数理解析研究所講究録
第1565巻 2007年 124-132



# 数理物理学と学習理論*

東京工業大学精密工学研究所
渡辺澄夫
〒226-8503 横浜市緑区長津田4259 メールボックス R2-5

平成19年5月31日

**概要**

ある確率分布から発生したサンプルが与えられたとき、サンプルから確率分布を推測することを、統計的推測あるいは学習といい、そのための理論を学習理論という。「学習理論」とは、もともとは心理学の言葉であり、人間や動物の学習に見られる現象や法則を研究する場合に用いられていたが、近年では、コンピュータによってサンプルから確率分布を推測する問題に対しても使われるようになってきた。本論は、後者の意味での学習理論について述べる。一方、数理物理学は、新しい数学的概念の創出により物理学を数学的に建設し、数学と物理学の両方を深く理解するための学問である。場の量子論や統計力学など広範な物理学の研究において数学的概念の創出が大切であることはよく知られている。本論では、数理物理学と学習理論との間にごくわずかながら数理的な関連がある可能性について紹介する。ここで述べる内容は、既に知られていることであり、新しいものではないが、数理物理学と学習理論が、あまりにも遠く隔たっているために、両者の間の小さな関係について紹介されることはほとんどないのが実情である。そこで、ここでは、その小さな関係について初めて出会う人に説明することを目的とする。

## 1 はじめに

場の量子論や統計力学では、無限次元空間上の確率分布およびその概念の拡張が中心的な問題になるのであるが、本論では、有限次元の空間の上の確率分布について考える。

$(\Omega, \mathcal{B}, P)$ を確率空間として、$X$ を $N$ 次元ユークリッド空間 $\mathbf{R}^N$ に値を取る確率変数とする。また $X_1, X_2, ..., X_n$ を $X$ と同じ確率分布に従う独立な確率変数とする。$N$ 次元の空間とは別に $d$ 次元のユークリッド空間 $\mathbf{R}^d$ を考える。この上に確率分布 $\varphi(w)dw$ が定義されているとする。また $\mathbf{R}^N \times \mathbf{R}^d$ 上の関数 $f(x,w)$ が定義されているものとする。学習理論で大切な役割を果たす概念は、数学的には次のような形をしている。

$$F = -\log \int \exp(-\sum_{i=1}^{n} f(X_i, w))\varphi(w)dw$$

関数 $f(x,w)$ と確率分布 $\varphi(w)dw$ に適切な数学的な条件を与えると $F$ は実数に値を取る確率変数になる。数学的な課題は次のようなものである。

問題「関数 $f(x,w)$ および $\varphi(w)$ に自然で適切な数学的条件を与えることにより、確率変数 $F$ の挙動を（特に $n \to \infty$ において）明らかにせよ」

この問題をより具体的に説明すると次のようになる。

問題「上記の問題を考察するとき、$f(x,w)$ の $w \in \mathbf{R}^d$ の代数的な構造として何を考察するのが自然であるか、その代数的構造を考察するために集合としての $\mathbf{R}^d$ に、どのような幾何学を入れるのが適切である

*この小論は、2006年12月に数理解析研究所において開催された研究会「量子解析におけるミクロマクロ双対性」（代表：小嶋泉先生）における招待講演の記録です。

か、その代数的および幾何学的な構造の洞察に基づいて平均 $\int \varphi(w)dw$ を実行して、確率変数 $F$ の挙動を導出せよ。」

ということになる。この問題は数学的には解決されていない部分も含むので、本論では、問題を提起するにとどめる。残りの文章では、「なぜ、この問題が産業・社会・科学の中で大切であるのか」を説明する。

**注意.** 量子情報を取り扱う場合には $f(X,w)$ は、非可換環に値を取るケースが重要になる可能性がある。その場合について考察する能力を著者は有していないので、本論では、可換環に値を取る場合についてだけ述べる。

**注意.** この問題は既に解決されていると感じる読者も多いかも知れないが、関数 $f(X,w)$ が $w$ について多項式である場合でも、$F$ の挙動は最近まで知られていなかった。その場合の解決においては「$\mathbf{R}^d$ について双有理同値な変換の中で、$f(x,w)$ を $w$ について正規交差になるようなものを見つけよ」という問題が、$F$ の挙動を解決する問題と密接な関係を有している。

