深層ニューラルネットの積分表現理論

Integral Representation Theory of Deep Neural Networks

2017年2月

早稲田大学大学院先進理工学研究科
電気・情報生命専攻 情報学習システム研究

園田　翔
Sho SONODA

目 次

記号一覧

$\mathbb{N}$	0 を含まない自然数
$\mathbb{N}_0$	0 を含む自然数
$\mathbb{R}$	実数
$\mathbb{C}$	複素数
$\mathbb{R}^m$	m 次元 Euclid 空間
$\mathbb{S}^{m-1}$	$m-1$ 次元単位球面
$\mathbb{R}_+$	開半直線
$\mathbb{H}$	開半空間
$\mathbb{Y}^{m+1}$	ニューラルネットの隠れ層パラメータ $(a,b) \in \mathbb{R}^m \times \mathbb{R}$ または $(u,\alpha,\beta) \in \mathbb{S}^{m-1} \times \mathbb{R}_+ \times \mathbb{R}$ の空間
$\overline{z}$	複素数 z の複素共役
$\widetilde{f}$	関数 f の反転（reflection）$\widetilde{f}(x) := f(-x)$
$f * g$	関数 f, g の畳み込み
$f \lesssim g$	関数 f, g に対し，ある正定数 $C \geqslant 0$ が存在して $f \leqslant Cg$ が成り立つ
$\partial_t f$	関数 $f(x,t)$ の時間微分 $\partial f(x,t)/\partial t$
∇f	関数 $f(x,t)$ の勾配 $\partial f(x,t)/\partial x$
$\triangle f$	関数 f の Laplacian
$f_\sharp \pi$	関数 f による確率測度 π の押出測度
id	恒等写像
$\mathbf{1}_A$	集合 A の指示関数

$\widehat{f}$	Fourier 変換
$\mathcal{H}$	Hilbert 変換
R	Radon 変換
$\mathrm{R}^\dagger$	双対 Radon 変換
Λ	逆投影作用素
$\mathcal{W}$	ウェーブレット変換
$\mathcal{W}^\dagger$	双対ウェーブレット変換
$\mathscr{R}$	リッジレット変換
$\mathscr{R}^\dagger$	双対リッジレット変換

表 1: 関数と超関数のクラス

関数空間	$\mathcal{A}$	双対空間	$\mathcal{A}'$
多項式関数	$\mathcal{P}$	–	
連続関数	C	–	
一様連続関数	UC	–	
無限遠で消滅する連続関数	C_0	–	
台がコンパクトな連続関数	C_c	–	
p 次可積分関数	L^p	可積分関数 $(1/p+1/q=1)$	L^q
局所可積分関数	L^1_{loc}	–	
滑らかな関数	$\mathcal{E}$	台がコンパクトな超関数	$\mathcal{E}'$
急減少関数	$\mathcal{S}$	緩増加超関数	$\mathcal{S}'$
台がコンパクトかつ滑らかな関数	$\mathcal{D}$	Schwartz 超関数	$\mathcal{D}'$
緩増加関数	$\mathcal{O}_{\mathcal{M}}$	–	
–		急減少超関数	$\mathcal{O}'_{\mathcal{C}}$
Lizorkin 関数	$\mathcal{S}_0$	Lizorkin 超関数	$\mathcal{S}'_0$
確率測度	$\mathscr{P}$	–	

表 2: クラスの包含関係

$$
\begin{array}{lccccccc}
\text{(関数)} & \mathcal{D}(\mathbb{R}^m) & \subset & \mathcal{S}(\mathbb{R}^m) & \subset & \mathcal{O}_{\mathcal{M}}(\mathbb{R}^m) & \subset & \mathcal{E}(\mathbb{R}^m) \\
 & \cap & & \cap & & \cap & & \cap \\
\text{(超関数)} & \mathcal{E}'(\mathbb{R}^m) & \subset & \mathcal{O}'_{\mathcal{C}}(\mathbb{R}^m) & \subset & \mathcal{S}'(\mathbb{R}^m) & \subset & \mathcal{D}'(\mathbb{R}^m)
\end{array}
$$

第1章　序論

深層ニューラルネットは，2012 年頃から機械学習や人工知能の分野で急速に発展を続けている学習機械である。深層ニューラルネットの快挙は，大画像に対する一般物体認識タスクで人間と同程度のスコアを記録し，囲碁では「人類最強」とも呼ばれる棋士イ・セドル氏に勝利するなど，枚挙に暇がない。ニューラルネットは，神経細胞が繋がり合って情報を処理する様子を抽象化した「脳の数理モデル」として，20 世紀半ばに登場し，これまでに二度のブームを引き起こしている。深層ニューラルネットは第三次ブームの立役者である。

「深層」という修飾語は，中間層の数が従来のニューラルネットよりも多いことを強調している。ニューラルネットを深層化することで，内部の情報表現が階層化され，情報処理が効率化されることは，以前から予想されていた。しかし，古典的な学習法であるバックプロパゲーション（backpropagation）では，深層ニューラルネットを学習させることができなかった。原因は様々だが，例えば，層が深くなるに連れて，学習に必要な誤差信号が減衰し，学習が極端に遅くなるためである。深層ニューラルネットを学習させる技術を総称して，深層学習という。深層学習が立て続けに成功し始めたのは，2006 年の Hinton や Bengio のプレトレーニングからである。

本研究では，深層ニューラルネットの中で何が起きているのか，なぜ深層にした方が良いのかという問題に対して，深層ニューラルネットの積分表現理論の開発を通じて問題解決を図る。深層ニューラルネットの内部では，タスクに有利な情報表現（特徴量写像）が獲得されていると考えられている。情報表現を自動的に獲得するという意味で，深層学習は表現学習とも呼ばれる。しかし，深層学習はヒューリスティクスを多く含むので，実際に獲得される特徴量の素性は分からないことも多い。そもそも，浅いニューラルネットは任意の関数を近似できるほど表現力が高い（万能関数近似器）のに，なぜ深層にする必要があるのだろうか。

本研究が拠り所とする積分表現は，ニューラルネットの中間層素子に

関する総和を積分に置き換えて得られる。これは中間層素子を積分核とする積分変換であり，双対リッジレット変換と呼ばれる。リッジレット変換はRadon変換やウェーブレット変換との関係が深く，幾何学的性質や解析的性質がよく調べられている。通常のニューラルネットは，積分表現の離散化を通じて理解できる。積分表現理論は90年代に起きた第二次ブームにおいて，浅いニューラルネットの表現能力を調べる過程で成立した。残念ながら，深層ニューラルネットの積分表現理論は今日までほとんど調べられていない。中間層が二層以上ある場合には，単に積分核が入れ子になるだけで，中間層同士の関係をうまく定式化できないためである。

本研究の結果は二つに分けられる：浅いニューラルネットの積分表現理論と，深層ニューラルネットの積分表現理論である。浅いニューラルネットの理論では，ReLUと呼ばれる活性化関数に対応するように積分表現理論を拡張し，ニューラルネットとRadon変換およびウェーブレット変換との関係を詳らかにし，さらに積分表現を離散化してニューラルネットを学習する方法を提案した。

深層ニューラルネットの理論では，デノイジング・オートエンコーダー（denoising autoencoder; DAE）と呼ばれるクラスに対して，DAEを輸送写像とみなす方法で，積分表現を構成した。また，輸送写像の極限を調べることで，無限層ニューラルネットに相当する連続DAEの性質を明らかにした。DAEはデータ分布のエントロピーを減らす方向に入力データを再配置する輸送作用があり，この作用は層を深くした方が顕著になることが分かった。従って，浅いDAEと深層DAEとでは抽出される特徴量が異なることから，DAEにおいては積極的に深層化すべきであると言える。本研究の結果を深層学習のアルゴリズムに反映する方法の開発は，今後の重要な課題である。

本論文の構成は，第1章が本研究の概要と論文の構成の説明，第2章が関連研究と先行研究のサーベイ，第3章と第4章が本論を展開するうえでの準備，第5章から第7章が本論，第8章が本研究の総括である。第2章以降の各章の詳細は次の通りである。

第2章では関連研究および先行研究について，深層ニューラルネットと浅いニューラルネットの二つの観点で整理する。まず深層ニューラルネットについては，最初に全体の動向を概観する。次に，本研究の主題の一つである「深層ニューラルネットの中では何が起きているか」について言及している研究を整理する。本研究で取り扱うReLUやDAEについ

ては独立に節を設けるほか，オートエンコーダーと対照的な表現学習の例として，畳み込みネットワークについても解説する。一方，浅いニューラルネットについては，まず 90 年代の結果を整理する。具体的には，万能関数近似能力を軸にして積分表現理論が登場するまでの経緯を説明する。続いて，積分表現理論以降に登場したリッジレット解析や学習理論について，その後の展開を整理する。

第 3 章では，本研究で用いる数学的な道具を整理する。具体的には，Fourier 変換や Radon 変換，ウェーブレット解析，拡散方程式，最適輸送理論の基本的な定理や公式を整理する。さらに，本論で展開される超関数や特異積分の計算について解説する。これらの計算には，これまでにまとまった解説が少なく，申請者が独自に計算した内容も含む。

第 4 章では積分表現理論について基本事項を説明する。本章は本論を展開するうえでの準備にあたるが，積分表現理論は本研究の要であり，申請者の考察も多く含むことから，独立に章を設けた。まず積分表現理論がリッジレット解析と等価であることを説明したあと，リッジレット変換が Radon 変換とウェーブレット変換の合成変換に分解できることを示す。これにより，リッジレット解析の幾何学的な意味付けが明らかとなる。最後に，リッジレット変換の離散化や，ベクトル値の場合の考え方を説明する。これにより，現実のニューラルネットと積分表現との関係が明らかとなる。

第 5 章では浅いニューラルネットの積分表現理論を展開する。まず，深層学習において，ReLU と呼ばれる非有界な活性化関数が用いられる背景を簡単に説明する。これにより，深層ニューラルネットの積分表現理論を展開するためには ReLU を含む超関数によるリッジレット解析が必要であることが分かる。本章の前半では，超関数によるリッジレット変換が存在すること，および適当な条件の下で再構成公式（逆変換）が成り立つことを理論的に示す。後半では，リッジレット変換の具体例を解析的に計算し，さらに再構成公式の数値例を計算することで，理論の実効性を確認する。

第 6 章では深層ニューラルネットの積分表現理論を展開する。まず, DAE が登場した背景と，DAE の学習アルゴリズムを簡単に説明し，Alain and Bengio の変分計算によって学習アルゴリズムの停留点が陽に求まることを示す。続いて，得られた DAE が輸送写像とみなせることを説明する。本章の前半では，浅い DAE による輸送の性質を調べる。後半では，三つの深層 DAE（積層 DAE，合成 DAE，連続 DAE）を導入し，深層 DAE に

よる輸送現象を軸として深層DAEの積分表現理論を展開する。積層DAEは深層学習の一種であるプレトレーニングで現れる形式だが，解析が難しい。合成DAEは浅いDAEの合成写像であり，これ自体も輸送写像なので解析は比較的容易である。連続DAEは合成DAEの連続極限であり，無限層のニューラルネットに相当する。本章の主結果は二つある。まず，連続DAEによる輸送に伴って変形されたデータ分布（押出測度）が，逆向きの拡散方程式に従うことを示す。つまり，連続DAEはデータ分布のエントロピーを減らすようにデータ点を再配置する連続力学系である。次に，積層DAEと合成DAEの等価性を示す。つまり，積層DAEから得られる特徴量は，ある線形写像によって適当な合成DAEから得られた特徴量に変換できる。二つの主結果の系として，合成DAEと積層DAEはいずれも，層を重ねるに連れて連続DAEと類似の振舞いをするようになることが分かる。最後に，深層DAEの積分表現は，層毎の積分表現を合成したものとして得る。

第7章では積分表現を離散化することでニューラルネットを学習させる方法を説明する。再構成公式を離散化することで学習済ニューラルネットが得られる。離散化は離散フーリエ変換のように規則的な格子に沿って行うこともできるが，本章ではサンプリングによる方法を提案する。これは，データからリッジレット変換を推定し，得られた変換を確率分布とみなして，パラメータをサンプリングする方法である。リッジレット変換をパラメータ空間上の確率分布とみなしたものをオラクル分布と呼ぶ。人工データおよび実データに対してアルゴリズムを適用し，バックプロパゲーションに依らない学習が行えることを確認した。

第8章では本研究を総括し，今後の展望について述べる。

なお付録Aでは，本文で省略した定理の証明を掲載している。また付録Bでは，情報やエントロピー，複雑性などの基本的かつ解釈の難しい概念について，諸分野での用例を元にして整理する。本付録の内容は，深層ニューラルネットの中でどのように情報を処理しているかについて考察を加えるための背景知識となるが，本論を展開するうえで必ずしも全て理解しておく必要はないため，付録に置いた。

第2章　本研究の位置づけ

深層学習と積分表現理論という二本の軸で本研究の立ち位置を整理する。ニューラルネットは今日の深層学習ブームを含めて，これまでに三度のブームを引き起こしている。第一次ブームは，ニューラルネットが登場して間もなく起きた。当時のニューラルネットは，形式ニューロンや線形パーセプトロンと呼ばれる中間層を持たないニューラルネットであった。第二次ブームは，多層パーセプトロンと呼ばれる中間層を備えたニューラルネットと，その学習法であるバックプロパゲーション（backpropagation）が主役であった。積分表現理論は第二次ブームに端を発する，浅いニューラルネットの理論である。第三次ブームは，さらに多くの中間層を備えた深層ニューラルネットと，その学習法である深層学習（deep learning）が主役である。

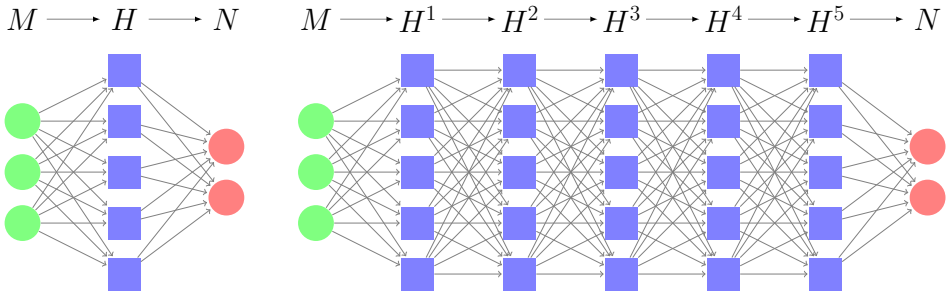


図 2.1: 本研究の対象となる浅いニューラルネット（左）と深層ニューラルネット（右）。緑は入力層（$M = \mathbb{R}^m$），青は中間層（$H = \mathbb{R}^J$），赤は出力層（$N = \mathbb{R}^n$）を表す。

ニューラルネットには，決定的ニューラルネット（deterministic）と，確率的ニューラルネット（probabilistic, generative）がある。確率的ニューラルネットは，Boltzmann Machine（BM）や Restricted Boltzmann Machine（RBM），およびそれらを深層化した Deep Belief Network（DBN）や Deep Boltzmann Machine（DBM）が代表的である。階層型ニューラルネットは，

入力層，複数の中間層，出力層から構成される。階層型でないニューラルネットには，リカレントニューラルネット（recurrent neural network）のように循環的な結合や，Boltzmann Machine のように相互結合をもつものがある。階層型のニューラルネットには，全結合型（full connected）と畳み込み型（convolution）がある。全結合型は多層パーセプトロン（multi layer perceptron）とも呼ばれる。

本研究では主に，決定的かつ階層型の全結合ニューラルネットを扱う。本論文を通じて，単にニューラルネットといえばこれを指す。また，中間層が一層のものを**浅いニューラルネット**（shallow neural network）と呼び，中間層が二層以上のものを**深層ニューラルネット**（deep neural network）と呼ぶ。なお，「深層学習」という言葉は，字義通りには，深層ニューラルネットの学習法の総称であるが，自然言語処理で用いられる再帰型ニューラルネット（recursive neural network）などは浅いモデルに相当するので，文脈に応じて判断されたい。

2.1　深層学習の理論

深層学習の発展は著しい。既に多くの解説 (Goodfellow et al., 2016; LeCun et al., 2015; Schmidhuber, 2015; 麻生英樹 et al., 2015; 得居誠也, 2014; Bengio, 2009; Bengio and Delalleau, 2011; Bengio et al., 2013a; Bengio, 2013) が公開されている一方で，トレンドは毎年のように変化している。以下では本研究に関係するトピックに焦点を絞り，網羅的な解説は他書に譲ることとする。

2.1.1　近年の動向

これまでの深層学習の展開は概ね三つの時代に分けられる。まず，Hinton et al. (2006); Bengio et al. (2007); Ranzato et al. (2007) に始まる事前学習（pre-training）の時代である。事前学習とは，目的のタスクを解く前に，教師なし学習によって深層ニューラルネットのパラメータを調整する操作である。それまで，深層ニューラルネットを学習させることは極めて困難であり，実質的にはほとんど不可能なのではないかとさえ思われていたが，事前学習後のパラメータを初期値として再度バックプロパゲーションを行うことで，目的のタスクを学習できるようになることが分かった。この時代には，深層ニューラルネットが従来の浅いニューラ

ルネットや SVM よりも性能が高いことを実験的に評価する報告も多く (Jarrett et al., 2009; Glorot and Bengio, 2010; Erhan et al., 2010; Coates, 2012)，深層ニューラルネットの実力が広く認知されて商業化が進んだ今日から見れば牧歌的である。事前学習の実体は，入力データに対して教師なし学習を適用して，特徴抽出器を学習することである。従って，事前学習は教師なし特徴学習（unsupervised feature learning）や表現学習（representation learning）とも呼ばれる。Restricted Boltzmann Machine（RBM）やオートエンコーダー（autoencoder），k-means や ICA などの教師なし特徴学習が盛んに研究された (Lee et al., 2008; Larochelle et al., 2007; Lee, 2010; Erhan et al., 2010; Le et al., 2011; Coates and Ng, 2011; Coates et al., 2011; Bengio and Delalleau, 2011; Coates, 2012; Bengio et al., 2013a)。1,000 台の PC クラスタを三日間稼働させたことでも話題になった “Google cat” (Le et al., 2012) などは，教師なし特徴学習の象徴例である。深層学習で標準的に用いられる ReLU (Nair and Hinton, 2010; Glorot et al., 2011) や DropOut (Srivastava et al., 2014) はこの時代に登場した。

次に，Krizhevsky et al. (2012) や Hinton et al. (2012) に始まる畳み込みネットワーク（convolutional networks; ConvNets, CNN）の時代である。畳み込みネットワークの登場により，事前学習を経由せずに深層ニューラルネットを学習できることが明らかとなり，「もはや事前学習は誰も使わない」という言葉も聞かれた。畳み込みネットワークの威力は凄まじく，画像認識や音声認識，化合物活性認識などのコンペで圧倒的な成績を収めた。特に，Krizhevsky et al. (2012) が登場した大画像一般物体認識コンペ（ImageNet Large Scale Visual Recognition Competition; ILSVRC, Russakovsky et al. (2015)）では，翌年からも毎年圧倒的な記録更新が続き，ついに人間と同等な認識精度を達成することとなった (Sermanet et al., 2014; Girshick et al., 2014; Szegedy et al., 2015; Simonyan and Zisserman, 2015; He et al., 2016)。畳み込みネットワークの台頭に伴い，MNIST や CIFAR-10 などの「小規模」データ・セットは，最早ベンチマークとしての意味を為さなくなった (Benenson, 2016)。この時代には機械学習だけでなく画像処理や音声信号処理，自然言語処理，ロボット工学，創薬などの応用分野の研究者も大量に参入して，特に畳み込みネットワークの研究が加速し，認識精度とネットワークの深さを追求する時代になった。Google や Facebook, Microsoft, Baidu などの巨大企業に深層学習の研究者が集中し，大量の GPU を一週間稼働させるといった力量作戦が当

たり前となっていった。畳み込み構造のように，ニューラルネットの構造に細工を加える方法が発達し，2015 年にはついに 1,000 層の深層ニューラルネットが登場した。深層学習で標準的に用いられる AdaGrad (Duchi et al., 2011) や RMSprop (Tieleman and Hinton, 2012), ADAM (Kingma and Ba, 2015) などの最適化技術はこの時期に発達した。

そして 2014 年頃からは，ポスト・パターン認識の時代である。画像や音声の認識タスクは既に飽和状態となり，巨大なニューラルネットを作製する研究は，莫大な予算と，潤沢かつ良質なデータ，そして人材の結集なくしては遂行が困難になった。一方，世間では人工知能が人類を超えるという技術的特異点（singularity）が話題となった。機械学習では比較的珍しいことだが，Nature のような商業誌にも深層学習が登場するようになり (Mnih et al., 2015; LeCun et al., 2015; Silver et al., 2016)，悪夢のような絵を生成する Inceptionism (Mordvintsev et al., 2015) や，人類最強の棋士イ・セドル氏に勝利した AlphaGo (Silver et al., 2016) のように，見た目に分かりやすいアウトプットが続いたことが背景にあると考えられる。基礎研究では，Bayes 法や統計物理学の背景をもつ研究者が参入して，生成モデルによる表現学習 (Bengio et al., 2014; Alain et al., 2016; Kingma and Welling, 2014; Goodfellow et al., 2014; Sohl-Dickstein et al., 2011, 2015) が再び脚光を浴びるようになった。そして，生成モデルやリカレントニューラルネットを用いたデータ生成 (Boulanger-Lewandowski et al., 2012; Graves, 2013; Goodfellow et al., 2014; Gatys et al., 2015; Gregor et al., 2015; Radford et al., 2016; van den Oord et al., 2016) や，ゲームで人間に勝つことを目指す強化学習 (Mnih et al., 2013, 2015)，機械翻訳（Neural Machine Translation; NMT）などに応用される attention (Mnih et al., 2014; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015) などが，深層学習の新たな挑戦的タスクとして登場した。バイナリ重み (Courbariaux et al., 2015) や ハッシュ (Chen et al., 2015) のように，学習時の誤差信号は著しく単純化されていても学習できることが分かってきた。

Why Now?

深層学習はなぜ最近までできなかったのだろうか。ニューラルネットを深層構造にする方が情報処理能力が向上するという仮説は，第二次ブームの頃からあった (Rumelhart et al., 1987; Hinton, 1989; White, 1990; Utgoff and Stracuzzi, 2002)。深層学習ブームの火付け役である Hinton

自身も階層的な情報表現の重要性を強く訴えていたし，Bengio もまたスパース分散表現の熱烈な信奉者であった。また，畳み込みネットワークを用いた深層学習は，既に Fukushima (1980) や LeCun et al. (1998) が実現させていたことも特筆すべきことである。

計算機の発達とデータの氾濫を理由に挙げる研究者は多い。深層学習が登場した 2006 年と，第二次ブームにあたる 1990 年前後との間には，15 年程度の時間差がある。Moore の法則に基づいて概算すると，CPU の性能は少なくとも $2^{15/1.5} \approx 1,000$ 倍ほど違う。また，1998 年に登場して以来標準的なベンチマークとして用いられている MNIST データセットは，28×28 のグレースケールの線画（0 から 9 までの手書きのアラビア数字）が 6 万枚という規模であるのに対して，2012 年の AlexNet で用いられた ImageNet データセットは 224×224 に縮小された 1,000 クラスのカラー画像 100 万枚であるから，単純にピクセル数を比較しても 1,000 倍オーダーの差がある。

興味深いことに，圧倒的な量の差をもってそれまでできなかったことができるようになると，理論も変化した。90 年代後半には，なぜ学習できないかを説明する理論が多かったのに対して，今日では，なぜ学習できるかを説明する理論が増えてきた。例えば，ニューラルネットの学習問題は高次元の非凸最適化問題なので，局所解に陥るためにうまく解けないという説明がある。Dauphin et al. (2014) は，次元が高い場合には，大多数の危点は鞍点であることを示し，うまく解けないのは Newton 法の使い方に問題がある可能性を指摘した。また，ニューラルネットの学習問題は学習時間が NP 困難なので，うまく解けない (Blum and Rivest, 1992) というように，計算複雑性理論による説明もある。これに対して Livni et al. (2014) は，（計算複雑性理論の延長である）統計的学習理論と，深層学習という現実とのギャップを埋めるべく，確率的勾配法の解析に取り組んでいる。

2.1.2 Why Deep?

そもそもニューラルネットは，中間層が一層あれば，任意の関数を近似できるほど高い表現能力を持つ。従って，関数近似や表現能力という観点では，深層ニューラルネットは単に冗長なだけですらある。疑い深い研究者の中には，深層ニューラルネットと同等以上の性能を示す浅いニューラルネットを構成してみせた人もあるほどである (Ba and Caruana, 2014)。

しかし，浅いニューラルネットに拘る研究者は，今や少数派である。

なぜ深層構造が良いのかについての研究は，表現能力の効率性を数学的な命題として示すものが多い (Håstad, 1986; Delalleau and Bengio, 2011; Montufar et al., 2014; Telgarsky, 2016; Eldan and Shamir, 2016; Cohen et al., 2016)。中間層を重ねることで，中間層素子の発火パターンの組み合わせや，さらにそれらの高次の組み合わせが扱えるようになるので，同じ数の中間層素子であれば，並列に浅く並べるよりも，直列に深く並べる方が，表現できるパターンの数は指数的に増す。従って，深層ニューラルネットの表現能力は，浅いニューラルネットと比較して指数関数的に効率が良く，パラメータ空間の次元も節約できる (Bengio et al., 2006a; Montufar et al., 2014)。単に表現能力ではなく，学習能力を評価した研究も多い。例えば Arora et al. (2014) はサンプル複雑性，Giryes et al. (2015a) は深層ニューラルネットの stability，Neyshabur et al. (2015) は深層ニューラルネットの Rademacher 複雑性を評価した。

表現能力や学習能力を評価する研究では，回路や多項式関数などを用いて，深層ニューラルネットと比較して，浅いニューラルネットで近似するためには中間層素子の数が指数関数オーダーで多く必要となるような状況を構成して，深層ニューラルネットの優位性を主張する。勿論，次元の呪いを思い出せば，次元削減は歓迎すべきことではある。一方で，この類の説明では，例として構成された回路や関数が現実に登場する可能性について説明を欠いていることに注意すべきである。

2.1.3　何を表現しているか

ある人が「お婆さん」を認識する時，その人の脳内では，お婆さんに反応する唯一の神経細胞が発火しているのだろうか。それとも，複数の神経細胞の発火パターンが，お婆さんに対応しているのだろうか。前者を局所表現と言い，後者を分散表現と言う。Hinton や Bengio は，分散表現の信奉者である。生物の系統図のように，体系化や階層化を通じて，知識はコンパクトに表現できる。同様に，中間層を重ねることで，発火パターンは組み合わせ的に複雑さを増し，中間層素子を一層に並べるよりも，効率的に情報を表現できる。

特徴量写像

深層ニューラルネットは，入力に近い方を特徴量写像，出力に近い方を予測器に分けて解釈するのが基本的である。本研究でもこの考え方に則って解析する。Cho and Saul (2009) や Montavon et al. (2011), Jawanpuria (2015) などは，深層ニューラルネットの特徴量写像から得られた中間表現に主成分分析をかけることで，カーネル主成分分析としての性能を評価した。Yosinski et al. (2014) は中間層特徴量がどの程度，転移学習に利用できるかを調べることで，層が深まるに連れてドメイン特異性が高くなることを発見した。また，Szegedy et al. (2014); Yosinski et al. (2015) は，学習済ニューラルネットに対して，訓練データとかけ離れた入力を与えると，ニューラルネットが誤動作することを実験的に示した。Sonoda and Murata (2016) でも，この立場からデノイジング・オートエンコーダーの特徴づけを行っている。

ランダム説

深層ニューラルネットのパラメータ数はあまりにも膨大なので，各パラメータはほとんどランダムでもネットワーク全体としては何らかの機能を果たしていると考えるのはもっともなことである。Saxe et al. (2011) や Cambria et al. (2013) は，パラメータがランダムであっても判別に有利な特徴量が得られることを実験的に報告をしている。Giryes et al. (2015b) は，圧縮センシングのアナロジーにより，パラメータが正規乱数であっても，活性化関数が ReLU であれば，似ている入力と異なる入力とを分離するような作用があることを示した。一方 Duvenaud et al. (2014) は，ランダムな特徴写像の合成を繰り返すと，カオス的に振舞うことを示した。一口にランダムといっても，ランダムネスの入れ方次第では互いに矛盾するような結論が導かれることに注意すべきである。

無限層

第 6 章では無限層のニューラルネットを解析する。勿論，無限層に言及するのは本研究が初めてのことではない。特定のカーネル関数の合成写像を解析的に計算して深層カーネルを構成する方法 (Cho and Saul, 2009) や，パラメータがランダムな場合の解析 (Saxe et al., 2011; Duvenaud et al., 2014; Giryes et al., 2015b)，ニューラルネットを Markov 連鎖や力

学系とみなす解析 (Sohl-Dickstein et al., 2015; Sonoda and Murata, 2016) などは，個別の中間層を特徴づける研究であり，無限層まで合成した結果を調べることができる。

2.1.4 ReLU

深層学習では，活性化関数として ReLU (Rectified Linear Unit) を用いるのが標準的である (Nair and Hinton, 2010; Glorot et al., 2011; Goodfellow et al., 2013; Dahl et al., 2013; Maas et al., 2013)。従来用いられてきたシグモイド関数や RBF と比較して，ReLU は学習を加速し，学習結果をスパースにする効果があることが経験的に知られている (Glorot et al., 2011; Jarrett et al., 2009; Krizhevsky et al., 2012; Zeiler et al., 2013; Maas et al., 2013)。シグモイド関数 σ を用いる場合，入力信号 x の値が閾値 b から乖離していると，出力 $\sigma(x-b)$ の入力 x に対する感度は鈍化する。そのため，学習時の誤差信号は層が深まるに連れて消滅 (vanishing gradient) しやすい。一方 ReLU $(\cdot)_+$ を用いる場合は，出力 $(x-b)_+$ の入力に対する感度は一定である。従って，学習時の誤差信号も減衰することなく伝わり，学習が素早く進むようになるためである (Glorot et al., 2011)。

2.1.5 デノイジング・オートエンコーダー

Vincent et al. (2008) は古典的なオートエンコーダー (Bourlard and Kamp, 1988; Baldi and Hornik, 1989) の修正版としてデノイジング・オートエンコーダー (denoising autoencoder; DAE) を導入した。古典的なオートエンコーダーは恒等写像 $x \mapsto x$ だが，DAE はわざとノイズを加えた入力 $\tilde{x}$ から雑音を除いた信号を取り出す雑音除去写像 $\tilde{x} \mapsto x$ として訓練される。当初，ノイズはロバスト性を高める目的で取り入れられた。

DAE の理論づけには，現時点で少なくとも五つないし六つの観点がある。まず，DAE はデータが分布している多様体を学習しているとする多様体学習説 (Rifai et al., 2011; Rifai and Dauphin, 2011; Alain and Bengio, 2014)，次に，データの生成モデルを学習しているとする生成モデル説 (Vincent et al., 2010; Bengio et al., 2013b, 2014), 入力データの情報損失が最小となるようなデータ表現を学習しているとする infomax 説 (Vincent et al., 2010) や, スパースな表現を獲得しているとするスパース説 (Arpit et al., 2016), 局所解を避けるような初期値になっているという

学習ダイナミクスについての報告 (Erhan et al., 2010), そして特定のエネルギー関数によるスコアマッチングがDAEの手続きと等価であるという発見 (Vincent, 2011; Kamyshanska and Memisevic, 2013, 2015) である。はじめの三つの観点はVincentの原論文の中でも触れられている。銘々の理屈付けで，DAEが学習しているものはそれぞれ以下のとおりである：多様体学習説ではデータが分布している多様体，生成モデル説ではデータを生成する確率分布のパラメータ，infomax説では入力データの情報損失が最小となるようなデータ表現，スパース説ではスパースなデータ表現，学習ダイナミクスでは局所解を避けるような初期値，スコアマッチングではデータ分布自体である。なお，これらは主に浅いDAEの解析である。

特に，最後のスコアマッチング説は，DAEが入力データの分布に関する情報を（少なくとも理論的には）全て保持していることを決定づけたという点で画期的であった。この発見を転機として，DAEの研究はそれまでの積層（stack）から，生成モデルを用いてデータ分布を積極的に推定するスタイルへとシフトしていった (Larochelle and Murray, 2011; Bengio et al., 2013b, 2014)。

生成モデルは，当初オートエンコーダーより優勢であったRBMやDBN, DBMとの類似性が高いこともあり，ベイズ計算を得意とする研究者が多く参入してきた。Kingma and Welling (2014) の変分オートエンコーダー（variational AE; VAE）や，Sohl-Dickstein et al. (2011, 2015) の最小確率流（minimum probability flow），Goodfellow et al. (2014) の敵対的ネットワーク（generative adversarial network; GAN）など，洗練されたアルゴリズムが矢継ぎ早に発表された。GANはゲーム理論に基づく方法であり，actor-criticとの関係も指摘されている。さらに，生成モデルの特徴を活かして，半教師付き学習 (Kingma et al., 2014; Rasmus et al., 2015) や，データ生成 (Radford et al., 2016) などの新しいタスクが登場した。

2.1.6 畳み込みネットワーク

畳み込みネットワークは，視覚系や信号処理・画像処理のコミュニティが参入してきたこともあり，この数年間でもっとも研究が進んだモデルと考えられる。従って，本研究で直接取り扱うことはないが，理論的に重要と思われる結果をいくつか列挙しておく。Lee et al. (2008) は，画像データに対してスパース制約付きRBMによる特徴抽出を行うと，二層目

に視覚系の V2 野と同様の特徴量が抽出されることを発見した。Wibisono et al. (2010); Mroueh et al. (2015); Anselmi et al. (2015) は，特徴量写像に不変性を要求すると畳み込み構造が現れることを示した。Saxe et al. (2011) は，畳み込み構造があればネットワークの重みはランダムでも機能することを指摘した。Bruna and Mallat (2013) は畳み込みネットワークを散乱変換（scattering transform）という変換としてモデル化し，プーリングが不変性の獲得に必要であることを指摘した。散乱変換は Wiatowski et al. (2016) が一般化する方向を検討している。Szegedy et al. (2015) や Srivastava et al. (2015), He et al. (2016) では，インセプション構造やスキップ接続を用いることが深化を成功させるために不可欠であることを示した。

2.2　浅いニューラルネットの理論

本研究の出発点となる積分表現理論は Murata (1996) に始まる。積分表現理論は，ニューラルネットの表現能力（関数近似能力）を調べる研究が盛んに行われていた時代に登場した。関数近似理論の研究者である Kůrková (2012) に曰く，積分変換の離散化を経由して関数を近似するテクニックは，関数近似理論の分野では古くから用いられてきた。ニューラルネットを積分変換の離散化とみなす例は，Irie and Miyake (1988) による Fourier 変換や，Carroll and Dickinson (1989) や Ito (1991) による Radon 変換などに始まる。Candès (1998) の用語に立脚すれば，積分表現は双対リッジレット変換である。リッジレット変換は Radon 変換とウェーブレット変換の合成変換とみなせることから，幾何学的な解釈性に優れており，調和解析の一分野を築くまでに発展した。本節では，積分表現理論を軸として深層学習以前のニューラルネットの理論を整理する。

2.2.1　関数近似理論

ニューラルネットは万能関数近似器（universal approximator）である。つまり，中間層素子の数を無限に増やすことで，任意の関数を近似できる。厳密には，ニューラルネットは二乗可積分関数の空間 $L^2(\mathbb{R}^m)$ や，連続関数の空間 $C(\mathbb{R}^m)$ にコンパクト開位相を入れた空間などで稠密である。

ニューラルネットの万能性に初めて言及したのは，おそらく Hecht-Nielsen (1987) である。Hecht-Nielsen は Kolmogorov-Arnold の表現定

理 (Kolmogorov, 1956a; Arnold, 1957; Sprecher, 1965) により，ニューラルネットが任意の関数を近似しうることを指摘した。ただし表現定理の形式は浅いニューラルネットよりも複雑であり，完全な証明にはならない。

浅いニューラルネットの万能性は，1988 年から 1989 年にかけて同時多発的に異なる方法で証明された。Cybenko (1989) は Hahn-Banach の拡張定理を用いたが，非構成的であるためにそれ以上先に進むことは難しかった。Hornik et al. (1989) は Stone-Weierstrass の定理，Irie and Miyake (1988) と Funahashi (1989) は Fourier 変換に帰着して万能性を示した。これらは構成的ではあるものの，ニューラルネットを多項式や三角関数に翻訳する方法なので，リッジレット解析ほどの解釈性はない。Carroll and Dickinson (1989) は Radon 変換に帰着して万能性を示した。Radon 変換はリッジレット変換と非常に近いが，Carroll and Dickinson の証明は手続き的であり，リッジレット解析ほどの解釈性はない。

活性化関数が有界でない場合の万能性は，まず Mhaskar and Micchelli (1992) が B-spline を使って示した。続く Leshno et al. (1993) の証明は近似単位元を使っており，簡潔なだけでなく実解析的な示唆に富んでいる。詳細は Pinkus (1999) を見よ。

表 2.1: $g(x) := \sum_{j=1}^{J} c_j \eta(a_j \cdot x - b_j)$ による $f(x)$ の近似可能性

	f	η	位相	証明方針
Irie and Miyake 1988	L^1	L^1	各点	Fourier 反転公式
Cybenko 1989	C	シグモイド	広義一様	Hahn-Banach
Hornik+ 1989	C	シグモイド	広義一様	Stone-Weierstrass
Funahashi 1989	C	有界連続かつ単調増加	広義一様	Fourier 反転公式
Carroll and Dickinson 1989	$\mathcal{D}$	tanh	L^2	Radon 反転公式
Mhaskar and Micchelli 1992	L^p	$\mathcal{S}_0'$	L^p	B-spline
Leshno+ 1993	C	$\mathcal{S}_0'$	広義一様	近似単位元

万能性が明らかになると，次は近似の効率が焦点となった。つまり，中間層素子数 n や入力の次元 m に対して，近似誤差がどのようなオーダーで減衰するかという評価である。ここで，近似誤差は推定誤差ではないことに注意せよ。従って，サンプルサイズ S やデータの分布 π は登場しない。Barron (1993) は，Fourier 変換を用いて，$O(1/\sqrt{n})$ という評価を導いた。これは，先行する Maurey (Pisier, 1981) と Jones (1992) の名前

を合わせて，Maurey-Jones-Barron 評価（MJB bound）と呼ばれている。MJB 評価は，中間層素子一個のニューラルネットから始めて，近似対象にもっとも近づくように中間層素子を追加していく貪欲法を理論的に行うことで導かれる。このアイデアは実際の学習アルゴリズムに忠実で素朴だが，強力である。今日でも最適化理論の文脈で用いられる。

MJB 評価は二つの側面で画期的であった。まず，$O(1/\sqrt{n})$ は多項式近似や三角関数近似あるいはスプライン近似よりも速い。これは，ニューラルネットの中間層素子が可変基底として機能するためである。つまり，多項式や三角関数のように基底関数系を固定すると，必ず近似に不利な点が存在するが，ニューラルネットの中間層素子は基底自体が適応できるので，効率的に部分空間が張れるためである (Barron, 1993)。そして，$O(1/\sqrt{n})$ には次元 m が含まれない。これは，「ニューラルネットは次元 m に影響されない」ことの根拠として持て囃された。もっとも，これは次元 m が近似対象の滑らかさ s と相殺しているためである。つまり，次元が上がるに連れて関数クラスが縮小しているのである。

Mhaskar (1996) は，関数の滑らかさ s を考慮した Jackson 型評価 $O(n^{-s/m})$ を導いた。Petrushev (1998) は Gegenbauer 多項式を用いて代数的に離散化する場合の誤差を評価した。Kůrková and Sanguineti (2001); Kůrková (2012) は MBJ 評価の精密化として変動ノルムによる評価を導いた。詳細は Kainen et al. (2013) を参照せよ。

2.2.2　学習能力の評価

ニューラルネットのパラメータ空間は高次元であるにも関わらず学習できたので，あたかもニューラルネットは次元の呪いとは無縁であるかのように持て囃された。統計的学習理論の大家である Vapnik (2006) が述懐するところでは，この問題に説明を付けるために，Fisher 統計学に依らない学習理論を模索し続け，ついに統計的学習理論にたどり着いたという。

Poggio and Girosi (1990); Girosi et al. (1995) はリッジ正則化付きバックプロパゲーションの停留点が Green 関数を用いて書けることを利用して，正則化ネットワークを開発し，汎化誤差を評価した。Bartlett (1998); Niyogi and Girosi (1999) はニューラルネットのサンプル複雑性を計算した。Bartlett and Mendelson (2002) や Bousquet and Elisseeff (2002) はニューラルネットの複雑性を計算するために Rademacher 複雑性や

stability を開発した。Cesa-Bianchi et al. (2004) は SGD のサンプル複雑性を評価した。最近では Bach (2014) が凸型ニューラルネット (Bengio et al., 2006b) の Rademacher 複雑性を評価している。

2.2.3 リッジレット解析

Murata (1996) や Candès (1999, 1998), Rubin (1998) は，ほぼ同時にリッジレット変換を発明した (Donoho, 1999, 2001; Rubin, 2004; Starck et al., 2010)。積分表現理論にリッジレット（ridgelet）という名前を付けたのは Candès (1998) である。Candès と Donoho は画像のような多次元信号に対する新しいウェーブレット理論を追究しており，やがて多重解像度幾何解析（Geometric Multiscale Analysis; GMA）という名称で組織的に研究を発展させた (Donoho, 2002)。Rubin は主に実解析的な側面に着目しており，Calderón 再生公式の一般化として発展させた。

再構成公式 $\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f = f$ も含めたリッジレット変換 $\mathscr{R}_\psi$ の枠組みは，四つの関数クラスによって特徴づけられる。すなわち，リッジレット変換の定義域（domain）$\mathcal{X}(\mathbb{R}^m)$ と値域（range）$\mathcal{Y}(\mathbb{Y}^{m+1})$, およびリッジレット変換と双対リッジレット変換に用いられる二つのリッジレット関数のクラス $\mathcal{Z}(\mathbb{R})$ と $\mathcal{W}(\mathbb{R})$ である。

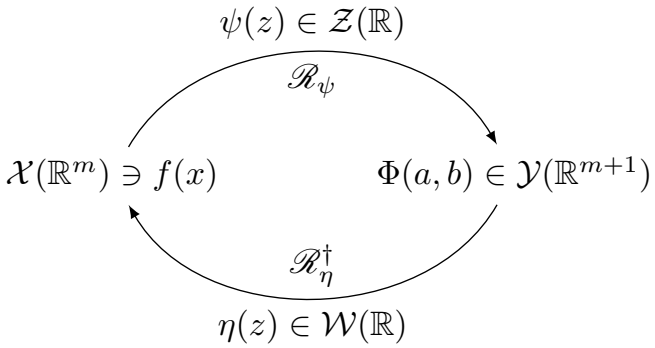


図 2.2: リッジレット解析に纏わる四つのクラスの関係

Murata (1996) による積分変換 T は $\mathcal{Z}, \mathcal{W} \subset L^1 \cap L^2$ の場合に相当する。ただし，この他に許容条件を含めた複数の付帯条件が課されている。Candès (1999, 1998) は $\mathcal{Z} = \mathcal{W} \subset \mathcal{S}$（急減少関数）の場合にリッジレット解析を展開した。Candès の流儀では，リッジレット関数と双対リッジレット関数には同一の関数を用いる。Rubin (1998) は $\mathcal{Z}, \mathcal{W}$ が Borel 測

度の場合にリッジレット変換だけでなく Radon 変換や k-plane 変換まで包括する Calderón 再生公式を示した。Kostadinova et al. (2014, 2015) はリッジレット変換の定義域が Lizorkin 超関数 $\mathcal{X} = \mathcal{S}_0'$ にまで拡張できることを示した。これは今日知られている時点で最大の定義域だが，リッジレット関数のクラスを Lizorkin 関数 $\mathcal{W} = \mathcal{Z} = \mathcal{S}_0 \subset \mathcal{S}$ に制限するという犠牲の上に実現されている。Sonoda and Murata (2015) では，$\mathcal{Z} \subset \mathcal{S}$ にとることで $\mathcal{W} \subset \mathcal{S}'$（緩増加超関数）まで拡張できることを示した。

表 2.2: リッジレット解析が展開できるクラス

	$\mathcal{X}$	$\mathcal{Y}$	$\mathcal{Z}$	$\mathcal{W}$
Murata 1996	$L^1 \cap L^p$ $(p \in [1, \infty])$	-	$L^1 \cap L^2$	$L^1 \cap L^2$
Candès 1998	L^1, L^2	-	$\mathcal{S}$	$\mathcal{S}$
Rubin 1998	L^2	-	Borel 測度	Borel 測度
Kostadinova+ 2014	$\mathcal{S}_0'$	-	$\mathcal{S}_0$	$\mathcal{S}_0$
Sonoda & Murata 2015	L^1, L^2	$\mathcal{S}'$	$\mathcal{S}$	$\mathcal{S}'$

第3章　数学的準備

実解析や関数解析については (猪狩惺, 1996; 柴田良弘, 2006; Rudin, 1991; Brezis, 2011; Yosida, 1995), 超関数については (垣田高夫, 1999; Schwartz, 1966; Trèves, 1967), Lizorkin 超関数や Besov 空間については (Yuan et al., 2010; Holschneider, 1995; 澤野嘉宏, 2011) を参考にした。Fourier 解析については (Grafakos, 2008), Radon 変換については (Helgason, 2011; Natterer, 2008; Quinto, 2006; Kuchment, 2014), ウェーブレット変換については (Daubechies, 1992; Mallat, 2009; Holschneider, 1995) を参考にした。拡散方程式や発展方程式については (伊藤清三, 1979; 小川卓克, 2013), 最適輸送理論と Wasserstein 幾何学については (Villani, 2009; Ambrosio et al., 2008; 桑江一洋 et al., 2015) を参考にした。

3.1　基礎空間

基本的な空間の位相と測度を整理する。

m 次元 Euclid 空間 $\mathbb{R}^m$ には Euclid 内積から定まる標準位相と Lebesgue 測度を入れる。関数ノルムと区別するため，$x \in \mathbb{R}^m$ の Euclid ノルムは単に $|x|_2$ ないし $|x|$ と書く。

$(m-1)$ 次元単位球面 $\mathbb{S}^{m-1}$ は，$\mathbb{R}^m$ の部分集合

$$\mathbb{S}^{m-1} = \{x \in \mathbb{R}^m \mid |x|_2 = 1\},$$

に $\mathbb{R}^m$ の誘導位相を入れたものとする。Lebesgue 測度から誘導された表面測度を $\mathrm{d}\sigma$ とする。特に断りのない限り，球面測度 $\mathrm{d}u$ といえば一様確率測度 $\mathrm{d}u := \Gamma(m/2)/2\pi^{m/2}\mathrm{d}\sigma$ を表すものとする。すなわち，以下を満たす

$$\int_{\mathbb{S}^{m-1}} \mathrm{d}u = 1.$$

半直線 $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x > 0\}$ には，$\mathbb{R}$ の誘導位相を入れる。半直線は通常の積によって位相群となるので，Haar 測度 $\mathrm{d}x/x$ が自然である。ただしウェーブレット変換やリッジレット変換の台空間の一部として扱う場合には，別の測度 $\mathrm{d}x/x^s$ を入れることもある。Haar 測度に限り，以下のスケール不変性が成り立つ

$$\int_{\mathbb{R}_+} \left|\psi\left(ar\right)\right|^p \frac{\mathrm{d}r}{r} = \int_{\mathbb{R}_+} \left|\psi\left(r\right)\right|^p \frac{\mathrm{d}r}{r}, \quad a > 0.$$

半直線の測度については 付録 A.1 を参照せよ。

半空間 $\mathbb{H} := \mathbb{R}_+ \times \mathbb{R}$ には，$\mathbb{R}^2$ の誘導位相を入れる。測度は Haar 測度どうしの直積測度である $\mathrm{d}x\mathrm{d}y/x$ を用いるが，文脈に応じて使い分ける。

3.2　関数と超関数

Euclid 空間上の関数および超関数の定義と位相を整理する。関数空間の記号法は (猪狩惺, 1996) および (Schwartz, 1966) に従う。

以下の包含関係は Schwartz (1966) による。

$$\begin{array}{lccccccc}
\text{(関数)} & \mathcal{D}(\mathbb{R}^m) & \subset & \mathcal{S}(\mathbb{R}^m) & \subset & \mathcal{O}_{\mathcal{M}}(\mathbb{R}^m) & \subset & \mathcal{E}(\mathbb{R}^m) \\
 & \cap & & \cap & & \cap & & \cap \\
\text{(超関数)} & \mathcal{E}'(\mathbb{R}^m) & \subset & \mathcal{O}'_{\mathcal{C}}(\mathbb{R}^m) & \subset & \mathcal{S}'(\mathbb{R}^m) & \subset & \mathcal{D}'(\mathbb{R}^m)
\end{array}$$

3.2.1　Euclid 空間上の関数

多項式関数の空間を $\mathcal{P}(\mathbb{R}^m)$ と書く。$\mathcal{P}(\mathbb{R}^m)$ には後述する緩増加超関数 $\mathcal{S}'$ の位相を入れる。

連続関数の空間を $C(\mathbb{R}^m)$ と書く。$C(\mathbb{R}^m)$ には広義一様収束による位相（コンパクト開位相）を入れる。

無限回微分可能関数（滑らかな関数, smooth functions）の空間を $C^\infty(\mathbb{R}^m)$ または $\mathcal{E}(\mathbb{R}^m)$ と書く。特に $\mathcal{E}(\mathbb{R}^m)$ と書くときは，以下で説明する位相が入っている。すなわち，$\mathcal{E}(\mathbb{R}^m)$ の点列 u_j が $\mathcal{E}$ の位相で u に収束する

$$u_j \to u \quad \text{in } \mathcal{E}$$

とは，任意のコンパクト集合 K と任意の多重指数 $\alpha \in \mathbb{N}_0^m$ に対して，

$$\partial^\alpha u_j \to \partial^\alpha u \quad \text{uniformly on } K$$

となることをいう。

コンパクト集合 K に台を持つ滑らかな関数の全体を $C^\infty(K)$ と書く。台がコンパクトかつ滑らかな関数の空間を $C_c^\infty(\mathbb{R}^m)$ または $\mathcal{D}(\mathbb{R}^m)$ と書く。すなわち，任意のコンパクト集合 K に対する合併の記号 $\bigcup_K$ を用いて以下のように書ける

$$\mathcal{D}(\mathbb{R}^m) := \bigcup_K C^\infty(K).$$

$\mathcal{D}(\mathbb{R}^m)$ における収束（$u_j \to u$ in $\mathcal{D}$）とは，任意の多重指数 $\alpha \in \mathbb{N}_0^m$ に対して，

$$\partial^\alpha u_j \to \partial^\alpha u \quad \text{uniformly}$$

となることをいう。

急減少関数の空間を $\mathcal{S}(\mathbb{R}^m)$ と書く。滑らかな関数 $u \in C^\infty(\mathbb{R}^m)$ のセミノルムを

$$\rho_{\alpha,\beta}(u) := \sup_x |x^\alpha \partial^\beta u(x)|, \quad \alpha, \beta \in \mathbb{N}_0^m$$

とおく。滑らかな関数 u が急減少関数であるとは，任意の多重指数 $\alpha, \beta \in \mathbb{N}_0^m$ に対してセミノルム $\rho_{\alpha,\beta}(u)$ が有界となることをいう

$$\mathcal{S}(\mathbb{R}^m) := \{u \in C^\infty(\mathbb{R}^m) \mid \forall \alpha, \beta \in \mathbb{N}_0^m : \rho_{\alpha,\beta}(u) < \infty\}.$$

$\mathcal{S}(\mathbb{R}^m)$ にはセミノルムによる位相を入れる。すなわち，$u_j \to u$ in $\mathcal{S}$ とは，任意の多重指数 $\alpha, \beta \in \mathbb{N}_0^m$ に対して，

$$\rho_{\alpha,\beta}(u_j - u) \to 0,$$

となることをいう。

緩増加関数の空間を $\mathcal{O}_\mathcal{M}(\mathbb{R}^m)$ と書く。滑らかな関数 $u \in \mathcal{E}(\mathbb{R}^m)$ が緩増加関数であるとは，任意の急減少関数 $f \in \mathcal{S}(\mathbb{R}^m)$ との積 fu が再び急減少関数になることをいう

$$\mathcal{O}_\mathcal{M}(\mathbb{R}^m) := \{u \in \mathcal{E}(\mathbb{R}^m) \mid \forall \phi \in \mathcal{S}(\mathbb{R}^m) : u\phi \in \mathcal{S}(\mathbb{R}^m)\}.$$

$\mathcal{O}_{\mathcal{M}}(\mathbb{R}^m)$ には $\mathcal{E}$ の位相を誘導する。ある関数が緩増加関数であるかどうかを調べるには，定義に従うよりも，以下の同値な条件を調べるほうが簡単である (Schwartz, 1966; Grafakos, 2008)。すなわち，$u \in \mathcal{E}(\mathbb{R}^m)$ が緩増加関数であることは，任意の多重指数 $\alpha \in \mathbb{N}_0^m$ に対して，ある多項式 $p \in \mathcal{P}(\mathbb{R}^m)$ があって，$|\partial^\alpha u| \leqslant p$ が成り立つことと同値である

$$\mathcal{O}_{\mathcal{M}}(\mathbb{R}^m) = \{u \in \mathcal{E}(\mathbb{R}^m) \mid \forall \alpha \in \mathbb{N}_0^m \, \exists p \in \mathcal{P}(\mathbb{R}^m) : |\partial^\alpha u| \leqslant p\}.$$

言い換えれば，任意の導関数 $|\partial^\alpha u(x)|$ の無限遠 $|x| \to \infty$ における増大速度が高々多項式程度（at most polynomial growth）のとき，u は緩増加関数である。

Lizorkin 関数の空間を $\mathcal{S}_0(\mathbb{R}^m)$ と書く。急減少関数 $u \in \mathcal{S}(\mathbb{R}^m)$ が Lizorkin 関数とは，任意の指数のモーメントが消滅することをいう

$$\mathcal{S}_0(\mathbb{R}^m) := \left\{ u \in \mathcal{S}(\mathbb{R}^m) \;\middle|\; \forall \alpha \in \mathbb{N}_0^m : \int x^\alpha u(x) \mathrm{d}x = 0 \right\}.$$

このような関数は無数に存在する。例えば，$u_0 \in \mathcal{S}(\mathbb{R}^m)$ として台が原点を含まないものをとると，その Fourier 変換 $\widehat{u_0}$ は $\mathcal{S}_0(\mathbb{R}^m)$ の元である。定義より，$\mathcal{S}_0(\mathbb{R}^m)$ は $\mathcal{S}(\mathbb{R}^m)$ の閉部分空間である。

3.2.2　Euclid 空間上の超関数

$\mathcal{A}(\mathbb{R}^m)$ は適当な位相が入った関数空間とする。つまり，$\mathcal{A}(\mathbb{R}^m)$ は線形位相空間（topological vector space）である。$\mathcal{A}(\mathbb{R}^m)$ の位相的双対空間（topological dual space）$\mathcal{A}'(\mathbb{R}^m)$ とは，$\mathcal{A}(\mathbb{R}^m)$ の有界線形汎関数 $f : \mathcal{A}(\mathbb{R}^m) \to \mathbb{C}$ の全体である

$$\mathcal{A}'(\mathbb{R}^m) := \{f : \mathcal{A}(\mathbb{R}^m) \to \mathbb{C} \text{ bounded linear functional}\}.$$

$\mathcal{A}'(\mathbb{R}^m)$ には（一般に複数の位相が入れられるが）汎弱位相（弱 $*$ 位相）を入れる。すなわち，$\mathcal{A}'(\mathbb{R}^m)$ の位相的双対 $\mathcal{A}''(\mathbb{R}^m)$ が再び $\mathcal{A}(\mathbb{R}^m)$ と同相になるような位相のうち，最も粗い位相である。

Schwartz 超関数の空間 $\mathcal{D}'(\mathbb{R}^m)$ は，$\mathcal{D}(\mathbb{R}^m)$ の位相的双対空間に汎弱位相を入れた空間である。単に超関数といえばこの空間の元を指す。$C_c^\infty(\mathbb{R}^m)$ 上の線形汎関数 T が超関数であること（$T \in \mathcal{D}'(\mathbb{R}^m)$）を示すには，以下の同値な条件 (猪狩惺, 1996, 定理 7.1) を調べる方が容易である

$$\forall K \text{ compact } \exists n \geqslant 0 \text{ s.t. } |\langle T, u \rangle| \lesssim p_{n,K}(u), \quad u \in C^\infty(K),$$

ただし

$$p_{n,K}(u) := \sup_{|\alpha|\leqslant n} \sup_{x\in K} |\partial^\alpha u(x)|$$

とする[1]。また，超関数 $T \in \mathcal{D}'(\mathbb{R}^m)$ は，コンパクト集合 K 上では可積分関数の導関数として表されることが知られている（Schwartz 超関数の構造定理 (猪狩惺, 1996, 定理 7.3)）

$$\exists f \in L^2(K), \alpha \in \mathbb{N}_0^m \text{ s.t. } \langle T, u\rangle = \int_K f(x)\partial^\alpha u(x)\mathrm{d}x, \quad u \in C^\infty(K).$$

緩増加超関数の空間 $\mathcal{S}'(\mathbb{R}^m)$ は，$\mathcal{S}(\mathbb{R}^m)$ の位相的双対空間に汎弱位相を入れた空間である。$\mathcal{S}(\mathbb{R}^m)$ 上の線型汎関数 T が緩増加超関数であること（$T \in \mathcal{S}'(\mathbb{R}^m)$）の必要十分条件 (猪狩惺, 1996, 定理 7.7) は，以下で与えられる

$$\exists n \geqslant 0 \text{ s.t. } |\langle T, u\rangle| \lesssim \sum_{|\alpha|,|\beta|\leqslant n} \rho_{\alpha,\beta}(u), \quad u \in \mathcal{S}(\mathbb{R}^m).$$

台がコンパクトな超関数の空間 $\mathcal{E}'(\mathbb{R}^m)$ は，$\mathcal{E}(\mathbb{R}^m)$ の位相的双対空間に汎弱位相を入れた空間である。

急減少超関数の空間を $\mathcal{O}'_C(\mathbb{R}^m)$ と書く。緩増加超関数 $u \in \mathcal{S}'(\mathbb{R}^m)$ が急減少超関数であるとは，任意の関数 $f \in \mathcal{D}(\mathbb{R}^m)$ との畳み込み $f * u$ が急減少関数になることをいう

$$\mathcal{O}'_C(\mathbb{R}^m) := \{u \in \mathcal{S}'(\mathbb{R}^m) \mid \forall \phi \in \mathcal{D}(\mathbb{R}^m) : u * \phi \in \mathcal{S}(\mathbb{R}^m)\}.$$

$T \in \mathcal{D}'(\mathbb{R}^m)$ が急減少超関数であること（$T \in \mathcal{O}'_C(\mathbb{R}^m)$）は，以下の各条件とそれぞれ同値である (Schwartz, 1966, Ch. 7 Th. 9)

(i) $\forall k \in \mathbb{N}_0 : (1 + |x|^2)^{k/2} T$ is bounded

(ii) $\forall u \in \mathcal{D}(\mathbb{R}^m) : T * u \in \mathcal{S}(\mathbb{R}^m)$.

