論　文

積分表現ニューラルネットとリッジレット変換

Integral Representation of Neural Networks and Ridgelet Transform

園田 翔

Sho Sonoda

1　はじめに

ニューラルネットは近年目覚ましい発展を遂げつつある AI 技術において中心的な役割を担ってきた学習機械である．広義のニューラルネットはニューロンと呼ばれるパラメータ付きの情報処理ユニットが縦横無尽に接続したネットワークの総称であり，その歴史は 1943 年の形式ニューロン [12] まで遡ることができる．中でも，今日の AI 技術で主に用いられているのは，**深層ニューラルネット**である．これはニューロンを並列に接続した**隠れ層**ないし**中間層**と呼ばれる部分ネットワーク構造を，さらに縦列に接続した層状のネットワーク構造をもつ（**図 1**）．深層ニューラルネットのパラメータを調整し，画像認識や機械翻訳といった目的の機能を獲得させる技術を総称して**深層学習**という．これまでに提案された学習法は多様だが，その原型の一つは**最小二乗法**である．一般に，ニューラルネットのパラメータは学習に用いる訓練データのサイズより桁違いに多い．例えば，Google Brain が開発した Switch Transformer [5] のパラメータは 1.6 兆次元にも及ぶ．従って，ニューラルネットのパラメータは冗長性が高く，同等の機能を実現するパラメータが無数に存在する．このため，一般に深層学習を成功させ

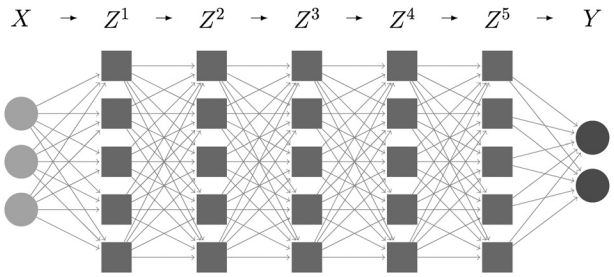

図 1　隠れ層（Z^i）を 5 つ備えた深層ニューラルネット．ただし本稿では隠れ 1 層の浅いニューラルネット（$X \to Z^1 \to Y$）を扱う．

[筆者紹介]

そのだ　しょう．理化学研究所革新知能統合研究センター（理研 AIP）深層学習理論チーム・研究員．JST さきがけ研究員（兼任）．2010 年早稲田大学理工学部電気・情報生命工学科卒業．2012 年同大学大学院修士課程修了後，パナソニックにて車載機器開発に従事，その後退職．2017 年早稲田大学大学院先進理工学研究科電気・情報生命専攻博士後期課程修了．博士（工学）．専門は機械学習の理論と応用．学位論文「深層ニューラルネットの積分表現理論」．学振特別研究員，早稲田大学助手を経て，2018 年より現職．主に深層学習の理論解析に取り組んでいる．趣味は微分から積分まで多岐にわたる．

ることは難しく，また学習済パラメータを解釈したり診断したりすることも難しい．深層ニューラルネットを最小二乗法で訓練するというアイデアは 1980 年代に多層パーセプトロン [19] という名称で提案されていたが，実用に足る深層学習技術が登場したのは 2010 年代になってからのことである．

ニューラルネットに限らず，現代的な学習機械の中身を人間が隅々まで理解したり制御したりすることは今日においても非常に難しい．**深層学習を理解し制御する**問題は，深層学習を通じて得られる典型的な解を特徴付ける問題に帰着される．現実に用いられる深層ニューラルネットは複雑であるため，今日の理論解析においては様々に単純化した数学模型が用いられている．本稿では，1 つの隠れ層の数学模型である**積分表現ニューラルネット**と，その疑似逆作用素（分解作用素）にあたる**リッジレット変換**について解説する．

2 基本設定

ニューラルネットの基本構成要素である**（全結合）ニューロン**とは，活性化関数 $\sigma : \mathbb{R} \to \mathbb{C}$ とパラメータ $(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$ を用いて定義される $\boldsymbol{x} \in \mathbb{R}^m$ の関数である:

$$\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b).$$

全結合のほかには「畳み込み」もあるが，本稿では割愛する．**図 1** において，ニューロンは 1 つの青四角ノードに対応し，ニューロンへの入力 $\boldsymbol{x}$ と出力 $\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ はそれぞれノードに入るエッジとノードから出るエッジに対応する．

本稿で扱う**隠れ 1 層の（全結合）ニューラルネット**とは，以下の関数である:

$$g(\boldsymbol{x}; \theta_p) := \sum_{i=1}^{p} c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i).$$

ただし $\theta_p := (\boldsymbol{a}_i, b_i, c_i)_{i=1}^p \in (\mathbb{R}^m \times \mathbb{R} \times \mathbb{R})^p$ がパラメータである．活性化関数は双曲線正接関数 $\sigma(b) = \tanh b$ や ReLU（rectified linear unit の略）と呼ばれる切断冪関数 $\sigma(b) = \max\{0, b\}$ にとることが多い．

本稿では学習問題として**正則化付き経験二乗誤差最小化問題**を想定する．つまり，有限個の訓練データ $(\boldsymbol{x}_i, y_i)_{i=1}^n \in (\mathbb{R}^m \times \mathbb{R})^n$ に対し，以下のような正則化付き二乗誤差の最小化問題として定式化する:

$$L(\theta_p) := \frac{1}{n} \sum_{i=1}^{n} |y_i - g(\boldsymbol{x}_i; \theta_p)|^2 + \Omega(\theta_p).$$

