解説

機械学習における物理が果たす役割
――量子機械学習とその現状

大関真之

東北大学大学院情報科学研究科

mohzeki@tohoku.ac.jp

機械学習では，入力と出力の関係がよくわからないものに対して，とりあえず自分たちが扱うことのできる簡素な関数を用意する．猫の姿を見て，猫だと認識するのは，入力が画像データであるのに対して，猫だという指摘をする結果をはじき出すのだからよくわからない入出力関係があるのは間違いない．しかしそれをコンピュータが実践できるように扱いがしやすい関数を用意する必要がある．ただし，その関数には変更可能なパラメータを複数含ませておき，できるだけ広い豊富な表現力をもつようにしておく．後で微調整を行い，実際に入力を与えた時に適切な出力が与えられるようにする．関数に変更を加えてなんとか辻褄を合わせるというのは変分法に相当する．変更可能なパラメータ以外にも機械学習では，非線形変換を伴う関数を利用して，その表現能力を増強することができる．ニューラルネットワークが深層化を経て，非常に複雑な非線形変換を獲得するのも，そうした目的があるためだ．

非線形変換というのは，ある種の飛躍であり，計算の難しさの象徴として現れる．例えば高校生のときに二次関数まで学んだのちに，三角関数をはじめ，指数関数と対数関数を学び，その豊富な数学の表現力に魅了される．しかし同時にその扱いに厄介さを覚えることもある．ただ慣れてくれば，その扱いもその存在も親しみをもって普段使いをする対象となる．人間というのはなかなか勝手なものである．

そうした機械学習の分野で一つ気になる用語が登場しつつある．量子機械学習である．その冠にある量子力学は理解をすんなりと許さない，厄介な対象となる代表例である．交換しない演算子を利用して微視的な自由度の変化を追う際に新しい計算方法や概念を学ぶため，私たちの感覚のアップデートを必要とするというのも大きい．この量子力学は，機械学習とは一切関係のなさそうなものである．しかし人類はそうしたものですら，より豊富な表現力を得るために，機械学習において利用する関数に取り込もうとしている．

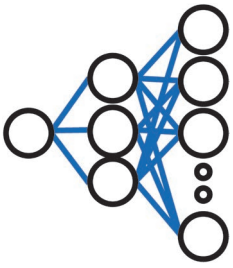
量子力学では，エネルギーの固有状態を調べるのに，原理的には高次元の行列の対角化を伴い，最終的には数値計算に頼る．また時間発展にしても，ハミルトニアンが指数関数化されて状態ベクトルにかかることにより，そう単純ではない変化を生み出し，やはり数値計算に頼る他なかった．問題設定そのものは単純に行えるとしても単純ではない変化を生み出すのが量子力学の難しさである．しかしこの部分を積極的に利用すれば，機械学習で利用されうる「複雑な非線形変換」を作り出せることが期待される．量子機械学習とはそういう営みであると換言できる．

他にもニューラルネットワークでなされる誤差逆伝播法から始まる勾配法による最適化を，運動方程式により駆動される系の変化と捉えれば，その運動方程式をシュレーディンガー方程式に置き換えるなどして量子力学の原理を導入することも考えられる．これも量子機械学習の一つのあり方と言える．

こうした自然に量子力学を取り入れようとする発想が出てきた理由は，近年注目される量子コンピュータを始めとして制御可能な量子デバイスの進展があり，実際に動作する量子デバイスを手にしているという現状にある．人類にとって，もはや量子力学は親しみをもって普段使いする対象になりつつあるのだ．

用語解説

ニューラルネットワーク： ニューラルネットワークは，入力された引数に対して非線形変換と線形変換を繰り返し非自明な関数を構築する．その繰り返しの回数に応じて，層の深さが決まり，深いほど関数の複雑度が増す．さて，その非線形変換を量子力学に現れるユニタリ変換に変更したらどうなるだろうか．



図に示すのはよく現れるニューラルネットワークの抽象的な図である．右端の丸印が入力の引数を表し，縦に並ぶ丸の数は入力の次元に対応する．青線で引かれて左側に移るたび，係数がかかり和を取るルールである．丸印を経ると任意の非線形変換が実行されて次第にその値を変えていき，最も左側ではある値に変わり出力となる．この青い線で行われることは線形変換，丸印で行われることは非線形変換であるが，それぞれの変換を変更する余地がある．それにより新しい形のモデルを作ることができる．

1. 機械学習に対するアプローチ

機械学習とは何か．たった一言で言えば，「非自明な関数を自明な関数に」コンピュータを用いて，変換することである．自明，非自明という言葉は，素性がわかっているかどうか，を表す．素性がわかっているのであれば，その関数を利用することは原理的に可能であり，その関数による入力と出力の関係を当然のことながら完璧に再現することができる．例えば犬か猫かを見た目から識別するという作業は，2次元画像を縦横に並んだ数値からなる行列またはそれを一列に並べたベクトルによる入力として，出力結果として2値を返す作業と言える．これは全くの非自明な関数による入出力関係である．それを自分が所有するコンピュータ上でプログラムされた自明な関数として利用できるようにしたいというわけだ．

1.1 モデルと損失関数

基本的な機械学習を大別すると，教師あり学習と教師なし学習に分類される．前者は入力$\mathbf{x}$に対して，出力tが出るという非自明な関数関係を仮定する．いわば背後にある真のモデルを置くことから始まる．

$$t=f(\mathbf{x}) \tag{1}$$

ここで$\mathbf{x}$はベクトルであり，tは簡単のためスカラーとした．もちろん出力をベクトルにする一般化は容易に行える．しかし上記のモデルはあくまで妄想だ．多分そうであろうくらいの意味で，全く知ることのできないものだ．我々が知ることのできるものは，入力$\mathbf{x}$に対して出力tがこのようにして与えられるという入出力関係を示すデータのみである．そのデータがいくつか与えられたデータセットをもつところから話が始まる．その意味でデータは機械学習の大前提である．