Lizorkin 超関数の空間 $\mathcal{S}'_0(\mathbb{R}^m)$ は，$\mathcal{S}_0(\mathbb{R}^m)$ の位相的双対空間に汎弱位相を入れた空間である。Lizorkin 超関数の空間 $\mathcal{S}'_0(\mathbb{R}^m)$ は，緩増加超関数の空間 $\mathcal{S}'(\mathbb{R}^m)$ を多項式関数で割った商空間と位相同型である (Yuan et al., 2010, Prop. 8.1)

$$\mathcal{S}'_0(\mathbb{R}^m) \cong \mathcal{S}'(\mathbb{R}^m)/\mathcal{P}(\mathbb{R}^m).$$

[1] 関数 f, g に対して $f \lesssim g$ とは，ある正定数 C が存在して $f \leqslant Cg$ となることをいう。

3.3　一般の空間上の関数と超関数

一般の空間上の関数と超関数のクラスについて整理する。各論に入る前に，位相ベクトル空間 （topological vector space）の一般論 (Trèves, 1967; Schwartz, 1966; 澤野嘉宏, 2011) を述べる。まず，領域 $\Omega(\subset \mathbb{R}^m)$ 上の関数空間 $\mathcal{D}(\Omega)$ とは，関数空間

$$\mathcal{D}(\Omega) := \bigcup_{K\subset\Omega} C^\infty(K),$$

に，次のセミノルムで位相を入れた空間である

$$p_\alpha(u) := \sup_{x\in\Omega} |\partial^\alpha u(x)|, \quad \alpha \in \mathbb{N}_0^m.$$

また, $\mathcal{A}$ が Banach 空間であれば, $\mathcal{A}'$ には作用素ノルム $\|T\| := \sup_{\|u\|\leqslant 1} |T(u)|$ で強位相を入れる。Frechét 空間の場合には，有界集合 B からノルムを定義する。特に局所凸分離位相ベクトル空間の場合には，弱位相 $\sigma(\mathcal{A}, \mathcal{A}')$ による弱有界集合（任意の $T \in \mathcal{A}'$ が集合 B 上で有界）と同値である（Mackey の定理)。

直積空間上のクラスに対しては, 核型定理が基本的である (Trèves, 1967; Hertle, 1983)。例えば $\mathbb{X}, \mathbb{Y} = \mathbb{R}^m, \mathbb{S}^{m-1}, \mathbb{S}^{m-1}\times\mathbb{R}$ のとき, $\mathcal{A} = \mathcal{S}, \mathcal{E}, \mathcal{E}', \mathcal{O}_C', \mathcal{S}', \mathcal{D}'$ に対して以下が成り立つ

$$\mathcal{A}(\mathbb{X}\times\mathbb{Y}) = \mathcal{A}(\mathbb{X})\hat{\otimes}\mathcal{A}(\mathbb{Y}) = \mathcal{A}(\mathbb{X}; \mathcal{A}(\mathbb{Y})).$$

ただし $\hat{\otimes}$ は（π ないし ε の意味で）完備化したテンソル積をあらわす。

3.3.1　球面上の関数と超関数

一般にコンパクト集合上では，無限遠（at infinity）における性質は定義しえない。従って，急減少関数 $\mathcal{S}$ や，急減少関数を急減少関数にうつす乗法作用素である緩増加関数 $\mathcal{O}_\mathcal{M}$ などは，単に $\mathcal{D}$ または $\mathcal{E}$ に帰着する。

球面 $\mathbb{S}^{m-1}$ 上の場合，Euclid 空間上で定義した関数クラスは $\mathcal{D}(\mathbb{S}^{m-1}) \subset \mathcal{D}(\mathbb{R}^m)$ と $\mathcal{E}'(\mathbb{S}^{m-1}) \subset \mathcal{E}'(\mathbb{R}^m)$ のいずれかに縮退する

$$\begin{aligned}\mathcal{D}(\mathbb{S}^{m-1}) &= \mathcal{S}(\mathbb{S}^{m-1}) = \mathcal{O}_\mathcal{M}(\mathbb{S}^{m-1}) = \mathcal{E}(\mathbb{S}^{m-1}),\\ \mathcal{E}'(\mathbb{S}^{m-1}) &= \mathcal{O}_C'(\mathbb{S}^{m-1}) = \mathcal{S}'(\mathbb{S}^{m-1}) = \mathcal{D}'(\mathbb{S}^{m-1}).\end{aligned}$$

なお，$\mathcal{D}(\mathbb{S}^{m-1})$ には球面上の Laplace-Beltrami 作用素 $\triangle_u$ を用いたセミノルムによる位相を入れる

$$\rho^\ell(\Phi) := \sup_u \left|\triangle_u^{\ell/2}\Phi(u)\right|, \quad \ell \in \mathbb{N}_0.$$

3.3.2 半空間上の関数と超関数

まず，$\mathcal{E}(\mathbb{H}) \subset \mathcal{E}(\mathbb{R}^2)$ と $\mathcal{D}(\mathbb{H}) \subset \mathcal{D}(\mathbb{R}^2)$ とする。$T \in \mathcal{E}(\mathbb{H})$ に対し，

$$\mathrm{D}_{s,t}^{k,\ell}T(\alpha,\beta) := (\alpha + 1/\alpha)^s\,(1+\beta^2)^{t/2}\partial_\alpha^k\partial_\beta^\ell T(\alpha,\beta), \quad s,t,k,\ell \in \mathbb{N}_0.$$

とおく。

半空間上の急減少関数 $\mathcal{S}(\mathbb{H})$ は，$T \in \mathcal{E}(\mathbb{H})$ であって，任意の $s,t,k,\ell \in \mathbb{N}_0$ に対してセミノルムが有限なものの全体である

$$\sup_{(\alpha,\beta)\in\mathbb{H}} \left|\mathrm{D}_{s,t}^{k,\ell}T(\alpha,\beta)\right| < \infty.$$

すなわち，β に関しては通常の急減少関数であり，α に関しては任意の有理式よりも早く減衰する関数を急減少関数と定義する。$\mathcal{S}(\mathbb{H})$ はウェーブレット解析で頻繁に登場する (Holschneider, 1995)。

半空間上の緩増加関数 $\mathcal{O}_\mathcal{M}(\mathbb{H})$ は，$T \in \mathcal{E}(\mathbb{H})$ であって，任意の $k,\ell \in \mathbb{N}_0$ に対して $s,t \in \mathbb{N}_0$ が存在して

$$\left|\mathrm{D}_{0,0}^{k,\ell}T(\alpha,\beta)\right| \lesssim (\alpha+1/\alpha)^s(1+\beta^2)^{t/2},$$

となるものの全体である。すなわち，急減少関数を急減少関数にうつす乗法作用素である。この定義と，次の同値条件は園田が独自に与えた。

定理 3.3.1. $U \in C^\infty(\mathbb{H})$ が $\mathcal{O}_\mathcal{M}(\mathbb{H})$ であることは，任意の $k,\ell \in \mathbb{N}_0$ に対して，ある $s,t > 0$ があって，以下が成り立つことと同値

$$|\partial_\alpha^k\partial_\beta^\ell U(\alpha,\beta)| \lesssim (\alpha+1/\alpha)^s(1+\beta^2)^{t/2}.$$

Proof. $T \in \mathcal{S}(\mathbb{H})$ を任意にとる。$U \in C^\infty(\mathbb{H})$ に対し，

$$\partial_\alpha^K\partial_\beta^L(TU) = \sum_{k,\ell}\binom{K}{k}\binom{L}{\ell}\partial_\alpha^{K-k}\partial_\beta^{L-\ell}T\partial_\alpha^k\partial_\beta^\ell U$$

なので，各 $U^{(k,\ell)} := \partial_\alpha^k\partial_\beta^\ell U$ が適当な $(\alpha+1/\alpha)^s(1+\beta^2)^t$ で押さえられれば，$\sup|T^{K-k,L-\ell}U^{k,\ell}| \lesssim \sup|T^{K-k,L-\ell}(\alpha+1/\alpha)^s(1+\beta^2)^t| < \infty$ が言える。 □

半空間上の Schwartz 超関数 $\mathcal{D}'(\mathbb{H})$ は，$\mathcal{D}(\mathbb{H})$ 上の有界線形汎関数 Φ であって，任意のコンパクト集合 $\mathrm{K} \subset \mathbb{H}$ に対して適当な $N \in \mathbb{N}_0$ をとって

$$\left| \int_{\mathrm{K}} T(\alpha,\beta)\Phi(\alpha,\beta)\frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha} \right| \lesssim \sum_{k,\ell \leqslant N} \sup_{(\alpha,\beta)\in\mathbb{H}} |\mathrm{D}_{0,0}^{k,\ell} T(\alpha,\beta)|, \quad \forall T \in \mathcal{D}(\mathrm{K}),$$

とできるものの全体である。ただし積分は Φ の作用の意味でとる。

半空間上の緩増加超関数 $\mathcal{S}'(\mathbb{H})$ は，$\Phi \in \mathcal{S}(\mathbb{H})$ であって，ある $N \in \mathbb{N}_0$ に対して

$$\left| \int_{\mathbb{H}} T(\alpha,\beta)\Phi(\alpha,\beta)\frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha} \right| \lesssim \sum_{s,t,k,\ell \leqslant N} \sup_{(\alpha,\beta)\in\mathbb{H}} \left|\mathrm{D}_{s,t}^{k,\ell} T(\alpha,\beta)\right|, \quad \forall T \in \mathcal{S}(\mathbb{H}).$$

が成り立つものの全体である。

3.3.3　直積空間上の関数と超関数

$\mathbb{S}^{m-1} \times \mathbb{R}$ 上の急減少関数は，(Helgason, 2011; Kostadinova et al., 2014) などに見られる

$$\mathcal{S}(\mathbb{S}^{m-1} \times \mathbb{R}) := \{\Phi \in C^\infty(\mathbb{S}^{m-1} \times \mathbb{R}) \mid \forall \rho_{p,q}^{k,\ell}(\Phi) < \infty\},$$

ただし

$$\rho_s^{k,\ell}(\Phi) := \sup_{u,p} (1+p^2)^{s/2} \left| \partial_p^k \triangle_u^{\ell/2} \Phi(u,p) \right|, \quad s,k,\ell \in \mathbb{N}_0.$$

$\mathbb{Y}^{m+1}$ 上の急減少関数は，(Holschneider, 1995; Kostadinova et al., 2014) などに見られる

$$\mathcal{S}(\mathbb{Y}^{m+1}) := \{T \in C^\infty(\mathbb{Y}^{m+1}) \mid \forall \rho_{p,q}^{j,k,\ell}(T) < \infty\},$$

ただし

$$\rho_{p,q}^{j,k,\ell}(T) := \sup_{u,\alpha,\beta} \left(\alpha^p + \frac{1}{\alpha^p}\right) (1+\beta^2)^{q/2} \left| \partial_\alpha^j \partial_\beta^k \triangle_u^{\ell/2} T(u,\alpha,\beta) \right|, \quad p,q,j,k,\ell \in \mathbb{N}_0.$$

以下の核型定理 (Kostadinova et al., 2014; Trèves, 1967) が成り立つ。

$$\begin{aligned}
\mathcal{S}(\mathbb{Y}^{m+1}) &= \mathcal{D}(\mathbb{S}^{m-1}) \hat{\otimes} \mathcal{S}(\mathbb{H}), \\
\mathcal{S}(\mathbb{S}^{m-1} \times \mathbb{R}) &= \mathcal{D}(\mathbb{S}^{m-1}) \hat{\otimes} \mathcal{S}(\mathbb{R}), \\
\mathcal{S}_0(\mathbb{S}^{m-1} \times \mathbb{R}) &= \mathcal{D}(\mathbb{S}^{m-1}) \hat{\otimes} \mathcal{S}_0(\mathbb{R}), \\
\mathcal{S}'(\mathbb{Y}^{m+1}) &= \mathcal{S}'(\mathbb{H}; \mathcal{D}'(\mathbb{S}^{m-1})).
\end{aligned}$$

3.4 超関数の畳み込み

一般に超関数の畳み込みは，非可換であり，

$$\phi * \psi \neq \psi * \phi,$$

また，結合的ですらない

$$\phi * (\psi * \eta) \neq (\phi * \psi) * \eta.$$

Schwartz (1966, Ch.6 Th.7, Ch.7 Th.7) によれば，$\mathcal{D}' * \mathcal{E}' * \mathcal{E}' * \cdots$ や $\mathcal{S}' * \mathcal{O}'_{\mathcal{C}} * \mathcal{O}'_{\mathcal{C}} * \cdots$ という組合せであれば，可換かつ結合的であることが保証される。

表 3.1 は超関数の意味での畳み込みが定義できる組合せと，畳み込みの値域の一覧である (Schwartz, 1966)。ここで正則化（regularization）とは，超関数を平滑化して関数を得る作用である。

表 3.1: 超関数の畳み込み

通称	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_1 * \mathcal{A}_2$
正則化	$\mathcal{D}$	$\mathcal{D}', \mathcal{E}'$	$\mathcal{E}, \mathcal{D}$
コンパクト台をもつ超関数	$\mathcal{E}'$	$\mathcal{E}', \mathcal{E}, \mathcal{D}'$	$\mathcal{E}', \mathcal{E}, \mathcal{D}'$
正則化	$\mathcal{S}$	$\mathcal{S}, \mathcal{S}'$	$\mathcal{S}, \mathcal{O}_{\mathcal{M}}$
Schwartz convolutor	$\mathcal{O}'_{\mathcal{C}}$	$\mathcal{S}, \mathcal{O}'_{\mathcal{C}}, \mathcal{S}'$	$\mathcal{S}, \mathcal{O}'_{\mathcal{C}}, \mathcal{S}'$

3.5 Fourier 解析

関数 $f : \mathbb{R}^m \to \mathbb{C}$ に対し，Fourier 変換と逆 Fourier 変換をそれぞれ以下で定義する

$$\widehat{f}(\xi) := \int_{\mathbb{R}^m} f(x) e^{-ix\cdot\xi} \mathrm{d}x, \quad \xi \in \mathbb{R}^m$$
$$\check{f}(x) := \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} f(\xi) e^{ix\cdot\xi} \mathrm{d}\xi, \quad x \in \mathbb{R}^m.$$

Fourier 変換は定義域に応じて存在の意味が異なり，反転公式の収束の意味も異なる。表 3.2 に Fourier 変換の一覧を示す。ただし UC_0 とは，一様連続かつ無限遠で減衰する関数（$\lim_{|x|\to\infty} f(x) = 0$）である。

Fourier 変換の基本形は L^1 関数に対する定義である。このとき Fourier 変換は各点 ξ ごとに絶対収束の意味で定義され，反転公式は（対象となる関数の連続点で）各点収束である。積分が絶対収束していれば，具体的な値を求めるには留数計算そのほかいかなる部分和の極限を使っても一意な値を求める事ができる。$\widehat{f}$ は各点で定義されているだけだが，f の可積分性から $\widehat{f}$ の連続性が言える。$f \in L^1(\mathbb{R}^m)$ に対し，以下が成り立つ（Riemann-Lebesgue の補題）

$$\|\widehat{f}\|_\infty \leqslant \|f\|_1.$$

また，$f, g \in L^1 \cap L^2(\mathbb{R}^m)$ に対し，以下が成り立つ

$$\int f(x)\overline{g(x)}\mathrm{d}x = \int \widehat{f}(\xi)\overline{\widehat{g}(\xi)}\mathrm{d}\xi \qquad \text{Plancherel's theorem}$$

$$\int |f(x)|^2\mathrm{d}x = \int |\widehat{f}(\xi)|^2\mathrm{d}\xi \qquad \text{Parseval's theorem.}$$

急減少関数の Fourier 変換は，急減少関数が可積分であることを用いて，L^1 関数に対する Fourier 変換を $\mathcal{S}(\mathbb{R}^m)$ に制限して定義する。まず，急減少関数 $f \in \mathcal{S}(\mathbb{R}^m)$ の Fourier 変換 $\widehat{f}$ は再び急減少関数になる。従って，Fourier 変換を $\mathcal{S}(\mathbb{R}^m)$ に制限したものは，線形作用素 $\mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$ を定める。この作用素は有界かつ全単射であることが知られている。

緩増加超関数の Fourier 変換は，急減少関数に対する Fourier 変換の双対作用素として定義する。まず，任意の緩増加超関数 $u \in \mathcal{S}'(\mathbb{R}^m)$ に対して，ある緩増加超関数 $\widehat{u} \in \mathcal{S}'(\mathbb{R}^m)$ で，任意の急減少関数 $f \in \mathcal{S}(\mathbb{R}^m)$ に対して $\widehat{u}(f) = u(\widehat{f})$ となるものが一意に存在する。緩増加超関数の Fourier 変換とは，$u \in \mathcal{S}'(\mathbb{R}^m)$ に前述の $\widehat{u}$ を対応付ける作用素 $\mathcal{S}'(\mathbb{R}^m) \to \mathcal{S}'(\mathbb{R}^m)$ である。この作用素もまた，有界かつ全単射であることが知られている。

L^2 関数の Fourier 変換は，$L^1 \cap L^2$ 関数が $L^2(\mathbb{R}^m)$ で稠密であることを用いて，$L^1 \cap L^2$ 関数に対する Fourier 変換の有界拡張（bounded extension, Grafakos (2008, 2.2.4)）によって定義する。まず，任意の $f \in L^2(\mathbb{R}^m)$ と，f に L^2 収束する任意の点列 $f_j \in L^1 \cap L^2(\mathbb{R}^m)$ に対して，点列 $\widehat{f_j}$ の L^2 極限 $\widehat{f}$ は点列の取り方に依らず一意に存在する。L^2 関数の Fourier 変換とは，$f \in L^2(\mathbb{R}^m)$ に前述の極限 $\widehat{f}$ を対応付ける作用素 $L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m)$ である。この作用素もまた，有界かつ全単射であることが知られている。

緩増加関数 $\mathcal{O}_\mathcal{M}(\mathbb{R}^m)$，急減少超関数 $\mathcal{O}'_\mathcal{C}(\mathbb{R}^m)$，多項式関数 $\mathcal{P}(\mathbb{R}^m)$，台がコンパクトな超関数 $\mathcal{E}'(\mathbb{R}^m)$ の Fourier 変換はいずれも，緩増加超関数の Fourier 変換を制限して定義する。特に，$\mathcal{O}_\mathcal{M}(\mathbb{R}^m) \to \mathcal{O}'_\mathcal{C}(\mathbb{R}^m)$ は全単射で

ある。また，原点に台をもつ超関数を $\mathcal{E}'(\{0\})$ として，$\mathcal{P}(\mathbb{R}^m) \to \mathcal{E}'(\{0\})$ は全単射である (Rudin, 1991, Ex.7.16)。

表 3.2: $\mathbb{R}^m$ 上の関数に対する Fourier 変換

定義域	定義	値域	反転公式の収束
L^1	絶対可積分	$L^\infty \cap UC_0$	f の連続点で各点収束
$\mathcal{S}$	L^1 の制限	$\mathcal{S}$	各点収束
$\mathcal{S}'$	双対作用素	$\mathcal{S}'$	$\mathcal{S}'$ の位相
L^2	$L^1 \cap L^2$ から有界拡張	L^2	L^2 収束
$\mathcal{O}_\mathcal{M}$	$\mathcal{S}'$ の制限	$\mathcal{O}'_\mathcal{C}$	$\mathcal{S}'$ の位相
$\mathcal{P}$	$\mathcal{S}'$ の制限	$\mathcal{E}'(\{0\})$	$\mathcal{S}'$ の位相

3.6 Hilbert 変換

関数 $f : \mathbb{R} \to \mathbb{C}$ に対し，Hilbert 変換を以下で定義する。

$$\mathcal{H}f(s) := \frac{i}{\pi}\,\mathsf{pv}\int_{-\infty}^{\infty}\frac{f(t)}{s-t}\mathrm{d}t, \quad s \in \mathbb{R}$$

ここで pv は Cauchy の主値積分を表す。以下が成り立つ。

$$\widehat{\mathcal{H}f}(\omega) = \operatorname{sgn}\omega \cdot \hat{f}(\omega), \quad \omega \in \mathbb{R}$$
$$\mathcal{H}^2 f(s) = f(s), \quad s \in \mathbb{R}.$$

3.7 Radon 変換

3.7.1 定義

関数 $f : \mathbb{R}^m \to \mathbb{R}$ の Radon 変換および関数 $\Phi : \mathbb{S}^{m-1} \times \mathbb{R} \to \mathbb{R}$ の双対 Radon 変換は以下で定義する。

$$\mathrm{R}f(u,p) := \int_{(\mathbb{R}u)^\perp} f(pu+y)\mathrm{d}y, \quad (u,p) \in \mathbb{S}^{m-1} \times \mathbb{R}$$
$$\mathrm{R}^\dagger\Phi(x) := \int_{\mathbb{S}^{m-1}} \Phi(u, u\cdot x)\mathrm{d}u, \quad x \in \mathbb{R}^m$$

ここで $(\mathbb{R}u)^\perp := \{y \in \mathbb{R}^m \mid y \cdot u = 0\}$ は方向ベクトル u が張る線形部分空間 $\mathbb{R}u \subset \mathbb{R}^m$ の直交補空間，$\mathrm{d}y$ は $(\mathbb{R}u)^\perp$ 上の Lebesgue 測度，そして $\mathrm{d}u$ は $\mathbb{S}^{m-1}$ 上の球面測度を表す。

3.7.2　逆投影フィルタ

関数 $\Phi : \mathbb{S}^{m-1} \times \mathbb{R} \to \mathbb{R}$ に対し逆投影フィルタ（backprojection filter）Λ^m を以下で定義する[2]

$$\Lambda^m \Phi(u,p) := \begin{cases} \partial_p^m \Phi(u,p), & m \text{ even} \\ \mathcal{H}_p \partial_p^m \Phi(u,p), & m \text{ odd.} \end{cases}$$

ここで $\mathcal{H}_p$ と ∂_p はそれぞれ，p に関する Hilbert 変換と偏微分である。p に関する Fourier 変換（$p \to \omega$）を用いて，以下の関係式が成り立つ

$$\widehat{\Lambda^m \Phi}(u,\omega) = i^m |\omega|^m \widehat{\Phi}(u,\omega).$$

3.7.3　諸性質

Fubini の定理の系

$$\int_{\mathbb{R}} \mathrm{R}f(u,p)\mathrm{d}p = \int_{\mathbb{R}^m} f(x)\mathrm{d}x, \quad \text{a.e. } u \in \mathbb{S}^{m-1}.$$

微分に纏わる公式

$$\begin{aligned} \mathrm{R}[\partial_i f](u,p) &= u_i \partial_p \mathrm{R}[f](u,p) \\ \partial_p^2 \mathrm{R} &= \mathrm{R}\triangle \\ \triangle \mathrm{R}^\dagger &= \mathrm{R}^\dagger \partial_p^2. \end{aligned}$$

投影切断面定理（Fourier slice theorem）

$$\widehat{f}(\omega u) = \int_{\mathbb{R}} \mathrm{R}f(u,p) e^{-ip\omega} \mathrm{d}p, \quad (u,\omega) \in \mathbb{S}^{m-1} \times \mathbb{R}$$

ただし左辺は m 次元 Fourier 変換（$x \to \xi = \omega u$）であり，右辺は p に関する 1 次元 Fourier 変換（$p \to \omega$）である。

[2] Λ^m は (Helgason, 2011) にも登場する古典的な作用素だが，定まった名前がない。小川卓克 (2013) は Riesz ポテンシャルと呼んでいる。ただし，狭義の Riesz ポテンシャルは Λ^m の逆に相当するので注意せよ。

3.7.4 反転公式

関数 $f \in L^1(\mathbb{R}^m)$ に対し，反転公式（Radon's inversion formula）が成り立つ

$$\mathrm{R}^\dagger \Lambda^{m-1} \mathrm{R} f = 2(2\pi)^{m-1} f.$$

逆変換を実行する古典的な方法には三通りある (Natterer, 2008)。逆投影定理を用いて Fourier 逆変換に帰着する方法，Radon の反転公式を数値積分によって計算する方法，Radon 変換の逆作用素を Kaczmarz 法などの線形計算によって代数的に求める方法（algebraic reconstruction techniques; ART）である。

3.7.5 幾何学的側面

広義の Radon 変換は逆問題や積分幾何学の主題の一つであり，物理学や幾何学的な背景をもつ。Helgason (2011) は Radon 変換を次のように定義している。$\mathbb{R}^m$ の超平面の全体を $\mathbb{P}^m$ とする。関数 $f : \mathbb{R}^m \to \mathbb{R}$ の Radon 変換は以下で定義される。

$$\mathrm{R} f(\xi) := \int_\xi f(x) \mathrm{d}m(x), \quad \xi \in \mathbb{P}^m,$$

ただし m は ξ 上の Lebesgue 測度である。

関数 $\Phi : \mathbb{P}^m \to \mathbb{R}$ の双対 Radon 変換は以下で定義される。

$$\mathrm{R}^\dagger \Phi(x) := \int_{x \in \xi} \Phi(\xi) \mathrm{d}\mu(\xi)$$

ここで μ は x を中心とする回転によって不変な位相群 $\{\xi | x \in \xi\}$ の Haar 確率測度である。

3.8 ウェーブレット変換

関数 $f \in L^2(\mathbb{R})$ のウェーブレット関数 $\psi \in L^2(\mathbb{R})$ による連続ウェーブレット変換を以下で定義する

$$\mathcal{W}_\psi f(a, b) := \int_{\mathbb{R}} f(x) \overline{\psi\left(\frac{x-b}{a}\right)} \frac{1}{a} \mathrm{d}x, \quad (a, b) \in \mathbb{R}_+ \times \mathbb{R}.$$

ψ はマザーウェーブレットまたは分解ウェーブレット（analyzing -）とも呼ばれる。

以下の許容条件（admissibility condition）が成り立つとする

$$K_\psi := 2\pi \int_{\mathbb{R}} \frac{|\widehat{\psi}(\xi)|^2}{|\xi|} \mathrm{d}\xi < \infty.$$

このとき，L^2 収束の意味で再構成公式が成り立つ

$$\int_0^\infty \left[\int_{\mathbb{R}} \mathcal{W}_\psi f(a,b) \psi \left(\frac{x-b}{a} \right) \frac{1}{a} \mathrm{d}b \right] \frac{\mathrm{d}a}{a} = K_\psi f(x),$$

また，$f(x)$ の連続点において各点収束する。

$\psi \in L^1(\mathbb{R})$ のとき，$\widehat{\psi}$ は連続関数である。従って，許容条件は $\widehat{\psi}(0) = \int_{\mathbb{R}} \psi(z)\mathrm{d}z = 0$ となることを要請している。逆に，$\int_{\mathbb{R}} \psi(z)\mathrm{d}z = 0$ かつ，ある $\alpha > 0$ に対して $\int_{\mathbb{R}} (1+|z|)^\alpha \psi(z)\mathrm{d}z < 0$ が成り立つとき，$|\widehat{\psi}(\zeta)| \lesssim |\zeta|^{\min(1,\alpha)}$ となって，許容条件が満たされる。

少し条件を変えると，再構成には分解ウェーブレットと異なるウェーブレット関数を使うこともできる。これを再構成ウェーブレット（reconstructing -）または合成ウェーブレット（synthesising -）と呼ぶ。多くの一般化ウェーブレットは Calderón (1964) による再生公式 (Rubin, 1998)

$$\int_{SO(m)} \int_0^\infty \frac{f * \mu_{k,t}}{t} \mathrm{d}t\mathrm{d}k = K_\mu f, \quad f \in L^2(\mathbb{R}^m)$$

に帰着できることが，後から再発見された。ただし μ は $\mathbb{R}^m$ 上の Borel 測度として，$\mu_{k,t}(x) := \mu(k^{-1}x/t)/t^m$ とし，

$$K_\mu = \int_{\mathbb{R}^m} \frac{\widehat{\mu}(u \cdot x)}{|\zeta|^m} \mathrm{d}\zeta,$$

である。さらに，ウェーブレット変換だけではなく，Radon 変換やリッジレット解析も Calderón 型の再生公式に帰着できることが分かっている (Rubin, 1998, 2004)。

3.9　関数近似の原理

関数 $f : \mathbb{R}^m \to \mathbb{C}$ の近似とは，収束列

$$f_j \to f \quad \text{as} \quad j \to \infty$$

のことである。ここで収束は文脈に応じて定めるものとする。例えば L^2 ノルムや一様ノルムを用いることが多い。

3.9.1 Dirac の δ 関数

関数近似の基本形は Dirac の δ 関数である。すなわち，急減少超関数 $f \in \mathcal{O}'_C(\mathbb{R}^m)$[3]に対して以下が成り立つ (Schwartz, 1966, Ch.7)

$$\delta * f = f.$$

ここでの畳み込みや等号は厳密には超関数の意味だが，多くの関数近似は極限でこの形式に帰着できる。なお，$L^1(\mathbb{R}^m)$ に畳み込み $*$ で積を定義した環には，単位元が存在しない。なぜならば，単位元 e があるとすれば任意の $f \in L^1(\mathbb{R}^m)$ に対して $e * f = f$ が成り立つが，このとき $(\widehat{e * f} =)\widehat{e}\widehat{f} = \widehat{f}$ より $\widehat{e} = 1 \notin L^1(\mathbb{R}^m)$ なので，$L^1(\mathbb{R}^m)$ 上の Fourier 変換の単射性によりそのような e は存在しない（$e = \delta$ 以外ありえない）ことが分かるためである。従って，$\mathbb{R}^m$ 上に Lebesgue 測度を入れた空間 $L^p(\mathbb{R}^m)$ で考える限り[4]，任意の関数[5]を有限の手続きで近似するスキームは存在しない。

3.9.2 近似単位元

$k \in L^1(\mathbb{R}^m)$ は

$$\int_{\mathbb{R}^m} k(x)\mathrm{d}x = 1,$$

を満たすとし，$t > 0$ に対し

$$k_t(x) := k\left(\frac{x}{t}\right)\frac{1}{t^m},$$

とおく。このとき $f \in L^p(1 \leqslant p < \infty)$ に対して以下が成り立つ (猪狩惺, 1996, Th.6.13)

$$\lim_{t\to 0} k_t * f = f \quad \text{in } L^p.$$

このような $\{k_t\}$ を近似単位元という。近似単位元が成り立つ原理は，形式的に以下が成り立つことから理解される

$$\lim_{t\to 0} k_t = \delta.$$

例えば $f_j := k_{1/j} * f$ とおくと，f_j は f の近似になる。

[3]急減少関数や，台がコンパクトな可積分関数など
[4]一般の Radon 測度や，畳み込み以外の演算を用いる場合などはこの限りではない。
[5]近似対象が有限個であれば，それらを予め辞書に持っておけば有限列近似ができる。

3.9.3　積分変換の反転公式

関数空間 V, W とし，積分変換 $T : V \to W$ と，その逆変換 $T^{-1} : W \to V$ を以下で定義する

$$T[f](\xi) := \int_{\mathbb{R}^m} f(x)\phi(x;\xi)\mathrm{d}x, \quad f \in V,\ \xi \in \Omega$$
$$T^{-1}[F](x) := \int_{\Omega} F(\xi)\phi^*(x;\xi)\mathrm{d}\xi, \quad F \in W,\ x \in \mathbb{R}^m$$

ただし Ω は変換後のパラメータ ξ の空間を表すものとし，積分核 ϕ, ϕ^* は諸々の積分が収束するようにとれているものとする[6]。このとき反転公式 $T^{-1}Tf = f$ の積分を形式的に順序交換すると以下を得る

$$\begin{aligned} f(x) &= \int_{\Omega} T[f](\xi)\phi^*(x;\xi)\mathrm{d}\xi \\ &= \int_{\mathbb{R}^m} f(x')K(x,x')\mathrm{d}x', \quad \text{a.e. } x \in \mathbb{R}^m \end{aligned}$$

ただし

$$K(x,x') := \int_{\Omega} \phi(x';\xi)\phi^*(x;\xi)\mathrm{d}\xi$$

とおいた。このような K を再生核という。特に T が Fourier 変換（すなわち $\phi(x;\xi) = e^{-ix\cdot\xi}$ かつ $\phi^*(x;\xi) = e^{ix\cdot\xi}$）のときは，

$$K(x,x') = \delta(x - x')$$

となり，この枠組みにも δ 関数が表れることが分かる。

反転公式を適当に有限和近似して f の近似が得られる。例えば，Ω の細分の列 $\bigsqcup_i \Omega_j^{(i)} (= \Omega)$ を用いて

$$f_j(x) := \sum_i \left[\int_{\Omega_j^{(i)}} T[f](\xi)\mathrm{d}\xi \right] \phi^*(x;\xi_i), \quad \xi_i \in \Omega_j^{(i)}$$

とおく。あるいは，$\mathbb{R}^m$ の細分の列 $\bigsqcup_i A_j^{(i)} (= \mathbb{R}^m)$ を用いて

$$f_j(x) := \sum_i \left[\int_{A_j^{(i)}} f(x')\mathrm{d}x' \right] K(x,x_i'), \quad x_i' \in A_j^{(i)}$$

とおけば，単調収束定理からいずれも $f_j \to f$ が従う。関数の基底展開はこの枠組みに帰着する。

[6] 再生核 Hilbert 空間論では $L^2(\mathbb{R}^m \times \Omega)$ にとる。

3.9.4 Calderón の再生公式

μ は $\mathbb{R}^m$ 上の動径 Borel 測度とし，

$$\int_{\mathbb{R}^m} \frac{\widehat{\mu}(\zeta)}{|\zeta|^m} \mathrm{d}\zeta = 1$$

とする。$t > 0$ に対して

$$\mu_t(x) := \mu\left(\frac{x}{t}\right) \frac{1}{t^m},$$

とおく。このとき以下が成り立つ (Rubin, 1998)

$$\int_0^\infty \frac{f * \mu_t}{t} \mathrm{d}t = f, \quad f \in L^2(\mathbb{R}^m)$$

これを Calderón の再生公式という。ウェーブレット変換や Radon 変換，リッジレット変換の反転公式や再構成公式などはこの形式に帰着できる (Rubin, 1998, 2004)。

Calderón 再生公式において $\mu_t = -t\partial_t k_t$ となるように μ をとる。ただし k_t は近似単位元とする。このとき形式的に，

$$-\int_0^\infty \frac{f * t\partial_t k_t}{t} \mathrm{d}t = -\int_0^\infty \partial_t (f * k_t) \mathrm{d}t = \lim_{t\to 0} f * k_t - \lim_{t\to\infty} f * k_t = f - 0$$

となることが理解される。

3.10 $1/x$ を含む積分

以下では $\mathbb{R}_\times := \mathbb{R}\backslash\{0\}, \mathbb{C}_\times := \mathbb{C}\backslash\{0\}$ と書く。

積分記号 $\int_{-\infty}^{\infty}$ には複数の意味がある。許容条件に現れる積分

$$\int_{-\infty}^\infty \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m} \mathrm{d}\zeta \tag{3.1}$$

が「一意」に定まるための条件を考える。分母 $\Omega(\zeta) := \overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)$ が $\mathbb{R}_\times$ 上の関数を定める時，かつ，ある $\alpha, \beta > 0$ に対して，

$$|\Omega(\zeta)| \lesssim |\zeta|^{m-1+\alpha}, \quad |\zeta| \in (0, 1]$$
$$|\Omega(\zeta)| \lesssim |\zeta|^{m-1-\beta}, \quad |\zeta| \in (1, \infty)$$

が成り立つ時，(3.1) は絶対収束する。すなわち，Lebesgue 積分の意味で可積分である。証明は比較定理による。

3.10.1　x^s の積分可能性

まず，関数 x^s $(s \in \mathbb{R})$ は $\mathbb{R}_\times$ で局所可積分かつなめらかな関数である

$$x^s \in L^1_{\mathrm{loc}} \cap C^\infty(\mathbb{R}_\times), \quad s \in \mathbb{R}.$$

一方，任意の $s \in \mathbb{R}$ に対して，$\int_0^\infty x^s \mathrm{d}x = \infty$ なので，$\mathbb{R}_\times$ 上，可積分でない

$$x^s \notin L^1(\mathbb{R}_\times), \quad s \in \mathbb{R}.$$

また，$s > -1$ のとき $\int_0^1 x^s \mathrm{d}x = 1/(1+p)$，$s < -1$ のとき $\int_1^\infty x^s \mathrm{d}x = -1/(1+p)$ である。すなわち，

$$x^s \in L^1([0,1]), \quad s > -1$$
$$x^s \in L^1([1,\infty)), \quad s < -1.$$

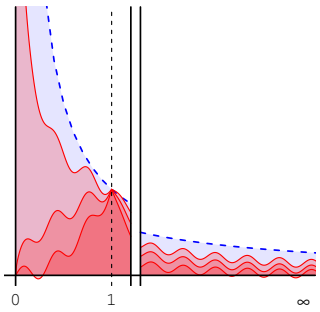


図 3.1: 原点での発散が $1/x$ より遅く，無限遠での減衰が $1/x$ より速いものは絶対可積分

3.10.2　Mellin 変換

関数 $f : [0,\infty) \to \mathbb{C}$ は，

$$|f(x)| \lesssim x^\alpha, \quad x \in (0,1]$$
$$|f(x)| \lesssim x^{-\beta}, \quad x \in (1,\infty)$$

を満たすとする。このとき Mellin 変換

$$\mathcal{M}f(s) := \int_0^\infty x^{s-1} f(x) \mathrm{d}x, \quad s \in \mathbb{C}$$

は，$\Sigma := \{s \in \mathbb{C} \mid \alpha < \Re s < \beta\}$ なる帯 (strip) 上で絶対可積分。特に，$\mathcal{M}f(s)$ は Σ 上の正則関数である。

逆に，$\mathcal{M}f(s)$ は Σ 上正則であることが分かっているとき，

$$\lim_{x\to 0+} x^s f(x) = 0, \quad \Re s > \alpha$$
$$\lim_{x\to \infty} x^s f(x) = 0, \quad \Re s < \beta.$$

3.10.3 Cauchy の主値

Cauchy の主値 $\mathsf{pv}\,\frac{1}{x}$ は $\mathbb{R}$ 上の緩増加超関数 ($\mathcal{S}'$) である。ただし，$\phi \in \mathcal{S}(\mathbb{R})$ に対し，

$$\mathsf{pv}\int_{-\infty}^{\infty} \frac{\phi(x)}{x}\mathrm{d}x := \lim_{\varepsilon\to 0}\int_{|x|>\varepsilon} \frac{\phi(x)}{x}\mathrm{d}x.$$

3.10.4 発散積分の正則化

発散積分の正則化 x^λ ($\lambda \in \mathbb{C}$) は，$\mathbb{R}_\times$ 上の超関数 ($\mathcal{D}'$) を定める (Gel'fand and Shilov, 1964)。例えば $\lambda = -1$ の場合の正則化は以下で与えられる。すなわち，$\phi \in \mathcal{D}(\mathbb{R})$ に対し，$a, b > 0$ として，

$$\langle x^{-1}, \phi\rangle := \int_{-\infty}^{-a} \frac{\phi(x)}{x}\mathrm{d}x + \int_{-a}^{b} \frac{\phi(x)-\phi(0)}{x}\mathrm{d}x + \int_{b}^{\infty} \frac{\phi(x)}{x}\mathrm{d}x.$$

特に，解析接続から導かれる発散積分 x^{-1} の正則化（Hadamard の有限部分）は，Cauchy の主値積分に一致する

$$\langle \mathsf{fp}\,\frac{1}{x}, \phi\rangle = \int_0^{\infty} \frac{\phi(x)-\phi(-x)}{x}\mathrm{d}x = \mathsf{pv}\int_{-\infty}^{\infty} \frac{\phi(x)}{x}\mathrm{d}x.$$

3.10.5 複素積分

複素関数 $\frac{1}{z}$ は $\mathbb{C}_\times$ で正則である。特に，線積分として以下が成り立つ。

$$\int_{|z|=1} \frac{\mathrm{d}z}{z} = \int_0^{2\pi} \frac{ie^{it}\mathrm{d}t}{e^{it}} = 2\pi i.$$

すなわち，$\mathbb{C}_\times$ 上不定積分を持たない。

3.10.6　特異積分

Hilbert 変換 $\mathsf{pv}\,\frac{1}{x}*$ は $\mathcal{S}(\mathbb{R})$ ないし $L^p(\mathbb{R})$ 上の特異積分作用素である。

$$\mathcal{H}f(s) := \frac{i}{\pi}\mathsf{pv}\int_{-\infty}^{\infty}\frac{f(t)}{s-t}\mathrm{d}t.$$

その高次元化として Riesz 変換が知られる。

3.11　拡散方程式

3.11.1　定義

次の放物型偏微分方程式を拡散方程式という

$$\partial_t u(x,t) = L(x,t)u(x,t) + f(u,x,t), \quad (x,t) \in \mathbb{R}^m \times \mathbb{R}_+$$

ただし

$$L(x,t) := \sum_{i,j} a_{ij}(x,t)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_i b_i(x,t)\frac{\partial}{\partial x_i} + c(x,t)$$

とし，(a_{ij}) は C^1 級正定値テンソル，b_i, c は Hölder 連続関数とする。f は非線形項と非斉次項をまとめた項である。

最も基本的な拡散方程式は熱方程式（$L = \triangle$）である

$$\partial_t u(x,t) = \triangle u(x,t), \quad (x,t) \in \mathbb{R}^m \times \mathbb{R}_+.$$

拡散係数を $D(x,t)$ とする $\mathbb{R}^m$ 上の拡散方程式は以下で与えられる

$$\partial_t u(x,t) = \nabla \cdot [D(x,t)\nabla u(x,t)], \quad (x,t) \in \mathbb{R}^m \times \mathbb{R}_+.$$

ここで $D(x,t)$ は各点で C^2 級かつ正定値対称なテンソルとする。D が定数のとき熱方程式に帰着する。

さらに，媒質自体が動いている場合には移流項 $\nabla \cdot (\mathbf{v}u)$，反応に伴う濃度変化がある場合には反応項 $c(u)$，濃度に依らず外界からの入出力がある場合には非斉次項 f を加えることがある

$$\begin{aligned}\partial_t u(x,t) = \nabla \cdot [D(x,t)\nabla u(x,t) - \mathbf{v}(x,t)u(x,t)] \\ + c(u(x,t)) + f(x,t), \quad (x,t) \in \mathbb{R}^m \times \mathbb{R}_+.\end{aligned}$$

このような方程式は移流反応拡散方程式と呼ばれる。

Riemann 多様体 (M, g) 上の Laplacian $\triangle_g$ は以下で与えられる

$$\triangle_g u := \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|} g^{ij} \frac{\partial}{\partial x_j} u \right), \quad u \in C^\infty(M).$$

ただし $|g|$ は g の行列式，g^{ij} は g の逆行列の (i, j) 成分を表す。これを用いて M 上の熱方程式は以下で与えられる

$$\partial_t u(x, t) = \triangle_g u(x, t), \quad (x, t) \in M \times \mathbb{R}_+.$$

3.11.2 熱核と熱半群

拡散方程式の基本解 $\mathrm{W}_t(x, y)$ を熱核（heat kernel）という。$\mathbb{R}^m$ 上の拡散方程式 $\partial_t u = Lu$ の場合，熱核 $\mathrm{W}_t(x, y)$ は以下を満たす

$$\begin{aligned} \partial_t \mathrm{W}_t(x, y) &= L\mathrm{W}_t(x, y), \quad x, y \in \mathbb{R}^m \\ \lim_{t \to 0} \mathrm{W}_t(x, y) &= \delta(x - y), \quad x, y \in \mathbb{R}^m \\ \lim_{|(x,y)| \to \infty} |\mathrm{W}_t(x, y)| &= 0, \quad t > 0. \end{aligned}$$

このような熱核は，例えば (a_{ij}) が C^1 級で，b_i, c が Hölder 連続のとき存在する (伊藤清三, 1979, Th.5.1)。

熱核を用いて，発展作用素を定義する

$$e^{tL} u_0 := \int_{\mathbb{R}^m} \mathrm{W}_t(x, y) u_0(y) \mathrm{d}y, \quad t > 0.$$

これを熱半群という。e^{tL} に対し，L を生成作用素（generator）と呼ぶこともある。名前が示すとおり，熱半群は（適当な正則条件を満たす）関数空間に作用する 1 パラメータ半群（C_0 半群）である。つまり以下が成り立つ

$$\begin{aligned} & e^{0L} u_0 = u_0, \\ & \lim_{t \to 0} e^{tL} u_0 = u_0, \quad t > 0 \\ & e^{tL} e^{sL} u_0 = e^{(t+s)L} u_0, \quad s, t > 0. \end{aligned}$$

熱半群の軌道 $e^{tL}u_0$ は，$u(x,0) := u_0(x)$ を初期値とする初期値問題 $\partial_t u = Lu$ の解である。形式的には，$e^{0L}u_0 = u_0$ であり，さらに

$$\partial_t[e^{tL}u_0] = L\partial_t[e^{tL}u_0]$$

が成り立つことから理解できる。古典的な証明は u_0 が有界連続関数の場合に示される。拡散方程式を弱形式で捉えると，$u_0 \in \mathcal{S}'(\mathbb{R}^m)$ でも同様の主張が成り立つ。

特に熱方程式（$L = \triangle$）の場合，熱核は Gauss 関数になる

$$\mathrm{W}_t(x,y) = (4\pi t)^{-m/2}\exp(-|x-y|^2/4t).$$

このとき熱半群は畳み込み作用素である

$$e^{t\triangle}u_0 = \mathrm{W}_t * u_0, \quad t > 0.$$

3.12　積分の変数変換に纏わる公式

3.12.1　確率変数の変数変換と押出測度

確率変数 X は $\mathbb{R}^m$ に値を取り，確率測度を μ とする確率分布に従うとする。可測写像 $f : \mathbb{R}^m \to \mathbb{R}^n$ を用いて $Y := f(X)$ と変換する。Y が従う確率分布（に対応する測度）を f による μ の押出測度（pushforward measure）と呼び，$f_\sharp\mu$ と書く。定義より，$f_\sharp\mu$ は以下を満たす

$$f_\sharp\mu(A) := \mu \circ f^{-1}(A), \quad A \in \mathcal{B}(\mathbb{R}^n)$$

ただし $\mathcal{B}$ は Borel 集合族を表す。

特に，f は Lipschitz 連続で，μ は確率密度関数 ν を持つとする。このとき重積分の変数変換の公式から以下が成り立つ

$$f_\sharp\nu \circ f(x) \cdot |\nabla f|(x) = \nu(x), \quad \mathsf{a.e.}\, x \in \mathbb{R}^m.$$

ただし $f_\sharp\nu$ は $f_\sharp\mu$ の確率密度関数であり，$|\nabla f|$ は f の Jacobian である。

3.12.2 多様体上の積分

$M \subset \mathbb{R}^n$ は Lipschitz 連続に埋め込まれた m 次元部分多様体とする。$U \subset \mathbb{R}^m$ と $f : U \to M$ を M の局所座標系（chart）とする。$f : M \to \mathbb{R}^n\,(m \leqslant n)$ の Jacobian $|\nabla f|$ は，$\nabla f = (\partial_i f_j)$ を $m \times n$ 行列として

$$|\nabla f| = \sqrt{|(\nabla f)^\top (\nabla f)|},$$

によって計算できる。ただし $|\cdot|$ は行列式を表す。さらに Borel 集合 $A \subset f(U)$ の体積は以下で計算できる

$$\mathsf{vol}(A) = \int_{f^{-1}(A)} |\nabla f| \mathrm{d}x.$$

詳細は (Evans and Gariepy, 2015, 3.3.4D) を見よ。

3.13 輸送問題と Wasserstein 幾何学

エントロピー勾配流までの流れを整理する。

3.13.1 最適輸送問題

Monge の問題

$D, E \subset \mathbb{R}^3$ は体積 1 の可測集合（＝砂山）とする。点 $x \in \mathbb{R}^3$ から点 $y \in \mathbb{R}^3$ まで砂を運ぶのにかかるコストは単位体積あたり $c(x, y)$ である。以下の条件を満たす写像 $T : D \to E$ を求めよ:

1. T は全単射
2. 任意の部分集合 $U \subset D$ に対して $T(U)$ と U の体積は等しい
3. 総コスト $C[T] := \int_D c(x, T(x))\mathrm{d}x$ を最小にする

Monge の問題の解 T を最適輸送写像 (an optimal transport map) と呼ぶ。Monge の問題は必ずしも解を持つとは限らず，また存在したとしても一意とは限らない。例えば，D が一点集合（特異測度）の場合には，輸送経路は分岐しなければならないので，T は写像として定式化できない。また，E が一点集合の場合には，T は存在したとしても単射にはできな

い。従って不良設定である。Monge がこの問題を提出したのは 1781 年と古いが，適当な正則条件の下で解の存在が示されたのは 1999 年以降である。なお，全単射性の制約を外した問題を一般化された Monge の問題と呼ぶ。

Monge-Kantorovich の問題

D, E の代わりに $\mathbb{R}^3$ 上の確率測度 μ, ν をとる。$\mathbb{R}^3 \times \mathbb{R}^3$ 上の確率測度で，周辺測度が μ と ν になるような確率測度の全体を $\Pi(\mu, \nu)$ と書く。$\Pi(\mu, \nu)$ の元 π で，以下の総コストを最小化するものを求めよ:

$$C[\pi] := \int_{\mathbb{R}^3 \times \mathbb{R}^3} c(x, y) \mathrm{d}\pi(x, y).$$

MK 問題の解 π_0 を最適輸送計画または最適カップリングと呼ぶ。

MK 問題の解の存在については, 次の定理が知られている。X, Y は Polish 空間とし，それぞれの空間上の確率測度を $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y)$ とする。コスト関数 $c : X \times Y \to \mathbb{R}_+$ は下半連続とする。結合分布 $\pi \in \Pi(\mu, \nu)$ で，総コスト $C[\pi] := \int_{X \times Y} c(x, y)\pi(\mathrm{d}x\mathrm{d}y)$ を有限にするものが存在すれば，MK 問題

$$\inf_{\pi \in \Pi(\mu, \nu)} \quad \int_{X \times Y} c(x, y)\pi(\mathrm{d}x\mathrm{d}y),$$

の解 π_0 が存在する。

MK 問題は確率論的には次の問題と等価である: $X \times Y$-値確率変数 (U, V) で, X 側の周辺分布が μ, Y 側の周辺分布が ν になるもののうち，$\mathbb{E}[c(U, V)]$ を最小化するものを求めよ。

Brenier の定理

$\mathscr{P}_2(\mathbb{R}^m)$ は $\mathbb{R}^m$ 上の連続確率分布で，二次モーメントをもつものの全体とし，$\mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m) \subset \mathscr{P}_2(\mathbb{R}^m)$ は Lebesgue 測度に対して絶対連続なものの全体とする。$\mathbb{R}^m \times \mathbb{R}^m$ 上の Borel 確率測度 π が $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^m)$ の結合分布 (coupling) であるとは，任意の Borel 集合 $B \in \mathcal{B}(\mathbb{R}^m)$ に対して

$$\pi(B \times \mathbb{R}^m) = \mu(B),$$
$$\pi(\mathbb{R}^m \times B) = \nu(B),$$

となることをいう。結合分布の全体を $\Pi(\mu,\nu)$ と書く。

$\mu \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$, $\nu \in \mathscr{P}_2(\mathbb{R}^m)$ とする。Brenier (1991) によれば，コスト関数 $c(x,y) = \frac{1}{2}|x-y|^2$ に対する MK 問題の解 $\pi \in \Pi(\mu,\nu)$ に対して，ある全単射写像 $T : \mathbb{R}^m \to \mathbb{R}^m$ で，$\pi(A \times B) = \mu(A \cap T^{-1}(B))\ \forall\, A, B \in \mathcal{B}(\mathbb{R}^m)$ となるものが存在する。さらに凸関数 $\phi : \mathbb{R}^m \to \mathbb{R}$ で，$T = \nabla\phi\ \mu$-a.e. となるものが存在する。

McCann (2001) はコンパクト Riemann 多様体の場合に，Figalli and Gigli (2010) は非コンパクト Riemann 多様体の場合に Brenier の定理を拡張した。ただし Riemann 多様体の場合，コスト関数は $c(x,y) = \frac{1}{2}d_g(x,y)^2$ となる。最適輸送写像 T に対して局所弱凸関数 ϕ が存在して $T = \exp(\nabla\phi)$ と表せる。ただし exp は Riemann 多様体の接空間に対して定義された指数写像である。また勾配ベクトル場に沿う輸送写像を

$$T_t(x) := \exp_x(t\nabla\phi), \quad x \in M,\ t \in [0,1]$$

とおくと，押出測度 $\mu_t := T_{t\sharp}\mu_0$ は μ_0 から μ_1 への Wasserstein 測地線になる。Wasserstein 測地線については後述する。

3.13.2 Wasserstein 幾何と Otto 解析

(M,g) を完備かつ連結な境界をもたない m 次元 Riemann 多様体とする。体積測度を vol_g とし，Riemann 距離関数を d と書く。M 上の $V \in C^\infty(M)$ による重み付き測度を

$$\mathfrak{m} := e^{-V}\mathsf{vol}_g,$$

と書く。$\mathfrak{m}$ に対して絶対連続な確率測度の全体を $\mathscr{P}^{\mathrm{ac}}(M)$ とし，さらに二次モーメントを持つものの全体を $\mathscr{P}_2^{\mathrm{ac}}(M)$ と書く。

Wasserstein 空間

(E,d) を Polish 空間，$p \in [1,\infty)$ とする。E 上の確率測度 μ,ν に対し，p-次 Wassserstein 距離は以下で定義される

$$W_p(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{E\times E} d^p(x,y)\pi(\mathrm{d}x\mathrm{d}y) \right\}^{1/p}.$$

特に $(\mathscr{P}_p(E), W_p)$ は可分な完備距離空間であり，$\mathscr{P}_p(E)$ に弱収束位相を入れた空間と同相である。つまり，$\mu_n \to \mu$ weak in $\mathscr{P}_p$ は $\lim_{n\to\infty} W_p(\mu_n,\mu) = 0$ と同値である。

Riemann 構造

Riemann 多様体 (M, g) 上の Wasserstein 空間 $(\mathscr{P}_2^{\mathrm{ac}}(M), W_2)$ には，Riemann 測地線が Wasserstein 距離と両立するような Riemann 構造 $\mathfrak{g} : T\mathscr{P}_2^{\mathrm{ac}}(M) \times T\mathscr{P}_2^{\mathrm{ac}}(M) \to \mathbb{R}$ が入る。すなわち，各点 $\mu \in \mathscr{P}_2^{\mathrm{ac}}(M)$ 上の接空間を

$$T_\mu \mathscr{P}_2^{\mathrm{ac}}(M) := \overline{\{\nabla\phi \mid \phi \in C_c^\infty(M)\}},$$

とし，その上の内積を

$$\mathfrak{g}_\mu(v, w) := \int_M g(v(x), w(x))\mu(\mathrm{d}x), \quad v, w \in T_\mu \mathscr{P}_2^{\mathrm{ac}}(M)$$

とおけばよい。ただし $C_c^\infty(M)$ は M 上の台がコンパクトかつ滑らかな関数を表す。閉包は上記内積から定まるノルムに関してとり，内積は連続的に拡張する。接空間の考え方は後述する。

上記設定のもとで，$\mu_0, \mu_1 \in \mathscr{P}_2^{\mathrm{ac}}(M)$ を結ぶ Wasserstein 測地線を μ_t として，以下が成り立つ

$$W_2(\mu_0, \mu_1)^2 = \int_0^1 \mathfrak{g}(\dot{\mu}_t, \dot{\mu}_t)\mathrm{d}t.$$

つまり，左辺は Wasserstein 空間の距離関数であり，右辺は Riemann 計量によって測った距離であって，等式は両者が一致することを意味している。

このように形式的な Riemann 計量によって確率空間を解析する方法は，創始者 Otto (2001) に因んで Otto 解析と呼ばれる。

接空間

$\mu \in \mathscr{P}_2^{\mathrm{ac}}(M)$ における接空間 $T_\mu \mathscr{P}_2^{\mathrm{ac}}(M)$ とは，μ を始点として $\mathscr{P}_2^{\mathrm{ac}}(M)$ に値をとる曲線 μ_t の初速ベクトル $\dot{\mu}_0$ を全て集めたベクトル空間である。

まず，μ_t に対応する M 上の輸送写像の速度ベクトルを v_t として，発散定理から次の連続の方程式が成り立つ

$$\frac{\partial}{\partial t}\mu_t = -\mathsf{div}_{\mathfrak{m}}(v_t\mu_t), \quad \mathfrak{m}\text{-a.e.}$$

ただし M 上のベクトル場 F に対し，$V \in C^\infty(M)$ を用いて

$$\mathsf{div}_{\mathfrak{m}}(F) := \mathsf{div}V - g(F, \nabla V)$$

と定義した。なお，μ_t を $\mathscr{P}_2^{\mathrm{ac}}(M)$ の点とみる場合には $\dot{\mu}_t$ と書き，M 上の測度とみる場合には $\frac{\partial}{\partial t}\mu_t$ と書いて区別している。この連続の方程式を通じて，$\mathscr{P}_2^{\mathrm{ac}}(M)$ 上の速度ベクトル $\dot{\mu}_t$ と M 上の速度ベクトル場 v_t を同一視できる。従って，$T_\mu\mathscr{P}_2^{\mathrm{ac}}(M)$ はベクトル場 v_0 の全体（と，線形同型）である。

一方，Riemann 多様体上の Brenier の定理によって，輸送写像の初速ベクトル v_0 は，ある関数 $\phi \in C_c^\infty(M)$ を用いて $v_0 = \nabla\phi$ で与えられる。従って，$T_\mu\mathscr{P}_2^{\mathrm{ac}}(M)$ は $\phi \in C_c^\infty(M)$ による勾配ベクトル場 $\nabla\phi$ の全体と同一視できることになる

$$T_\mu\mathscr{P}_2^{\mathrm{ac}}(M) \cong \overline{\{\nabla\phi \mid \phi \in C_c^\infty(M)\}}.$$

3.13.3 エントロピー勾配流

エントロピー汎関数

$\mu \in \mathscr{P}^{\mathrm{ac}}(M)$ の $\mathfrak{m}$ に対する相対エントロピーを

$$\mathsf{Ent}_{\mathfrak{m}}[\mu] := \int_M \rho\log\rho\,\mathrm{d}\mathfrak{m},$$

と定義する。ただし ρ は $\mu = \rho\mathfrak{m}$ なる関数（Radon-Nykodim 導関数）である。

Shannon 情報理論における標準的な定義との関係を調べる。簡単のため，$M = \mathbb{R}^m$ とし，$\mu, \mathfrak{m}$ はいずれも Lebesgue 測度 $\mathrm{d}x$ に対して絶対連続で，$\mu = p\,\mathrm{d}x, \mathfrak{m} = q\,\mathrm{d}x$ と書けるとする。

$$\begin{aligned}\mathsf{Ent}_{\mathfrak{m}}[\mu] &= \int_{\mathbb{R}^m} \frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \log \frac{\mathrm{d}\mu/\mathrm{d}x}{\mathrm{d}\mathfrak{m}/\mathrm{d}x}\,\mathrm{d}\mathfrak{m} \\ &= \int_{\mathbb{R}^m} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x \\ &= KL(p \parallel q).\end{aligned}$$

従って特に，$\mathfrak{m} = \mathrm{d}x\,(V \equiv 0)$ の場合には，負のエントロピーになる

$$\mathsf{Ent}_{\mathrm{d}x}[\mu] = -H[\mu].$$

エントロピーの勾配

$\mu_t = \rho_t \mathfrak{m}$ を $\mathscr{P}_2^{\mathrm{ac}}(M)$ 上の任意の Wasserstein 測地線とする。勾配の定義から以下が成り立つ

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Ent}_{\mathfrak{m}}[\mu_t] = \mathrm{d}\mathsf{Ent}_{\mathfrak{m}}(\dot{\mu}_t) = \mathfrak{g}_{\mu_t}(\mathsf{grad}\,\mathsf{Ent}_{\mathfrak{m}}, \dot{\mu}_t).$$

一方，左辺を直接計算して以下を示せる

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Ent}_{\mathfrak{m}}[\mu_t] = \frac{\mathrm{d}}{\mathrm{d}t}\int_M \rho_t(x)\log\rho_t(x)\mathrm{d}\mathfrak{m}(x) = \mathfrak{g}_{\mu_t}(\nabla\log\rho_t, \dot{\mu}_t).$$

従って，μ_t の任意性から，一般の $\mu = \rho\mathfrak{m} \in \mathscr{P}_2^{\mathrm{ac}}(M)$ に対して以下が導ける

$$\mathsf{grad}\,\mathsf{Ent}_{\mathfrak{m}}(\mu) = \nabla\log\rho.$$

エントロピー勾配流

$\mathsf{Ent}_{\mathfrak{m}}$ による $(\mathscr{P}_2^{\mathrm{ac}}(M), W_2)$ 上の勾配流を考える。すなわち，μ_t を $(\mathscr{P}_2^{\mathrm{ac}}(M), W_2)$ の未知曲線として，

$$\dot{\mu}_t = -\mathsf{grad}\,\mathsf{Ent}_{\mathfrak{m}}(\mu_t)$$

を満たすものを求める。

関係式 $\mathsf{grad}\,\mathsf{Ent}_{\mathfrak{m}}(\mu) = \nabla\log\rho$ に注意して，連続の方程式に $v_t = -\nabla\log\rho_t$ を代入すると，弱形式の熱方程式を得る

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_M h(x)\mathrm{d}\mu_t(x) = \int_M Lh(x)\mathrm{d}\mu_t(x), \quad h \in C_c^\infty(M)$$

ただし

$$Lh := \triangle h - g(\nabla V, \nabla h).$$

つまり，μ_t は熱方程式の弱解であることが分かる。

特に Euclid 空間の場合 $(M = \mathbb{R}^m, V \equiv 0)$ は，$L = \triangle$ になる。

第4章　ニューラルネットの積分表現理論

積分表現理論の基礎事項を整理する。本章では，単にニューラルネットといえば浅いニューラルネット（三層パーセプトロン）を指すものとする。

4.1　ニューラルネットの積分表現

ある関数 $f : \mathbb{R}^m \to \mathbb{C}$ をニューラルネット g で近似する。活性化関数を $\eta : \mathbb{R} \to \mathbb{C}$ として，J 個の中間層素子をもつニューラルネット g は以下の式で与えられる

$$g(x) := \sum_j^J c_j \, \eta(a_j \cdot x - b_j), \quad (a_j, b_j, c_j) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{C}$$

ここで，(a_j, b_j) を中間層パラメータまたは隠れ層パラメータ，c_j を係数または出力層パラメータと呼ぶ。中間層パラメータの空間 $\mathbb{R}^m \times \mathbb{R}$ を $\mathbb{Y}^{m+1}$ で表す。

ニューラルネットの積分表現は，通常のニューラルネットにおいて中間層素子数の総和を積分に置き換えたものである

$$g(x) := \int_{\mathbb{Y}^{m+1}} \mathrm{T}(a, b) \eta(a \cdot x - b) \mathrm{d}\mu(a, b),$$

ここで，$\mathrm{T} : \mathbb{Y}^{m+1} \to \mathbb{C}$ は出力層パラメータの連続版に相当し，μ は $\mathbb{Y}^{m+1}$ 上の適当な測度を表す。通常のニューラルネットは，積分表現を適当な方法で離散化して得られたものと考える。形式上は，測度 μ として適当な特異測度

$$\mu(a, b) := \sum_j^J w_j \delta(a - a_j) \delta(b - b_j), \quad w_j \in \mathbb{C}$$

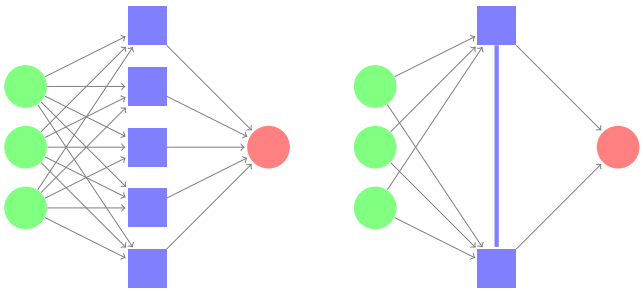


図 4.1: 通常のニューラルネット（左）と積分表現（右）

をとることで有限和もまた積分表現の一種とみなせる。

積分表現の式は，リッジレット関数 η による関数 T の**双対リッジレット変換** $\mathscr{R}_\eta^\dagger \mathrm{T}$ として知られる。積分表現の係数である関数 T として，目的関数 f のリッジレット関数 ψ によるリッジレット変換 $\mathscr{R}_\psi f$ をとると，再構成公式が成り立つ

$$\int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi f(a,b)\eta(a\cdot x-b)\mathrm{d}a\mathrm{d}b = f(x), \quad x\in\mathbb{R}^m.$$

