思ひ至るのは自然の成り行きだと思われる。その考えが経験則として正しい（であろう）ことが、機械学習の研究の歴史が示唆している。次章では、より素朴な微積分計算からそういった示唆を深掘りしたい。

3 超球面を近似する DNN の具体的な構成

ここでは、入力層空間内に生成される超球面をうまく近似する DNN（合成関数）を具体的に構成しよう。なお、一般の初期重みバイアスに対する勾配降下によって、一般の超曲面に収束するか否かの洞察は今後の課題である。まず、 h を ReLU 関数とする。すなわち $h(x) = \max(x, 0)$ とし、 $N = 2^n$ ($n \in \mathbb{Z}_{\geq 0}$) を層の幅とする。そして $2n+1$ を層の深さとする（下で述べる通り、偶奇の層に対して構造を入れるので、このような層の深さとなる）。

第1層：入力データ $x \in [-1, 1]^d$ 、重み行列 $w := \{w_j\}_{j=1}^{2^n} \in \mathbb{R}^{2^n \times d}$ 、バイアス $b := \{b_j\}_{j=1}^{2^n} \in \mathbb{R}^{2^n}$ に対して、第1層 $z^1 := \{z_j^1\}_{j=1}^{2^n} \in \mathbb{R}^{2^n}$ を

$$z^1 := z^1(x) := h(wx + b) := \begin{pmatrix} h(w_1 \cdot x + b_1) \\ \vdots \\ h(w_{2^n} \cdot x + b_{2^n}) \end{pmatrix}$$

と定義する。

偶数層：第 $2k$ 層（偶数層） $z^{2k} = \{z_j^{2k}\}_{j=1}^{2^n} \in \mathbb{R}^{3 \cdot 2^{n-k}}$ を以下のように定義する： $J = 1, 2, \dots, 2^{n-k}$ に対して

$$\begin{aligned} z_{3J-2}^{2k} &= h\left(\frac{1}{2}z_{2J-1}^{2k-1} + \frac{1}{2}z_{2J}^{2k-1}\right), \\ z_{3J-1}^{2k} &= h\left(\frac{1}{2}z_{2J-1}^{2k-1} - \frac{1}{2}z_{2J}^{2k-1}\right), \\ z_{3J}^{2k} &= h\left(-\frac{1}{2}z_{2J-1}^{2k-1} + \frac{1}{2}z_{2J}^{2k-1}\right). \end{aligned} \tag{1}$$

この偶数層の合成関数は、領域を分割する役割を担っている。

奇数層：第 $2k+1$ 層（奇数層） $z^{2k+1} \in \mathbb{R}^{2^{n-k}}$ を以下のように定義する。

$$z_J^{2k+1} = z_{3J-2}^{2k} - z_{3J-1}^{2k} - z_{3J}^{2k}. \tag{2}$$

特に、最終層 $n = k$ のときは、スカラー値 $z^{2n+1} = z_1^{2n+1} \in \mathbb{R}$ となる。

この奇数層は、大雑把に言って、偶数層で分割した領域の取舍選択を行っている。このような層構造により、超曲面の凸形状をうまくとらえることができる。しかし、この種の微積分計算を行ってみるとよく分かるが、凹凸が入り混じった超曲面を近似する DNN（合成関数）を具体的に構成することは案外難しく、今後の課題となる。

このような合成関数を経て得られる超多面体 Ω_N （ファセットに囲まれている）を次のように定義する：

$$\Omega_N := \{x : z^{2n+1}(x) > 0\}.$$

[5] では、このニューラルネットワークに対する勾配降下も計算しているが、ここでは簡単のため、 Ω_N の形状がどうなっているかを問うにとどめておこう。

Theorem 3.1. Ω_N は凸である。

証明の概略を読めばわかる通り、凸超多面体になることの証明はそこまで難しくはない。しかし、こういった類の微積分計算を行ってみるとよく分かるが、凹凸曲面がターゲットとなると、それを近似する DNN（合成関数）を具体的に構成することは想像以上に難しい。なお、和と差を使えば、そういった凹凸超多面体は簡単に構成できると思われる。実際の機械学習の実装において、そういった凹凸超多面体が合成関数構造で生成されているのか、或いは和と差によって生成されているのか、その点を詳細に調べることも今後の課題となる。