ここで注意をすると，真のモデルは，データを通してのみ知ることのできるブラックボックスなのだ．つまり全く非自明な関数というわけだ．そのブラックボックスのヒントを与えてくれるデータセットに対して，私たちの妄想を具体的に書き下す．真のモデルの形は知る由もないから$f(\mathbf{x})$とは全く異なる$g(\mathbf{x})$で表現しよう．

$$t=g(\mathbf{x})+\epsilon \tag{2}$$

ϵはいわば真のモデルと自分で用意したモデルとの誤差である．ここで問題となるのは，入力$\mathbf{x}$に対しての出力$g(\mathbf{x})$は決してデータセットの通りに出力されることはないということだ．しかしできるだけ近い方が望ましい．そこで$g(\mathbf{x})$を変化させながら$t\approx g(\mathbf{x})$となることを目指す．この過程を学習とよぶ．ここで$g(\mathbf{x})$を用意することをモデル化とよぶ．ここで自前で用意した関数をモデルとよび，モデル$g(\mathbf{x})$は自前で用意しているという意味では，「自明な」関数であると言える．自分で扱いのしやすい形にしておくことで，後々の面倒を避けるためだ．この自前の自明な関数$g(\mathbf{x})$が真のモデルからの出力tと矛盾ができるだけないようにするためには，次のような損失関数を用意する．

$$L[g]=\frac{1}{2}\sum_{i=1}^{N}(t_i-g(\mathbf{x}_i))^2 \tag{3}$$

ここでN個の異なるデータについて和を取り，データセットの中にあるデータ全てについて和を取ることを意味する．学習のために用意されたデータセットを特に訓練データセットとよぶ．この損失関数は，モデル$g(\mathbf{x})$に対する関数であるから汎関数である．この$g(\mathbf{x})$の変分を取り，損失関数が最小となれば，少なくとも目の前にあるデータセットにはあった自前の自明な関数$g(\mathbf{x})$を得ることができる．おそらくその関数は，データセットから，真のモデルの$f(\mathbf{x})$の構造を反映したものになっているのだから，試したことのない入力$\mathbf{x}$に対しても，出力tを精度よく予測できるだろうと期待できる．その期待がどれだけ信頼できるものか裏切られるものかを示す指標を汎化性能とよぶ．機械学習では，その汎化性能ができるだけ高くなるように，ありとあらゆる手段を講じる．その方法の一つが正則化である．目の前に与えられたデータセットに合わせすぎることなくモデルの形に対しての罰金項や別の項を付け加えることにより，汎化性能を保つ．

教師なし学習の場合についても同様である．教師なし学習では，$\mathbf{x}$という出力のみがあり，その出力の背後には，とりあえず真のモデルがあるとしよう．そしてそのモデルは確率分布関数からなる生成モデルとよばれ，出力をとにかくするものとする．

$$\mathbf{x}\sim P(\mathbf{x}) \tag{4}$$

ここで波線は，確率分布関数$P(\mathbf{x})$に従って，データ$\mathbf{x}$が出力されるという意味だ．いくつかのデータを引き取って，データセットをもつところから始まる．このデータには，きっと何がしかの傾向があるだろうと期待して，私たちが妄想する傾向を示す自前の自明な確率分布関数$Q(\mathbf{x})$をもってこよう．この確率分布関数から生じたものであるならば，その尤もらしさを示す尤度関数，ないしは次のような対数尤度関数が大きくなると考えられる．

$$L[Q]=\sum_{i=1}^{N}\log Q(\mathbf{x}_i) \tag{5}$$

再び汎関数が現れて，データセットと矛盾のない確率分布関数を探してこいという問題設定が浮上する．

1.2 モデルの用意

機械学習の具体的な手順においては，その汎関数の最小化ないしは最大化において，関数$g(\mathbf{x})$または$Q(\mathbf{x})$をいくつかのパラメータにより特徴付ける．教師あり学習において，単純な方法は，次の線形モデルである．

$$g(\mathbf{x})=\sum_{k}a_k\phi_k(\mathbf{x}) \tag{6}$$

この係数としてかかるa_kが変分パラメータとなり，機械学習の分野では単純にパラメータとよばれる．$\phi_k(\mathbf{x})$は既知の関数であることが多い．これは自分で設定して構わな

い．パラメータの数が多いほど，モデルの複雑さが増す．その複雑さでもって多様なデータに合わせることが可能となる．パラメータの数を無限大として，和の代わりにヒルベルト空間上での内積を採用しても差し支えない．

$$g(\mathbf{x}) = \langle \mathbf{a} | \boldsymbol{\phi}(\mathbf{x}) \rangle \tag{7}$$

ここで $|\mathbf{a}\rangle = |a_1, a_2, \cdots\rangle$ であり，$|\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \cdots\rangle$ である．この膨大なパラメータを学習する，つまり最適化をする手続きについてはやや不安を覚えることだろう．しかし二乗誤差を損失関数として，正則化としてパラメータの大きさを用意して，最もデータに合ったものを求めると，

$$|\mathbf{a}\rangle = \sum_{i=1}^{N} w_i |\boldsymbol{\phi}(\mathbf{x}_i)\rangle \tag{8}$$

という形をもつことを示すことができる（表現定理）．再びこれをモデルに戻すと，

$$g(\mathbf{x}) = \sum_{i=1}^{N} w_i \langle \boldsymbol{\phi}(\mathbf{x}_i) | \boldsymbol{\phi}(\mathbf{x}) \rangle \tag{9}$$