ただし ψ と η は後述する**許容条件**を満たしているものとする。すなわち，関数 f のリッジレット変換を係数とするニューラルネット（$\mathrm{T}\leftarrow\mathscr{R}_\psi f$）は，関数 f として振る舞うことを意味する。積分表現 $\mathscr{R}_\eta^\dagger\mathscr{R}_\psi f$ を離散化することで，関数 f を近似するニューラルネットが得られる。

4.2　リッジレット解析

関数 $f:\mathbb{R}^m\to\mathbb{C}$ の $\psi:\mathbb{R}\to\mathbb{C}$ によるリッジレット変換 $\mathscr{R}_\psi f$ は以下の形式的な積分で与えられる

$$\mathscr{R}_\psi f(a,b) := \int_{\mathbb{R}^m} f(x)\overline{\psi(a\cdot x-b)}|a|^s\mathrm{d}x, \quad (a,b)\in\mathbb{Y}^{m+1} \text{ and } s>0.$$

指数 s の取り方は，研究者によって異なる。本研究では $s=1$ を用いる。このとき例えば 定理 5.2.1 の見た目が簡単になる。Murata (1996) によるオリジナルの定義は $s=0$ に相当する。このときユークリッド座標による定式化が簡単になる。この他，Candès (1998) は $s=1/2$, Rubin (1998) は $s=m$, Kostadinova et al. (2014) は $s=1$ を用いている。

$f \in L^1(\mathbb{R}^m)$ かつ $\psi \in L^\infty(\mathbb{R})$ のとき，リッジレット変換 $\mathscr{R}_\psi f(a,b)$ は各点 $(a,b) \in \mathbb{Y}^{m+1}$ で絶対収束する。証明は Hölder の不等式から直ちに従う

$$\int_{\mathbb{R}^m} \left| f(x)\overline{\psi(a \cdot x - b)} \right| |a|^s \mathrm{d}x \leqslant \|f\|_{L^1(\mathbb{R}^m)} \cdot \|\psi\|_{L^\infty(\mathbb{R})} |a|^s < \infty.$$

特に $s = 0$ のときは a に依らない評価になるので，$\mathscr{R}_\psi f \in L^\infty(\mathbb{Y}^{m+1})$ が言える。さらに，$\mathscr{R}$ は有界双線形作用素 $L^1(\mathbb{R}^m) \times L^\infty(\mathbb{R}) \to L^\infty(\mathbb{Y}^{m+1})$ である。

関数 $\mathrm{T} : \mathbb{Y}^{m+1} \to \mathbb{C}$ の $\eta : \mathbb{R} \to \mathbb{C}$ による双対リッジレット変換 $\mathscr{R}_\eta^\dagger \mathrm{T}$ は以下の形式的な積分で与えられる

$$\mathscr{R}_\eta^\dagger \mathrm{T}(x) := \int_{\mathbb{Y}^{m+1}} \mathrm{T}(a,b)\eta(a \cdot x - b)|a|^{-s} \mathrm{d}a\mathrm{d}b, \quad x \in \mathbb{R}^m.$$

$\eta \in L^\infty(\mathbb{R})$ かつ $\mathrm{T} \in L^1(\mathbb{Y}^{m+1}; |a|^{-s}\mathrm{d}a\mathrm{d}b)$ のとき，双対リッジレット変換 $\mathscr{R}_\eta^\dagger \mathrm{T}(x)$ は各点 $x \in \mathbb{R}^m$ で絶対収束する

$$\int_{\mathbb{Y}^{m+1}} \left| \mathrm{T}(a,b)\eta(a \cdot x - b) \right| |a|^{-s} \mathrm{d}a\mathrm{d}b \leqslant \|\mathrm{T}\|_{L^1(\mathbb{Y}^{m+1}; |a|^{-s}\mathrm{d}a\mathrm{d}b)} \cdot \|\eta\|_{L^\infty(\mathbb{R})} < \infty,$$

このとき，$\mathscr{R}^\dagger$ は有界双線形作用素 $L^1(\mathbb{Y}^{m+1}; |a|^{-s}\mathrm{d}a\mathrm{d}b) \times L^\infty(\mathbb{R}) \to L^\infty(\mathbb{R}^m)$ である。適当な条件のもとで，η による双対リッジレット変換 $\mathscr{R}_\eta^\dagger$ は，η によるリッジレット変換 $\mathscr{R}_\eta$ の双対作用素である。つまり，適当な双対積 $\langle \cdot, \cdot \rangle$ の下で，以下の等式が成り立つ

$$\langle \mathscr{R}_\eta f, \mathrm{T} \rangle = \langle f, \mathscr{R}_\eta^\dagger \mathrm{T} \rangle.$$

関数 ψ と η は以下の積分 $K_{\psi,\eta}$ が有界かつ非零のとき許容的であるという

$$K_{\psi,\eta} := (2\pi)^{m-1} \int_{-\infty}^{\infty} \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m} \mathrm{d}\zeta.$$

ψ, η が許容条件を満たすとき，適当な正則性条件の下で再構成公式が成り立つ

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f = K_{\psi,\eta} f.$$

4.3　リッジレット変換の幾何学的解釈

リッジレット変換は Radon 変換とウェーブレット変換の合成変換に分解できる。この事実から，しばしば「リッジレット変換は Radon 領域のウェーブレット解析」と言われる (Starck et al., 2010; Kostadinova et al., 2014)。

4.3.1　リッジレット変換の分解

まず，パラメータ空間の極座標を次のようにとる

$$u := a/|a|, \quad \alpha := 1/|a|, \quad \beta := b/|a|.$$

ただし，ウェーブレット変換との繋がりを強調するために，スケール変数 α はあえて逆数にとった。混乱の恐れが無い限り，パラメータ空間は座標系の取り方に依らず $\mathbb{Y}^{m+1}$ と書く。極座標表示のもとで，リッジレット変換は以下のようになる

$$\mathscr{R}_\psi f(u,\alpha,\beta) = \int_{\mathbb{R}^m} f(x)\overline{\psi\left(\frac{u\cdot x-\beta}{\alpha}\right)}\frac{1}{\alpha^s}\mathrm{d}x.$$

パラメータ $(u,\alpha,\beta) \in \mathbb{Y}^{m+1}$ を固定する。ψ から生成されるウェーブレット関数を以下のように書く

$$\psi_\alpha(p) := \psi\left(\frac{p}{\alpha}\right)\frac{1}{\alpha^s}.$$

u と平行な元 $pu\,(p\in\mathbb{R})$，u と直交する元 $y \in (\mathbb{R}u)^\perp$ として，$x = pu+y$ と直交分解すると，リッジレット変換の積分核は以下のように分解できる

$$\begin{aligned}\psi\left(\frac{u\cdot x-\beta}{\alpha}\right)\frac{1}{\alpha^s} &= \psi\left(\frac{u\cdot(pu+y)-\beta}{\alpha}\right)\frac{1}{\alpha^s}\\ &= \psi_\alpha(p-\beta)\otimes 1(y).\end{aligned}$$

つまり，積分核は $(\mathbb{R}u)^\perp$ 上では定数関数 1，$\mathbb{R}u$ 上ではウェーブレット関数 ψ_α として振る舞う。

この分解に従って積分の順序を変更すると，リッジレット変換の分解を得る。

$$\begin{aligned}\mathscr{R}_\psi f(u,\alpha,\beta) &= \int_{\mathbb{R}} \left(\int_{(\mathbb{R}u)^\perp} f(pu+y)\mathrm{d}y \right) \overline{\psi\left(\frac{p-\beta}{\alpha}\right)} \frac{1}{\alpha^s}\mathrm{d}p \\ &= \int_{\mathbb{R}} \mathrm{R}f\,(u,p)\,\overline{\psi_\alpha(p-\beta)}\mathrm{d}p \\ &= \int_{\mathbb{R}} \alpha^{1-s}\mathrm{R}f\,(u,\alpha z+\beta)\,\overline{\psi(z)}\mathrm{d}z \\ &= \left(\mathrm{R}f(u,\cdot) * \widetilde{\overline{\psi_\alpha}}\right)(\beta).\end{aligned}$$

ただし，R は Radon 変換を表す。Fubini の定理により，いずれか一つの積分が絶対収束することが示されれば，すべての等号が成り立つ（つまり，積分の順序変更は有効である）。右辺の式はいずれも，リッジレット変換が Radon 変換とウェーブレット変換の合成変換であることを表している。

さらに，畳み込み形式のリッジレット変換に対して Fourier 変換の恒等式 $\mathcal{F}^{-1}\mathcal{F} = \mathsf{id}$ を適用することで，**リッジレット変換に対する投影切断面定理**（Fourier slice theorem）を得る

$$\mathscr{R}_\psi f(u,\alpha,\beta) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\omega u)\overline{\widehat{\psi}(\alpha\omega)}\alpha^{1-s}e^{i\omega\beta}\mathrm{d}\omega.$$

4.3.2 双対リッジレット変換の分解

リッジレット変換の場合と同様に極座標表示することで，双対リッジレット変換は双対ウェーブレット変換の双対 Radon 変換であることが分かる。積分の順序を変更するため，双対リッジレット変換は絶対収束しているものとする。

$$\begin{aligned}\mathscr{R}_\eta^\dagger \mathrm{T}(x) &= \int_{\mathbb{R}^m}\int_{\mathbb{R}} \mathrm{T}(a,b)\eta(a\cdot x-b)|a|^{-s}\mathrm{d}b\,\mathrm{d}a \\ &= \int_0^\infty \int_{\mathbb{S}^{m-1}}\int_{\mathbb{R}} \mathrm{T}(ru,b)\eta(ru\cdot x-b)\mathrm{d}b\,\mathrm{d}u\,r^{m-s-1}\mathrm{d}r \\ &= \int_{\mathbb{S}^{m-1}}\int_0^\infty \int_{\mathbb{R}} \mathrm{T}\left(\frac{u}{\alpha},\frac{\beta}{\alpha}\right)\eta\left(\frac{u\cdot x-\beta}{\alpha}\right)\frac{\mathrm{d}\beta\mathrm{d}\alpha\mathrm{d}u}{\alpha^{m-s+2}} \\ &= \int_{\mathbb{S}^{m-1}}\int_0^\infty \int_{\mathbb{R}} \mathrm{T}\left(u,\alpha,u\cdot x-\alpha z\right)\eta(z)\frac{\mathrm{d}z\mathrm{d}\alpha\mathrm{d}u}{\alpha^{m-s+1}}.\end{aligned}$$

積分の変数変換は二段階で行う。まず二番目の式では $(r,u) \leftarrow (|a|, a/|a|)$ と変換して，極座標変換に対する余面積公式を用いる。次に三番目の式では $(\alpha,\beta) \leftarrow (1/r, b/r)$ と変換して，Fubini の定理によって積分の存在を保証する。なお四番目の式では簡単のため記号を濫用して $\mathrm{T}(u,\alpha,\beta) := \mathrm{T}(u/\alpha, \beta/\alpha)$ と書いている。

4.4　離散化の考え方

積分表現の離散化の考え方を整理する。

4.4.1　辞書と係数

活性化関数 η から生成される全ての中間層素子 $h(x;a,b) := \eta(a\cdot x - b)$ を集めた辞書を $\mathscr{D}$ と書く

$$\mathscr{D} := \{h(\,\cdot\,;a,b) \mid (a,b) \in \mathbb{Y}^{m+1}\}.$$

関数 $\mathrm{T} : \mathbb{Y}^{m+1} \to \mathbb{C}$ と，部分辞書 $D \subset \mathscr{D}$ との内積を以下で定義する

$$D\mathrm{T}(x) := \int_{\phi(D)} \mathrm{T}(a,b)h(x;a,b)\mathrm{d}\mu(a,b), \quad x \in \mathbb{R}^m$$

ただし，$\phi : D \to \mathbb{Y}^{m+1}$ は関数 $h(\,\cdot\,;a,b)$ をパラメータ (a,b) に対応付ける局所座標系（local coordinate system）とする。測度 μ は文脈に応じて適宜定めるものとする。特に，D が有限辞書の場合は有限和を表すものと考える

$$D\mathrm{T}(x) = \sum_{\phi(D)} \mathrm{T}(a,b)h(x;a,b),$$

これは通常のニューラルネットである。また，辞書 $\mathscr{D}$ 自体による内積は，測度 μ を適当に定めたうえで，双対リッジレット変換と等しい

$$\mathscr{D}\mathrm{T}(x) = \mathscr{R}_\eta^\dagger \mathrm{T}(x), \quad x \in \mathbb{R}^m.$$

4.4.2 離散化の評価

辞書 $\mathscr{D}$ から高々可算濃度の辞書 $D \subset \mathscr{D}$ を選出することを**離散化** (discretization) という。一般に辞書 $\mathscr{D}$ を離散化する方法は一通りではない。例えば $\overline{\mathsf{span}\,\mathscr{D}}$ ($\bar{\cdot}$ は閉包を表す) が Hilbert 空間になる場合には，D として正規直交基底がとれる。ただし Donoho (2001) なども指摘する通り，リッジレット変換は一般に L^1 関数に対してのみ定義されるので，'正規直交' リッジレットなるものは存在しない。従って，実際にはアトム分解や分子分解を求めることになる。

離散化に伴って，辞書の表現能力は劣化する。関数近似という観点では，適当な関数ノルムで辞書の表現能力を評価するのが基本的である。例えばバックプロパゲーションによる関数近似は，データの分布で重み付けられた L^2 ノルムによって評価していることに相当する。一方，数値解析では，正則性を評価するために Sobolev ノルム (H^1_0 など) や，最悪評価の観点から L^∞ ノルムを用いることもある。また，機械学習では，単に表現能力だけではなく，学習後の汎化能力や停留点の安定性，学習の速さやサンプル複雑性など，複数の観点から総合的に評価する必要がある。

4.4.3 離散化の考え方

基底やフレームは，目的の関数族に含まれる全ての関数を一様に近似するための離散化である。一方，機械学習ではよくある設定だが，ある特定の関数 f_0 を近似することが求められている場合には，基底やフレームは冗長である。つまり，辞書 D は，単に f_0 を近似するのに必要な元を保持していれば十分である。このような辞書は，任意の関数を近似するためには全く不十分であっても構わない。バックプロパゲーションや辞書学習は，フレームですらない辞書を選出する方法である。

バックプロパゲーション

ニューラルネットのバックプロパゲーション学習は，以下の最適化問題を勾配法で解くことと等価である

$$\begin{aligned} &\text{minimize} \quad \mathbb{E}|D\mathrm{T}(X) - Y|^2 \quad \text{w.r.t. } D \text{ and } \mathrm{T} \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{s.t. } |D| \leqslant n \text{ and } |\mathrm{T}|_p \leqslant \lambda. \end{aligned} \tag{4.1}$$

ただし $|D|$ は辞書 $D(\subset \mathscr{D})$ の濃度を表し，この制約により中間層素子が予め有限個に固定されている状態を表す。また $|\mathrm{T}|_p \leqslant \lambda$ は係数に関する ℓ^p-正則化を表す。

Maurey-Jones-Barron 評価

辞書 $\mathscr{D}$ から高々 n 個以下の元を取り出して得られる凸包の全体を $\overline{\mathsf{conv}_n \mathscr{D}}$ と書く。ただし $\bar{\cdot}$ は閉包を表す。次の近似誤差の評価が知られている (Kůrková, 2012, Cor.5.4)

$$\inf_{g \in \overline{\mathsf{conv}_n \mathscr{D}}} \|\mathscr{D}\mathrm{T} - g\|_2 \leqslant \sqrt{\frac{\sup_{h \in \mathscr{D}} \|h\|_2^2 \cdot \|\mathrm{T}\|_1^2 - \|\mathscr{D}\mathrm{T}\|_2^2}{n}},$$

ただし，関数ノルムはいずれも有界領域においてとる。これを Maurey-Jones-Barron（MJB）評価と言う (Kainen et al., 2013)。証明は，古典的には逐次貪欲近似法（Matching Pursuit）を理論的に実行する方針で示せる (Jones, 1992)。Kůrková (2012) はさらに精密な評価の系として導いている。

MJB 評価において $\mathrm{T} = \mathscr{R}_\psi f$ とし，$\mathscr{D}\mathrm{T} = f$ とおくと，系として以下を得る

$$\inf_{g \in \overline{\mathsf{conv}_n \mathscr{D}}} \|f - g\|_2 \leqslant \frac{1}{\sqrt{n}} \sup_{h \in \mathscr{D}} \|h\|_2 \cdot \|\mathscr{R}_\psi f\|_1.$$

これはニューラルネットの中間層素子数を見積もるための手がかりとなる。また，$\|\mathscr{R}_\psi f\|$ は ψ を選択するための評価関数となることが分かる。

辞書学習

辞書学習は以下のように定式化できる

$$\text{minimize} \quad \mathbb{E}_f \left[\inf_{|\mathrm{T}|_p \leqslant \lambda} \|D\mathrm{T} - f\|_2 \right] \quad \text{w.r.t. } D \quad \text{s.t. } |D| \leqslant n.$$

すなわち，よく現れる関数 f に対して重点的に対応できるように，辞書を学習する。関数の分布に対する汎化誤差を最小化することになるので，各点近似と一様近似の中間的な方法である。辞書学習のサンプル複雑性は Vainsencher et al. (2010); Maurer and Pontil (2010); Seibert et al. (2014); Gribonval et al. (2015) が調べている。

オラクル・サンプリング

測度 μ を，確率密度関数を持つ確率測度とする。以下では，μ の確率密度関数と確率分布を区別せず μ と書く。

係数 T を以下を満たすようにとる

$$\mathscr{R}_\psi f(a,b) = \mathrm{T}(a,b)\mu(a,b).$$

確率分布 μ に従う n 個のサンプルを $(a_j, b_j)(j = 1, \cdots, n)$ として，これらのサンプルから生成される辞書を

$$D_n := \{h(\,\cdot\,; a_j, b_j) \mid (a_j, b_j) \sim \mu,\, j = 1, \cdots, n\}$$

とおく。このとき大数の法則により，以下が成り立つ

$$\begin{aligned}\frac{1}{n} D_n \mathrm{T}(x) &= \frac{1}{n}\sum_{j=1}^{n} \mathrm{T}(a_j, b_j) h(x; a_j, b_j) \\ &\to^p \int_{\mathbb{Y}^{m+1}} \mathrm{T}(a,b) h(x; a, b) \mathrm{d}\mu(a,b) \quad \text{as} \quad n \to \infty \\ &= \mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) = f(x).\end{aligned}$$

このように，リッジレット変換 $\mathscr{R}_\psi f$ を確率測度 μ と係数 T に分解して，μ からサンプリングする方法で辞書 D_n が得られる。こうして得られる測度 μ を**オラクル分布**（oracle distribution）と呼ぶ (Sonoda and Murata, 2014)。

4.4.4 バックプロパゲーションとの関係

積分表現を離散化したものは通常のニューラルネットである。では逆に，バックプロパゲーションによって学習したニューラルネットは必ず積分表現の離散化とみなせるのであろうか。つまり，近似対象の関数 f を既知として，任意の (4.1) 極小点 g に対して，あるリッジレット関数 ψ と離散化アルゴリズム $\mathcal{A}$ が存在して，$\mathcal{A}[\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f] = g$ とできるのだろうか。

図 4.2 に示す数値実験の結果から，この問に対する回答はおそらく肯定的である。まず左図は，$f(x) = \sin 2\pi x,\ x \in [-1, 1]$ に対して，中間層素子が 10 個のニューラルネット $g(x) = \sum_{j=1}^{10} c_j \eta(a_j \cdot x - b_j)$ を 1,000 体学習して，学習後の中間層パラメータ $\{(a_j, b_j)\}_{j=1}^{10}$ をパラメータ空間 $\mathbb{Y}^2$ に 1,000 体分（= 10,000 点）プロットした散布図である。一方，右図は，

同じ $f(x)$ のリッジレット変換 $\mathscr{R}_\psi f(a,b)$ を数値的に計算したものである。$\mathscr{R}_\psi f(a,b)$ の値が高い位置にあるパラメータ (a,b) は，実際のニューラルネットでも使われやすいことが分かる。従って，バックプロパゲーションによるパラメータは，確率的にはある ψ による積分表現の離散化になっていることが期待される。

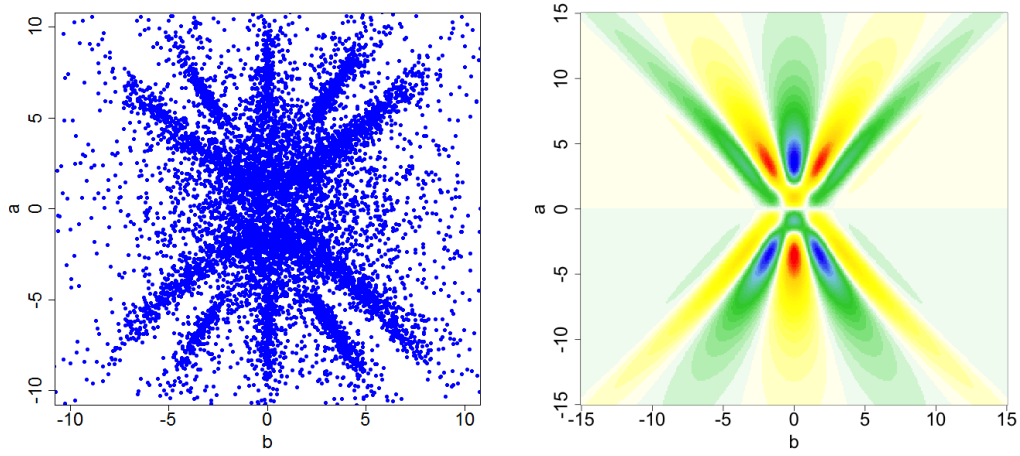


図 4.2: バックプロパゲーションによって学習されたパラメータとリッジレット変換の比較

4.5　ベクトル値の情報共有

出力がベクトル値 $(f_1, \cdots, f_k)$ の場合，成分毎に学習する方法が基本的である。しかし，カラー画像を写像 $\mathbb{R}^2 \to \mathbb{R}^3$ として近似する場合や，教師データの一部に欠損値があって補間する必要がある場合は，共通の辞書 D を利用する方が効率が良い。

ベクトル値関数の学習問題は，マルチタスク学習や転移学習として定式化する。なお，仮に元のデータが行列形式で与えられている場合は，PCA や ICA, NMF などの行列分解に帰着することが多い。また，マルチカーネル学習（Multiple Kernel Learning; MKL）や，カーネル関数を共分散行列とする Gauss 過程（Gaussian Process; GP）として扱う方法もある。

4.5.1 グループ正則化によるマルチタスク学習

本節に限り, ベクトル値変数を強調して太字で表す。ベクトル値のニューラルネット

$$\mathbf{g}(\mathbf{x}) = \sum_{j=1}^{J} \mathbf{c}_j \, \eta(\mathbf{a}_j \cdot \mathbf{x} - b_j),$$

に対し，グループ正則化項 $\Omega[\mathbf{g}]$ は以下で与えられる

$$\Omega[\mathbf{g}] := \sum_{j=1}^{J} |\mathbf{c}_j|_p.$$

ただし典型的には $p = 2$ にとる。つまり，積分表現では次のようになる

$$\Omega[\mathbf{g}] = \int_{\mathbb{Y}^{m+1}} |\mathbf{T}(a,b)|_p \mathrm{d}\mathbf{a}\mathrm{d}b = \||\mathbf{T}|_p\|_1.$$

ベクトル値のニューラルネットのグループ正則化付きバックプロパゲーション学習は，辞書学習と形式的に等価である

$$\text{minimize} \quad \mathbb{E}_{\mathbf{f}} \left[\inf_{\||\mathbf{T}|_p\|_1 \leqslant \lambda} \|D\mathbf{T} - \mathbf{f}\|_2 \right] \quad \text{w.r.t. } D \quad \text{s.t. } |D| \leqslant n.$$

4.5.2 輸送写像の場合

Brenier の定理により，任意の最適輸送写像 $f : \mathbb{R}^m \to \mathbb{R}^m$ は，ある凸関数 $\mathcal{E}$ を用いて

$$f(x) = x + \nabla \mathcal{E}(x) = \nabla \mathcal{F}(x)$$

と書ける。ただし $\mathcal{F} := \frac{1}{2}|\cdot|^2 + \mathcal{E}$ とおいた。以下ではリッジレット変換を計算する都合上，f および $\mathcal{E}, \mathcal{F}$ はコンパクト集合 K 上でのみ値をとるものとする。$\frac{\mathrm{d}}{\mathrm{d}z}\Psi(z) = \psi(z)$ なるリッジレット関数を用いて，輸送写像 f のリッジレット変換は以下のように計算できる

$$\begin{aligned} \mathscr{R}_\Psi f(a,b) &= \int_K \nabla \mathcal{F}(x) \overline{\Psi(a \cdot x - b)} \mathrm{d}x \\ &= -a \int_K \mathcal{F}(x) \overline{\psi(a \cdot x - b)} \mathrm{d}x = -a \, \mathscr{R}_\psi \mathcal{F}(a,b). \end{aligned}$$

すなわち，輸送写像の成分毎のリッジレット変換は，ポテンシャル関数 $\mathcal{F}$ の情報を共有していることを示唆している。特に，グループ正則化項は以下で与えられる

$$\left\| \left| \mathscr{R}_{\Psi} f \right|_p \right\|_1 = |a|_p \left\| \mathscr{R}_{\psi} \mathcal{F} \right\|_1.$$

第5章　有界でない活性化関数のための積分表現理論

5.1　はじめに

深層ニューラルネットでは，活性化関数 $\eta(z)$ として ReLU（rectified linear unit）z_+ を用いるのが標準的である。ところが，従来のリッジレット解析 (Murata, 1996; Candès, 1998) では，ReLU のように有界でない活性化関数（リッジレット関数）は想定されていない。従来用いられてきたシグモイド関数や RBF などと比較して，ReLU は (1) 実装が `if` 文一つで済むので計算コストが削減できる (2) 誤差信号が飽和しないので学習が加速する (3) 学習結果がスパースになるというメリットがあるため，現在では標準的に用いられている。深層時代のニューラルネットを扱うために，(Sonoda and Murata, 2015) では，活性化関数が Lizorkin 超関数の場合にもリッジレット解析が展開できることを示した。

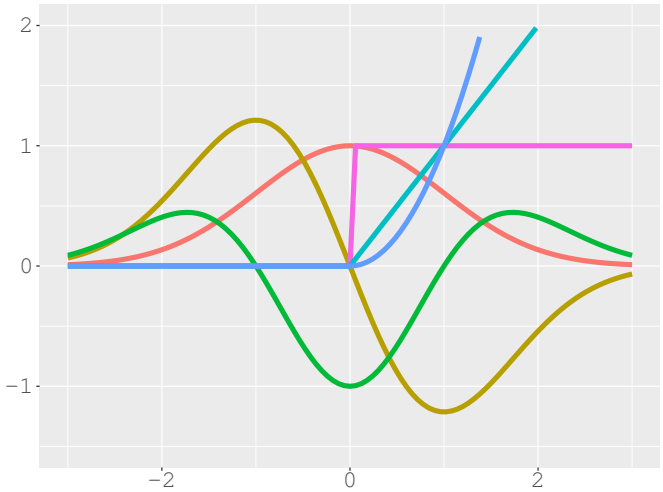


図　5.1:　Lizorkin 超関数の例。Gauss 関数 $G(z)$ およびその導関数 $G'(z), G''(z)$，切断べき関数 z_+^0, z_+, z_+^2 が含まれる。

表 5.1: Lizorkin 超関数に属する活性化関数の例

活性化関数	$\eta(z)$	$\mathcal{W}$
非有界関数		
切断べき関数	$z_+^k := \begin{cases} z^k & z > 0 \\ 0 & z \leqslant 0 \end{cases}, \quad k \in \mathbb{N}_0$	$\mathcal{S}_0'$
ReLU	$z_+ (= z_+^1)$	$\mathcal{S}_0'$
ソフトプラス関数	$\sigma^{(-1)}(z) := \log(1 + e^z)$	$\mathcal{O}_{\mathcal{M}}$
可積分でない有界関数		
ステップ関数	z_+^0	$\mathcal{S}_0'$
シグモイド関数	$\sigma(z) := (1 + e^{-z})^{-1}$	$\mathcal{O}_{\mathcal{M}}$
双曲線正接関数	$\tanh(z)$	$\mathcal{O}_{\mathcal{M}}$
隆起 (bump) 関数		
RBF	$\mathrm{G}(z) := (2\pi)^{-1/2} \exp(-z^2/2)$	$\mathcal{S}$
シグモイド関数の導関数	$\sigma'(z)$	$\mathcal{S}$
Dirac's δ	$\delta(z)$	$\mathcal{S}_0'$
振動的 (oscillatory) 関数		
RBF の導関数	$\mathrm{G}^{(k)}(z)$	$\mathcal{S}$
シグモイド関数の導関数	$\sigma^{(k)}(z)$	$\mathcal{S}$
Dirac's δ の導関数	$\delta^{(k)}(z)$	$\mathcal{S}_0'$

Lizorkin 超関数

Lizorkin 超関数 $\mathcal{S}_0'(\mathbb{R})$ は，ReLU z_+ だけでなく，ReLU を含む切断べき関数（truncated power functions）z_+^k，そして無限遠での増大度が高々多項式オーダーの関数を含む広大なクラスである（表 5.1）。$\mathcal{S}_0'(\mathbb{R})$ の非零元とは，緩増加超関数 $\mathcal{S}'(\mathbb{R})$ であって，多項式関数でないものである[1]。活性化関数 η が多項式関数の場合にはリッジレット変換の再構成公式が成り立たないので，多項式を除くことは本質的である。

[1] 多項式関数は定数関数 0 と同一視されるので，多項式関数が含まれないわけではない。

5.1.1 方針

ReLU $\eta(z) = z_+$ をうまく組み合わせると有界関数になる。例えば，$h > 0$ に対して

$$\eta^\star(z) := \eta(z+h) - \eta(z),$$

とおけば良い。右辺は差分作用素 $[\Delta^h f](z) := f(z+h) - f(z)$ を用いて

$$\eta^\star = \Delta^h \eta,$$

と書けることに注意する。$\eta^\star$ は有界関数なので，従来のリッジレット解析の意味で再構成公式が成り立つ。

$$f(x) = \int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi f(a,b)\eta^\star(a\cdot x - b)\mathrm{d}a\mathrm{d}b. \tag{5.1}$$

ここで，積分表現が b に関して畳み込みの形をしていることに着目する。畳み込みは差分作用素と可換なので，形式的に以下が成り立つことが期待される。

$$\begin{aligned}(5.1) &= \int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi f(a,b)\Delta^h[\eta](a\cdot x - b)\mathrm{d}a\mathrm{d}b \\ &= \int_{\mathbb{Y}^{m+1}} \Delta^h[\mathscr{R}_\psi f](a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b.\end{aligned}$$

最後の式は ReLU による積分表現である。すなわち，ReLU を用いるためには

$$\begin{aligned}\mathrm{T}(a,b) &:= \Delta^h[\mathscr{R}_\psi f](a,b) \\ &= -\int_{\mathbb{R}^m} f(x)\overline{[\Delta^h\psi](a\cdot x - b)}\mathrm{d}x \\ &= -\mathscr{R}_{\Delta^h\psi} f(a,b).\end{aligned}$$

となるように係数をとればよい。本章では，差分作用素を微分作用素に置き換えたうえで，上記の積分が存在することを示す。

5.2 超関数によるリッジレット変換

前章で導いたリッジレット変換の極座標表示において，積分を超関数の作用とみなして，超関数によるリッジレット変換を定義する

5.2.1　超関数によるリッジレット変換の定義と存在

定義 5.2.1. $f \in \mathcal{X}(\mathbb{R}^m)$ の $\psi \in \mathcal{Z}(\mathbb{R})$ によるリッジレット変換 $\mathscr{R}_\psi f$ とは，各点 $(u, \alpha, \beta) \in \mathbb{Y}^{m+1}$ に対して値

$$\mathscr{R}_\psi f(u, \alpha, \beta) = \left(\mathrm{R}f(u, \cdot) * \widetilde{\overline{\psi_\alpha}}\right)(\beta), \quad (u, \alpha, \beta) \in \mathbb{Y}^{m+1} \tag{5.2}$$

を対応させる写像である。ただし $\psi_\alpha(p) := \psi(p/\alpha)/\alpha$ とする。

この定義において，畳み込み $(\cdot * \cdot)$, スケーリング $(\cdot_\alpha)$, 反転 $(\widetilde{\cdot})$, 複素共役 $(\overline{\cdot})$ はすべて超関数の意味である。つまり，次のように定義することと同値である

$$\mathscr{R}_\psi f(u, \alpha, \beta) := \int \overline{\mathrm{R}f(u, \alpha z + \beta)} \psi(z) \mathrm{d}z, \quad (u, \alpha, \beta) \in \mathbb{Y}^{m+1} \tag{5.3}$$

ただし “積分” $\int_{\mathbb{R}} \cdot \psi(z) \mathrm{d}z$ は汎関数 ψ の作用の意味であって，必ずしも Lebesgue の意味の積分とは限らない。作り方から明らかに，ψ が局所可積分関数 (L^1_{loc}) であれば，超関数 ψ によるリッジレット変換は，従来の関数 ψ によるリッジレット変換と一致する。

定理 5.2.1. 表 5.2 に示す $\mathcal{X}, \mathcal{Z}$ の組み合わせにおいて，$\mathscr{R} : \mathcal{X}(\mathbb{R}^m) \times \mathcal{Z}(\mathbb{R}) \to \mathcal{Y}(\mathbb{Y}^{m+1})$ は $\mathbb{Y}^{m+1}$ の各点 (u, α, β) で存在し，双線形写像である。

表 5.2: リッジレット変換が定義できる $\mathcal{X}$ と $\mathcal{Z}$ の組み合わせおよび対応する値域。$\mathcal{B}, \mathcal{A}, \mathcal{Y}$ は $\mathscr{R}_\psi f(u, \alpha, \beta)$ をそれぞれ $\beta, (\alpha, \beta), (u, \alpha, \beta)$ の関数とみなした場合のクラスを表す。

$f(x)$	$\mathrm{R}f(u, p)$	$\psi(z)$	$\mathscr{R}_\psi f(u, \alpha, \beta)$		
$\mathcal{X}(\mathbb{R}^m)$	$\mathcal{X}(\mathbb{S}^{m-1} \times \mathbb{R})$	$\mathcal{Z}(\mathbb{R})$	$\mathcal{B}(\mathbb{R})$	$\mathcal{A}(\mathbb{H})$	$\mathcal{Y}(\mathbb{Y}^{m+1})$
$\mathcal{D}$	$\mathcal{D}$	$\mathcal{D}'$	$\mathcal{E}$	$\mathcal{E}$	$\mathcal{E}$
$\mathcal{E}'$	$\mathcal{E}'$	$\mathcal{D}'$	$\mathcal{D}'$	$\mathcal{D}'$	$\mathcal{D}'$
$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}'$	$\mathcal{O}_{\mathcal{M}}$	$\mathcal{O}_{\mathcal{M}}$	$\mathcal{O}_{\mathcal{M}}$
$\mathcal{O}'_{\mathcal{C}}$	$\mathcal{O}'_{\mathcal{C}}$	$\mathcal{S}'$	$\mathcal{S}'$	$\mathcal{S}'$	$\mathcal{S}'$
L^1	L^1	$L^p \cap C$	$L^p \cap C$	$\mathcal{S}'$	$\mathcal{S}'$

証明は各クラスの定義に従って “積分” の収束性を確認する。詳細は 付録 A.2 を見よ。表において $\mathcal{Z}$ の大きさは畳み込み $\mathcal{B} = \mathcal{X} * \mathcal{Z}$ が存在す

る範囲で最大のものをとった。畳み込みにより， $\mathcal{X}$ と $\mathcal{Z}$ の大小関係にはトレードオフの関係がある。

5.2.2 リッジレット変換の性質

連続性

命題 5.2.2 ($L^1(\mathbb{R}^m) \to L^\infty(\mathbb{Y}^{m+1})$ の連続性)**.** $\psi \in \mathcal{S}(\mathbb{R})$ を固定する。このときリッジレット変換 $\mathscr{R}_\psi : L^1(\mathbb{R}^m) \to L^\infty(\mathbb{Y}^{m+1})$ は有界作用素である。

Proof. $f \in L^1(\mathbb{R}^m)$ と $\psi \in \mathcal{S}(\mathbb{R})$ を任意にとって固定する。$\mathscr{R}_\psi f(u, \alpha, \beta)$ は各点で絶対収束している。畳み込み形式に対して Young の不等式を適用して，以下を得る

$$\begin{aligned}\operatorname*{ess\,sup}_{(u,\alpha,\beta)} \left| \left(\mathrm{R}f(u,\cdot) * \widetilde{\psi_\alpha} \right)(\beta) \right| &\leqslant \|f\|_{L^1(\mathbb{R}^m)} \cdot \operatorname*{ess\,sup}_{(\alpha,\beta)} |\psi_\alpha(\beta)| \\ &\leqslant \|f\|_{L^1(\mathbb{R}^m)} \cdot \operatorname*{ess\,sup}_{(r,\beta)} |r \cdot \psi(r\beta)| < \infty.\end{aligned}$$

はじめの不等式では $\int_{\mathbb{R}} |\mathrm{R}f(u,p)| \mathrm{d}p \leqslant \|f\|_1$ を用いた。二番目の不等式では $r \leftarrow 1/\alpha$ によって変数を変換した。ψ は急減少関数なので，有界性がいえる。 □

単射性

リッジレット関数 ψ が許容的であれば，再構成公式が成り立つので，ψ によるリッジレット変換 $\mathscr{R}_\psi$ は適当な関数空間上の作用素として単射である。

リッジレット関数が多項式の場合，リッジレット変換は必ずしも単射にならない。例えば，$\psi(z) = z + 1$ の場合を考える。このとき $\psi^{(2)} \equiv 0$ が成り立つことに注意して，適当な急減少関数 g を用いて $f := \triangle g$ とお

くと，$\mathscr{R}_\psi f = 0$ である

$$\begin{aligned}\mathscr{R}_\psi f(u,\alpha,\beta) &= \left(\mathrm{R}\triangle g(u,\cdot) * \widetilde{\psi_\alpha}\right)(\beta)\\ &= \left(\partial^2 \mathrm{R}g(u,\cdot) * \widetilde{\psi_\alpha}\right)(\beta)\\ &= \left(\mathrm{R}g(u,\cdot) * \partial^2\widetilde{\psi_\alpha}\right)(\beta)\\ &= \left(\mathrm{R}g(u,\cdot) * 0\right)(\beta)\\ &= 0.\end{aligned}$$

ただし，二番目の式では Radon 変換の公式 (Helgason, 2011)

$$\mathrm{R}\triangle g(u,p) = \partial_p^2 \mathrm{R}g(u,p),$$

を用いた。

5.2.3　双対リッジレット変換の定義と存在

定義 5.2.2. 関数 $\mathrm{T} \in \mathcal{Y}(\mathbb{Y}^{m+1})$ の超関数 $\eta \in \mathcal{W}(\mathbb{R})$ による双対リッジレット変換 $\mathscr{R}_\eta^\dagger \mathrm{T}$ とは，各点 $x \in \mathbb{R}^m$ に対して値

$$\mathscr{R}_\eta^\dagger \mathrm{T}(x) = \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \mathrm{T}\left(\frac{u}{\alpha}, \frac{\cdot}{\alpha}\right) * \eta_\alpha(u\cdot x)\frac{\mathrm{d}\alpha \mathrm{d}u}{\alpha^m}, \quad x \in \mathbb{R}^m$$

を対応させる写像である。ただし $\eta_\alpha(p) := \eta(p/\alpha)/\alpha$ とする。

この定義もまた，畳み込みは超関数の意味で理解する。従って，次のように定義することと同値である

$$\mathscr{R}_\eta^\dagger \mathrm{T}(x) = \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \int_{\mathbb{R}} \mathrm{T}\left(u,\alpha,u\cdot x - \alpha z\right)\eta(z)\frac{\mathrm{d}z\mathrm{d}\alpha\mathrm{d}u}{\alpha^m}, \quad x \in \mathbb{R}^m.$$

ただし $\int_{\mathbb{R}} \cdot \eta(z)\mathrm{d}z$ は汎関数 η の作用の意味で理解する。

双対リッジレット変換 $\mathscr{R}_\eta^\dagger$ は，存在すればリッジレット変換 $\mathscr{R}_\eta$ の双対作用素 (Yosida, 1995) である。

定理 5.2.3. $\mathcal{X}, \mathcal{Z}$ を 表 5.2 の組み合わせの中から選び，$\eta \in \mathcal{Z}$ を任意にとって固定する。$\mathscr{R}_\eta : \mathcal{X}(\mathbb{R}^m) \to \mathcal{Y}(\mathbb{Y}^{m+1})$ は単射かつ，$\mathscr{R}_\eta^\dagger : \mathcal{Y}'(\mathbb{Y}^{m+1}) \to \mathcal{X}'(\mathbb{R}^m)$ が存在すると仮定する。このとき $\mathscr{R}_\eta^\dagger$ は $\mathscr{R}_\eta$ の双対作用素 $(\mathscr{R}_\eta)' : \mathcal{Y}'(\mathbb{Y}^{m+1}) \to \mathcal{X}'(\mathbb{R}^m)$ である。

Proof. 仮定より $\mathscr{R}_\eta$ は $\mathcal{X}(\mathbb{R}^m)$ 上稠密に定義されていて，単射である。従って，古典的な定理 (Yosida, 1995, VII. 1. Th. 1, pp.193) より，双対作用素 $(\mathscr{R}_\eta)' : \mathcal{Y}'(\mathbb{Y}^{m+1}) \to \mathcal{X}'(\mathbb{R}^m)$ はただ一つ存在する。一方，$f \in \mathcal{X}(\mathbb{R}^m)$ と $\mathrm{T} \in \mathcal{Y}(\mathbb{Y}^{m+1})$ に対して以下が成り立つ

$$\left\langle \mathscr{R}_\eta f, \mathrm{T} \right\rangle_{\mathbb{Y}^{m+1}} = \int_{\mathbb{R}^m \times \mathbb{Y}^{m+1}} f(x)\overline{\eta(a \cdot x - b)\mathrm{T}(a,b)}\mathrm{d}x\mathrm{d}a\mathrm{d}b = \left\langle f, \mathscr{R}_\eta^\dagger \mathrm{T} \right\rangle_{\mathbb{R}^m}.$$

従って，双対作用素の一意性により $(\mathscr{R}_\eta)' = \mathscr{R}_\eta^\dagger$ である。 □

5.3 再構成公式

超関数によるリッジレット変換の再構成公式を与える。まず，許容条件を定義し，その同値な言い換えとして構造定理を述べる。次に，L^1 関数に対する再構成公式を二通りの方法で証明する。最後に，L^2 関数に対する再構成公式を二通りの方法で証明する。

5.3.1 許容条件

許容条件とは，リッジレット変換の再構成公式が成り立つための十分条件である。活性化関数 $\eta \in \mathcal{W}(\mathbb{R})$ が超関数であっても意味を為すように，従来の許容条件を修正する必要がある。

定義 5.3.1. $(\psi, \eta) \in \mathcal{S}(\mathbb{R}) \times \mathcal{S}'(\mathbb{R})$ が許容的 *(admissible)* であるとは，原点 0 のある近傍 $\Omega \subset \mathbb{R}$ から原点を除いた領域において $\widehat{\eta}$ が局所可積分 $(\widehat{\eta} \in L^1_{\mathrm{loc}}(\Omega \backslash \{0\}))$ で，しかも次の積分が零でない値に収束することをいう

$$K_{\psi,\eta} := (2\pi)^{m-1} \left(\int_{\Omega \backslash \{0\}} + \int_{\mathbb{R} \backslash \Omega} \right) \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m} \mathrm{d}\zeta. \tag{5.4}$$

ここで $\int_{\Omega\backslash\{0\}}$ と $\int_{\mathbb{R}\backslash\Omega}$ はそれぞれ Lebesgue 積分と $\mathbb{R}\backslash\Omega$ に制限した汎関数 $\widehat{\eta}$ の作用の意味で理解する。

Remarks

許容条件の一部に $\eta \in \mathcal{W}(\mathbb{R})$ の Fourier 変換 $\widehat{\eta}$ を使うために，$\mathcal{W} \subset \mathcal{S}'$ を仮定した。($\eta \in \mathcal{D}'$ に対する Fourier 変換 $\widehat{\eta}$ は，一般にはうまく定義できない。)

第一項 $\int_{\Omega\setminus\{0\}}$ の収束性は Ω の取り方によらない。二つの近傍 Ω, Ω' に対して，二つの積分の差 $\int_{\Omega\setminus\Omega'}$ は常に有限だからである。また，第二項 $\int_{\mathbb{R}\setminus\Omega}$ は常に収束する。$|\zeta|^{-m} \in \mathcal{O}_{\mathcal{M}}(\mathbb{R}\setminus\Omega)$ なので $|\zeta|^{-m}\overline{\widehat{\psi}(\zeta)}$ は急減少関数であり，一方 $\widehat{\eta} \in \mathcal{S}'(\mathbb{R})$ なので，その作用である第二項は収束する。従って，$K_{\psi,\eta}$ の収束性も Ω の取り方によらない。

積分区間からは原点 0 を除去してある。$|\zeta|^{-m}$ を原点に特異点をもつ超関数とみなす場合，$\widehat{\eta}$ が同じく原点で特異的な超関数（$\widehat{\eta} = \delta$ など）となった場合に，超関数の積が不定になるためである。幸い $|\zeta|^{-m}$ の特異点は原点に限るので，原点 0 を除去し，$|\zeta|^{-m}$ を $\Omega\setminus\{0\}$ 上の関数とみなして Lebesgue 積分を計算する条件に修正することで許容条件は一意に定まる。特に，Lebesgue 積分において一点の差は零なので，被積分 '超関数' が原点において関数になる場合には，$\int_{\mathbb{R}\setminus\{0\}} = \int_{\mathbb{R}}$ が成り立つ。再構成公式は修正後の許容条件で成り立つ。

$\widehat{\eta}$ が原点 $\{0\}$ のみに台をもつ場合，ψ に依らず常に $K_{\psi,\eta} = 0$ となって，η は許容的にはなりえない。Rudin (1991, Ex. 7.16) によれば，$\mathsf{supp}\ \widehat{\eta} = \{0\}$ は η が多項式関数であることと同値である。従って，$\mathcal{W}$ は Lizorkin 超関数 $\mathcal{S}'/\mathcal{P} \cong \mathcal{S}'_0$ にとるのが自然である。 多項式関数 Q に対して

$$K_{\psi,\eta} = K_{\psi,\eta+Q}.$$

が成り立つので，$K_{\psi,\eta}$ は $\mathcal{S}'_0(\mathbb{R})$ 上 well-defined である。

原点における不定性

許容条件において原点を除去することの必要性は，次の例から理解される。

例 5.3.1 ((Schwartz, 1966, Ch.5 Th.6) 改)**.** $\eta(z) = z$, $\psi(z) = \Lambda\mathrm{G}(z)$ とする。ただし $\mathrm{G}(z) := \exp(-z^2/2)$ である。このとき *Fourier* 変換は以下で与えられる

$$\widehat{\eta}(\zeta) = \delta(\zeta) \quad \textit{and} \quad \overline{\widehat{\psi}(\zeta)} = |\zeta| \cdot \mathrm{G}(\zeta).$$

このとき，二つの超関数の積は結合的でない

$$\int_{\mathbb{R}} \mathsf{pv}\ \frac{1}{|\zeta|} \times (|\zeta| \cdot \mathrm{G}(\zeta) \times \delta(\zeta))\, \mathrm{d}\zeta = 0,$$
$$\int_{\mathbb{R}} \left(\mathsf{pv}\ \frac{1}{|\zeta|} \times |\zeta| \cdot \mathrm{G}(\zeta)\right) \times \delta(\zeta)\mathrm{d}\zeta = \mathrm{G}(0) \neq 0.$$

一方，原点を除去した許容条件 (5.4) であれば，値は一意に定まる

$$K_{\psi,\eta} = \int_{0<|\zeta|<1} \frac{|\zeta| \cdot \mathrm{G}(\zeta)}{|\zeta|} \cdot 0 \mathrm{d}\zeta + \int_{1\leqslant|\zeta|} \frac{|\zeta| \cdot \mathrm{G}(\zeta)}{|\zeta|} \delta(\zeta) \mathrm{d}\zeta = 0.$$

5.3.2 許容リッジレット関数の構造定理

許容条件の被積分関数を形式的に

$$\widehat{\phi}(\zeta) := \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m},$$

とおく。$\widehat{\phi}$ は形式的に以下を満たす

$$|\zeta|^m \widehat{\phi}(\zeta) = \overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta).$$

従って形式的に Fourier 逆変換をとることで以下を得る

$$\Lambda^m \phi = \overline{\widetilde{\psi}} * \eta.$$

ただし厳密には，$\widehat{\eta}$ は原点で特異的な超関数かもしれないので，約分は不定になる可能性がある。また $\widehat{\phi}$ の Fourier 変換の存在も，一般には仮定しなければ分からない。

定理 5.3.1 (許容リッジレット関数の構造定理)**.** $(\psi, \eta) \in \mathcal{S}(\mathbb{R}) \times \mathcal{S}'(\mathbb{R})$ を任意にとって固定する。ある $k \in \mathbb{N}_0$ と $c_j \in \mathbb{C}$ に対して

$$\widehat{\eta}(\zeta) = \sum_{j=0}^{k} c_j \delta^{(j)}(\zeta), \quad \zeta \in \{0\}.$$

が成り立つとする。また，$\widehat{\eta}$ は 0 のある近傍 Ω から 0 を除いた集合上で連続（$\widehat{\eta} \in C(\Omega \backslash \{0\})$）とする。このとき ψ と η が許容的であることは，次の条件と同値。すなわち，ある $\phi \in \mathcal{O}_{\mathcal{M}}(\mathbb{R})$ が存在して

$$\Lambda^m \phi = \overline{\widetilde{\psi}} * \left(\eta - \sum_{j=0}^{k} c_j z^j \right) \quad \textit{and} \quad \int_{\mathbb{R}\backslash\{0\}} \widehat{\phi}(\zeta) \mathrm{d}\zeta \neq 0$$

が成り立つ。さらに，$\lim_{\zeta \to +0} |\widehat{\phi}(\zeta)| < \infty$ かつ $\lim_{\zeta \to -0} |\widehat{\phi}(\zeta)| < \infty$.

特に，ψ が k 次以上の vanishing moment をもつとき，

$$\int_{\mathbb{R}} \psi(z) z^j \mathrm{d}z = 0, \quad j(\leqslant k) \in \mathbb{N}_0$$

条件は次のように緩和できる

$$\Lambda^m \phi = \widetilde{\overline{\psi}} * \eta, \qquad \left| \int_{\mathbb{R}} \phi(z) \mathrm{d}z \right| < \infty \quad \text{and} \quad \int_{\mathbb{R}} \widehat{\phi}(\zeta) \mathrm{d}\zeta \neq 0.$$

証明は 付録 A.3 を見よ。
定理の系として次の構成法を得る

系 5.3.2 (許容的リッジレット関数の構成法)**.** $\eta \in \mathcal{S}_0'(\mathbb{R})$ は与えられたものとする。0 の適当な近傍 Ω と $k \in \mathbb{N}_0$ を選んで，$\zeta^k \cdot \widehat{\eta}(\zeta) \in C(\Omega)$ にできるものとする。$\psi_0 \in \mathcal{S}(\mathbb{R})$ として

$$\int_{\mathbb{R}} \zeta^k \overline{\widehat{\psi_0}(\zeta)} \widehat{\eta}(\zeta) \mathrm{d}\zeta \neq 0.$$

となるものをとる。このとき

$$\psi := \Lambda^m \psi_0^{(k)},$$

とおくと，ψ と η は許容的である。

証明は $\phi := \widetilde{\overline{\psi_0^{(k)}}} * \eta$ が 定理 5.3.1 の条件を満たすことから明らか。

5.3.3　L^1 再構成公式

Fourier 変換に帰着する方法と，Radon 変換に帰着する方法の二通りを示す。

定理 5.3.3 (Fourier 変換を経由する再構成公式)**.** 可積分関数 $f \in L^1(\mathbb{R}^m)$ は *Fourier* 変換 $\widehat{f}$ もまた可積分とする。$(\psi, \eta) \in \mathcal{S}(\mathbb{R}) \times \mathcal{S}_0'(\mathbb{R})$ は許容的であるとする。このときほとんどいたるところの $x \in \mathbb{R}^m$ で再構成公式が成り立つ

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) = K_{\psi,\eta} f(x), \quad \text{a.e. } x \in \mathbb{R}^m.$$

特に，f の連続点 x において等式が成り立つ。

Proof. 再構成公式の左辺は次の特異積分に帰着する

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) = \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \mathrm{R}f(u,\cdot) * \lambda_\alpha(u\cdot x)\frac{\mathrm{d}\alpha \mathrm{d}u}{\alpha^m}.$$

ただし $\lambda_\alpha(p) := (\overline{\widetilde{\psi}} * \eta)(p/\alpha)/\alpha$ とおいた。投影切断面定理 $\mathrm{R}f(u,\cdot)*\lambda_\alpha(\beta) = (2\pi)^{-1}\int_{\mathbb{R}} \widehat{f}(\omega u)\widehat{\lambda}(\alpha\omega)|\alpha\omega|^m e^{i\omega\beta}\mathrm{d}\omega$ により，

$$\begin{aligned}
\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) &= \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \frac{1}{2\pi} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \int_{\mathbb{R}} \widehat{f}(\omega u)\widehat{\lambda}(\alpha\omega)|\omega|^m e^{i\omega u\cdot x}\mathrm{d}\omega\mathrm{d}\alpha\mathrm{d}u,\\
&= \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \frac{1}{2\pi} \int_{\mathbb{S}^{m-1}} \int_{\mathbb{R}} \int_{\varepsilon\leqslant\frac{\zeta}{\omega}\leqslant\delta} \widehat{f}(\omega u)\widehat{\lambda}(\zeta)|\zeta|^{-m}\mathrm{d}\zeta|\omega|^{m-1} e^{i\omega u\cdot x}\mathrm{d}\omega\mathrm{d}u, \quad \alpha \leftarrow \frac{\zeta}{\omega}\\
&= \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \frac{1}{2\pi} \int_{\mathbb{R}^m} \left[\int_{\varepsilon\leqslant\frac{|\zeta|}{|\xi|}\leqslant\delta} \widehat{\lambda}(\zeta)|\zeta|^{-m}\mathrm{d}\zeta\right] \widehat{f}(\xi) e^{i\xi\cdot x}\mathrm{d}\xi, \quad \xi \leftarrow \omega u\\
&= \frac{1}{(2\pi)^m} \int_{\mathbb{R}\setminus\{0\}} \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m}\mathrm{d}\zeta \int_{\mathbb{R}^m} \widehat{f}(\xi) e^{i\xi\cdot x}\mathrm{d}\xi = K_{\psi,\eta} f(x), \quad \text{a.e. } x\in\mathbb{R}^m.
\end{aligned}$$

ただし極限と積分の交換は超関数 $\widehat{\lambda} \in \mathcal{O}_{\mathcal{C}}'(\mathbb{R})$ の作用が連続であることから従う。L^1 関数に対する Fourier 変換の反転公式は概収束なので，再構成公式も概収束の意味で成り立つ。 □

Radon 変換を経由する方法では，リッジレット変換のウェーブレットに相当する部分が，Radon 変換で用いられる逆投影フィルタ Λ の役割を果たすことが陽に示される。言い換えれば，許容条件とは，ウェーブレット部分が逆投影フィルタを為すための条件である。

定理 5.3.4 (Radon 変換を経由する再構成公式). $f \in L^1(\mathbb{R}^m)$ とし，$(\psi,\eta) \in \mathcal{S}(\mathbb{R})\times\mathcal{S}'(\mathbb{R})$ に対して実数値関数 $\phi \in L^1\cap C^\infty(\mathbb{R})$ が存在して，以下を満たすとする

$$\Lambda^m\phi = \overline{\widetilde{\psi}} * \eta \quad \textit{and} \quad \int_{\mathbb{R}} \widehat{\phi}(\zeta)\mathrm{d}\zeta = -1.$$

このとき $\mathscr{R}_\eta^\dagger\mathscr{R}_\psi f$ は *Radon* の反転公式に帰着する

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) = \mathrm{R}^\dagger\Lambda^{m-1}\mathrm{R}f(x) = 2(2\pi)^{m-1} f(x), \quad \text{a.e. } x\in\mathbb{R}^m.$$

証明は 付録 A.4 を見よ。許容条件の構造定理（定理 5.3.1）により，定理の仮定は許容条件である。従って，許容条件は逆投影フィルタを構成するための条件であることが分かる。

Radon 変換の公式 (Helgason, 2011, Lem.2.1, Th.3.1, Th.3.7)

$$(-\triangle)^{\frac{m-1}{2}}\mathrm{R}^\dagger = \Lambda^{m-1}\mathrm{R}, \quad \text{and} \quad \mathrm{R}(-\triangle)^{\frac{m-1}{2}} = \mathrm{R}^\dagger\Lambda^{m-1}.$$

と組み合せて，以下の系を得る

系 5.3.5.

$$\mathscr{R}_\eta^\dagger\mathscr{R}_\psi = \mathrm{R}^\dagger\Lambda^{m-1}\mathrm{R} = (-\triangle)^{\frac{m-1}{2}}\mathrm{R}^\dagger\mathrm{R} = \mathrm{R}^\dagger\mathrm{R}(-\triangle)^{\frac{m-1}{2}}.$$

5.3.4　L^2 有界拡張

Fourier 変換の場合と同様の手続きに則り，$L^1 \cap L^2$ 関数のリッジレット変換を拡張して L^2 関数のリッジレット変換が定義できる。以下では $\mathbb{Y}^{m+1}$ の測度を $\alpha^{-m}\mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}u$ にとる。

ψ が**自己許容的**とは，ψ が自分自身（$\eta = \psi$）と許容条件を満たすことを言う。リッジレット変換の双対性から Parseval の公式と Plancherel の公式が成り立つ

定理 5.3.6. $(\psi, \eta) \in \mathcal{S}\times\mathcal{S}'$ は許容条件を満たすとし，簡単のため $K_{\psi,\eta} = 1$ とする。$f, g \in L^1 \cap L^2(\mathbb{R}^m)$ に対し以下が成り立つ

$$(\mathscr{R}_\psi f, \mathscr{R}_\eta g) = \left(\mathscr{R}_\eta^\dagger\mathscr{R}_\psi f, g\right) = (f, g)\,.$$

特に ψ が自己許容的であるとき，

$$\|\mathscr{R}_\psi f\|_2 = \|f\|_2.$$

命題 5.2.2 により，リッジレット変換は $L^1(\mathbb{R}^m)$ 上の有界作用素である。従って，$\psi \in \mathcal{S}(\mathbb{R})$ が自己許容的であるとき，有界拡張（bounded extension）の手続きに則って，リッジレット変換を $L^2(\mathbb{R}^m)$ 上に拡張できる (Grafakos, 2008, 2.2.4)。

定理 5.3.7 (リッジレット変換の L^2 有界拡張)**.** $\psi \in \mathcal{S}(\mathbb{R})$ は自己許容的とし，簡単のため $K_{\psi,\psi} = 1$ とする。このとき $L^1 \cap L^2(\mathbb{R}^m)$ 上のリッジレット変換は $L^2(\mathbb{R}^m)$ 上の作用素として一意に拡張でき，このとき $\|\mathscr{R}_\psi f\|_2 = \|f\|_2$ である。

Proof. 関数 $f \in L^2(\mathbb{R}^m)$ と，f に L^2 収束する点列 $f_n \in L^1 \cap L^2(\mathbb{R}^m)$ を任意にとる。このとき

$$\|f_n - f_m\|_2 = \|\mathscr{R}_\psi f_n - \mathscr{R}_\psi f_m\|_2, \quad \forall n, m \in \mathbb{N}.$$

右辺は $n, m \to \infty$ のとき右辺は $L^2(\mathbb{Y}^{m+1})$ の Cauchy 列である。従って L^2 空間の完備性により，$\mathscr{R}_\psi f_n$ の極限 $\mathrm{T}_\infty \in L^2(\mathbb{Y}^{m+1})$ が一意に存在する。有界拡張の手続きに従い，この極限 T_∞ を f のリッジレット変換と定義する $\mathscr{R}_\psi f := \mathrm{T}_\infty$。 □

(ψ, η) と $(\psi^\star, \eta^\star)$ が**同値**であるとは，(ψ, η) と $(\psi^\star, \eta^\star)$ がそれぞれ許容条件を満たし，さらにそれぞれの畳み込み積分が一致することをいう

$$\overline{\widetilde{\psi}} * \eta = \overline{\widetilde{\psi^\star}} * \eta^\star.$$

このとき直ちに以下が成り立つ

$$(\mathscr{R}_\psi f, \mathscr{R}_\eta g) = (\mathscr{R}_{\psi^\star} f, \mathscr{R}_{\eta^\star} g)\,.$$

許容条件的な組 (ψ, η) が**許容的分解可能**とは，自己許容的な二つの組 $(\psi^\star, \psi^\star)$ と $(\eta^\star, \eta^\star)$ があって，$(\psi^\star, \eta^\star)$ と (ψ, η) が同値になることをいう。このとき Schwartz の不等式により以下が成り立つ

$$(\mathscr{R}_\psi f, \mathscr{R}_\eta g) \leqslant \|\mathscr{R}_{\psi^\star} f\|_2 \|\mathscr{R}_{\eta^\star} g\|_2.$$

ψ が自己許容的でなく，従って $\mathscr{R}_\psi$ が $L^2(\mathbb{R}^m)$ 上定義できない場合であっても，η の選び方次第では，再構成作用素 $\mathscr{R}_\eta^\dagger \mathscr{R}_\psi$ が定義できる。

定理 5.3.8 (L^2 再構成公式)**.** 関数 $f \in L^2(\mathbb{R}^m)$ とし，$(\psi, \eta) \in \mathcal{S} \times \mathcal{S}'$ は許容的分解可能とする。また，簡単のため $K_{\psi,\eta} = 1$ とする。このとき

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f \to f, \quad in\ L^2.$$

証明は 付録 A.5 を見よ。

5.3.5 Calderón 再生公式

Calderón 再生公式に帰着することもできる。

定理 5.3.9 (Calderón 再生公式を経由する L^2 再構成公式). $f \in L^2(\mathbb{R}^m)$ かつ $(\psi, \eta) \in \mathcal{S} \times \mathcal{S}'(\mathbb{R})$ は $\overline{\widetilde{\psi}} * \eta$ が *Borel* 測度であるとする。このとき L^2 再構成公式が成り立つ

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f = f, \quad in\ L^2.$$

Proof. $\lambda_\alpha(p) := \overline{\widetilde{\psi}} * \eta(p/\alpha)/\alpha$, $\mu_{u,\alpha}(x) := \overline{\widetilde{\psi}} * \eta(u \cdot x/\alpha)/\alpha^m$ とおく。

$$\begin{aligned}
&\mathrm{R}f(u,\cdot) * \lambda_\alpha(u \cdot x)\frac{1}{\alpha^m} \\
&\quad = \int_{\mathbb{R}} \int_{(\mathbb{R}u)^\perp} f(pu+y)\lambda\left(\frac{u\cdot(x-pu-y)}{\alpha}\right)\frac{1}{\alpha^{m+1}}\mathrm{d}y\mathrm{d}p \\
&\quad = \int_{\mathbb{R}^m} f(x')\lambda\left(\frac{u\cdot(x-x')}{\alpha}\right)\frac{1}{\alpha^{m+1}}\mathrm{d}x' \\
&\quad = f * \mu_{u,\alpha}(x)\frac{1}{\alpha}.
\end{aligned}$$

なので，Calderón 再生公式によって

$$\begin{aligned}
\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) &= \lim_{\substack{\delta\to\infty \\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \mathrm{R}f(u,\cdot) * \lambda_\alpha(u\cdot x)\frac{\mathrm{d}\alpha\mathrm{d}u}{\alpha^m} \\
&= \lim_{\substack{\delta\to\infty \\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta f * \mu_{u,\alpha}(x)\frac{\mathrm{d}\alpha\mathrm{d}u}{\alpha} = K_{\psi,\eta} f(x).
\end{aligned}$$

□

5.4　許容的なリッジレット関数の構成例

系 5.3.2 に基いて，許容的なリッジレット関数を具体的に構成する。以下では，活性化関数 $\eta \in \mathcal{S}_0'(\mathbb{R})$ に対して，リッジレット関数 $\psi \in \mathcal{S}(\mathbb{R})$ の候補を Gauss 関数

$$\mathrm{G}(z) := \exp\left(-z^2/2\right),$$

から作る。

まず G の高階導関数

$$\psi_0 = \mathrm{G}^{(\ell)},$$

のなかで,

$$\int_{\mathbb{R}\setminus\{0\}} \overline{\widehat{\psi_0}(\zeta)}\widehat{\eta}(\zeta)\mathrm{d}\zeta \neq 0, \pm\infty$$

を満たす ψ_0 をさがす。このとき

$$\psi := \Lambda^m \psi_0,$$

とおくと，系 5.3.2 により (ψ, η) は許容条件を満たす。

ここで，Gauss 関数の Hilbert 変換は以下で与えられる

$$\mathcal{H}\mathrm{G}(z) = \frac{2i}{\sqrt{\pi}} F\left(\frac{z}{\sqrt{2}}\right),$$

ただし $F(z)$ は Dawson 関数 $F(z) := \exp(-z^2)\int_0^z \exp(w^2)\mathrm{d}w$ である。

例 5.4.1. z_+^k $(k \in \mathbb{N}_0)$ と $\psi = \Lambda^m \mathrm{G}^{(\ell+k+1)}$ $(\ell \in \mathbb{N}_0)$ は ℓ が偶数のとき許容条件を満たす。奇数の場合には $K_{\psi,\eta} = 0$.

