P-10

# ニューラルネットワークにおけるリーマン計量構造について
**Riemannian metrical structure in Neural Networks**

○伊藤龍二 [1], 青柳美輝 [2]
Ryuji Ito, Miki Aoyagi

Abstract: This study deals with neural networks in the sense of geometric transformations acting on the coordinate representation of the underlying data manifold which the data is sampled from. In this paper, we introduce a formalized general theory of neural networks in the setting of Riemannian geometry.

## 1. はじめに

本稿では，ニューラルネットワークのデータサンプリング表現において，それぞれの層における座標変換をデータ多様体の幾何学的変換とみなすことにより，ニューラルネットワーク構造のリーマン幾何学による一般的な理論構築を目指す．

## 2. ニューラルネットワーク

本稿ではアインシュタインの縮約記法を使用する．$x^{(l)}$ は $l$ 番目の座標系，$\varphi^{(l)}$ は $l$ 番目の座標変換を表す．括弧のない添字の，上付きの添字はベクトルの成分，下付きの添字は余ベクトルの成分を表し，同じ項で上付きと下付きの添字が同じ場合はその添字での和をとることを意味する．また，テンソル $A^{a\cdot}_{\cdot b}$ のような添字内のピリオドはどの指標が 1 番目か 2 番目かを表す記号である．

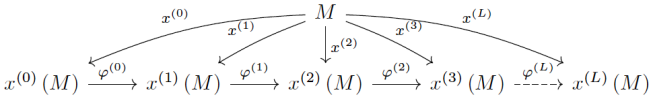


図 1: ニューラルネットワークの幾何学的表現 ([1])

$M$ をハウスドルフ，パラコンパクト位相空間とし，局所的に $\mathbb{R}^{\dim M}$ と同相となる多様体とする．開集合 $U \subset M$ に対し，同相写像 $x : U \to x(U) \subseteq \mathbb{R}^{\dim M}$ を $U$ 上の座標系とする．フィードフォワードネットワークにおいては，座標変換 $\varphi^{(l)} : x^{(l)}(M) \to (\varphi^{(l)} \circ x^{(l)})(M)$ を学習することにより，新しい座標 $x^{(l+1)} := \varphi^{(l)}(x^{(l)}) : M \to x^{(l+1)}(M)$ を決定する．ただし，$l = 0$ のときはデカルト座標 $x^{(0)} : M \to x^{(0)}(M)$ である．データ多様体上の点 $q \in M$ は，それぞれの座標系でスカラーベクトルとして表現される．すなわち，$l+1$ 層において，点 $q$ の座標は，層による合成 $x^{(l+1)}(q) := (\varphi^{(l)} \circ \ldots \circ \varphi^{(1)} \circ \varphi^{(0)} \circ x^{(0)})(q)$ として表される．出力層における座標表現を $x^{(L)}(M) \subseteq \mathbb{R}^{\dim M}$ とする．ReLU や tanh などの活性化関数 $f$ に対して，標準的なフィードフォワードネットワークでは，座標変換は $x^{(l+1)} := \varphi^{(l)}(x^{(l)}) := f(x^{(l)}; l)$ となる．ReLU は全単射ではないので，適切な座標表現とはならない．残差ネットワークを用いて座標変換 $x^{(l+1)} := \varphi^{(l)}(x^{(l)}) := x^{(l)} + f(x^{(l)}; l)$ 及び，$f$ を ReLU とすれば，全単射になる．

座標系 $x^{(l+1)} := \varphi^{(l)} \circ \ldots \circ \varphi^{(1)} \circ \varphi^{(0)} \circ x^{(0)}$ は，ニューラルネットワークによって学習された座標変換 $\varphi^{(l)} : x^{(l)}(M) \to (\varphi^{(l)} \circ x^{(l)})(M)$ を表現している．