という形に直すことができる．無限個のパラメータの代わりに，有限のデータセットの個数分のパラメータによりモデルを定めることができるようになる．しかもその意味は明快である．w_i という重みにより各データが新たな入力 $\mathbf{x}$ に対してどの程度寄与をするのか，それを考慮して線形結合によりモデルを表現している．その際に新たな入力 $\mathbf{x}$ とこれまでの入力に相当する各データ $\mathbf{x}_i$ の「距離」を定めるのが $K(x', x) = \langle \boldsymbol{\phi}(\mathbf{x}') | \boldsymbol{\phi}(\mathbf{x}) \rangle$ である．この部分は用意した非線形変換の形により元々は決まるものであるが，モデルを決定するためには，要するに距離として適切な対称なグラム行列を用意すればよい．重み w_i を計算するには基本的には $N \times N$ 次元の逆行列の計算を必要とするだけで単純である．そこでこのグラム行列を切り替えて（カーネルトリック）モデルを用意する方法として，カーネル法がある．

この線形モデルを階層的に複雑にしていったものが，いわゆるニューラルネットワークだ．このニューラルネットワークも至言すれば，線形変換と非線形変換により構成された複雑な非線形変換 $\phi_k(\mathbf{x})$ に対して最後は線形結合を取った形 (6) をしている．

さて教師なし学習においては，$Q(\mathbf{x})$ として，

$$Q(\mathbf{x}) \propto \exp(-\beta E(\mathbf{x})) \tag{10}$$

というギブスボルツマン分布を置き，このエネルギー関数 $E(\mathbf{x})$ にデータの特性を加味した様々な形を仮定する．例えば $\mathbf{x}$ はごく少数のパターンの組み合わせで生じたと仮定すれば，

$$E(\mathbf{x}) = \frac{1}{2} \| \mathbf{x} - B\mathbf{z} \|^2 \tag{11}$$

として，いくつかのパターンを有する行列 B とどのパターンに起因しているかを示すベクトル $\mathbf{z}$ で $\mathbf{x}$ を記述することを期待する．この際 $\mathbf{z}$ の次元は $\mathbf{x}$ よりも少ない場合に元のデータ $\mathbf{x}$ に対して次元圧縮を行い，より簡素な表現を得ることを目的としていると見ることができる．時には統計力学の分野で利用されてきたイジング模型を利用することもある．その場合は特にボルツマン機械学習とよぶ．

$$E(\mathbf{x}) = -\sum_{i \neq j} J_{ij} x_i x_j - \sum_i h_i x_i \tag{12}$$

この自明な関数 $g(\mathbf{x})$ ないしは $Q(\mathbf{x})$ の「自明な」という文言をここでさらに細分化する必要が出てくる．$g(\mathbf{x})$ の構成は線形モデルにせよ，ニューラルネットワークにせよ，非線型変換と線形変換の単純な繰り返しとなっている．$\mathbf{x}$ を入力して，出力がどんなものになるのかを調べるのは非常に容易である．$\mathbf{x}$ が示すものが非常に高次元なデータであり，ニューラルネットワークの場合に何度も変換を経る場合には，その計算量が大きなものとなってくるが，それに対応した専用計算機を用いる現代ではさほど大きな問題とはならない．一方でボルツマン機械学習の場合には，自前で用意した確率分布関数 $Q(\mathbf{x})$ は，形こそ自明であれ，その結果を出力する際には，マルコフ連鎖モンテカルロ法に代表されるサンプリング計算に時間を要する．そのためボルツマン機械学習は，その表現力の豊かさの一方でイマイチ流行を迎えるということはなかった．つまり自前で作ったモデルであり，その素性がよくわかっていたとしても扱いがしやすいかどうかは全く別問題であり，計算量という観点が重要となってくる．その意味で自明といえども簡単に実行できることは意味していない．線形モデルやニューラルネットワークは，自明で簡単なものと言える．しかしボルツマン機械学習は設定は自明だけどその結果を得ることはそれほど簡単なものではない．そのため学習は闇雲な設計をしてはならない．うまい計算上の工夫や実現可能性を踏まえて設計をする必要がある．逆に言えばどんな関数を用意できるかは計算能力の向上とともに変化していくのだ．これは機械学習そのものの発展にも過去も現在でも言えることであり，そして未来についても言えることだ．

さらにモデルとデータによっては，学習においてその計算量が問題となってくる．パラメータの最適化が必要で，その最適化にかかる計算時間が膨大となってくるのだ．近年ではその計算時間の削減に計算機の性能の向上とアルゴリズムの改良により大きく進歩があった．

1.3 学習における最適化手法

さて自明な関数 $g(\mathbf{x})$ などを用意したら，最適化問題を解くことで機械はそのデータセットについて，自前で用意した関数の範囲で学ぶこと，すなわち学習を達成することができる．自明な関数に含まれるパラメータを動かし，最適なパラメータを探索するために，素朴な勾配法を利用する．まずは最も単純なものは最急降下法を利用したものだ．

$$\mathbf{a} = \mathbf{a} - \eta \mathbf{m} \tag{13}$$

ここで $\mathbf{m}$ は勾配であり，

$$\mathbf{m} = \frac{\partial L}{\partial \mathbf{a}} \tag{14}$$

と計算される．ここでηは学習率とよばれ，勾配に応じてどれだけ更新をするのかを定める．物理屋の視点では，この学習率を微小時間であると見なせば，上記の更新式はオーダーダンプなランジュバン方程式で温度0極限を取ったものである．もちろんアンダーダンプなランジュバン方程式で温度0極限を取ったものもあり，モーメント勾配法とよび，現在の勾配に前回利用した勾配を足したものを採用するものである．これにより，前の更新時の勢いを利用する．