Proof. 以下の公式 (Gel'fand and Shilov, 1964, § 9.3) から分かる

$$\widehat{z_+^k}(\zeta) = \frac{k!}{(i\zeta)^{k+1}} + \pi i^k \delta^{(k)}(\zeta), \quad k \in \mathbb{N}_0. \qquad \square$$

例 5.4.2. $\eta(z) = \delta^{(k)}(z)$ $(k \in \mathbb{N}_0)$ と $\psi = \Lambda^m \mathrm{G}$ は k が偶数のとき許容的を満たす。奇数のときは $K_{\psi,\eta} = 0$.

例 5.4.3. $\eta(z) = \mathrm{G}^{(k)}(z)$ $(k \in \mathbb{N}_0)$ と $\psi = \Lambda^m \mathrm{G}$ は k が偶数のとき許容条件を満たす。奇数のときは $K_{\psi,\eta} = 0$.

例 5.4.4. $\eta(z) = \sigma^{(k)}(z)$ $(k \in \mathbb{N}_0)$ と $\psi = \Lambda^m \mathrm{G}$ は k が奇数のとき許容条件を満たす。偶数のときは $K_{\psi,\eta} = 0$. また $\sigma^{(-1)}$ と $\psi = \Lambda^m \mathrm{G}''$ は許容条件を満たす。

5.5 数値例

一次元信号と二次元信号（画像）に対する再構成公式を数値積分することで，理論との整合性を確認する。表 5.3 は前節の理論的な "診断" に基いて，許容条件の成否をまとめた表である。'+' は許容的，'0' は $K_{\psi,\eta} = 0$ によって非許容的，そして '∞' は $|K_{\psi,\eta}| = \infty$ によって非許容的となることを表している。

表 5.3: リッジレット関数 $\psi = \Lambda^m \psi_0$ と活性化関数 η に対する許容条件の成否

活性化関数	η	$\psi = \Lambda^m \mathrm{G}$	$\psi = \Lambda^m \mathrm{G}'$	$\psi = \Lambda^m \mathrm{G}''$
シグモイド関数の導関数	σ'	$+$	0	$+$
シグモイド関数	σ	∞	$+$	0
ソフトプラス関数	$\sigma^{(-1)}$	∞	∞	$+$
Dirac's δ	δ	$+$	0	$+$
ステップ関数	z_+^0	∞	$+$	0
ReLU	z_+	∞	∞	$+$
線形関数	z	0	0	0
RBF	G	$+$	0	$+$

5.5.1　正弦波

一次元信号の例として閉区間 $x \in [-1, 1]$ 上の正弦波 $f(x) = \sin 2\pi x$ をとりあげる。

信号は区間 $[-1, 1]$ から等間隔（$\Delta x = 1/100$）にサンプルした。次の再構成公式を数値積分によって計算した

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \mathscr{R}_\psi f(a, b) \eta(ax - b) \frac{\mathrm{d}a\mathrm{d}b}{|a|}.$$

数値積分は領域 $(a, b) \in [-30, 30] \times [-30, 30]$ を等間隔（$\Delta a = \Delta b = 1/10$）に離散化して行った

$$\mathscr{R}_\psi f(a, b) \approx \sum_{n=0}^{N} f(x_n) \overline{\psi(a \cdot x_n - b)} |a| \Delta x, \quad x_n = x_0 + n\Delta x$$

$$\mathscr{R}_\eta^\dagger \mathscr{R} f(x) \approx \sum_{(i,j)=(0,0)}^{I,J} \mathscr{R}_\psi f(a_i, b_j) \eta(a_i \cdot x - b_j) \frac{\Delta a \Delta b}{|a_i|}, \quad a_i = a_0 + i\Delta a,\ b_j = b_0 + j\Delta b$$

ただし $x_0 = -1,\ a_0 = -30,\ b_0 = -30,\ N = 200, (I, J) = (600, 600)$.

図 5.2 はリッジレット変換 $\mathscr{R}_\psi f(a, b)$ の結果を示す。リッジレット関数 ψ の取り方に応じて，リッジレット変換の像は変化する。リッジレット関数 $\psi = \Lambda \mathrm{G}^{(\ell)}$ の階数が上がるに連れて，$\mathscr{R}_\psi f$ は局所性が強くなる。そし

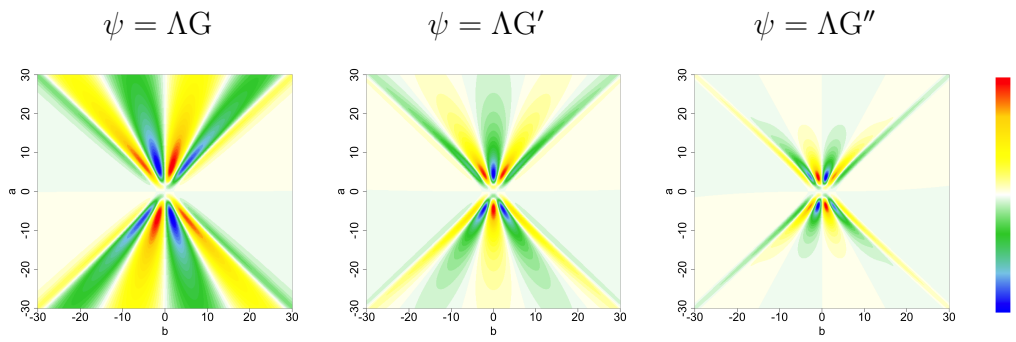


図 5.2: 閉区間 $[-1, 1]$ 上定義された正弦波 $f(x) = \sin 2\pi x$ の ψ によるリッジレット変換 $\mathscr{R}_\psi f(a, b)$

て図 5.3 に示すとおり，いずれのリッジレット変換 $\mathscr{R}_\psi f$ も，適当な活性化関数 η の組み合わせで再構成可能である。

図 5.3 は RBF とステップ関数，そして ReLU による再構成結果を示す。実線は再構成結果，破線は元の信号を表す。許容条件を満たすと診断された組み合わせのセルは，ほぼ完全に再構成結果ができている。左下の ReLU のセルは，非許容的 (‘∞’) のはずだが，不完全ながら像が現れているように見える。ReLU の Fourier 変換 $\widehat{z_+}(\zeta)$ は極 ζ^{-2} を持ち，$\sigma^{(-1)} * \Lambda G$ ではその極が残るために，積分器のような働きをしたと考えられる。

5.5.2 Shepp-Logan phantom

二次元信号の例として *Shepp-Logan phantom* (Shepp and Logan, 1974) を取り上げる。

原画像は 256×256 ピクセルのグレースケール画像であり，これを二次元信号 $f : [-1, 1]^2 \to [0, 1]$ とみなす。次の再構成公式を数値積分によって計算する

$$\int_{\mathbb{R}} \int_{\mathbb{R}^2} \mathscr{R}_\psi f(a, b) \eta(a \cdot x - b) \frac{\mathrm{d}a\mathrm{d}b}{|a|},$$

ただし $(a, b) \in [-300, 300]^2 \times [-30, 30]$ を $\Delta a = (1, 1)$ と $\Delta b = 1$ の間隔で離散化した。

図 5.4 に再構成結果を示す。一次元信号の場合と同様，許容的と診断されたセルでは概ね明確に再構成結果が現れた。また，左下の非許容的なセルには，ローパスがかけられたように曇った像が現れた。

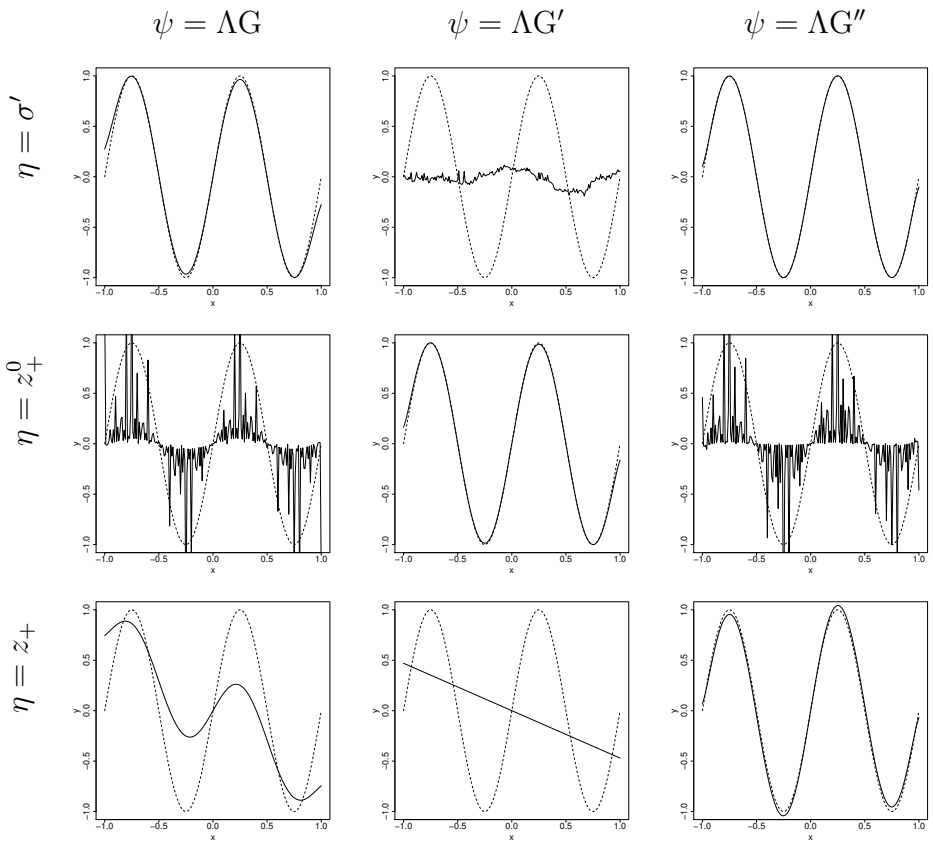


図 5.3: 活性化関数 σ', z_+^0, z_+ による再構成結果 (実線)。点線は元の信号。

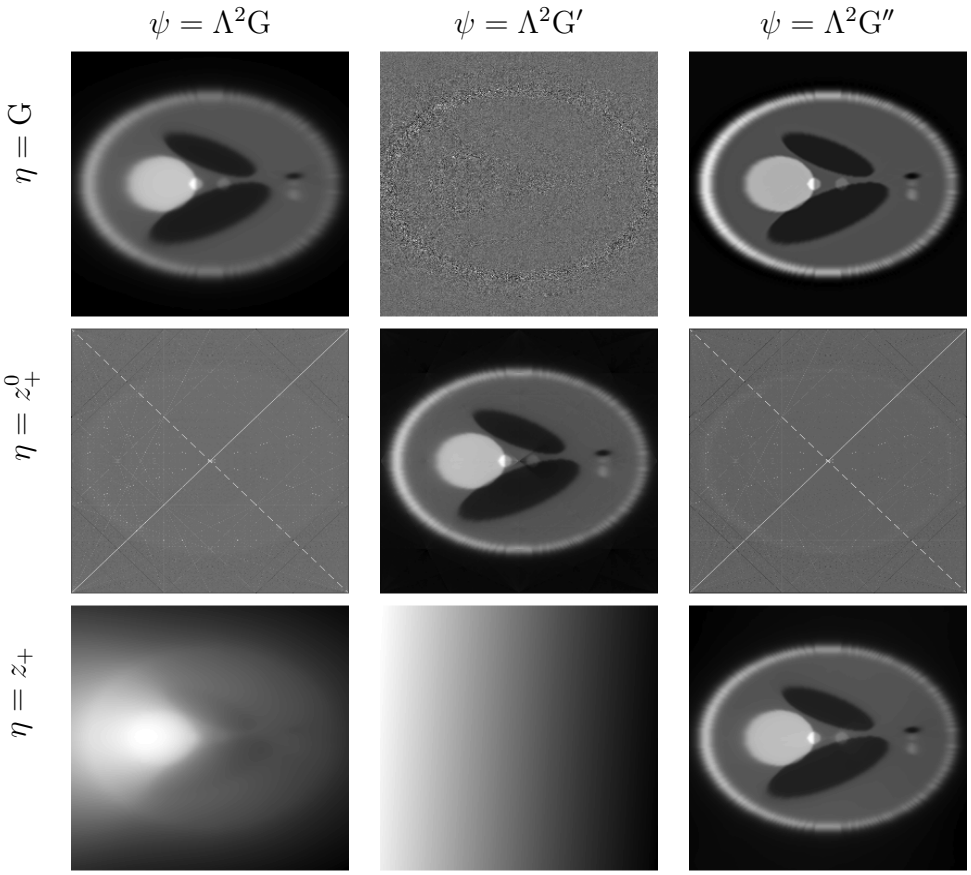


図 5.4: 活性化関数 G, z_+^0, z_+ による再構成結果

5.6　まとめ

ReLU などの有界でない活性化関数をもつニューラルネットを扱うために，非有界活性化関数によるニューラルネットのための積分表現理論を整備した。浅いニューラルネットの積分表現は双対リッジレット変換であり，活性化関数はそのリッジレット関数に相当する。特に ReLU は緩増加超関数や Lizorkin 超関数に属する関数であり，これまでに利用されてきた多くの活性化関数がこれらのクラスに属する（表 5.1）。従って，課題は超関数によるリッジレット解析を構築することである。

まず，リッジレット変換の畳み込み形式を超関数の意味で捉え，超関数によるリッジレット変換を定義した。定理 5.2.1 では，拡張したリッジレット変換の一意存在性を確認し，関数 f のクラス $\mathcal{X}$ とリッジレット関数 ψ のクラス $\mathcal{Z}$ のサイズがトレードオフ関係にあることを明らかにした（表 5.2）。また，命題 5.2.2 では $\mathscr{R}_\psi : L^1(\mathbb{R}^m) \to \mathcal{S}'(\mathbb{Y}^{m+1})$ が有界作用素であることを示した。この性質は後に L^2 拡張で必要になる。一方，定理 5.2.3 では，双対リッジレット変換がリッジレット変換の双対作用素であることを通じて，一意性を確認した。

拡張の過程で最も慎重を期すポイントは，許容条件である。従来のリッジレット解析で用いられてきた許容条件をそのまま超関数の意味に拡張すると，超関数の積 $\widehat{\eta}(\zeta) \times |\zeta|^{-m}$ が現れる。例 5.3.1 にも示した通り，一般に超関数の積は不定であり，従来の形式の許容条件は ill-defined である。超関数の積が問題になるのは原点に限ることに着目して，許容条件を修正した。許容リッジレット関数の構造定理（定理 5.3.1）は，許容条件を実空間で述べなおした定理である。構造定理の系として許容的関数を構成する方法を導いた（系 5.3.2）

L^1 関数に対するリッジレット変換の再構成公式は，Fourier 変換を経由する方法（定理 5.3.3）と，Radon 変換を経由する方法（定理 5.3.4）の二通りで示した。リッジレット変換は Radon 変換と関係が深いが，前者は Radon 変換を投影切断面定理で戻す方法に対応し，後者は Radon 変換を Radon の反転公式で戻す方法に対応している。再構成のための正則性条件は前者の方がやや緩い。一方，後者は実空間で閉じており，解釈性が高い。特に，許容条件は逆投影フィルタを構成するための条件であることが分かる。

さらに，L^2 有界拡張の手続きに則って，L^1 関数に対するリッジレット変換を，L^2 関数に拡張した。まず $L^1 \cap L^2$ 関数の場合に Parseval の定理が成り立つこと（定理 5.3.6）を示し，リッジレット変換が有界作用素で

あること（命題 5.2.2）を用いて L^2 関数までリッジレット変換を拡張した（定理 5.3.7）。この他に，L^2 有界拡張とは別の方法として，Calderón の再生公式を経由する方法（定理 5.3.9）を示した。

§ 5.4 では，切断べき z_+^k などの具体的な関数に対して許容条件を満たす例を計算した。続く数値実験では，再構成公式を数値積分することで，信号が再構成できることを視覚的に確認した。これらの具体的な構成を通じて，ReLU を活性化関数にもつニューラルネットが万能関数近似能力を持つことを確認した。

ニューラルネットが学習しているものは，近似対象の関数のリッジレット変換である。許容条件の構造定理や，再構成公式の計算例からも分かる通り，同一の関数を近似するためのリッジレット変換は無数に存在する。実際，与えられた活性化関数 η に対して，許容条件の $K_{\psi,\eta}$ が有限になる ψ の全体 $\mathcal{F}_\eta$ と，零になる ψ の全体 $\mathcal{N}_\eta$ は，それぞれ線形空間をなす

$$\mathcal{F}_\eta := \left\{ \psi \in \mathcal{S}(\mathbb{R}) \;\middle|\; \int_{-\infty}^{\infty} \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|} \mathrm{d}\zeta \text{ is finite} \right\},$$

$$\mathcal{N}_\eta := \left\{ \psi \in \mathcal{S}(\mathbb{R}) \;\middle|\; \int_{-\infty}^{\infty} \frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|} \mathrm{d}\zeta = 0 \right\}.$$

従って，η に対して許容条件を満たす活性化関数の全体 $\mathcal{A}_\eta$ は $\mathcal{F}_\eta$ と $\mathcal{N}_\eta$ の差として与えられる

$$\mathcal{A}_\eta = \mathcal{F}_\eta \backslash \mathcal{N}_\eta.$$

バックプロパゲーション学習では，そのうちの一つのリッジレット関数を暗に探し求めていることになる。許容的なリッジレット関数の中で，何が最良の選択なのかを調べることは，今後の重要な課題である。

第6章　深層ニューラルネットの積分表現理論

深層ニューラルネットの中間層は特徴量写像とみなせる。(Sonoda and Murata, 2016) では，中間層の特徴量写像を解析するために，ニューラルネットを輸送写像とみなす方法を提案した。特にデノイジング・オートエンコーダー（DAE）の場合は輸送写像が陽に求まる。DAE を複数合成した合成 DAE や連続 DAE の中では，入力データのエントロピーを減らす逆拡散現象が起きていることを示す。また，輸送解釈によって，これまで解釈が難しかった積層 DAE も，合成 DAE に帰着できることを示す。最後に，合成 DAE を経由して深層 DAE の積分表現を構成する。

6.1　はじめに

オートエンコーダーでは，ニューラルネットに恒等写像

$$x \mapsto x,$$

を学習させ，学習済の中間層を特徴量として用いる。恒等写像は最も基本的な輸送写像である。なお，オートエンコーダーの語源となった「コード」とは，こうして得られた特徴量のことである。

本章では，デノイジング・オートエンコーダー（denoising autoencoder; DAE）(Vincent et al., 2008) による輸送写像を計算する。DAE とは，訓練データにわざとノイズを付加し，元の値を推定させる訓練法である。つまり，ノイズ除去機能を学習させるので，デノイジングという修飾語がついている。このような操作は，オートエンコーダーに頑健性を賦与するためのヒューリスティクスとして登場した。Alain and Bengio (2014) は DAE の手続きの結果として得られる写像が，次の式で陽に書けること

を発見した

$$x \mapsto \frac{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)(x-\varepsilon)]}{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)]}.$$

ただし ε は DAE で人為的に加えるノイズを表し，π_0 はデータ x が従う確率分布を表す。Sonoda and Murata (2016) は，この結果を変形して，DAE による輸送写像を見出した

$$x \mapsto x + t\nabla \log[\mathrm{W}_{t/2} * \pi_0](x).$$

ただし t はノイズの分散，W_t は熱核，π_0 は入力データ x の確率分布である。つまり，DAE の輸送写像はオートエンコーダー $x \mapsto x$ に補正項 $x \mapsto f(x)$ を加えた形式 $x \mapsto x + f(x)$ になる。以下では，このように輸送写像という観点で深層 DAE を解析する。

6.2　浅い DAE

まず，浅いデノイジング・オートエンコーダー（DAE）の学習法を説明する。次に，Alain and Bengio (2014) の変分計算によって DAE が陽に求まることを示す。さらに，これが輸送写像と見なせることを示す。得られた輸送写像を改めて「浅い DAE」と定義する。

6.2.1　DAE の学習アルゴリズム

入力 x を $\mathbb{R}^m$ に値をとり確率分布 π_0 に従う確率変数とする。x に分散 tI の正規ノイズを加えたものを $\widetilde{x}$ とする

$$\widetilde{x} := x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, tI).$$

DAE は，$\widetilde{x}$ から x を復元するように訓練した浅いニューラルネット g である。なお，DAE は**学習時と使用時で入力するものが異なる**ので混乱しないように注意せよ。学習時にはノイズを付加した入力 $\widetilde{x}$ を与えるのに対し，使用時にはノイズを付加していない入力 x を入力する。

DAE の学習アルゴリズムは，次の最適化問題と等価である

$$\text{minimize} \quad L[g] := \mathbb{E}_{x,\widetilde{x}}|g(\widetilde{x}) - x|^2 \quad \text{w.r.t} \quad g.$$

こうして得られる g は $\tilde{x}$ の関数だが，先に注意したとおり，DAE の使用時には $\tilde{x}$ ではなく x を入力する。なお本章を通じて，g は高々有限個の中間層素子をもつ通常のニューラルネットで，しかも最小点からの誤差は無視できるほど小さいものとする。

浅いニューラルネットとして実現された DAE g において，中間層と出力層に相当する写像をそれぞれ h と k と書く。つまり $g = k \circ h$ という

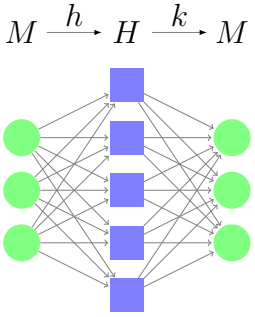


図 6.1: DAE の中間層 h をエンコーダー，出力層 k をデコーダーと呼ぶ。

関係が成り立つ。慣習に倣い，中間層 h をエンコーダー，出力層 k をデコーダーと呼び，入力 x に対して $z := h(x)$ を x の特徴量あるいはコードと呼ぶ。

6.2.2 Alain and Bengio の最適解

Alain and Bengio (2014, Theorem 1) は DAE の最適解が次のように陽に求まることを発見した。

$$g(x) = \frac{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)(x-\varepsilon)]}{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)]}, \quad \text{a.e. } x. \tag{6.1}$$

Proof. 証明は標準的な変分計算によって示せる。まず目的関数を次のように変形する

$$\begin{aligned} L[g] &= \int_{\mathbb{R}^m} \mathbb{E}_\varepsilon |g(x+\varepsilon) - x|^2 \pi_0(x) \mathrm{d}x \\ &= \int_{\mathbb{R}^m} \mathbb{E}_\varepsilon [|g(x') - x' + \varepsilon|^2 \pi_0(x'-\varepsilon)] \mathrm{d}x', \quad x' \leftarrow x + \varepsilon. \end{aligned}$$

従って，任意の関数 h に対して h 方向の変分 $\delta L[h]$ は以下のように計算できる

$$\begin{aligned}\delta L[h] &= \frac{d}{dt} L[g+th]\Big|_{t=0} \\ &= \int_{\mathbb{R}^m} \frac{\partial}{\partial t} \mathbb{E}_\varepsilon[|g(x)+th(x)-x+\varepsilon|^2 \pi_0(x-\varepsilon)]\mathrm{d}x\Big|_{t=0} \\ &= 2\int_{\mathbb{R}^m} \mathbb{E}_\varepsilon[(g(x)-x+\varepsilon)\pi_0(x-\varepsilon)]h(x)\mathrm{d}x.\end{aligned}$$

目的関数の停留点（最適解）では，任意の h に対して $\delta L[h] \equiv 0$ を満たす。このとき，変分法の基本補題によって $\delta L[h]$ の被積分関数はほとんどいたるところ 0 である

$$\mathbb{E}_\varepsilon[(g(x)-x+\varepsilon)\pi_0(x-\varepsilon)] = 0, \quad \text{a.e. } x.$$

これを g について解いて，Alain の最適解を得る。 □

6.2.3　輸送解釈と輸送表現

Sonoda and Murata (2016) では，DAE の最適解が輸送写像になることを発見した。

定理 6.2.1. 最適化問題 $\min_g \mathbb{E}_{x,\tilde{x}}|g(\tilde{x})-x|^2$ の最適解は以下で与えられる

$$g(x) = x + t\nabla \log[\mathrm{W}_{t/2} * \pi_0(x)]. \tag{6.2}$$

ただし ∇ は x についての微分（勾配）を表す。

Proof. 証明は Alain の最適解を変形すればよい

$$\begin{aligned}(6.1) &= x - \frac{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)\varepsilon]}{\mathbb{E}_\varepsilon[\pi_0(x-\varepsilon)]} \\ &= x - \frac{\int_{\mathbb{R}^m} \varepsilon \mathrm{W}_{t/2}(\varepsilon)\pi_0(x-\varepsilon)\mathrm{d}\varepsilon}{\mathrm{W}_{t/2} * \pi_0(x)} \\ &= x + \frac{t\nabla \mathrm{W}_{t/2} * \pi_0(x)}{\mathrm{W}_{t/2} * \pi_0(x)} \\ &= x + t\nabla \log[\mathrm{W}_{t/2} * \pi_0(x)], \end{aligned} \tag{6.3}$$

ただし W_t は熱核 $\mathrm{W}_t(\varepsilon) = (4\pi t)^{-m/2}\exp(-|\varepsilon|^2/4t)$ であり，二番目の変形は関係式

$$\nabla \mathrm{W}_{t/2}(\varepsilon) = -(\varepsilon/t)\mathrm{W}_{t/2}(\varepsilon) \tag{6.4}$$

から従う。 □

この式から直ちに，DAE は恒等写像 id と，ノイズ除去に伴う補正項 $t\nabla \log[\mathrm{W}_{t/2} * \pi_0]$ に分解できることが分かる。特に $t = 0$ のときは補正項が消滅するので，DAE は単に恒等写像を学習させる古典的なオートエンコーダーに帰着する。また，補正項に表れる $\mathrm{W}_{t/2} * \pi_0$ は，熱方程式 ($L = \frac{1}{2}\triangle$) に対する熱半群 $e^{t/2\triangle}$ の作用であることに注意せよ。つまり，熱方程式に従って π_0 を一旦なまし，なましを戻す方向 ($-\nabla \log e^{t/2\triangle}\pi_0$) に補正すると解釈できる。

以降では，(6.2) を輸送写像とみなす。すなわち，位置 x にある質点を，x から補正項の分だけ移動させる写像という意味に解釈する。輸送に伴う軌道の具体例は § 6.3.4 を参照せよ。(6.2) を一般化して次の輸送写像を導入する。

定義 6.2.1. 楕円形作用素 L による異方性 *DAE*（*anisotropic DAE*）を以下で定義する

$$\Phi_t(x; L) := x + t\nabla \log e^{tL}\pi_0(x), \quad x \in \mathbb{R}^m \tag{6.5}$$

$$= x + t\nabla \log \left[\int_{\mathbb{R}^m} \mathrm{W}_t(x, y; L)\pi_0(y)\mathrm{d}y\right], \quad x \in \mathbb{R}^m. \tag{6.6}$$

ただし $\mathrm{W}_t(x, y; L)$ は拡散方程式 $\partial_t u = Lu$ に対する熱核である。

(6.2) は $L = \triangle$ の場合に相当する。異方性 DAE のことを，浅い DAE の輸送表現，または，単に「浅い DAE」とも呼ぶ。特に明示する必要がない場合には作用素 L を省略する。

Remarks

輸送写像 (6.2) の形式は，縮小推定量の分野では Brown's representation of posterior (George et al., 2006) として知られている。実際，DAE は平均値の推定量なので，この形式が現れるのは自然なことである。

Alain の最適解から輸送の式を導くには，(6.4) のような代数的な関係式が活躍した。関係式 (6.4) のように，微分と多項式の積が置き換えられる関数は正規分布に限る。このことは Stein's characterizing operator of

normal distribution $Tf(x) := f'(x) - xf(x),\ f \in C^1(\mathbb{R})$ の性質から従う。すなわち，$Tf \equiv 0$ となる f は標準正規分布に限ることが知られている (Stein, 1972)。

6.2.4　初速ベクトル

浅い DAE を $\Phi_t = \mathsf{id} + t\nabla \log e^{tL}\pi_0$ とする。各点 $x \in \mathbb{R}^m$ を始点とする DAE の輸送軌道 $t \mapsto \Phi_t(x)$ は，$\mathbb{R}^m$ 内の曲線を描く。$t \to 0$ の極限をとると，輸送の初速ベクトルはスコアで与えられることが分かる

$$\frac{\partial}{\partial t}\Phi_t(x; L)\Big|_{t=0} = \lim_{t\to 0}\frac{\Phi_t(x; L) - x}{t} = \nabla \log \pi_0(x). \tag{6.7}$$

極限は半群の性質 $\lim_{t\to 0} e^{tL}\pi_0 = \pi_0$ から直ちに従う。Alain and Bengio (2014) は (6.1) の漸近展開を通じて，$L = \triangle$ の場合と同じ式を導いた。

DAE Φ_t による輸送に伴って，データ分布 π_0 も変形する。変形された分布 π_t は押出測度といい，

$$\pi_t := \Phi_{t\sharp}\pi_0,$$

と書く。$\mathbb{R}^m$ 上の確率測度の空間を $\mathscr{P}(\mathbb{R}^m)$ とする。測度の時間発展に伴う軌道 $t \mapsto \pi_t$ は，$\mathscr{P}(\mathbb{R}^m)$ 内の曲線を描く。輸送に伴う初速ベクトルがスコアで与えられることから，$\mathscr{P}(\mathbb{R}^m)$ における π_t の初速ベクトルは逆拡散の向きになることが示せる。

定理 6.2.2. 浅い *DAE* Φ_t による押出測度を $\pi_t := \Phi_{t\sharp}\pi_0$ とする。このとき

$$\partial_t \pi_{t=0} = -\triangle \pi_0. \tag{6.8}$$

つまり，(6.8) は $\mathscr{P}(\mathbb{R}^m)$ 上の速度ベクトル場とみなせる。ベクトル場と軌道の例は § 6.4.3 を参照せよ。確率測度の空間上のベクトル場についての厳密な扱いについては § 3.13.2 を参照せよ。浅い DAE の場合，一般の $t > 0$ に対しては $\partial_t \pi_t \neq -\triangle \pi_t$ である。

Proof. 重積分の変数変換の公式により，以下の恒等式が成り立つ

$$\pi_0 = \pi_t \circ \Phi_t \cdot |\nabla \Phi_t|.$$

ただし $|\cdot|$ は Jacobian を表す。

まず両辺の対数をとり，t で偏微分する

$$0 = \frac{\nabla \pi_t \circ \Phi_t \cdot \partial_t \Phi_t + \partial_t \pi_t}{\pi_t \circ \Phi_t} + \mathsf{tr}\left[(\nabla \Phi_t)^{-1} \nabla \partial_t \Phi_t\right].$$

ただし右辺第二項は $\log|\cdot|$ の微分に関する公式 (Petersen and Pedersen, 2012, (43))

$$\partial \log|J| = \mathsf{tr}\left[J^{-1}\partial J\right],$$

を用いた。

次に $t = 0$ を代入して整理する

$$\begin{aligned} 0 &= \frac{\nabla \pi_0 \cdot \nabla \log \pi_0 + \partial_t \pi_{t=0}}{\pi_0} + \mathsf{tr}\left[\nabla^2 \log \pi_0\right] \\ &= \frac{|\nabla \pi_0|^2}{\pi_0^2} + \frac{\partial_t \pi_{t=0}}{\pi_0} + \frac{\pi_0 \triangle \pi_0 - |\nabla \pi_0|^2}{\pi_0^2}. \end{aligned}$$

ただし $\Phi_0 = \mathsf{id}$ および $\partial_t \Phi_{t=0} = \nabla \log \pi_0$ を用いた。これを整理して (6.8) を得る □

6.3 合成 DAE と連続 DAE

前節の浅い DAE の解析を元にして，深層 DAE を解析する。まず合成 DAE $\Phi_{0:L}^t$ を導入する。合成 DAE では，合成した時刻 t_k において $\partial_t \pi_{t=t_k} = -\triangle \pi_{t_k}$ が成り立つ。続いて合成 DAE の極限写像として連続 DAE φ_t を導入する。連続 DAE では，各時刻で逆拡散方程式 $\partial_t \pi_t = -\triangle \pi_t$ が成り立つ。数値例では，合成の時間間隔 τ を短くするに連れて，合成 DAE が連続 DAE に収束する様子を可視化する。

6.3.1 合成 DAE

入力データ x_0 は $\mathbb{R}^m$ 上の確率分布 π_0 に従うとする。π_0 に対して計算された DAE を $\Phi_0 : \mathbb{R}^m \to \mathbb{R}^m$ とし，x_0 に Φ_0 を適用して得られる点を $x_1 := \Phi_0(x_0)$ と書く。x_1 は再び $\mathbb{R}^m$ の点であり，確率分布 $\pi_1 := \Phi_{0\sharp}\pi_0$ に従う確率変数である。この操作を繰り返して，π_ℓ に従うデータ x_ℓ から $\pi_{\ell+1} := \Phi_{\ell\sharp}\pi_\ell$ に従うデータ $x_{\ell+1} := \Phi_\ell(x_\ell)$ を得る。合成写像

$$\Phi_{0:L}^t := \Phi_L \circ \cdots \circ \Phi_0$$

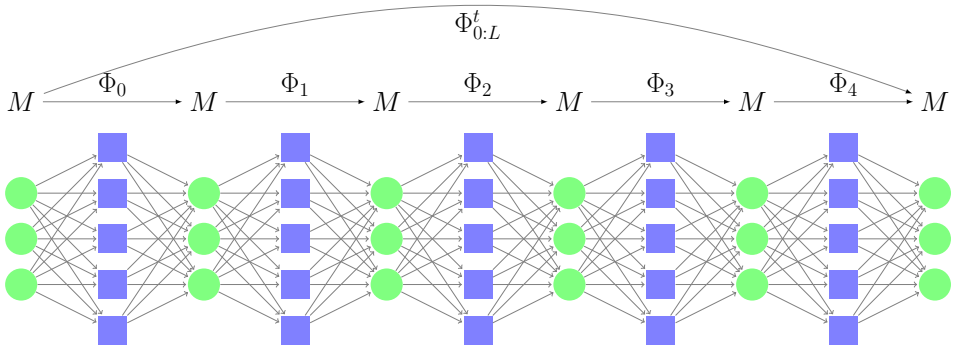


図 6.2: 合成 DAE

を合成 DAE（composition of DAEs）と呼ぶ。ここで，各要素 DAE Φ_ℓ で用いられたノイズの分散を $\tau_\ell I$ として，τ_ℓ を輸送に伴う時間とみなしたときの総時間を $t := \tau_0 + \cdots + \tau_L$ とおいた。また，Φ_ℓ までの累積 '時間' を $t_\ell := \tau_0 + \cdots + \tau_\ell$ と書く。作り方から，合成 DAE は初速だけでなく，各時点 t_ℓ $(\ell = 0, \cdots, L-1)$ において，その速度ベクトルがスコア $\nabla \log \pi_\ell$ と一致する。

6.3.2　連続 DAE

合成 DAE $\Phi^t_{0:L}$ において，総時間 $t = \tau_0 + \cdots + \tau_L$ を固定したまま層数 $L \to \infty$ の極限をとると，速度ベクトルは各時刻でスコアに一致するようになることが期待される。そこで，次のように定義する。

定義 6.3.1. π_0 は $\mathbb{R}^m$ 上の確率測度とする。次の連続力学系の解作用素 (*flow*) $\varphi_t : \mathbb{R}^m \to \mathbb{R}^m$ を連続 *DAE* と呼ぶ

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = \nabla \log \pi_t(x(t)), \quad t \geqslant 0 \tag{6.9}$$

ただし $\pi_t := \varphi_{t\sharp}\pi_0$。

定理 6.3.2 では，適当な正則条件のもとで合成 DAE の極限が連続 DAE に収束することを示す。つまり，連続 DAE は無限層ニューラルネットである。そして，合成 DAE による輸送軌道 $t \mapsto \Phi^t_{0:L}(x_0)$ は，連続 DAE の Euler 式の折れ線近似に相当する。

定義より明らかに，連続 DAE は以下の作用素方程式と同値である。

$$\varphi_0 = \mathsf{id}, \tag{6.10}$$
$$\partial_t \varphi_t = \nabla \log[\varphi_{t\sharp}\pi_0 \circ \varphi_t], \quad t \geqslant 0 \tag{6.11}$$

また，同じことだが，積分方程式に書き換えることもできる。

$$\varphi_t = \mathsf{id} + \int_0^t \nabla \log[\varphi_{s\sharp}\pi_0 \circ \varphi_s]\mathrm{d}s. \tag{6.12}$$

そして，φ_t は非線形半群である。

$$\varphi_{t\to s} \circ \varphi_{0\to t} = \varphi_{0\to s}, \quad 0 \leqslant t \leqslant s.$$

連続 DAE φ_t を一つのニューラルネットとして実現することを考えると，半群性は二つのニューラルネットの合成が一つのニューラルネットとして表されることを表している。この操作は何度も繰り返すことができるので，深層ニューラルネットを浅いニューラルネットに変換することも可能である。

定理 6.2.2 から直ちに次の重要な定理が従う。

定理 6.3.1. $\mathbb{R}^m$ 上の確率測度 π_0 に対する連続 *DAE* を φ_t とする。このとき押出測度 $\pi_t := \varphi_{t\sharp}\pi_0$ は以下の逆拡散方程式（*backward heat equation*）による初期値問題の解である

$$\partial_t \pi_t = -\triangle \pi_t, \quad \pi_{t=0} = \pi_0. \tag{6.13}$$

逆拡散方程式の解釈は次節で加える。

6.3.3 極限の存在と一意性

定理 6.3.2. $\mathbb{R}^m$ 上の確率測度 π_0 を固定する。$\Phi_{0:L}^t$ を L 層からなる総時間 t の合成 *DAE* とし，φ_t を連続 *DAE* とする。ある開集合 Ω が存在して，$\log \pi_0$ はこの中で *Lipschitz* 連続とする。このとき各点 $x \in \Omega$ において，

$$\lim_{L\to\infty} \Phi_{0:L}^t(x) = \varphi_{0\to t}(x). \tag{6.14}$$

Proof. 簡単のため τ_ℓ は ℓ に依らず共通（$\tau_\ell \equiv \tau$）とする。半群の性質により，初速 $\partial_t \Phi_0(x; L)$ は楕円形作用素 L に依らない。従って，合成 DAE の要

素は全て等方的（$L = \triangle$）と仮定して一般性を失わない。仮定より $\log \pi_0$ は Lipschitz 連続なので，軌道の乖離度 $|\Phi^t_{0:L}(x) - \varphi_{0\to t}(x)|$ を $\tau = t/L$ によって上から評価できる．従って，極限 $L \to \infty$ において乖離は 0 に収束する。作り方から，極限関数 $\lim_{L\to\infty} \Phi^t_{0:L}$ は各時刻 t で (6.11) を満たす。従って，常微分方程式の解の一意性により，得られた極限関数は連続 DAE である。 □

6.3.4　数値例

共通の入力データ分布 π_0 に対して，浅い DAE Φ_t，合成 DAE $\Phi^t_{0:L}$, 連続 DAE φ_t を計算し，輸送軌道の違いを視覚的に確かめる。

二次元正規分布

多次元正規分布 $\mathcal{N}(\mu_0, \Sigma_0)$ を初期分布とする浅い DAE Φ_t と連続 DAE φ_t は，以下のように解析的に求められる

$$\Phi_t(x) = (I + t\Sigma_0^{-1})^{-1}x + (I + t^{-1}\Sigma_0)^{-1}\mu_0, \tag{6.15}$$

$$\Phi_{t\sharp}\mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}(\mu_0, \Sigma_0(I + t\Sigma_0^{-1})^{-2}), \tag{6.16}$$

$$\varphi_t(x) = \sqrt{I - 2t\Sigma_0^{-1}}(x - \mu_0) + \mu_0, \tag{6.17}$$

$$\varphi_{t\sharp}\mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}\left(\mu_0, \Sigma_0 - 2tI\right). \tag{6.18}$$

また，合成 DAE $\Phi^t_{0:L}$ は，浅い DAE を繰り返し合成して計算できる。

図 6.3 は，データ分布を

$$\pi_0 = \mathcal{N}\left(\mu_0 = [0, 0], \Sigma_0 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

とした場合に，上記の解析解に基いて数値計算した輸送軌道を示す。各点は格子点（灰色）および π_0 に従って生成された点（有色）から始まり，画面中央に集まる方向に輸送されている。浅い DAE では $t \to \infty$ の極限で原点に収束するのに対し，連続 DAE では $t = 1/2$ で x 軸上に広がったアトラクタに収束する。合成 DAE は，時間の刻みが小さい方が連続 DAE の軌道に近いことが分かる。

混合正規分布

初期分布 π_0 が混合正規分布の場合，浅い DAE の輸送写像は定義に従って解析的に計算できるが，連続 DAE は微分方程式 $\dot{x} = \nabla \log \pi_t(x)$ を数値的に解いて求める。図 6.4, 6.5, 6.6 は，それぞれデータ分布を

$$\pi_0 = 0.5\mathcal{N}\left([-1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.5\mathcal{N}\left([1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

$$\pi_0 = 0.2\mathcal{N}\left([-1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.8\mathcal{N}\left([1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

$$\pi_0 = 0.2\mathcal{N}\left([-1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.8\mathcal{N}\left([1,0], \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

とした場合に，上記の解析解に基いて数値計算した輸送軌道を示す。いずれも浅い DAE では $t \to \infty$ の極限で原点に収束し，合成 DAE では時間の刻みが小さいほど連続 DAE の軌道に近づく。連続 DAE では混合分布の二つのクラスタ中心に向かって輸送する効果があることが分かる。要素分布の重みを変えると，アトラクタの引き込み領域が変化する。特に図 6.5 の連続 DAE では，軌道が交差している。このことは速度場 $\partial_t \varphi_t$ が時間変化していることを反映している。つまり，連続 DAE が非線形半群であることの表れである。

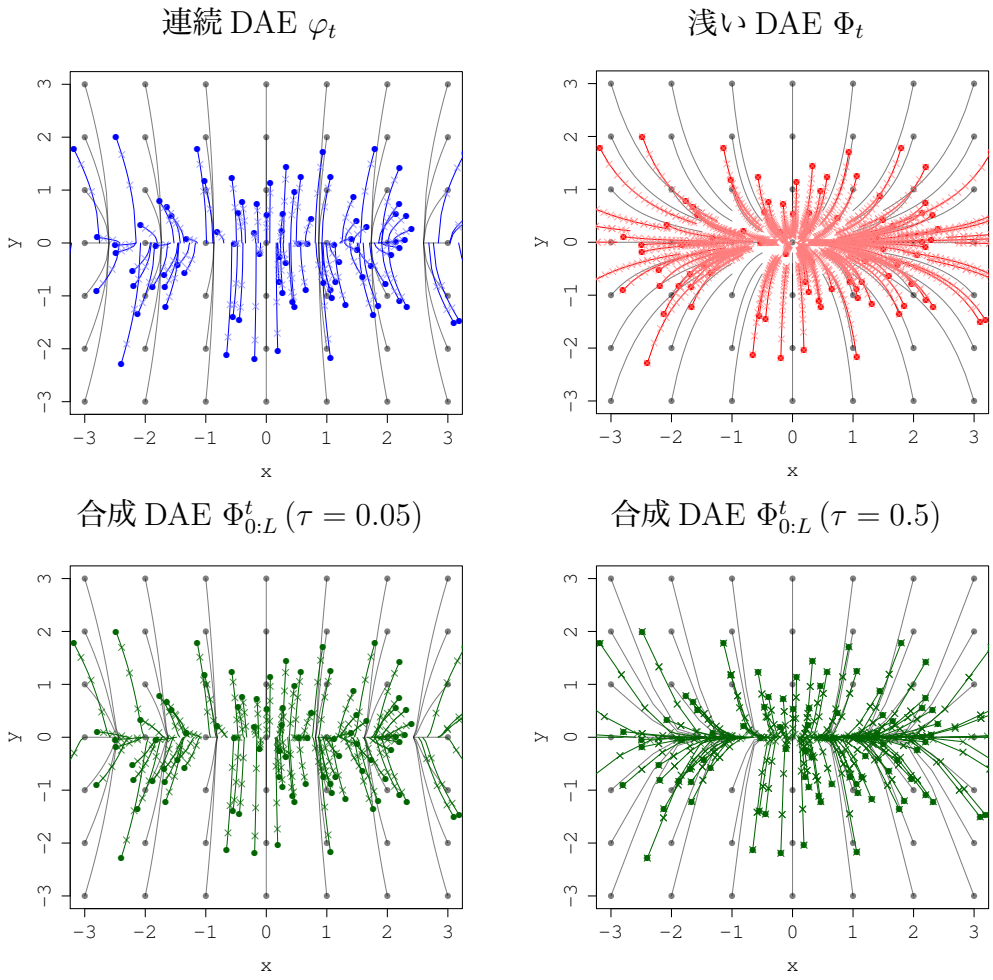


図 6.3: 共通のデータ分布（横長の二次元正規分布）に対して計算された輸送軌道

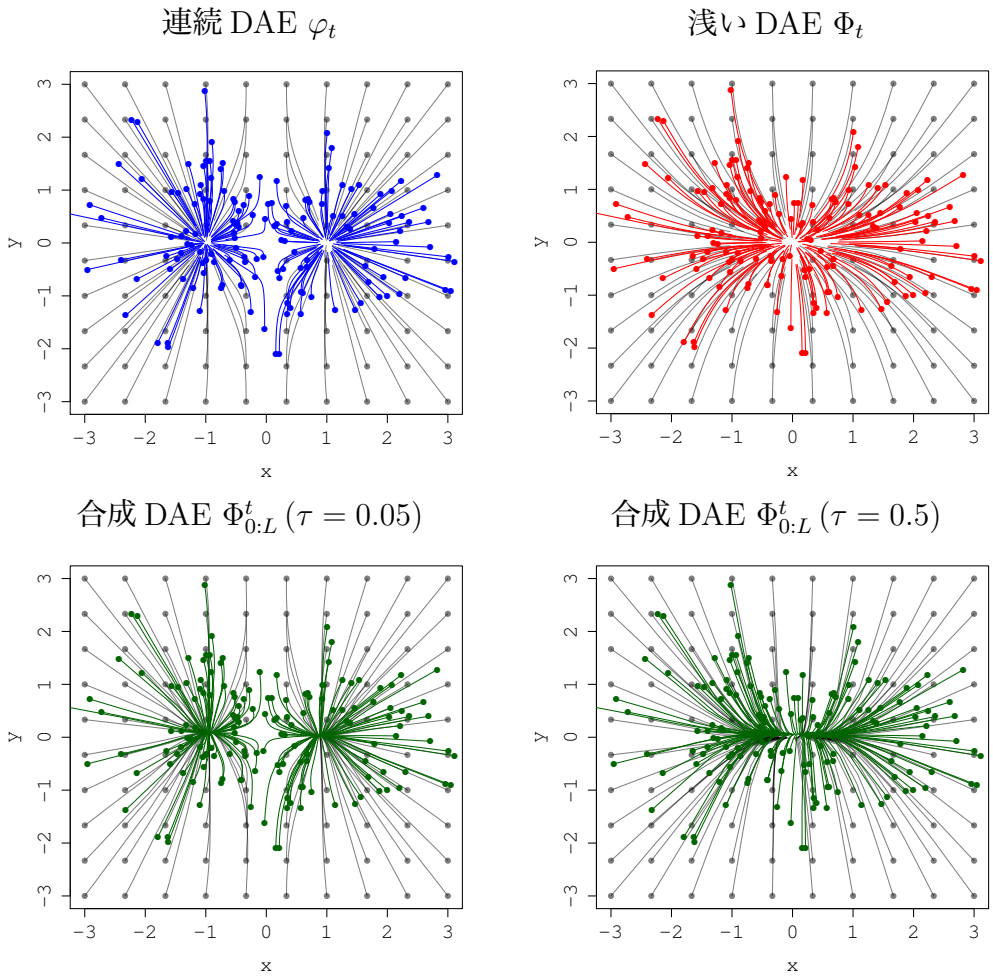


図 6.4: 共通のデータ分布（重みの等しい 2 混合正規分布）に対して計算された輸送軌道

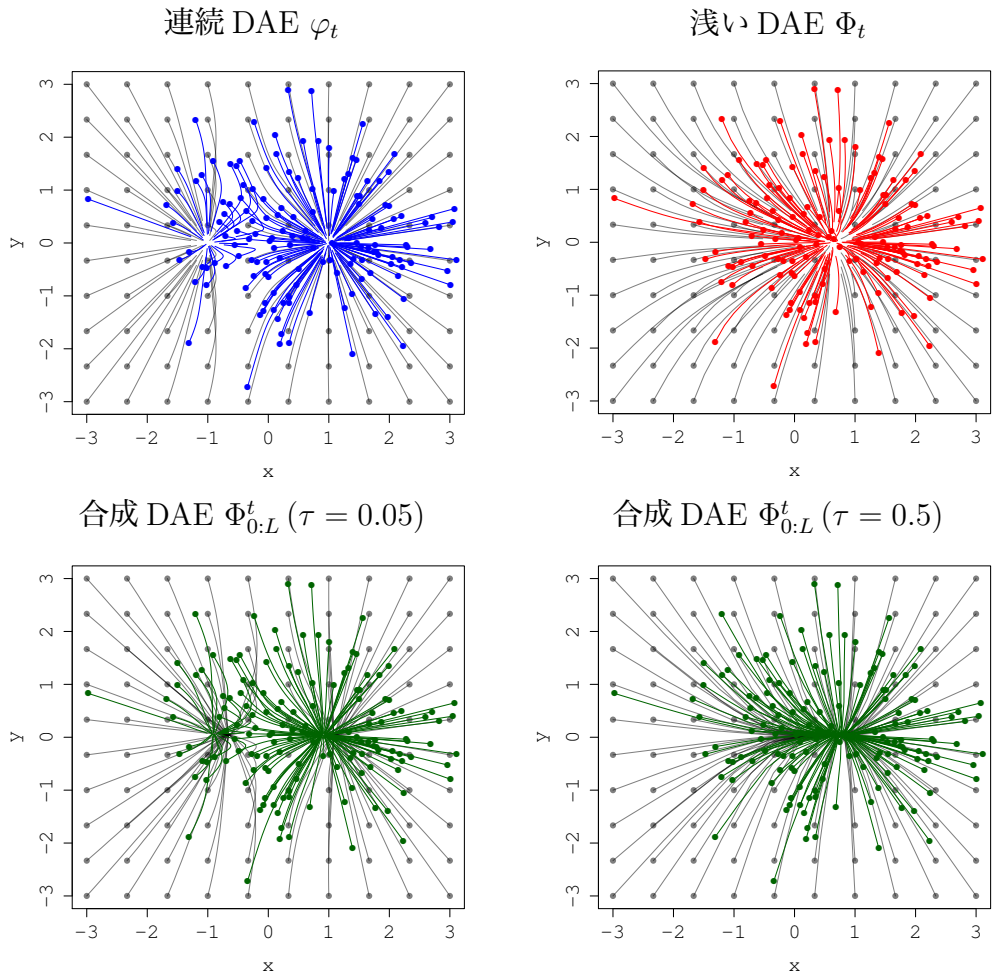


図 6.5: 共通のデータ分布（重みの異なる 2 混合正規分布）に対して計算された輸送軌道

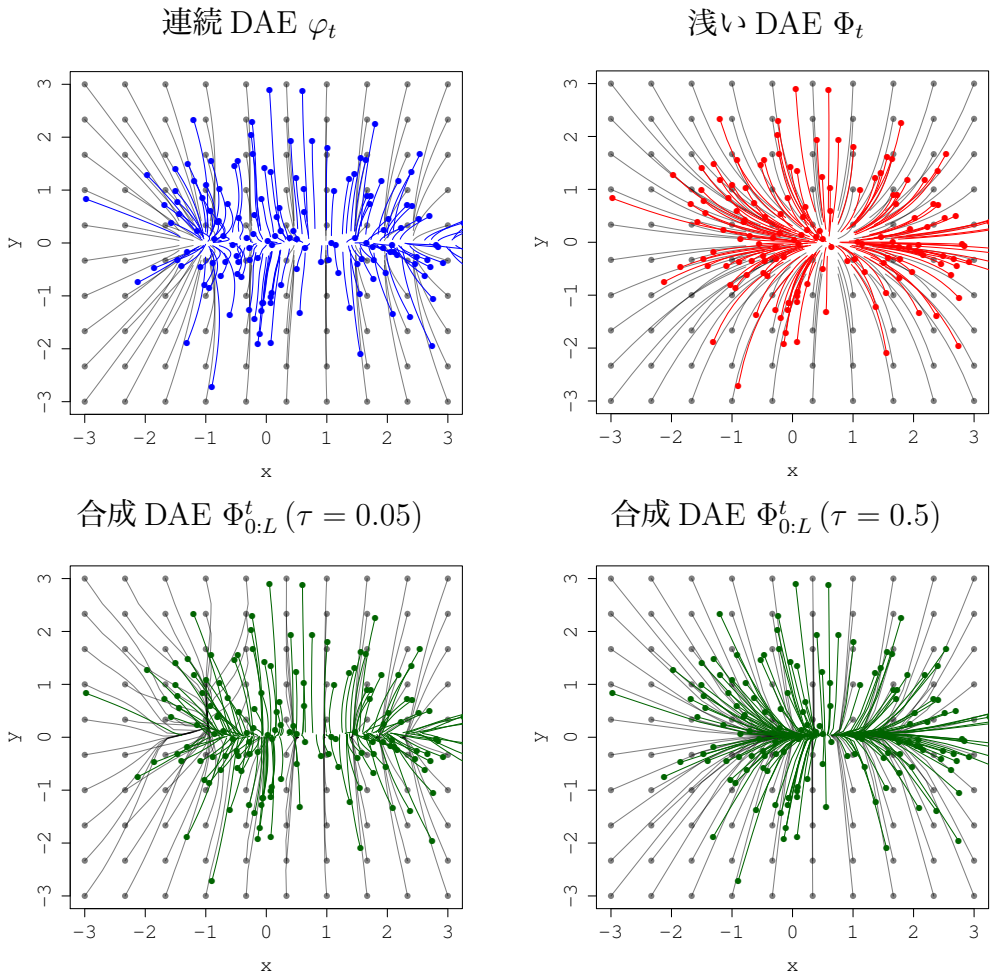


図 6.6: 共通のデータ分布（重みと分散の異なる 2 混合正規分布）に対して計算された輸送軌道

6.4　逆拡散方程式の解釈

定理 6.3.1 により，連続 DAE φ_t の輸送に伴うデータ分布 π_t の時間発展は，逆拡散方程式に従うことが分かった。このことに解釈を加え，数値例を通じて理解を深める。

6.4.1　最終値問題

逆拡散方程式の初期値問題は，通常の拡散方程式の最終値問題と同値である

$$\partial_t u_t = \triangle u_t, \quad u_{t=T} = \pi_0 \quad \text{for some } T$$

ここで u_t は $\mathbb{R}^m$ 上の確率測度である。実際,

$$\pi_t = u_{T-t},$$

とおくと，π_t は (6.13) の解になる。つまり，逆拡散方程式は時間を遡る方向に進む拡散過程を記述している。何らかの方法で最終値 π_T が分かっている場合には，熱核を用いて

$$\pi_t = \mathrm{W}_{T-t} * \pi_T, \quad 0 \leqslant t \leqslant T$$

と書ける。例えばこれを積分方程式 (6.12) と組み合わせることで，連続 DAE が従う別の積分方程式が得られる。

$$\varphi_t = \mathsf{id} + \int_0^t \nabla \log[\mathrm{W}_{T-s} * \pi_T] \circ \varphi_s \mathrm{d}s.$$

6.4.2　エントロピー勾配流

最適輸送理論によれば，拡散方程式はエントロピーを増大させる抽象的勾配流 (abstract gradient flow) である[1] (Otto and Villani, 2000), (Villani, 2009, Th. 23.19)。従って，逆拡散方程式はエントロピーを減少させる抽象的勾配流である

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t = -\mathsf{grad}\, H[\pi_t]. \tag{6.19}$$

[1] § 3.13.2 にアウトラインをまとめた。

ただし，π_t は確率測度の空間 $\mathscr{P}(\mathbb{R}^m)$ 上の点とみなし，方程式は $\mathscr{P}(\mathbb{R}^m)$ 上の常微分方程式（発展方程式）として理解する。$H[\pi] := \mathbb{E}_\pi[-\log \pi]$ はエントロピー汎関数，grad は $\mathscr{P}(\mathbb{R}^m)$ 上に Wasserstein 計量の意味で定義された勾配である。また，実空間 $\mathbb{R}^m$ に立ち戻ってみると，連続 φ_t は各時刻 t で π_t のエントロピーを減らす方向に輸送する写像であることが分かる。

抽象的勾配流のポテンシャル汎関数が明らかになったことで，連続 DAE の軌道や収束先を幾何学に把握できるようになった。例えば，エントロピーを計算することで π_T の見当が付けられる。また，一般にエントロピー汎関数 $H[\pi]$ は下に凸なので，符号を反転した勾配流は不安定になることが予想される。このことは拡散過程の逆問題が一般に不良設定であることとも符合している。

6.4.3 数値例

確率測度の空間 $\mathscr{P}(\mathbb{R}^m)$ でみても，合成 DAE $\Phi^t_{0:L}$ は，連続 DAE φ_t に対する Eulers 式折れ線近似であることが分かった。数値例を用いてこの違いを視覚的に確かめる。

二次元確率測度の空間 $\mathscr{P}(\mathbb{R}^2)$ の部分空間として，次のような二次元正規分布

$$\mathcal{N}\left(\mu_0 = [0,0], \Sigma_0 = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right),$$

の空間 $\mathscr{P}_N(\mathbb{R}^2)$ をとる。(6.16) と (6.18) から分かる通り，浅い DAE と合成 DAE，連続 DAE の軌跡はいずれも，この空間の中で閉じている。$\mathscr{P}_N(\mathbb{R}^2)$ 上でエントロピー汎関数は以下で与えられる

$$H(\sigma_1, \sigma_2) = \frac{1}{2} \log \det \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} + C$$

$\mathscr{P}_N(\mathbb{R}^2)$ におけるエントロピー勾配流と浅い DAE および合成 DAE の軌跡を図 6.7 に示す。いずれも分散が小さくなる方向に流れていることが分かる。浅い DAE や，合成 DAE の軌跡は次第に勾配流から外れ，一点 $(\sigma_1, \sigma_2) = (0, 0)$ に収束することが分かる。なお，$\mathscr{P}_N(\mathbb{R}^2)$ は (σ_1, σ_2) 座標系で平坦である。つまり，図 6.7 における線分の長さは，線分に対応す

る Wasserstein 距離の定数倍に等しい。実際，正規分布同士の Wasserstein 距離は以下で与えられる (Takatsu, 2011, Theorem 2.2)

$$W_2(\mathcal{N}(m,U),\mathcal{N}(n,V))^2 = |m-n|^2 + \mathrm{tr}\, U + \mathrm{tr}\, V - 2\mathrm{tr}\sqrt{U^{1/2}VU^{1/2}},$$

ので，$\mathscr{P}_N(\mathbb{R}^2)$ に制限すると以下が得られるためである

$$W_2\left(\mathcal{N}\left([0,0],\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}\right),\mathcal{N}\left([0,0],\begin{bmatrix}\tau_1^2 & 0\\ 0 & \tau_2^2\end{bmatrix}\right)\right)^2 = (\sigma_1-\tau_1)^2+(\sigma_2-\tau_2)^2.$$

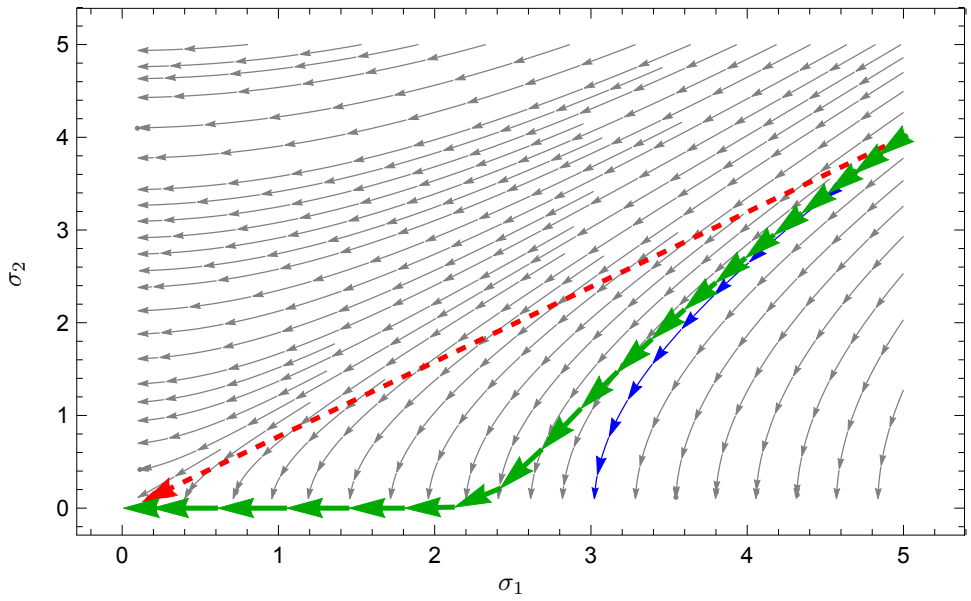


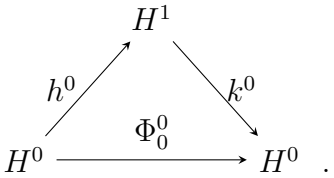
図 6.7: 二次元正規分布の空間における DAE の軌跡。細線はエントロピー勾配流（連続 DAE)，破線と太線はそれぞれ $(\sigma_1,\sigma_2)=(5,4)$ を初期値とする浅い DAE （$t=1000$）と合成 DAE（$\tau=0.8$)。

6.5 積層 DAE と合成 DAE の等価性

輸送解釈を応用して，積層 DAE（stacked DAE）を解析する。積層 DAE は，深層ニューラルネットを初期化する方法（pre-training）の一種として提案された。積層 DAE では，DAE の中間層から得られる特徴量に対して再度 DAE を適用することで，高次の特徴量を得る。従って，積層 DAE は合成 DAE とは異なる写像になる。ところが，積層 DAE を輸送写像とみなすと，線形変換（デコーダー）によって合成 DAE に変換できることが分かる。つまり，積層 DAE から得られる特徴量写像は，合成 DAE として実現できることを示す。

6.5.1 積層 DAE

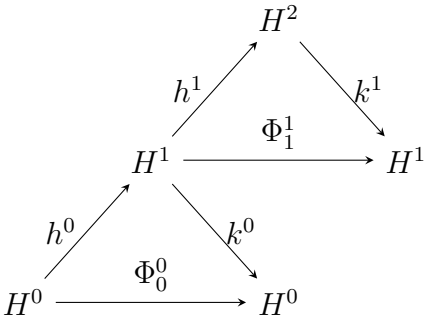
記法を揃える都合上，以下では入力を z^0，入力空間を $H^0 = M_0^0 = \mathbb{R}^m$，入力データが従う確率分布を π_0^0 と書く。入力空間 H^0 上の DAE を $\Phi_0^0 : H^0 \to H^0$, Φ_0^0 のエンコーダー（中間層）を $h^0 : H^0 \to H^1$ とし，デコーダー（出力層）を $k^0 : H^1 \to H^0$ とする。



ただし H^1 は Φ^0 の中間層が値をとる有限次元の Euclid 空間とする。z^0 に h^0 を適用して得られる像を $z^1 := h^0(z^0)$ と書く。

特徴量 z^1 に対して再び DAE の手続きを施すと，高次の DAE $\Phi_1^1 : H^1 \to H^1$ が得られる。ただしノイズは H^1 上で等方的な正規乱数 $\varepsilon \sim \mathcal{N}(0, tI_{H^1})$ を用いる。得られた Φ^1 のエンコーダー h^1 を用いて，高次の特徴量 $z^2 :=$

$h^1(z^1)$ が得られる。



.