ただし Ω は適当な正則化項を表す．

■記法 これ以降では，異なる変数に関する Fourier 変換が混在することがあるので，入力 $\boldsymbol{x} \in \mathbb{R}^m$ に関する Fourier 変換を $\widehat{f}(\boldsymbol{\xi}) := \int_{\mathbb{R}^m} f(\boldsymbol{x}) \exp(-i\boldsymbol{x} \cdot \boldsymbol{\xi}) \mathrm{d}\boldsymbol{x}$ のように $\widehat{\cdot}$ で表し，バイアス $b \in \mathbb{R}$ に関する Fourier 変換を $\sigma^\sharp(\omega) := \int_{\mathbb{R}} \sigma(b) \exp(-ib\omega) \mathrm{d}b$ のように $\cdot^\sharp$ で表す．急減少関数を $\mathcal{S}$，緩増加超関数を $\mathcal{S}'$ で表す．$\mathcal{S}'$ は $\mathcal{S}$ の位相的双対空間である．

3 積分表現ニューラルネット

ニューラルネットを扱う上での難しさは，パラメータが非線形関数 σ の中にあることに由来する．この状況を改善する方法の一つは，以下の積分表現のように**パラメータ $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ を無限次元空間のベクトル γ として取り直す**ことである:

定義 1. $\mathcal{V} := \mathbb{R}^m \times \mathbb{R}$ とおく．$\boldsymbol{x} \in \mathbb{R}^m$ を入力とし，$\sigma : \mathbb{R} \to \mathbb{C}$ を**活性化関数**，$\gamma : \mathcal{V} \to \mathbb{C}$ を**パラメータ分布**とする**積分表現ニューラルネット**を以下の形式的な積分で定義する:

$$S[\gamma](\boldsymbol{x}) := \int_{\mathcal{V}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b.$$

混乱を防ぐため，$(\boldsymbol{a}, b, c)$ をパラメータと呼び，γ を**パラメータ分布**と呼んで区別する．積分表現は隠れ 1 層の全結合ニューラルネットの模型である．パラメータ空間 $\mathcal{V}$ 全体にわたる積分は，全てのニューロン $\{\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b) \mid (\boldsymbol{a}, b) \in \mathcal{V}\}$ を足し合わせることを意味する．つまり，積分表現は**連続ニューラルネット（連続モデル）**を表している．また，**図 1** では 1 つの隠れ層の幅を無限に広くすることに対応する．ただし，$\mathcal{V}$ 上の Dirac 測度 δ を用いて特異測度 $\gamma_p := \sum_{i=1}^p c_i \delta_{(\boldsymbol{a}_i, b_i)}$ を作れば，以下のように**有限ニューラルネット（有限モデル）**も表せる:

$$S[\gamma_p](\boldsymbol{x}) = \sum_{i=1}^p c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i).$$

結局，積分表現では連続モデルと有限モデルを同じ形式で表現できるということである．

積分表現にすることで，ニューラルネットは**線形化**される．つまり，ベクトル $(\boldsymbol{a}, b, c)$ に代わって関数 γ をパラメータとみなすと，パラメータとニューラルネットとの対応関係 $\gamma \mapsto S[\gamma]$ は次の意味で線形になる: 任意の 2 つのパラメータ分布 γ, γ' とスカラー $\alpha, \alpha' \in \mathbb{R}$ に対し，

$$S[\alpha\gamma + \alpha'\gamma'] = \alpha S[\gamma] + \alpha' S[\gamma'].$$

そしてこれに伴い，学習問題は**凸化**される．すなわち，凸関数 $\ell : \mathbb{R} \to \mathbb{R}$ が定める損失関数を $L[\gamma] = \ell(S[\gamma])$ として，任意の $t \in [0, 1]$ に対し，以下が成り立つ:

$$L[t\gamma + (1-t)\gamma'] \leqslant tL[\gamma] + (1-t)L[\gamma'].$$

このように，有限次元パラメータ $(\boldsymbol{a}, b, c)$ をうまく無限次元パラメータ γ に持ち上げることで，非線形・非凸問題を線形・凸問題に変えられる．

積分表現はニューラルネットの表現能力を調べる関数近似理論の文脈で度々採用されてきた方法である [1][2][14][18]．特に線形化・凸化の仕組みは近年の平均場理論（学習過程に対する大数の法則）[3][13][15][17][22] や ReLU ネットのレプレゼンター定理 [16][20][29] などで利用されている．

4 リッジレット変換

積分表現を導入する最も積極的な動機は**リッジレット変換（ridgelet transform）**である．リッジレット変換は積分表現作用素 S の右逆作用素である．一般に，与えられた作用素の逆が陽に書けることは稀だが，リッジレット変換は閉形式で与えられることが 1990 年代に発見された [2][11][14][18][28]．まず本節でリッジレット変換を天下り式に与えて概要を説明し，次節以降でリッジレット変換の導出について説明する．

定義 2. 各点 $(\boldsymbol{a}, b) \in \mathcal{V}$ において，関数 $f : \mathbb{R}^m \to \mathbb{C}$ のリッジレット関数 $\rho : \mathbb{R} \to \mathbb{C}$ に関する**リッジレット変換**を以下で定義する:

$$R[f; \rho](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}.$$

ここで，リッジレット関数 ρ は活性化関数 σ と独立に選べる．

定義 3. $\sigma, \rho : \mathbb{R} \to \mathbb{C}$ の *Fourier* 変換をそれぞれ $\sigma^\sharp, \rho^\sharp$ と書く．入力次元を $m \in \mathbb{N}$ として，σ, ρ のスカラー積を次のように定義する*:*

$$(\!(\sigma, \rho)\!) := (2\pi)^{m-1} \int_{\mathbb{R}} \frac{\overline{\rho^\sharp(\omega)}\sigma^\sharp(\omega)}{|\omega|^m} \mathrm{d}\omega.$$