機械学習においては，データ，モデル，最適化手法の3要素を駆使して，データセットにうまく合わせること，汎化性能をよくすることを目標とする．その方針で進化してきたために，物理的なメカニズムとの接点を意識するよりも，とにかく汎化性能をよくするものを求めてきた経緯がある．例えばよく利用されているのは鞍点やプラトーを抜け出すようなメカニズムを導入したAdam（Adaptive Momentum）というものであり，パラメータの更新則は以下の通り，少し変更が加えられている．

$$\mathbf{a}=\mathbf{a}-\eta\frac{\mathbf{m}}{\sqrt{\mathbf{v}}+\epsilon} \tag{15}$$

また勾配は一次と二次の項に注目して以下のような量を計算する．

$$\mathbf{m}=\beta_1\mathbf{m}+(1-\beta_1)\frac{\partial L}{\partial \mathbf{a}} \tag{16}$$

$$\mathbf{v}=\beta_2\mathbf{v}+(1-\beta_2)\frac{\partial L}{\partial \mathbf{a}}\odot\frac{\partial L}{\partial \mathbf{a}} \tag{17}$$

全く物理では見たこともない更新式であろう．ここで$\odot$はベクトルの成分ごとの積を表す．式(15)のように分母に勾配の大きさを入れておくことで，あまりうまくパラメータが更新されていない方向については勾配を実効的に大きくする作用がある．これにより，鞍点やプラトーの抜け出しを効率的に行う．

それでは物理学で見られるようなダイナミクスに基づく更新則は，機械学習では議論されないのであろうか．もちろん存在する．有限温度のオーバーダンプなランジュバン方程式を利用することにより，最適なモデルの周辺や，極値で特徴付けられるモデルをサンプリングする手法[1]が提案されている．

$$\mathbf{a}=\mathbf{a}-\eta\mathbf{m}+\sqrt{2T\eta}\epsilon \tag{18}$$

ここでϵは平均0，分散1のガウス分布に従う確率変数である．他にも局所エントロピーという量を導入して，暫定的なモデルの周辺におけるエントロピーを加味して，損失関数とともに自由エネルギーを最小化する手法も提案されている．エントロピー勾配降下法とよばれる．[2] まず仮想的な時間発展

$$\mathbf{b}_{t+1}=\mathbf{b}_t-\eta\mathbf{m}+\sqrt{2T\eta}\epsilon+\gamma(\mathbf{a}-\mathbf{b}_t) \tag{19}$$

をT時刻だけ進めて，その結果を元に

$$\mathbf{a}=\mathbf{a}-\eta'\gamma(\mathbf{a}-\langle\mathbf{b}\rangle_T) \tag{20}$$

とパラメータを更新する．ここで$\langle\mathbf{b}\rangle_T$は$T$時間における経験平均$\langle\mathbf{b}\rangle_T=\sum_{t=1}^{T}\mathbf{b}_t/T$である．この$\mathbf{b}_t$の更新により$\mathbf{a}$の周辺で典型的な状態へ到達させて，その結果を受けて$\mathbf{a}$を更新することからエントロピー勾配降下法という名称がついている．より詳しくは原論文に記載されているが，局所エントロピーを定義して，それぞれ勾配を計算すると，上記の更新則が自然に得られる．

2. 量子機械学習へ

それでは機械学習に量子力学を導入したらどうだろうか．いわば量子機械学習というものである．発想の飛躍がありそうに映るかもしれないが，割と自然な発想である．機械学習の構成要素は，データはもちろんのこと，モデル，損失関数とその最適化手法である．データが量子系そのものであるという対象は，多くの研究があるわけではない．そこで後の2つについて残る誌面で紹介することにしよう．まずは損失関数とその最適化手法に量子性を導入してみよう．

2.1 量子勾配法

素朴にはトンネル効果により，学習の途中に現れる損失関数の極小を抜け出すようなメカニズムがうまく働くのではないだろうか．パラメータを自由度として，それに対して非可換な運動量を導入する．損失関数に運動エネルギーを追加する．その際に量子系を導入して，その密度演算子を

$$\hat{\rho}\propto\exp\left(-\beta L[\hat{g}]-\frac{\hat{p}^2}{2\lambda}\right) \tag{21}$$

とする．ここで$\hat{g}$はモデルに含まれるパラメータ$\mathbf{a}$が演算子$\hat{\mathbf{a}}$となったことを示している．これに対して，運動量演算子$\hat{p}$とは$[\hat{a},\hat{p}]=i$という交換関係をもつ．ここで量子系をシミュレートするために演算子からc数にするために経路積分表示をする．その結果として密度演算子から虚時間方向に並んだ巨大な系の確率分布関数を得る．

$$P(\mathbf{a}_{t=0},\mathbf{a}_{t=1},\cdots,\mathbf{a}_{t=T})\propto\prod_{t=1}^{T}\exp\left(-\beta L[g^t]-\frac{\lambda}{2}\|\mathbf{a}_t-\mathbf{a}_{t-1}\|_2^2\right) \tag{22}$$

この指数の肩の部分を実効的な損失関数と見立て，有限温度（逆温度β）でのダイナミクスを考えると，その損失関数をポテンシャルエネルギーとして，オーバーダンプなランジュバン方程式を構成すれば，以下のような勾配法を構築することができる．

$$\mathbf{a}_t=\mathbf{a}_t-\eta\mathbf{m}+\lambda(2\mathbf{a}_t-\mathbf{a}_{t-1}-\mathbf{a}_{t+1})+\sqrt{2T\eta}\epsilon\,. \tag{23}$$