また，それぞれの座標系での計量 $g_{x^{(l)}(M)}(X, Y) := g_{(\varphi^{(l)} \circ x^{(l)})(M)}(\varphi^{(l)}_* X, \varphi^{(l)}_* Y)$ を接空間の間の写像 $\varphi^{(l)}_* : Tx^{(l)}(M) \to T(\varphi^{(l)} \circ x^{(l)})(M)$ によって，第 $l+1$ 層から第 $l$ 層に引き戻すことによって定義される引き戻し計量とする．

標準的なフィードフォワードニューラルネットワークは以下のように定義される．

$$x^{(l+1)} := f(x^{(l)}; l) \tag{1}$$

また，残差ネットワークは次のように定義される．

$$x^{(l+1)} = x^{(l)} + f(x^{(l)}; l) \tag{2}$$

本稿では，最終的には $L \to \infty$, $l \in [0, 1] \subset \mathbb{R}$ とし，解析を行うことを目標とする．従って，残差ネットワークを次のように表す．

$$x^{(l+1)} \simeq x^{(l)} + f(x^{(l)}; l)\Delta l \tag{3}$$

ここで，$\Delta l = 1/L$ とする．

## 3. ニューラルネットワークにおけるリーマン計量テンソル

可微分幾何学の視点から考察すれば，データ多様体は同じままだが，ニューラルネットワークの層を重ねるごとに，データ多様体の座標表現が非線形活性化のたびに，変化すると見ることができる．ニューラルネットワークの目的は，異なるクラスの集合が超平面によって直線的に分離できるような，データ多様体の座標表現を見つけることである．

1: 日大理工・院 (前)・数学　2: 日大理工・教員・数学

**定義 1 (リーマン多様体)** $M$ を $n$ 次元可微分多様体とする．$M$ の接空間上で滑らかに変化する正定値計量テンソル $g$ が与えられた多様体 $(M, g)$ をリーマン多様体と定義する．

ネットワークが分類器として十分に訓練されている場合，ユークリッド距離では，同じクラスの 2 つの入力座標が遠く離れていても，出力座標では近くにある場合がある．同様に，異なるクラスの 2 つの点は，入力座標で表されたときには互いに近く，出力座標では遠く離れていることがある．実際に距離を測定するのは，出力座標であり，教師なしの場合であっても，データ多様体を平坦化した表現になる傾向がある．したがって，出力座標における計量はユークリッド計量になる．

$$g(x^{(L)})_{a_L b_L} := \eta_{a_L b_L} \tag{4}$$

計量テンソルの要素は，座標変換を伴うテンソルとして変換される．

$$g(x^{(l)})_{a_l b_l} = \left(\frac{\partial x^{(l+1)}}{\partial x^{(l)}}\right)^{a_{l+1}\cdot}_{\cdot a_l} \left(\frac{\partial x^{(l+1)}}{\partial x^{(l)}}\right)^{b_{l+1}\cdot}_{\cdot b_l} g(x^{(l+1)})_{a_{l+1} b_{l+1}} \tag{5}$$

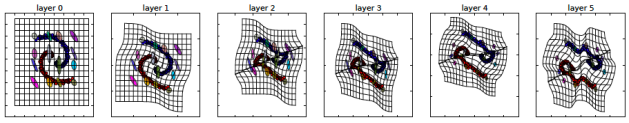


図 2: 計量変化の様子 [1]

図 2 は，螺旋状の多様体を分離する双曲接線活性化関数をもつ残差ネットワークのそれぞれの座標変換を表している．また，一定半径 $ds = \sqrt{g_{a_l b_l}(x^{(l)}) dx^{(l)a_l} dx^{(l)b_l}}$ をもつ球をいろいろな点及びそれぞれの座標系で描いたものである．出力座標では，式 (4) で表される標準的なユークリッド距離で測定されるため，円は丸いものになる．計量テンソルの座標表現は，式 (5) および式 (6) によってネットワークを介して入力座標に引き戻される．したがって，データ多様体上では，入力層に対応するデカルト座標上でみれば，円は丸くならない．上記の再帰的な式は出力層から入力層に向かっており，計量テンソルの座標表現は，出力から入力へとネットワークを介してバックプロパゲーションされたものになる．