4 定理の証明の概略

証明の方針を一言で述べると「誤差逆伝播法に基づいた素直な微積分計算」となる。以下、 x に対する $z = z(x)$ の偏微分を

$$\partial_x z := \frac{\partial z}{\partial x}$$

と略記して、 $\partial_x z_1^{2n+1}$ の計算を行う。 $D_{k,J}^0$ および $D_{k,J}^1$ を以下のように定義する。

$$D_{k,J}^0 := \left\{ x : \frac{1}{2} z_{2J}^{2k-1}(x) \geq \frac{1}{2} z_{2J-1}^{2k-1}(x) \right\}, \quad (3)$$

$$D_{k,J}^1 := \left\{ x : \frac{1}{2} z_{2J}^{2k-1}(x) < \frac{1}{2} z_{2J-1}^{2k-1}(x) \right\}. \quad (4)$$

上の定義 $D_{k,J}^0$ or 1 は、領域を分割・統合する際の基本概念に位置づけられる。この基本概念を出発点として、DNN が持つ領域の分割・統合構造を（微積分計算をもとに）詳細に調べていくこととなる。そのとき、 z_J^{2k+1} の微分を以下のように直接求められる：

$$\begin{cases} \partial_{z_{2J-1}^{2k-1}} z_J^{2k+1} = 1 & \text{if } x \in D_{k,J}^0, \\ \partial_{z_{2J-1}^{2k-1}} z_J^{2k+1} = 0 & \text{if } x \in D_{k,J}^1, \\ \partial_{z_{2J}^{2k-1}} z_J^{2k+1} = 0 & \text{if } x \in D_{k,J}^0, \\ \partial_{z_{2J}^{2k-1}} z_J^{2k+1} = 1 & \text{if } x \in D_{k,J}^1. \end{cases} \quad (5)$$

実際、

$$\begin{aligned} \partial_{z_{2J-1}^{2k-1}} z_J^{2k+1} &= \partial_{z_{2J-1}^{2k-1}} h \left(\frac{1}{2} z_{2J-1}^{2k-1} + \frac{1}{2} z_{2J}^{2k-1} \right) - \partial_{z_{2J-1}^{2k-1}} h \left(\frac{1}{2} z_{2J-1}^{2k-1} - \frac{1}{2} z_{2J}^{2k-1} \right) \\ &\quad - \partial_{z_{2J-1}^{2k-1}} h \left(-\frac{1}{2} z_{2J-1}^{2k-1} + \frac{1}{2} z_{2J}^{2k-1} \right) \\ &= \partial h \left(\frac{1}{2} z_{2J-1}^{2k-1} + \frac{1}{2} z_{2J}^{2k-1} \right) - \partial h \left(\frac{1}{2} z_{2J-1}^{2k-1} - \frac{1}{2} z_{2J}^{2k-1} \right) \\ &\quad + \partial h \left(-\frac{1}{2} z_{2J-1}^{2k-1} + \frac{1}{2} z_{2J}^{2k-1} \right) \\ &= \begin{cases} 1 & \text{if } x \in D_{k,J}^0, \\ 0 & \text{if } x \in D_{k,J}^1. \end{cases} \end{aligned}$$

ただし $\partial h(\cdot)$ を以下のように定義する:

$$\partial h(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

式 (5) によって、 z_1^{2n+1} の微分は以下の通りである:

$$\begin{aligned} \partial_x z_1^{2n+1} &= \partial_{z_1^{2n-1}} z_1^{2n+1} \cdot \partial_x z_1^{2n-1} + \partial_{z_2^{2n-1}} z_1^{2n+1} \cdot \partial_x z_2^{2n-1} \\ &= \begin{cases} \partial_x z_1^{2n-1} & \text{if } x \in D_{n,1}^0, \\ \partial_x z_2^{2n-1} & \text{if } x \in D_{n,1}^1. \end{cases} \end{aligned}$$