**注意.** 上記で述べたことは、古典統計力学において有限次元の空間上のランダムハミルトニアン

$$H(w) = -\sum_{i=1}^{n} f(X_i, w)$$

から定まる自由エネルギーが従う確率分布の挙動を調べよ、という問題とよく似ている。本論の「数理物理学と学習理論のわずかな関係」という言及は、どちらにおいても $F$ が似た形をした関数であるという点だけに立脚しているたいへんに弱いものである。学習理論で現れる $F$ は、数理物理学で現れる自由エネルギーとは、次の２つの点が大きく異なる。
(1) 学習理論では $F$ の $n$ による挙動の変化が問われる。$F$ を $n$ の関数として $F(n)$ と書くことにすると、確率変数 $F(n+1)-F(n)$ の $n \to \infty$ における漸近挙動が極めて重要である。大偏差原理と類似する方法で、ある $w_0 \in \mathbf{R}^d$ が存在して、確率収束

$$\frac{F(n)}{n} \to E[f(X, w_0)]$$

を示すことには意義があるが、応用上は、より高次の挙動が解明されることが望ましい。$F(n)$ の $n$ 未満のオーダーの項が、$X$、$f$、$\varphi$ のどのような数学的性質によって定まるかを明らかにすることが、応用上も数学上も大切なことである。
(2) 統計力学では、気体の状態方程式、スピンの相転移、アモルファスの磁性など、多種の自然現象が考察されている。自然現象で重要になるハミルトニアンの形と、学習理論において重要になるハミルトニアンの形は、まったく似ていないことが普通であって、自然科学の中で普遍的に見られた現象（相転移など）が学習理論の中でも同じように生じるわけではない。上記で $F$ を「自由エネルギー」と呼んでいるものの、物理学における分配関数の計算例で似ているものはあまりないようである。（例えば、格子ゲージ理論やスピンシステム理論では、考察している格子空間の次元が大切になることが多いため、『学習理論では何次元の問題を考えるのか』という質問が多くなされるが、学習理論で考察する問題は、その意味での次元は陽には出てこないように思われる。）。

## 2 準備

この文章は、数理物理学の研究をしている先生方への紹介として書いているので、$F$ を自由エネルギーと呼ぶことにするが、上記で述べた理由により、自由エネルギーは、数理物理学の自由エネルギーと形が少しだけ似ているものにすぎず、自然科学としての意味は持っていない。以下では、物理現象ではない問

題において、この自由エネルギーがどんな意味を持つかについて紹介する。なお、以下で紹介する内容は、著者の独創ではなくて、統計学や情報理論では、よく知られていることである。

**相対エントロピー.** 有限次元のユークリッド空間上に定義された二つの正値で連続な確率密度関数 $p_1(x)$, $p_2(x)$ について、相対エントロピー $D(p_1||p_2)$ を

$$D(p_1||p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx$$

と定義する（積分が有限確定値になる場合だけを考える)。このとき、

$$S(t) = \log t + \frac{1}{t} - 1 \quad (0 < t < \infty)$$

と定義すると $S(t) \geq 0$ で、$S(t) = 0 \Leftrightarrow t = 1$ であり

$$D(p_1||p_2) = \int p_1(x) S(p_1(x)/p_2(x)) dx$$

であることから $D(p_1||p_2) = 0$ は $p_1(x)$ と $p_2(x)$ が等しい関数のときのみ0であり、それ以外では正の値になることがわかる。

