深層ニューラルネットワークの解剖
——統計力学によるアプローチ

吉 野　元　〈大阪大学サイバーメディアセンター　yoshino@cmc.osaka-u.ac.jp〉

深層ニューラルネットワーク（Deep Neural Network, DNN）を用いた機械学習は，深層学習とよばれ，画像認識，機械翻訳などで身近なものとなった．しかしその高い学習能力のメカニズムはよくわかっておらず，ブラックボックスとして使われている面が無視できない．最先端の応用では様々なノウハウが駆使されるが，単純化した状況設定から考える物理学の発想がこのブラックボックスにメスを入れるのに役立つであろう．ニューラルネットワークを用いた機械学習はスピングラスに端を発するランダム系の統計力学，情報統計力学において伝統的に重要なテーマである．

Nビットの入力を，Nビットの出力に変換する「関数」を，DNNでデザインすることを考えてみよう．このNをDNNの「幅」とよぶことにする．入出力を含めて，ネットワークには多数のニューロンがある．あるニューロンの状態を変数S_iで表そう．これが入力信号$h=\sum_j J_{ij}S_j$の関数として$S_i=f(h)$で決まるとする．ここでS_jは隣接する，上流側，すなわち入力層に近い方の層にあるニューロンの状態でJ_{ij}はシナプス結合とよばれる．$f(h)$は活性化関数とよばれる．このDNN（このさき機械とよぶ）は多くの調節可能なシナプス結合J_{ij}をもち，これを調節してデザインできる機械の全体集合をΩ_0としよう．

統計力学的には次のような問いが立つ．M個の異なる入出力データの組が訓練データ（境界条件）として与えられたとして，これに完全に適合する機械は，シナプス結合J_{ij}を色々変えて，何通り作ることができるか？　この「正解の集合」をΩとし，その統計力学を考えるのである．

学習の問題で重要なのは，訓練データである．人工的だがシンプルなシナリオとして，(1) ランダムな入出力データ，(2) Ω_0から無作為に選んだ一つの「教師機械」にランダムなデータを入力し，対応する出力を取り出し，この組を「生徒機械」の訓練データとする，というものがある．(1)はガラス・ジャミング系の統計力学に深く関係する．他方，(2)はいわば結晶（隠された「教師機械」）を推定する統計力学である．

DNNの構成要素として最も単純なのは，符号を取り出す関数$f(h)=\mathrm{sgn}(h)$を活性化関数とするもので，ニューロンの状態はイジング変数$S_i=\pm 1$になる．これはいわゆるパーセプトロンの一つである．単体の場合は(1)(2)のシナリオともに深く理解されている．しかしこれを多数組み合わせたDNNの理論解析は困難とされてきた．

この困難は次のように克服できる．まず，全パーセプトロンの入出力関係が満足されることを拘束条件として導入することにより，シナプス結合J_{ij}のほかにニューロンS_iも力学変数に加えることができる．これによって，入力と出力を多段階の非線形写像で結ぶ問題が，局所的な相互作用をもつ多体系の統計力学として捉え直される．

得られた系には入出力層以外にランダムネスはない．ここで重要なヒントとなるのは，無限大次元の剛体球ガラスなど，近年急速に発展したガラス・ジャミング系の平均場理論である．そこではハミルトニアンにランダムネスがない系に対してもスピングラスなどランダム系で用いられたレプリカ法が強力なツールとなることが明らかになっている．

レプリカ法で理論を構成して解析した結果，熱力学極限N（幅），M（データ数）$\to\infty$で，比$\alpha=M/N$の増大とともに(1) レプリカ対称性の破れを伴うガラス転移，(2) 結晶化が，ネットワークの両端から逐次的に起こって解空間Ωが狭くなること，ネットワークが十分深ければ中央部に「遊び」（液体領域）が残されることがわかった．これはある種の濡れ転移とみなせる．現実的には幅Nは有限であり，転移はクロスオーバーとなり，系は深さ方向にダイナミックスが変化する複雑な液体となる．

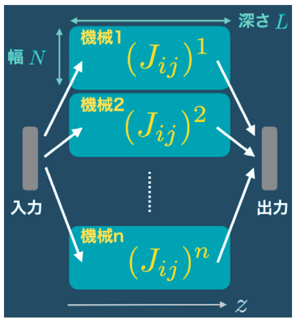
用語解説

深層ニューラルネットワーク（DNN）：
多数の層からなる層状のニューラルネットワーク．様々なタイプがあるが，ここでは多くの層からなる中間層（隠れ層）を通して，入力層から出力層まで信号が逆戻りせずに伝搬するネットワークを考える．

スピングラス：
強磁性と反強磁性相互作用がランダムに混在した磁性体．

ガラス・ジャミング系：
ガラス系とはガラス状態，すなわち乱れたパターンを（準）安定な固体状態としてもつ系である．例えば，ある箱にN個の球を入れるとして，球同士を重なりを許さない「剛体の制約を充足する」配置の位相空間Ωでの統計力学が考えられる．これは剛体球系の液体–結晶転移，結晶を押しかためた最密充填とともに，過冷却状態でのガラス転移，ガラスを押し固めたジャミング（ランダム充填）を研究する舞台である．

レプリカ法：
ガラス系の統計力学における代表的な理論手法．DNNの問題では，下図のように，同じ訓練データのもとで学習している複数の機械をレプリカとする．これらを互いに比べることによって，どの程度似通った機械になっているのか，またその深さ方向zでの変化を探ることができる．



