解説：ミニ特集

学習理論って何？

情報幾何と機械学習

赤穂　昭太郎*

*独立行政法人 産業技術総合研究所 脳神経情報研究部門
茨城県つくば市梅園 1-1-1 中央第 2
*The National Institute of Advanced Industrial Science and Technology, Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki, Japan
*E-mail: s.akaho@aist.go.jp

キーワード：微分幾何（differential geometry），双対性（duality），平坦空間（flat space），射影（projection），確率モデル（probabilistic model），統計的推定（statistical inference）．
JL 0005/05/4405-0299 ©2005 SICE

1. はじめに：なぜ情報幾何なのか

幾何学は視覚に訴える学問である．だから，難しい理論的な話も幾何を用いて視覚的に説明すれば，初心者にも直感的に理解することができる．

では機械学習を幾何的に説明するとどのようになるだろうか．一言で言えば，機械学習とは，データが与えられたとき，そのデータにうまくあてはまるモデルを見つけるという操作である．これは，分野によってシステム同定，統計的推定などと呼ばれるものと基本的に同じである．

この操作を絵で描けば，図1のようになる．候補となるモデルの集合は，なんらかのパラメータで表わされる空間をなしている．一方，データの方は必ずしもモデルに完全にフィットするわけではないのでその外の空間の点であらわそう．すると，データに最もよくあてはまるモデルを見つけるには，データ点からモデルの空間にまっすぐ射影を下ろしてやればよい．モデルの空間が平らならば射影もやさしいだろうし，ぐにゃぐにゃと曲がっていれば射影を下ろすのも大変だろう．

以上が，機械学習の幾何的解釈の大ざっぱな説明である．しかしながら，図に書いた空間に「構造」を入れてやらないと，それ以上深い議論ができない．われわれに最も身近なのはユークリッド空間である．それで済めば話は簡単だが，それではいろいろ不都合が出てくる．たとえば，既存のシステムや統計モデルの推定法は残念ながらユークリッド空間では解釈できない．

そこで登場するのが情報幾何というわけである．情報幾何は確率分布の空間に（非ユークリッド的だが）「自然な」構造を導入する．すると，確率分布に基づくいろいろな分野，たとえば統計学・情報理論・システム理論の問題を統一的に扱うことができ，既存の推定法を説明したり，異なる分野の関係を明らかにしたりできるようになる．そういう意味で，情報幾何は異分野間の共通言語的な役割をもつことができる可能性がある．しかしながら，工学分野の人間にはなじみの薄い微分幾何という数学がベースになっているため，実際にはなかなかしきいが高いというのが現実であろう．そこで本稿では，情報幾何の概要を，数学的厳密性はある程度犠牲にして，できるだけ直感に訴える形で説明していきたい．

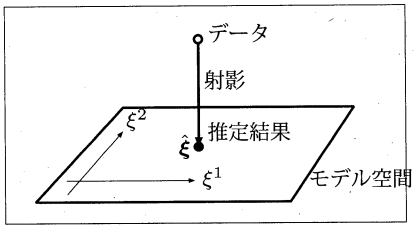

図1　機械学習の幾何的イメージ

2. 情報幾何とは何か

情報幾何は微分幾何に基づいて構築された枠組みだから，ある程度微分幾何の概念に慣れておく必要がある．われわれが慣れ親しんでいるユークリッド空間では，「まっすぐ」「平ら」などの概念はほとんど自明で，特に意識する必要はない．ところが，一般の空間ではこれらをきちんと定めてやる必要がある．

2.1 確率分布の空間

情報幾何の出発点は，n 次元の実数パラメータ $\boldsymbol{\xi} = (\xi^1, \ldots, \xi^n)$ をもつ確率変数 X の確率分布モデル $f(x;\boldsymbol{\xi})$ である(注1)．$\boldsymbol{\xi}$ を座標系と考えると，確率分布モデル全体はこの座標系の張るなめらかな空間（幾何の言葉で言うと**多様体**）とみなすことができ，1つ1つの確率分布はその空間中の1点として表わされる．

例 1 (離散分布) X が離散変数で $\{x_0, x_1, \ldots, x_n\}$ を取るとし，$\mathrm{Prob}(X = x_i) = q_i (> 0)$ とおく．$\sum_{i=0}^{n} q_i = 1$ だから，独立なパラメータの個数は n 個で，たとえば $q_1, \ldots, q_n$ を取れば，n 次元のパラメータ空間となる．

例 2 (正規分布) X を1次元実数とし，その確率密度を $f(x;\mu,\sigma^2) = \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$ とする．これは μ, σ によって規定される2次元空間である．

ちなみに，上の例を考えればわかるように，パラメータは一般に実数空間全体に定義されるわけではなく，その部分集合（$q_i > 0, \sigma > 0$ など）が定義域となっている．

(注1) $f(x;\boldsymbol{\xi})$ は X が離散変数なら確率値関数であり，連続変数なら確率密度関数である．幾何を考える都合上，$f(x;\boldsymbol{\xi})$ は定義域の上で正の値を取ると仮定する．