相対エントロピーの性質から、次のことがわかる。ある確率変数 $X$ が与えられたとき、確率分布 $p(x)$ の汎関数

$$L(p) = E[-\log p(X)]$$

を考えると、$L(p)$ は、$p(x)$ が $X$ が従う確率分布であるときに限り最小値をとる。情報科学や統計学においては、$X$ のサンプル $x_1, x_2, ..., x_n$ に対して、経験エントロピー

$$L_n(p) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i)$$

を定義して、この値から $p(x)$ を推論することがしばしば行われる。

**注意.** 「データから確率分布を推測する」ということは、古典的には情報科学的な事柄であって、物理現象を意味しないことの方が多いようである。実際、古典的な問題を考えている際には、それは、物理現象ではなくて、概念的な操作に過ぎないと考えても差し支えないように感じられる。本論は、古典的な想念を前提に書かれている。しかしながら、本研究会の中心的な課題である「量子解析におけるミクロマクロ双対性」という問題においては、「データから確率分布を推測する」ということは、おそらく物理現象そのものなのであり、その問題を考えるときに、本論でいうところの自由エネルギー（を非可換に拡張できたとして拡張した概念）が意味があるのかどうか、は、今のところ、著者には、まったくわからない。ただ、著者は、古典的な場合の応用研究を通して、統計学でしばしば考察される「最尤推定量」よりもこの自由エネルギーの方が応用上重要であり、また数学的に自然なものであると感じているということをお伝えしたいと考えている。

## 3 統計的推測

統計的推測において、非常にしばしば表れる問題は次のようなものである。

**問題設定 (1).** ユークリッド空間 $\mathbf{R}^d$ 上の確率分布 $\varphi(w)dw$ とこれに従う確率変数 $W$ を考える。$W$ の実現値 $w$ がひとつ得られたとき、これを固定する。$p(x|w)$ に従う独立なサンプル $X_1, X_2, ..., X_n, ...$ が

得られるものとする。$X_1, X_2, , , .X_n$ の実現値 $x_1, x_2, ..., x_n$ が得られたとき、その実現値を元に、$X_{n+1}$ の分布を

$$r(x_{n+1}; x_1, x_2, ..., x_n)$$

であると予想するものとする（$r$ は $x_{n+1}$ の確率密度関数であり、$x_1, x_2, .., x_n$ をパラメータとして持つものである)。相対エントロピー

$$G(w, x_1, x_2, ..., x_n) = \int dx_{n+1} p(x_{n+1}|w) \log \frac{p(x_{n+1}|w)}{r(x_{n+1}; x_1, x_2, ..., x_n)}$$

の平均値

$$E[G] = \int dw\varphi(w) \int dx_1 \cdots dx_n p(x_1|w)p(x_2|w) \cdots p(x_n|w)G(w, x_1, x_2, ..., x_n)$$

を最小にするためには、$r(x_{n+1}; x_1, x_2, ..., x_n)$ をどのように設計すればよいだろうか。また $E[G]$ の最小値は、どうなるだろうか。

**問題の意味.** この問題設定は情報学的に次のような課題を考えていることに対応している。$p(x|w)$ が真の分布であって、ここから独立に $n$ 個のサンプルが得られたとき、このサンプルを用いて $(n+1)$ 個目を予測する。予測の方法としては、様々なものが考察の対象となるが、確率分布 $r(x_{n+1}; x_1, ..., x_n)$ で表されるようなものであればどんなものでもかまわない。予測の精度を相対エントロピー $G$ で比べることとするとき（$G$ が小さければ小さいほど精度のよい予測であると考えることにした、ということである)、最も予測の精度がよくなるのは確率分布 $r(x_{n+1}; x_1, ..., x_n)$ として、どのようなものをえらんだときであるか。ただし $w$ は、確率分布 $\varphi(w)dw$ に従うものと仮定する。

この問題は相対エントロピーの性質から解答を得ることができる。

$$p(x_{n+1}; x_1, ..., x_n) = \frac{\int dw\varphi(w) \prod_{i=1}^{n+1} p(x_i|w)}{\int dw\varphi(w) \prod_{i=1}^{n} p(x_i|w)} \tag{1}$$

と定義すると、これは $x_{n+1}$ の確率分布であり、

$$\begin{aligned} E[G] &= \int dw\varphi(w) \prod_{i=1}^{n} \Big\{ \int dx_i p(x_i|w) \Big\} \Big[ \int dx_{n+1} p(x_{n+1}; x_1, ..., x_n) \log \frac{p(x_{n+1}; x_1, ..., x_n)}{r(x_{n+1}; x_1, ..., x_n)} \Big] \\ &+ \int dw\varphi(w) \int dx_{n+1} p(x_{n+1}|w) \log p(x_{n+1}|w) \\ &- \int dw\varphi(w) \prod_{i=1}^{n+1} \Big\{ \int dx_i p(x_i|w) \Big\} \log p(x_{n+1}; x_1, ..., x_n) \end{aligned}$$