レプリカ法の概念図．

濡れ転移：
例えば気相と液層が共存するとき，温度の低下とともに壁の表面から液相が厚みを増していく現象．

1. はじめに

深層ニューラルネットワークは画像や音声の認識，機械翻訳などで大きな成功を収めている．多くの物理学者も興味をもち，研究に活用するアイディアが活発に議論されている．[1,2] 非常に表現能力の高い機械であることは間違いない．しかし，その学習のメカニズムはよくわかっておらず，ブラックボックスのまま使われている面が無視できない．[*1]

機械学習の性能は，機械だけで議論できず，機械とデータのセットで決まる．標準的なDNNによる学習では，多数の入力・出力データのセットを「訓練データ」とし，個々の入力に対して対応する期待された出力が出てくるようにシナプス結合J_{ij}などのパラメータを調節する．このプロセスが（教師あり）学習である．

現実のデータがどのような性質をもっているのかは難しい問題である．統計力学的な解析では，現実のデータの代わりに，解析の見通しがクリアな単純な人工データのモデル（以下ではシナリオとよぶ）のアンサンブルでまず考察するのが常套手段である（図1）．[3–5] 当然，人工データと現実のデータの性質の違いは問われ，シナリオを見直し，ステップアップする努力が求められる．

1.1 シナリオ（1）：丸暗記—clustering/SAT-UNSAT転移，ガラス転移/ジャミング

もっとも素朴なシナリオでは，入出力の各ビットが，互いに相関のない全くランダムなデータを用いる．これは現実のデータからはるかに遠い状況ではあるが，機械の「丸暗記力」がわかる．これも一つの能力ではあり，DNNがどれほどの丸暗記力をもつのかも，一つの基本問題である．

統計力学的視点からは，これは「制約充足問題」の一つと捉えられ，物理のガラス転移，ジャミングの問題と深く関わるところも大変興味深い点である．[*2]

1.2 シナリオ（2）：教師–生徒機械—結晶化

次に，ただやみくもに暗記する能力とは違う，「真似する能力」を測ろうというのが「教師–生徒機械」の設定である．ある「教師機械」を考え，これにランダムなデータを入力して，その出力を取り出し，この入力・出力のセットを「生徒機械」の学習に使う．学習データの数を増やしていくと，生徒機械のシナプス結合J^{student}が，教師機械のJ^{teacher}に次第に近づいていくことが期待される．[*3]

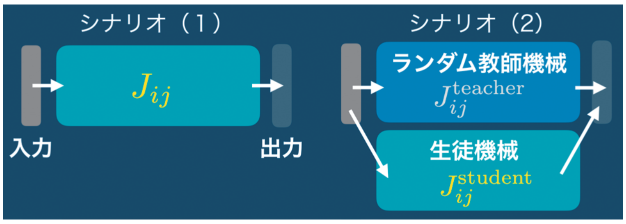


図1　人工データに基づく学習のための2つのシナリオ．

この「教師機械」は，現実のデータの背後にあるもの，例えば物理法則の役割をしていると考えればよいだろう．そうすると，どのような教師機械を考えるかが重要である．一番素朴にはアーキテクチュアは生徒機械と同じとしたうえで，ランダムなシナプス結合J^{teacher}を生成して教師機械としてしまう．膨大な数のパラメータが互いに無相関に決められているランダム教師と，ごく少数のパラメータで記述される物理法則とはずいぶん異なる．したがってこれは単純すぎる設定だが，まずはこれから始める．

このシナリオは「統計的推定」[*4]の一種である．「教師–生徒機械」のシナリオはベイズの定理に基づいた統計的推定を説明するのによく用いられる．[6] 我々の設定は，「ベイズ最適」[*5]とよばれる最も理想化されたベイズ推定に相当する．隠された教師のJ^{teacher}を，与えられた訓練データから推定するというこの設定は，隠れた結晶（教師）を探す，結晶化の統計力学ともみなせる．

1.3 現実にもう一歩近づくには

統計力学では熱力学極限$N\to\infty$を取ることが自然である．さらに，本稿で議論するDNNでは層間の結合が密な結合（図2参照）であるため幅$N\to\infty$で∞次元系になり，平均場理論が正確になる，という利点がある．しかし現実の系で幅Nは有限であり，相転移はすべてクロスオーバーになる．実際，数値シミュレーションを行うと幅Nに依存した有限の緩和時間τ_Nをもつ「液体」として振る舞うことがわかる．[11] $N\to\infty$での相転移現象は，τ_Nよりも短い時間スケールでの振る舞いに反映されると期待される．

また上のシナリオではNビットの訓練データにおける各ビット間に相関があるとは考えておらず，データはN次元の情報をもつ．しかし，現実のデータはNビットだったと

[*1] 例えば次のような未解決問題がある．DNNではしばしば，訓練データの数を遥かに上回るパラメータ数で学習を行うという異常なモデリング（over-parametrization）になる．これは物理法則が自然界の膨大なデータを，極めて少数のパラメータで表現できる（と信じられている）こととと比べると，センスが真逆である．実験データのデータ数よりもパラメータ数が多いフィッティング曲線など，常識的には受け入れられるものではないだろう．では，over-parametrizationの状況でも，単なる丸暗記でない，意味のある学習がDNNで実現しているように見える理由は何か？