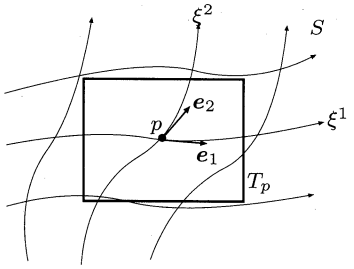

図2　曲がった空間も局所的には線形空間

2.2　点の近く：ユークリッド空間

さて，この空間 S に構造を入れてやろう．その流れを大まかに言うと，まず各点の近傍ではユークリッド空間で近似し，計量という量でその構造を決める．さらにその近傍同士のつなぎかたを接続という量で決めてやることにより，S 全体の構造が決まる．以下ではまず，S のある点 p をまっすぐに動かすという操作を通じてこれらの概念を説明していこう．以下点 $p \in S$ の $\boldsymbol{\xi}$ 座標を $\boldsymbol{\xi}(p)$ と書くことにする．

どんなに曲がった空間でも，p の近くでは，われわれのよく知っているユークリッド空間で近似できる（**図2**）．これを T_p と書こう（原点を点 p におく）．ユークリッド空間ならば，点をまっすぐ動かすことは簡単で，T_p 内の任意の方向に直線的に進めばよい．

しかしこれが通用するのは p の近くだけで，実際には無限小しか進むことはできない．したがって，このユークリッド空間で考えたまっすぐな方向は，運動の軌跡の接線方向（**接ベクトル**という）を定めたにすぎない．T_p はいろいろな向きの接ベクトルの集合だから**接空間**と呼ばれる．

もっと長い距離をまっすぐ進むためには次節で導入する接続の概念を使う必要があるが，ここではもう少し接空間の構造を考えよう．S の座標軸 $\xi^1,\ldots,\xi^n$ のそれぞれの方向に対応する基底を $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ と書けば，T_p の点はその線形和 $\sum_{i=1}^n a_i \boldsymbol{e}_i$ で表わせる(注2)．T_p の構造を決めるには $\boldsymbol{e}_i$ と $\boldsymbol{e}_j$ の間の内積

$$g_{ij}(\boldsymbol{\xi}) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle \qquad (1)$$

を定めてやればよい（角度や長さが計算できる）．$g_{ij}(\boldsymbol{\xi})$ を（**リーマン**）**計量**という．これを ij 成分とする行列を G とおくと，G は正定値対称である必要はあるが，それを満たせば任意に取ってよく，$\boldsymbol{\xi}$ に依存して変化してもよい．

さて，情報幾何では**フィッシャー情報行列**

$$g_{ij}(\boldsymbol{\xi}) = \mathrm{E}_{\boldsymbol{\xi}}\left[(\partial_i l)(\partial_j l)\right] \qquad (2)$$

を計量とする．ただし簡略化のため $\partial_i = \partial/\partial\xi^i$，$l = \log f(x;\boldsymbol{\xi})$ とおいた．また，$E_{\boldsymbol{\xi}}[\ ]$ は，$f(x;\boldsymbol{\xi})$ に関する期待値

$$\mathrm{E}_{\boldsymbol{\xi}}[g(x)] = \int f(x;\boldsymbol{\xi}) g(x) dx \qquad (3)$$

を表わすとする(注3)．

フィッシャー情報行列を選ぶのにはいくつかの必然性があるが，直感的にわかりやすいのは，統計的推定の基本的な不等式である**情報量不等式**（クラメール・ラオ不等式）との関係である．N 個の独立なサンプルからなんらかの推定法によって推定したパラメータを $\hat{\boldsymbol{\xi}}$ とおくと，これはサンプルの出方によってゆらぐ確率変数となる．$\hat{\boldsymbol{\xi}}$ の期待値が真のパラメータ $\boldsymbol{\xi}^*$ に一致するとき，$\hat{\boldsymbol{\xi}}$ の分散は，フィッシャー情報行列を G として，

$$\mathrm{Var}[\hat{\boldsymbol{\xi}}] \geq \frac{1}{N} G^{-1} \qquad (4)$$

を満たす(注4)．これを情報量不等式という．最尤推定量などの「良い」推定量では，漸近的にはこの不等式の等号が成立する．したがって，フィッシャー情報行列は推定量の散らばり具合の逆数になっており，これを距離尺度として取るのは自然なことである．

例 3　正規分布の場合，$(\xi^1,\xi^2) = (\mu,\sigma)$ を座標系に取ると，$\log f(x;\boldsymbol{\xi}) = (x-\mu)^2/(2\sigma^2) - \{\log(2\pi\sigma^2)\}/2$ なので，フィッシャー情報行列は以下のように計算できる．

$$G = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \qquad (5)$$

これを使うと，たとえば μ,σ を $d\mu, d\sigma$ 微小に動かしたときの，変化の大きさは $(d\mu^2 + 2d\sigma^2)/\sigma^2$ となる．σ が小さいときは微小な変動でも分布としての変化が大きく，σ が大きいところでは変化は少ないことを反映している．