となる。この式の第 1 項は、$p(x_{n+1}; x_1, ..., x_n)$ と $r(x_{n+1}; x_1, ..., x_n)$ の相対エントロピーの平均であり、この二つの確率分布が等しいときに最小値 0 を取る。第 2 項と第 3 項は、確率分布 $r(x_{n+1}; x_1, ..., x_n)$ に依存しない。従って、

$$r(x_{n+1}; x_1, ..., x_n) = p(x_{n+1}; x_1, ..., x_n)$$

のとき、$E[G]$ は最小値をとることがわかる。また

$$S(w) = -\int p(x|w) \log p(x|w) dx$$

と定義し、

$$F_n(x_1,...,x_n) = -\log \int dw' \varphi(w') \prod_{i=1}^{n} p(x_i|w') \tag{2}$$

と定義して

$$E[\,\cdot\,] = \int dw \varphi(w) \prod_{i=1}^{n+1} \Big\{ \int dx_i p(x_i|w) \Big\} [\,\cdot\,]$$

と書くことにすれば、最小値は

$$E_{min} = E[F_{n+1}(x_1,...,x_{n+1}) - F_n(x_1,...,x_n) - S(w)]$$

となる。これはまた

$$f(x,w,w') = \log \frac{p(x|w)}{p(x|w')}$$

と定義して、$F_n(x_1,...,x_n)$ の正規化

$$F_n^*(x_1,...,x_n) = -\log \int dw' \varphi(w') \exp(-\sum_{i=1}^{n} f(x_i,w,w')) \tag{3}$$

を考えると、

$$E_{min} = E[F_{n+1}^*(x_1,...,x_{n+1}) - F_n^*(x_1,...,x_n)]$$

と書くことができる。

**注意.** 式 (1) で表される確率分布 $p(x_{n+1};x_1,...,x_n)$ をベイズ予測分布という。パラメータ $w$ が確率分布 $\varphi(w)dw$ に従って発生している場合には、ベイズ予測分布が相対エントロピーの意味で最良の予測を与えることがわかった。また、パラメータ $w$ を発生している確率分布に依存せずに最良の予測を与える予測は、一般には存在しないこともわかった。

**注意.** 式 (3) は、本論で述べている自由エネルギーであり、予測精度を算出することは、この自由エネルギーを算出することと同じである。統計力学において自由エネルギーが算出されれば、物理現象について多くのことが解明されるのと同様に、学習理論においても自由エネルギーが解明されれば、学習システムについて多くのことが解明できることになる。式 (3) の挙動の解明は、(数学的な難易度は不明であるものの)、確率論の研究者にも、「数学的な課題である」と感じられるものではないかと思われる。

**注意.** 予測分布 (式 (1)) の構成において、真の分布 $\varphi(w)dw$ とは別の仮に準備した確率分布 $\psi(w)dw$ を用いて計算を行ったとき、

$$F_{\psi,n}^*(x_1,...,x_n) = -\log \int dw' \psi(w') \exp(-\sum_{i=1}^{n} f(x_i,w,w')) \tag{4}$$

と定義すれば、汎化誤差 $E[G]$ について

$$E[G] = E[F_{\psi,n+1}^*(x_1,...,x_{n+1}) - F_{\psi,n}^*(x_1,...,x_n)]$$

が成り立つ (この平均の $E[\,\cdot\,]$ は、真の $\varphi(w)dw$ で算出する)。真の分布は多くの場合において不明であるが、仮の確率分布を用いたとき、どれだけの汎化誤差のずれがあるかについて理論計算を行っておくと、次のような実問題に役立つことが知られている。得られたデータ $(x_1,...,x_n$ を用いて、$\psi(w)dw$ に対して $F_{\psi,n}^*(x_1,...,x_n)$ を数値計算して、理論値と比較することにより、仮に用いている $\psi(w)dw$ の正当性や、二つ以上の $\psi(w)dw$ のうちのどちらがより適切であるかを数値的に比較することができる。これが統計的モデル選択である。