ただし虚時間の長さだけの並列したモデルを用意するため，実際上の応用を考慮した手法としては非合理的であるが，量子力学的要素がもたらす効果を見たいということで試してみる．その結果，図1に示すように汎化性能が上がる．[3] 図1に示す例は，有名な手書き文字認識のデータセットMNISTに対する識別を例にしたものである．ただし

MNISTの識別は現在では，最も簡単なタスクの一つであり，それこそAdamによる最適化を多層のニューラルネットワークに対して行えば，非常に高い識別率を示す．この実験では2層のニューラルネットワークに対して，訓練データセットを少なめに用意して，MNISTであっても識別率を向上させるのに骨が折れるような状況にしている．そのような条件設定で，量子揺らぎの導入により汎化性能の向上をみた．全く同じ勾配率で，MNIST以外にも顔認識のデータセットによる検証も行われ，同様の結果を得ている．[3] 量子ゆらぎによる項を単純に正則化として捉えることもできるが，物理学由来のメカニズムによる理解をすることができる．

量子効果を完全に手で追うのは困難なため，古典解を求め，そこからのゆらぎを調べる．どのような古典解が選択されているのかを調べるためだ．素朴にはその古典解の間を遷移するのが，量子勾配法のダイナミクスとなる．各虚時間ごとに古典解$\mathbf{a}$からのゆらぎを$\mathbf{a}_t=\mathbf{a}-\mathbf{b}_t$として表す．ここで古典解とは$\mathbf{a}_t=\mathbf{a}$とした場合に損失関数を最小化する解である．$\mathbf{b}_t$は大きさが非常に小さい，古典解からのゆらぎを示す．そこから量子効果を示す部分をフーリエ変換を駆使して対角化を行うと，そのゆらぎについての独立な2次形式を得ることができる．それぞれ$\mathbf{b}_t$について積分をして，$\mathbf{a}$に関する周辺確率を取り出す．解として選ばれるのは，この周辺確率が最大となるところであるから，この周辺確率の下限を代わりに最大化することを考える．代理関数法とよばれる手法だ．

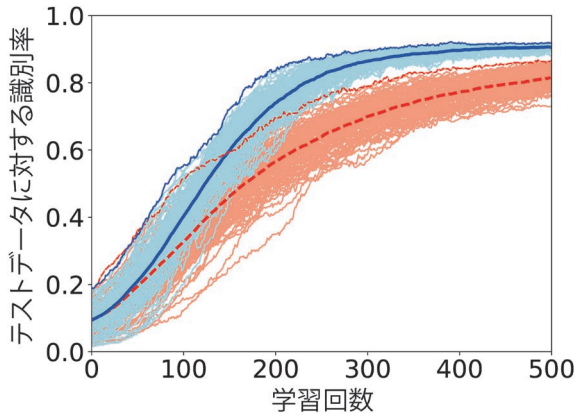


図1　量子勾配法による汎化性能の様子．赤がAdam，青が量子ゆらぎを導入したAdamによる学習を経た場合の識別率である．多数の実験による性能を重ね打ちしている．平均を青の太い線，赤の太い点線で示した．

$$P(\mathbf{a})\geq\prod_{t=1}^{T}\int d\mathbf{b}_t\,\exp\left(-\beta L[g^t]-\frac{\gamma}{2}(\mathbf{a}-\mathbf{b}_t)^2\right)\qquad(24)$$

ここでγは元の$P(\mathbf{a})$で2次形式をなす行列の最大固有値から決まる定数である．各虚時間ごとに右辺の指数関数の肩の部分をポテンシャルエネルギーとしたオーバーダンプなランジュバン方程式を，$\mathbf{a}_t$について微分をとり構築すると，式(19)を得る．ここでT個のアンサンブル平均の代わりに，ランジュバン方程式における時間平均を採用することにする．そして$\mathbf{a}$の更新においても勾配を計算すれば，式(20)を得る．$\mathbf{a}$の更新を繰り返すことにより$P(\mathbf{a})$に対する下限を更新して，周辺確率の最大化が近似的に達成される．残念ながらモデルによっては，非凸な構造をもつため極大化にとどまる．つまり量子勾配法の古典解周りの挙動は近似的に，エントロピー勾配降下法の挙動と一致することがわかる．エントロピー勾配降下法では，局所エントロピーと損失関数からなる，いわば自由エネルギーの最小化を経て，平坦な極小解に落ち込みやすいという性質がある．量子勾配法は，そのような平坦な極小解を求めて，トンネル効果により谷同士を飛び越えるという性質をもつことがわかる．

さて平坦な極小解に陥ることは機械学習への応用を考えた際に顕著な性質が得られる．汎化性能に対する影響である．汎化性能とは，パラメータの最適化に用いられた「訓練」用のデータセットとは異なる，「テスト」用のデータセットに対して，同様のタスクを実施した際にどれだけ良好な性能をもつかを示す指標である．基本的には訓練用のデータセットとテスト用のデータセットにはある程度の相関があり，似たものであるはずだ．図1にあるような，幅に違いのある極小の谷を2つ考える．訓練用のデータセットによる損失関数で狭い極小の谷に落ち込んだ場合は，そのパラメータのままでテスト用のデータセットによる損失関数においては極小の谷とはよべないところに引っかかってしまう．しかし広い谷に落ち込んだ場合には，訓練用でもテスト用の損失関数であっても，さほど大きな差が生じない（図2）．つまり汎化性能を引き上げるためには，広い谷に落とすような工夫が必要となる．量子勾配法及びエントロピー勾配降下法は，平坦な極小を好むことから汎化性能を引き上げるメカニズムを有していることがわかる．ただし量子勾配法そのものは虚時間にわたる並列的なモデルを用意する必要があり，実用上は有用ではない．もしも量子系があれば話は別なのだが．