S に別の座標系 $\boldsymbol{\theta}$ を取ったとき，$\boldsymbol{\xi}$ から $\boldsymbol{\theta}$ への変換がどれだけ非線形でも，1点 p の近くで考えれば線形変換で近似できる．具体的には p における $\partial\theta^i/\partial\xi^j$ を ij 成分にもつヤコビ行列 B である．だから，T_p の点の表現は基底 $\boldsymbol{e}_i$ と係数 a_i を B で変換してやれば，$\boldsymbol{\xi}$ 座標系から $\boldsymbol{\theta}$ 座標系に容易に変換できる（同様に計量の変数変換も B を使って変換できる）．これは，接空間や計量という概念が座標系の取り方に本質的には不変であることを示している．幾何ではこの「不変性」というのを非常に大事にしている．

2.3　ユークリッド空間をつなぐ

S の点 p は接空間 T_p を考えることにより，接ベクトルの方向に微小距離 $d\boldsymbol{\xi}$ だけはまっすぐ動くことができた．ここではそれをもっと延長していこう．

(注2) 基底の表現法には $\partial/\partial\xi^i$ などいろいろな取り方があり，座標変換などを考える際には便利であるが，本稿では特に必要がないので $\boldsymbol{e}_i$ としたまま扱う．

(注3) x が離散変数を含んでいればその部分は総和にする

(注4) 不等号は左辺から右辺を引いたものが正定値になるという意味である．

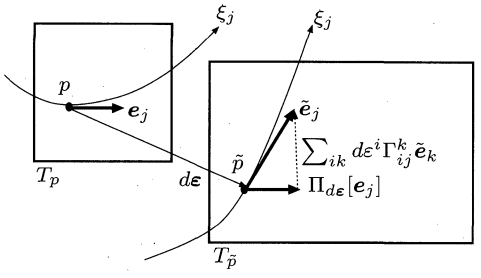

図3　接続は接空間同士のつながり方を決める

新しく動いた点 $\boldsymbol{\xi}(\tilde{p}) = \boldsymbol{\xi}(p) + d\boldsymbol{\xi}$ では，新たな接空間 $T_{\tilde{p}}$ で考える必要がある．その新しい空間で，最初に動いた $d\boldsymbol{\xi}$ と「同じ向き」のベクトル $d\boldsymbol{\xi}'$ を定めてやれば，さらにそこから微小に $d\boldsymbol{\xi}'$ だけ動かしてやることができる．この操作を積み重ねていけば，点をまっすぐ長い距離動かせるようになる．

より一般に，点 p を $d\boldsymbol{\varepsilon} = (d\varepsilon^1, \ldots, d\varepsilon^n)$ だけ微小変化させて点 $\tilde{p}$ に移したとき，T_p のベクトル $d\boldsymbol{\xi}$ が $T_{\tilde{p}}$ に移った先のベクトルを $\Pi_{d\boldsymbol{\varepsilon}}[d\boldsymbol{\xi}]$ と書き，これを**平行移動**という（**図3**）．これは $d\boldsymbol{\varepsilon}$ が微小ならば線形変換であらわすことができる．具体的には，まず T_p の基底 $\boldsymbol{e}_j$ の平行移動を

$$\Pi_{d\boldsymbol{\varepsilon}}[\boldsymbol{e}_j] = \tilde{\boldsymbol{e}}_j - \sum_{i,k} d\varepsilon^i \Gamma^k_{ij} \tilde{\boldsymbol{e}}_k \qquad (6)$$

と書こう（ただし $\tilde{\boldsymbol{e}}_j$ は $T_{\tilde{p}}$ の基底）．この式の Γ^k_{ij} を**接続（係数）**という．直感的には，接ベクトルは移動量に比例して接続係数の分だけ方向を変える．一般の接ベクトル $d\boldsymbol{\xi} = \sum_{j=1}^n a_j \boldsymbol{e}_j$ は，$\sum_{j=1}^n a_j \Pi_{d\boldsymbol{\xi}}[\boldsymbol{e}_j]$ に移ることになる．

点のまっすぐな移動は，$d\boldsymbol{\xi}' = \Pi_{d\boldsymbol{\xi}}[d\boldsymbol{\xi}]$ によって接ベクトルをそれ自身の方向に平行移動させる操作を連続的に繰り返せばよい．こうして得られた軌跡はまっすぐな線を定義するが，これはたまたま取った座標系 $\boldsymbol{\xi}$ で見たときに直線になっているとは限らないので，**測地線**という別の名前がついている．