**注意.** 物理学において自然現象を考察するとき、その自然現象に対してハミルトニアンを与えることでモデルを作り、そのモデルに対して経路積分や分配関数を計算して挙動を理論的に予言し、その結果を自然現象と比較する、というプロセスが実行されることになる。学習理論においては、与えられたデータに対して、モデル $p(x|w)$ や $\psi(w)$ を与えることで、ハミルトニアンと、ハミルトニアンを積分する測度とを与えることでモデルを作り、自由エネルギーや汎化誤差を理論的に予言し、その予言と実験値とを比較することで、モデルの正当性を吟味することになる。

## 4 統計的検定

学習あるいは統計的推測において自由エネルギーが重要な役割を果たしていることを上で述べたが、統計的検定においても自由エネルギーが大切であることを次に紹介する。

**問題設定 (2).** データ $x^n = (x_1, x_2, ..., x_n)$ が得られているとする。パラメータの集合 $\mathbf{R}^d$ 上のある確率分布から発生された $w$ が固定され、その $w$ によって定まる確率分布 $p(x|w)$ から独立にデータ $x^n$ が得られたということは仮定する。このとき、パラメータ $w$ を発生した確率分布について二つの仮説を考える。
(1) 帰無仮説 $\varphi_0(w)dw$
(2) 対立仮説 $\varphi_1(w)dw$
与えられたデータから、どちらの仮説を選ぶかは、次のアルゴリズムによって定めるものとする。「ある関数 $S(x^n)$ を計算して、$S(x^n) > a$ ならば帰無仮説を取り、そうでないとき対立仮説を取る」。このアルゴリズムによって、データから仮説を決めるとき、帰無仮説が正しいときに対立仮説を選ぶ確率を危険率という。また、対立仮説が正しいとき、対立仮説を選ぶ確率を検出力という。一般に危険率を小さくすると、検出力も小さくなる。同じ危険率のもとで、検出力を最大にする検定法を最強力検定という。最強力検定を与える関数 $S(x^n)$ は何だろうか。また、そのとき、例えば危険率 0.05 になるような検定アルゴリズムを作るにはどうしたらよいだろうか。

結論から述べると最強力な検定を与える関数は次のものである。

$$L(x^n) = \frac{\int dw \varphi_1(w) \prod_{i=1}^n p(x_i|w)}{\int dw \varphi_0(w) \prod_{i=1}^n p(x_i|w)}$$

この関数を用いて、「危険率 0.05 の検定」を作るためには、帰無仮説のもとで、事象 $L(x^n) > a$ の確率が 0.05 になるような $a$ を定める必要がある。すなわち、$P(L(x^n) > a) = 0.05$ となるような $a$ を決める必要がある。そのためには、確率変数 $L(x^n)$ が従う確率分布を定める必要がある。それは確率変数 $-\log L(x^n)$ が従う確率分布を定めることと同値であるが、

$$-\log L(x^n) = -\log \int dw \varphi_1(w) \prod_{i=1}^n p(x_i|w) + \log \int dw \varphi_0(w) \prod_{i=1}^n p(x_i|w)$$

これは、本論で述べているところの自由エネルギーの差である。つまり、自由エネルギーの確率分布を求めることができれば、そこから最強力検定を作ることができる。

**例.** 例えば、確率モデルとして

$$p(x|a) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-a)^2}{2})$$

を考え、二つの仮説として

$$\begin{aligned}\varphi_0(a) &= \delta(a)\\ \varphi_1(a,\sigma) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{a^2}{2})\end{aligned}$$