[*2] 訓練データの数Mの増大によって制約が厳しくなると解空間Ωは狭くなっていく．それだけでなく，制約が厳しくなることによってΩが複数の部分空間に分裂するclustering転移が，グラフの彩色問題，k-SAT問題などの「制約充足問題」ではしばしば見られる．[6,7] 制約をさらに厳しくしていくと，クラスターの位相体積があるところで0になり，SAT-UNSAT転移，すなわちそれ以上多い制約を満たせなくなる限界に達する．

熱運動する剛体球の集団（例えば硬いコロイド粒子）を箱に閉じ込めたうえで箱を潰していくと圧力が上昇する．急激に圧縮すると系は結晶化するチャンスを逃してガラス化が起こる．[8] これがclustering転移に対応する．さらに圧縮すると完全に身動きができない，ランダムなパッキング状態，ジャミングに至る．これがSAT-UNSAT転移に対応する．

[*3] この訓練のあと，いわゆる汎化能力を次のように測定できる．すなわち，生徒機械と教師機械に，訓練には使わなかった新しいランダムな入力を入れてその出力を取り出し，生徒機械が教師機械とどの程度同じ出力を未知のデータに対して出せるのかを測ればよい．

[*4] 誤り訂正符号，圧縮センシングなど多くの問題がある．[4–6]

[*5] すなわち，教師–生徒機械が全く同じアーキテクチュアで，生徒は教師のシナプス結合J^{teacher}の値はそのものは知らないものの，教師に関するそれ以外のすべて（J^{teacher}の分布，入力・出力は何か）を完全に知ったうえでJ^{teacher}を推定しようとしている．これはスピングラス[9]では西森線上にいることに相当する．[4,10]

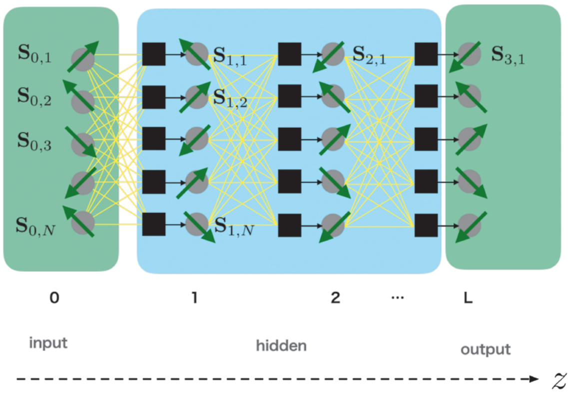


図2　多層パーセプトロンネットワークの概念図.[11] 各層にN個のニューロンがある. 入力を0層, 出力をL層とし, その他$l=1, 2, ..., L-1$を隠れ層とよぶ. 各ニューロンには, M個の訓練データ$\mu=1, ...M$に応じた発火パターン$S^{\mu}=\pm 1$があり, M成分のベクトルスピン$\mathbf{S}=(S^1, S^2, ...S^M)$（緑矢印）として表せる. パーセプトロンは■, シナプス結合は黄色線で表している. 隣接する層間の結合は密なので平均場模型的（∞次元系）, これが1次元的に連なったある種の$1+\infty$次元系である.

してもビット間に相関があり, 埋め込まれている情報の有効的な次元は低いことが多い. *6

以上から有限幅N効果が重要であると考えられる. また教師にある種の有効的次元$D(<N)$をもたせるシナリオ[12]が考えられる. これらは今後の課題として最後に議論する.

2. 深層パーセプトロンネットワーク

最も単純なDNNとして, 図2のような, 幅N, 深さLの多層パーセプトロンのネットワークを考える. 各層$l=0, 1, ..., L$でN個のベクトルスピン$\mathbf{S}_{l,i}=(S^1_{l,i}, S^2_{l,i}, ..., S^M_{l,i})$ $(i=1, 2, ..., N)$がある. 各成分$\mu=1, 2, ..., M$はM組の訓練データに対応し, その値は入出力層$l=0, L$では訓練データそのものを表し, $l=1, 2, ..., L-1$では隠れ層にあるニューロンの状態を表す.

パーセプトロン$\blacksquare=(l, i)$の出力$S^{\mu}_{\blacksquare}$は

$$S^{\mu}_{\blacksquare}=\mathrm{sgn}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N} J^k_{\blacksquare} S^{\mu}_{\blacksquare(k)}\right) \qquad \mu=1,2,\ldots,M. \tag{1}$$

のように決まる. ニューロンSは± 1の値をとるイジング変数となる. ここで$\blacksquare(k)=(l-1, k)$で, $J^k_{\blacksquare}$は■と$\blacksquare(k)$を結ぶシナプス結合である. シナプス結合は$\sum_{k=1}^{N}(J^k_{\blacksquare})^2=N$のように規格化されているとする. 図2のように隣接する層のニューロン間は密に結合している.