このように積層（stack）とは，特徴量 z^ℓ に対して DAE $\Phi_\ell^\ell : H^\ell \to H^\ell$ を学習し，得られたエンコーダー $h^\ell : H^\ell \to H^{\ell+1}$ を用いて高次の特徴量 $z^{\ell+1} := h^\ell(z^\ell)$ を得る操作である。

6.5.2　積層 DAE の輸送写像

記述を簡単にするため，以下の記法を導入する

$$h^{0:L} := h^L \circ \cdots \circ h^0,$$
$$k^{L:0} := k^0 \circ \cdots \circ k^L.$$

特に $h^{0:L}$ を積層 DAE と呼ぶ。図 6.8 に示すとおり，積層 DAE $h^{0:L}$ にデコーダー $k^{L:0}$ を作用すると，H^0 の中での合成 DAE になる（定理 6.5.1）。この操作を復号（decoding）と呼ぶことにする。本節では，定理を述べるためにまず記号を整理する。

まず，積層 DAE $h^{0:\ell} : H^0 \to H^{\ell+1}$ は高次元空間への写像だが，元の入力は $H^0 = \mathbb{R}^m$ の点なので，実際には $H^{\ell+1}$ に埋め込まれた高々 m 次元多様体 $M_{\ell+1}^{\ell+1}$ に値をとる

$$M_{\ell+1}^{\ell+1} := h^{0:\ell}(M_0^0), \quad \ell = 0, \cdots, L.$$

従って，対応する確率分布 $\pi_{\ell+1}^{\ell+1}$ もこの多様体の中に台をもつ

$$\pi_{\ell+1}^{\ell+1} := h_\sharp^{0:\ell}\pi_0^0, \quad \ell = 0, \cdots, L.$$

次に，得られた空間に対してデコーダー k^ℓ を作用させて得られる空間にも記号を付ける。

$$M_{\ell+1}^n := k^{\ell:n}(M_{\ell+1}^{\ell+1}), \quad n = 0, \cdots, \ell$$
$$\pi_{\ell+1}^n := k_\sharp^{\ell:n}\pi_{\ell+1}^{\ell+1}, \quad n = 0, \cdots, \ell.$$

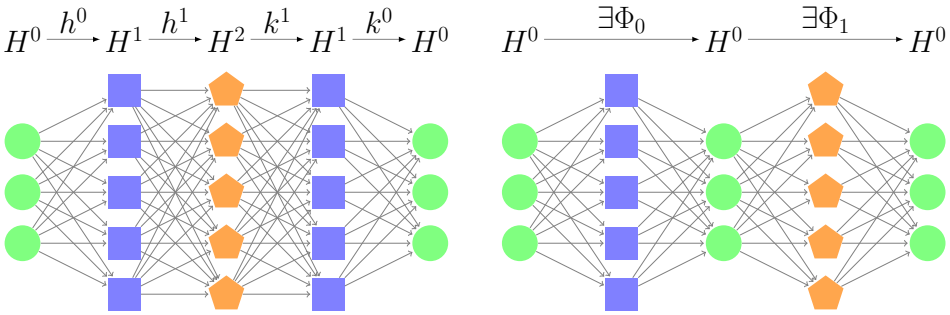


図 6.8: 積層 DAE を復号したもの（左）は，合成 DAE（右）になる。

作り方から M_n^ℓ は高次元 Euclid 空間 H^ℓ に埋め込まれた高々 m 次元の多様体である。π_n^ℓ の台は M_n^ℓ に含まれる。

最後に，M_n^ℓ と M_{n+1}^ℓ を結ぶ写像を Φ_n^ℓ と書く。すなわち $k^{n:\ell} \circ h^{0:n} : M_0^0 \to M_{n+1}^\ell$ を用いて，

$$\Phi_n^\ell := (k^{n:\ell} \circ h^{0:n}) \circ (k^{(n-1):\ell} \circ h^{0:(n-1)})^{-1} : M_n^\ell \to M_{n+1}^\ell,$$

とおく。後述する定理 6.5.2 により，$\Phi_n^{\ell+1}$ が異方性 DAE であれば，図式を可換にする写像 Φ_n^ℓ が存在して，しかも異方性 DAE である。従って，Φ_n^ℓ は押し並べて異方性 DAE と仮定してよい。

ここまで定義した記号を図 6.9 に示す。大三角形の左側の斜辺が積層 DAE $h^{0:L}$, 右側の斜辺が復号に用いるデコーダー $k^{L:0}$, 底辺が復号の結果として得られる合成 DAE $\Phi_{0:L}$ に相当する。定理 6.5.2 で示す図式の可換性を使うと，積層 DAE と合成 DAE の等価性を主張する次の定理が従う。

定理 6.5.1. 各 ℓ に対してエンコーダーの制限写像 $h^\ell|_{M_\ell^\ell}$ は連続単射とし，さらに各 n に対してデコーダーの制限写像 $k^\ell|_{M_n^{\ell+1}}$ は単射とする。このとき積層 *DAE* を復号したものは合成 *DAE* である。

$$k^{L:0} \circ h^{0:L} = \Phi_L^0 \circ \cdots \circ \Phi_0^0.$$

Proof. 定理 6.5.2 に示す位相共役性

$$k^\ell \circ \Phi_n^{\ell+1} = \Phi_n^\ell \circ k^\ell$$

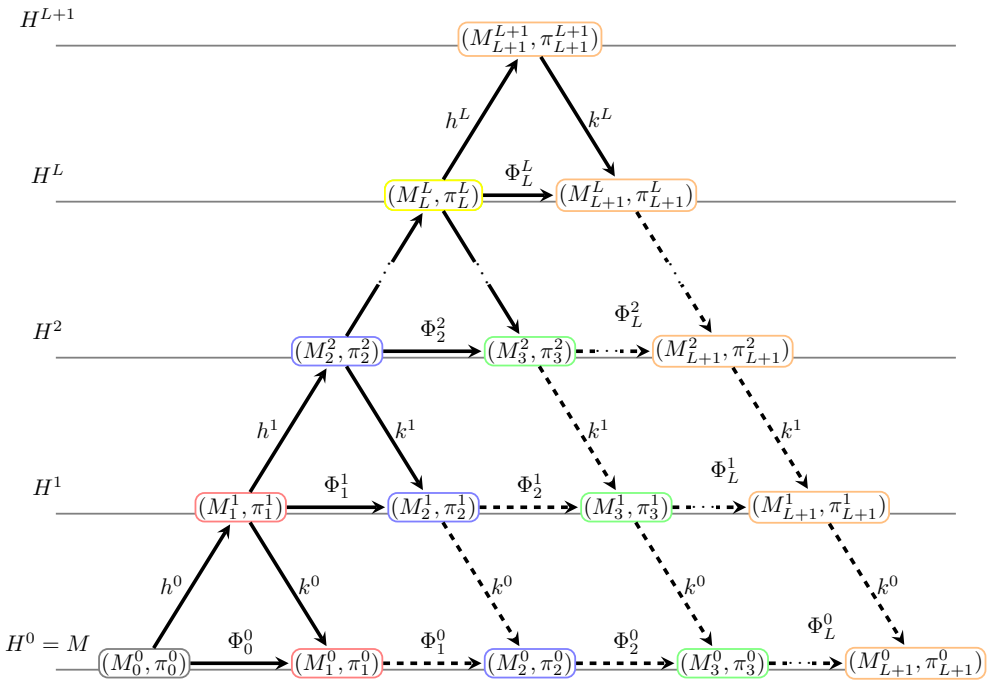


図 6.9: 積層 DAE を合成 DAE に復号する過程を表す可換図式

を再帰的に適用することで，以下のように示せる

$$
\begin{aligned}
& k^{L:0} \circ h^{0:L} \\
&= k^{(L-2):0} \circ k^{L-1} \circ \Phi_L^L \circ h^{L-1} \circ h^{0:(L-2)} \\
&= k^{(L-2):0} \circ \Phi_L^{L-1} \circ k^{L-1} \circ h^{L-1} \circ h^{0:(L-2)} \\
&= k^{(L-2):0} \circ \Phi_L^{L-1} \circ \Phi_{L-1}^{L-1} \circ h^{0:(L-2)} \\
&\cdots \\
&= \Phi_L^0 \circ \Phi_{L-1}^0 \circ \cdots \circ \Phi_0^0.
\end{aligned}
$$

□

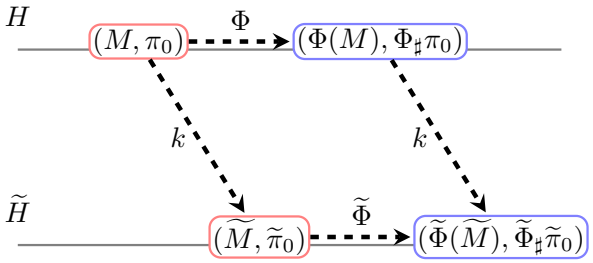
図式の可換性からは，さらに以下のような関係式も見出だせる。

$$
h^\ell = \left(k^{(\ell+1):0}|_{M_{\ell+1}^{\ell+1}} \right)^{-1} \circ \Phi_n^0 \circ k^{\ell:0}.
$$

この式は，高次元空間どうしの写像である $h^\ell : H^\ell \to H^{\ell+1}$ が，低次元多様体である M への写像 $k^{\ell:0}$ と，M の中での変換 Φ_n^0，M からの写像

$\left(k^{(\ell+1):0}|_{M_{\ell+1}^{\ell+1}}\right)^{-1}$ に因子分解できることを示している。つまり，積層 DAE の低ランク性を反映している。

6.5.3 位相共役性



主張を述べる前に，記号を整理する。

まず，深層ニューラルネットの隣接する二つの中間層の中間層素子が値をとる空間を $H := \mathbb{R}^J, \widetilde{H} := \mathbb{R}^I$ とする。$\nabla, \nabla^2 (\widetilde{\nabla}, \widetilde{\nabla}^2)$ はそれぞれ $H(\widetilde{H})$ 上の勾配および Hesse 行列を表し，便宜的に $J(I)$ 次元の縦ベクトルおよび $J \times J(I \times I)$ 行列として扱う。

M は H に埋め込まれた滑らかな m 次元多様体とし，π_0 は M 上の C^2 級確率密度関数とする。π_0 は入力データの分布を表す。$C^2(H)$ 上の楕円型作用素を

$$\begin{aligned} L_t u(z) := a(z,t)^\top \nabla^2 u(z) a(z,t) \\ + b(z,t)^\top \nabla u(z) + c(z,t)^\top u(z), \quad u \in C^2(H) \end{aligned}$$

と定義する。ただし a, b, c はそれぞれ H に値をとるテンソルとし，縦ベクトルとみなす。L_t による DAE $\Phi : H \to H$ を

$$\Phi = \mathsf{id}_H + tK \nabla \log e^{tL_t} \pi_0,$$

とする。ただし K は $J \times J$ 正定値対称行列である。

線形写像 $k : H \to \widetilde{H}$ を固定する。k による M の像空間を $\widetilde{M} := k(M)$ とし，π_0 の押出測度を $\widetilde{\pi}_0 := k_\sharp \pi_0$ とする。$C^2(\widetilde{H})$ 上の楕円形作用素を

$$\begin{aligned} \widetilde{L}_t \widetilde{u}(x) := \widetilde{a}(x,t)^\top \widetilde{\nabla}^2 \widetilde{u}(x) \widetilde{a}(x,t) \\ + \widetilde{b}(x,t)^\top \widetilde{\nabla} \widetilde{u}(x) + \widetilde{c}(x,t)^\top \widetilde{u}(x), \quad \widetilde{u} \in C^2(\widetilde{H}) \end{aligned}$$

と書く。ただし $\tilde{a}, \tilde{b}, \tilde{c}$ はそれぞれ $\widetilde{H}$ に値をとるテンソルとする。$\widetilde{L}_t$ による DAE $\widetilde{\Phi} : \widetilde{H} \to \widetilde{H}$ を

$$\widetilde{\Phi} := \mathsf{id}_{\widetilde{H}} + t\widetilde{K}\,\widetilde{\nabla} \log e^{t\widetilde{L}_t} \tilde{\pi}_0$$

と書く。ただし $\widetilde{K}$ は $I \times I$ 正定値対称行列である。

定理 6.5.2. (M, π_0) および L_t による *DAE* Φ が与えられているものとする。線形写像 $k : H \to \widetilde{H}$ は，M への制限写像 $k|_M$ が単射であるとする。このとき $C^2(\widetilde{H})$ 上の楕円形作用素 $\widetilde{L}_t$ が存在して，$(\widetilde{M}, \tilde{\pi}_0)$ および $\widetilde{L}_t$ による *DAE* $\widetilde{\Phi} : \widetilde{H} \to \widetilde{H}$ に対して以下が成り立つ

$$k \circ \Phi|_M = \widetilde{\Phi} \circ k|_M. \tag{6.20}$$

証明は付録 A.6 を見よ。

6.5.4　数値例

二次元のスイスロールデータに対して二段の積層 DAE $h^1 \circ h^0$ と二段の合成 DAE $\Phi_1 \circ \Phi_0$ を訓練し，獲得された輸送写像で入力データ（黒）を輸送させた結果を示す。赤が一段目，青が二段目の輸送結果である。ただし積層 DAE の輸送結果は一段目を $k^0 \circ h^0(x)$, 二段目を $(k^0 \circ k^1) \circ (h^1 \circ h^0)(x)$ として $\mathbb{R}^2$ に引き戻した。いずれも輸送が進むに連れて，細い線状に収束することが分かる。つまり，積層 DAE を復号したものと，合成 DAE とは，類似の働きを示す。

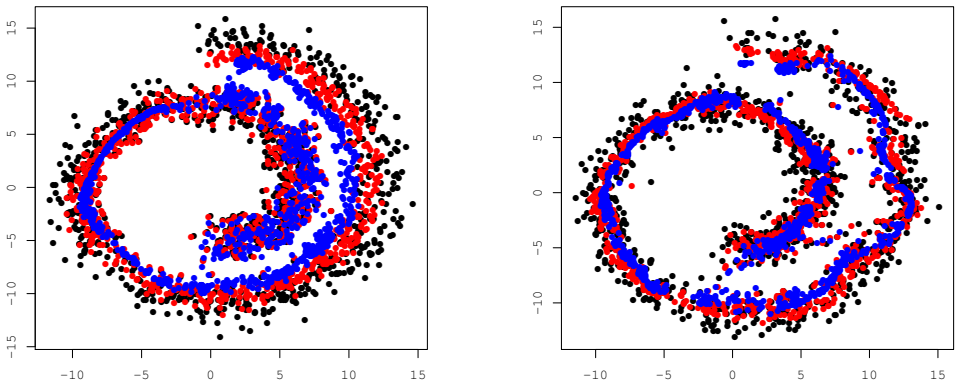


図 6.10: スイスロールに対して積層 DAE（左）と合成 DAE（右）による輸送を適用した結果

6.6　深層 DAE の積分表現

前節で導入した復号によって，積層 DAE $h^{0:L}$ は合成 DAE $\Phi_{0:L} = k^{L:0} \circ h^{0:L}$ に変換できることが分かった。合成 DAE の各要素 DAE $\Phi_\ell = \mathsf{id} + t\nabla \log e^{tL_\ell}\pi_\ell$ は浅いニューラルネットに相当するので，要素毎に積分表現 $g_\ell := \mathscr{R}_\eta^\dagger \mathscr{R}_\psi \Phi_\ell$ にすることで，深層 DAE の積分表現

$$[\mathscr{R}_\eta^\dagger \mathscr{R}_\psi \Phi_L] \circ \cdots \circ [\mathscr{R}_\eta^\dagger \mathscr{R}_\psi \Phi_0],$$

が構成できる。

要素 DAE Φ_ℓ のリッジレット変換を計算するため，以下では輸送写像を適当なコンパクト集合 K 上の外側では零として考える。記述を簡単にするため $\mathcal{F}_\ell := \frac{1}{2}|\cdot|^2 + t \log e^{tL_\ell}\pi_\ell$ とおく。すなわち，次の関係式が成り立つ

$$\Phi_\ell = \nabla \mathcal{F}_\ell.$$

要素 DAE のリッジレット変換は，部分積分を用いて以下のように計算できる

$$\begin{aligned}\mathscr{R}_\psi \Phi_\ell(a,b) &= \int_K \nabla \mathcal{F}_\ell(x) \overline{\psi(a \cdot x - b)} \mathrm{d}x \\ &= -a \int_K \mathcal{F}_\ell(x) \overline{\psi'(a \cdot x - b)} \mathrm{d}x = -a \mathscr{R}_{\phi'} \mathcal{F}_\ell(a,b).\end{aligned}$$

ただし $\psi'(z) := \frac{\mathrm{d}}{\mathrm{d}z}\psi(z)$ と書いた。興味深いことに，要素 DAE を成分毎にリッジレット変換して得られる “情報” は，ポテンシャル関数 $\mathcal{F}_\ell$ のリッジレット変換として集約できることが分かる。

これを用いて，要素 DAE の積分表現は以下で与えられる

$$\begin{aligned}g_\ell(x) &= \mathscr{R}_\eta^\dagger \mathscr{R}_\psi \Phi_\ell(x) \\ &= -\int_{\mathbb{Y}^{m+1}} a \mathscr{R}_{\psi'} \mathcal{F}_\ell(a,b) \eta(a \cdot x - b) \mathrm{d}a\mathrm{d}b.\end{aligned}$$

従って，例えば二つの輸送写像を合成した場合の積分表現は次のように計算できる

$$\begin{aligned}&g_{\ell+1} \circ g_\ell(x) \\ &= -\int_{\mathbb{Y}^{m+1}} a \mathscr{R}_{\psi'} \mathcal{F}_{\ell+1}(a,b) \eta \left(\int_{\mathbb{Y}^{m+1}} (a \cdot a' \mathscr{R}_{\psi'} \mathcal{F}_\ell)(a',b') \eta(a' \cdot x - b') \mathrm{d}a'\mathrm{d}b' - b \right) \mathrm{d}a\mathrm{d}b \\ &= -\int_{\mathbb{Y}^{m+1}} a \mathscr{R}_{\psi'} \mathcal{F}(a,b) \eta(a \cdot x - b) \mathrm{d}a\mathrm{d}b.\end{aligned}$$

ただし第二式は $\Phi_{\ell+1} \circ \Phi_\ell$ を一つの輸送写像 $\Phi = \nabla \mathcal{F}$ とみた場合の式である。Brenier の定理により，このようなポテンシャル $\mathcal{F}$ は必ず存在する。このように，輸送ないし力学系の半群性を利用して，入れ子の問題が解消できる。

6.7 まとめ

デノイジング・オートエンコーダー（DAE）が輸送写像とみなせることを示し，輸送解釈を通じて深層 DAE の性質を調べた。

DAE の学習アルゴリズムは，目的関数 $\mathbb{E}_{\tilde{x},x}|x - g(\tilde{x})|^2$ の最小化問題と等価である。Alain and Bengio (2014) は，変分計算により DAE の最適解 $g(x)$ を導いた。定理 6.2.1 では，この $g(x)$ が輸送写像になることを示した。そこで，輸送写像を一般化した異方性 DAE を，浅い DAE の輸送表現 Φ_t として定義した。

(6.7) に示した通り，浅い DAE の輸送に伴う初速ベクトル $\partial_t \Phi_{t=0}$ は，データ分布のスコア $\nabla \log \pi_0$ で与えられる。このことを用いて，定理 6.2.2 では，浅い DAE による押出測度 $\pi_t := g_{\sharp t}\pi_0$ の初速ベクトル $\partial_t \pi_{t=0}$ が負のラプラシアン $-\triangle \pi_0$ で与えられることを示した。つまり，DAE の押出測度は拡散方程式を遡る向きに発展するのである。この性質は初速ベクトルに限るが，浅い DAE を合成すると，合成の度にこの性質が表れる。そこで，合成を無限に繰り返した極限写像として連続 DAE φ_t を導入した。厳密には，φ_t は極限写像の性質として期待される常微分方程式 $\dot{x} = \nabla \log \pi_t(x)$ の解作用素として定義し，定理 6.3.2 において合成 DAE の極限写像が連続 DAE に収束することを示した。§ 6.3.4 では，π_0 が二次元（混合）正規分布の場合に，輸送軌道を数値的に計算して，浅い DAE と合成 DAE および連続 DAE の軌道の違いを視覚的に示した。

作り方から，連続 DAE による押出測度 π_t は，各時刻で逆拡散方程式に従う（定理 6.3.1）。逆拡散方程式の初期値問題は，拡散方程式の最終値問題と等価である。Wasserstein 幾何学によれば，拡散方程式はエントロピー汎関数の勾配流なので，逆拡散方程式もまたエントロピーを減らす向きの勾配流である。一般に，拡散方程式の逆問題は不良設定なので，連続 DAE の意味は慎重に吟味する必要がある。統計的には，DAE Φ_t は平均値 x の推定量 $\hat{x}$ なので，エントロピーを減らすことが了解される。§ 6.4.3 では，正規分布の空間の部分空間においてエントロピー勾配流を計算し，浅い DAE と合成 DAE および連続 DAE の確率測度の空間における軌道

の違いを視覚的に示した。

合成 DAE は，連続 DAE の Euler 式折れ線近似とみなせる。合成 DAE は高々有限層の深層ニューラルネットであり，連続 DAE は連続無限層のニューラルネットに相当する。このように，ニューラルネットを輸送写像としてモデル化することで，深層構造の性質が調べられる。深層ニューラルネットを輸送写像として解釈する方法の応用例として，積層 DAE と合成 DAE の等価性を証明した（定理 6.5.1）。定理は DAE 間の位相共役性（定理 6.5.2）を繰り返し適用して示す。積層 DAE の中間層写像の構成は複雑だが，いかに複雑でも中間層は入力データの輸送写像であることに着目すると，積層 DAE は合成 DAE に変換できることが示せる。

本研究を通じて，深層 DAE は合成 DAE として書けることが分かったので，合成 DAE を要素毎に積分表現にすることで，深層ニューラルネットの積分表現を導いた。輸送写像に見られる $x \mapsto x + f(x)$ という形式は，ResNet や GoogLeNet, Highway Network などの大規模なネットワークを学習させるためのヒューリスティクスとして用いられている。従って，多くの深層ニューラルネットは輸送写像として理解できることが期待される。特に教師あり学習の場合には，多成分拡散方程式のような構造が表れると予想できる。個別の輸送写像に対してリッジレット変換を計算する方法は今後の重要な課題である。

深層 DAE の中間層が表すもの

深層 DAE から得られる特徴量は，輸送写像のリッジレット変換を離散化したものである。おそらく，深層 DAE から得られた特徴量を眺めてみても，特徴量の意味を解読することは困難であろう。深層 DAE は，特定の文法に従ったコードというよりは，輸送写像という機能を学習しているからである。

深層化の極限である連続 DAE の性質を鑑みると，深層 DAE にはエントロピーを低減する作用がある。エントロピーを減らすことは，次元削減や多様体学習につながる。この作用は層を深めるほど強調されるので，エントロピーを減らす効果を期待するのであれば，積極的に深層化すべきである。ただし，DAE は教師なし学習なので，全てのタスクに対して有効な表現が得られるわけではない。使い方を誤れば過学習も起こしうることに注意せよ。

二段階学習の定式化

深層ニューラルネットを輸送写像 $h_t(x)$ と線形出力 $k(z)$ に分ける。初期の深層学習で行われていた二段階学習は，次のような交互最適化とみなせる

$$\text{minimize} \quad \mathbb{E}|k \circ h_{t=T}(x) - f(x)|^2 + \lambda \cdot \int_0^T \mathbb{E}|h_t(x) - \varphi_t(x)|^2 \mathrm{d}t \quad \text{w.r.t. } h_t, k.$$

ただし φ_t は連続 DAE，f は本来の近似対象である。交互最適化は各項一回ずつであり，プレトレーニングが第二項の最適化，ファインチューニングが第一項の最適化に相当する。f を近似するという目的を達成するためには第一項のみで十分だが，深さパラメータ t から見れば第一項は境界条件に過ぎず，第一項だけでは問題の自由度が高すぎることが分かる。従って，正則化項として第二項が加えられている。第二項は中間層を連続 DAE に近づけるという意味であり，これは積層 DAE の極限が連続 DAE になるという，本研究の解析結果を反映している。連続 DAE は必ずしも関数 f を近似するために有利な正則化であるとは限らないが，例えば短時間（$T \to 0$）のときはオートエンコーダー（恒等写像）になるので，少なくとも情報の損失は起こさないような正則化になっていることが分かる。

第7章　積分表現の離散化による学習法

積分表現を離散化することで，学習済のニューラルネットが得られる。Sonoda and Murata (2015) では，これをバックプロパゲーションの初期値として利用することで，一様乱数などの汎用的な初期値よりも収束が速くなることを実験的に示した。

7.1　はじめに

ニューラルネットのバックプロパゲーション学習は非凸最適化問題であるため，パラメータの初期値が最適解から離れていると，学習過程で局所解やプラトー領域に捕われ，学習が停滞する。

近似対象 $f(x)$ に対して，リッジレット変換の絶対値 $|\mathscr{R}_\psi f(a,b)|$ は，ニューラルネットによって $f(x)$ を近似するための各パラメータ (a,b) の有用度を表している。従って，$|\mathscr{R}_\psi f(a,b)|$ の値が高い (a,b) を抽出することで，バックプロパゲーションの初期値にできる。$|\mathscr{R}_\psi f(a,b)|$ を正規化して確率分布 $\mu(a,b)$ とみなしたものをパラメータのオラクル分布 (oracle distribution) と呼ぶ

$$\mu(a,b) := \frac{|\mathscr{R}_\psi f(a,b)|}{\int_{\mathbb{Y}^{m+1}} |\mathscr{R}_\psi f(a,b)| \mathrm{d}a\mathrm{d}b}.$$

オラクル分布からサンプリングしたパラメータは，汎用的な正規分布や一様分布から生成されたパラメータよりも有利にバックプロパゲーションを開始できる。特に低次元の問題では，中間層パラメータをオラクルサンプルで学習し，出力層パラメータを線形回帰によってフィッティングすることで，バックプロパゲーションを経ずに高い精度を得ることができる。

7.1.1　関連研究

バックプロパゲーションのランダム初期化を効果的に行う方法として，活性化関数が直線的に動作する線形領域に初期値を振る方法が知られている (LeCun et al., 2012)。線形領域の外側は飽和領域と呼ばれ，訓練誤差が消滅するため学習が緩慢になる領域である。例えば Bengio ら (LeCun et al., 2012) は，入力次元 m に対して平均 0，標準偏差 $m^{-1/2}$ となるようにランダムサンプリングすることで，パラメータの対称性が破れ，学習が効果的に進むことを示した 。

深層学習では，活性化関数として ReLU を使うことが標準的である。ReLU の場合，活性領域が飽和しないことがメリットの一つとして挙げられる。線形領域に注目した方法は簡便かつ実用的だが，パラメータが万遍なく分布するために高次元では非効率である。提案手法ではオラクル分布に従ってサンプリングすることで効率的に初期値を選ぶことができる。

パラメータを初期化する別の方法として，教師なし学習を用いる方法も多く提案されている。代表的な入力ベクトルに対して選択的に反応するように初期化する方法 (Denoeux and Lengelle, 1993) や，k-means などのクラスタリングを利用する方法は古典的である (Coates, 2012)。深層学習においては，過学習を回避するためにプレトレーニング (pre-training) が用いられる。これはオートエンコーダなどの教師なし学習によってパラメータを初期化する方法である。

サンプリングによる学習という観点では，逐次モンテカルロ法 (SMC) (Doucet et al., 2001) などのベイズ学習も利用される。SMC は汎用的な初期分布から反復法によってパラメータの分布を推定するのに対し，オラクル分布は初めからパラメータの分布が計算できる。

本研究のように積分表現理論を学習に応用する研究の例は少数だが，例えば Sprecher (1996, 1997) が手がけている。

7.2 オラクル分布とサンプリング学習

7.2.1 問題設定

教師ありデータ $\{(x_s, y_s)\}_{s=1}^{S} \subset \mathbb{R}^m \times \mathbb{R}$ を用いてニューラルネットを学習させる。ただしデータは，真の関数 $f : \mathbb{R}^m \to \mathbb{R}$ があって，

$$y_s = f(x_s) + \text{noise}$$

の形で与えられるものとする。本章では，中間層素子数が n 個の浅いニューラルネットを

$$g(x) = \sum_{j=1}^{n} c_j \cdot \eta\left(a_j \cdot x - b_j\right),$$

と書く。ここで $\eta : \mathbb{R} \to \mathbb{R}$ を活性化関数と呼び，$(a_j, b_j) \in \mathbb{R}^m \times \mathbb{R}$ を中間層パラメータ，$c_j \in \mathbb{R}$ を出力層パラメータと呼ぶ。ニューラルネットの学習問題は以下のように定式化できる

$$\text{minimize} \quad \frac{1}{S}\sum_{s=1}^{S} \left| y_s - g\left(x_s; \{a_j, b_j, c_j\}_{j=1}^{n}\right) \right|^2 \quad \text{w.r.t. } \{a_j, b_j, c_j\}_{j=1}^{n}.$$

7.2.2 オラクル分布によるサンプリング学習

近似対象の関数 $f(x)$ を固定する。以下では簡単のため，リッジレット関数 ψ は活性化関数 η に対して許容的かつ，$\|\mathscr{R}_\psi f\|_1 := 1$ となるように選ばれているものとする。このとき，近似対象の関数 $f(x)$ に対し，オラクル分布 $\mu(a, b)$ と係数 $\mathrm{T}(a, b)$ を以下で定義する

$$\mu(a, b) := |\mathscr{R}_\psi f(a, b)|, \quad (a, b) \in \mathbb{Y}^{m+1},$$
$$\mathrm{T}(a, b) := \frac{\mathscr{R}_\psi f(a, b)}{|\mathscr{R}_\psi f(a, b)|}, \quad (a, b) \in \mathbb{Y}^{m+1}.$$

オラクル分布から独立に生成されたサンプルを $\{(a_j, b_j)\}_{j=1}^{n}$ とおく

$$(a_j, b_j) \sim \mu(a, b), \quad (j = 1, \cdots, n).$$

各点 $x \in \mathbb{R}^m$ において，大数の法則により以下が成り立つ (Murata, 1996)

$$\frac{1}{n}\sum_{j=1}^{n} c_j \cdot \eta(a_j \cdot x - b_j) \to^p f(x) \quad \text{as} \quad n \to \infty.$$

サンプリング学習では，オラクル分布 $\mu(a, b)$ に従ってサンプルを生成し，最小二乗法によって c_j を決定する。

7.3　オラクル分布からのサンプリング法

7.3.1　オラクル分布の計算

リッジレット変換 $\mathscr{R}_\psi f(a,b)$ はデータ $\{(x_s, y_s)\}_{s=1}^S$ からモンテカルロ積分によって推定する

$$\begin{aligned}\mathscr{R}_\psi f(a,b) &= \frac{1}{Z}\int_{\mathbb{R}^m} f(x)\overline{\psi(a\cdot x - b)}\mathrm{d}x \\ &\approx \frac{1}{SZ}\sum_{s=1}^{S} y_s \cdot \psi(a\cdot x_s - b),\end{aligned} \tag{7.1}$$

ただし $Z = K_{\psi,\eta}\|\mathscr{R}_\psi f\|_1$ は理論上必要な正規化係数である。後にオラクル分布からのサンプリングに用いる棄却法や MCMC などの汎用的なサンプリング法では，確率分布の定数倍を無視することができるので，Z を具体的に求める必要はない。

7.3.2　リッジレット関数の計算

Sonoda and Murata (2015) では，ReLU を含む様々な活性化関数に対してリッジレット関数の例を計算した。以下では，活性化関数としてシグモイド関数を用いる場合の例 (Murata, 1996) を説明する。

活性化関数 η として標準的なシグモイド関数 $\sigma(z) := \frac{1}{1+\exp(-z)}$ を足しあわせたシグモイド対を用いる

$$\eta(z) := \frac{1}{H}\left\{\sigma(z+h) - \sigma(z-h)\right\}, \quad (h>0),$$

ここで $H := \sigma(h) - \sigma(-h)$ は活性化関数の最大値を 1 に正規化する定数である。

リッジレット関数 ψ は標準軟化子

$$\rho(z) = \exp\frac{1}{z^2-1}\mathbf{1}_{[-1,1]}(z)$$

を用いて，以下のようにとれる

$$\psi(z) = \begin{cases}\rho^{(m)}(z) & m \text{ even} \\ \rho^{(m+1)}(z) & m \text{ odd}\end{cases}.$$

標準軟化子 $\rho(z)$ の高階導関数 $\rho^{(k)}(z)$ は以下のように解析的に計算できる

$$\rho^{(k)}(z) = \frac{P_k(z)}{(z^2-1)^{2k}}\rho(z) \quad (k = 0, 1, 2, \cdots),$$

ここで $P_k(z)$ は z の多項式であり，以下の漸化式によって求められる

$$\begin{aligned} P_0(z) &\equiv 1, \\ P_{k+1}(z) &= P_k'(z)(z^4 - 2z^2 + 1) + P_k(z)\left\{-4kz^3 + 2(2k-1)z\right\}. \end{aligned}$$

一般に $\rho^{(k)}(z)$ は k が増大するに連れて台区間 $[-1, 1]$ の両端近辺で激しく振動する。この性質のため，入力次元 m が高い場合には $\mathscr{R}_\psi f(a, b)$ は数値的に不安的である。

7.3.3 高次元入力への対応

入力次元が低い場合，$\mu(a, b)$ からのサンプリングは棄却法を用いて遂行できる。一方，入力次元が高い場合は微分の高階化に伴って $\mu(a, b)$ の計算が数値的に不安定になるだけでなく，サンプリングの効率も低下するため，効率を改善するための措置が必要である。

この問題に対処するため，$\mu(a, b)$ を以下のように上から評価したものを用いる

$$\begin{aligned} \mu(a, b) &\approx \frac{1}{Z}\Big|\sum_{s=1}^{S} y_s \psi(a \cdot x_s - b)\Big|, \\ &\leqslant \frac{1}{Z}\sum_{s=1}^{S} |y_s||\psi(a \cdot x_s - b)|, \\ &\propto \sum_{s=1}^{S} w_s \mu_s(a, b), \end{aligned}$$

ここで $\mu_s(a, b) \propto |\psi(a \cdot x_s - b)|$ は成分分布を表し，$w_s := |y_s| / \sum_{t=1}^{S} |y_t|$ は各成分分布 $\mu_s(a, b)$ の混合比を表す。こうして得られた近似分布は $\mu(a, b)$ を混合分布の形でなましたものとみなすことができる。これを**混合近似分布**と呼ぶ。

さらに，個々の成分分布 $\mu_s(a, b)$ はベータ分布を用いて近似できる。一般に，微分の階数 k が十分高い場合，$\rho^{(k)}(z)$ の振幅は台区間 $[-1, 1]$ の

両端付近で最大値をとり，原点周辺では相対的に小さな値をとる。従って，$\rho^{(k)}(z)$ は閉区間上の確率分布であるベータ分布を用いて近似できる ($\mathrm{Beta}(z; 100, 3)$ など)。ベータ分布近似を行うことで，簡便かつ高速なサンプリングが期待できる。

混合近似分布 $\sum_{s=1}^{S} w_s \mu_s(a, b)$ からのサンプリングは，段階的に行う。まず混合比 w_s に従って一つの成分分布 $\mu_s(a, b)$ を選択する。続いて，選ばれた成分分布 $\mu_s(a, b)$ から (a, b) をサンプリングする。

$\mu_s(a, b)$ からのサンプリングは，まず $z \sim \mathrm{Beta}(z; \alpha, \beta)$ を生成し，次に制約式 $z = a \cdot x_s - b$ を満たすように (a, b) をサンプリングする。ここで，与えられた z に対して $z = a \cdot x_s - b$ となる (a, b) は無数に存在する。これは $\mu(a, b)$ を混合近似したために発生した偽の自由度である。提案手法では，新たに二つの制約条件を設けた。

1. a は x_s に平行
2. ノルム $|a|$ は二つの入力ベクトルどうしの距離の逆数 $1/|x_s - x_t|$ と同程度のスケールでばらつく。

まず， 入力ベクトル x_s に対して a は $a \cdot x_s$ の形式で現れるため，x_s に平行な成分以外は 0 と仮定した。また，$1/|a|$ は入力空間において当該神経細胞が選択的に反応する領域の広さを規定する。すなわち，$1/|a|$ が小さすぎる場合には，中間層素子 $\eta(a \cdot x - b)$ はある一つの入力ベクトル x_s にしか反応しなくなる。このような孤立化を防ぐため， $1/|a|$ は少なくとも二つの入力ベクトルをカバーする程度まで大きくとる必要がある。そこで本研究では， ランダムに選択された二つの入力ベクトル間の距離 $|x_s - x_t|$ を計算し， これを $1/|a|$ とした。このように a を定めた後， b は a および予めサンプリングされていた z から $b = a \cdot x_s - z$ によって計算する。

ベータ分布の形状パラメータ α, β として， 一回のサンプリングプロセスは Algorithm 1 のようにまとめられる。各ステップは入力次元 m および必要なサンプル数 s に対して線形にスケールする。さらに， 混合近似のために， 訓練データ集合のサイズに依存しないアルゴリズムになっている。

Algorithm 1 混合近似 $\sum_{s=1}^{S} w_s \mu_s(a, b)$ からのサンプリング

データ番号 s, t を混合比 w_s に従って抽出
$\zeta \sim \mathrm{Beta}(\zeta; \alpha, \beta)$ と $\gamma \sim \mathrm{Bernoulli}(\gamma; p = 0.5)$ を生成
$z \leftarrow (-1)^{\gamma}\zeta$
$1/|a| \leftarrow |x_s - x_t|$
$a \leftarrow |a| x_s / |x_s|$
$b \leftarrow a \cdot x_s - z$
return (a, b)

7.4 実験

人工データおよび実データに対し，三つの初期化法 (表 7.1) の性能を比較した。

表 7.1: 実験に用いる初期化法

Method	Hidden (a, b)	Output c	BP training
SR	Oracle Sampling	Linear Regression	(w/ for MNIST)
SBP	Oracle Sampling	Random Sampling	w/ BP
BP	Random Sampling	Random Sampling	w/ BP

7.4.1 人工データを用いた回帰問題

まず一次元のフィッティング問題を取り上げた。目的関数として位相幾何学者の正弦曲線（Topologist's Sine Curve; TSC）$f(x) = \sin \frac{2\pi}{x}, (f(0) = 0)$ を用いた。TSC は原点に近づくに連れて振動数を増す曲線であり，フィッティングが難しい。訓練データは区間 $[-1, 1]$ から等間隔に抽出された 201 点の関数値を用いた。中間層素子は，それぞれの場合において 100 個（シグモイド対としては 50 個）とし，出力関数は線形出力とした。BP および SBP では $\mathcal{N}(0, 1)$ に従って初期値を与え，BFGS によるバッチ学習を行った。

図 7.1 に訓練誤差の推移，図 7.2 にフィッティングの結果を示す。SR はバックプロパゲーションを行わないため，図 7.1 では定数としてプロットしてある。SR はバックプロパゲーションを経ずに最高精度を達成し，

周波数の変化に追従できている。SBP はフィッティング結果にノイズが入るものの，概ね概形を捉えられている。一方，BP は周波数の変化に追従できず，誤差が高止まりしている。

図 7.3 にオラクル分布 $\mu(a, b)$ に従ってサンプリングされたパラメータのプロットを示す。この図が示すように，一般に $\mu(a, b)$ は (a, b) 座標系では歪んだ形状をしている。予備実験では，$\mu(a, b)$ の歪みを緩和するような座標変換を施すことで，サンプリング効率を向上できることが分かっている。

この実験では，複雑な曲線をフィッティングさせるタスクを取り上げ，提案手法の性能を調べた。実験結果は，オラクル分布が正規分布よりも有利な初期値を与えられることを示している。特に SR では，バックプロパゲーションを経由せずに良いフィッティング結果を得ることができ，サンプリング学習が独立した学習法ともなりうることを示唆している。

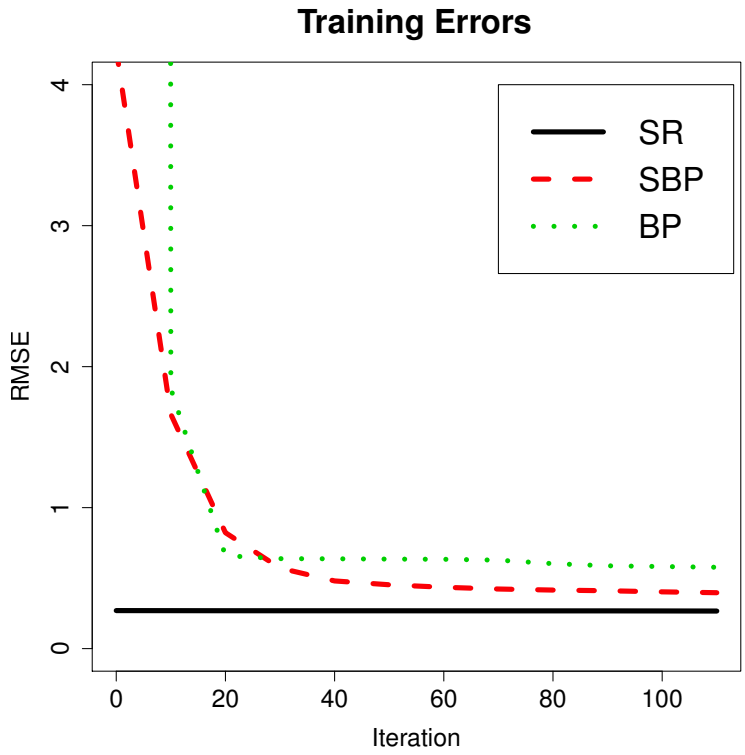


図 7.1: TSC フィッティングに対する訓練誤差。SR はバックプロパゲーションを経ずに最も良い精度を達成している。

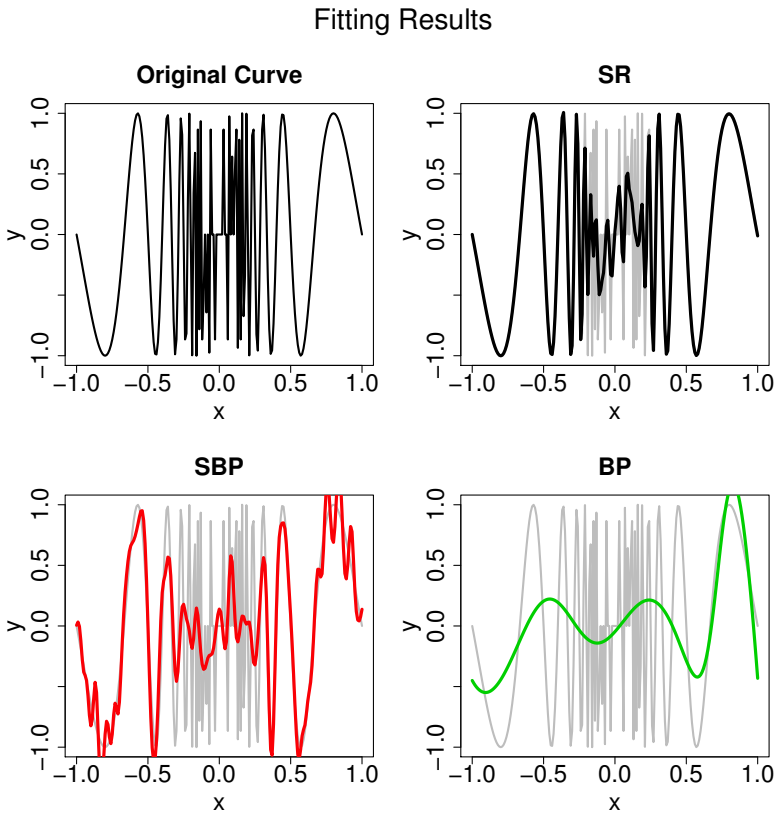


図 7.2: フィッティングの結果。オリジナルの TSC(左上) $\sin\frac{2\pi}{x}$ は原点に近づくに連れて周波数を増す特徴をもつ。

7.4.2 実データによるクラス判別

次に手書き文字データセット MNIST (LeCun and Cortes, 1998) を用いて高次元の実データに対する性能を評価した。MNIST は 0 から 9 までの 10 個の数字のいずれかを手書きした 28×28 ピクセルのグレースケール画像データであり， $60,000$ 点の訓練データと $10,000$ 点のテストデータをもつ。各ラベルは 1 と 0 を等確率に並べた 10 次元ランダムバイナリベクトルとして表現し，ネットワークの出力は 10 次元とした。

中間層素子は，LeCun et al. (1998) が用いた構成と同じ 300 個（シグモイド対としては 150 個）とした。LeCun の報告では誤答率 4.7% が記録されている。出力はシグモイド関数とし， クロスエントロピーを損失関数としてバックプロパゲーション学習を行った。BP および SBP のためのランダム初期化パラメータは， 正規分布 $\mathcal{N}(0, 1/28)$ に従って生成した。

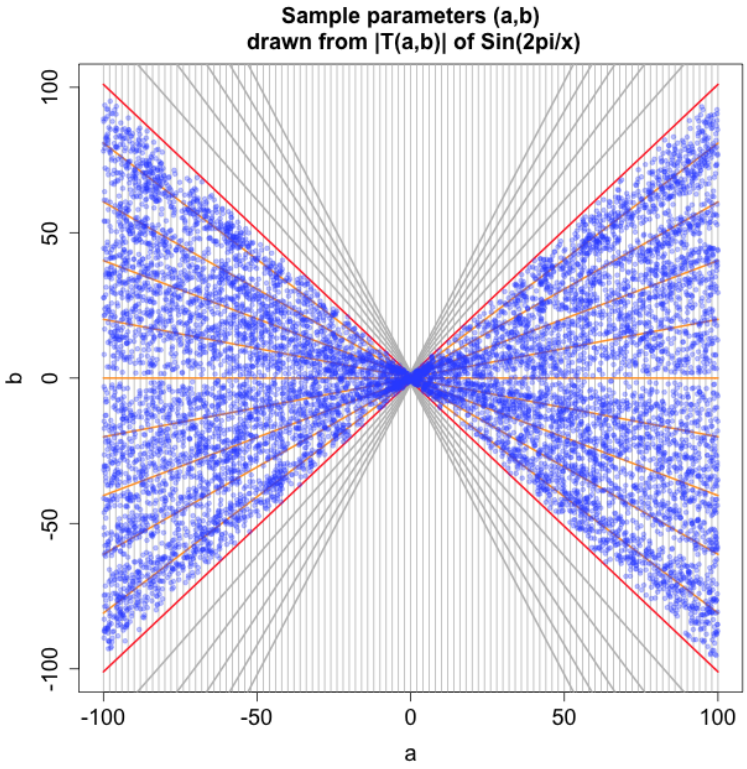


図 7.3: TSC に対するオラクル分布 $\mu(a, b)$ から生成されたサンプル

オラクル分布からのサンプリングには混合近似のテクニックを用いた。バックプロパゲーションはヘッシアンの対角近似を用いた確率的勾配降下法（Stochastic Gradient Descent; SGD）によって行った (LeCun et al., 2012)。実験は 50GB メモリ，2.8GHz Xeon X5660 プロセッサを搭載した計算機上で行い，実装には R を用いた。

図 7.4 は，テストデータに対する判別誤差率の推移を示している。ただし SR 単体では十分な精度を達成することができなかったため (23.0%)，SR に対しても全パラメータのバックプロパゲーション学習を行った。SR の収束値は 9.94% であった。SR の誤差率は単調減少ではなく， これは SR がオーバーフィットしている可能性を示唆している。SBP は三つの中で収束が最も速く， 最も高い精度（8.30%）に収束した。BP の収束値は 8.77% であった。

表 7.2 は， 学習時間の内訳を示している。オラクル分布からのサンプリング時間（約 0.01 秒）は正規分布からのサンプリング時間と同程度に

なることは注目に値する。これは混合近似によって計算が簡単化されたことによる。一方， SR の回帰ステップにはより多くの時間（2.60 秒）が必要とされている。

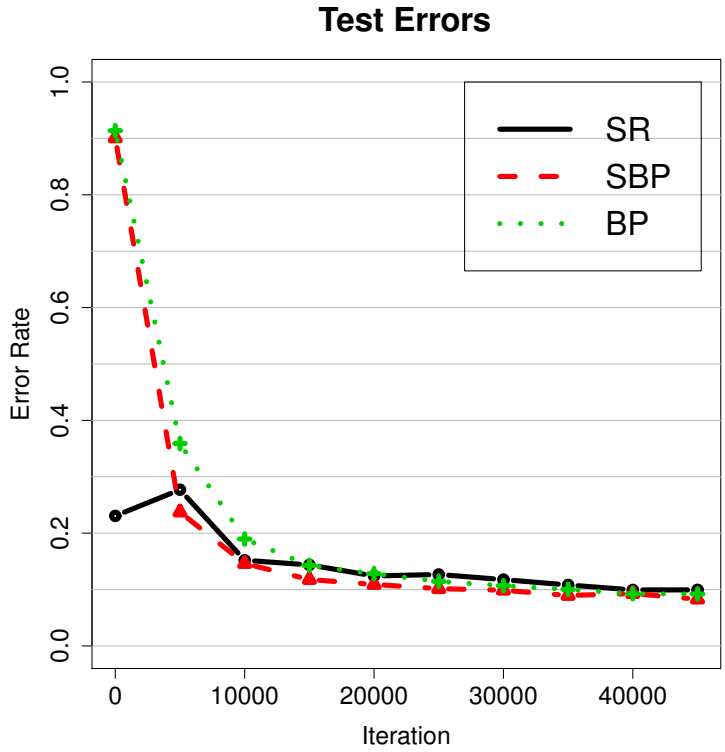


図 7.4: テストデータ ($n = 10,000$) に対する判別誤差率

この実験では， 提案手法が高次元の実データに対しても適用できることを示した。SR はバックプロパゲーションの初期で一旦誤差率が上昇しているので，やや過学習する傾向がある。一方，SBP は BP よりも速く誤差率を低下させることができた。混合近似によってサンプリング時間とサンプリング効率を改善するだけでなく，オラクル分布から有効な情報を取り出せることが示された。

7.5 まとめ

ニューラルネットのサンプリング学習法を新たに提案した。積分表現理論 (Murata, 1996; Sonoda and Murata, 2015) に基づき，データからオ

表 7.2: MNIST に対する学習時間

Method	Sampling [s]	Regression [s]	BP training [s]
SR	1.15×10^{-2}	2.60	2.00×10^{3}
SBP	1.14×10^{-2}	-	2.31×10^{3}
BP	1.15×10^{-2}	-	2.67×10^{3}

ラクル分布を計算する方法を開発した。高次元入力の問題においてサンプリング効率が低下する問題を解決するため，混合近似分布を開発した。

これまで，ニューラルネットの積分表現は理論を中心に発展してきたが，学習のためのアルゴリズムとして実装する研究はほとんど行われてこなかった。オラクル分布から実際にサンプリングするアルゴリズムを構築した点で本研究は新規性が高い。

数値実験では人工データ TSC と実データ MNIST を用いて三つの方法 (SR, SBP, BP) を比較した。TSC に対する回帰タスクでは，SBP はオーバーフィットの可能性を示す一方，SR はバックプロパゲーションを経ずに最も良い精度を達成することができた。一方，MNIST に対する判別タスクでは， SR は局所解に陥る傾向を示す一方， SBP は BP よりも高速に収束することができた。これらの結果は， バックプロパゲーションの初期値としてオラクル分布を利用する場合，出力層パラメータはフィッティングするよりも乱数初期化するべきであることを示唆している。

サンプリング学習では， 本来必要な数よりも冗長なパラメータをサンプリングする必要がある。したがってニューラルネットの中間層素子を枝刈りする手法と組み合わせることでより実用的な学習法になることが期待される。

第8章　結論

深層ニューラルネットの中では何が起きているのだろうか。また，なぜ深層にした方が良いのだろうか。本研究では，深層ニューラルネットの積分表現理論の開発を通じて，これらの問題解決に取り組んだ。

ニューラルネットを積分表現にすることで，ニューラルネットの幾何学的性質や解析的性質が調べられる（第4章)。まず，積分表現は双対リッジレット変換である。リッジレット変換はRadon変換とウェーブレット変換の合成変換なので，ニューラルネットは，その逆変換として理解できる。また，積分表現の離散化に伴う近似誤差は，Maurey-Jones-Barron評価やJackson型の評価として詳細に調べられている。さらに，リッジレット変換からオラクル分布を計算することで，バックプロパゲーションによらない学習もできる（第7章)。

これまで，深層ニューラルネットの積分表現理論はほとんど調べられてこなかった。中間層が二層以上ある場合には，積分核が入れ子になるためである。本研究では，深層ニューラルネットの特徴量写像を輸送写像とみなす方針で，深層ニューラルネットの積分表現を開発した。

本研究着手時点で，課題は大きく二つあった。まず，本研究が拠り所とする積分表現理論（リッジレット解析）は，深層学習で標準的に用いられるReLUのような，非有界な活性化関数を想定していなかった。そこで，超関数によるリッジレット変換の理論を構築し，深層ニューラルネットの積分表現理論を展開するための基礎を築いた（第5章)。

次に，深層ニューラルネットを基底と係数に分解する方法が不明であった。浅いニューラルネットの場合は，中間層$h(x)$を基底関数，出力層$k(z)$を係数に対応付けることで自然に積分表現が現れた。一方，深層ニューラルネットの場合は，特徴量写像$h(x)$は複数の中間層$h^\ell\,(\ell=1,\cdots,L)$の合成写像

$$h(x)=h^L\circ\cdots\circ h^1(x),$$

であり，浅い構造と同じように基底$h(x)$と係数$k(z)$に分解する方法では，個別の中間層h^ℓの振舞いを調べることができない。このような入れ

子構造の問題があるため，“深層ニューラルネットの積分表現” なるものは，これまで提案されてこなかった。

本研究では，基底 h を輸送写像 Φ_t に置き換える方針を検討した

$$g(x) = \Psi \circ \Phi_{t=T}(x).$$

ただし Ψ は出力層に相当する線形写像とし，$t = 0$ のとき輸送写像は恒等写像 $\Phi_0 = \mathsf{id}$ とする。つまり，各中間層 h^ℓ を入力点 x の空間に作用する輸送写像（力学系）とみなす方針である。輸送写像は何度合成しても輸送写像であるから，特徴量写像全体の輸送の性質が分かれば，個別の中間層は中継ぎの輸送写像として理解できる。あるいは逆に，個別の中間層の輸送の性質が分かれば，特徴量写像はそれらを順に辿る輸送写像として理解できる。特に，輸送経路を無限に細かく分割したり，輸送写像を際限なく合成することで，無限層ニューラルネットが考えられるようになる。

第 6 章では，特にデノイジング・オートエンコーダー（DAE）が輸送写像とみなせることに着目して，DAE の性質を解析した。まず，DAE はエントロピーを減らすように入力データ分布を変形する輸送写像であることが分かった。さらに，この作用は浅い DAE よりも深層 DAE で顕著になることも分かった。つまり，深層 DAE は浅い DAE と本質的に異なる挙動をすることが明らかになった。これはニューラルネットを積極的に深くする動機付けともなる。そして，深層ニューラルネットの積分表現は，個別の輸送写像の積分表現として構成した。このような輸送解釈の方法は，データの座標系や，パラメータの取り方に依存しないノンパラメトリックな解析手法であり，解釈性が高い。また，後述する通り，DAE に限らず多くの深層ニューラルネットの解析に応用できると考えられる。入れ子の問題を輸送現象によって解決するアイデアは，水と油が自発的に分離する現象に着想を得た。

深層ニューラルネットの積分表現

図 8.1 上段は，中間層（青）を 8 層持つ深層ニューラルネットである。例えば，積層 DAE によって事前学習した後，出力層（赤）を付けてファインチューニングしたものは，この構造になる。図 8.1 中段は，入力層（緑）から入力層への写像を 8 回合成した深層ニューラルネットである。例えば，合成 DAE はこの構造をもつ。上段のように中間層どうしが合成

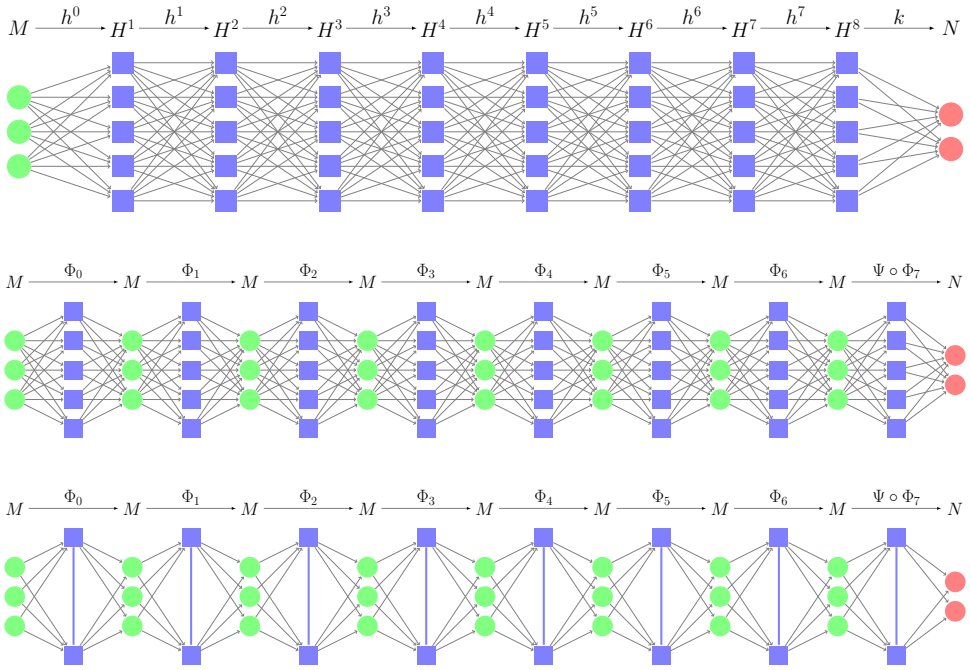


図 8.1: 要素毎に積分表現にして深層ニューラルネットの積分表現を得る。

された構造では，積分表現にする術がない。一方，中段のように合成写像の構造であれば，要素毎に積分表現にして，図 8.1 下段に示すような深層ニューラルネットの積分表現が得られる。

輸送解釈によって，積分表現の入れ子の問題はどのように解決されたのだろうか。輸送写像はそれ自体，積分表現にできる

$$\Phi_t(x) = \int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi \Phi_t(a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b, \quad t \in [0,T]$$

これを用いて，深層ニューラルネットの積分表現は以下で与えられる

$$\begin{aligned} g(x) &= \Psi \circ \Phi_{t=T}(x) \\ &= \int_{\mathbb{Y}^{m+1}} (\Psi \circ \mathscr{R}_\psi \Phi_{t=T})(a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b. \end{aligned}$$

輸送写像の最大の特徴は，半群性 $\Phi_{t+s} = \Phi_t \circ \Phi_s$ である。これにより，以下のように写像の合成をキャンセルして全ての情報を係数に集約できる

$$\begin{aligned} &\int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi \Phi_{t+s}(a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b \\ &= \int_{\mathbb{Y}^{m+1}} \mathscr{R}_\psi \Phi_t(a,b)\eta\left(\int_{\mathbb{Y}^{m+1}} (a\cdot \mathscr{R}_\psi \Phi_s)(a',b')\eta(a'\cdot x - b')\mathrm{d}a'\mathrm{d}b' - b\right)\mathrm{d}a\mathrm{d}b. \end{aligned}$$

積分表現として見ると，左辺は中間層が一層のニューラルネットを表すのに対し，右辺は中間層が二層のニューラルネットを表している。

Why Deep?