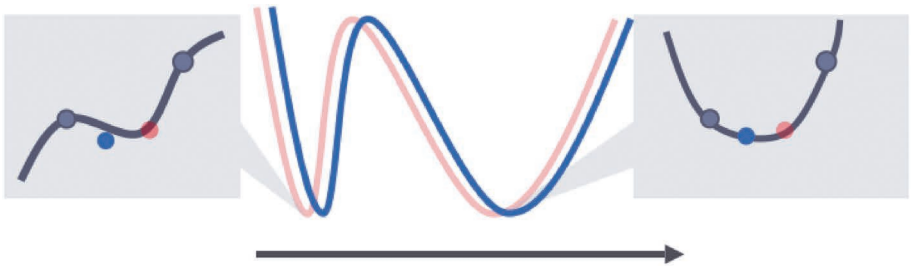


図2　平坦な極小は汎化性能が高い．赤が訓練データ，青がテストデータであり，グレーは両者に共通するデータであるとする．赤の訓練データに合わせる際に利用する損失関数の振る舞いを中央部に赤線で示した．青のテストデータに対する損失関数を重ねている．広い極小の谷は訓練データとテストデータに対する合わせ方として両者にうまく合う形であり，狭い極小の谷では訓練データに過剰に合わせている．

2.2　モデルの量子化

勾配法，すなわち最適化手法に量子性を導入するという量子機械学習の型があれば，モデルに量子系を導入するという型もある．まずはその背景について紹介しよう．量子系を扱う研究は，数値計算による手法が主流である．原理はシンプルな割に，その時間発展をつぶさに追うのが難しいためである．しかし量子力学に従う実験系を作ってしまえば，そのような計算量困難とは無縁となる．耳にしたことがあるかもしれないが，最近これらの量子力学に従う実験系として，量子デバイスの開発が急速に進んでいる．その究極的な形が量子コンピュータであろう．量子コンピュータでは任意のユニタリ変換を施すことが可能なデバイスである．

任意のユニタリ時間発展演算子を実行することができれば，シュレーディンガー方程式に基づく時間発展をシミュレートすることができる．これまではデジタルコンピュータ上で，そのシミュレートをすることで，我々は量子系の振る舞いについて予言を行い，実際の実験とともに比較して，その理解につなげていた．逆に言えば，これが計算量の壁を感じさせる源である．量子系をデジタルコンピュータ上でシミュレートしようとしているから計算量の壁を感じるのだ．むしろ，デジタルコンピュータ上でのシミュレートではなく，“量子”コンピュータ上でのシミュレートを行うことでシュレーディンガー方程式通りの時間発展をするとしたらどうだろう．

現状の量子コンピュータでは，限られたゲート操作回数で動作するのみであり，そして操作ごとに生じるエラーを訂正する機構を実装することなく出力を得るものになっている．理論通りの量子コンピュータと区別して，Noisy Intermediate-Scale Quantum computer（NISQ）とよばれる．[4] そうは言ってもせっかくできた操作のできる量子系だ．これを使わない手はない．

全く非自明な量子系の基底状態のエネルギーを求めたいとしよう．ハミルトニアン自体$\hat{H}$は自明である．形はわかっている．しかしその固有状態$|\Psi\rangle$は非自明ということで，非自明な量子系と称した．その量子系の基底状態のエネルギーを求めるにはどうしたらよいだろうか．機械学習のアプローチと同様に，自明な関数をもってくればよい．そこでもち出してくるのは，自前の自明な量子系をもってくるという方法だ．その自明な量子系は，いくつかのユニタリ時間発展を通して自明な量子系を用意するとしよう．

$$|\Psi(\mathbf{a})\rangle = \prod_{t=1}^{T} U(\mathbf{a}_t)|\psi\rangle \tag{25}$$

ここで$\mathbf{a}_t$は各ステップで実行されるユニタリ変換を特徴づけるパラメータである．単一パラメータにより特徴づけされるユニタリ変換もあれば，2つのパラメータで特徴づけされるものもある．$|\psi\rangle$は比較的用意するのが容易な量子状態にして，そこから順次ユニタリ変換を実行することで様々な量子系を作ることを想定している．これはちょうどニューラルネットワークに入力をしたのち，線形変換と非線形変換を繰り返すことと同様である．つまり量子系を作るためにどのようなユニタリ変換を構成するのか，というものがモデルを作る部分に相当する．[5]

さてこのように用意した自明な量子系のパラメータを動かして，非自明な量子系の基底状態のエネルギーを探る．自明な量子系によるハミルトニアンの期待値を以下のように定義する．

$$E(\mathbf{a}) \equiv \frac{\langle\Psi(\mathbf{a})|\hat{H}|\Psi(\mathbf{a})\rangle}{\langle\Psi(\mathbf{a})|\Psi(\mathbf{a})\rangle} \tag{26}$$

この期待値が最小となるような自明な量子系を探せというのが問題となる．自明な量子系は自らが構成してパラメータを用意しているから，上記の期待値をパラメータに関する微分を行い，パラメータの更新を行う勾配法を設計することは容易である．ユニタリ変換の順番や構成が量子コンピュータにおける回路の構成に相当するので，この方法を変分量子固有状態法（Variational Quantum Eigensolver）[6] とよび（図3），量子機械学習の一つのトレンドを作っている．