とすれば、$L(x^n)$ を用いることにより、帰無仮説 $a=0$ に対する対立仮説を検定することができる。対立仮説が異なれば、最強力検定も異なることになる。

上記で述べた $L(x^n)$ が最強力検定を与えることを示す。これは、ネイマン・ピアソンの補題と呼ばれるものである。帰無仮説が正しいものもとので確率を $P(\cdot|\varphi_0)$ と書き、対立仮説が正しいもとでの確率を $P(\cdot|\varphi_1)$ と書くことにする。上記の関数 $L(x^n)$ を用いた検定と任意の関数 $S(x^n)$ を用いた検定の危険率が等しいと仮定して、検出力の間に不等式が成り立つことを示せばよい。仮定は

$$P(L(x^n)>b|\varphi_0)=P(S(x^n)>a|\varphi_0) \tag{5}$$

である。

$$P^* \equiv P(L(x^n)>b|\varphi_1)-P(S(x^n)>a|\varphi_1)$$

と定義して $P^*\geq 0$ であることを示す。二つの事象 $B, A$ を

$$\begin{aligned}B &= \{x^n; L(x^n)>b\}\\ A &= \{x^n; S(x^n)>a\}\end{aligned}$$

と定義すると

$$\begin{aligned}P^* &= \int \varphi_1(w)dw\Big[\int_B-\int_A\Big]\int\prod_{i=1}^n p(x_i|w)\\ &= \int \varphi_1(w)dw\Big[\int_{B\cap A^c}-\int_{A\cap B^c}\Big]\int\prod_{i=1}^n p(x_i|w)\end{aligned}$$

ここで $L(x^n)$ の定義と、式 (5) を使うと

$$\begin{aligned}P^* &\geq \int \varphi_0(w)dw\Big[\int_{B\cap A^c}-\int_{A\cap B^c}\Big]\int\prod_{i=1}^n p(x_i|w)\\ &\geq \int \varphi_0(w)dw\Big[\int_B-\int_A\Big]\int\prod_{i=1}^n p(x_i|w)\\ &= 0\end{aligned}$$

が得られる。

## 5 自由エネルギーの挙動

以上で、本論で述べる自由エネルギーが学習理論において極めて重要な量であることを述べた。ここでは統計的推測および統計的検定について説明したが、与えられたデータの符号化の問題においても、もしもパラメータの分布が $\varphi(w)dw$ に従い、データが $p(x|w)$ から独立に得られたのであれば、そのときの符号長の最小値が自由エネルギーで与えられることが知られている。

本論で述べた「自由エネルギー」は、統計学においては「対数周辺尤度」、情報理論では「ベイズ符号長」、学習理論では「確率的複雑さ」と呼ばれている。本論で、特に「自由エネルギー」という言葉を用いたのは、自由エネルギーの理論や近似理論においては、自由エネルギーの計算のために物理学において長年に渡って培われてきた方法があり、そのことは意外に、統計学や情報学においては知られていないからである。もしも $n$ が非常に大きいとき

$$P(w) \propto \exp(-\sum_{i=1}^{n} f(X_i, w))$$

がある $w_0$ を平均とする正規分布に近づくと考えてよい場合には、$w_0$ の周りで、ハミルトニアンを2次まで展開することで、自由エネルギーの近似値を求めることができる。この方法は統計学でも知られている方法であり、BIC（ベイズ情報量規準）あるいはMDL(最小記述長）を与えるものである。ハミルトニアンを2次式で近似できるのは、物理学においては、考察している系が漸近的に自由場になる場合であって、最も容易なケースである。すなわち、この近似法は統計的正則モデル（フィッシャー情報行列が正則）という極めて限られたモデルにおいてのみ応用が可能であり、多くの学習モデル、例えば、混合正規分布、神経回路網、隠れマルコフモデル、ベイズネットワーク、縮小ランク回帰などにおいては、自由場による近似は、正しい結果を与えないことが知られている。2次式で近似する方法のほかに $P(w)$ を成分毎に独立なもの