ここで議論するのは, 完璧に学習できた「正解」の空間の統計力学である. *7 単体のパーセプトロンの場合にはE. Gardner[3]が解析しており, 「正解」の空間体積はガードナー体積とよばれる. 拘束条件となる訓練データの入出力を$\mathbf{S}_{0,i}$, $\mathbf{S}_{L,i}$ $(i=1, 2, ..., N)$として, DNNのガードナー体積は

$$V(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\}) = \left(\prod_{\blacksquare=(1,1)}^{(L,N)} \mathrm{Tr}_{\mathbf{J}_{\blacksquare}}\right)\left(\prod_{\blacksquare=(1,1)}^{(L-1,N)} \mathrm{Tr}_{\mathbf{S}_{\blacksquare}}\right) \mathrm{e}^{-\beta\sum_{\mu}\sum_{\blacksquare} v(r^{\mu}_{\blacksquare})} \tag{2}$$

と表せる. ここで和$\mathrm{Tr}_{\mathbf{J}_{\blacksquare}}$はパーセプトロン■への入力に参加するすべてのシナプス結合に関する状態和であり, 和$\mathrm{Tr}_{\mathbf{S}}$はすべての訓練データのパターン$\mu=1, 2, ..., M$に関するニューロンの状態和である.

ガードナー体積には「正解」だけを勘定に入れる. 「正解」の条件は, すべての訓練データのパターン$\mu=1, 2, ..., M$についてすべてのパーセプトロン$\blacksquare=(1, 1), ..., (L, N)$, が入出力関係式(1)を満たしていることである. これを境界条件とするために式(2)に入れてあるのが「ボルツマン因子」

$$\mathrm{e}^{-\beta v(r)}=\theta(r) \tag{3}$$

である. $\theta(r)$は階段関数で$r>0$の場合のみ状態和に寄与する. 式(21)で次の変数を導入した.

$$r^{\mu}_{\blacksquare} \equiv S^{\mu}_{\blacksquare}\sum_{i=1}^{N}\frac{J^i_{\blacksquare}}{\sqrt{N}} S^{\mu}_{\blacksquare(i)} \tag{4}$$

力学変数としてシナプス結合$J^i_{\blacksquare}$のみとすると, 多数回畳み込まれた非常に複雑な長距離相互作用系の問題になってしまう. ここではニューロンも力学変数に入れたことによってこの畳み込みが解かれ, 系全体の有効ハミルトニアンが$\sum_{\blacksquare}\sum_{\mu} v(r^{\mu}_{\blacksquare})$のように局所的なハミルトニアンの和として表せているのが重要なポイントである. これでDNNのガードナー体積をある$1+\infty$次元の系の分配関数とみなすことができるようになり, 格段に解析しやすくなった.

我々の問題は$1+\infty$次元空間の剛体球系の統計力学に近い. 剛体相互作用する2つの球のボルツマン因子はまさに式(3)の形であり, rが向き合う2つの球の表面間の距離(gap)である. そのため式(4)をギャップ(gap)変数とよぶ.

ガードナーが解析した単体のパーセプトロン[3]を含めて, 連続自由度の制約充足問題[7]その逆問題としての統計的推定問題は, 剛体球系の統計力学と関係が深い. *8

3. レプリカ理論の構成

ガードナー体積$V(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\})$は訓練データ$(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\})$に適合するDNNの集合の大きさである. 訓練データの数Mを増やしていくと, これは小さくなっていく. そのとき, この位相空間の内部にどのような変化が起こるのかが興味深い問題である. この問題を解析するためにレプリカ法を

*6 例えばアルファベットの手書き文字画像を集めた公開データであるMNISTでは$N=28\times 28=784$ピクセルのデータに$D\sim 14$次元位の情報しかないと推定されている.[12]

*7 標準的な学習は次のように行われるであろう. まず, シナプス結合を適当に初期化しておき, 各訓練データμの入力データ$S^{\mu}_{0,i}$ $(i=1, 2, ..., N)$に対して, このパーセプトロンネットワークが出す出力を求める. その出力と, 訓練データの出力データ, すなわち望まれるスピン配位$S^{\mu}_{L,j}$ $(j=1, 2, ..., N)$とは食い違っている. そこで, このズレに対して適当なコスト関数(loss function)を定義し, これが小さくなっていくようにシナプス結合を変化させていく. この際, 今のモデルでは活性化関数が符号関数なので通常の逆誤差伝搬法は使えないが, モンテカルロ法は使える.

*8 剛体球系では, エネルギーは0か∞しかなく, その統計力学はエントロピーがすべてを支配する. 例えば, 高密度の剛体球系では乱れた液体状態にいるよりも結晶配置をとった方が粒子がラットリング運動でうごきまわるスペース(自由体積)が大きくなり, エントロピー的に有利になって結晶化(Alder転移)が起こる.

用いる.[11] その意義を説明する.

同じ訓練データ $(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\})$ を用いて独立に学習している n 個の機械があると想像してみよう.(リードページの図参照)レプリカ $a=1, 2, ..., n$ はそれぞれ異なる機械である.しかし,同じ訓練データで学習しているのであるから,互いに多少とも似通ったものになっているのではないかと思われる.あるパーセプトロン■への入力に寄与するシナプス結合,またそのパーセプトロンの出力が2つの機械 a と b でどの程度似ているかは,それらの間の重なり

$$Q_{ab,\blacksquare}=\frac{1}{N}\sum_{i=1}^{N}(J^{i}_{\blacksquare})^{a}(J^{i}_{\blacksquare})^{b} \qquad q_{ab,\blacksquare}=\frac{1}{M}\sum_{\mu=1}^{M}(S^{\mu}_{\blacksquare})^{a}(S^{\mu}_{\blacksquare})^{b} \tag{5}$$

で特徴づけられる(規格化より $Q_{aa,\blacksquare}=q_{aa,\blacksquare}=1$).レプリカ a, b の配位が遠ければこれらの重なりは0となり,逆に近ければ有限の大きさになるであろう.