これは周波数領域において $|\omega|^{-m}$ で重み付けられた L^2 内積を定める．$(\!(\cdot, \cdot)\!)$ を内積とする Hilbert 空間を $L^2_m(\mathbb{R})$ と書く．

特に，$(\!(\sigma, \rho)\!) \neq 0, \infty$ のとき，σ と ρ は許容的（admissible）であるという．

定理 1（再構成公式, [28], Theorem 5.6）． $\sigma \in \mathcal{S}'(\mathbb{R})$（緩増加超関数），$\rho \in \mathcal{S}(\mathbb{R})$（急減少関数），関数 $f \in L^1(\mathbb{R}^m) \cap L^2(\mathbb{R}^m)$ に対し，以下が成り立つ：

$$S[R[f; \rho]] = (\!(\sigma, \rho)\!) f.$$

ただし等号はノルム収束および概収束の意味で成り立ち，特に f の任意の連続点 x_c で各点収束する．

緩増加超関数（$\mathcal{S}'$）は Gaussian $\sigma(b) = \exp(-|b|^2/2)$ やステップ関数 $\sigma(b) = b^0_+$，シグモイド関数 $\sigma(b) = \tanh(b)$，ReLU $\sigma(b) = \max\{b, 0\} = b_+$ など，考えうるほとんど全ての活性化関数を含む広いクラスである．

σ と ρ が許容的であるとき，$(\!(\sigma, \rho)\!) = 1$ となるように ρ を正規化できる．従って，再構成公式は（与えられた σ に対して許容的な ρ による）リッジレット変換 $R[\,\cdot\,; \rho]$ が積分表現作用素 S の**右逆作用素**であることを示している．

再構成公式はまた，連続ニューラルネットの**構成的な普遍近似定理**とも解釈できる．任意の $f \in L^1(\mathbb{R}^m) \cap L^2(\mathbb{R}^m)$ に対し，パラメータ分布を $\gamma_f := R[f]$ とおけば $S[\gamma_f] = f$ とでき，これは連続ニューラルネット $S[\gamma_f]$ が関数 f を表現していると解釈できるためである．

リッジレット変換は発見的に得られたため，オリジナルの論文 [2][14][18] を読んだ読者は以下のような疑問を抱くだろう:

Q1. リッジレット変換の必要性．つまり，$S[\gamma] = f$ を満たす γ は常にリッジレット変換で与えられるか

Q2. リッジレット関数 ρ の自然な選び方は何か

Q3. 多様なニューラルネットに対してリッジレット変換を導出できるか

これらの疑問については，筆者らの一連の研究 [23][24][25][26][27] により近年大幅に理解が進んだ．次節以降では，その研究結果のさわりを紹介する．

5 積分方程式 $S[\gamma] = f$ の解き方

関数 $f : \mathbb{R}^m \to \mathbb{C}$ が与えられたとして，パラメータ分布 $\gamma : \mathcal{V} \to \mathbb{C}$ を未知関数とする積分方程式を考える:

$$S[\gamma] = f. \tag{1}$$

データ $y = f(\boldsymbol{x})$ を生成する関数 $f : \mathbb{R}^m \to \mathbb{C}$ の情報が全て与えられているため，この方程式は学習問題の原始形とみなせる．明らかに，リッジレット変換 $\gamma = R[f; \rho]$ は特殊解である．以下では [23] に

沿って $S[\gamma] = f$ の解き方を説明し，その過程でリッジレット変換が自然に導出されることを見る．

簡単のため，積分や級数の順序を自由に交換する．これらの操作は S の有界性（連続性）から保証される．順序交換の正当化，Fourier 表示や 補題 1（後述）の証明は [23] を参照されたい．

■ **Step 1: Fourier 表示**　まず，積分表現を Fourier 表示にする．

$$\begin{aligned} S[\gamma](\boldsymbol{x}) &= \int_{\mathbb{R}^m} [\gamma(\boldsymbol{a},\cdot) *_b \sigma](\boldsymbol{a}\cdot\boldsymbol{x})\mathrm{d}\boldsymbol{a} \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^m\times\mathbb{R}} \gamma^\sharp(\boldsymbol{a},\omega)\sigma^\sharp(\omega)e^{i\omega\boldsymbol{a}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{a}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left[\int_{\mathbb{R}^m}\gamma^\sharp(\boldsymbol{\xi}/\omega,\omega)e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}\right]\sigma^\sharp(\omega)|\omega|^{-m}\mathrm{d}\omega. \end{aligned} \tag{2}$$