$$p(w_1)p(w_2)\cdots p(w_d)$$

の中で相対エントロピーの意味で最小の誤差になるもの用いる方法がある。これは物理学においては平均場近似と呼ばれ、物理学の計算法が情報学において応用された典型的な例である（情報学において独立に発見されたのではなく、平均場近似という方法が応用された)。1999年に学習理論においても応用されて、変分ベイズ法という名称で知られるようになった。統計物理学においてよく知られているように、平均場近似は、自由エネルギーの解析において、必ずしも正しい結果を与えない。学習理論においては、平均場近似は、自由場による近似（フィッシャー情報行列が正則であるとする近似）よりは良い近似を与えるものの、やはり正しい結果を導かないことが知られている。なお、学習理論においては、自由エネルギーそのものよりも、自由エネルギーのサンプルに対する増分（すなわち汎化誤差）や、二つの事前分布における自由エネルギーの差（すなわち検定量）が重要となるのであるが、それらの問題において平均場近似を用いると、自由エネルギーとしては、ほぼ正しい結果を与えている場合でも、増分や差においては、大きな違いを与えることが多いことが近年になって知られるようになってきた。

数理物理学において良く知られた結果として、自由エネルギーなどの物理学的な量を厳密に算出しようとすると、様々な数学的な概念が大切な役割を果たすことが多い。学習理論においても、ハミルトニアンがパラメータ $w$ に関する可換環としての性質を考察することが重要であることは近年になって明らかになってきた。数理物理学と学習理論とは、考察する対象としては、互いに関係のないものを扱っているように思われあが、そこに現れてくる数理の中に、両者の協力が成り立つ場面がわずかな可能性ながら、ありうるかもしれない。

## 6 結論

数理物理学と学習理論に現れる自由エネルギーの概念を説明して、両者の間にあるかもしれないわずかな関連について説明を行った。数式の形式が少しだけ似ているということが、どれだけ実質的な関係と対応するのか、ということについては現時点では決して明らかでないが、将来、両者の間に、より深い関連が発見され、両方の分野が進展することがあればすばらしいと思われる。

数理物理学とは大きな隔たりのある内容であるにも関らず、発表の機会を与えてくださった数理解析研究所の小嶋泉先生に心から感謝申し上げます。

## 付録

本論では、自由エネルギーと形式的に似ている概念が統計学・情報理論・学習理論において大切な役割を果たすことを述べたが、これらの事柄に関する歴史について簡単に紹介する。

統計学において自由エネルギーに相当する概念の重要性を最初に指摘したのは [1] のようである。分配関数の値を証拠 (evidence) と呼んでいる。有限次元の正規分布に近い場合に鞍点法で近似する方法を最初に考案したのは誰かよくわからない。その結果として得られる値でモデルの選択をすることを提案したのは [2] のようであるが、鞍点法で近似すること自体は、遥か昔から統計学でも行われていたと思われる。この正規分布による近似によって符号長が計算されているということを述べたのが [3] である。自由エネルギーと汎化誤差の関係について最初に言及したのは [4] ではないかと考えられる。この時代まで、自由エネルギーと汎化誤差の間の関係が知られていなかったというのは信じがたいことである。自由エネルギーの計算に平均場近似を導入することは、ほぼ同時期に多くの研究者が提案しているようであるが、混合分布において平均場近似が有用であることを述べた [5] をあげておく。最後に著者は次のような研究をしている。与えられたサンプルから、サンプルを発生している確率分布の構造を抽出するために統計的推測や統計的検定を行うと集合

$$\{w \in \mathbf{R}^d \; ; \; E[f(X,w)] = 0\}$$

が代数多様体や解析的集合となるが、自由エネルギーの挙動は

$$\zeta(z) = \int E[f(X,w)]^z \varphi(w) dw$$

を有理型関数に解析接続したときの極を調べることで解明することができる [6]。

## 参考文献

[1] Good,I.J.(1950) Probability and the Weighing of Evidence. Charles Griffin, London.

[2] Schwarz,G. (1976) Estimating the dimension of a model. Annals of statistics,6(2),461-464.

[3] Rissanen,J.(1984) Universal coding, information, prediction. and estimation," IEEE Trans. on Information Theory, 30,629-636.

[4] Levin,E. Tishby,N. and Solla,A.(1990) A statistical approach to learning and generalization in layered neural networks, Proc. of IEEE, 78(10),1568-1574.

[5] Attias,H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. Proc. of UAI, 21-30.

[6] Watanabe, S. (1999) Algebraic analysis for singular statistical estimation. Lecture Notes in Computer Sciences, Springer, 1720, 39-50.