レプリカ系の有効ハミルトニアンは単に独立な n レプリカのハミルトニアンの和である.ハミルトニアンだけを考えると次の2つの対称性がある:(i) シナプス結合,ニューロンが正負の値を取る確率の対称性 (ii) レプリカ対称性,すなわちすべてのレプリカは同等で,添字 $a=1, 2, ..., n$ の置換に関して系は不変.訓練データによって (i) の対称性は境界で破られているが,同じ境界条件がすべてのレプリカに働くので (ii) レプリカ対称性は破られていない.$N, M\to\infty$ では,上に導入した $n\times n$ 行列 $Q_{ab,\blacksquare}$ と $q_{ab,\blacksquare}$,が秩序変数の役割をし,上の対称性 (i)(ii) の(自発的な)破れがもし起これば検知できる.ガードナーが単体のパーセプトロン■を解析した際[3] は1つの $Q_{ab,\blacksquare}$ で済んでいたのが,大幅に拡張されていることになる.

さて,訓練データをランダムに生成すると,様々なものが得られ,ガードナー体積もそれぞれに異なるであろう.一つの興味は,「典型的なサンプル」の振る舞いである.典型的なサンプルの振る舞いは,訓練データ $\mathbf{S}_0, \mathbf{S}_L$ の生成に関して平均 $\overline{\cdots}^{\mathbf{S}_0,\mathbf{S}_L}$ をとった自由エネルギー

$$\begin{aligned} -\beta F &= \overline{\log V(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\})}^{\mathbf{S}_0,\mathbf{S}_L} \\ &= \partial_n \overline{V^n(\{\mathbf{S}_{0,i}, \mathbf{S}_{L,i}\})}^{\mathbf{S}_0,\mathbf{S}_L}\Big|_{n=0} \end{aligned} \tag{6}$$

から得られる.ここで V^n は n 個のレプリカのガードナー体積の単純な積である.レプリカ数 n を実数に解析接続し,テイラー展開 $V^n=1+n\log V+O(n^2)$ を用いた.

自由エネルギーは,先に導入した秩序パラメータの関数

$$F=F[\{Q_{ab}(l), q_{ab}(l)\}] \tag{7}$$

となる.ここでレプリカ間の重なりは同じ層の中では均一で,層のラベル $l=0, 1, 2, ..., L$ にのみ依存するとしている.

上の自由エネルギーを導出する詳細[11] は省略するが,重要なポイントを述べる.まず,2種類の秩序パラメータ式(5) に共役な外場をそれぞれ導入し,これらに関する自由エネルギーを求める.これをルジャンドル変換することによって,秩序パラメータを変数とする自由エネルギー汎関数 $F[\{Q_{ab}(l), q_{ab}(l)\}]$ が得られる.この手続きが解析的

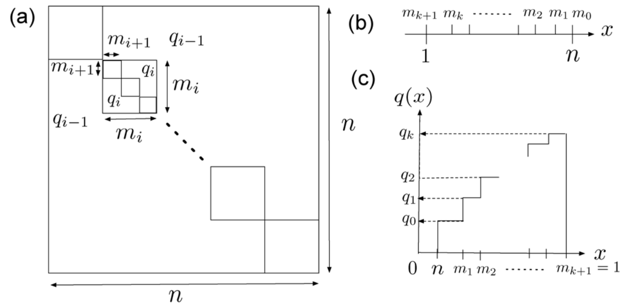


図3 レプリカ間の重なりについてのParisi仮説[14] の概念図.[11] $n\times n$ の q_{ab} 行列において階層的なブロック構造を仮定.小ブロック群の大きさは (b),対応する行列成分は (c) のようになっている.$k\to\infty$,$n\to 0$ 極限では $0<x<1$ の定義域をもつ関数 $q(x)$ によって行列の成分が表される.Q_{ab} 行列についても同じ.シナリオ (2) 教師–生徒機械の場合はレプリカ $a=0$ を追加し,教師–生徒間の重なり $Q_{0a}=R$, $q_{0a}=r$ もパラメータとする.

に実行可能であるのは,この模型において隣接する層間の結合が密で,平均場模型的(∞次元系)であるからである.*9

強磁性ならば秩序パラメータである磁化 m に共役な磁場 h を無限小に弱く掛けることによって熱力学極限 $N\to\infty$ における自発的対称性の破れを $\lim_{h\to 0}\lim_{N\to\infty}m$ によって検出できる.レプリカ間の重なりに共役な外場を導入する理由もこれと同じであり,ある種のランダム磁場を掛ける操作とも解釈できる.このことはParisi–Virasoro[13] によってスピングラスの場合について最初に議論された.自由エネルギー $F[\{Q_{ab}(l), q_{ab}(l)\}]$ を最小化することは,熱力学極限をとったのち外場を切ることに相当する.この方法によって,スピングラス[4, 9, 14] のような外的なランダムネスがない構造ガラス[8] や乱れのないスピン系[7] などにレプリカ法を導入できる.今のDNNの問題でも入出力層以外において外的なランダムネスはないので有効である.