繰り返すことで以下が得られる:

$$\begin{aligned} \partial_x z_1^{2n+1} &= \begin{cases} \partial_{z_1^{2n-3}} z_1^{2n-1} \cdot \partial_x z_1^{2n-3} + \partial_{z_2^{2n-3}} z_1^{2n-1} \cdot \partial_x z_2^{2n-3}, & \text{if } x \in D_{n,1}^0, \\ \partial_{z_3^{2n-3}} z_2^{2n-1} \cdot \partial_x z_3^{2n-3} + \partial_{z_4^{2n-3}} z_2^{2n-1} \cdot \partial_x z_4^{2n-3}, & \text{if } x \in D_{n,1}^1, \end{cases} \\ &= \begin{cases} \partial_x z_1^{2n-3} & \text{if } x \in D_{n,1}^0 \cap D_{n-1,1}^0, \\ \partial_x z_2^{2n-3} & \text{if } x \in D_{n,1}^0 \cap D_{n-1,1}^1, \\ \partial_x z_3^{2n-3} & \text{if } x \in D_{n,1}^1 \cap D_{n-1,2}^0, \\ \partial_x z_4^{2n-3} & \text{if } x \in D_{n,1}^1 \cap D_{n-1,2}^1. \end{cases} \end{aligned}$$

上記の操作を何回も繰り返すことで以下の等式が得られる:

$$\partial_x z_1^{2n+1} = w_j, \quad \text{if } x \in D_j.$$

ただし

$$D_j := D_{n,1}^{0 \text{ or } 1} \cap D_{n-1,1}^{0 \text{ or } 1} \cap \dots$$

と定義される。よって、微分積分学の基本定理から次の式が得られる:

$$z_1^{2n+1}(x) = \sum_{j=1}^N (h(w_j \cdot x + b_j) \chi_{D_j}(x)).$$

この D_j の幾何学的形状に着目することで主定理が証明される (この D_j を深掘りする詳細な微積分計算、勾配降下の微積分計算等に興味がある読者は [5] を参照されたい)。

謝辞

発表の機会および会場の提供をして頂いた京都大学に深く感謝いたします。また、研究集会の運営をして頂いた中央大学の澤野嘉宏氏および防衛大学校の瀬戸道夫氏にも感謝いたします。

References

- [1] 甘利俊一, 深層学習と統計神経力学, SGC ライブラリ (サイエンス社) 2023.
- [2] エリアス・M・スタイン, ラミ・シャカルチ (新井仁之, 杉本充, 高木啓行, 千原浩之訳), フーリエ解析入門, 日本評論社, 2007.
- [3] G. E. Hinton, S. Osindero, Simon and Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation, vol. 118, (2006), 157-1544.

- [4] J. Hopfield, The energy relay: a proofreading scheme based on dynamic cooperativity and lacking all characteristic symptoms of kinetic proofreading in DNA replication and protein synthesis, *Proceedings of the National Academy of Sciences*, vol. 77(9), (1980), 5248–5252.
- [5] R. Kawasumi and T. Yoneda, Pointwise convergence of Fourier series and deep neural network for the indicator function of d-dimensional ball, preprint, arXiv:2304.08172v5.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, In *Proceedings of the 25th International Conference on Advances in neural information processing systems*, vol. 25, (2012) 1097–1105.
- [7] S. Kuratsubo, E. Nakai and K. Ootsubo, Generalized Hardy identity and relations to Gibbs-Wilbraham and Pinsky phenomena, *J. Funct. Anal.*, vol. 259, (2010), 315–342.
- [8] W. S. McCulloch, and W. Pitts, A logical calculus of the ideas immanent in nervous activity, vol. 5 (4), (1943), 115–133.
- [9] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, vol. 65 (6), (1958), 386–408.
- [10] D. E. Rumelhart, J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.
- [11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, *nature*, vol. 529, (7587), (2016), 484–489.