ニューラルネットは浅くても万能関数近似器であるのに，なぜ深層ニューラルネットの方が学習能力が高いのだろうか。連続 DAE をプレトレーニングとする二段階学習は，逆拡散方程式の最終値問題とみなせる

$$\begin{aligned}&\partial_t \Phi_{t\sharp}\pi_0 = -\triangle \Phi_{t\sharp}\pi_0,\\&\Phi_{t=0} = \mathsf{id},\\&\Psi \circ \Phi_{t=T} = f.\end{aligned}$$

一方，素朴なバックプロパゲーション学習は次の最適化問題と同値である

$$\text{minimize} \quad \mathbb{E}|\Psi \circ \Phi_{t=T}(x) - f(x)|^2 \quad \text{w.r.t.} \quad \Phi_{t=T}, \Psi.$$

最終値問題から見れば，これは $t = T$ のみを規定する条件（最終値条件）にすぎない。つまり，途中の輸送経路 Φ_t は自由ということになる。一方，二段階学習では各時刻 t で Φ_t の振る舞いまで規定されるので，探索すべき関数空間は相対的に小さい。つまり，学習問題としては二段階学習の方が簡単である。このように，輸送写像は問題の複雑性を緩和していると考えられる。従来の浅いニューラルネットの理論では，深層ニューラルネットの中間層に相当する写像は単に特徴量写像として扱っていたので，この違いが表現できていなかったのである。

図 8.2 は，学習前のニューラルネットと，学習後のニューラルネットにおいて，中間層の発火パターンを PCA によって可視化したものである。同じクラスの点は同じ色で示されている。学習前には，異なるクラスの点が混ざり合っている様子が分かる。これに対して学習後は，層が上がるに連れて分離が進んでいく様子が見て取れる。つまり，層が深まるに連れて判別問題の難しさが緩和しているのである。

深層ニューラルネット g を，輸送写像 Φ_t と判別器 Ψ_t に分解する。ただし，Ψ_t は単なる線形出力層ではなく，任意の関数を近似できるクラスとする。$g = \Psi_t \circ \Phi_t$ に写像 f を学習させる最適化問題は，次のように変

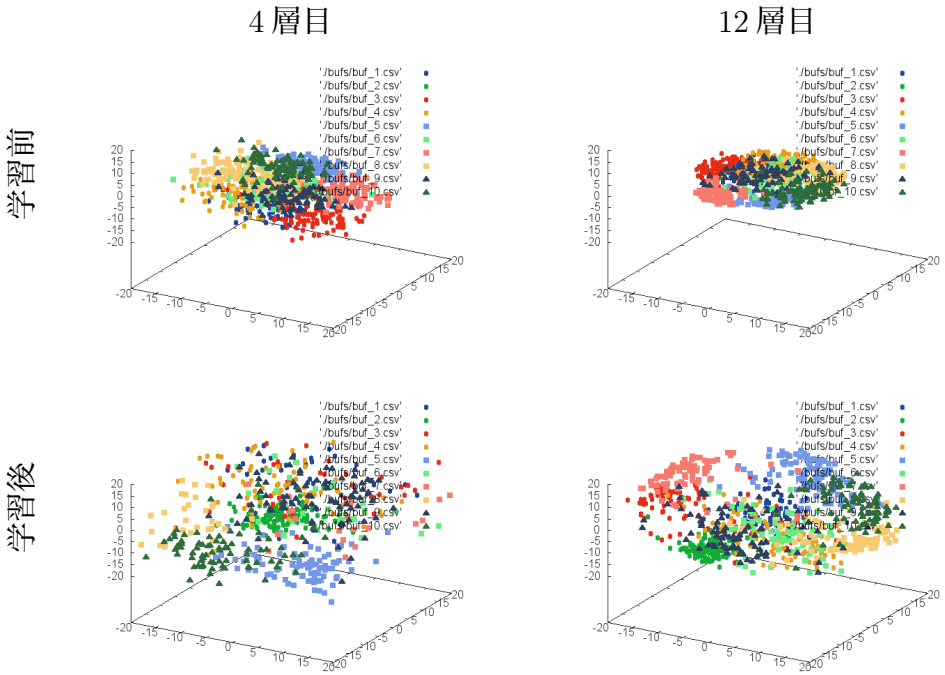


図 8.2: 中間層特徴量の PCA による可視化例。同じクラスは同じ色で表されている。実験及び図の作製は松原拓央氏の協力による。

形できる

$$\begin{aligned}
&\min_g \int_{\mathbb{R}^m} |f(x) - g(x)|^2 \pi_0(x) \mathrm{d}x \\
&= \min_{\Phi_t, \Psi_t} \int_{\mathbb{R}^m} |f(x) - \Psi_t \circ \Phi_t(x)|^2 \pi_0(x) \mathrm{d}x \\
&= \min_{\Phi_t, \Psi_t} \int_{\mathbb{R}^m} |f \circ \Phi_t^{-1}(x) - \Psi_t(x)|^2 \pi_t(x) \mathrm{d}x.
\end{aligned}$$

ただし二番目の変形では $\pi_0 \circ \Phi_t^{-1}(x) |\nabla \Phi_t^{-1}(x)| = \pi_t(x)$ を用いた。ここで，$f \circ \Phi_t^{-1}$ は輸送によって緩和された f を表す。$T = 0$ のときは，$\Phi_t = \mathrm{id}$ なので，輸送を介さず Ψ_t のみで f を近似することになる。このとき Ψ_t の複雑さは，MJB 評価によって $\|\mathscr{R}_\psi f\|_1$ で表される。一方，輸送を介したときは $\Psi_t = f \circ \Phi_t^{-1}$ が成り立つので，Ψ_t が担う複雑さは $\|\mathscr{R}_\psi (f \circ \Phi_t^{-1})\|_1$ になる。輸送写像を適切に選べば，$\|\mathscr{R}_\psi (f \circ \Phi_t^{-1})\|_1 \leqslant \|\mathscr{R}_\psi f\|_1$ となることが期待できる。このように，輸送写像 Φ_t を挟むことで f の複雑さを緩和することが期待される。

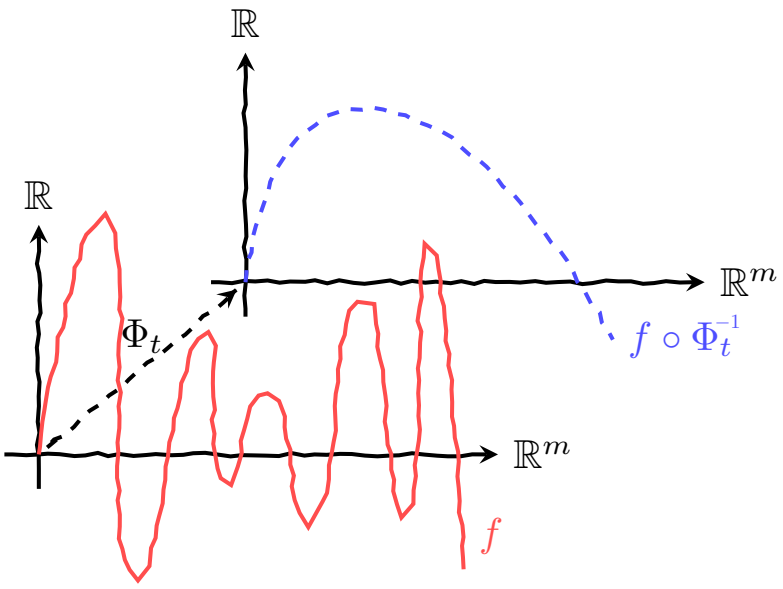


図 8.3: 輸送によって近似対象 $f(x)$ の複雑性が緩和されることの概念図

今後の展望

慣習に従い，学習機械の汎化誤差を近似誤差と推定誤差に分けて考える。近似誤差を評価するのは関数近似理論，推定誤差を評価するのは統計学や学習理論である。積分表現理論は専ら関数近似の理論であり，データの存在は希薄である。しかし今後は，統計的な解析にも取り組んでいく必要がある。深層学習を利用すると，どのような構造にすれば良いのか，どの活性化関数を使えば良いのかといったモデル選択の問題や，どうすれば学習が上手くいくのかといった最適化の問題に直面する。このような問題を解析するためには，データとニューラルネットを対応付ける規則，すなわち学習アルゴリズムを解析する必要がある。また，深層ニューラルネットのパラメータは数十億個にのぼり，データサイズから見てもほとんど無限と思われるほど大量にあるにも関わらず，学習できるのはなぜか。このような問題は一般的な新 NP 問題を解決する糸口とも捉えられるので，今後の重要な課題と言える。

謝辞

学部生の時から一貫して指導して頂いた村田昇教授に心から感謝いたします。また，本論文の審査を快く引き受けて頂いた渡邊亮教授，井上真郷教授，浜田道昭准教授，そして産業技術総合研究所の赤穂昭太郎博士に深く感謝いたします。また，予備審査や公聴会の場で貴重なご意見を賜った内田健康教授，小川哲司准教授，筑波大学の日野英逸准教授に深く感謝いたします。とりわけ，村田研究室の先輩でもある日野英逸准教授には，研究生活の助言だけでなく，日本学術振興会特別研究員の申請書を何度も丁寧かつ親身に指導して頂き，無事採用されることができました。重ねてお礼申し上げます。

本研究遂行にあたり，研究会等で建設的な議論をさせて頂いた名古屋工業大学の松添博教授，本谷秀堅教授，大阪教育大学の芦野隆一教授，大阪電気通信大学の萬代武史教授，日本大学の原一之教授，福岡大学の藤木淳准教授，産業技術総合研究所の兼村厚範博士，そして早稲田大学の矢田部浩平氏，田中一成氏，森岡幹氏，鈴木知彦先輩，東京大学の本武陽一博士，唐木田亮氏，小宮山純平博士，矢野恵佑氏，南賢太郎氏，黒田淳夫氏に感謝いたします。なかでも矢田部浩平氏は私が博士後期課程に復学して最初の友人であり，ともに学振に採用され，氏の結婚式二次会の司会まで務めさせていただきました。ありがとうございます。また共同研究では，貴重なデータの提供だけでなく，分野を跨ぐ有意義な議論を繰り返しさせていただいた東京大学の加藤真平准教授，筑波大学の川崎真弘助教，京都大学の加納学教授，新日鐵住金株式会社の北田宏博士に感謝いたします。

研究生活を共にした村田研究室の諸先輩方と同期，後輩の皆様，特に一緒に研究をさせて頂いた村田研究室の関光宏先輩，金田有紀さん，嶋田達之介君，關翼人君，松原拓央君，中村圭太君，研究会でも度々ご一緒させて頂いている竹内孝先輩には大変お世話になりました。

積分表現理論は「脳の情報表現について数理的に研究したい」という私の曖昧な要望に応えて，村田昇教授から頂いたテーマです。当時学部

３年生だった私は，渡された論文 (Cybenko, 1989; Hornik et al., 1989; White, 1990; Barron, 1993; Murata, 1996) の内容が断片的にしか分からず，ここから何を読み取れば良いのか，ここに何を付け足せば新しい研究になるのか，そもそも自分にそんなことができるのか，皆目検討も付かず，途方に暮れていたことを覚えています。学部４年生のときに，当時修士２年生だった関光宏先輩が積分表現理論をやり始めて，(Sonoda and Murata, 2014) の元になる研究が進み，大変心強く思いました。関さん，どうもありがとうございます。積分表現理論は 90 年代によく研究されていたので，今思い返しても工学系の学部生には無茶なテーマのように思いますが，その後今日まで関数解析を勉強するための強力なベンチマークになりました。地道な勉強とサーベイを進めるうちに，見よう見まねで自分でも定理が証明できたときには，心底嬉しかった記憶があり，自信にも繋がりました。噛めば噛むほど面白い論文を紹介して頂いた村田先生，どうもありがとうございます。

最後に，私の研究生活を支えて頂いた家族，友人，関係者の皆様に感謝いたします。

付 録A　証明等

A.1　半直線上の測度

$\psi : \mathbb{R}_+ \to \mathbb{C}$ に対して，

$$\|\psi\|^p_{L^p_s(\mathbb{R}_+)} := \int_{\mathbb{R}_+} |\psi(r)|^p \frac{\mathrm{d}r}{r^s}, \quad s \in [0, \infty)$$

とおく。$s = 1$ のとき測度は乗法群の Haar 測度になり，このときノルムはスケール不変である。

Haar 測度を用いる場合

Haar 測度に限り，以下のスケール不変性が成り立つ。

$$\int_{\mathbb{R}_+} \left|\psi\left(ar\right)\right|^p \frac{\mathrm{d}r}{r} = \int_{\mathbb{R}_+} \left|\psi\left(r\right)\right|^p \frac{\mathrm{d}r}{r}, \quad a > 0.$$

Proof. 単に $ar = r'$ と変換して $\mathrm{d}r/r = \mathrm{d}r'/r'$ を用いることもできるが，$r = \exp\rho$ と変換して $\mathrm{d}r/r = \mathrm{d}\rho$ を用いるほうが群論的である。

$$\|\psi\|^p_{L^p(\mathbb{R}_+)} = \int_{\mathbb{R}_+} |\psi(r)|^p \frac{\mathrm{d}r}{r} = \int_{\mathbb{R}} |\psi(\exp\rho)|^p \mathrm{d}\rho = \|\psi \circ \exp\|^p_{L^p(\mathbb{R})}.$$

ここから，$f \in L^1(\mathbb{R}^m) \Rightarrow f \circ \log \in L^1(\mathbb{R}_+)$ が分かる。この変数変換によってスケーリング ar はシフト $\log a + \log r$ に写る。これは，指数関数が加法群と乗法群の群同型であることを思い出せば当然である。従って，Lebesgue 積分のシフト不変性により，ノルムはスケール不変であることが分かる。 □

群準同型を用いると，スケール畳み込み (scale convolution) の可換性も分かる

$$\psi *_s \phi(s) := \int_{\mathbb{R}_+} \psi\left(\frac{s}{r}\right) \phi(r) \frac{\mathrm{d}r}{r} = \phi *_s \psi.$$

また，Young の不等式が示せる

$$\begin{aligned}\|\psi *_s \phi\|_{L^r(\mathbb{R}_+)} &= \|\psi \circ \exp * \phi \circ \exp \|_{L^r(\mathbb{R})} \\ &\leqslant \|\psi \circ \exp \|_{L^p(\mathbb{R})} \|\phi \circ \exp \|_{L^q(\mathbb{R})} = \|\psi\|_{L^p(\mathbb{R}_+)} \|\phi\|_{L^q(\mathbb{R}_+)}.\end{aligned}$$

Haar 測度でない場合

スケール不変性は成り立たなくなる

$$\int_{\mathbb{R}_+} |\psi(ar)|^p \frac{\mathrm{d}r}{r^s} = a^{s-1} \int_{\mathbb{R}_+} |\psi(r)|^p \frac{\mathrm{d}r}{r^s}.$$

また，スケール畳み込みは非可換であり，

$$\int_{\mathbb{R}_+} \psi\left(\frac{u}{r}\right) \phi(r) \frac{\mathrm{d}r}{r^s} = u^{1-s} \int_{\mathbb{R}_+} \psi(r) \phi\left(\frac{u}{r}\right) \frac{\mathrm{d}r}{r^{2-s}}.$$

ノルムも複雑になる。

$$\int_{\mathbb{R}_+} \left| \int_{\mathbb{R}_+} \psi\left(\frac{u}{r}\right) \phi(r) \frac{\mathrm{d}r}{r^s} \right|^p \frac{\mathrm{d}u}{u^s} = \int_{\mathbb{R}} \left| \int_{\mathbb{R}} \psi \circ \exp(x-y) \phi \circ \exp(x) e^{(1-s)x} \mathrm{d}x \right|^p e^{(1-s)y} \mathrm{d}y.$$

A.2 定理 5.2.1 の証明

定義により，リッジレット変換 $\mathscr{R}_\psi f(u, \alpha, \beta)$ は Radon 変換 $\mathrm{R}f(u, p)$ と，スケーリングされた超関数 $\psi_\alpha(\beta)$ との畳み込みである

$$f(x) \mapsto \mathrm{R}f(u, p) \mapsto \left(\mathrm{R}f(u, \cdot) * \widetilde{\psi_\alpha}\right)(\beta) = \mathscr{R}_\psi f(u, \alpha, \beta).$$

以下では，この分解に沿って段階的に証明を行う。

Step 1 まず $\mathrm{R}f(u, p)$ のクラス $\mathcal{X}(\mathbb{S}^{m-1} \times \mathbb{R})$ を調べる（表 5.2 の 2 列目）。Hertle (1983, Th 4.6, Cor 4.8) により，$\mathcal{X} = \mathcal{D}, \mathcal{E}', \mathcal{S}, \mathcal{O}'_C, L^1$ に対して Radon 変換は連続単射であることが分かっている

$$\mathrm{R} : \mathcal{X}(\mathbb{R}^m) \longhookrightarrow \mathcal{X}(\mathbb{S}^{m-1} \times \mathbb{R})$$

従って $\mathcal{X}$ の選び方はこの中に制限される。

Trèves (1967, § 51) により $\mathcal{X} = \mathcal{D}, \mathcal{E}', \mathcal{S}, \mathcal{O}'_C, L^1$ はそれぞれ核型なので，次の核型定理が成り立つ

$$\mathcal{X}(\mathbb{S}^{m-1} \times \mathbb{R}) \cong \mathcal{X}(\mathbb{S}^{m-1}) \hat{\otimes} \mathcal{X}(\mathbb{R}).$$

従って，$(\alpha,\beta)\in\mathbb{H}$ と $u\in\mathbb{S}^{m-1}$ のクラスは独立に考えて良い。
Step 2 次に $\mathscr{R}_\psi f(u,\alpha,\beta)$ において，$u\in\mathbb{S}^{m-1}$ と $\alpha>0$ を固定して β のみを変数とした場合のクラス $\mathcal{B}(\mathbb{R})$ を調べる（表 5.2 の 3, 4 列目）。Schwartz の結果 表 3.1 により，各 $\phi\in\mathcal{X}(\mathbb{R})$ に対して超関数の意味での畳み込み積分 $\phi*\psi$ が定義できる $\psi\in\mathcal{Z}(\mathbb{R})$ の最大のクラスの組み合わせと，畳み込みの結果得られる関数 $\phi*\psi$ のクラス $\mathcal{B}(\mathbb{R})$ は，表 5.2 のようになる。ただし $\phi*\psi$ の正則性を担保するために，$\mathcal{X}=L^1$ に対しては $\mathcal{Z}=L^p\cap C$ とした。明らかに各 $\mathcal{Z}=\mathcal{D}',\mathcal{S}',L^p\cap C$ に対して，$\psi\in\mathcal{Z}(\mathbb{R})$ ならば $\psi_\alpha\in\mathcal{Z}(\mathbb{R})$ である。
Step 3 $\mathscr{R}_\psi f(u,\alpha,\beta)$ において，$u\in\mathbb{S}^{m-1}$ を固定して (α,β) を変数とした場合のクラス $\mathcal{A}(\mathbb{H})$ を調べる（表 5.2 の 5 列目）。以下では $\phi(p):=\mathrm{R}f(u,p)$ とおいて，

$$\mathcal{W}[\psi;\phi](\alpha,\beta):=\int_{\mathbb{R}}\phi(\alpha z+\beta)\overline{\psi(z)}\mathrm{d}z=\phi*\widetilde{\overline{\psi_\alpha}}(\beta),$$

と書くことにする。つまり $\mathscr{R}_\psi f(u,\alpha,\beta)=\mathcal{W}[\psi;\phi](\alpha,\beta)$ である。核型定理により $f\in\mathcal{X}(\mathbb{R}^m)$ のとき $\phi\in\mathcal{X}(\mathbb{R})$ である。
Case 3a ($\mathcal{X}=\mathcal{D}$ かつ $\mathcal{Z}=\mathcal{D}'$ のとき $\mathcal{A}=\mathcal{E}$) 直接計算により，各 $k,\ell\in\mathbb{N}_0$ に対して以下が成り立つことが分かる

$$\partial_\alpha^k\partial_\beta^\ell\mathcal{W}[\psi;\phi](\alpha,\beta)=\mathcal{W}[z^k\cdot\psi;\phi^{(k+\ell)}](\alpha,\beta).$$

$\phi\in\mathcal{D}(\mathbb{R})$ のとき $\phi^{(k+\ell)}\in\mathcal{D}(\mathbb{R})$ かつ，$\psi\in\mathcal{D}'(\mathbb{R})$ のとき $z^k\cdot\psi\in\mathcal{D}'(\mathbb{R})$ なので，$\partial_\alpha^k\partial_\beta^\ell\mathcal{W}[\psi;\phi](\alpha,\beta)$ は各点 $(\alpha,\beta)\in\mathbb{H}$ で存在する。従って，$\phi\in\mathcal{D}(\mathbb{R})$ かつ $\psi\in\mathcal{D}'(\mathbb{R})$ ならば $\mathcal{W}[\psi;\phi]\in\mathcal{E}(\mathbb{H})$ である。
Case 3b（$\mathcal{X}=\mathcal{E}'$ かつ $\mathcal{Z}=\mathcal{D}'$ のとき $\mathcal{A}=\mathcal{D}'$）コンパクト集合 $\mathrm{K}\subset\mathbb{H}$ を任意にとって固定する。ある $N\in\mathbb{N}_0$ が存在して以下が成り立つことを示す

$$\left|\int_{\mathrm{K}}\mathrm{T}(\alpha,\beta)\mathcal{W}[\psi;\phi](\alpha,\beta)\frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha}\right|\lesssim\sum_{k,\ell\leqslant N}\sup_{(\alpha,\beta)\in\mathbb{H}}|\partial_\alpha^k\partial_\beta^\ell\mathrm{T}(\alpha,\beta)|,\quad\forall\mathrm{T}\in\mathcal{D}(\mathrm{K}).$$

K に台をもつ滑らかな関数 $\mathrm{T}\in\mathcal{D}(\mathrm{K})$ を任意にとる。二つのコンパクト集合 $\mathrm{A}\subset\mathbb{R}_+$ と $\mathrm{B}\subset\mathbb{R}$ を $\mathrm{K}\subset\mathrm{A}\times\mathrm{B}$ となるようにとる。$k,\ell\in\mathbb{N}_0$ を以下

を満たすようにとって固定する。

$$\left|\int_{\mathbb{R}} u(z)\phi(z)\mathrm{d}z\right| \lesssim \sup_{z\in\mathsf{supp}\,\phi} |u^{(k)}(z)|, \quad \forall u \in \mathcal{E}(\mathbb{R}) \tag{A.1}$$

$$\left|\int_{\mathbb{R}} v(z)\overline{\psi(z)}\mathrm{d}z\right| \lesssim \sup_{z\in\mathbb{R}} |v^{(\ell)}(z)|, \quad \forall v \in \mathcal{D}(B). \tag{A.2}$$

各 $\alpha > 0$ に対して $\mathrm{T}(\alpha,\cdot) * \widetilde{\phi} \in \mathcal{D}'(\mathbb{R})$ が成り立つ。従って，(A.1) と (A.2) を逐次適用して，以下が成り立つ

$$\begin{aligned}
&\left|\int_{\mathbb{R}} \mathrm{T}(\alpha,\beta)\int_{\mathbb{R}} \phi(\alpha z+\beta)\overline{\psi(z)}\mathrm{d}z\frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha}\right| \\
&\quad\leqslant \int_0^\infty \left|\int_{\mathbb{R}}\int_{\mathbb{R}} \mathrm{T}(\alpha,\beta-\alpha z)\overline{\psi(z)}\mathrm{d}z\cdot\phi(\beta)\mathrm{d}\beta\right|\frac{\mathrm{d}\alpha}{\alpha} \\
&\quad\lesssim \int_0^\infty \sup_{\beta\in\mathsf{supp}\,\phi}\left|\int_{\mathbb{R}} \partial_\beta^k \mathrm{T}(\alpha,\beta-\alpha z)\psi(z)\mathrm{d}z\right|\frac{\mathrm{d}\alpha}{\alpha} \\
&\quad\lesssim \int_0^\infty \sup_{\beta\in\mathsf{supp}\,\phi}\sup_z\left|\partial_\beta^{k+\ell}\mathrm{T}(\alpha,\beta-\alpha z)\right|\alpha^{\ell-1}\mathrm{d}\alpha \\
&\quad= \int_{\mathrm{A}} \sup_{\beta\in\mathrm{B}}\left|\partial_\beta^{k+\ell}\mathrm{T}(\alpha,\beta)\right|\alpha^{\ell-1}\mathrm{d}\alpha \\
&\quad\leqslant \sup_{(\alpha,\beta)\in\mathrm{K}}\left|\partial_\beta^{k+\ell}\mathrm{T}(\alpha,\beta)\right|\cdot\int_{\mathrm{A}}\alpha^{\ell-1}\mathrm{d}\alpha.
\end{aligned}$$

つまり，$\mathcal{W}[\psi;\phi] \in \mathcal{D}'(\mathbb{H})$ である。ただし三番目の式は $\partial_z[\mathrm{T}(\alpha,\beta-\alpha z)] = (-\alpha)\partial_\beta \mathrm{T}(\alpha,\beta-\alpha z)$ を繰り返し適用して従う。四番目の式は T の台のコンパクト性から従う。従って，$N = k+\ell$ にとれば，T の任意性より $\mathcal{W}[\psi;\phi] \in \mathcal{D}'(\mathbb{H})$ が結論される。

Case 3c（$\mathcal{X} = \mathcal{S}$ かつ $\mathcal{Z} = \mathcal{S}'$ のとき $\mathcal{A} = \mathcal{O}_{\mathcal{M}}$）まず，任意の $k,\ell \in \mathbb{N}_0$ に対して $\phi^{(k+\ell)} \in \mathcal{S}(\mathbb{R})$ かつ $z^k\cdot\psi \in \mathcal{S}'(\mathbb{R})$ である。従って Case 3a と同様の議論によって $\mathcal{W}[\psi;\phi] \in \mathcal{E}(\mathbb{H})$ である。以下では任意の $k,\ell\in\mathbb{N}_0$ に対してある $s,t\in\mathbb{N}_0$ が存在して以下が成り立つことを示す

$$\left|\partial_\alpha^k\partial_\beta^\ell \mathcal{W}[\psi;\phi](\alpha,\beta)\right| \lesssim (\alpha+1/\alpha)^s(1+\beta^2)^{t/2}.$$

$\partial_\alpha^k\partial_\beta^\ell\mathcal{W}[\psi;\phi](\alpha,\beta) = \partial_\alpha^0\partial_\beta^0\mathcal{W}[\phi^{(k+\ell)};z^k\cdot\psi](\alpha,\beta)$ かつ，$\phi^{(k+\ell)}\in\mathcal{S}(\mathbb{R})$，$z^k\cdot\psi\in\mathcal{S}'(\mathbb{R})$ なので，$k=\ell=0$ の場合のみを示せば十分である。$\psi\in\mathcal{S}'(\mathbb{R})$

より，ある $N \in \mathbb{N}_0$ をとって次のようにできる

$$\left| \int_{\mathbb{R}} u(z)\overline{\psi(z)}\mathrm{d}z \right| \lesssim \sum_{s,t\leqslant N} \sup_{z\in\mathbb{R}} |z^s u^{(t)}(z)|, \quad \forall u \in \mathcal{S}(\mathbb{R}).$$

$u(z) \leftarrow \phi(\alpha z + \beta)$ と代入して，

$$\begin{aligned}
\left| \int_{\mathbb{R}} \phi(\alpha z+\beta)\overline{\psi(z)}\mathrm{d}z \right| &\lesssim \sum_{s,t\leqslant N} \sup_{z\in\mathbb{R}} |z^s \partial_z^t \phi(\alpha z+\beta)| \\
&= \sum_{s,t\leqslant N} \sup_{p\in\mathbb{R}} \left| \left(\frac{p-\beta}{\alpha}\right)^s \alpha^t \phi^{(t)}(p) \right| \\
&\lesssim \sum_{s,t\leqslant N} \alpha^{t-s}\beta^s \sup_{p\in\mathbb{R}} |p^s \phi^{(t)}(p)| \\
&\lesssim (\alpha + 1/\alpha)^N (1+\beta^2)^{N/2}.
\end{aligned}$$

つまり，$\mathcal{W}[\psi;\phi] \in \mathcal{O}_{\mathcal{M}}(\mathbb{H})$ である。ただし二番目の式は $p \leftarrow \alpha z + \beta$ とした。四番目の式は仮定 $\phi \in \mathcal{S}(\mathbb{R})$ から常に $\sup_p |p^s \phi^t(p)|$ が有限となることから従う。

Case 3d（$\mathcal{X} = \mathcal{O}'_{\mathcal{C}}$ かつ $\mathcal{Z} = \mathcal{S}'$ のとき $\mathcal{A} = \mathcal{S}'$）ある $N \in \mathbb{N}_0$ があって以下が成り立つことを示す

$$\left| \int_{\mathbb{H}} \mathrm{T}(\alpha,\beta)\mathcal{W}[\psi;\phi](\alpha,\beta)\frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha} \right| \lesssim \sum_{s,t,k,\ell\leqslant N} \sup_{\alpha,\beta\in\mathbb{H}} \left|\mathrm{D}_{s,t}^{k,\ell}\mathrm{T}(\alpha,\beta)\right|, \quad \forall \mathrm{T} \in \mathcal{S}(\mathbb{H}) \tag{A.3}$$

ただし

$$\mathrm{D}_{s,t}^{k,\ell}\mathrm{T}(\alpha,\beta) := (\alpha+1/\alpha)^s (1+\beta^2)^{t/2} \partial_\alpha^k \partial_\beta^\ell \mathrm{T}(\alpha,\beta).$$

とおいた。$\mathrm{T} \in \mathcal{S}(\mathbb{H})$ を任意にとって固定する。$\psi \in \mathcal{S}'(\mathbb{R})$ なので，ある $s,t \in \mathbb{N}_0$ をとって以下のようにできる

$$\left| \int_{\mathbb{R}} u(z)\overline{\psi(z)}\mathrm{d}z \right| \lesssim \sup_z |z^t u^{(s)}(z)|, \quad \forall u \in \mathcal{S}(\mathbb{R}).$$

任意の $\alpha > 0$ に対して $\mathrm{T}(\alpha, \cdot) * \widetilde{\phi} \in \mathcal{S}(\mathbb{R})$ である。従って，以下が成り立つ

$$
\begin{aligned}
&\left| \int_{\mathbb{H}} \mathrm{T}(\alpha,\beta) \int_{\mathbb{R}} \phi(\alpha z+\beta)\overline{\psi(z)}\mathrm{d}z \frac{\mathrm{d}\alpha\mathrm{d}\beta}{\alpha} \right| \\
&\quad \leqslant \int_0^\infty \left| \int_{\mathbb{R}}\int_{\mathbb{R}} \mathrm{T}(\alpha,\beta)\phi(\alpha z+\beta)\mathrm{d}\beta \cdot \overline{\psi(z)}\mathrm{d}z \right| \frac{\mathrm{d}\alpha}{\alpha} \\
&\quad \lesssim \int_{\mathbb{R}} \sup_z \left| z^t \int_{\mathbb{R}} \mathrm{D}^{0,0}_{s,0}\mathrm{T}(\alpha,\beta)\phi^{(s)}(\alpha z+\beta)\mathrm{d}\beta \right| \frac{\mathrm{d}\alpha}{\alpha} \\
&\quad \lesssim \int_{\mathbb{R}} \sup_p \left| p^t \int_{\mathbb{R}} \mathrm{D}^{0,0}_{s+t,0}\mathrm{T}(\alpha,\beta)\phi^{(s)}(p+\beta)\mathrm{d}\beta \right| \frac{\mathrm{d}\alpha}{\alpha} \\
&\quad \leqslant \int_{\mathbb{R}}\int_{\mathbb{R}} \sup_p \left|p^t\phi^{(s)}(p+\beta)\right| \left|\mathrm{D}^{0,0}_{s+t,0}\mathrm{T}(\alpha,\beta)\right| \frac{\mathrm{d}\beta\mathrm{d}\alpha}{\alpha} \\
&\quad \lesssim \int_{\mathbb{R}}\int_{\mathbb{R}} \sup_p \left|(1+|p+\beta|^2)^{t/2}\phi^{(s)}(p+\beta)\right| \left|\mathrm{D}^{0,0}_{s+t,t}\mathrm{T}(\alpha,\beta)\right| \frac{\mathrm{d}\beta\mathrm{d}\alpha}{\alpha} \\
&\quad \lesssim \int_{\mathbb{H}} \left|\mathrm{D}^{0,0}_{s+t,t}\mathrm{T}(\alpha,\beta)\right| \frac{\mathrm{d}\beta\mathrm{d}\alpha}{\alpha} \\
&\quad \leqslant \sup_{(\alpha,\beta)\in\mathbb{H}} \left|\mathrm{D}^{0,0}_{s+t+\varepsilon,t+\delta}\mathrm{T}(\alpha,\beta)\right| \int_{\mathbb{H}} (\alpha+1/\alpha)^{-\varepsilon}\,(1+\beta^2)^{-\delta/2} \frac{\mathrm{d}\beta\mathrm{d}\alpha}{\alpha},
\end{aligned}
$$

ただし二番目の式は $\partial_z[\phi(\alpha z+\beta)] = \alpha\cdot\phi'(\alpha z+\beta)$ と $\alpha \lesssim \alpha + 1/\alpha$ を繰り返し適用して得られる。三番目の式は $p \leftarrow \alpha z$ と変数変換して $(\alpha+1/\alpha)^s\cdot\alpha^{-t} \lesssim (\alpha+1/\alpha)^{s+t}$ を適用する。五番目の式は $|p| \lesssim (1+p^2)^{1/2}$ と Peetre の不等式 $1+p^2 \lesssim (1+\beta^2)(1+|p+\beta|^2)$ から従う。六番目の式は任意の t に対して $(1+p^2)^{t/2}\phi(p)$ が有界であることから従う。最後の式は Hölder の不等式から従い，この積分は $\varepsilon > 0$ かつ $\delta > 1$ のとき有限の値を取る。従って $\mathcal{W}[\psi;\phi] \in \mathcal{S}'(\mathbb{H})$ である。

Case 3e（$\mathcal{X} = L^1$ かつ $\mathcal{Z} = L^p \cap C$ のとき $\mathcal{A} = \mathcal{S}'$）$\psi$ の連続性により，$\phi * \psi$ もまた連続である。一方，Lusin の定理により，ほとんど至る所の点 $x \in \mathbb{R}$ で $\phi^\star(x) = \phi(x)$ となるような連続関数 $\phi^\star$ が存在する。従って $\phi * \psi$ の連続性により，以下が成り立つ

$$
\phi^\star * \psi(x) = \phi * \psi(x), \quad \text{for every } x \in \mathbb{R}.
$$

$\phi^\star$ と ψ は連続な可積分関数なので，ある $s, t \in \mathbb{R}$ をとって以下のように

できる

$$|\phi^\star(x)| \lesssim (1+x^2)^{-s/2}, \quad s>1,$$
$$|\psi(x)| \lesssim (1+x^2)^{-t/2}, \quad tp>1.$$

従って，以下が成り立つ

$$\begin{aligned}
&\left|\int_{\mathbb{R}} \phi(x)\overline{\psi\left(\frac{x-\beta}{\alpha}\right)}\frac{1}{\alpha}\mathrm{d}x\right| \\
&\quad\lesssim \left|\int_{\mathbb{R}} (1+x^2)^{-s/2}\left(1+\left(\frac{x-\beta}{\alpha}\right)^2\right)^{-t/2}\mathrm{d}x\right|\alpha^{-1} \\
&\quad\lesssim \left|\int_{\mathbb{R}} (1+x^2)^{-s/2}\left(1+(x-\beta)^2\right)^{-t/2}\mathrm{d}x\right|(1+\alpha^2)^{t/2}\alpha^{-1} \\
&\quad\lesssim (1+\beta^2)^{-\min(s,t)/2}(\alpha+1/\alpha)^{t-1}.
\end{aligned}$$

つまり，$\mathcal{W}[\psi;\phi]$ は局所可積分かつ無限遠での増大度が高々多項式程度の関数である。特に，$(t-1)p<m-1$ のとき $\mathcal{W}[\psi;\phi]\in L^p(\mathbb{H};\alpha^{-m}\mathrm{d}\alpha\mathrm{d}\beta)$ である。

Step 4 最後に $\mathscr{R}_\psi f(u,\alpha,\beta)$ のクラス $\mathcal{Y}(\mathbb{Y}^{m+1})$ を調べる（表 5.2 の 6 列目）。$\mathcal{Y}(\mathbb{Y}^{m+1})$ は $\mathcal{X}(\mathbb{S}^{m-1})\hat{\otimes}\mathcal{A}(\mathbb{H})$ から決まる。球面 $\mathbb{S}^{m-1}$ はコンパクトなので，$\mathcal{D}=\mathcal{S}=\mathcal{O}_\mathcal{M}=\mathcal{E}$ また $\mathcal{E}'=\mathcal{O}'_\mathcal{C}=\mathcal{S}'=\mathcal{D}'$ である。これを鑑みて表 5.2 を得る。

A.3 定理 5.3.1 の証明

十分性. 条件を超関数の意味で Fourier 変換して $|\zeta|^m\widehat{\phi}(\zeta)=\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)$ $(\zeta\neq 0)$ を得る。まず $\Omega\backslash\{0\}$ の範囲では，仮定により $\widehat{\eta}$ は連続関数なので，$\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)|\zeta|^{-m}$ は通常の関数の意味での積になり，これは $\widehat{\phi}(\zeta)$ と等しい。一方 $\mathbb{R}\backslash\Omega$ の範囲では，$|\zeta|^{-m}$ は急減少関数 $(\mathcal{O}_\mathcal{M})$ である。従って $\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)|\zeta|^{-m}$ は高々一つの緩増加超関数を含む積 $(\mathcal{S}\cdot\mathcal{S}'\cdot\mathcal{O}_\mathcal{M})$ であって，これは結合的かつ可換なので一意的で，$\widehat{\phi}(\zeta)$ と等しい。従って，

$$\frac{K_{\psi,\eta}}{(2\pi)^{m-1}}=\left(\int_{\Omega\backslash\{0\}}+\int_{\mathbb{R}\backslash\Omega}\right)\frac{\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)}{|\zeta|^m}\mathrm{d}\zeta=\int_{\mathbb{R}\backslash\{0\}}\widehat{\phi}(\zeta)\mathrm{d}\zeta\neq 0.$$

必要性. $\Omega_0 := \Omega \cap [-1,1]$, $\Omega_1 := \mathbb{R}\backslash\Omega_0$ と書く。許容条件により $\int_{\Omega_0\backslash\{0\}} \overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta)|\zeta|^{-m}\mathrm{d}\zeta$ は絶対収束していて，かつ $\widehat{\eta}$ は $\Omega_0\backslash\{0\}$ で連続なので，ある $v_0 \in L^1(\mathbb{R}) \cap C(\mathbb{R}\backslash\{0\})$ で，$\Omega_0\backslash\{0\}$ への制限が以下を満たすものが存在する

$$\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta) = |\zeta|^m v_0(\zeta), \quad \zeta \in \Omega_0\backslash\{0\}.$$

連続性と可積分性から $v_0 \in L^\infty(\mathbb{R})$ なので, 特に $\lim_{\zeta\to+0} v_0(\zeta)$ と $\lim_{\zeta\to-0} v_0(\zeta)$ はいずれも有限である。

一方，$|\zeta|^{-m} \in \mathcal{O}_{\mathcal{M}}(\Omega_1)$ である。また，$\overline{\widehat{\psi}} \cdot \widehat{\eta} \in \mathcal{O}'_{\mathcal{C}}(\mathbb{R})$ である。（なぜならば $\psi * \eta \in \mathcal{O}_{\mathcal{M}}(\mathbb{R})$ であり，Fourier 変換は $\mathcal{O}_{\mathcal{M}}(\mathbb{R})$ と $\mathcal{O}'_{\mathcal{C}}(\mathbb{R})$ の全単射だからである。）従って，ある $v_1 \in \mathcal{O}'_{\mathcal{C}}(\mathbb{R})$ で以下の超関数の意味での等式を満たすものが存在する

$$\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta) = |\zeta|^m v_1(\zeta), \quad \zeta \in \Omega_1.$$

v_0 と v_1 を用いて

$$v := v_0 \cdot \mathbf{1}_{\Omega_0} + v_1 \cdot \mathbf{1}_{\Omega_1}.$$

とおく。$v_0 \cdot \mathbf{1}_{\Omega_0} \in \mathcal{E}'(\mathbb{R})$ かつ $v_1 \cdot \mathbf{1}_{\Omega_1} \in \mathcal{O}'_{\mathcal{C}}(\mathbb{R})$ なので，$v \in \mathcal{O}'_{\mathcal{C}}(\mathbb{R})$ である。従って $\widehat{\phi} = v$ となる $\phi \in \mathcal{O}_{\mathcal{M}}(\mathbb{R})$ が存在して，特に以下が成り立つ

$$\overline{\widehat{\psi}(\zeta)}\widehat{\eta}(\zeta) = |\zeta|^m \widehat{\phi}(\zeta), \quad \zeta \in \mathbb{R}\backslash\{0\}.$$

許容条件により

$$\int_{\mathbb{R}\backslash\{0\}} \widehat{\phi}(\zeta)\mathrm{d}\zeta = \int_{\Omega_0\backslash\{0\}} v_0(\zeta)\mathrm{d}\zeta + \int_{\Omega_1} v_1(\zeta)\mathrm{d}\zeta \neq 0.$$

$\mathbb{R}$ 全体では，原点での特異性を反映して次のようになる

$$\overline{\widehat{\psi}(\zeta)}\left(\widehat{\eta}(\zeta) - \sum_{j=0}^{k} c_j \delta^{(j)}(\zeta)\right) = |\zeta|^m \widehat{\phi}(\zeta), \quad \zeta \in \mathbb{R}.$$

従って超関数の意味で Fourier 変換をとって以下を得る

$$\left[\overline{\widetilde{\psi}} * \left(\eta - \sum_{j=0}^{k} c_j z^j\right)\right](z) = \Lambda^m \phi(z), \quad z \in \mathbb{R}.$$

A.4 定理 5.3.4 の証明

特異積分

$$\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f(x) = \lim_{\substack{\delta\to\infty \\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} \int_\varepsilon^\delta \mathrm{R}f(u,\cdot) * \lambda_\alpha(u\cdot x)\frac{\mathrm{d}\alpha \mathrm{d}u}{\alpha^m}$$

において $\lambda_\alpha(p) = (\Lambda^m\phi)(p/\alpha)/\alpha$ であり，さらに $(\Lambda\phi)(p/\alpha)/\alpha = \alpha\Lambda[\phi(p/\alpha)/\alpha]$ を繰り返し適用して次のように変形できる

$$\begin{aligned}
\int_\varepsilon^\delta (\Lambda^m\phi)\left(\frac{p}{\alpha}\right)\frac{\mathrm{d}\alpha}{\alpha^{m+1}} &= \Lambda^{m-1}\left[\int_\varepsilon^\delta (\Lambda\phi)\left(\frac{p}{\alpha}\right)\frac{\mathrm{d}\alpha}{\alpha^2}\right] \\
&= \Lambda^{m-1}\left[\frac{1}{p}\int_{p/\delta}^{p/\varepsilon}(\Lambda\phi)(z)\mathrm{d}z\right], \quad z \leftarrow p/\alpha \\
&= \Lambda^{m-1}\left[\frac{1}{p}\mathcal{H}\phi\left(\frac{p}{\varepsilon}\right) - \frac{1}{p}\mathcal{H}\phi\left(\frac{p}{\delta}\right)\right] \\
&= \Lambda^{m-1}[k_\varepsilon(p) - k_\delta(p)].
\end{aligned}$$

ただし

$$k(z) := \frac{1}{z}\mathcal{H}\phi\,(z) \quad \text{and} \quad k_\gamma(p) := \frac{1}{\gamma}k\left(\frac{p}{\gamma}\right) \quad \text{for } \gamma = \varepsilon, \delta.$$

とおいた。従って，

$$\int_\varepsilon^\delta \mathrm{R}f(u,\cdot) * \lambda_\alpha \frac{\mathrm{d}\alpha}{\alpha^m} = \Lambda^{m-1}\mathrm{R}f(u,\cdot) * (k_\varepsilon - k_\delta).$$

$k \in L^1 \cap L^\infty(\mathbb{R})$ かつ $\int_{\mathbb{R}} k(z)\mathrm{d}z = 1$ となることを示す。このとき k_γ は近似単位元であり，$\lim_{\gamma\to 0} k_\gamma = \delta$ がいえる。まず $k \in L^1(\mathbb{R})$ を示す。そのために，ある $s, t > 0$ が存在して

$$\begin{aligned}
|k(z)| &\lesssim |z|^{-1+s} \quad \text{as} \quad |z| \to 0 \\
|k(z)| &\lesssim |z|^{-1-t} \quad \text{as} \quad |z| \to \infty.
\end{aligned}$$

が成り立つことを示す。一つ目の条件は $\mathcal{H}\phi(0) = 0$ から従う。実際，ϕ は実数値なので $\widehat{\phi}$ は偶関数で，以下が成り立つ。

$$\mathcal{H}\phi(0) = \int_{\mathbb{R}} \operatorname{sgn}\zeta \cdot \widehat{\phi}(\zeta)\mathrm{d}\zeta = \int_{(-\infty,0]} \widehat{\phi}(\zeta)\mathrm{d}\zeta - \int_{(0,\infty)} \widehat{\phi}(\zeta)\mathrm{d}\zeta = 0.$$

二つの条件は $\phi \in L^1(\mathbb{R})$ なので ϕ と $\mathcal{H}\phi$ がともに無限遠で減衰することから従う。以上で可積分性が示された。有界性は k が連続かつ可積分であることから自動的に従う。次に $\int_{\mathbb{R}} k(z)\mathrm{d}z = 1$ を示す。こちらは仮定 $\int_{\mathbb{R}} \widehat{\phi}(\zeta)\mathrm{d}\zeta = -1$ を用いて次のように計算で示せる

$$\int_{\mathbb{R}} k(z)\mathrm{d}z = -\int_{\mathbb{R}} \frac{\mathcal{H}\phi(z)}{0-z}\mathrm{d}z = -\phi(0) = 1.$$

一方，$\lim_{\delta\to\infty} k_\delta = 0$ である。実際，$k \in L^\infty(\mathbb{R})$ なので $\Phi \in L^1(\mathbb{S}^{m-1}\times\mathbb{R})$ に対して以下が成り立つ

$$\|\Phi * k_\delta\|_{L^\infty(\mathbb{S}^{m-1}\times\mathbb{R})} \leqslant \delta^{-1}\|\Phi\|_{L^1(\mathbb{S}^{m-1}\times\mathbb{R})}\|k\|_{L^\infty(\mathbb{R})}.$$

最後に，Vitalli の優収束定理を用いて $\mathbb{S}^{m-1}$ 上の積分と極限の交換が示せる。

$$\begin{aligned}\mathscr{R}_\eta^\dagger\mathscr{R}_\psi f(x) &= \lim_{\substack{\delta\to\infty\\ \varepsilon\to 0}} \int_{\mathbb{S}^{m-1}} [J(u,\cdot) * (v_\varepsilon - v_\delta)](u\cdot x)\mathrm{d}u \\ &= \int_{\mathbb{S}^{m-1}} J(u, u\cdot x)\mathrm{d}u, \quad \text{a.e. } x \in \mathbb{R}^m \\ &= \mathrm{R}^\dagger\Lambda^{m-1}\mathrm{R}f(x).\end{aligned}$$

A.5 定理 5.3.8 の証明

(ψ,ψ) と (η,η) は自己許容的であるとして一般性を失わない。以下では

$$I[f;(\varepsilon,\delta)](x) := \int_{\mathbb{S}^{m-1}}\int_\varepsilon^\delta\int_{\mathbb{R}} \mathscr{R}_\psi f\left(\frac{u}{\alpha}, \frac{u\cdot x-\beta}{\alpha}\right)\eta\left(\frac{\beta}{\alpha}\right)\frac{\mathrm{d}\beta\mathrm{d}\alpha\mathrm{d}u}{\alpha^{m+1}}.$$

とし，$\Omega[\varepsilon,\delta] := \mathbb{S}^{m-1}\times[\mathbb{R}_+\backslash(\varepsilon,\delta)]\times\mathbb{R} \subset \mathbb{Y}^{m+1}$ と書く。

$$\begin{aligned}\big\|f - I[f;(\varepsilon,\delta)]\big\|_2 &= \sup_{\|g\|_2=1} \big|\,(f - I[f;(\varepsilon,\delta)], g)\,\big| \\ &= \sup_{\|g\|_2=1} \big|\,(\mathscr{R}_\psi f, \mathscr{R}_\eta g)_{\Omega[\varepsilon,\delta]}\,\big| \\ &\leqslant \sup_{\|g\|_2=1} \big|\mathscr{R}_\psi f\big|_{L^2(\Omega[\varepsilon,\delta])}\big\|\mathscr{R}_\eta g\big\|_{L^2(\mathbb{Y}^{m+1})} \\ &= \sup_{\|g\|_2=1} \big|\mathscr{R}_\psi f\big|_{L^2(\Omega[\varepsilon,\delta])}\|g\|_2 \\ &\to 0\cdot 1, \quad \text{as} \quad \varepsilon\to 0 \text{ and } \delta\to\infty\end{aligned}$$

三番目の式は Schwartz の不等式を用いた。最後の極限は $\Omega[\varepsilon,\delta]\to\varnothing$ に従って $\|\mathscr{R}_\psi f\|_{L^2(\Omega[\varepsilon,\delta])}$ が消滅することから従う。

A.6 定理 6.5.2 の証明

仮定より $k|_M$ は単射なので, 左逆写像 $\phi : \widetilde{M} \to M$ が存在して, $\phi \circ k|_M = \mathsf{id}_M$ を満たす。ϕ は $\widetilde{M}$ と M の間の微分同相である。

任意の関数 $f \in C^2(M)$ の ϕ による引き戻しを $\phi^* f (:= f \circ \phi)$ と書く。各点 $x \in \widetilde{M}$ で以下が成り立つ[1]

$$(\nabla f) \circ \phi(x) = k^\top \widetilde{\nabla} \phi^* f(x), \tag{A.4}$$

$$(\nabla^2 f) \circ \phi(x) = k^\top \widetilde{\nabla}^2 \phi^* f(x) k. \tag{A.5}$$

また π_0 の引き戻し $\phi^*\pi_0$ について，確率変数の変数変換の公式から次の関係式が成り立つ

$$|k|^{-1} \phi^* \pi_0(x) = \widetilde{\pi}_0(x), \quad x \in \widetilde{M}.$$

記述を簡潔にするため,

$$\mathcal{E} := \log e^{tL_t} \pi_0$$

と書く。このとき各点 $x \in \widetilde{M}$ で

$$\begin{aligned} \Phi \circ \phi(x) &= \phi(x) + tK(\nabla \mathcal{E}) \circ \phi(x) \\ &= \phi(x) + tKk^\top \widetilde{\nabla} \phi^* \mathcal{E}(x). \end{aligned}$$

従って，$\widetilde{K} := kKk^\top$ として以下を得る

$$k \circ \Phi \circ \phi(x) = x + t\widetilde{K} \widetilde{\nabla} \phi^* \mathcal{E}(x), \quad x \in \widetilde{M}. \tag{A.6}$$

(A.6) の右辺が $\widetilde{\Phi}$ に一致することを示す。まず,

$$\begin{aligned} u(z,t) &:= e^{tL_t} \pi_0(z), \quad z \in H \\ \widetilde{u}(\cdot, t) &:= |k|^{-1} \phi^* u(\cdot, t) \end{aligned}$$

とおく。定義から直ちに $u(z,0) = \pi_0(z)$, $\widetilde{u}(x,0) = \widetilde{\pi}_0(x)$ および

$$\partial_t u(z,t) = L_t u(z,t), \quad z \in H$$

[1] 左辺はそれぞれ，M 上の勾配と Hess 行列の引き戻しなので，$\phi^*(\nabla f)$, $\phi^*(\nabla^2 f)$ とも書ける。

が成り立つ。さらに，各点 $x \in \widetilde{M}$ において

$$\partial_t \tilde{u}(x,t) = \widetilde{L}_t \tilde{u}(x,t), \tag{A.7}$$

が成り立つ。ただし $\widetilde{L}_t$ の係数は以下で与えられる

$$\begin{aligned}
\tilde{a}(x,t) &:= ka(\phi(x),t),\\
\tilde{b}(x,t) &:= kb(\phi(x),t),\\
\tilde{c}(x,t) &:= c(\phi(x),t).
\end{aligned}$$

実際，直接計算によって，以下のように示せる

$$\begin{aligned}
\partial_t \tilde{u}(x,t) &= |k|^{-1}\partial_t u(\cdot,t)\circ\phi(x)\\
&= |k|^{-1}L_t u(\cdot,t)\circ\phi(x)\\
&= |k|^{-1}a(\phi(x),t)^\top \nabla^2 u(\cdot,t)\circ\phi(x)a(\phi(x),t)\\
&\quad + |k|^{-1}b(\phi(x),t)^\top \nabla u(\cdot,t)\circ\phi(x) + |k|^{-1}c(\phi(x),t)^\top u(\phi(x),t)\\
&= |k|^{-1}a(\phi(x),t)^\top k^\top \widetilde{\nabla}^2 \phi^* u(x,t) ka(\phi(x),t)\\
&\quad + |k|^{-1}b(\phi(x),t)^\top k^\top \widetilde{\nabla}\phi^* u(x,t) + |k|^{-1}c(\phi(x),t)^\top u(\phi(x),t)\\
&= \tilde{a}(x,t)^\top \widetilde{\nabla}^2 \tilde{u}(x,t)\tilde{a}(x,t) + \tilde{b}(x,t)^\top \widetilde{\nabla}\tilde{u}(x,t) + \tilde{c}(x,t)^\top \tilde{u}(x,t)\\
&= \widetilde{L}_t \tilde{u}(x,t).
\end{aligned}$$

よって，$\tilde{u}$ は $\tilde{\pi}_0$ を初期データとする拡散方程式 (A.7) の解なので，初期値問題の解の一意性により

$$\tilde{u}(x,t) = e^{t\widetilde{L}_t}\tilde{\pi}_0(x), \quad x \in \widetilde{M}$$

が従う。つまり，対数微分によって定数倍の差が無視できることに注意して，以下が成り立つ

$$\widetilde{\nabla}\phi^*\mathcal{E}(x) = \widetilde{\nabla}\log e^{t\widetilde{L}_t}\tilde{\pi}_0(x).$$

これを (A.6) に代入して

$$k\circ\Phi\circ\phi(x) = \widetilde{\Phi}(x), \quad x \in \widetilde{M}.$$

$\widetilde{M}$ 上，ϕ は全単射なので，次の位相共役性を得る

$$k\circ\Phi(z) = \widetilde{\Phi}\circ k(z), \quad z \in M.$$

(A.4) の導出

成分計算によって示す。まず連鎖律により，

$$\frac{\partial f \circ \phi}{\partial x_i}(x) = \sum_{q=1}^{J} \frac{\partial f}{\partial z_q}(\phi(x)) \frac{\partial \phi_q}{\partial x_i}(x).$$

次に両辺に k_{ip} をかけて和をとると，

$$\begin{aligned}
\sum_{i=1}^{I} k_{ip} \frac{\partial f \circ \phi}{\partial x_i}(x) &= \sum_{q=1}^{J} \frac{\partial f}{\partial z_q}(\phi(x)) \sum_{i=1}^{I} k_{ip} \frac{\partial \phi_q}{\partial x_i}(x) \\
&= \sum_{q=1}^{J} \frac{\partial f}{\partial z_q}(\phi(x)) \sum_{i=1}^{I} \frac{\partial k_i}{\partial z_p}(\phi(x)) \frac{\partial \phi_q}{\partial x_i}(x) \\
&= \sum_{q=1}^{J} \frac{\partial f}{\partial z_q}(\phi(x)) \delta_{pq} = \frac{\partial f}{\partial z_p}(\phi(x)).
\end{aligned}$$

(A.5) の導出

まず (A.4) の両辺を微分して，

$$\sum_{i=1}^{I} k_{ip} \frac{\partial^2 f \circ \phi}{\partial x_j \partial x_i}(x) = \frac{\partial f_p \circ \phi}{\partial x_j}(x).$$

ただし $f_p(z) := \frac{\partial f}{\partial z_p}(z)$ と書いた。従って，(A.4) を再び適用して，

$$\sum_{j=1}^{I} \sum_{i=1}^{I} k_{jq} k_{ip} \frac{\partial^2 f \circ \phi}{\partial x_j \partial x_i}(x) = \frac{\partial f_p}{\partial z_q}(\phi(x)) = \frac{\partial^2 f}{\partial z_q \partial z_p}(\phi(x)).$$

付 録B　背景知識

本研究の課題の解決と結果の考察にあたり，背景となる知識についてまとめる。まず，ニューラルネットの中ではどのように情報を表現し，処理しているのだろうか。この問題を理解するために，情報やデータ表現について整理する。また，ニューラルネットはどのように設計すべきか。この問題を理解するために，複雑性やモデル選択について整理する。そして，ニューラルネットの中では，どのように秩序が形成されているのだろうか。この問題をアナロジーとして理解するために，水と油が分離する原理の考え方を整理する。