ここで，第 2 式の $*_b$ は b に関する畳み込み積分を表す．第 3 式は関数 ϕ に対する恒等式（Fourier 反転公式）$\phi(b) = \frac{1}{2\pi}\int_{\mathbb{R}}\phi^\sharp(\omega)e^{i\omega b}\mathrm{d}\omega\ (\forall b\in\mathbb{R})$ において $\phi(b) = [\gamma(\boldsymbol{a},)*_b\sigma](b)$, $b = \boldsymbol{a}\cdot\boldsymbol{x}$ として得られる．第 4 式は変換変換 $(\boldsymbol{a},\omega) = (\boldsymbol{\xi}/\omega,\omega)$, $\mathrm{d}\boldsymbol{a}\mathrm{d}\omega = |\omega|^{-m}\mathrm{d}\boldsymbol{\xi}\mathrm{d}\omega$ から得られる．Fourier 表示により，枠 $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)\mathrm{d}\boldsymbol{a}\mathrm{d}b$ が $\exp(i\boldsymbol{\xi}\cdot\boldsymbol{x})\mathrm{d}\boldsymbol{\xi}\otimes\sigma^\sharp(\omega)|\omega|^{-m}\mathrm{d}\omega$ に変換された．

■ **Step 2: 特殊解**　Fourier 表示 (2) の $[\cdots]$ の中身は $\boldsymbol{\xi}$ に関する Fourier 逆変換であることに注意する．このことから，γ として次のような変数分離形を仮定する:

$$\gamma^\sharp_{f,\rho}(\boldsymbol{\xi}/\omega,\omega) := \widehat{f}(\boldsymbol{\xi})\overline{\rho^\sharp(\omega)}. \tag{3}$$

ここで f は (1) で与えられた関数であり，ρ は任意の関数である．このとき，$\gamma_{f,\rho}$ は特殊解である．実際，

$$\begin{aligned} S[\gamma](\boldsymbol{x}) &= \left[(2\pi)^{m-1}\int_{\mathbb{R}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega\right]\left[\frac{1}{(2\pi)^m}\int_{\mathbb{R}^m}\widehat{f}(\boldsymbol{\xi})e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}\right] \\ &= (\!(\sigma,\rho)\!)f(\boldsymbol{x}). \end{aligned} \tag{4}$$

すなわち，$(\!(\sigma,\rho)\!)\neq 0$ のとき ρ を適切に正規化することで，変数分離形 (3) は積分方程式 (1) の特殊解である．

■ **（補足）リッジレット変換の導出**　実は，変数分離形 (3) の正体はリッジレット変換の Fourier 表示である．特殊解 (3) を γ に関して解いてみよう．まず変数変換により，$\gamma^\sharp_{f,\rho}(\boldsymbol{a},\omega) = \widehat{f}(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}$ である．これの逆 Fourier 変換を計算するとリッジレット変換になることが分かる:

$$\begin{aligned} \gamma_{f,\rho}(\boldsymbol{a},b) &= \frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)e^{-i\omega b}}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^m\times\mathbb{R}} f(\boldsymbol{x})\overline{\rho^\sharp(\omega)e^{i\omega(\boldsymbol{a}\cdot\boldsymbol{x}-b)}}\mathrm{d}\omega \\ &= \int_{\mathbb{R}^m\times\mathbb{R}} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a}\cdot\boldsymbol{x}-b)}\mathrm{d}\boldsymbol{x}. \end{aligned}$$

■ **Step 3: 一般解**　一般解は特殊解と零空間 $\ker S = \{\gamma\in\mathcal{G}\mid S[\gamma]=0\}$ の和で与えられるので，以下では零空間の表示を計算する．

変数分離形 (3) と再構成公式 (4) の関係を見比べると，$(\!(\sigma,\rho_0)\!) = 0$ となる $\rho_0\in L^2_m(\mathbb{R})$ を用いて変数分離形を作ると，f によらず $S[\gamma_{f,\rho_0}] = (\!(\sigma,\rho_0)\!)f = 0$ となることが分かる．結局，以下の補題から一般解はこのように非自明な零成分の無限和で書けることが示せる．

補題 1（[23], Lemma 11）． $\{e_i\}_{i\in\mathbb{N}}$ と $\{\rho_j\}_{j\in\mathbb{N}}$ はそれぞれ $L^2(\mathbb{R}^m)$ と $L^2_m(\mathbb{R})$ の正規直交系とする．このとき，任意の $\gamma \in L^2(\mathbb{R}^m \times \mathbb{R})$ は以下のように一意に展開できる*:*

$$\gamma = \sum_{ij} \langle \gamma, R[e_i, \rho_j] \rangle_{L^2(\mathbb{R}^m \times \mathbb{R})} R[e_i; \rho_j].$$

補題 1 において基底の取り方は自由なので，$L^2_m(\mathbb{R})$ の正規直交基底 $\{\rho_j\}_{j\in\mathbb{N}\cup\{*\}}$ として

$$(\!(\sigma, \rho_*)\!) = 1; \quad (\!(\sigma, \rho_j)\!) = 0 \ (j \in \mathbb{N}),$$

となるものをとる．このとき，一般解は $(c_{ij}) \in \ell^2(\mathbb{N}^2)$ を任意径数として次のように書ける:

$$\gamma = R[f; \rho_*] + \sum_{ij} c_{ij} R[e_i; \rho_j]. \tag{5}$$

実際，作り方から，

$$S[\gamma] = (\!(\sigma, \rho_*)\!) f + \sum_{ij} c_{ij} (\!(\sigma, \rho_j)\!) e_i = 1 \cdot f + \sum_{ij} 0 \cdot e_i = f$$

となり，(5) は解であることが分かる（十分性）．一方，解がこれで尽くされていることは 補題 1 から従う（必要性）．

■（補足）自然な ρ の選び方 $(\!(\sigma, \rho_*)\!) = 1$ を満たす $\rho_* \in L^2_m(\mathbb{R})$ は無数に存在するので，(5) の第 1 項 $R[f; \rho_*]$ には任意性がある．自然な代表元の選び方として，共役作用素 S^* が取れる．