上記はいわば参考になるデータが存在しないという意味では教師なし学習である．

それでは量子機械学習の発想で，教師あり学習に相当することをどのように行えばよいだろうか．要するに自明なモデルとして量子系を用意すればよいと考えると発想は次々と出てくるであろう．ユニタリ変換により非線形変換を行うモデルを用意するのが素直な発想である．

$$|\Psi(\mathbf{x})\rangle = U(\mathbf{x})|\psi\rangle \tag{27}$$

まずデータxに対して決められたユニタリ変換を用意する．これが非線形変換に相当するということに注目すれば，その内積により決まる次の量は，カーネル法におけるグラム行列に相当することがわかる．

$$K(x', x) = \langle\Psi(\mathbf{x}')|\Psi(\mathbf{x})\rangle \tag{28}$$

図3　変分量子回路の様子．

既存のカーネル法と同様にカーネルトリックによりグラム行列として量子系の内積を利用すればよいという素朴な発想が浮上する．

ただしその利用には注意があり，データの数が大きくなればなるほど，グラム行列の行列要素を準備するのに時間がかかり，さらにパラメータの決定に必要な逆行列計算に伴う時間については別に方法を考える必要がある．幸いなことに後者の逆行列計算は，ユニタリ変換を自在に操る量子コンピュータを用いることで高速化できる．将来にはその点の問題を解消することがある程度期待できる．[7] 実際にグラム行列を伴うサポートベクターマシンの学習における逆行列演算を効率的に行うアプローチも提案されている．[8] また逆行列計算については，量子コンピュータを利用した計算を参考にして，脱量子化，すなわち古典的なアプローチでも高速に逆行列計算が実行できることが明らかとなっている．[9] そうは言っても量子機械学習という仰々しい名前がある割には，変わったグラム行列を用意するのみであるといった使い方では芸がない．非線形変換に続けて，変分可能な（学習可能な）ユニタリ変換をさらに付け足したアプローチが提案されている．[10]

$$W(\mathbf{a})|\Psi(\mathbf{x})\rangle \tag{29}$$

最終的に出力の基底による観測を通して，出力結果の確率分布を得る．

$$P(y) = |\langle y|W(\mathbf{a})|\Psi(\mathbf{x})\rangle|^2 \tag{30}$$

これはちょうど非線形変換の後に線形結合によりモデルとする式(7)の形への帰着とも言える．入力$\mathbf{x}$に対して非線形変換を通して，重み$W(\mathbf{a})$で線形結合を行う2層のニューラルネットワークと捉えることもできる．多層化を行うことも容易であり，その誤差逆伝搬法を構築することも系統的に行うことができる．

量子系で自明なモデルを用意するということであれば，ギブスボルツマン分布を作るというのも手である．ただしその場合は密度行列がモデルとなる．例えば量子アニーリングでよく用いられる横磁場をもつイジング模型をハミルトニアンとして，密度行列$\hat{Q}$を

$$\hat{Q} = \frac{1}{Z}\exp\left(-\beta E(\hat{\sigma}^z) + \beta\Gamma\sum_{i=1}^{N}\hat{\sigma}_i^x\right) \tag{31}$$

としよう．ここで$E(\hat{\sigma}^z)$は式(12)であり，$\hat{\sigma}_i^z$及び$\hat{\sigma}_i^x$はパウリ行列のz成分，x成分である．この量子ボルツマンマシンともよぶべき自明なモデルが相手にする非自明なモデルは確率分布関数$P(x)$である必要はなく，より一般の量子系の密度行列$\hat{P}$である．これらの密度行列の距離として，量子相対エントロピーを用いる．

$$S(\hat{P}|\hat{Q}) = \mathrm{Tr}\left(\hat{P}\log\hat{P} - \hat{P}\log\hat{Q}\right) \tag{32}$$

この量子相対エントロピーの最小化を通して，できるだけ$\hat{P}$に近い$\hat{Q}$を求めることが目的となる．これまでと同様にパラメータに関する微分をとり，勾配法により更新を行い最適なパラメータの探索をする．例えば相互作用J_{ij}に関する微分は次の関係を与える．

$$\frac{\partial}{\partial J_{ij}}S(\hat{P}|\hat{Q}) = \mathrm{Tr}\left(\sigma_i^z\sigma_j^z\hat{P}\right) - \mathrm{Tr}\left(\sigma_i^z\sigma_j^z\hat{Q}\right) \tag{33}$$

量子系の期待値の差をできるだけなくすことで目標が達成することが見て取れる．前者は対象となる非自明な量子系に関する観測結果である．後者は自らが用意した自明な量子系に関する期待値で，計算をすることが求められる．しかしながら量子系の期待値を計算することは容易ではない．特に有限温度の期待値を計算するためには膨大な計算量を必要とする．ここで量子デバイスの登場というわけだ．例えば量子アニーリングマシンとして知られるD-Wave 2000Q（現在はD-Wave Advantageも利用可能）を用いるなどが候補となる（図4）．しかし残念ながら量子アニーリングマシンは現在のところパウリ行列のz方向の基底のみで観測できるものであり，任意の方向の観測が可能ではない．そのため例えば横磁場に関する微分で現れる物理量について，その観測を行うことができないため，正しく横磁場の値について学習を行うことができない．そこで横磁場を固定されたパラメータとして，ある種の正則化として捉えることがある．その効果により，横磁場のないボルツマンマシンに対して，量子ボルツマンマシンの方がよい性能を与える可能性があることがいくつかの研究で指摘されているが，それが本質的に量子系ならではの結果かどうかは，決定的なことはまだ言えない．また用意できる量子系が横磁場をもつイジング模型のみであり，モデルの範囲が制限されている．実際のデバイスについてはそういう状況ではあるものの，量子ボルツマンマシンは量子系を用いることで，指数関数的加速を示す例[11]であり，対象となる非自明なモデルが量子系ではないものであっても，グローバーの探