2.4　α-接続

さて，接続係数はどのように決めたらよいのだろうか．2つの接ベクトル $d\boldsymbol{\xi}_1, d\boldsymbol{\xi}_2$ を平行移動させたとき，通常はその幾何的な関係が変わってほしくない．具体的には，平行移動させる前の内積と，平行移動させた後の内積は同じ値であってほしい．この制約下では，接続係数は計量 g_{ij} に依存して一意に決まってしまう(注5)．これをリーマン接続（またはレビチビタ接続）という(注6)．だから，普通の微分幾何では空間の構造は計量だけから決まってしまう．

ところが，後で述べるように統計的な立場からは，むしろ内積を保存しない接続の方が意味をもつ場合がある．といっても何でもいいわけではなく，ある種の統計的不変性を仮定すると，接続係数はつぎのように自由パラメータ α をもつものに限定される．便宜上接続係数 Γ^k_{ij} を計量 g_{ij} で変換したものを $\Gamma_{ij,k} = \sum_h \Gamma^h_{ij} g_{hk}$ とおくと，

$$\Gamma^{(\alpha)}_{ij,k} = \mathrm{E}_{\boldsymbol{\xi}}\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right] \qquad (7)$$

となる．これを **α-接続**という．$\alpha = 0$ の場合がリーマン接続となるが，情報幾何ではつぎの節で見るようにむしろ $\alpha = \pm 1$ の場合が特に重要である．

2.5　平坦な空間

接続係数は，微小な距離にある接空間の間の「ずれ」を表わしている．もし，ある座標系 $\boldsymbol{\xi}$ を取ったとき，その α-接続の接続係数が全部 0 だったらそのずれも当然 0 である．このような座標系は存在するとは限らないが，もし存在するなら，**α-（アファイン）座標系**といい，その空間は**α-平坦**であるという．

α-平坦な空間では，測地線は α-座標系での直線として表わされる（**α-測地線**）．これは感覚的にはユークリッド空間にかなり近いまっすぐな構造をもつ空間である（計量が場所によって違うのでユークリッド空間とは異なるが）．ほかにも α-平坦な空間はいろいろと便利な性質があり，工学的に有用な多くの応用例では α-平坦な空間の場合を考える．

例 4　指数分布族と呼ばれる

$$f(x;\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^n \theta^i F_i(x) - \psi(\boldsymbol{\theta}) + C(x)\right) \qquad (8)$$

という形の分布族は $\boldsymbol{\theta}$ をアファイン座標系として 1-平坦である．この分布族は統計の情報幾何において中心的役割を果たすもので，1-接続，1-平坦などのことを特に e-接続，e-平坦などと呼ぶ（e:exponential）．なお，正規分布は指数分布族の形をしており，$F_1(x) = x, F_2(x) = x^2$ とおくと，その e-座標系は $\theta^1 = \mu/\sigma^2, \theta^2 = -1/(2\sigma^2)$ となる．

例 5　確率分布 $F_i(x)$ の線形和で定義される**混合分布族**

$$f(x;\boldsymbol{\theta}) = \sum_{i=1}^n \theta^i F_i(x) + (1 - \sum_{i=1}^n \theta^i) F_0(x) \qquad (9)$$

は $\boldsymbol{\theta}$ をアファイン座標系として -1-平坦である．したがって，-1 接続，-1-平坦のことを特に m-接続，m-平坦と呼ぶ（m:mixture）．

例 6　より一般的に $\alpha \neq 1$ をパラメータとして

$$f(x;\boldsymbol{\theta}) \propto (\sum_{i=1}^n \theta^i F_i(x))^{2/(1-\alpha)} \qquad (10)$$

という形の分布族（α-分布族）を考える．これは $\alpha \neq -1$ を除いて一般に α-平坦ではない(注7)．このように，一般に

(注5) ただし対称性 $\Gamma^k_{ij} = \Gamma^k_{ji}$ を仮定する．

(注6) リーマン接続のもとでは，測地線は2点を結ぶ最小距離の曲線になっていることも言える．

(注7) ただし，確率の総和が1という条件を外して拡大した空間では α-平坦になる．拡大した空間については 3.4 も参照．

確率分布で考えている限りは $\alpha = \pm 1$ の場合だけが特別なので，応用上もほとんどが ±1-接続 つまり e-接続か m-接続を扱う．

2.6 双対座標

互いに符号が反対の接続，α-接続と $-\alpha$-接続はいろいろな意味でペアになっている．そのうちでも最も基本的な性質は，ある空間が α-平坦なら，同時に $-\alpha$-平坦でもあるということである（**双対平坦**）．ただし，それぞれアファイン座標系は別のものになる．

双対平坦な空間 S の α-座標系を $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$，$-\alpha$-座標系を $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ で表わすことにしよう(注8)．これらは以下のルジャンドル変換と呼ばれる関係によって相互に変換される．ルジャンドル変換とは，ポテンシャル関数 $\psi(\boldsymbol{\theta}), \varphi(\boldsymbol{\eta})$ が存在し，

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum_{i=1}^{n} \theta^i \eta_i = 0, \tag{11}$$

$$\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\eta}, \quad \frac{\partial \varphi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \boldsymbol{\theta} \tag{12}$$