以下の補題に現れる $\mathcal{A} \subset \mathcal{S}'(\mathbb{R}), \mathcal{G} \subset L^2(\mathbb{R}^m \times \mathbb{R})$ はそれぞれ活性化関数 σ とパラメータ分布 γ が属するある Hilbert 空間（重み付き Sobolev 空間）である．紙面の都合上，詳細は [23] を参照されたい．

補題 2（[23], Lemma 9）． 与えられた活性化関数 $\sigma \in \mathcal{A}$ に対し，共役作用素 $S^* : L^2(\mathbb{R}^m) \to \mathcal{G}$ はある $\sigma_* \in L^2_m(\mathbb{R})$ を用いて $S^*[f] = R[f; \sigma_*]$ と書ける．さらに，任意の $f \in L^2(\mathbb{R}^m)$ に対して次が成り立つ：（再構成公式）$SS^*[f] = \|\sigma\|_{\mathcal{A}} f$，（*Plancherel*）$\|\sigma\|_{\mathcal{A}} \|f\|_{L^2(\mathbb{R}^m)} = \|S^*[f]\|_{\mathcal{G}}$，（直交射影）$S^*S = \mathsf{proj}_{\mathcal{G} \to (\ker \mathsf{S})^\perp}$.

つまり，共役作用素は非自明な零成分を含まないリッジレット変換である．特に，その像は零空間に（$\mathcal{G}$ の内積で）直交している．零成分を含まないパラメータ分布は L^2 正則化付き学習問題の解として自然に現れることが知られている [27][24]．以上の議論をまとめて以下を得る．

定理 2（[23], Theorem 10）． $f \in L^2(\mathbb{R}^m), \sigma \in \mathcal{S}'(\mathbb{R}), \gamma \in L^2(\mathbb{R}^m \times \mathbb{R})$ に対し，積分方程式 $S[\gamma] = f$ の一般解は以下の形式で一意に表示される*:*

$$\gamma = S^*[f] \oplus_{\mathcal{G}} \sum_{ij} c_{ij} R[e_i; \rho_j].$$

6 零空間の役割

前節の議論により, ニューラルネットの零空間の構造が明らかとなった．この零空間は仕事をするのだろうか．直交基底 $\{e_i\}_{i\in\mathbb{N}}$ の代わりに $L^2(\mathbb{R}^m)$ の任意の列 $F := \{f_i\}_{i\in\mathbb{N}}$ を用いて $\gamma_F := \sum_{i=1}^{\infty} R[f_i; \rho_i]$ とおくと，$\gamma_F \in \ker S$ となり，実空間には影響を及ぼさないように思われる．実は，各 $i \in \mathbb{N}$ に対し，$(\!(\sigma_i, \rho_i)\!) = 1$ となる σ_i を活性化関数とするニューラルネットを S_i として，枠変換を $A_i := S^* S_i$ とおけば，$S[A_i[\gamma_F]] = f_i$ として零空間の情報 f_i が読み出せる．このように，ニューラルネットは

1 つの関数 f だけではなく関数列 $\{f_i\}_{i\in\mathbb{N}}$ を記憶する能力がある．このことはパラメータ分布の空間 $L^2(\mathbb{R}^m\times\mathbb{R})$ がテンソル積 $L^2(\mathbb{R}^m)\otimes L^2(\mathbb{R})$ と同型であり，さらに $L^2(\mathbb{R})$ が数列空間 $\ell^2(\mathbb{N})$ と同型であり，従って $L^2(\mathbb{R}^m\times\mathbb{R})$ は $L^2(\mathbb{R}^m)$ の関数列の空間 $\ell^2(\mathbb{N};L^2(\mathbb{R}^m))$ と同型である事実とも整合している．例えば，勾配法による学習過程のように，パラメータ分布 γ が直接変換を受けるような場合には，零空間に保持された情報が実空間に影響を及ぼしうることが分かる．

[23] ではさらに，（連続モデルではなく）$\gamma_p:=\sum_{k=1}^p\gamma(\boldsymbol{a}_k,b_k)\delta_{(\boldsymbol{a}_k,b_k)}$ のような有限モデルも $\mathcal{G}$ に埋め込むことで同様の構造をもつことを示した．また，深層 ReLU ネットを対象とするノルムベースの汎化誤差評価において，従来の議論を慎重に見直すことで，汎化誤差評価は零成分に依存しないことを示した．そして，怠惰学習（lazy learning）のように零成分を許す学習法は，ノルム正則化解のように零成分を許さない学習法よりも速く解に到達することを指摘した．

7 多様体上のニューラルネット

5 節で示した Fourier 表示の方法を応用すると，多様体上の全結合ネット [25] や信号空間上の群畳み込みネット [26] に対してもリッジレット変換を導出できる．本稿の残りの部分では，双曲空間上の全結合ネットの例を説明する．