訓練データに関する平均 $\overline{\cdots}^{\mathbf{S}_0,\mathbf{S}_L}$ について補足する.シナリオ (1) 丸暗記の場合,これは文字通りランダムな境界条件に関する平均である.またシナリオ (2) 教師–生徒機械においては,教師機械の入出力を生徒機械の境界条件として使う.教師機械を表すためにもう一つレプリカ $a=0$ を追加で導入しておき,生徒機械 $a=1, 2, ..., n$ との重なり Q_{0a}, q_{0a} も定義しておく.この場合に得られる自由エネルギーはFranz–Parisiポテンシャル[15] とよばれる.

レプリカ間の重なりは図3に示した階層構造をもつとするParisi仮説[14] を仮定する.自由エネルギー (6) を求めるために $n\to 0$ 極限を考えるが,このとき図3(c) のように,行列成分は行列の対角線からの距離に関係した変数 x の関数となる.同じ固定境界条件をすべてのレプリカが受けるため $q(x, 0)=q(x, L)=1$ である.このもとで自由エネルギー (7) の変分方程式(Parisi方程式)を数値的に解き,各層 l での秩序パラメータ関数 $Q(x, l)$, $q(x, l)$ を求めた.シ

*9 個々のシナプス結合の強さは $O(1/\sqrt{N})$ と小さい.このため自由エネルギーにおける相互作用の寄与をcumulant展開で評価でき,その際,複雑な高次項は無視できる.さらにシナプス結合の数を c,その強さを $O(1/\sqrt{c})$ としておき,$N\gg c\gg 1$ と考えると,層をまたぐループの寄与も無視できる.また前述のように,入出力層にある訓練データにおいても異なるビット間に相関がないとしている.このとき,tree近似に基づく上の自由エネルギーの表式 (7)[11] が厳密になる.

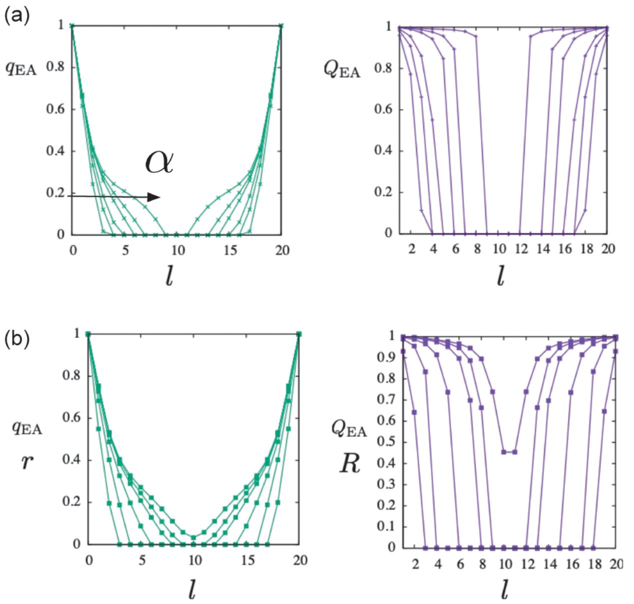


図4　DNNのデザイン空間において濡れ転移が訓練データの増大とともに進行する様子.[11] Parisi方程式の解. $L=20$の系. (a) シナリオ (1) $\alpha=50$, 100, 200, 1,000, 2,000, 4,000. (b) シナリオ (2) 教師–生徒機械 $\alpha=25$, 100, 250, 500, 625, 714. 訓練データの増大とともにガードナー体積が減少し, Edwards–Anderson秩序パラメータ $q_{\mathrm{EA}}(l)=q(1,l)$, $Q_{\mathrm{EA}}(l)=Q(1,l)$ が有限の領域が両端の入出力層近傍からネットワーク内側に成長していく. なおシナリオ (2) はベイズ最適な推定になっているため教師–生徒機械の重なりは生徒–生徒機械の重なりと等しく $r=q_{\mathrm{EA}}$, $R=Q_{\mathrm{EA}}$ である. スピングラスの西森線上でも同種の恒等式が成り立つ.[4,10]

ナリオ (2) ではこれに加え, 教師–生徒の重なり R, r を求めた.

4. DNNのデザイン空間における「濡れ転移」

まず秩序パラメータ関数 $Q(x,l)$, $q(x,l)$ のうち, 行列の対角線にもっとも近い非対角成分である $x=1$ での値, $q_{\mathrm{EA}}(l)=q(1,l)$ と $Q_{\mathrm{EA}}(l)=Q(1,l)$ に注目する. これはスピングラスにおいてEdwards–Anderson (EA) 秩序パラメータとよばれているものである.[4,9,14] 図4に示すように, シナリオ (1) 丸暗記, (2) 教師–生徒機械どちらでも訓練データ数の増大によって $\alpha=M/N$ が増大していくと, EA秩序パラメータが有限の領域が, 両端の入出力層近傍からネットワーク内側に向かって成長していくことがわかる. これはいわゆる「濡れ転移」[16] の様相である (用語解説参照). ネットワークの中央部には, EA秩序パラメータが0の「液体」領域がある. ここは解空間が広く, 同じ訓練データで学習している異なる機械が, 互いに非常に異なったものになっている. 一方, 両端付近にあるEA秩序パラメータが有限の領域は「固体」であり, 解空間が狭まっている. *10, *11

より詳しく見ると, α の増大に伴う「固体領域」の増大は, 一層一層, 逐次転移によって起こっていること, その一つ一つは連続転移であることがわかった. すなわちEA秩序パラメータが連続的に立ち上がる2次転移になっている. またシナプスとニューロンのEA秩序パラメータは各層で同時に転移している. 相転移が2次転移であることは, 自由エネルギーランドスケープの変化が滑らかであることを意味し, 学習アルゴリズムの観点からは朗報である.