という関係が成り立つことをいう．ちなみに，$\boldsymbol{\theta}$ 座標に対する計量を g_{ij}，$\boldsymbol{\eta}$ 座標に対する計量を g^{ij} と書くと，

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}, \quad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij}, \tag{13}$$

という関係があるので，g_{ij} および g^{ij} は計量であると同時に，局所的な座標変換のヤコビ行列となっている(注9)．また，接空間 T_p の α 座標での基底 $\boldsymbol{e}_i$ と $-\alpha$ 座標での基底 $\boldsymbol{e}^j$ の間に

$$\langle \boldsymbol{e}_i, \boldsymbol{e}^j \rangle = \delta_i^j \tag{14}$$

という双直交の関係が成立する．最後の関係は，後で出てくる直交射影と深く関係している．直交性を見るには1つの座標系だけで見るよりも双対座標とペアにして見た方がわかりやすい．

例 7 双対平坦という関係から，指数分布族は 1-平坦（e-平坦）であると同時に −1-平坦（m-平坦）でもある．これに対応する m-座標系は $\eta_i = \mathrm{E}_{\boldsymbol{\theta}}[F_i(x)]$ となり，これは十分統計量の空間である（3.1 参照）．したがって観測されたデータから十分統計量を計算すれば，それを e-座標を用いて S の点として扱うことができる．

たとえば，正規分布（例 4）の場合は，$\eta_1 = \mathrm{E}[x] = \mu$，$\eta_2 = \mathrm{E}[x^2] = \mu^2 + \sigma^2$ となり，観測データはそのサンプル平均 $\hat{\mu}$ とサンプル分散 $\hat{\sigma}^2$ を用いて空間の点 $\boldsymbol{\eta} = (\hat{\mu}, \hat{\mu}^2 + \hat{\sigma}^2)$ として表わせる．また，ポテンシャル関数 $\psi(\boldsymbol{\theta})$ は (8) 式の $\psi(\boldsymbol{\theta})$ そのものであり，$\varphi(\boldsymbol{\eta})$ は (11) 式から求まる．

一方，混合分布族は 1-平坦（e-平坦）でもある．これに対応する e-座標系は，指数分布族のように単純な形をしていない．したがって，双対平坦ではあるが混合分布族よりも指数分布族の方が統計的推定との関連がつけやすい．

2.7 部分空間と射影

本稿の一番最初に述べたように，機械学習の幾何的意味というのは観測されたデータをモデルの空間に射影することである．情報幾何では，データとモデルの両方を含む大きな確率分布の空間 S は，双対平坦なもの（指数分布族など）を考え，モデルをその部分空間で，データを経験分布に対応する S の点として位置づける．以下では部分空間の性質と，射影について説明する．

ユークリッド空間でも，平らな部分空間への射影は曲がった部分空間への射影よりもやさしい．情報幾何でも平坦な部分空間は重要な概念である．双対平坦な空間 S があったとき，その α-座標系での平らな部分空間（つまり線形部分空間）M を **α-平坦な部分空間**という(注10)．ここで注意を要するのは，S 自身の平坦性と異なり，α-平坦な部分空間だからといって $-\alpha$-平坦とは限らないことである．

さて，部分空間への射影を考える際に重要な概念がダイバージェンスである．双対平坦な空間の 2 点 p,q の間 の **α-ダイバージェンス**はルジャンドル変換の（11）式に類似した以下の式で定義される．

$$D^{(\alpha)}(p\|q) = \psi(\boldsymbol{\theta}(p)) + \varphi(\boldsymbol{\eta}(q)) - \sum_{i=1}^{n} \theta^i(p)\eta_i(q) \tag{15}$$

これは点の間の隔たりを表わすものであるが，数学的な「距離」ではない．なぜなら対称性や三角不等式が満たされないからである．ではなぜこんなものを考えるかというと，アファイン座標系と相性がいいのと，距離ではないとはいっても距離の重要な性質を多く受け継いでいるというのがその理由である．具体的には $D^{(\alpha)}(p\|q) \geq 0$ であり，等号は $p = q$ のときに限り成り立つ．また，p と q が非常に近いときは距離に一致する．ちなみに，双対となる $-\alpha$-ダイバージェンスは $D^{(-\alpha)}(p\|q) = D^{(\alpha)}(q\|p)$ となる．

特に，指数分布族を考えると，その $\alpha = 1$ での e-ダイバージェンスは 2 つの分布 $f(x)$ と $g(x)$ のカルバックダイバージェンス

$$K(f\|g) = \int f(x)[\log f(x) - \log g(x)]dx \tag{16}$$

に一致し，双対の $\alpha = -1$ での m-ダイバージェンスは $K(g\|f)$ となる．

(注8) 本稿では詳しく説明しないが，上付き添え字と下付き添え字を区別して双対関係を記述すると便利である．詳しくはテンソルに関する文献[18]を参照のこと．また，すでに述べたように，α-測地線は $\boldsymbol{\theta}$ 座標での直線，$-\alpha$-測地線は $\boldsymbol{\eta}$ 座標での直線となる．