自然言語処理やロボティクスなどでは，双曲空間 $\mathbb{H}^m$ 上のデータや正定値対称行列のなす Riemann 多様体（SPD 多様体 $P(m)$）上のデータが現れることが知られている．これに呼応して，双曲空間上のニューラルネット（hyperbolic neural network; HNN）[6][21][30] や SPD 多様体上のニューラルネット（SPDNet）[4][7][10] を定式化する研究が展開されている．しかし，$\mathbb{R}^m$ 上の全結合ニューロンを構成するスカラー積 $\boldsymbol{a}\cdot\boldsymbol{x}$ やバイアス $-b$，要素毎の活性化 $\sigma(\cdot)$ を多様体上で定義する方法は自明でなく，幾何学的考察と実装上の都合を折衷したネットワークが提案されてきた．[25] では，双曲空間や SPD 多様体を具体例として含む**非コンパクト対称空間 $\boldsymbol{X}=\boldsymbol{G}/\boldsymbol{K}$** 上で連続ニューラルネット $S[\gamma]$ を定式化し，X 上の Fourier 変換である **Helgason-Fourier 変換** [8][9] を用いて，5 節と同様に Fourier 表示を経由する方法でリッジレット変換 $R[f;\rho]$ を導出した．リッジレット変換（再構成公式）により，X 上の連続ニューラルネットの普遍近似定理が自動的に従う．

7.1 全結合ニューロンの幾何学的再定式化

まず Euclid 空間上の全結合ニューロンを幾何学的に定式化し直す（**図 2**）．$r>0,\boldsymbol{u}\in\mathbb{S}^{m-1}$ を用いて $\boldsymbol{a}=r\boldsymbol{u}$ と極座標表示にする．このとき，Euclid 空間上のニューラルネットに現れる $\boldsymbol{u}\cdot\boldsymbol{x}-b$ は，点 $\boldsymbol{x}\in\mathbb{R}^m$ と超平面 $\xi(\boldsymbol{y},\boldsymbol{u}):=\{\boldsymbol{x}\in\mathbb{R}^m\mid\boldsymbol{u}\cdot(\boldsymbol{x}-\boldsymbol{y})=0\}$（ただし $\boldsymbol{u}\cdot\boldsymbol{y}=b$）との符号付距離であることが分かる．一方，非線形関数 σ を様々なスケール r で畳み込む形式 $\sigma(r\cdot)$ は，σ が生成するウェーブレット関数と解釈できる．

この観察から，ニューラルネットは空間 $X(=\mathbb{R}^m$ or $G/K)$ の点 x と図形 ξ の符号付距離 $d(x,\xi)$ に活性化関数 σ のスケール畳み込みを作用した幾何学的なウェーブレット変換

$$\int_{\mathbb{R}\times\Xi}\gamma(a,\xi)\sigma(ad(x,\xi))\mathrm{d}a\mathrm{d}\xi$$

として一般化できる．ただし Ξ は図形 ξ の集合であり，$\mathrm{d}\xi$ はその上の適当な測度である．

非コンパクト対称空間 X の基本的な図形 ξ として，**測地線**や**球**，**ホロ球**がある．ホロ球とは，半径無限大の球である．Euclid 空間上の超平面は測地線の集合であり，同時に半径無限大の Euclid 球でも

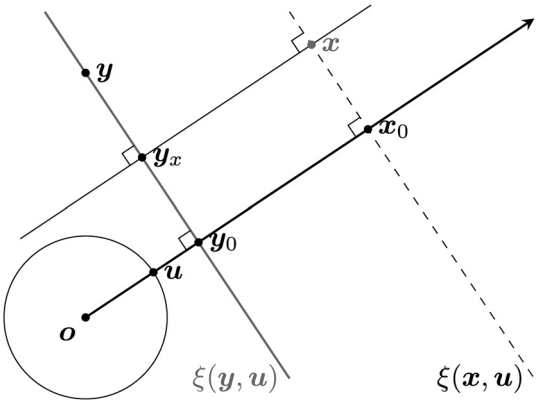

図 2 Euclid 空間上の全結合ニューロン $\sigma(a\cdot x-b)$ は，点 x と超平面 $\xi(y,u)$ との符号付き距離 $d(x,\xi)$ にウェーブレット関数 σ を適用したもの.

ある．従って，測地線集合とホロ球はいずれも Euclid 超平面の対応物とみなせる．[6][21] では測地線集合，[25]（本研究）ではホロ球をそれぞれ図形 ξ として採用している．

7.2 Poincaré 円盤モデル

非コンパクト対称空間の典型的な具体例として双対空間 $\mathbb{H}^2$ の Poincaré 円盤モデル $\mathbb{B}^2$ を説明する．$\mathbb{B}^2:=\{\boldsymbol{x}\in\mathbb{R}^2\mid|\boldsymbol{x}|<1\}$ を $\mathbb{R}^2$ の単位開円盤とする．各点 $\boldsymbol{x}\in\mathbb{B}^2$ における接ベクトル $\boldsymbol{u},\boldsymbol{v}\in T_{\boldsymbol{x}}\mathbb{B}^2$ に対し，$\mathfrak{g}_{\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u},\boldsymbol{v})/(1-|\boldsymbol{x}|^2)^2$ により Riemann 計量を入れる．ここで $(\cdot,\cdot)$ は $\mathbb{R}^2$ の Euclid 内積である．$\mathbb{B}^2$ は双曲平面 $\mathbb{H}^2$ の **Poincaré 円盤モデル**と呼ばれる．$\partial\mathbb{B}^2:=\{\boldsymbol{u}\in\mathbb{R}^2\mid|\boldsymbol{u}|=1\}$ を $\mathbb{B}^2$ の境界とする．$\partial\mathbb{B}^2$ には一様確率測度 $\mathrm{d}\boldsymbol{u}$ を入れる．