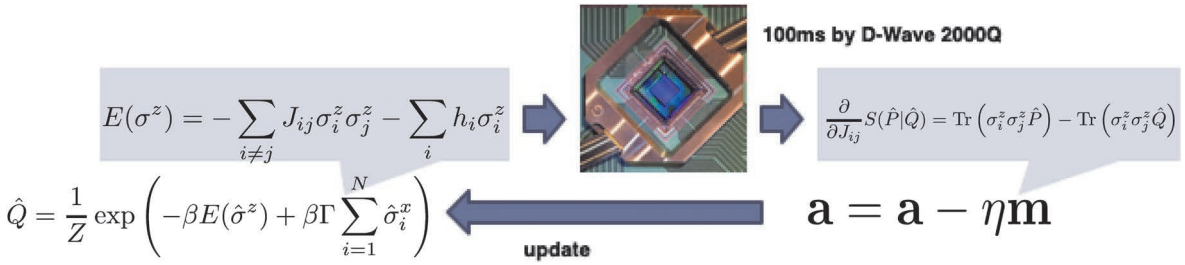


図4　量子ボルツマンマシンの学習の様子．

索アルゴリズムと同様な二乗の加速を示す[12]ことが知られている．その意味では，今後量子デバイスの開発が進み，用意できる量子系の範囲が広がることで量子ボルツマンマシンの有効性が向上することが期待される．

3. 機械学習と物理学の関係

最後に機械学習はデータに根差した一手法にとどまらず，そのものの概念が，物理学と非常に馴染みのあるものであるということを強調しておきたい．特に量子力学では，そもそも非自明な波動関数を自明な関数で表現しようと変分的アプローチを取ってきたし，他にもフーリエ変換が行っていることも非自明な関数を自前で用意した三角関数で理解しようとする営みだ．機械学習ではその自明な関数には多くのパラメータが含まれるため，計算を通してデータに合わせるというスタイルを取る．物理学ではごく少数のパラメータをもつ洗練されたモデルが考案され，実験をいかにうまく広く説明するのかにより判定され，いくつも提案されて廃れての繰り返しを続けてきた．その手続きを系統的に行うことができれば，どれだけ効率的になるだろうか．ただ系統的に行うのはよいとしても，その解釈や起源をどこに求めたらよいのだろうか．次なる悩みも抱える．そうした悩みは尽きないにしても，量子機械学習のような物理学と機械学習がマージした格好で発展しつつある分野の登場は，物理学者としてはなかなかに味わい深いものではないだろうか．物理としての最大の武器である「自然に聞けばわかる」というところをよりどころにしている点は注目に値する．人工的に構築した量子デバイスは，その自然に問いかける実験装置である．そうした意味では量子機械学習はまさに物理学の営みである．物理学における数値シミュレーションが実験の代わりを果たすようになり，次第に非自明な振る舞いを引き出して，その価値を見出されたように，量子機械学習も今後その価値を強く示すようになるだろう．また機械学習の機械の意味が，デジタルコンピュータである時代から，量子コンピュータを始め，もっと広い機械を相手にした意味合いに変わりつつある．量子デバイスの活用が機械学習の方法論の一つに組み込まれ，物理学そのものだけでなく機械学習の発展にも寄与することとなる．そうした両分野を刺激するのが，量子機械学習という分野であり，単なるブームや流行り言葉で終わることはないだろう．短いながらも一人でも多くの読者が共感したり刺激を受けることを願って，とりあえずここまでとしよう．

参考文献

1）M. Welling and Y. W. Teh, in *ICML'11: Proceedings of the 28th International Conference on Machine Learning* (Omnipress, 2011) p. 681.
2）P. Chaudhari et al., J. Stat. Mech. **2019**, 124018 (2019).
3）M. Ohzeki et al., Sci. Rep. **8**, 9950 (2018).
4）E. Grumbling, M. Horowitz 編，西森秀稔訳，『米国科学・工学・医学アカデミーによる量子コンピュータの進歩と展望』(共立出版，2019).
5）P. J. J. O'Malley, Phys. Rev. X **6**, 031007 (2016).
6）A. Peruzzo et al., Nat. Comm. **5**, 5213 (2014).
7）H. A. W. Harrow, A. Hassidim and S. Lloyd, Phys. Rev. Lett. **103**, 150502 (2009).
8）P. Rebentrost, M. Mohseni, and S. Lloyd, Phys. Rev. Lett. **113**, 130503 (2014).
9）E. Tang: in *STOC 2019: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (ACM, 2019) p. 217.
10）M. Schuld and N. Killoran, Phys. Rev. Lett. **122**, 040504 (2019).
11）M. Kieferovà and N. Wiebe, arXiv:1612.05204 (2016).
12）N. Wiebem, A. Kappor, and K. M. Svore, arXiv:1412.3489 (2014).

著者紹介

大関真之氏： 量子アニーリングをはじめ，量子力学と統計力学を利用して情報科学の問題に取り組む．飽きっぽい性格なので，次はどの分野で遊んでみようかと常に考えている．

（2020年5月30日原稿受付）

Role of Physics in Machine Learning —Quantum Machine Learning and Its Current Status

Masayuki Ohzeki

abstract: Machine learning consists of three essential parts as data, model, and optimization. We review these in short and then consider several generalizations into their quantum versions, namely quantum machine learning. In optimization, we consider a generalization of the loss function with a noncommutable operator. The result demonstrates the superiority in generalization performance when we add this noncommutable operator. Next, we discuss quantum machine learning with the quantum model. The implementation of the quantum model is now realizable by various quantum devices appearing so far. How to use the quantum device and compute several necessary quantities are described.