$$g(x^{(l)})_{a_l b_l} = \prod_{l'=L-1}^{l} \left[ \left(\frac{\partial x^{(l'+1)}}{\partial x^{(l')}}\right)^{a_{l'+1}\cdot}_{\cdot a_{l'}} \left(\frac{\partial x^{(l'+1)}}{\partial x^{(l')}}\right)^{b_{l'+1}\cdot}_{\cdot b_{l'}} \right] \eta_{a_L b_L} \tag{6}$$

ネットワークを残差を用いて式 (3) により定義すれば，座標変換のヤコビアンが得られる．また，$\delta^{a_{l+1}\cdot}_{\cdot a_l}$ をクロネッカーのデルタ表示とすれば，

$$\left(\frac{\partial x^{(l+1)}}{\partial x^{(l)}}\right)^{a_{l+1}\cdot}_{\cdot a_l} = \delta^{a_{l+1}\cdot}_{\cdot a_l} + \left(\frac{\partial f(x^{(l)}; l)}{\partial x^{(l)}}\right)^{a_{l+1}\cdot}_{\cdot a_l} \Delta l \tag{7}$$

となる．つまり，バックプロパゲーションによる計量テンソルの座標表現は，出力から入力への行列積によって，任意の層 $l$ で定義できる．

$$P^{a_L\cdot}_{\cdot a_l} \equiv \prod_{l'=1}^{L-1} \left[ \delta^{a_{l+1}\cdot}_{\cdot a_l} + \left(\frac{\partial f(z^{l'+1}; l')}{\partial z^{(l'+1)}}\right)^{a_{l'+1}\cdot}_{\cdot e_{l'+1}} \left(\frac{\partial z^{(l'+1)}}{\partial x^{l'}}\right)^{e_{l'+1}\cdot}_{\cdot a_{l'}} \Delta l \right] \tag{8}$$

ここで，$z^{(l+1)} := W^{(l)} \cdot x^{(l)} + b^{(l)}$ とする．これにより出力計量を標準的なユークリッド計量 $\eta_{ab}$ とすれば，線素は任意の層 $l$ の座標系で次のように表される．

$$ds^2 = \eta_{ab} P^{a\cdot}_{\cdot a_l} P^{b\cdot}_{\cdot b_l} dx^{a_l} dx^{b_l} \tag{9}$$

これまでの解析では，層ごとの次元が一定であると仮定してきたが，応用上用いられるニューラルネットワークは，ノード数が頻繁に変化する．この場合は，引き戻し計量によって処理すればよい．プッシュフォワード (写像の微分) ヤコビアン行列のランクがすべての $p \in M$ に対して一定である限り，多様体は低次元および高次元の空間に埋め込みすることができる．データの基礎となるデータ多様体の次元は，ニューラルネットワークの最小のボトルネック層である層の次元と考えられる．すなわち，$\dim M := \min_l \dim x^{(l)}(M)$ となり，他のすべての高次元の層は，この最小次元の表現の埋め込み表現になる．実際，応用上，ニューラルネットワークの次元を変更することは必要と思われる．まず，tanh，$\sigma$，ReLU のようなニューラルネットワークが利用できる非線形座標変換の種類が限られている状況の下では，ニューラルネットワークが存在する多種多様な多様体を分離する能力が制限される．例えば，tanh 形式の座標変換しか利用しない場合，図 2 の単純な螺旋を線形的に分離することは困難である．もし，極座標のような螺旋に適した座標系を使えば簡単にデータを分離することができる．また，ネットワークを高次元化することで，データの分離が容易になる．

## 4. 参考文献

[1] Michael Hauser and Asok Ray: ”Principles of Riemannian Geometry in Neural Networks”,Neural Information Processing Systems (NIPS),2017,P1–10.