Poincaré モデルにおいて，次のことが知られている: (1) 測地線は境界 $\partial\mathbb{B}^2$ に直交する Euclid 円弧である．(2) ホロ円は境界 $\partial\mathbb{B}^2$ に接する Euclid 円である．従って特に，任意のホロ円 ξ は，ξ が通る点 $\boldsymbol{x}\in\mathbb{B}^2$ と境界 $\partial\mathbb{B}^2$ に接する点 $\boldsymbol{u}\in\partial\mathbb{B}^2$ によって指定できる．(3) 原点 $\boldsymbol{o}\in\mathbb{B}^2$ からホロ円 $\xi(\boldsymbol{x},\boldsymbol{u})$ への符号付距離は $\langle\boldsymbol{x},\boldsymbol{u}\rangle=\log\frac{1-|\boldsymbol{x}|^2}{|\boldsymbol{x}-\boldsymbol{u}|^2}$．(4) 特に，$\mathbb{B}^2$ 上の関数 f の Helgason-Fourier 変換とその反転公式は以下で与えられる:

$$\widehat{f}(\lambda,\boldsymbol{u}):=\int_{\mathbb{B}^2}f(\boldsymbol{x})e^{(-i\lambda+1)\langle\boldsymbol{x},\boldsymbol{u}\rangle}\mathrm{dvol}_{\mathfrak{g}}(\boldsymbol{x})$$
$$f(\boldsymbol{x})=\frac{1}{4\pi^2}\int_{\mathbb{R}\times\mathbb{S}^1}\widehat{f}(\lambda,\boldsymbol{u})e^{(i\lambda+1)\langle\boldsymbol{x},\boldsymbol{u}\rangle}\frac{\mathrm{d}\lambda\mathrm{d}\boldsymbol{u}}{|\boldsymbol{c}(\lambda)|^2}.$$

ただし $\boldsymbol{c}$ は Harish-Chandra の $\boldsymbol{c}$-関数であり，$|\boldsymbol{c}(\lambda)|^{-2}=\frac{\pi\lambda}{2}\tanh\frac{\pi\lambda}{2}$ である．なお，Helgason-Fourier 変換の積分核は X 上の Laplace-Beltrami 作用素の固有関数である．

7.3 ホロ円 HNN

$\mathbb{B}^2$ 上の連続ニューラルネットを以下のように定義する: 各点 $\boldsymbol{x}\in\mathbb{B}^2$ に対し，

$$S[\gamma](\boldsymbol{x})=\int_{\mathbb{R}\times\mathbb{S}^1\times\mathbb{R}}\gamma(a,\boldsymbol{u},b)\sigma(a\langle\boldsymbol{x},\boldsymbol{u}\rangle-b)e^{\langle\boldsymbol{x},\boldsymbol{u}\rangle}\mathrm{d}a\mathrm{d}\boldsymbol{u}\mathrm{d}b.$$

ただし技術的な制約により，重み関数 $\exp(\langle\boldsymbol{x},\boldsymbol{u}\rangle)$（Poisson 核）を導入した（**図 3** も参照）．このとき，

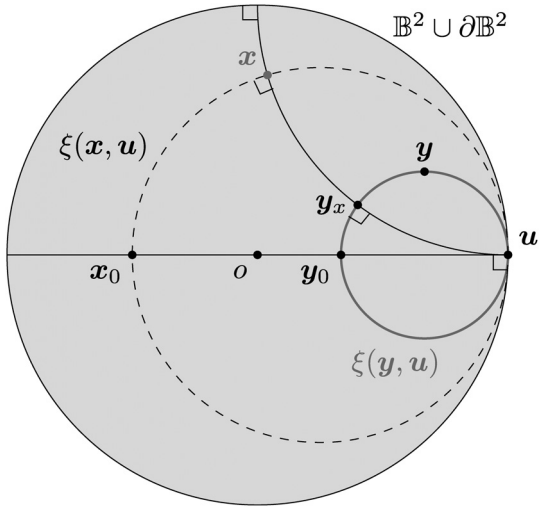

図 3　Poincaré 円盤 $\mathbb{B}^2$ 上のニューラルネットに現れる $\langle x,u\rangle - b$（ただし $a=1$ とする）は，点 $x\in\mathbb{B}^2$ とホロ円 $\xi(y_0,u)$（ただし $\langle y_0,u\rangle = b$）との符号付距離である．

リッジレット変換は以下で与えられる: 各点 $(a,\boldsymbol{u},b)\in\mathbb{R}\times\mathbb{S}^1\times\mathbb{R}$ に対し，

$$R[f;\rho](a,\boldsymbol{u},b)=\int_{\mathbb{B}^2}\boldsymbol{c}[f](\boldsymbol{x})\overline{\rho(a\langle\boldsymbol{x},\boldsymbol{u}\rangle-b)}e^{\langle\boldsymbol{x},\boldsymbol{u}\rangle}\mathrm{dvol}_{\mathfrak{g}}(\boldsymbol{x}).$$

ただし各点 $\boldsymbol{x}\in\mathbb{B}^2$ に対し $\boldsymbol{c}[f](\boldsymbol{x}):=\int_{\mathbb{R}\times\mathbb{S}^1}\widehat{f}(\lambda,\boldsymbol{u})e^{(i\lambda+1)\langle\boldsymbol{x},\boldsymbol{u}\rangle}\frac{\mathrm{d}\lambda\mathrm{d}\boldsymbol{u}}{|\boldsymbol{c}(\lambda)|^4}$ である．特に，$(\!(\sigma,\rho)\!):=\frac{1}{2\pi}\int_{\mathbb{R}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-1}\mathrm{d}\omega$ として，次の再構成公式が成り立つ:

$$S[R[f;\rho]]=(\!(\sigma,\rho)\!)f.$$

証明は 5 節と同様に Fourier 表示を経由し，Helgason-Fourier 反転公式を利用する（[25] 参照）．

参 考 文 献

[1] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, Vol. 39, No. 3, pp. 930–945, 1993.