シナリオ (1) で最初の層が相転移する臨界点は $\alpha_{\mathrm{g}}(1)\simeq 2.03$ であり, その後, l 層目が相転移する臨界点は $\ln\alpha_{\mathrm{g}}(l)\propto l$ のように l に対して指数関数的に増大する. シナリオ (2) でも同様である. これらのことは, 典型的な訓練データに対するDNNの記憶容量が, 層の数 L に対して指数関数的に増大することを意味している. つまりDNNの「丸暗記能力」は非常に高いと言える. 層の数 L が有限であれば, 図4(b) のように十分大きな α によって液体領域は消失する. 中央部に高い自由度が残されることは, 学習ダイナミックスにおいて中央部が系の緩和を助ける可能性を示唆する.

5. 階層性のくりこみ

前章ではEA秩序パラメータに注目した. 系のハミルトニアンにおいてはシナプス結合, ニューロンが正負の値をとる対称性は破れていないが, 境界条件で破れている. その影響が, α とともに拘束条件がきつくなることによってネットワークの内部に伝わる濡れ転移を引き起こした.

ここではもう一つの対称性, レプリカ対称性の破れ (RSB)[14] を議論する. Parisi仮説で秩序パラメータ関数が x 依存性をもつとき, レプリカ対称性, すなわちレプリカの添字に関する置換対称性が自発的に破れていることになる. シナリオ (2) はベイズ最適であるため, RSBは起こらず, 前章で得た「固体」は「結晶」(教師機械そのもの) である.

一方, シナリオ (1) では以下のようにRSBが起こり, 前章の「固体」は「ガラス」であることがわかる. 2つの秩序パラメータ関数の逆関数から, レプリカ間の重なり分布関数[14] が次のように得られる.

$$p(q,l)=\frac{\mathrm{d}x(q,l)}{\mathrm{d}q}\qquad P(Q,l)=\frac{\mathrm{d}x(Q,l)}{\mathrm{d}Q}\tag{8}$$

これは同じ訓練データで学習している任意に選んだ2つの機械の重なりの分布関数である.

図5に示すように, 秩序パラメータ関数は逐次相転移を反映して, 河岸段丘のような関数になる. 今, ある l 層がガラス転移するとき, ある $(x_{l-1}<)x_l$ で立ち上がる1段の階段ができる. このとき同時に, すでにガラスになっている1, 2, .., $l-1$ 層目でも同じ場所 x_l に階段が追加される. これを反映してレプリカ間の重なり分布関数は, 逐次相転移のたびにデルタピークが分裂して数を増やしていく. *12

この状況を概念的にまとめたのが図6である. RSBは解空間がクラスターに分裂 (ガラス転移, エルゴード性の破

*10 前に述べたようにこの問題は $1+\infty$ 次元の剛体球系に類似している. 両端付近が先に固化するのも, その方が系全体のエントロピーとして有利になるためと考えられる.

*11 α の増大とともにEA秩序パラメータの値も増大する. ジャミング (SAT/UNSAT転移) では上限の1に達する.

*12 これは学習ダイナミックスにおいて, 少しずつ訓練データを増やしながら学習させるアニーリングが有効に働く可能性を示唆する.

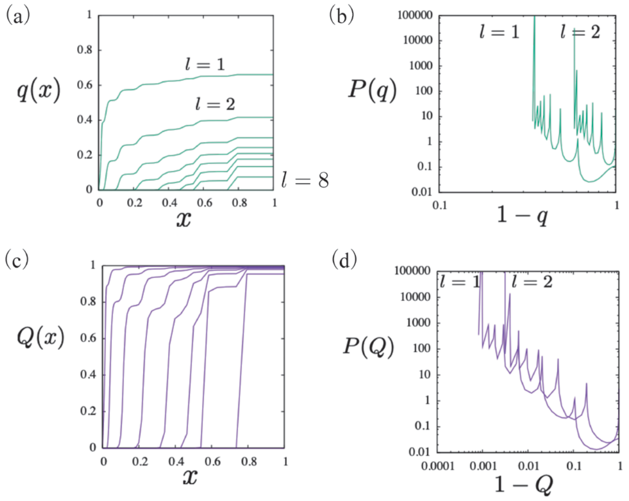


図5　シナリオ（1）での秩序パラメータ関数および重なり分布関数.[11] $L=20$, $\alpha=4{,}000$. (a), (c)：入力層側の$l=1, 2, ..., 8$での$q(x, l)$, $Q(x, l)$が表示されている（出力層側での$l=18, ..., 12, 11$はこれと同じ）. (b), (d)：これから得られる重なり分布関数（式(8)参照）$P(q, l)$, $P(Q, l)$のうち$l=1, 2$を例示.