(注9) すぐわかるように g_{ij} と g^{ij} は互いに逆行列の関係にある．

(注10) 空間自体の平坦性と区別するために α-自己平行部分空間と呼ぶこともある．

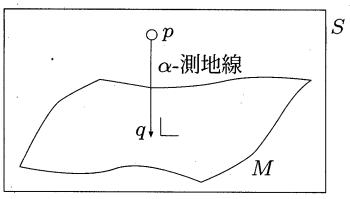

図4　射影はダイバージェンスの停留点

ユークリッド空間での射影が簡単な理由の1つは，ある点から部分空間内の点への距離が直交方向への距離成分と部分空間内の距離成分に分解できることにある（ピタゴラスの定理）．情報幾何の場合も，つぎのように拡張されたピタゴラスの定理が成り立つ．

定理 1 (拡張ピタゴラスの定理) 双対平坦空間 S の点 p,q,r に対し，p と q を α-測地線で結び，q と r を $-\alpha$-測地線で結ぶ．この2つの測地線の q における接ベクトルが直交するとき，以下の関係式が成り立つ：

$$D^{(\alpha)}(p\|r) = D^{(\alpha)}(p\|q) + D^{(\alpha)}(q\|r). \qquad (17)$$

ここで，S の点 p から部分空間 M に引いた α-測地線が点 q で M と直交しているとき **α-射影**とよぶことにする．ピタゴラスの定理から，部分空間への α-射影と α-ダイバージェンスとの関係が導かれる．

定理 2 (射影定理) 双対平坦空間 S の点 p から，部分空間 M への α-射影 q は，α-ダイバージェンス $D^{(\alpha)}(p\|q)$ の停留点である．特に，M が $-\alpha$-平坦な部分空間なら，射影は一意的に存在し，$D^{(\alpha)}(p\|q)$ の最小値をとる．

S は双対平坦だから，ピタゴラスの定理と射影定理は α と $-\alpha$ を入れ替えても成り立つ．

射影定理により，M が $-\alpha$-平坦な部分空間の場合，α-射影を取るのが自然である．その場合，以下のように，M の中と外とで α-座標と $-\alpha$-座標を分けて取る方が，皆まっすぐな世界になるのでわかりやすい．

M が k 次元の $-\alpha$-平坦な部分空間のとき，座標成分を最初の k 個と残りの $n-k$ 個に分けて，$(\boldsymbol{\theta}^{\mathrm{I}}, \boldsymbol{\theta}^{\mathrm{II}})$, $(\boldsymbol{\eta}_{\mathrm{I}}, \boldsymbol{\eta}_{\mathrm{II}})$ とおこう．あらかじめ $\boldsymbol{\eta}$ に適当に線形変換を施しておくことにより，M は $\boldsymbol{\eta}_{\mathrm{II}} = \hat{\boldsymbol{\eta}}_{\mathrm{II}}$（定数）を満たす線形部分空間となるようにできる（**図5**）．ここで新たに，$(\boldsymbol{\theta}^{\mathrm{I}}; \boldsymbol{\eta}_{\mathrm{II}})$ という**混合座標系**という2つの座標系を混ぜたものを考える．S の任意の点はこの混合座標を用いても一意的に表現される．混合座標を用いると，$(\boldsymbol{\theta}^{\mathrm{I}}; \boldsymbol{\eta}_{\mathrm{II}})$ から M への α-射影は単に後半を $\hat{\boldsymbol{\eta}}_{\mathrm{II}}$ でおきかえた $(\boldsymbol{\theta}^{\mathrm{I}}; \hat{\boldsymbol{\eta}}_{\mathrm{II}})$ で求められ，α-射影の具体的な表示が得られる．

3. 機械学習の情報幾何

前章まで見てきたように，情報幾何では双対平坦な空間（特に e-平坦，m-平坦）が幾何的に単純な構造をもつ．そ

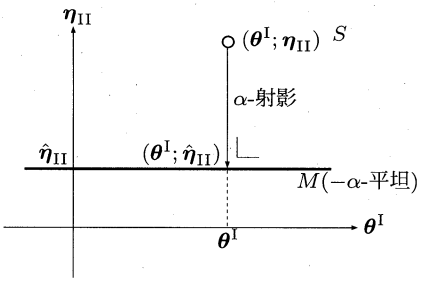

図5　混合座標系で書けばまっすぐに見える

して実際，以下で述べる多くの問題が平坦な空間の性質を生かした学習モデル，学習アルゴリズムを扱っている．

3.1 統計的推定

例7で述べたように，統計的な扱いやすさから，ここでは S として指数分布族を仮定しよう．その際，仮定したモデルを含むような十分広いものを選ぶ必要がある．すると，モデルは S の部分空間 M として表現される．これを**曲指数分布族**という．