B.1　情報とは何か

情報とは何だろうか。数学的に定義されているものに限っても，Shannon 情報量と Fisher 情報量のように，情報には複数の異なる定義がある。情報は文脈に応じて，記号，信号，メッセージ，データ，記録，事実，証拠，暗号，秘密，プライバシー，意味，（系に対する）入出力，価値，知識，構造，法則，真実，特徴量，メディア，エントロピーといった言葉の代わりに使われることもある。

本節では，ニューラルネットの中で起きている情報処理や，情報の表現様式を理解するために，Shannon 情報量を軸として古典的な情報概念を整理する。まず，Shannon の意味での情報について基本的な理解を説明したあと，情報理論，統計学，信号処理における情報について詳細な性質を概観する。そして，集合代数を用いて Shannon 情報が事象の生起に纏わる情報であることを説明する。ここまでで取り上げるのはいずれも Shannon 情報と親和性の高い概念である。続いて，Shannon 情報と関連するが異なる情報概念をいくつか取り上げ，Shannon 情報について批判的に理解する。最後に，Shannon 情報を取り巻く概念の変遷を時系列で整理する。

B.1.1 基本的な理解

Shannon の意味での情報とは**我々の知識を確実にするための手がかりや，事象を絞り込んで特定するためのヒント**のことである。例えば帰り道の電車で，今夜の晩ごはんが何か考えているとしよう。考えられる候補は，ビーフカレーか野菜カレーかカツカレーであることが分かっているとする。そこにメールが届いて，「今日はお肉が安かった」とあれば，候補はビーフかカツに絞り込まれる。このメールの内容が，晩ごはんを当てる問題を解くうえでの情報である。

相互情報量を使うと，このメールの情報量を測ることができる。そのためにまず，晩ごはんを当てる問題の難しさを測る尺度として Shannon 情報量（エントロピー）から説明する。

Shannon 情報量（エントロピー）

まず晩ごはんの候補

$$\mathcal{X} = \{\text{ビーフ}, \text{野菜}, \text{カツ}\}$$

とし，簡単のため $\mathcal{X}$ を標本空間 Ω と同一視して，確率変数 $X : \Omega \to \mathcal{X}$ を考える。X が従う確率分布を P として，離散確率変数 X の**エントロピー**を以下で定義する

$$H[X] := -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x).$$

この量は，後述する解釈により，晩ごはんを当てる問題の難しさを定量的に評価している。(X, P) は晩ごはんを当てる問題を表しており，情報源と呼ばれる。

相互情報量

次に，可能なメールのメッセージの全体を $\mathcal{Y}^*$ とし，$\mathcal{Y}^*$ から Ω の生起に関する内容のみを抽出した集合を改めて $\mathcal{Y}$ として，メールの内容を表す確率変数を $Y : \Omega \to \mathcal{Y}$ とする[1]。メールのメッセージは晩ごはんの組

[1] Y の代わりに $\mathcal{Y}^*$ に値をとる確率変数 $Y^* : \Omega^* \to \mathcal{Y}^*$ を考える方が直接的であり，自然である。ただし $\Omega^* = \mathcal{Y}^*$ とする。このとき Y^* は Ω 上の確率変数とはみなせない。なぜならばメッセージの数は明らかにカレーの候補の数よりも多いので，$\Omega \to \mathcal{Y}^*$ は全射にならないからである。X についての情報を取り出すためにはまず $\Omega \times \Omega^*$ 上の結合分布 $P(X, Y)$ を導入し，ベイズの公式を用いて条件付き分布 $p(X|Y)$ を計算する。

合せを全て尽くすことができると考えられるので，$\mathcal{Y} = 2^{\Omega}$ と仮定してよい。「今日はお肉」というメールの内容に対応する事象を

$$A = \{\text{ビーフ}, \text{カツ}\}$$

とする。メールを受信する前のエントロピーは $H[X]$ であったのに対して，メールを受信した後のエントロピーは条件付き確率 $P(X|A)$ によるエントロピー $H[X|A]$ に変化する。従って，メールの情報量は，二つのエントロピーの差 $H[X] - H[X|A]$ で測ることができる。ただし一般にこの量は負の値をとることもある[2]が，その期待値である**相互情報量**

$$I[X;Y] := H[X] - H[X|Y]$$

は，常に非負の値をとる[3]。すなわち，平均的には，情報を得ることによって問題の難しさは減少すると言える。

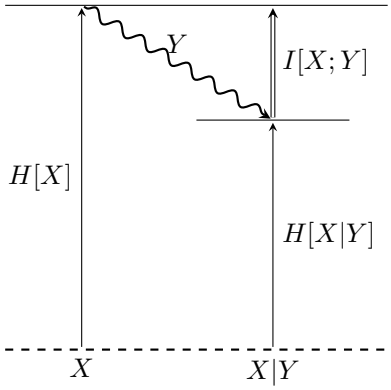


図 B.1: 相互情報量

Kullback-Leibler 情報量（KL ダイバージェンス）

相互情報量の場合には X が従う真の確率分布 P を既知としていたが，現実には適当なモデル Q を想定することになる。このとき「私にとって

[2]例えば，$P(X = \text{ビーフ}) = 0.1, P(X = \text{カツ}) = 0.1, P(X = \text{野菜}) = 0.8$ のように確率が偏っている場合に，少数派で条件付けた場合には $H[X] \approx 0.92$ に対して $H[X|A] = 1.0$ のようにエントロピーを増すことがある。

[3]KL 情報量の正値性から導く

の」問題の難しさは**交差エントロピー**

$$H[X;Q] := -\sum_{X\in\mathcal{X}} P(X)\log Q(X).$$

によって測られることになる。基本的な不等式により $H[X] \leqslant H[X;Q]$ が成り立つ。従って，真の確率分布 P を知っている「神様」から見ると，交差エントロピーは問題の難しさを過大評価していることになる。この差分

$$KL[P\|Q] := H[X;Q] - H[X].$$

を *Kullback-Leibler* **情報量（*KL* ダイバージェンス）**と呼ぶ。

このとき相互情報量は

$$\begin{aligned} H[X;Q] - H[X;Q|_Y] &= H[X] + KL[P\|Q] - (H[X|Y] + KL[P|_Y\|Q|_Y]) \\ &= I[X;Y] + KL[P\|Q] - KL[P|_Y\|Q|_Y] \end{aligned}$$

となる。ただし Y で条件付けた P, Q をそれぞれ $P|_Y, Q|_Y$ と書いた。

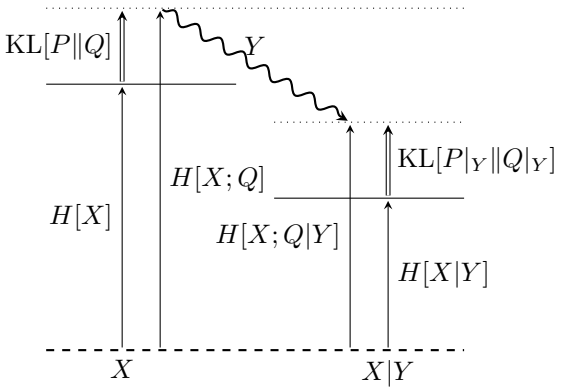


図 B.2: モデル Q を想定する場合の相互情報量

Fisher 情報量

メールのやり取りを N 晩繰り返すと，毎晩どのカレーが出たかという実績データ

$$\mathcal{D} = \{X_1, \ldots, X_N\},$$

が得られる。ただし X_n は第 n 日目のカレーを表す。カレーの出方は共通の多項分布 $Q(\theta)$ に従って毎晩独立に表れるものとして，データ $\mathcal{D}$ から多項分布 $Q(\theta)$ のパラメータ

$$\theta = (\theta_{\text{ビーフ}}, \theta_{\text{野菜}}, \theta_{\text{カツ}})$$

を推定したい。ただし θ の各成分はそれぞれ，ビーフ，野菜，カツが出現する確率を表す。

パラメータ θ の推定量は無数に考えられるが，どのように推定量 $\widehat{\theta}$ を作っても，真のパラメータ θ との間にはほぼ必ず推定誤差が生じる。推定誤差を可能な限り小さくするには，どのような推定をしたら良いだろうか。Cramér-Rao の定理によれば，推定誤差の分散 $\mathsf{var}\,\widehat{\theta}_N$ は，*Fisher* **情報行列**

$$J_{ij}(\theta) := \mathbb{E}_\theta[\partial_i \log Q(\theta) \partial_j \log Q(\theta)]$$

の逆行列によって下から評価されることが分かっている。つまり，どのような推定量を用いても，この精度を（平均的に）下回ることはできない。このように，Fisher 情報行列（情報量）は，データから取り出せる限界の情報量を表している。

例えば多項分布の場合は，

$$\mathsf{var}\,\widehat{\theta}_N \geqslant (NJ(\theta))^{-1} = \frac{1}{N} \begin{bmatrix} \theta_{\text{ビーフ}} & 0 & 0 \\ 0 & \theta_{\text{野菜}} & 0 \\ 0 & 0 & \theta_{\text{カツ}} \end{bmatrix}$$

となる。そして，多項分布の場合には，最尤推定量

$$\widehat{\theta}_N := \left(\frac{N_{\text{ビーフ}}}{N}, \frac{N_{\text{野菜}}}{N}, \frac{N_{\text{カツ}}}{N} \right),$$

によって下限を達成できることが知られている。ここで N_x は各 $x \in \mathcal{X}$ の登場回数を表す。

事前知識と十分統計量

多項分布の最尤推定量において，データ $\mathcal{D}$ は実績の累積数のみを記録しておけば十分であり，カレーの順番のようなデータは冗長であること

が分かる。このように，統計モデル $Q(\theta)$ のパラメータ θ を推定するのに十分なデータを**十分統計量**と呼ぶ。

メールからカレーを当てる問題や，カレーの履歴データからカレーの出現確率を推定する問題では，標本空間 $\Omega = \mathcal{X}$ やメールの空間 $\mathcal{Y}$，カレーの分布 $Q(\theta)$ などを予め自分たちで設定した。このように，何かを推定するには，まず問題を定式化するためのモデルを設定する必要があり，このときにデータからモデルを定めるというメタ推定問題を解いている。これを**事前知識**と呼ぶ。

B.1.2 情報理論における情報

Cover and Thomas (2006, Ch.3.1, Ch.11.10, Ch.17.7) に曰く，「Shannon 情報量は典型集合の体積に関連する概念であり，Fisher 情報量は典型集合の表面積に関連する概念である。」つまり，AEP により Shannon 情報量は典型集合の体積の対数に漸近することと，Fisher 情報量は KL ダイバージェンスの Hesse 行列であることが本質的だという。

確率変数 $X : \Omega \to \mathcal{X}$ にエントロピー汎関数 $H[X]$ を導入したら，情報理論になるだろうか。そうだとすれば情報理論の数学的構造は確率構造を決定した時点で全て決まることになる。否，情報理論は単なる確率論ではなく，通信理論，符号理論，暗号理論，信号理論，統計学や物理学の対象を，エントロピーという観点から横断的に取り上げる総合的な科学である。

Shannon 情報量

Shannon 情報量 $H[X]$ は自己情報量 $-\log P(X)$ の期待値

$$H[X] := \mathbb{E}_X[-\log P(X)],$$

である。

事象 A に対して，自己情報量 $-\log P(A)$ は，事象 A の生起確率が小さく稀であるほど大きな値をとる。従って，自己情報量は事象 A が生起した時の驚きの度合いや，事象 A の生起を当てる問題の難しさを表している。その期待値である $H[X]$ は，X を予測する問題の平均的な難しさを表している。なお，結合分布の Shannon 情報量は劣モジュラ関数である。

Shannon 情報量は統計力学における Boltzmann エントロピー（$S = k \log W$）と同じ式で表されることから，情報論的エントロピーとも呼ばれる。このアナロジーにより，確率変数 X を物理量とみなし，確率分布 P を物理量の状態とみなすと，$H[X]$ は系 (X, P) の乱雑さを表しているとみなせる。

情報源符号化定理によれば，$H[X]$ は X を瞬時符号によって符号化する場合の平均符号長の下限になる。つまり，$H[X]$ は符号長としての実体をもつ。最小記述長（Minimum Description Length; MDL）は，符号長の下限を出発点として定義された複雑性の一種である。MDL は Kolmogorov 複雑性や BIC と近似的に等価である。

大数の弱法則により，$X_i \sim P$ を iid 系列として，

$$-\frac{1}{n} \log P(X_1, \cdots, X_n) \to^P H[X]$$

が成り立つ。すなわち，系列のなかで典型的なものが生起する確率は等しく $2^{-nH[X]}$ に漸近する。この性質を漸近等分割性（Asymptotic Equipartition Property; AEP）という。

総和 $-\sum_x P(x) \log P(x)$ を積分 $-\int p(x) \log p(x) \mathrm{d}x$ に置き換えることで，Shannon 情報量は連続確率変数に対して拡張できる。これを微分エントロピーという。有限回の質問で実数を特定することは不可能なので，情報源を量子化しない限り符号長として微分エントロピーを解釈することは不可能である。一方，AEP は成り立つ。また微分エントロピーは一般に x の座標変換によって不変ではないので，微分幾何学的な量ではない。一方，その差である KL ダイバージェンスは座標変換によって不変である特に，分散を固定した確率分布で，微分エントロピー最大の元は正規分布であり，これは正規分布の特徴付けになっている。

確率過程のエントロピーは AEP を拡張する形で定義する。すなわち，“単位時間” あたりのエントロピー生成速度の極限 $\lim_{n\to\infty} -(1/n) \log P(X^n)$ が存在するとき，これをエントロピーレートと呼び，確率過程のエントロピーと考える。

位相エントロピーは，測度論的な量である Shannon エントロピーの位相版である。まず，準距離空間 (T, d) が全有界な場合には，有限個の開集合の族によって全空間を覆うことができる（有限開被覆）。半径 ε の開被覆による被覆数の下限を $N(T, d, \varepsilon)$ とすると，$-\log N(T, d, \varepsilon)$ は開被覆を根源事象とする一様分布を入れた場合のエントロピーとみなせる。これを被覆エントロピーという。被覆エントロピーと同値なエントロピーで

ある ε-エントロピーは，Vapnik-Chervonenkis 次元に繋がる概念である。

相互情報量

相互情報量 $I[X;Y]$ は系 X を観測して観測値 Y を得た時の，系 $X|Y$ のエントロピーの減少度

$$I[X;Y] := H[X] - H[X|Y],$$

である[4]。後述する KL 情報量を用いると，

$$I[X;Y] = KL(p(X,Y)\|p(X)p(Y))$$

と書ける。従って X と Y が統計的独立であれば相互情報量は 0 になる。つまり，相互情報量は独立性の尺度である。

マルコフ過程 $X \to Y \to Z$ において，Z は先行する Y 以上に X についての情報を持つことはない

$$I[X;Y] \geqslant I[X;Z].$$

これをデータ処理不等式という。

送信者が送るメッセージを確率変数 X とし，受信者が受け取るメッセージを Y として，通信路を条件付き確率 $p(Y|X)$ でモデル化する。通信路符号化定理によれば，通信路容量 $C := \max_{X\sim P} I[X;Y]$ を超えない任意の伝送レートが達成可能である。

Kullback-Leibler 情報量

$KL[P\|Q]$ は真の分布 P から見たモデル Q のエントロピーの差分

$$\begin{aligned} KL[P\|Q] &:= H[P;Q] - H[P] \\ &= \mathbb{E}_P[\log P/Q] \end{aligned}$$

であり，対数尤度比の期待値である。従って KL 情報量は適合度の尺度とみなせる。Csiszár はこの観点を一般化した f-ダイバージェンスを研究した。特に多項分布の場合には，KL 情報量は χ^2-統計量に漸近する。

[4] 具体的に Y から $X|Y$ を求める操作が統計的推定や学習ないし最適化である。

情報不等式により，$KL[P\|Q] \geqslant 0$ であり，等号は $P = Q$ の時に限る。従って KL 情報量は確率分布の距離とみなせる。ただし距離の公理を満たさないのでダイバージェンスと呼ばれることが多い。情報幾何学では KL ダイバージェンスを拡張して多様なダイバージェンスを検討する。統計モデル $P(\theta)$ において，KL 情報量の二階微分（Hesse 行列）は，Fisher 情報行列である (Jeffreys, 1946)

$$\frac{\partial^2}{\partial t \partial s} KL[P(\theta)\|P(\theta + t\xi + s\eta)]\Big|_{(t,s)=(0,0)} = J(\theta)(\xi, \eta).$$

このことから，Fisher 情報行列の行列式は統計多様体の主曲率とみなせる。

KL ダイバージェンスの一般化は様々な方面から検討されてきた。α-ダイバージェンスは統計多様体の双対幾何を導く基本的な例として注目を集めた。双対構造は凸構造に起因し，一般に Bregman-ダイバージェンスに対して導ける。f-ダイバージェンスは尤度比 p/q の形式を拡張したものである。NMF では β-ダイバージェンス，音声信号処理では Itakura-Saito ダイバージェンスなど，タスクに特化したダイバージェンスもある。

確率分布の距離には他にも，全変動距離や Wasserstein 距離などがあり，Pinsker 不等式

$$\|P - Q\|_{TV} \leqslant \sqrt{\frac{1}{2} KL[P\|Q]},$$

や，Q が標準正規分布のとき Talagrand 不等式

$$W_2(P, Q) \leqslant \sqrt{2KL[P\|Q]},$$

が成り立つ。Wasserstein 距離と KL 情報量の平方根の不等式を輸送不等式という[5]。

B.1.3 統計学における情報

データ，情報，知識

Rao (2010) はデータと情報，知識および知恵を区別した。まず，データがどのようにして取得されたかという知識や，専門家の見解などを含

[5] 全変動距離は Hamming 距離に対する Wasserstein 距離なので，Pinsker 不等式は輸送不等式の一種である。

めてデータと呼んだ。このデータからモデルを特定し，仮説が妥当かどうかを統計的に検証することで，仮説の不確実さを評価する。こうして得られた仮説と不確実さのセットを情報と呼んだ。仮説の修正を繰り返すことによって，価値の高い情報が生まれ，知識となる。そして，このようにデータから知識を獲得するための知識を知恵と呼んだ。

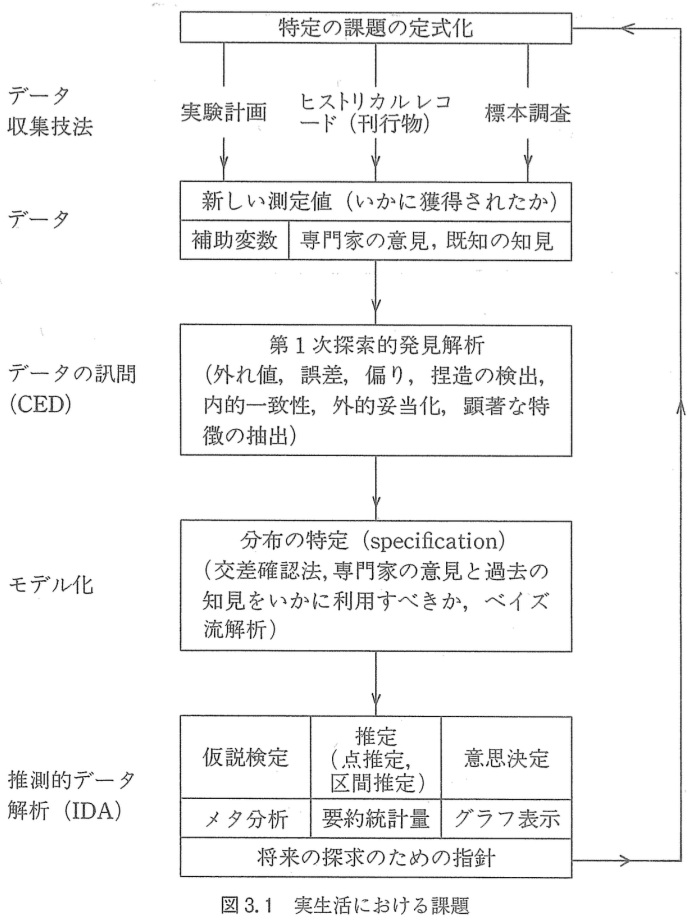


図 3.1 実生活における課題

図 B.3: Rao (2010) による統計解析フロー

十分統計量

Fisher (1925) は**十分統計量**を導入した。十分統計量とはデータ X の関数 $T(X)$ で，X が従う確率モデル P_θ のパラメータ θ を推定するために十分なもののことである。ここで十分であるとは，条件付き確率 $P_\theta[X|T=t]$

がθに依存しないという意味である。統計量$T(X)$が十分統計量であることと，確率変数Yの任意の確率分布に対して，$I[Y;X] = I[Y;T(X)]$となることは同値である。Xのうち十分統計量に寄与しないデータは，“ゴミデータ”である。つまり，あるデータが情報なのかどうかは，モデルの設定次第という考え方である。

事前知識

統計的推測にあたっては，データに対して必ず統計的モデルを想定しなければならない。モデルを想定することをFisherは特定化（specification）と呼んだ。より一般に**逆問題**という観点でも，有限のデータから関数を特定するためには，関数空間を制限するための先験的な情報が必要である。例えば**正則化項や罰則項**なども関数空間を制限するための事前知識である。

Bayes推定では，**事前分布**を用いて事前知識を統計的推測のプロセスに組み込むことができる。Fisher統計学の全盛時代には，事前分布は恣意性があるので客観的ではないという批判があった（主観確率）。これを承けて，無情報事前分布や共役事前分布などの「客観的な」事前分布を設定する方法も開発されてきた。現代的に見ると，事前分布が主観的ないし恣意的とみなされるのは，唯一無二の「真の構造」なるものを想定しているためである。赤池統計学や機械学習，知識発見を目的とするデータマイニングなどでは，真の構造ではなく，予測や汎化を第一の目的とし，恣意的かどうかよりも，過学習を懸念する。つまり，真の構造は必ずしも推定できなくても構わず，想定する必要すらないのである。従って，予測や知識発見といった目的を達成する限りにおいては事前分布も迷わず試してみるというのが現代的な考え方といえる。

Fisher情報量

パラメータ付けられた統計モデルp_θに対して，Fisher情報行列は

$$J_{ij}(\theta) := \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log p_\theta(X) \, \frac{\partial}{\partial \theta_j} \log p_\theta(X) \right]$$

で定義される。ただし $\mathbb{E}_\theta$ は p_θ による期待値を表す。適当な正則条件のもとでスコアの期待値は常に 0 である

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log p_\theta(X) \right] \equiv 0$$

なので,

$$J_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(X) \right]$$

が成り立つ。ノンパラメトリックの場合にも，スコアを方向微分の意味で定義して同様に定義できる。

Cramér-Rao の不等式により，Fisher 情報行列の逆行列は不偏推定量の共分散行列の下限である。つまり，Fisher 情報量は，統計的推定において，一つのサンプルによって減少させることのできるパラメータの不確実性の下限である。図 B.4 は，正規分布 $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ から独立に生成した $N = 10$ 個の正規乱数 $x_n \, (n = 1, \cdots, N)$ に対して，分散 σ^2 を既知として平均 μ の不偏推定量 $\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$ を繰り返し計算した例である。Cramér-Rao の定理により推定誤差分散の平均的な下限は $\sigma^2/N = 0.1$ である。ヒストグラムの実線は推定誤差の標準偏差，点線は真値を示す。Fisher 情報量（の逆数）とは，点線から実線までの幅に相当する量である。

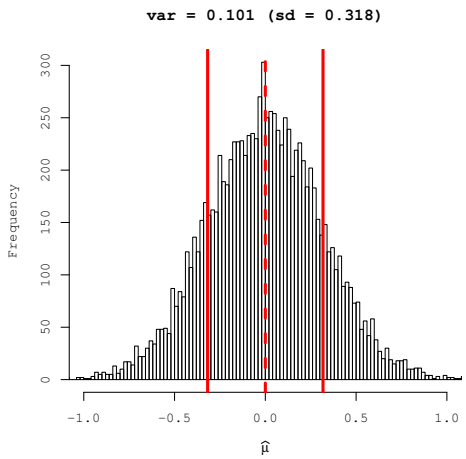


図 B.4: Fisher 情報量は推定誤差分散の逆数の下限

既に述べた通り，Fisher 情報行列は KL ダイバージェンスの Hesse 行列である。つまり，KL ダイバージェンスを Taylor 展開によって二次近似し

た場合の，第二次項である。また，接ベクトルは Fisher スコア $\partial_\theta \log p_\theta$ である。そして，適当な正則条件のもとで Fisher 情報行列は正定値である。従って，統計モデル $\{p_\theta\}$ を θ で座標付けられた微分可能多様体とみなすと，Fisher 情報行列はそのうえの Riemann 計量となる。これを Fisher 計量とよび，統計モデルに Fisher 計量を入れたもの（厳密にはさらに α-接続を入れる）を統計多様体という。Cramér-Rao 不等式は Fisher 計量に関する Cauchy-Schwartz の不等式として得られる。

B.1.4 信号処理における情報

ここで信号とは，関数ないし確率過程のこと，すなわち生体信号や画像データ，時空間データのことをいう。AEP にも現れている通り，Shannon 情報量は有限性の強い概念である。従って，信号のように連続無限クラスの対象に対して Shannon 情報量を定義する方法は自明ではない。Shannon は信号を連続情報源と呼び，信号を情報理論の枠組み（即ち確率空間にエントロピー汎関数を入れたもの）で扱う方法を検討した。拡張の方針は，アナログ信号のデジタル化であり，その成果の一つは標本化定理として知られる。Shannon による信号の離散化は，情報をいかに表現するかという一般的な問題の先駆的な事例でもある。

信号 $s : X \to Y$ の定義域 X を離散化することを標本化といい，値域 Y を離散化することを量子化という。Shannon の標本化定理は，信号を等間隔に標本化しても帯域制限された空間の中から信号を特定できることを保証する定理である。一方，レート歪み理論では，量子化 $X \to \widehat{X}$ に伴う歪み（distortion）を情報損失 $I[X;\widehat{X}]$ で測り，歪みを最低限に抑える方法として k-means やベクトル量子化が適切であることを主張している。

標本化

標本化とは，有限個の数字の組を座標（成分ベクトル）とみなして，関数空間上の点に対応付けることである。標本化定理は，座標に対応する基底として三角関数の列をとると，関数空間の大きさが Nykist 周波数で測れることを主張している。関数を基底（frame, atom）と係数に分解する方法を調べる学問は，調和解析と呼ばれる[6]。調和解析では，係数を表現（representation）と呼ぶ。

[6]古典調和解析は Fourier 変換論ともいえるが，一般に積分変換は関数空間上の座標変換なので，座標の取り方を研究する学問と言い換えられる。現代に目を向けると，ウェー

量子化

一方，量子化とは，ユークリッド空間を適当な集合族に分割することである。例えば k-means ではユークリッド距離とデータの分布を使って空間を分割する方法である。集合族を通じて空間の構造を調べる方法は実解析的[7]ともいえるが，量子化やクラスタリングの技術は専ら情報理論や統計学によるところが大きい。

B.1.5 集合代数としての情報

Shannon の情報は，事象が生起したかどうかについての情報である。つまり，標本空間を Ω として「生起が分かる Ω の事象」の全体を $\mathcal{F}$ と定義すると，$\mathcal{F}$ は σ-代数（完全加法族）になる。まず，事象 $A \subset \Omega$ の生起が分かるとしよう。このとき $\mathcal{F}$ の定義より $A \in \mathcal{F}$ である。また，余事象 A^c が生起しなかったことも分かるので，$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ である。そして，二つの事象 $A, B \subset \Omega$ の生起が分かるとしよう。このとき「A または B」や「A かつ B」が生起したかどうかも分かるので，$A, B \in \mathcal{F} \Rightarrow A \cup B, A \cap B \in \mathcal{F}$ である。最後に，何も起きていないことを表す空事象 $\varnothing$ と，何かが起きたことを表す全事象 Ω も $\mathcal{F}$ に含めることにすると，$\mathcal{F}$ は有限加法族である。合併に関する条件を可算個の事象 A_i の合併 $\bigcup_i A_i \in \mathcal{F}$ まで認めることにすると，$\mathcal{F}$ は σ-代数になる。

二つの事象 $A, B \in \mathcal{F}$ の間に $A \subset B$ が成り立つとき，A は B よりも多くの情報を持つと言える。この包含関係により，$\mathcal{F}$ には順序構造が入る。事象の自己情報量 $I(A) := -\log P(A)$ は，この順序関係の準同型（$A \subset B \Rightarrow I(A) \geqslant I(B)$）になっている。ただし一般に包含関係は半順序だが，事象の体積に基づく自己情報量の大小関係は包含関係がなくても定義できる全順序なので，必ずしも同型にはならない。

マルチンゲール理論では，部分 σ-代数の包含関係 $\mathcal{G} \subset \mathcal{F}$ に対して，$\mathcal{F}$ は $\mathcal{G}$ よりも多くの情報を持つと言う。σ-代数は考えうる事象を全て集めたものなので，部分族である $\mathcal{G}$ の方が想定が甘く，$\mathcal{G}$ の事象から期待される情報は少ないということである。確率変数 X の Shannon 情報量

ブレットの理論は多重解像度解析を経てフレーム理論，アトム分解，基底学習へと発展した。また抽象調和解析も関数を通じて表現論として発達した。

[7]現代数学において，関数解析，実解析，調和解析の区別は曖昧だが，古典的には Hilbert 空間と線形作用素の理論，Euclid 空間上の実数値関数の理論，Fourier 級数の理論のことである。

$H[X] := \mathbb{E}[-\log P(X)]$ は，この順序関係の準同型になっている。つまり，$Y := \mathbb{E}[X|\mathcal{G}]$ を $\mathcal{G}$ による X の条件付き期待値とすると，$H[X] \geqslant H[Y]$ である。

このように，Shannon 情報は事象の生起に纏わる情報であり，σ-代数とつながりが深い。一方で，Shannon 情報や σ-代数では，「一を聞いて十を知る」というような，複数の事実から演繹して得られる新たな情報は，基本的にはカウントしない。勿論，σ-代数においても，$A \subset B$ のような包含関係がある場合には，$A \Rightarrow B$ が成り立つ。しかし，情報の意味を考えれば当然 $A \Rightarrow B$ が言える時であっても，$A \subset B$ という関係がない限り，A の生起が分かったからといって B の生起まで分かるということは保証されないのである。

B.1.6 情報の演繹と計算

Brillouin (1962, § 19) に曰く，「計算機はけっして新しい情報を創造しないが，既知の情報の価値ある変換を遂行する。」データ処理不等式を思い出せば，Brillouin が言わんとすることは尤もである。つまり，事象 X についての情報 Y が得られたとして，Y を計算機に入力して得られた出力を Z とすると，どのような Z であっても $I[X;Y] \geqslant I[X;Z]$ が成り立つので，「けっして新しい情報を創造しない」ことが分かる。

それでは，次のような例はどうだろうか。例えば，ヒエログリフが読めない現代人にとって，ロゼッタ・ストーンから情報を得ることはできない。しかし，おそらく古代エジプト人であれば，容易にロゼッタ・ストーンから情報を得ることができるだろう。同じ情報を目にしているはずなのに，両者の違いはどこに生じるのだろうか。ヒエログリフの内容を情報 Y として，古代エジプト人に翻訳してもらった内容を Z とすれば，翻訳文 Z は原文 Y よりも遥かに「情報量が多い」のではないか。

データ処理不等式の教えるところによれば，答えは否である。この場合であっても，Z は Y を元にして得られた以上，$I[X;Y] \geqslant I[X;Z]$ が成り立つ。ロゼッタ・ストーンから得られる情報 Y とは，ヒエログリフという文字の系列であって，その意味内容ではない。そして，ロゼッタ・ストーンを翻訳したことで増えたように感じられたものは，情報ではなく，情報の意味や価値なのである。このように，計算機は「既知の情報の価値ある変換を遂行する。」

σ-代数としての構造からも分かる通り，Shannon の情報理論では，複

数の情報から三段論法によって得られる情報を，情報としてカウントしない。プライバシーの理論では演繹に相当する情報まで考慮することがあるが，現時点では極めて限定的である。

B.1.7 情報の意味と価値

Shannon の情報理論では，公理的確率論を用いて，情報の意味や価値という側面が慎重かつ巧妙に捨象されている。Wiever は次のような警句さえ述べている:「通信理論において，情報という言葉は特別な意味で用いられており，それを日常的な用法と混同してはならない。特に，情報を意味と混同してはならない。」勿論，ここで一般的な意味論を展開するつもりはない。しかし，確率構造がもう少し複雑になれば，すぐに情報の意味や価値に相当するものが現れる。

意思決定

例えば，Markov 決定過程（Markov decision process; MDP）のように，行動 a に応じて状態 s が動的に変化し，同時に報酬 r を受け取る構造を想定する。MDP は確率空間を内包するので，Shannon の情報理論が展開できる。そして次のように意味や価値を見出せる。まず，MDP において意思決定者の Markov 氏は，手元の状態（情報）s に基いて行動 a を選択する。このとき選ばれた行動 a は，Markov 氏にとっての情報 s の意味とみなせる。そして，S が受け取った報酬 r は情報の価値である。このような構造はロボットの制御から経済学までいたるところで見出せる。

ゲーム理論

ゲーム理論の主題は，ゲームに勝つ（あるいは負けを最低限に抑える）ための戦略である。ゲーム理論において情報とは，プレイヤーが戦局を見極め，行動を選択するための材料である。従って，MDP の例と同様に，選択された行動と，行動の結果得られる報酬は，行動選択の元になった情報の意味と価値に対応する[8]。実は，価値に着目することで，測度論に依らない確率論が構成できる。これをゲーム理論的確率論 (Shafer and Vovk,

[8]MDP は一人ゲームとして定式化できるので，この相似関係は半ば必然的である。

2001) という[9]。つまり極端に言えば，情報理論は測度論を用いずに，ゲーム理論のみを使って構成しうるのである。

さらに，ゲーム理論では複数のプレイヤーが登場するので，「「誰が何を知っているか」について誰が何を知っているか」などの，組み合わせ的な情報構造が現れる。例えば完備情報ゲームとは，全てのプレイヤーが，ゲームのルール，誰が参加しているか，各プレイヤーの取りうる行動，それらの行動に伴う利得を全て知っていることが保証されたゲームである。あるいは完全情報ゲームとは，全てのプレイヤーが，ゲームの過去の展開を全て知っていることが保証されたゲームである。また経済学において，市場において売り手と買い手が保有する情報に差がある状態は，情報の非対称性と呼ばれる。

このように，ゲーム理論は価値を主とする情報の理論である。古典統計学や Shannon の情報理論では専ら「正しい情報」を抜き出すことが主題だが，今日では暗号理論やプライバシーの理論のように情報を隠蔽したり，歪めたり，匿名化したりすることも主題となる。このような問題にはゲーム理論的な定式化が自然である。また機械学習においても，強化学習や GAN (Goodfellow et al., 2014) などは，確率論ではなくゲーム理論的な定式化によって成功した好例である。

B.1.8 情報理論小史

情報（information）の現代的な意味は，Shannon (1948) の情報理論に負うところが大きい。もっとも，情報という日本語や information という英単語は Shannon 以前からあり，現在に至るまで複数の意味で使われている。日本語の情報は元々，機密情報など，何らかの価値をもった情報のことを主に指していたとされる (小野厚夫, 2005; 高橋秀俊, 1952)。情報という概念は，20 世紀を通じた情報技術の発展とともに変化してきた。

Shannon の情報理論は，形式的には確率論と調和解析のうえに展開されているが，その思想背景に目を向ければ通信技術や熱力学の発展だけ

[9] Kolmogorov 流の公理的確率論とゲーム理論との対立は，Fermat と Pascal の往復書簡（賭博の理論）にまでルーツを辿れる。賭け金の分配（確率変数）の公平性に着目する Pascal の考え方をゲーム理論的，カードの組合せの数（確率分布）に着目する Fermat の考え方を確率論的とみなせる。賭け金に着目する「公平な賭け」の考え方は von Mises の確率論やマルチンゲール理論，そしてゲーム理論的確率論でも採用されている。もっとも，ゲーム理論的確率論の創始者の 1 人である Vovk は Kolmogorov の弟子であり，Kolmogorov 自身，ランダムネスのモデルとして公理的確率論には必ずしも満足していなかったと言われている。

でなく，統計学や逆問題などの帰納哲学からも影響を受けている。そういう意味では，情報理論が開花する土壌は 19 世紀から少しずつ醸成されていた。Shannon は通信路を舞台として情報理論を展開したが，メッセージを記号列として抽象化する考え方は 19 世紀末の Boole 代数や Peirce の記号論に根差している。実際，Shannon の修士論文 (Shannon, 1940) は電気回路が Boole 代数とみなせることを指摘したものであって，Shannon は早くから記号論に造詣が深かったことが分かる。20 世紀初頭には Nykist の信号理論や Hartley の通信理論などが Shannon の情報理論に先駆けて展開されており，それぞれ Shannon の標本化定理や Shannon 情報量の原型となった。また Shannon 情報量の元となる，エントロピー増大の法則を “証明” した H 定理 (Boltzmann, 1872) や，Shannon 情報量と関係の深い Fisher 情報量 (Edgeworth, 1908) なども，20 世紀前半に研究が進んだ概念である。

Vapnik (2006, Ch.4) に言わせれば，Kolmogorov, Fisher, Popper が活躍した “The great 1930s” は，60 年代，90 年代と並んで帰納推論に革命的な進展のあった時代である。まず現代のような形式的な数学としての確率論は Kolmogorov (1933) による確率論の公理化にはじまり，1940 年代までに Wiener による確率過程論，Ito の確率積分，Cramér の大偏差原理などが出揃う[10]。一方，統計学は帰納の哲学なので批判や論争は免れないが，Fisher を中心とする古典統計学の体系化は (Fisher, 1922) にはじまる。Popper と Carnap による科学哲学論争と並行して，確率の解釈（von Mises のコレクティフ，Keynes の論理確率，Ramsey らの主観確率）が出揃ったのもこの時期である。このような機運の高まりを受けつつ，終戦後ついに天才 Shannon が登場し，鋭い洞察力でもって情報の体系化を果たしたのである。

情報理論が登場した 1940 年代から 50 年代にかけては 3C[11] による情報革命の時代である。すなわち，Shannon の通信理論（Communication）と並び，Wiener のサイバネティクス（Cybernetics/Control），Turing や von Neumann による計算理論（Computer）が一斉に開花したルネサンス的時代である。当時の熱狂と混乱ぶりは Brillouin (1962) のトピックの多様さからも感じられる。1946 年に世界初のコンピュータである ENIAC が登場し，1950 年には既に Shannon がチェスのプログラムを題材にして世

[10] なお確率論の現代化（公理化）は解析学の中では最後発であり，当時既に関数解析（Hilbert 空間論，線形作用素論）や調和解析（表現論）などは完成しつつあり，また 1931 年には Hilbert の夢を終焉させた不完全性定理が発表されている。

[11] (梅垣壽春 et al., 1983) の序による

界初の AI 研究とされる論文を書いている。このような技術的躍進の背景には第二次世界大戦があり，多くの情報技術が戦時中に開発され，終戦を待って公開されたのである。オペレーションズ・リサーチもまたそのような技術の一つである。後の冷戦時代にはゲーム理論が注目されたこともある。人工知能研究の原点とされるダートマス会議が開催されたのは 1956 年のことである。ニューラルネットの原型である McCulloch and Pitts (1943) の形式ニューロンや，Rosenblatt (1958) のパーセプトロンもこの時期に登場した。バイオインフォマティクスの舞台となる分子生物学も同時期に開花した。Avery (1944) によって遺伝子の正体が DNA であることがほぼ明らかとなり，Watson (1953) によって DNA の二重らせん構造が示された。続く 1960 年代は Fano (1961) をして既に「情報理論の研究は確立され，残るは実用化研究のみで，おそらく障害物もないであろう」と言わしめた。実際，情報理論の本流と目される符号理論や通信理論にとって，60 年代から 70 年代は冬の時代であった。60 年代末には (Minsky and Papert, 1969) の中で線形パーセプトロンの限界が示され，第一次ニューラルネットブームも終焉を迎えた。

一方，情報理論の周辺に目を向ければ，60 年代は様々な新しい科学が興った時代である。まず，Kolmogorov (1963) の悲願であるアルゴリズム複雑性が完成し，Solomonoff (1964) がアルゴリズム確率を定義するなど，圧縮不可能性の理論が展開した。そして，Lindley や Savage, Good らがベイジアン論争を展開した。逆問題では Tikhonov (1963) の正則化法が成功を収めたことで，方程式から最適化問題へのパラダイム・シフトが起きた。制御理論でも Kálmán (1960) による状態空間モデルが登場し，現代制御理論の道が拓けた。Vapnik-Chervonenkis 理論の土台となる Hoeffding (1963) の確率不等式や Kolmogorov (1956b) の ε-エントロピー（容量，metric entropy），Dudley らの確率的一様収束など，関数空間上の確率論もこの時期に発達した。Bellman (1961) は制御理論の文脈で「次元の呪い（curse of dimensionality）」という言葉を提唱した。加えて，Chomsky の生成文法や，Tarski や Montague による形式意味論など，新しい科学が成立した時代でもある。分子生物学では mRNA が発見され，セントラルドグマが定着した。人工知能では Robinson が定理の自動証明のための導出原理を創案した。これらと関連して，Raiffa and Schlaifer (1961) は情報の価値について論じ，Arrow (1963) や Akerlof (1970) は市場における情報の非対称性を論じ，Bar-Hillel (1964) や MacKay (1969) は意味論的情報理論を展開した。

1970 年代後半から 80 年代にかけては，POS やパーソナルコンピューターの登場によって多変量解析と時系列解析が大衆化した時代である。統計学では，60 年代までのイデオロギー論争からは一変して，Nelder and Wedderburn (1972) による一般化線形モデルや Akaike (1973) による AIC，Rubin (1974) による因果推論，Efron (1979) によるブートストラップ法，Huber (1981) のロバスト統計学，Andersen and Gill (1982) によるマルチンゲール統計学，Geman and Geman (1984) に始まる MCMC，Pearl (1988) のベイジアンネットワークなど，新しい分野が次々と登場した。情報理論でも，多元情報理論，Ziv and Lempel (1977) のユニバーサル符号化，Rissanen (1978) の MDL，Csiszár and Körner (1981) のタイプ理論，Amari (1985) の情報幾何学など，枚挙に暇がない。信号処理では Morlet et al. (1982) や Grossmann and Morlet (1984) を先駆として，Daubechies (1988) のフレーム理論，Mallat (1989) の多重解像度解析に至るウェーブレットの理論が急激に発達した。スパース信号処理もこの時期に同時多発的に発見された。長い歴史をもつ関数近似理論もこの時期には de Boor が B-spline を普及させたほか，1968 年には関数近似理論の専門誌である JAT が創刊した。集中不等式 (McDiarmid, 1989; Hoeffding, 1963; Azuma, 1967; Bernstein, 1924) の新時代が始まったのもこの時期である。人工知能では Feigenbaum (1977) の知識工学によって「知識の時代」に入り，Minsky (1975) のフレーム理論がその中核におかれた。Solomonoff のアルゴリズム学習理論が最盛期を迎え，Vapnik and Chervonenkis (1971); Sauer (1972); Shelah (1972) の VC 理論，Valiant (1984) の PAC 学習によってアルゴリズムの学習可能性を評価する計算論的学習理論が始まった。分子生物学では遺伝子組換え技術が発達し，1990 年にはヒトゲノム計画が始まる。これはバイオインフォマティクスの本格始動といえる。1973 年には金融派生商品のモデルとして Black-Scholes 方程式が発表され，金融工学が始まった。Rumelhart et al. (1986) が三層パーセプトロンのバックプロパゲーションを再発見し，第二次ニューラルネットブームが始まるのもこの時代である。

1990 年代は情報爆発が顕著になった時代である。インターネットの普及だけでなく，バイオインフォマティクスの興隆も背景にある。KDD（Knowledge Discovery in Databases）, IoT（Internet of Things）という言葉が普及した。情報の信頼度を見定め，情報を取捨選択するための，情報リテラシーという考え方もこの時期に登場した。演繹的な記号操作を主とする人工知能の勢いが弱まり，データからの帰納推論を主とする機

械学習やデータマイニングが発達していった。機械学習やデータマイニングは，伝統的な推測統計学とは目的を異にする。つまり，伝統的な推測統計学では，唯一無二の真の構造があることを前提としてモデルを設定し，推定されたモデルの真贋を吟味して現象自体の理解を図るのに対し，機械学習やデータマイニングでは，真のモデルの存在は仮定せずに，予測精度の高いものや，新たな知識獲得に貢献するものを良いモデルと考えるのである。Vapnik (2006) によれば，サポートベクターマシン (Cortes and Vapnik, 1995) や統計的学習理論は，ニューラルネットの動作原理を追究する過程で誕生したという。現在用いられている多くの非線形凸最適化法も 90 年代に発達した。Rubinstein et al. (2010) は，信号処理もまたこの時期の機械学習の影響を強く受けたと述べている。スパース表現 (Olshausen and Field, 1997; Chen et al., 1998; Tibshirani, 1996) や辞書学習 (Mallat and Zhang, 1993) は信号処理が演繹的な調和解析から帰納的な機械学習へとパラダイム・シフトを果たした結果とみなせる。

2000 年代には，ドットコム・バブルが崩壊する一方で，インターネットの常時接続が普及し，Web2.0 と呼ばれる時代になった。例えば Google や amazon, Wikipedia, YouTube, Facebook などが登場し，情報を発信するコストが劇的に低下した。スマートフォンが人々の日常を一変させたことも記憶に新しい。大量生産大量消費から多品種少量生産へと移行したことで，サンプルサイズ N よりも説明変数の数 P の方が圧倒的に多い問題が散見されるようになった。これを計算量理論の NP 困難にひっかけて，「新 NP 問題」と呼ぶことがある。アルゴリズム取引に代表される超高頻度データが登場し，数理ファイナンスの発展を背景として予測可能性に基づくゲーム理論的確率論 (Shafer and Vovk, 2001) も登場した。クラウドソーシングやヒューマンコンピュテーション (von Ahn, 2005) のように，インターネットを介した情報処理が現実のものとなり，データだけでなくノイズの種類も一層多様化した。2003 年にはヒトゲノム計画が完了し，バイオインフォマティクスは高次の遺伝情報を扱うポストゲノム時代に入った。これらの情報爆発を受けて，2010 年代にはビッグデータやデータサイエンティストという言葉も登場する。今日では，プライバシーの理論 (Dwork, 2006) や，忘れられる権利などの情報倫理，またそれらを包括する Floridi (2010) の情報哲学（Philosophy of Information; PI）なども展開している。

B.2 データ表現の観点

機械学習手法の性能は，データの表現に大きく依存する。データの表現は，属性（attribute）や特徴量（feature）とも呼ばれる。言うまでもなく，画像処理や音声信号処理，自然言語処理などの技術者は，データ表現すなわち特徴量の設計に心血を注いでいる。Bengio et al. (2013a, §1) によれば，表現学習（representation learning）では，判別や予測に有用な情報が容易に取り出せるようなデータ表現をデータから学習することを目指している。Box-Cox 変換に代表される変数変換や，PCA に代表される次元削減は，古典的な表現学習とみなせる。これらは教師なし学習に分類される。これに対し，一般化線形モデルや k-近傍法に代表されるノンパラメトリック回帰は，教師ありの表現学習とみなせる。

本節においてデータの表現とは，データを計算可能な形式に変換する決定的または確率的な法則とする。まず，表現とは確率変数または写像であって，その実現値や像のことではないので注意せよ。そして，データを計算機で処理するという大前提により，データの表現は計算機で扱えるものに制限した。また，写像を表現と呼ぶ方法は，調和解析や表現論での用語に倣った。何かを表現するには媒体が必要である。例えば自分の位置を知らせるには，座標系を固定したうえで座標を知らせれば良い。あるいは特定のベクトルについて言及するには，基底を定めて係数を読めば良い。このとき，座標系や基底は媒体であり，自分の位置を座標に変換する手続きや，ベクトルを係数に変換する写像が，表現である。

図 B.5 にデータ表現の観点を示す。良いデータ表現の要件は何だろうか。ノーフリーランチ定理が示唆する通り，全てのタスクに有利な表現はない。データ表現は計算可能かつ単射であれば，十分に思われる。実際には，全射性の方が重要な場合もある。確率論や統計学では標本空間は単に集合であったが，信号処理やデータ解析，パターン認識，機械学習などの応用分野では，標本空間は信号や群，グラフ，文書，学習機械などの構造を持った対象の族である。従ってデータ表現はこれらの構造と両立（compatible）した準同型であることが望ましい。

B.2.1 計算可能性

データ表現は計算可能でなければ，計算機で扱うことはできない。

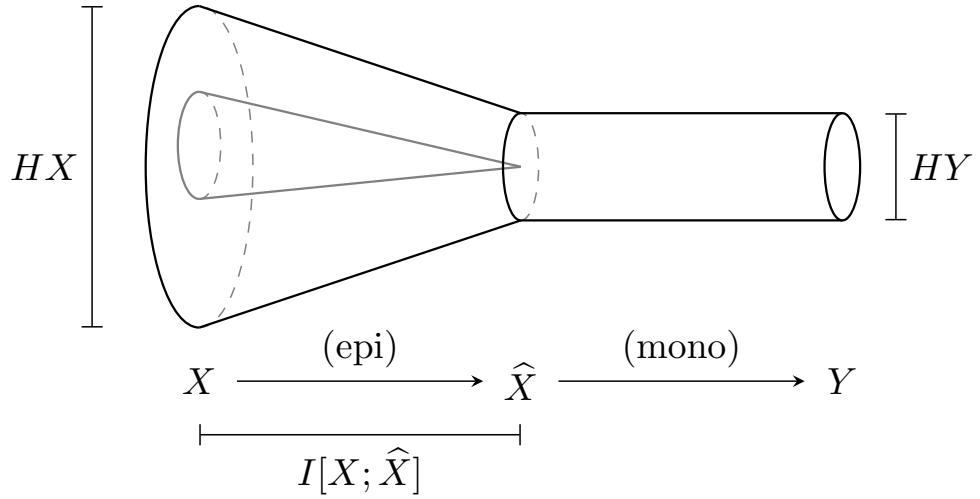


図 B.5: データ表現の観点

B.2.2 単射性，忠実性，モノ

データ表現が写像 $\widehat{X} = \psi(X)$ で与えられる場合，ψ が単射であれば元の X を特定できる。例えば帯域制限された信号 X を標本化する場合，ψ はサンプリング作用素であり，サンプリング周波数が Nykist 周波数の二倍を上回っている限り，単射である。逆に，単射でない場合には元の信号は損なわれる。例えばサンプリング周波数が足りない場合には，元の信号を特定できなくなる。積分変換の反転公式は，ψ として積分変換をとった場合に ψ が単射であるための条件を示している。

データ表現が一般の確率変数 $\widehat{X}$ の場合，単射性に相当する概念は忠実性である。忠実性は相互情報量 $I[X;\widehat{X}]$ によって測ることができる。レート歪み理論では相互情報量に基いて連続情報源の最適な量子化を議論する。ICA や一部の表現学習では $I[X;\widehat{X}]$ を最大化（infomax）することで適切な表現を獲得する。

B.2.3 全射性，十分性，エピ

あえて情報を落とすことが得策な場合がある。例えば画像認識においては各ピクセルの輝度値を全て区別できる表現よりも，同じクラスラベル同士では同じ表現になる方が，続く判別ステップは簡単になる。つまり，確率変数 X を別の確率変数 Y に変換するタスクでは，Y を特定する情報

量を損なわない範囲 $I[Y;X] = I[Y;\hat{X}]$ で忠実性を落とすほうが良い。このようにして得られる概念は，十分統計量である。

忠実性を犠牲にすべき場面は他にもある。例えば計算量やメモリコストを減らすためには符号化器は小さいほうが良い（MDL 原理）。また高次元データを扱う場合にも，次元削減をすることはよくある。ノイズに対するロバスト性を上げるためには表現空間を小さくするほうが良い（バイアス・バリアンス分解）。一般に，可能な限り小さなモデルを選ぶことをオッカムの剃刀原理といい，モデルの大きさは様々な複雑性（complexity）で測られる。

B.2.4 準同型性，合目的性

表現はデータの構造 $\mathbf{S}$ を反映していることが望ましい[12]。従って，表現 $\hat{X}$ の良し悪しは定義域 X との相互情報量 $I[X;\hat{X}]$ だけでなくタスク $X \to Y$ に応じても変わる。表現 $\hat{X}$ が構造 $\mathbf{S}$ をどの程度保存しているかを測る量 $C[\hat{X}]$ があれば，$I[X;\hat{X}]$ と合わせて考慮すべきである。例えば高次元データの統計的推定を行うのであれば，次元削減を検討することはよくある。このとき削減後のデータの次元は，$\mathbf{S}$ として線形構造をとった場合の $C[\hat{X}]$ の例である。計算機上では，画像 X はバイナリベクトル $\tilde{X}$ として表現されているが，Fisher ベクトルなどの特徴量ベクトル $\hat{X}$ は，元のデータ X の画像としての構造 $\mathbf{S}$ を再現すべく，創意工夫を重ねて作られている。畳み込みネットワークでは，$\tilde{X}$ から，構造 $\mathbf{S}$ を強調した特徴量 $\hat{X}$ を自動的に作成することが期待されている。

B.2.5 統計的性質

データ表現は不偏性や有効性，頑健性などの推定量としての統計的性質においても優れているに越したことはない。例えば Vincent et al. (2010) は，単に目的のタスクにおける性能の善し悪しだけではなく，学習の速さも重要であるという。これは人間が初めて見る問題に対してすぐに適応できるのに対し，現行の機械学習では大量のデータと反復を必要とすることのギャップを念頭に置いている。

[12] ただし，ほとんどの数学的構造（距離や測度，線形性，多様体構造など）は形式的には導入できるので，タスクに必要なものに限定すべきである。

B.3 複雑性の測り方

情報と同様に，複雑性（complexity, 複雑さ，複雑度）もまた 20 世紀を通じて多様な文脈のもとで定義が検討されてきた。

以下で扱う複雑性は，二種類に分けられる。すなわち，ある対象 X 自体の複雑さを測るものと，その対象 X が属するクラス $\mathcal{C}$ の大きさを測るものである。

前者の例は，特定の確率分布に対するエントロピーや，特定の系列の最小記述長などがある。一方，後者の例は，関数空間の位相や，統計モデルのパラメータ数などがある。もっとも，前者は各点評価ということなので，sup や inf を用いて一様評価にして，クラスの複雑性に拡張できることも多い。例えばアルゴリズムの計算量や関数ノルムなどは，特定のアルゴリズムや関数に対して計算することもできるし，アルゴリズムの族や関数族の中で sup をとったものをクラスの複雑性とみなすこともできる。しかし，位相やパラメータ数のように，対象の族に対してしか定義できない複雑性もある。

本節では特に，モデル選択のための複雑性に興味がある。学習理論の成果によって，汎化誤差はモデル複雑性によって評価できることが分かっている。また，AIC に代表される情報量規準の罰則項は，モデル複雑性を反映していることが多い。直観的には，仮説空間が大きいと，それだけ一点を特定する問題は難しくなるので，モデル複雑性はモデル選択問題の難しさを反映している。従って，ニューラルネットの ‘複雑性’ を評価したという場合には，クラスの複雑性の意味であることが多い。一方，MDL 原理のように，個別のモデルの複雑性に基いてモデル選択を行うこともある。

B.3.1 クラスの複雑性

関数ノルム

伝統的に, 関数空間の表現能力は関数ノルムを用いて測る。例えば Sobolev ノルムの指数は関数の滑らかさを表す。関数は滑らかであるほど ‘少なく’ なる[13]ので，Sobolev 指数は Sobolev 空間の大きさを反映している。また，ニューラルネットの万能関数近似能力を示す場合は，一様ノルムや L^2 ノルムを用いるが，これはニューラルネットの空間が連続関数や連続関数

[13]少ないとはいえ非可算無限個の元があるので，この表現は便宜的に用いている。

や L^2 関数の空間と同等の大きさをもつことを示す意図がある。数値解析では精度の観点から一様ノルムが好まれる。一方，信号処理では周波数解析を行うために L^2 ノルムが好まれる。Kůrková (2012) はニューラルネットの中間層素子の族 G の複雑性として G-variation を導入し，離散化に伴う近似誤差を評価した。フレーム理論では任意の関数を完全再構成するための離散化の良さを測るために L^2 ノルムを用いる。これはフレームの複雑性とみなせる。

正則化項

最適化問題において制約条件は解空間の複雑性を制御するための条件である。Lagrange 未定乗数法により，制約付きの最適化問題は，正則化付き目的関数の無制約最適化問題に変換できる。従って，正則化項は解空間の複雑性を表している。例えば L^p 正則化項は線形モデルの関数ノルムであるから，モデルの複雑性を図っているとみなせる。逆に，目的関数から解しか含まない項を分離すると，それは正則化項とみなせる。例えば AIC の罰則項などは KL 情報量の推定値から対数尤度を差し引いた残りとみなすことができる。

モデル複雑性

学習理論により，汎化誤差はサンプル複雑性や仮説空間の複雑性を用いて制御できることが分かっている。Valiant は，仮説空間 $\mathcal{H}$ の濃度 $|\mathcal{H}|$ を用いて汎化誤差を評価した。Vapnik は仮説空間の VC 次元を用いて Valiant の評価を改善した。Rademacher 複雑性を用いると，VC 理論の一様性の仮定をゆるめた設定で汎化誤差を評価できる。

計算複雑性

計算理論の目的は，計算機（あるいは言語や論理体系）の計算の性質のなかで，実装に依存しない性質を見出すことである。例えば，古典的な計算モデルである有限オートマトン，文脈依存文法，Turing 機械は，互いに表現力の異なるクラスであることが分かっている。また，Church-Turing のテーゼにより，アルゴリズムは実装に依存しない数学的対象であることが分かっている。アルゴリズムをもつ関数を計算可能関数と呼ぶ。

計算可能関数は Blum の公理に基づく複雑性によって分類できる。そのような複雑性の典型例が時間計算量および空間計算量であり，いわゆる P や NP などのクラスを導く。これらのクラスはもちろん，計算機の実装に依存しない性質である。

B.3.2 個別の対象の複雑性

アルゴリズム複雑性

情報源符号化定理により，エントロピーは符号長という実体的な意味をもつ。つまり，圧縮の限界値によって情報量が定義できるのである。この観点を敷衍して符号化器自体を符号化したものは，アルゴリズムの複雑性とみなせる。アルゴリズム複雑性は，Occum の剃刀にもとづくモデル選択に利用できる。歴史的には Solomonoff, Kolmogorov, Chaitin が先駆となり，Rissanen が MDL から確率的複雑性（stochastic complexity）にいたる一連の仕事をした。

文字列 X の計算機 U における Kolmogorov 複雑性 $K_U(X)$ とは，文字列 X を出力する計算機 U のプログラム π の記述長の中で最も短いもの $\min_{U(\pi)=x} l(\pi)$ のことである。任意の万能チューリングマシンは同等な表現能力を持つことを示す Church の定立により，K_U は存在すれば定数の差を除いて全ての計算機で一意に定まることが言えるので，この定義は意味を為す。Kolmogorov 複雑性は，Kolmogorov がランダムネスの定義を模索する上でたどり着いた概念であり，文字列 X の生起確率 P を仮定せずに定義されている。文字列 X が iid の場合には，エントロピーに確率収束する。逆に，符号長が Kolmogorov 複雑性と同値になるような X の確率測度をユニバーサル確率という。

文字列 X を生成する確率分布の推定量を $\widehat{P}$ とする。複数の推定量が考えられる場合には，最も「シンプル」な推定量を採用するのが Occum の剃刀の教えるところである。Rissanen は，モデルのシンプルさを測る量として，推定量 $\widehat{P}$ 自体の記述長 $l(\widehat{P})$ と，$\widehat{P}$ による X の記述長 $l(X;\widehat{P})$ の合算である MDL を採用した。正確には，これは二段階符号化（coding, compression）と呼ばれる MDL の一形態であり，このままでは適用範囲が限られる。Rissanen は，MDL が最尤符号長 $\widehat{p}(\mathcal{D}|\mathcal{M})$ に一致することを利用して，MDL を一般化を展開した。一般化された MDL を確率的複雑性（stochastic complexity）と呼ぶ。適当な設定のもと，MDL は汎化誤

差と近似誤差の合算を評価していることになり，また Kolmogorov 複雑性や BIC と近似的に等価であることが示される。

ランダムネス

ランダムネスは計算可能な記号列に対して定義される。ランダムネスを定義する試みは von Mises (1919) にはじまる。von Mises は確率論を定式化するためにランダムネスの定義を考察していた。今日，標準的と目されているのは Martin-Löf (1966) による定義である。Martin-Löf ランダムネスは「わずかな列しか持たないような特別な性質を持たない」という典型性についての考察から出発して，典型性の確率論的な検定（test）を設定し，検定に合格（pass）したものをランダムとみなす考え方である。Martin-Löf ランダムネスは，Kolmogorov 複雑性（圧縮不可能性）に基づくランダムネスや，マルチンゲール（予測不可能性）に基づくランダムネスと同値であることが分かっている。実数列に対するランダムネスや，位相空間におけるランダムネスの定義は今日もなお議論が続けられている。

カオス

Poincaré が創始した力学系の理論では，決定的でありながら予測不可能な系をカオスと呼ぶ。今日に至るまでカオスの定義は定まっていないが，しばしば筆頭に挙げられるのが初期値鋭敏性である。初期値鋭敏性は Lyapunov 指数や Kolmogorov-Sinai エントロピー，あるいは位相エントロピーで測るのが古典的である。力学系は決定論的な対象でありながら，確率論的な概念であるエントロピーが定義できるというのは驚くべきことである。カオス理論は複雑系科学として発展し，フラクタル次元など多くの複雑性の指標が開発された。