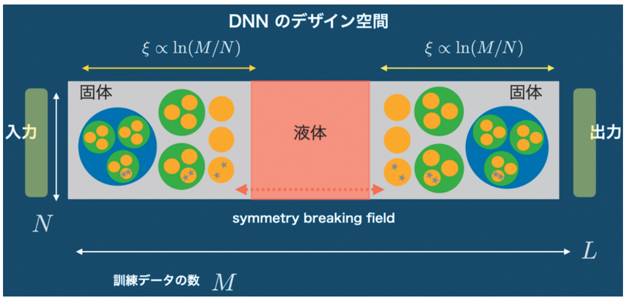


図6　DNNのデザイン空間における濡れ転移によって形成される階層的クラスターとその空間変化.[11] クラスターの大きさは含まれる解の数の多さを表す（解の総数はすべての層で同じであるが，それは表現されていない）. シナリオ（2）ではRSBは起こらず，各層に教師機械の配位を中心とした結晶（例えば黄色で表示したクラスター1つ）があると考えればよい.

れに対応）することを意味する.*13 DNNではこのクラスターが階層構造をもち，かつこれがネットワークの両端から内部に向かって段階的に，簡単なものに繰り込まれる.

6.　有限幅効果—複雑な液体

幅Nが有限の現実の系では，濡れ転移はすべてクロスオーバーとなり，熱力学的には系はつねに液体状態にある. また各層内でのニューロン■の置換に関して系は不変であるのでレプリカ間の重なり式(5)は十分長時間観測すれば0となる. これは回転対称性のある強磁性ベクトルスピン模型で，系全体の磁化が回転してしまうことに似ている.

しかし，シナリオ（1）の学習ダイナミックスの数値シミュレーションを行うとダイナミックスに特徴的な層構造が見られることがわかった.[11] ネットワークの両端付近で最もダイナミックスが遅く，複雑な多段階緩和になっており，逆に中央にいくほどダイナミックスが速く，緩和の構造も単純である. これは$N\to\infty$での理論が示唆する，空間的変化する自由エネルギー地形を反映していると考えられる.

7.　おわりに

現実にもう一歩近づくには1.3節で触れたように有限サイズNの効果の考察が重要である. また教師機械に隠れた有効次元$D(<N)$をもたせるシナリオ[12]も興味深い. これらの効果は，層間をまたぐ相互作用のループ補正[11]を評価することによって考察できると考えられる.*14

本稿で議論したDNNデザイン空間のように「端における拘束」が重要な問題は，様々な分野にある. 例えば生物においては，外界に対する望ましい入出力関係を実現するように系を進化させる遺伝子制御ネットワークやタンパク質のアロステリック効果などがある.[17] またガラスの物理においても，蒸着法を用いて層状に成長させることによって得られるultrastable-glassとよばれる熱的，力学的に非常に安定なガラス状態が注目されている.[18] 今後これらの問題にも本研究のアプローチが展開できる可能性がある.

本研究は科研費基盤研究（B）19H01812の助成を受けて行われたものです.

参考文献

1）G. Carleo et al., Rev. Mod. Phys. **91**, 045002 (2019).
2）田中章詞，富谷昭夫，橋本幸士，『ディープラーニングと物理学』（講談社，2019）.
3）E. Gardner, J. Phys A: Math. Gen. **21**, 257 (1988); E. Gardner and B. Derrida, J. Phys A: Math. Gen. **22**, 1983 (1989).
4）西森秀稔，『スピングラス理論と情報統計力学』（岩波書店，1999）.
5）小渕智之，樺島祥介，日本物理学会誌**76**, 140 (2021).
6）L. Zdeborová and F. Krzakala, Adv. Phys. **65**, 453 (2016).
7）H. Yoshino, SciPost Phys. **4**(6), 040 (2018).
8）G. Parisi, P. Urbani, and F. Zampoini, *Theory of simple glasses: Exact solutions in infinite dimensions* (Cambridge Univ. Press, 2020).
9）高山　一，『スピングラス』（丸善，1991）.
10）Y. Iba, J. Phys A: Math. Gen. **32**, 3875 (1999).
11）H. Yoshino, SciPost Phys. Core **2**, 005 (2020).
12）S. Goldt et al., Phys. Rev. X **10**, 041044 (2020).
13）G. Parisi and M. A. Virasoro, J. Phys. **50**, 3317 (1989).
14）M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond* (World Scientific, 1987).
15）S. Franz and G. Parisi, J. Phys. I, **5**, 1401 (1995).
16）M. E. Fisher, in *Statistical mechanics of membranes and surfaces*, edited by D. Nelson, T. Piran, and S. Weinverg (World Scientific, 1989). また同書のD. Nelsonによる1章，Fig. 2a.
17）A.Wagner, *Arrival of the fittest: How nature innovates* (Penguin Random House, 2014). https://youtu.be/aD4HUGVN6Ko
18）M. Ediger, J. Chem. Phys. **147**, 210901 (2017).

（2021年4月2日原稿受付）

Anatomy of Deep Neural Networks—A Statistical Mechanics Approach

Hajime Yoshino

abstract:　Statistical mechanics of a deep neural network, feedforward network of perceptrons, is studied using the replica method. We considered two scenarios 1) random scenario 2) teacher-student scenario in a Bayes optimal setting. The analysis performed in the thermodynamics limit revealed characteristic wetting transitions in the solution space.

*13 DNNを「デザインする位相空間」でのガラス転移であることに注意しておく. ここで想定しているDNNはすべての拘束条件（訓練データの入出力関係）を完全に満たした「正解」であり，それ自体はガラスではない.

*14 入出力層にある訓練データに相関をもたせると，ループ補正を通じ，学習しているネットワーク内部でも相関が生まれる.