一方，指数分布族では情報を落とすことなくデータを十分統計量に集約できる．十分統計量は N 個のサンプル $x_1,\ldots,x_N$ が観測されたとき，$F_i(x)$ のサンプル平均 $r_i = \sum_{j=1}^{N} F_i(x_j)/N$ で計算される．この r_i を η_i 座標成分として，データ点を S の点 $\boldsymbol{\eta} = \boldsymbol{r}$ で表わすことができる．

モデル M が S そのものであれば，座標値そのものが答えなのだから，$\boldsymbol{\eta}$ から $\boldsymbol{\theta}$ に座標に変換すればモデルパラメータが求まる．だが，一般の場合は，$\boldsymbol{\eta} = \boldsymbol{r}$ は M の外の点なので，射影を取らなくてはならない．統計的推定で用いられる最尤推定は，m-射影を取っていることに相当している．m-射影は e-平坦な部分空間に対しては非常に単純になる．

3.2 線形システム

本稿の読者にはシステム制御理論をご専門とされる方も多いであろう．正規ノイズを入力とする最小位相の線形システムは，パワースペクトルで特徴付けられる．対応する確率モデルは，システムのイノベーションの周波数成分がパワースペクトルを分散とする（一般には無限次元の）正規分布となる．実はこのパワースペクトルの空間はすべての α に関して α-平坦となっている[4]．

AR モデルや MA モデルはこのパワースペクトル空間の部分空間として特徴付けられるが，AR モデルは e-平坦，MA モデルは m-平坦な部分空間となっており，推定が単純であるが，ARMA モデルは AR と MA の両方を合わせたような空間になっているため，どちらに関しても平坦ではなく，一般に推定は難しい（**図6**）．

また，フィードバックシステムなどの安定性を議論する際には，行列の固有値が重要な役割を果たす．その中でも正定値行列の空間が基本的で，これは正規分布の分散の空

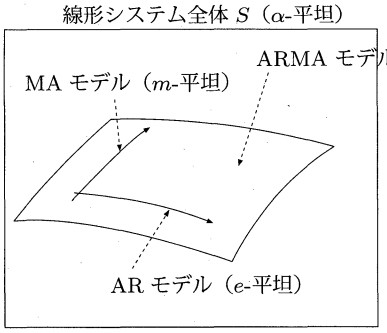

図6　線形システムの空間

間とみなすことができるので，平坦な部分空間として扱うことができる[25),29)]．

3.3　隠れ変数モデル

統計的推定において，確率変数 X のうち一部の成分だけが観測され，残りは観測できない状況を考えよう[1),10),30)]．この場合は，データは十分統計量のうち一部だけしか与えられないので，$\boldsymbol{\eta}$ 座標の1点として表わすことはできない．簡単のため，十分統計量が $\boldsymbol{r}=(\boldsymbol{r}_V,\boldsymbol{r}_H)$ と分けられると仮定し，データが $\boldsymbol{r}_V$ だけを規定するとしよう[(注11)]．各データは $\boldsymbol{\eta}_V=\boldsymbol{r}_V$ で規定され $\boldsymbol{\eta}_H$ は任意の値を取りうる部分空間 Q として表わされる．これは，S が指数分布族なら m-平坦な部分空間である．

データが1点では表わせないので，データの部分空間 Q に最も近いモデルの部分空間 M の点を見つけるということを考えよう．適当な初期値 $p\in M$ から初めて，つぎの2つのステップを繰り返すアルゴリズムが考えられる（**図7**）．

1. $p\in M$ から Q に e-射影を取り $q\in Q$ とする．
2. $q\in Q$ から M に m-射影を取り $p\in Q$ とする．

このアルゴリズムは e-射影と m-射影の頭を取って em-アルゴリズムと名づけられている．ここで都合がいいことに，M から Q へは e-射影で，反対向きの Q から M へは m-射影を取っている．双対接続でのダイバージェンスは $D^{(-\alpha)}(p\|q)=D^{(\alpha)}(q\|p)$ という関係にあるので，いずれの射影も M と Q の関係で見れば同じ評価基準を最小化しているものであることがわかる．もし M が e-平坦で，Q が m-平坦なら，各ステップでの射影は一意的となり，幾何的に単純となる．また，一般に em アルゴリズムは，2つの部分空間の間のダイバージェンスの極小値に収束することがわかっている．

一方，それより以前から知られているアルゴリズムに EM アルゴリズムがある[(注12)]．EM アルゴリズムでは E ステップで対数尤度の条件付き期待値を計算するが，それは em-

[(注11)] 実はこれは十分一般的な仮定で，ほとんどの場合適当な線形変換によりこの形にできる．

[(注12)] 詳しくは本ミニ特集の上田氏の記事を参照．EM は expectation-maximization の頭文字で em は exponential-mixture の頭文字で，偶然同じになっている．