[2] E. J. Candès. Ridgelets: Theory and applications. PhD thesis, Standford University, 1998.

[3] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, Vol. 32, pp. 3036–3046, 2018.

[4] Z. Dong, S. Jia, C. Zhang, M. Pei, and Y. Wu. Deep manifold learning of symmetric positive definite matrices with application to face recognition. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, pp. 4009–4015. AAAI Press, 2017.

[5] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint: 2101.03961, 2021.

[6] O. Ganea, G. Becigneul, and T. Hofmann. Hyperbolic neural networks. *Advances in Neural Information Processing Systems*, Vol. 31, 2018.

[7] Z. Gao, Y. Wu, X. Bu, T. Yu, J. Yuan, and Y. Jia. Learning a robust representation via a deep network on symmetric positive definite manifolds. *Pattern Recognition*, Vol. 92, pp. 1–12, 2019.

[8] S. Helgason. *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*. American Mathematical Society, 1984.

[9] S. Helgason. *Geometric Analysis on Symmetric Spaces: Second Edition*. American Mathematical Society, 2008.

[10] Z. Huang and L. V. Gool. A Riemannian network for SPD matrix learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, pp. 2036–2042. AAAI Press, 2017.

[11] S. Kostadinova, S. Pilipović, K. Saneva, and J. Vindas. The ridgelet transform of distributions. *Integral Transforms and Special Functions*, Vol. 25, No. 5, pp. 344–358, 2014.
[12] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp. 115–133, 1943.
[13] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, Vol. 115, No. 33, pp. E7665–E7671, 2018.
[14] N. Murata. An integral representation of functions using three-layered betworks and their approximation bounds. *Neural Networks*, Vol. 9, No. 6, pp. 947–956, 1996.
[15] A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. arXiv preprint: 1712.05438, 2017.
[16] R. Parhi and R. D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, Vol. 22, No. 43, pp. 1–40, 2021.
[17] G. Rotskoff and E. Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. *Advances in Neural Information Processing Systems*, Vol. 31, pp. 7146–7155, 2018.
[18] B. Rubin. The Calderón reproducing formula, windowed X-ray transforms, and radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications*, Vol. 4, No. 2, pp. 175–197, 1998.
[19] D. E. Rumelhart, J. L. McClelland, and P. R. Group. *Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, 1986.
[20] P. Savarese, I. Evron, D. Soudry, and N. Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the 32nd Conference on Learning Theory*, pp. 2667–2690, 2019.
[21] R. Shimizu, Y. Mukuta, and T. Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, https://openreview.net/forum?id=Ec85b0tUwbA, 2021.
[22] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, Vol. 130, No. 3, pp. 1820–1852, 2020.
[23] S. Sonoda, I. Ishikawa, and M. Ikeda. Ghosts in neural networks: Existence, structure and role of infinite-dimensional null space. arXiv preprint: 2106.04770, 2021.
[24] S. Sonoda, I. Ishikawa, and M. Ikeda. Ridge regression with over-parametrized two-layer networks converge to Ridgelet spectrum. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021*, Vol. 130, pp. 2674–2682, 2021.
[25] S. Sonoda, I. Ishikawa, and M. Ikeda. Fully-connected network on noncompact symmetric space and ridgelet transform based on Helgason-Fourier analysis. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:20405–20422, 2022.
[26] S. Sonoda, I. Ishikawa, M. Ikeda. Universality of group convolutional neural networks based on ridgelet analysis on groups. To appear in Advances in *Neural Information Processing Systems*, Vol. 35, 2022.
[27] S. Sonoda, I. Ishikawa, M. Ikeda, K. Hagihara, Y. Sawano, T. Matsubara, and N. Murata. The global optimum of shallow neural network is attained by ridgelet transform. arXiv preprint: 1805.07517, pp. 1–14, 2018.
[28] S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, Vol. 43, No. 2, pp. 233–268, 2017.
[29] M. Unser. A representer theorem for deep neural networks. *Journal of Machine Learning Research*, Vol. 20, No. 110, pp. 1–30, 2019.
[30] M.-X. Wang. Laplacian eigenspaces, horocycles and neuron models on hyperbolic spaces, https://openreview.net/forum?id=ZglaBL5inu, 2021.

[**Abstract**]

Characterization of the typical deep learning solutions is crucial to understanding and controlling deep learning. Due to the complex structure of real deep neural networks (NNs), various simplified mathematical models are employed in conventional theoretical analysis. In this study, we describe a mathematical model of a single hidden layer in an NN, which is an integral representation of NNs, and its right inverse operator (or analysis operator), the ridgelet transform. Furthermore, while the classical ridgelet transform was obtained heuristically, we had recently developed a natural technique to derive it. As an application, we succeeded in developing an NN on manifolds (noncompact symmetric spaces) and deriving the associated ridgelet transform.

Keyword: geometric deep learning, integral representation of neural networks, ridgelet transform, null space, noncompact symmetric space
キーワード：幾何学的深層学習，積分表現ニューラルネット，リッジレット変換，零空間，非コンパクト対称空間