B.4 モデル選択の考え方

ニューラルネットのアーキテクチャは，学習に先立って使用者が指定するメタパラメータである。アーキテクチャは汎化誤差がもっとも小さくなるモデルを選択するのが基本的な考え方である。

B.4.1 バイアス・バリアンス分解

回帰分析や教師あり学習では，平均二乗予測誤差に基づいてモデルを推定し，評価する。汎化誤差のバイアス・バリアンス分解は回帰分析のRSS にルーツを持つ古典的な考え方である

$$\mathbb{E}|Y - \widehat{f}(X)|^2 = \mathsf{bias}\,[\widehat{f}]^2 + \mathsf{var}\,[\widehat{f}] + \sigma^2.$$

一般に，仮説空間が大きくなると，バイアスは低減し，バリアンスは増大するので，バイアスとバリアンスはトレードオフ関係にある。ここで，仮説空間の大きさは，モデルがパラメトリックな場合にはパラメータ数，ノンパラメトリックの場合には有効パラメータ数や VC 次元などを用いる。従って，バイアスとバリアンスの総和である汎化誤差を最小化するモデルを選択せよというのが，バイアス・バリアンス分解に基づくモデル選択の考え方である。

バイアスを近似誤差，バリアンスを推定誤差と呼ぶこともある。近似誤差とは，「真の構造」をモデル化したときに生じるモデル化誤差のことである。関数近似理論などでは近似対象の関数を所与とするので近似誤差を評価できることもあるが，機械学習や統計的推定の場合にはデータの生成構造は未知とするので近似誤差は「本質的に減らすことのできない」量として無視される。一方，推定誤差とは，統計的推定に基づく誤差である。こちらは統計的な正則条件を満たす理想的な場合には Cramér-Rao のような定理が成り立つので，オーダーが評価できる。

Bottou and Bousquet (2008) は，近似誤差と推定誤差に加えて，最適化誤差を含めるべきだと主張する。複雑な学習モデルをビッグデータに適用する場合，学習アルゴリズムの計算量は無視することができず，アルゴリズムの収束条件を満たす前に計算を打ち切ることが頻繁に行われる。この打ち切りによって生じる誤差は，近似誤差でもなければ，推定誤差でもない，アルゴリズムの打ち切りによって生じる誤差である。これを最適化誤差と呼ぶ。学習問題をバイアス・バリアンスに最適化誤差も含めた三者のトレードオフ関係とみなすと，統計的な不偏性を損なってでも評価関数の凸性や正則性を優先するほうが，結果として得られる解の汎化誤差が小さくなるケースも報告されている。

B.4.2 統計的モデル選択

多変量解析や時系列解析において，統計モデルを選択する方法には大きく三つの観点がある。(1) 当てはまりの良さ（goodness of fit）ないし説明力に着目する場合と，(2) 予測の良さ（goodness of prediction）に着目する場合，そして (3) モデルの事後確率に着目する場合である。歴史的には，(1) から (3) の順番で登場した。もっとも，これらの観点は，解析の目的や手法，あるいは解析者の科学観に根ざすものであって，どの観点が絶対的に優れているというものではない。

当てはまりの良さとは，複数のモデルを最尤推定した場合にモデル同士の対数尤度を比較する方法や，モデルパラメータが零であることを帰無仮説として仮説検定にかける方法である。K. Pearson の χ^2-適合度検定（test for fit）に始まり，線形回帰における回帰係数の t-検定や残差（RSS）の分散分析（ANOVA），一般化線形モデル（GLM）における逸脱度（deviance）の検定などがこれに相当する。

予測の良さとは，複数のモデルを最尤推定した場合に，モデル同士の対数尤度の期待値を比較する方法であり，期待値をとる操作が「未知のデータ」に対する対数尤度を評価することに相当するので，「予測」と表現される。この観点は AIC の提唱者である赤池によって体系的に展開されたほか，Mallows' C_p, クロスバリデーション（CV）なども予測誤差を推定する方法である。AIC は尤度に拘る点で限定的だが，汎化誤差の理論である学習理論の源流の一つでもある。AIC について特筆すべき点は，AIC はしばしば真のモデルよりも小さいモデルを選択することになるが，赤池はこれを良しとしたことである。すなわち，真のモデルと一致することよりも予測能力を重視したのである。

一方，モデルの事後確率とは，複数のモデルを Bayes 推定した場合に，モデル同士の事後確率（Bayes 因子）を比較する方法である。Schwartz の BIC が典型的だが，これは情報理論的なモデル選択規準である MDL と漸近的に一致することが知られている。Bayes 推定における Bayes 因子は最尤推定における尤度比検定に相当するが，予測分布に用いて予測誤差を比較する方法（ABIC）もある。

B.4.3 Occum の剃刀

汎化誤差ではなく，モデルの複雑性自体を主として，もっとも単純なモデルを選択するという考え方もある。汎化誤差の選び方にも恣意性は

残るので，絶対的にどちらが正しいといえるものではない。例えば物理学は，シンプルな方が良いという考え方を突き詰めた結果，成功した例といえる。モデル複雑性の測り方には，MDL や VC 次元が挙げられる。MDL は Kolmogorov 複雑性と等価であることが知られているので，モデル空間の測度に由来する複雑性である。一方，VC 次元はメトリックエントロピーと等価であることが知られているので，モデル空間の位相に由来する複雑性である。従って，少なくとも MDL と VC 次元は相異なる尺度であることが推察される。VC 理論に対する批判にもある通り，モデル複雑性は「何に使うためのモデルか」という観点を不問にするので，いわば教師なし学習のようなもので，予測や推定というようにタスクが明確な場合には不十分である可能性がある。

B.4.4 逆問題と正則化

逆問題の典型的な形式は Fredholm の第 1 積分方程式

$$g(x) = \int_{\Omega} k(x, y) f(y) \mathrm{d}y$$

である。すなわち，有限個のデータとして g が与えられ，変換規則 k を既知として f を求める問題である。有限個の点から関数空間の点を特定することは不可能 (ill-posed) なので，先験知識を入れる必要がある。Tikhonov (1963) は，関数空間を制限する方法として正則化が有効であることを示した。それまで，逆問題は無限次元空間の方程式として，逆作用素を近似する方針で解くことが一般的であったが，Tikhonov をきっかけとして，関数空間に汎関数を導入して最適化問題として解くことが主流となった。

カーネル法

レプレゼンター定理により，凸正則化付き経験リスク最小化問題の停留点は $\sum \alpha_j k(\cdot, x_j)$ の形式で表現できる。また正則化項 $\Omega[f]$ が適当な表現空間におけるノルム $\Omega[f] = \|Tf\|$ で与えられる場合，変換 T とカーネル k は対応している。このように，正則化理論はカーネル法との相性が良い。

Bayes 推定

本来，逆問題は統計的推測とは無関係だが，正則化項 $J[f]$ を事前分布の対数，正則化項と誤差項の和 $\lambda J[f] + E[f]$ を事後分布の対数とみなすことで，最適化問題は事後分布の最頻値を求める操作とみなせる。

B.4.5 スパース正則化

いわゆる ℓ^1-正則化によって信号を分離するテクニックは，地球科学や信号処理の分野で同時多発的に発見された。動物の視覚系がスパース信号処理をしているという発見もこの時期にあった。Donoho and Stark (1989) によれば，理論面での体系的な研究は Logan (1965) に始まる。スパースモデリングを信号処理や統計的な観点で理論付けをしたのは Chen et al. (1998) や Tibshirani (1996) である。また同時期にスパースコーディング Olshausen and Field (1997) も登場した。2000 年代には Bayes や SVM と融合し，スパース正則化の学習理論や圧縮センシング (Candès et al., 2006) が登場した。

ℓ^2-正則化と異なり，ℓ^1-正則化は座標に依存する正則化項である。これは座標変換不変性という幾何学的な観点で「美しくない」だけでなく，画像や言語データのように，個別の座標の意味が必ずしも明確でないような場合に，学習結果の解釈が難しくなるというデメリットもある。しかし，例えば信号分離のようなタスクでは，線形変換とフィルタリングという古典的なアプローチよりも高性能な情報抽出ができることも事実である。実は，スパース制約は ℓ^0-正則化の緩和と説明されることも多いが，低ランク性という，別の幾何学的な観点から導くこともできる。他の事前知識と同様に，スパース制約の妥当性はタスクに応じて検討すべき事項である。

B.4.6 学習理論

Valiant は PAC（Probably Approximately Correct）学習という枠組みで学習可能性を定義して学習理論を創始した。狭義の計算論的学習理論とは Valiant の PAC 学習を指すが，広義には Valiant に先行する Gold のアルゴリズム学習理論や，Valiant に続く Vapnik の統計的学習理論を含めた総称である。

PAC 学習と VC 次元

集合 X を可測集合，$Y \subset \mathbb{R}$ を閉集合とし，$Z := X \otimes Y$ を積空間とする。関数族 $\mathcal{H} \subset Y^X$ を仮説空間または学習モデルと呼び，関数 $\ell : \mathcal{H} \times Z \to [0, \infty)$ を損失関数と呼ぶ。Z の確率分布を D として，期待リスク関数を $R_D[h] := \mathbb{E}_D[\ell(h, z)]$ によって定義する。

仮説空間 $\mathcal{H}$ が損失関数 ℓ に対して agnostic PAC 学習可能であるとは，以下の条件を満たすアルゴリズム $A : \bigcup_{n \in \mathbb{N}} Z^n \to \mathcal{H}$ が存在することを言う。すなわち，任意の Z の確率分布 D と任意の $\varepsilon, \delta > 0$ に対し，$m(\varepsilon, \delta)$ 個以上の元からなる D の任意の iid サンプル $S = \{z_n \mid z_n \sim D\}$ に対し，少なくとも確率 $1 - \delta$ で

$$R_D[A(S)] - \inf_{h \in \mathcal{H}} R_D[h] \leqslant \varepsilon,$$

が成り立つ。$m(\varepsilon, \delta)$ の下限を $\mathcal{H}$ を学習するためのサンプル複雑性と呼び，アルゴリズム A の計算量 $T(\varepsilon, \delta)$ を計算複雑性と呼ぶ。

Valiant は 0-1 損失関数による二値判別問題の場合に，仮説空間の濃度 $|\mathcal{H}|$ が有限ならば

$$m(\varepsilon, \delta) \leqslant \log(|\mathcal{H}|/\delta)/\varepsilon$$

が成り立つことを示した。すなわち，二値判別問題の有限仮説空間は常に学習可能である。

Vapnik は仮説空間の濃度が無限大の場合には，仮説空間 $\mathcal{H}$ の VC 次元 d に対して，定数 C_1, C_2 が存在して，

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leqslant m(\varepsilon, \delta) \leqslant C_2 \frac{d + \log(1/\delta)}{\varepsilon^2},$$

が成り立つことを示した。特に，このときの学習アルゴリズムは経験リスク最小化（empirical risk minimization; ERM）で良いことが分かっている。すなわち，二値判別問題において VC 次元が有限であることと PAC 学習可能であることは同値である。さらに，二乗損失による回帰問題の場合にも，shattering 次元を用いて同様の主張が成り立つ。

PAC 学習の定義に戻ると，VC 次元を用いて汎化誤差を評価できる。

$$R_D[A(S)] - \inf_{h \in \mathcal{H}} R_D[h] \leqslant \sqrt{C_2 \frac{d + \log(1/\delta)}{m}},$$

ただしこの評価は一般に非常に緩いことが知られている。

Valiant と Vapnik による結果は，最小二乗法による回帰や 0-1 損失による二値判別など，特定の教師ありバッチ学習に限定されていた。多値判別や，スパース正則化など一般の評価関数，オンライン学習，教師なし学習などへの拡張は後進の業績であり，現在まで未解決の問題も多い。また，k-means のように VC 次元が発散して理論が適用できないことも早くから知られていた。VC 次元の有限性と PAC 学習可能性は同値なので，k-means の問題は PAC 学習の定義が強すぎることを示唆している。Kearns et al. (1997) は，データの分布に依存しない一様評価は必ずしも最適な指標ではないことを指摘している。

Rademacher 複雑性

Bartlett and Mendelson (2002) による，データの分布に依存する複雑性の一つ。一様性を放棄したことで汎化誤差の評価を改善した。

仮説空間 $\mathcal{H}$ のデータ分布 D による Rademacher 複雑性は，

$$\mathfrak{R}_m[\mathcal{H}] := \mathbb{E}_{x_m \sim D}\left[\mathbb{E}_{\xi \in \{\pm 1\}^m}\left[\sup_{f \in \mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i h(x_i)\right|\right]\right]$$

で与えられる。損失関数が有界な場合，少なくとも $1-\delta$ の確率で

$$R_D[h] \leqslant R_D[A(S)] + \mathfrak{R}_m[\mathcal{H}] + \sqrt{\frac{8\log(2/\delta)}{m}},$$

が成り立つ。Bartlett and Mendelson (2002) では，R_n を計算するための公式と，木，ニューラルネット，カーネル法に対する複雑性が計算されている。

Stability

Bousquet and Elisseeff (2002) は感度解析に基づき，stability を提案した。Stability は Rogers and Wagner (1978), Devroye and Wagner (1979a,b) による leave-one-out（LOO）誤差の評価を発展させた方法である。VC 次元は stability の上限を与えることが分かっている (Kearns and Ron, 1999)。Shalev-Shwartz et al. (2010) は learnability や strong convexity との関係まで含めて論じている。VC 理論は仮説空間の複雑性のみに着目するのに対し，Stability は学習アルゴリズムのデータに対する連続性に着目するので，例えば正則化付き最小化（regularized loss minimization; RLM）や k-近傍法など VC 理論では扱えないものが扱える。

B.5 水と油はなぜ分離するか

水と油は自発的に分離する。このような現象は**相分離**と呼ばれる。相分離現象はあたかも，エントロピー増大の法則に逆らって，自発的にエントロピーが減少しているかのようである。しかし実際には，分離系のエントロピーは，混合系のエントロピーよりも高いので，相分離においてもエントロピー増大の法則が成立している[14]。（と，考えられている。）水分子と油分子を分離させる力のように，エントロピー増大に伴う相互作用を**エントロピー的な力**（entropic force）と呼ぶ。エントロピー的な力は，物性科学や生命科学で基本的な役割を果たす作用であり，これらの分野を中心にして精力的に調べられている。また例えばミセルの相分離現象は 1920 年代に熱力学的な議論が交わされているなど，歴史も長い。相分離現象ないしエントロピー的な力を記述する物理モデルはミクロ・メソ・マクロの三つのスケールに分類できる。以下では，相分離現象の物理モデルを整理する。

B.5.1 基本的な理解

相分離において，分離系のエントロピーは，混合系のエントロピーよりも高い。モデルの各論に先立って，この仕組みを説明する。まず，疎水性溶質分子（油）の近傍で水素結合による水の構造化が起こる。この構造を iceburg structure という。Iceberg structure はエネルギー的に安定だが，水分子の回転運動が制限されるためにエントロピー的に不安定であり，系全体の Gibbs 自由エネルギーは不安定になる。そして，Iceberg structure が崩壊する過程では，疎水性溶質分子間に引力相互作用が誘起されるので，疎水性溶質分子が凝集する。ただし今日では，iceberg structure ほどの秩序が実在するかどうかは，エネルギー変化の実験値と計算値の不整合から疑問視されている。いずれにせよ，水分子の回転を束縛するような構造が崩壊することにより，水がエントロピー利得を得るために，疎水性溶質分子は秩序構造を形成するにも関わらず，系全体としてはエントロピーは増大しているという理解である。

[14] 南極の氷の下では，低温高圧条件のために混合系の方が安定となり，相分離は起きないことが知られている。

B.5.2 マクロスケールモデル

マクロスケールモデルでは，系を基本単位とする。すなわち，水と油の化学ポテンシャルに基づき，系の Gibbs エネルギー

$$G = H - TS$$

を計算する。ただし H, T, S はそれぞれ系のエンタルピー，温度，エントロピーである。歴史的には，熱力学的は Newton 物理学とは異なる「新しい」物理学として発達した。純粋に熱力学的な観点では，分子を想定する必要はなく，系の状態すなわち熱，エントロピー，圧力のみを基本量として理論が展開できる。つまり，マクロスケールとミクロスケールとは，全く別のモデルであり，両者が整合する論理的な必然性はないのである。

B.5.3 ミクロスケールモデル

ミクロスケールモデルでは，量子や分子を基本単位とする。代表例として，分子動力学法（Molecular Dynamics; MD）と液体積分方程式理論がある。MD では分子を極性を持つ粒子として表現し，粒子が従う運動方程式を解く。一方，液体積分方程式理論では液体分子の存在密度関数を統計力学的に計算する。具体的には，MOZ 方程式（Molecular Ornstein-Zernike）

$$h(r_1, r_2) = c(r_1, r_2) + \int c(r_1, r)\rho(r)h(r, r_2)\mathrm{d}r,$$

を基礎方程式として，それを近似計算するための RISM（Reference Interaction Site Model）理論が展開されている。ただし r_i は分子の位置と配向，ρ は分子の存在密度，h は二分子の存在密度関数 $\rho_2(r_1, r_2)$ と一分子の平均密度 $\overline{\rho}$ から定義される全相関関数 $h(r_1, r_2) := \rho_2(r_1, r_2)/\overline{\rho}^2 - 1$，$c$ はグランドカノニカル分布から導かれる直接相関関数である。

MD と RISM はいずれも巨大な系を計算するには大量の計算機資源を要するほか，現実には実験値とあうように近似モデルを作り込む必要もある。この制約から，ミクロスケールモデルの対象は主に，分子および分子系の構造・エネルギー・機能・反応過程であり，生命科学でいえば，細胞内化学反応に相当する。例えば 3D-RISM ではタンパク質の活性部位へのリガンドの選択的結合（分子認識）がシミュレーションできる。

B.5.4 メソスケールモデル

メソスケールモデルでは，分子の集合体を基本単位とする。代表例として界面張力と界面の運動方程式，格子モデル，Ginzburg-Landau（GL）理論がある。界面の方程式は，分子ではなく境界面に着目する方法の総称であり，界面の位置エネルギーから導く方法や，GL 理論から導く方法など，複数の異なるモデルがある。格子モデルは，粒子ではなく場に着目する方法である。流体力学の用語で言えば，粒子に着目するミクロモデルが Lagrange 型であるのに対し，場に着目する格子モデルは Euler 型である。特に格子モデルの代表例であるイジングモデルには，ミクロモデルよりも正確なシミュレーションに成功したという歴史的経緯もある。GL 理論はもともと超電導の分野で相分離現象を説明するための理論として成功を収め[15]，今日では相分離現象を説明するための普遍的な基礎理論として位置づけられている。GL 理論は格子モデルの連続極限として得ることができ，界面方程式の元となるエネルギーは GL エネルギーから Allen-Cahn 方程式または Cahn-Hilliard 方程式を介して導かれる。従って，これらも Euler 型である。

GL 理論では，相分離の状態を秩序パラメータ（order parameter）$u(x)$ で表現する。秩序パラメータは実数に値をとる空間の関数（場）であり，各点での秩序の程度を表す。例えば，位置 x において $u(x)$ が 1 に近ければ，位置 x は水である可能性が高く，-1 に近ければ油である可能性が高い，という具合に設計する。相分離現象は，GL 自由エネルギー

$$F[u] := \int_{\mathbb{R}^3} [W(u(x)) + \frac{K}{2}|\nabla u(x)|^2]\mathrm{d}x.$$

を最小化する過程と説明される。ただし W は二重井戸ポテンシャル

$$W(u) := \frac{1}{4}(u^2 - 1)^2$$

であり，K は定数である。GL 自由エネルギーの第一項は秩序パラメータを 1 または -1 に近づける働きを表し，第二項は拡散項なので秩序パラメータを滑らかにする働きを表す。二重井戸ポテンシャルは意図的に設計されたわけではなく，第一原理計算によって導ける。

GL エネルギー最小化問題に対応する Euler-Lagrange 方程式は，相分離の方程式として知られる Allen-Cahn 方程式である

$$\frac{\partial u}{\partial t} = M\left(K\triangle u - W'(u)\right).$$

[15] Ginzburg はこの業績で 2003 年に Nobel 物理学賞を受賞している。

AC 方程式の解は一般に非保存量だが，水と油のように $u(x)$ を密度として設定する場合には，質量保存の式と組み合わせた Cahn-Hilliard 方程式

$$\frac{\partial u}{\partial t} = -M\triangle\left(K\triangle u - W'(u)\right).$$

も用いられる。

さらに，GL エネルギーにおいて界面厚さ 0 の極限（特異極限）をとると，界面のエネルギー（表面張力）が導ける。このとき，AC 方程式は平均曲率流

$$V(x,t) = H(x,t), \quad x \in \Gamma(t),$$

という界面の発展方程式に帰着する。ただし $\Gamma(t)$ は時刻 t における界面を表し，V と H はそれぞれ (x,t) における界面の成長速度と，法線方向の平均曲率を表す。同様に，CH 方程式の特異極限としては Mullins-Sekerka 方程式が得られることが分かっている。

参考文献

I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, Nature 521 (7553) (2015) 436–444.

J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks 61 (2015) 85–117.

麻生英樹, 安田宗樹, 前田新一, 岡野原大輔, 岡谷貴之, 久保陽太郎, ダヌシカボレガラ, 神嶌敏弘, 人工知能学会, 深層学習, 近代科学社, 2015.

得居誠也, Deep Learning 技術の今, URL http://www.slideshare.net/beam2d/deep-learning20140130, 2014.

Y. Bengio, Learning Deep Architectures for AI, Foundations and Trends® in Machine Learning 2 (1) (2009) 1–127.

Y. Bengio, O. Delalleau, On the Expressive Power of Deep Architectures, in: Algorithmic Learning Theory (ALT) 2011, Springer Berlin Heidelberg, 18–36, 2011.

Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013a) 1798–1828.

Y. Bengio, Deep Learning of Representations: Looking Forward, in: Statistical Language and Speech Processing (SLSP) 2013, 1–37, 2013.

G. E. Hinton, S. Osindero, Y.-W. W. Teh, A Fast Learning Algorithm for Deep Belief Nets, Neural Computation 18 (7) (2006) 1527–1554.

Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy Layer-Wise Training of Deep Networks, in: Advances in Neural Information Processing Systems 19, MIT Press, Vancouver, BC, 153–160, 2007.

M. Ranzato, C. Poultney, S. Chopra, Y. LeCun, Efficient Learning of Sparse Representations with an Energy-Based Model, in: Advances In Neural Information Processing Systems 19, MIT Press, Vancouver, BC, 1137–1144, 2007.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2146–2153, 2009.

X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of The 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, vol. 9, JMLR W&CP, Chia Laguna Resort, Sardinia, Italy, 249–256, 2010.

D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why Does Unsupervised Pre-training Help Deep Learning?, Journal of Machine Learning Research 11 (2010) 625–660.

A. Coates, Demystifying unsupervised feature learning, Ph.D. thesis, Stanford University, 2012.

H. Lee, C. Ekanadham, A. Y. Ng, Sparse deep belief net model for visual area V2, in: Advances in Neural Information Processing Systems 20, Curran Associates, Inc., Vancouver, BC, 873–880, 2008.

H. Larochelle, D. Erhan, A. Courville, J. Bergstra, Y. Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, in: Proceedings of The 24th International Conference on Machine Learning (ICML 2007), Omnipress, Corvallis, OR, 473–480, 2007.

H. Lee, Unsupervised feature learning via sparse hierarchical representations, Ph.D. thesis, Stanford University, 2010.

Q. V. Le, A. Karpenko, J. Ngiam, A. Y. Ng, ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning, in: Advances in Neural Information Processing Systems 24, Curran Associates, Inc., Granada, Spain, 1017–1025, 2011.

A. Coates, A. Ng, The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization, in: Proceedings of The 28th International Conference on Machine Learning (ICML-11), ACM, Belleview, WA, USA, 921–928, 2011.

A. Coates, A. Arbor, A. Y. Ng, An Analysis of Single-Layer Networks in Unsupervised Feature Learning, in: Proceedings of The 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, vol. 15, JMLR W&CP, Fort Lauderdale, FL, USA, 215–223, 2011.

Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the 29th International Conference on Machine Learning (ICML-12), Omnipress, Edinburgh, Scotland, UK, 81–88, 2012.

V. Nair, G. E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), Omnipress, Haifa, Israel, 807–814, 2010.

X. Glorot, A. Bordes, Y. Bengio, Deep Sparse Rectifier Neural Networks, in: Proceedings of The 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, vol. 15, JMLR W&CP, Fort Lauderdale, FL, USA, 315–323, 2011.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout : A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25, Curran Associates, Inc., Lake Tahoe, 1097–1105, 2012.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine 29 (6) (2012) 82–97.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, in: International Conference on Learning Representations 2014, Banff, BC, 1–16, 2014.

R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 580–587, 2014.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, C. Hill, A. Arbor, Going Deeper with Convolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, 1–9, 2015.

K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: International Conference on Learning Representations 2015, San Diego, California, 1–14, 2015.

K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, Nevada, 770–778, 2016.

R. Benenson, Who is the best at X?, URL `http://rodrigob.github.io/are_we_there_yet/build/`, 2016.

J. Duchi, E. Hazan, Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, Journal of Machine Learning Research 12 (2011) 2121–2159.

T. Tieleman, G. E. Hinton, Lecture 6.5 - RMSprop, COURSERA: Neural Networks for Machine Learning, Tech. Rep., University of Toronto, 2012.

D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations 2015, San Diego, California, 1–15, 2015.

V. Mnih, K. Kavukcuoglu, D. Silver, A. a. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

A. Mordvintsev, C. Olah, M. Tyka, Inceptionism: Going Deeper into Neural Networks, URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, 2015.

Y. Bengio, É. Thibodeau-Laufer, G. Alain, J. Yosinski, Deep Generative Stochastic Networks Trainable by Backprop, in: Proceedings of The 31st International Conference on Machine Learning, vol. 32, JMLR W&CP, Beijing, China, 226–234, 2014.

G. Alain, Y. Bengio, L. Yao, J. Yosinski, E. Thibodeau-Laufer, S. Zhang, P. Vincent, GSNs : Generative Stochastic Networks, Information and Inference 5 (2) (2016) 210–249.

D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, in: International Conference on Learning Representations 2014, Banff, BC, 1–14, 2014.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 2672–2680, 2014.

J. Sohl-Dickstein, P. Battaglino, M. R. DeWeese, Minimum Probability Flow Learning, in: Proceedings of The 28th International Conference on Machine Learning (ICML-11), ACM, Belleview, WA, USA, 905–912, 2011.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep Unsupervised Learning using Nonequilibrium Thermodynamics, in: Proceedings of The 32nd International Conference on Machine Learning, vol. 37, JMLR W&CP, Lille, France, 2256–2265, 2015.

N. Boulanger-Lewandowski, P. Vincent, Y. Bengio, Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription, in: Proceedings of the 29th International Conference on Machine Learning (ICML-12), Omnipress, Edinburgh, Scotland, UK, 1159–1166, 2012.

A. Graves, Generating Sequences with Recurrent Neural Networks, Tech. Rep., University of Toronto, 2013.

L. A. Gatys, A. S. Ecker, M. Bethge, A Neural Algorithm of Artistic Style, Tech. Rep., 2015.

K. Gregor, I. Danihelka, A. Graves, W. D. Rezende Danilo Jimenez, DRAW: A Recurrent Neural Network for Image Generation, in: Proceedings of The 32nd International Conference on Machine Learning, JMLR W&CP, Lille, France, 1462–1471, 2015.

A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations 2016, San Juan, Puerto Rico, 1–15, 2016.

A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel Recurrent Neural Networks, in: Proceedings of The 33rd International Conference on Machine Learning, vol. 48, JMLR W&CP, New York, 1747–1756, 2016.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari with Deep Reinforcement Learning, in: NIPS 2013 Deep Learning Workshop, 1–9, 2013.

V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent Models of Visual Attention, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 2204–2212, 2014.

I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 3104–3112, 2014.

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, Doha, Qatar, 1724–1734, 2014.

D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation By Jointly Learning To Align and Translate, in: International Conference on Learning Representations 2015, San Diego, California, 1–15, 2015.

M. Courbariaux, Y. Bengio, J.-P. David, BinaryConnect: Training Deep Neural Networks with binary weights during propagations, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., Montreal, BC, 3123–3131, 2015.

W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, Y. Chen, Compressing Neural Networks with the Hashing Trick, in: Proceedings of The 32nd International Conference on Machine Learning, vol. 37, JMLR W&CP, Lille, France, 2285–2294, 2015.

D. E. Rumelhart, J. L. McClelland, P. R. Group, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models, MIT Press, 1987.

G. E. Hinton, Connectionist learning procedures, Artificial Intelligence 40 (1-3) (1989) 185–234.

H. White, Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings, Neural Networks 3 (5) (1990) 535–549.

P. E. Utgoff, D. J. Stracuzzi, Many-Layered Learning, Neural Computation 14 (2002) 2497–2539.

K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics 36 (4) (1980) 193–202.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE 86 (1998) 2278–2324.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 2933–2941, 2014.

A. Blum, R. L. Rivest, Training a 3-node neural net is NP-complete, Neural Networks 5 (1992) 117–127.

R. Livni, S. Shalev-Shwartz, O. Shamir, On the Computational Efficiency of Training Neural Networks, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 855–863, 2014.

L. J. Ba, R. Caruana, Do deep nets really need to be deep?, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 2654–2662, 2014.

J. Håstad, Almost Optimal Lower Bounds for Small Depth Circuits, in: 18th annual ACM Symposium on Theory of Computing (STOC '86), 6–20, 1986.

O. Delalleau, Y. Bengio, Shallow vs. Deep Sum-Product Networks, in: Advances in Neural Information Processing Systems 24, Curran Associates, Inc., Granada, Spain, 666–674, 2011.

G. Montufar, R. Pascanu, K. Cho, Y. Bengio, On the Number of Linear Regions of Deep Neural Networks, in: Advances In Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 2924–2932, 2014.

M. Telgarsky, Benefits of depth in neural networks, in: 29th Annual Conference on Learning Theory, 1–23, 2016.

R. Eldan, O. Shamir, The Power of Depth for Feedforward Neural Networks, in: 29th Annual Conference on Learning Theory, vol. 49, 1–30, 2016.

N. Cohen, O. Sharir, A. Shashua, On the Expressive Power of Deep Learning: A Tensor Analysis, 29th Annual Conference on Learning Theory 49 (2016) 1–31.

Y. Bengio, O. Delalleau, N. L. Roux, The Curse of Highly Variable Functions for Local Kernel Machines, in: Advances in Neural Information Processing Systems 18, MIT Press, Vancouver, BC, 107–114, 2006a.

S. Arora, A. Bhaskara, R. Ge, T. Ma, Provable Bounds for Learning Some Deep Representations, in: Proceedings of The 31st International Conference on Machine Learning, vol. 32, JMLR W&CP, Beijing, China, 584–592, 2014.

R. Giryes, G. Sapiro, A. M. Bronstein, On the Stability of Deep Networks, in: International Conference on Learning Representations 2015, San Diego, California, 1–4, 2015a.

B. Neyshabur, R. Tomioka, N. Srebro, Norm-Based Capacity Control in Neural Networks, in: Proceedings of The 28th Conference on Learning Theory, vol. 40, JMLR W&CP, Paris, France, 1–26, 2015.

Y. Cho, L. K. Saul, Kernel Methods for Deep Learning, in: Advances in Neural Information Processing Systems 22, Curran Associates, Inc., Vancouver, BC, 342–350, 2009.

G. Montavon, M. L. Braun, K.-R. Müller, Kernel Analysis of Deep Networks, Journal of Machine Learning Research 12 (2011) 2563–2581.

P. Jawanpuria, Generalized Hierarchical Kernel Learning, The Journal of Machine Learning Research 16 (2015) 617–652.

J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 3320–3328, 2014.

C. Szegedy, W. Zaremba, I. Sutskever, Intriguing properties of neural networks, in: International Conference on Learning Representations 2014, Banff, BC, 1–10, 2014.

J. Yosinski, J. Clune, A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, 427–436, 2015.

S. Sonoda, N. Murata, Decoding Stacked Denoising Autoencoders, 2016.

A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, A. Y. Ng, On Random Weights and Unsupervised Feature Learning, in: Proceedings of The 28th International Conference on Machine Learning (ICML-11), ACM, Belleview, WA, USA, 1089–1096, 2011.

E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. Leung, L. Feng, Y.-S. Ong, M.-H. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B.-S. Oh, J. Jeon, K.-A. Toh, A. B. J. Teoh, J. Kim, H. Yu, Y. Chen, J. Liu, Extreme Learning Machines, IEEE Intelligent Systems 28 (6) (2013) 30–59.

R. Giryes, G. Sapiro, A. M. Bronstein, Deep Neural Networks with Random Gaussian Weights : A Universal Classification Strategy?, IEEE Transactions on Signal Processing 64 (13) (2015b) 3444–3457.

D. Duvenaud, O. Rippel, R. P. Adams, Z. Ghahramani, Avoiding pathologies in very deep networks, in: Proceedings of The 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, vol. 33, JMLR W&CP, Reykjavik, Iceland, 202–210, 2014.

I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout Networks, in: Proceedings of The 30th International Conference on Machine Learning, vol. 28, JMLR W&CP, Atlanta, Georgia, USA, 1319–1327, 2013.

G. E. Dahl, T. N. Sainath, G. E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 8609–8613, 2013.

A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing, Atlanta, 1–6, 2013.

M. D. Zeiler, M. Ranzato, R. Monga, M. Z. Mao, K. Yang, Q. Viet Le, P. Nguyen, A. W. Senior, V. Vanhoucke, J. Dean, G. E. Hinton, On rectified linear units for speech processing, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Vancouver, BC, 3517–3521, 2013.

P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, in: Proceedings of The 25th International Conference on Machine Learning (ICML-08), ACM, Helsinki, Finland, 1096–1103, 2008.

H. Bourlard, Y. Kamp, Auto-Association by Multilayer Perceptrons and Singular Value Decomposition, Biological Cybernetics 59 (1988) 291–294.

P. Baldi, K. Hornik, Neural networks and principal component analysis: Learning from examples without local minima, Neural Networks 2 (1) (1989) 53–58.

S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: explicit invariance during feature extraction, in: Proceedings of The 28th International Conference on Machine Learning (ICML-11), ACM, Belleview, WA, USA, 833–840, 2011.

S. Rifai, Y. Dauphin, The Manifold Tangent Classifier, in: Advances in Neural Information Processing Systems 24, Curran Associates, Inc., Granada, Spain, 2294–2302, 2011.

G. Alain, Y. Bengio, What Regularized Auto-Encoders Learn from the Data Generating Distribution, Journal of Machine Learning Research 15 (2014) 3743–3773.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, Journal of Machine Learning Research 11 (2010) 3371–3408.

Y. Bengio, L. Yao, G. Alain, P. Vincent, Generalized denoising auto-encoders as generative models, in: Advances in Neural Information Processing Systems 26, Curran Associates, Inc., Lake Tahoe, 899–907, 2013b.

D. Arpit, Y. Zhou, H. Ngo, V. Govindaraju, Why Regularized Auto-Encoders learn Sparse Representation?, in: Proceedings of The 33rd International Conference on Machine Learning, vol. 48, JMLR W&CP, New York, 136–144, 2016.

P. Vincent, A connection between score matching and denoising autoencoders, Neural Computation 23 (7) (2011) 1661–1674.

H. Kamyshanska, R. Memisevic, On autoencoder scoring, in: Proceedings of The 30th International Conference on Machine Learning, vol. 28, JMLR W&CP, Atlanta, Georgia, USA, 720–728, 2013.

H. Kamyshanska, R. Memisevic, The potential energy of an autoencoder, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (6) (2015) 1261–1273.

H. Larochelle, I. Murray, The Neural Autoregressive Distribution Estimator, in: Proceedings of The 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, vol. 15, JMLR W&CP, Fort Lauderdale, FL, USA, 29–37, 2011.

D. P. Kingma, S. Mohamed, J. D. Rezende, M. Welling, Semi-supervised learning with deep generative models, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montreal, BC, 3581–3589, 2014.

A. Rasmus, H. Valpola, M. Honkala, M. Berglund, T. Raiko, Semi-supervised learning with ladder networks, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., Montreal, BC, 3546–3554, 2015.

A. Wibisono, J. Bouvrie, L. Rosasco, T. Poggio, Learning and Invariance in a Family of Hierarchical Kernels, Tech. Rep., MIT, 2010.

Y. Mroueh, S. Voinea, T. Poggio, Learning with Group Invariant Features: A Kernel Perspective, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., Montreal, BC, 1558–1566, 2015.

F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio, Unsupervised learning of invariant representations, Theoretical Computer Science (2015) 1–10.

J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1872–1886.

T. Wiatowski, M. Tschannen, A. Stanic, P. Grohs, H. Bölcskei, Discrete Deep Feature Extraction: A Theory and New Architectures, in: Proceedings of The 33rd International Conference on Machine Learning, JMLR W&CP, New York, 2149–2158, 2016.

R. K. Srivastava, K. Greff, J. Schmidhuber, Training Very Deep Networks, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., Montreal, BC, 2377–2385, 2015.

N. Murata, An integral representation of functions using three-layered betworks and their approximation bounds, Neural Networks 9 (6) (1996) 947–956.

V. Kůrková, Complexity estimates based on integral transforms induced by computational units, Neural Networks 33 (2012) 160–167.

B. Irie, S. Miyake, Capabilities of three-layered perceptrons, in: IEEE International Conference on Neural Networks, IEEE, 641–648, 1988.

S. M. Carroll, B. W. Dickinson, Construction of neural nets using the Radon transform, in: International Joint Conference on Neural Networks 1989, vol. 1, IEEE, 607–611, 1989.

Y. Ito, Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory, Neural Networks 4 (3) (1991) 385–394.

E. J. Candès, Ridgelets: theory and applications, Ph.D. thesis, Standford University, 1998.

R. Hecht-Nielsen, Kolmogorov's Mapping Neural Network Existence Theorem, in: The 1st International Conference on Neural Networks, IEEE, New York, 1987.

A. N. Kolmogorov, On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables, Proceedings of the USSR Academy of Sciences 108 (1956a) 179–182.

V. I. Arnold, On functions of three variables, Proceedings of the USSR Academy of Sciences 114 (1957) 679–681.

D. A. Sprecher, On the structure of continuous functions of several variables, Transactions of the American Mathematical Society 115 (1965) 340–355.

G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems (MCSS) 2 (4) (1989) 303–314.

K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (5) (1989) 359–366.

K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural Networks 2 (3) (1989) 183–192.

H. Mhaskar, C. A. Micchelli, Approximation by superposition of sigmoidal and radial basis functions, Advances in Applied Mathematics 13 (3) (1992) 350–373.

M. Leshno, V. Y. Lin, A. Pinkus, S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Networks 6 (6) (1993) 861–867.

A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numerica 8 (1999) 143–195.

A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Transactions on Information Theory 39 (3) (1993) 930–945.

G. Pisier, Remarques sur un résultat non publié de B. Maurey, Séminaire d'Analyse Fonctionnelle 1980-1981 I (12) (1981) 1–12.

L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, The Annals of Statistics 20 (1) (1992) 608–613.

H. N. Mhaskar, Neural Networks for Optimal Approximation of Smooth and Analytic Functions, Neural Computation 8 (1996) 164–177.

P. P. Petrushev, Approximation by Ridge Functions and Neural Networks, SIAM Journal of Mathematical Analysis 30 (1) (1998) 155–189.

V. Kůrková, M. Sanguineti, Bounds on Rates of Variable-Basis and Neural-Network Approximation, IEEE Transactions on Information Theory 47 (6) (2001) 2659–2665.

P. C. Kainen, V. Kůrková, M. Sanguineti, Approximating multivariable functions by feedforward neural nets, in: M. Bianchini, M. Maggini, L. C. Jain (Eds.), Handbook on Neural Information Processing, vol. 49 of *Intelligent Systems Reference Library*, Springer Berlin Heidelberg, 143–181, 2013.

V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer New York, 2006 edn., 2006.

T. Poggio, F. Girosi, Networks for approximation and learning, Proceedings of the IEEE 78 (9) (1990) 1481–1497.

F. Girosi, M. Jones, T. Poggio, Regularization Theory and Neural Networks Architectures, Neural Computation 7 (1) (1995) 219–269.

P. L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Transactions on Information Theory 44 (2) (1998) 525–536.

P. Niyogi, F. Girosi, Generalization bounds for function approximation from scattered noisy data, Advances in Computational Mathematics 10 (1999) 51–80.

P. P. L. Bartlett, S. Mendelson, Rademacher and Gaussian Complexities: Risk Bounds and Structural Results, Journal of Machine Learning Research 3 (2002) 463–482.

O. Bousquet, A. Elisseeff, Stability and Generalization, Journal of Machine Learning Research 2 (2002) 499–526.

N. Cesa-Bianchi, A. Conconi, C. Gentile, On the generalization ability of on-line learning algorithms, IEEE Transactions on Information Theory 50 (9) (2004) 2050–2057.

F. Bach, Breaking the Curse of Dimensionality with Convex Neural Networks, Tech. Rep., INRIA Paris, 2014.

Y. Bengio, N. Le Roux, P. Vincent, O. Delalleau, P. Marcotte, Convex neural networks, in: Advances in Neural Information Processing Systems 18, MIT Press, Vancouver, BC, 123–130, 2006b.

E. J. Candès, Harmonic analysis of neural networks, Applied and Computational Harmonic Analysis 6 (2) (1999) 197–218.

B. Rubin, The Calderón reproducing formula, windowed X-ray transforms, and radon transforms in L^p-spaces, Journal of Fourier Analysis and Applications 4 (2) (1998) 175–197.

D. L. Donoho, Tight frames of k-plane ridgelets and the problem of representing objects that are smooth away from d-dimensional singularities in R^n, Proceedings of the National Academy of Science of the United States of America (PNAS) 96 (5) (1999) 1828–1833.

D. L. Donoho, Ridge functions and orthonormal ridgelets, Journal of Approximation Theory 111 (2) (2001) 143–179.

B. Rubin, Convolutionâbackprojection method for the k-plane transform, and Calderón's identity for ridgelet transforms, Applied and Computational Harmonic Analysis 16 (3) (2004) 231–242.

J.-L. Starck, F. Murtagh, J. M. Fadili, The ridgelet and curvelet transforms, in: Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity, Cambridge University Press, 89–118, 2010.

D. L. Donoho, Emerging applications of geometric multiscale analysis, Proceedings of the ICM, Beijing 2002 I (2002) 209–233.

S. Kostadinova, S. Pilipović, K. Saneva, J. Vindas, The ridgelet transform of distributions, Integral Transforms and Special Functions 25 (5) (2014) 344–358.

S. Kostadinova, S. Pilipović, K. Saneva, J. Vindas, The Ridgelet Transform and Quasi-asymptotic Behavior of Distributions, Operator Theory: Advances and Applications 245 (2015) 185–197.

S. Sonoda, N. Murata, Neural network with unbounded activation functions is universal approximator, Applied and Computational Harmonic Analysis .

猪狩惺, 実解析入門, 岩波書店, 1996.

柴田良弘, ルベーグ積分論, 内田老鶴圃, 2006.

W. Rudin, Functional Analysis, Higher Mathematics Series, McGraw-Hill Education, 2 edn., 1991.

H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Universitext, Springer-Verlag New York, 1 edn., 2011.

K. Yosida, Functional Analysis, Springer-Verlag Berlin Heidelberg, 6 edn., 1995.

垣田高夫, シュワルツ超関数入門, 日本評論社, 1999.

L. Schwartz, Théorie des Distributions, Hermann, Paris, nouvelle edn., 1966.

F. Trèves, Tological Vector Spaces, Distributions and Kernels, Academic Press, 1967.

W. Yuan, W. Sickel, D. Yang, Morrey and Campanato Meet Besov, Lizorkin and Triebel, Lecture Notes in Mathematics, Springer Berlin Heidelberg, 2010.

M. Holschneider, Wavelets: An Analysis Tool, Oxford mathematical monographs, The Clarendon Press, 1995.

澤野嘉宏, ベゾフ空間論, 日本評論社, 2011.

L. Grafakos, Classical Fourier Analysis, Graduate Texts in Mathematics, Springer New York, 2 edn., 2008.

S. Helgason, Integral Geometry and Radon Transforms, Springer-Verlag New York, 2011.

F. Natterer, X-ray tomography, in: L. L. Bonilla (Ed.), Inverse Problems and Imaging, vol. 1943, chap. 2, Springer-Verlag Berlin Heidelberg, 17–34, 2008.

E. Quinto, An introduction to X-ray tomography and Radon transforms, in: Proceedings of Symposia in Applied Mathematics: The Radon Transform, Inverse Problems, and Tomography, 1–23, 2006.

P. Kuchment, The Radon Transform and Medical Imaging, SIAM, 2014.

I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.

S. Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, Academic Press, 2009.

伊藤清三, 拡散方程式, 紀伊國屋書店, 1979.

小川卓克, 非線型発展方程式の実解析的方法, 丸善出版, 2013.

C. Villani, Optimal Transport: Old and New, vol. 338, Springer-Verlag Berlin Heidelberg, 2009.

L. Ambrosio, N. Gigli, G. Savaré, Gradient Flows in Metric Spaces and in the Space of Probability Measures, Lectures in Mathematics ETH Zürich, Birkhäuser, 2nd edn., 2008.

桑江一洋, 塩谷隆, 太田慎一, 高津飛鳥, 楽田和正, 最適輸送理論とリッチ曲率, in: 中央大学理工学部数学教室 (Ed.), 第 63 回 ENCOUNTER with MATHEMATICS, Tokyo, 115, 2015.

A. Hertle, Continuity of the radon transform and its inverse on Euclidean space, Mathematische Zeitschrift 184 (2) (1983) 165–192.

A. P. Calderón, Intermediate spaces and interpolation, the complex method, Studia Mathematica 24 (2) (1964) 113–190.

I. M. Gel'fand, G. E. Shilov, Generalized Functions, Vol. 1: Properties and Operations, Academic Press, New York, 1964.

L. C. Evans, R. F. Gariepy, Measure Theory and Fine Properties of Functions, CRC Press, revised edn., 2015.

Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Communications on Pure and Applied Mathematics 44 (4) (1991) 375–417.

R. J. McCann, Polar factorization of maps on Riemannian manifolds, Geometric And Functional Analysis 11 (3) (2001) 589–608.

A. Figalli, N. Gigli, Local semiconvexity of Kantorovich potentials on non-compact manifolds, ESAIM: Control, Optimisation and Calculus of Variations 17 (3) (2010) 648–653.

F. Otto, The Geometry of Dissipative Evolution Equations: The Porous Medium Equation, Communications in Partial Differential Equations 26 (2001) 101–174.

D. Vainsencher, S. Mannor, A. M. Bruckstein, The Sample Complexity of Dictionary Learning, Journal of Machine Learning Research 12 (2010) 3259–3281.

A. Maurer, M. Pontil, K-dimensional coding schemes in Hilbert spaces, IEEE Transactions on Information Theory 56 (11) (2010) 5839–5846.

M. Seibert, M. Kleinsteuber, R. Gribonval, R. Jenatton, F. Bach, On the sample complexity of sparse dictionary learning, in: IEEE Workshop on Statistical Signal Processing Proceedings, 5, 244–247, 2014.

R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, M. Seibert, Sample complexity of dictionary learning and other matrix factorizations, IEEE Transactions on Information Theory 61 (6) (2015) 3469–3486.

S. Sonoda, N. Murata, Sampling hidden parameters from oracle distribution, in: 24th International Conference on Artificial Neural Networks (ICANN) 2014, vol. 8681, Springer International Publishing, Hamburg, Germany, 539–546, 2014.

L. A. Shepp, B. F. Logan, The Fourier reconstruction of a head section, IEEE Transactions on Nuclear Science 21 (3) (1974) 21–43.

E. I. George, F. Liang, X. Xu, Improved minimax predictive densities under KullbackâLeibler loss, Annals of Statistics 34 (1) (2006) 78–91.

C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 2 (1972) 583–602.

K. B. Petersen, M. S. Pedersen, The Matrix Cookbook, Version: November 15, 2012, Tech. Rep., Technical University of Denmark, 2012.

F. Otto, C. Villani, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality, Journal of Functional Analysis 173 (2) (2000) 361–400.

A. Takatsu, Wasserstein geometry of Gaussian measures, Osaka Journal of Mathematics 48 (4) (2011) 1005–1026.

Y. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, Efficient BackProp, in: G. Montavon, G. B. Orr, K.-R. Muller (Eds.), Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2 edn., 9–48, 2012.

T. Denoeux, R. Lengelle, Initializing Back Propagation Networks With Prototypes, Neural Networks 6 (3) (1993) 351–363.

A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science, Springer New York, 1 edn., 2001.

D. A. Sprecher, A numerical implementation of Kolmogorov's superpositions, Neural Networks 9 (5) (1996) 765–772.

D. A. Sprecher, A Numerical Implementation of Kolmogorov's Superpositions II, Neural Networks 10 (3) (1997) 447–457.

Y. LeCun, C. Cortes, The MNIST database of handwritten digits, URL `http://yann.lecun.com/exdb/mnist/`, 1998.

T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, second edn., 2006.

H. Jeffreys, An Invariant Form for the Prior Probability in Estimation Problems, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 186 (1007) (1946) 453–461.

C. R. Rao, 統計学とは何か —偶然を生かす, 筑摩書房, 2010.

R. A. Fisher, Theory of Statistical Estimation, Mathematical Proceedings of the Cambridge Philosophical Society 22 (5) (1925) 700–725.

L. Brillouin, Science and Information Theory, Academic Press, New York, second edn., 1962.

G. Shafer, V. Vovk, Probability and Finance: It's Only a Game!, Wiley, New York, 2001.

C. E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal 5 (3) (1948) 3.

小野厚夫, 45 周年記念特別寄稿:情報という言葉を尋ねて（1）, 情報処理 46 (4) (2005) 347–351.

高橋秀俊, Information Theory, 日本物理学会誌 7 (1) (1952) 8–16.

C. E. Shannon, A Symbolic Analysis of Relay and Switching Circuits, thesis (M.S.), Tech. Rep., Massachusetts Institute of Technology, 1940.

L. Boltzmann, Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen, Sitzungsberichte Akademie der Wissenschaften 66 (1872) 275–370.

F. Y. Edgeworth, On the Probable Errors of Frequency-Constants, Journal of the Royal Statistical Society 71 (2) (1908) 381–397.

A. N. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer Berlin Heidelberg, 1933.

R. A. Fisher, On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 222 (1922) 309–368.

梅垣壽春, 大矢雅則, 北川敏男, 確率論的エントロピー—情報理論の函数解析的基礎 1—, 共立出版, 1983.

W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathematical Biophysics 5 (4) (1943) 115–133.

F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain., Psychological review 65 (6) (1958) 386–408.

R. M. Fano, Transmission of Information: A Statistical Theory of Communication, MIT Press, new edn., 1961.

M. Minsky, S. Papert, Perceptrons: An Introduction to Computational Geometry, MIT Press, Cambridge, MA, USA, 1969.

A. N. Kolmogorov, On tables of random numbers, SankhyÄ. The Indian Journal of Statistics. Series A 25 (4) (1963) 369–376.

R. Solomonoff, A formal theory of inductive inference, Information and Control 7 (1964) 1–22.

A. N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, Soviet Mathematics 4 (1963) 1035–1038.

R. E. Kálmán, A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engineering 82 (35) (1960) 35–45.

W. Hoeffding, Probability Inequalities for Sums of Bounded Random Variables, Journal of the American Statistical Associationssociation 58 (301) (1963) 13–30.

A. N. Kolmogorov, On certain asymptotic characteristics of completely bounded metric spaces (In Russian), Doklady Akademii Nauk SSSR 108 (3) (1956b) 385–388.

R. E. Bellman, Adaptive control processes: a guided tour, Princeton University Press, 1961.

H. Raiffa, R. Schlaifer, Applied Statistical Decision Theory, MIT Press, 1961.

K. Arrow, Uncertainty and the welfare economics of medical care, The American Economic Review 58 (1963) 941–973.

G. Akerlof, The market for lemons: quality uncertainty and the market mechanism, The Quarterly Journal of Economics 84 (3) (1970) 488–500.

Y. Bar-Hillel, Language and Information, Addison-Wesley, Jerusalem, 1964.

D. M. MacKay, Information, Mechanism and Meaning, MIT Press, 1969.

J. A. Nelder, R. W. M. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society. Series A (General) 135 (3) (1972) 370–384.

H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Proceedings of The 2nd International Symposium on Information Theory, 267–281, 1973.

D. B. Rubin, Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology 66 (5) (1974) 688–701.

B. Efron, Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics 7 (1) (1979) 1–26.

P. J. Huber, Robust Statistics, John Wiley & Sons, Inc., New York, 1981.

P. K. Andersen, R. Gill, Cox's Regression Model for Counting Processes: A Large Sample Study, The Annals of Statistics 10 (4) (1982) 1100–1120.

S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.

J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988.

J. Ziv, A. Lempel, A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory 23 (3) (1977) 337–343.

J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.

I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic Press, 1981.

S.-i. Amari, Differential-Geometrical Methods in Statistics, Springer New York, 1985.

J. Morlet, G. Arens, E. Fourgeau, D. Glard, Wave propagation and sampling theory, Geophysics 47 (1982) 203–236.

A. Grossmann, J. Morlet, Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape, SIAM Journal on Mathematical Analysis 15 (4) (1984) 723–736.

I. Daubechies, Orthonormal bases of compactly supported bases, Communications on Pure and Applied Mathematics 41 (7) (1988) 909–996.

S. G. Mallat, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693.

C. McDiarmid, On the method of bounded differences, 1989.

K. Azuma, Weighted sums of certain dependent random variables, Tohoku Mathematical Journal 19 (3) (1967) 357–367.

S. Bernstein, On a modification of Chebyshev's inequality and of the error formula of Laplace, Annals Science Institute SAV Ukraine, Sect. Math 1 (4) (1924) 38–49.

E. A. Feigenbaum, The Art of Artificial Intelligence. 1. Themes and Case Studies of Knowledge Engineering, in: International Joint Conference on Artificial Intelligence (IJCAI) V, 1014–1028, 1977.

M. Minsky, A framework for representing knowledge, in: P. H. Winston (Ed.), The Psychology of Computer Vision, McGraw-Hill, New York, 211–277, 1975.

V. N. Vapnik, A. Y. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, Theory of Probability and Its Applications 16 (2) (1971) 264–280.

N. Sauer, On the Density of Families of Sets, Journal of Combinatorial Theory, Series A 13 (1) (1972) 145–147.

S. Shelah, A combinatorial problem; stability and order for models and theories in infinitary languages, Pacific Journal of Mathematics 41 (1) (1972) 247–261.

L. G. Valiant, A theory of the learnable, Communications of the ACM 27 (11) (1984) 1134–1142.

D. E. Rumelhart, J. L. McClelland, P. R. Group, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol. 1: Foundations, MIT Press, 1986.

C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20 (3) (1995) 273–297.

B. R. Rubinstein, A. M. Bruckstein, M. Elad, Dictionaries for Sparse Representation Modeling, Proceedings of the IEEE 98 (6) (2010) 1045–1057.

B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1?, Vision Research 37 (23) (1997) 3311–3325.

S. S. Chen, D. L. Donoho, M. a. Saunders, Atomic Decomposition by Basis Pursuit, SIAM Journal on Scientific Computing 20 (1) (1998) 33–61.

R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B 58 (1) (1996) 267–288.

S. G. Mallat, Z. Zhang, Matching Pursuits With Time-Frequency Dictionaries, IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.

L. von Ahn, Human Computation, Ph.D. thesis, Carnegie Mellon University, 2005.

C. Dwork, Differential Privacy, in: Proceedings of the International Colloquium on Automata, Languages and Programming, Part II (ICALP), 1–12, 2006.

L. Floridi, Information: A Very Short Introduction, Oxford University Press, 2010.

R. von Mises, Grundlagen der Wahrscheinlichkeitsrechnung, Mathematische Zeitschrift 5 (1919) 52–99.

P. Martin-Löf, The Definition of Random Sequences, Information and Control 9 (6) (1966) 602–619.

L. Bottou, O. Bousquet, The Tradeoffs of Large Scale Learning, in: Advances in Neural Information Processing Systems 20, Curran Associates, Inc., Vancouver, BC, 161–168, 2008.

D. L. Donoho, P. B. Stark, Uncertainty Principles and Signal Recovery, SIAM Journal on Applied Mathematics 49 (3) (1989) 906–931.

B. F. Logan, Properties of high-pass signals, Ph.D. thesis, Columbia University, 1965.

E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509.

M. Kearns, Y. Mansour, A. Y. Ng, D. Ron, An Experimental and Theoretical Comparison of Model Selection Methods, Machine Learning 27 (1) (1997) 7–50.

W. H. Rogers, T. J. Wagner, A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules, The Annals of Statistics 6 (3) (1978) 506–514.

L. P. Devroye, T. J. Wagner, Distribution-Free Inequalities for the Deleted and Hold-out Error Estimates, IEEE Transactions on Information Theory 25 (2) (1979a) 202–207.

L. P. Devroye, T. J. Wagner, Distribution-Free Performance Bounds for Potential Function Rules, IEEE Transactions on Information Theory 25 (5) (1979b) 601–604.

M. Kearns, D. Ron, Algorithmic stability and sanity-check bounds for leave-one-out cross-validation., Neural Computation 11 (6) (1999) 1427–1453.

S. Shalev-Shwartz, O. Shamir, N. Srebro, K. Sridharan, Learnability, Stability and Uniform Convergence, Journal of Machine Learning Research 11 (2010) 2635–2670.

研究業績

*は本研究と直接関係のある業績を表す。

雑誌論文

* S. Sonoda, N. Murata, “Neural Network with Unbounded Activation Functions is Universal Approximator”, to appear in Applied and Computational Harmonic Analysis, Elsevier, (2015).

  S. Sonoda, N. Murata, H.Hino, H.Kitada, M.Kano, “A Statistical Model for Predicting the Liquid Steel Temperature in Ladle and Tundish by Bootstrap Filter”, ISIJ Int’l., 52(6):1086-1091, (2012).

国際会議

* S. Sonoda, N. Murata, “Sampling Hidden Parameters from Oracle Distribution”, The 24th International Conference on Artificial Neural Networks (ICANN2014), Hamburg, Germany, September 15-19, 2014.（査読付）

* S. Sonoda, N. Murata, “Ridgelet Analysis of ReLU Network”, The 29th Machine Learning Summer School (MLSS2015), Kyoto, Japan, August 23 - September 4, 2015.

  N. Murata, S. Sonoda, H.Hino, H.Kitada, M.Kano, “Sensitivity Analysis for Controlling Molten Steel Temperature in Tundish”, 2012 IFAC Workshop on Automation in the Mining, Mineral and Metal Industries, MMM 2012, Gifu, Japan, September 10 - 12, 2012.（査読付）

国内会議

* 園田翔, 村田昇, “深層デノイジング・オートエンコーダーの輸送理論解釈”, 第 19 回 情報論的学習理論ワークショップ (IBIS2016), 京都, 2016 年 11 月.

  嶋田達之介, 園田翔, 村田昇, 加藤真平, “Saliency Map を用いた CNN ブレーキシーン判別器の解析”, 第 19 回 情報論的学習理論ワークショップ (IBIS2016), 京都, 2016 年 11 月.

嶋田達之介, 松原拓央, 園田翔, 村田昇, パトリシアオータル, 加藤真平, “LiDAR 深度データを用いた CNN ブレーキシーン認識”, 第 18 回 情報論的学習理論ワークショップ (IBIS2015), 筑波, 2015 年 11 月.

* 園田翔, 村田昇, “ReLU ネットワークの積分表現理論”, 2015 年度 科学研究費シンポジウム「大規模複雑データの理論と方法論：最前線の動向」, 筑波, 2015 年 11 月.

* 園田翔, “深層学習のリッジレット解析にむけた取組み”, 2015 RIMS 共同研究 「ウェーブレット解析と信号処理」, 京都, 2015 年 11 月.

* 園田翔, 村田昇, “オラクル分布を用いたサンプリング学習アルゴリズム”, 第 18 回 IBISML 研究会（情報論的学習理論と機械学習研究会）, 筑波, 2014 年 9 月.

  金田有紀, 園田翔, 日野英逸, 村田昇, “複数粒子フィルタとモデル選択を用いた EEG データの電流ダイポール推定”, 第 17 回 IBISML 研究会（情報論的学習理論と機械学習研究会）, 沖縄, 2014 年 6 月.

  園田翔, 村田昇, 日野英逸, 進藤史裕, 北田宏, 加納学, “ブートストラップフィルタによる溶鋼温度分布の予測と制御”, 日本鉄鋼協会 第 162 回 秋季講演大会, 大阪, 2011 年 9 月.

招待講演等

* 園田翔, “深層ニューラルネットの積分表現理論”, 第 29 回 科研費新学術領域「多元計算解剖学」セミナー, 東京, 2016 年 11 月.

* 園田翔, “ニューラルネットの積分表現理論”, 第 2 回 産総研人工知能セミナー 「機械学習の理論的側面」, 台場, 2015 年 11 月.

受賞

* IBIS2016 学生最優秀プレゼンテーション賞, “無限層デノイジング・オートエンコーダーの輸送理論解釈”, 園田翔, 村田昇, 2016 年 11 月.

  日本鉄鋼協会 計測・制御・システム研究賞, “物理・統計的モデリングによる取鍋内溶鋼温度の高度予測技術”, 園田翔, 大倉才昇, 村田昇, 日野英逸, 加納学, 北田宏, 2012 年 3 月.

プレプリント

* S. Sonoda, N. Murata, “Decoding Stacked Denoising Autoencoders”, arXiv:1605.02832, (2016).