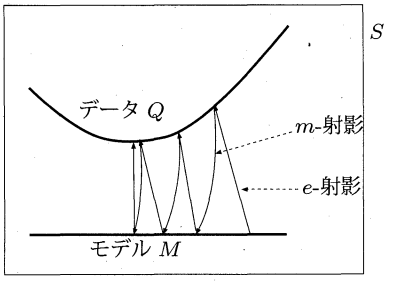

図7　em アルゴリズム（Q が m-平坦，M が e-平坦なら各射影は一意的）

アルゴリズムの第1ステップを

1. $p\in M$ から $q\in Q$ への写像として，$\boldsymbol{\eta}_H(q)=\mathrm{E}_p[\boldsymbol{r}_H\mid \boldsymbol{r}_V]$ を取る[(注13)]．

におきかえることに相当する．多くの場合どちらのアルゴリズムも一致するが複雑な問題設定では異なる場合もある[(注14)]．

3.4　集団学習

三人寄れば文殊の知恵ということわざがあるが，複数の学習モデルを組み合わせることによって高い性能を実現する手法を集団学習あるいはアンサンブル学習という．たとえば，入力 x が -1 か 1 かを識別するような識別器 $h_1(x),\ldots,h_n(x)$ を組み合わせて，$\theta^i\geq 0$ で重み付けた多数決

$$y=\sum_{i=1}^{n}\theta^i h_i(x) \tag{19}$$

の符号を最終的な出力とする．その際できるだけ性能の高い θ^i を求めることが問題となる．集団学習の中でもブースティングと呼ばれるアルゴリズムは非常にうまくいくことがわかっており，その幾何的な解釈も研究されている[14),15),21)~23)]．

ここでは x を入力して y を出力するという入出力型なので，条件付き確率 $f(y\mid x)$ をモデル化する．まず，確率分布を積分すると1になるという制限を外してより広く拡張した空間 $\tilde{S}$ で考える．ブースティングは，$\tilde{S}$ の中でデータ点からモデルの空間 M への射影としてとらえることができる．

モデル $M\subset\tilde{S}$ は次の正規化項のない指数分布型モデル

$$m(y\mid x;\boldsymbol{\theta})=\exp\left(\sum_{i=1}^{n}\theta^i F_i(x,y)+C(x,y)\right) \tag{20}$$

[(注13)] これは点 p のパラメータ $\boldsymbol{\theta}(p)$ で決まる十分統計量の条件付き分布 $f(\boldsymbol{r}_H\mid\boldsymbol{r}_V;\boldsymbol{\theta}(p))$ での期待値

$$\int f(\boldsymbol{r}_H\mid\boldsymbol{r}_V;\boldsymbol{\theta}(p))\boldsymbol{r}_H d\boldsymbol{r}_H \tag{18}$$

を表わす．

[(注14)] S を確率分布全体の空間に取れば一般的に等価性が言える．また，異なる場合もサンプル数が増えれば差が小さくなる．

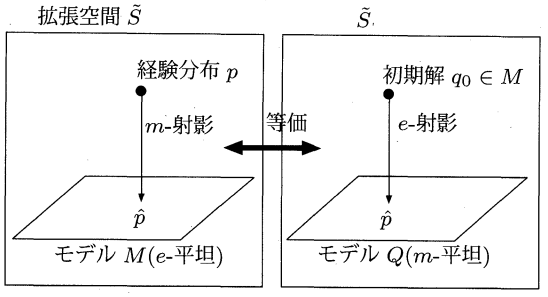

図8　ブースティング．実際には右の最適化問題を逐次的に解く．

を取る．ただし，$F_i(x,y)$ は

$$F_i(x,y) = \frac{1}{2}\{yh_i(x) - \mathrm{E}_{\mathrm{emp}}[yh_i(x) \mid x]\} \qquad (21)$$

とする(注15)．

M は $\tilde{S}$ の中の e-平坦な部分空間なので，m-射影が一意に求まる．ただし，それを直接求めることは難しいので，まずそれを等価な問題におきかえる．

具体的には，データ集合 $\{(x_j,y_j)\}_{j=1}^N$ が与えられたとき，以下の条件を満たす $m(y \mid x)$ の集合 $Q \subset \tilde{S}$ を考える．

$$\sum_{j=1}^{N} m(y_j \mid x_j)F_i(x_j) = 0, \quad \forall i = 1,\ldots,n. \qquad (22)$$

これは m に関する線形制約で，$\tilde{S}$ の中の m-平坦な部分空間になっている(注16)．先に述べたデータ点から M への m-射影は，$q_0(y \mid x) = \exp(C(x,y)) \in M$ という関数から Q への e-射影に一致する（**図8**）ことが証明できる．ブースティングアルゴリズムは，$q_0(y \mid x)$ を初期解として，$\theta^1,\ldots,\theta^n$ を逐次的に求めていくことにより，最終的にこの射影を求めていると解釈できる．

3.5　平均場近似・変分ベイズ法

確率変数の間の関連性をグラフの形で記述したモデルをグラフィカルモデルといい，その汎用性からさまざまな分野で広がりつつある．その構造の入れ方によってベイジアンネットワーク，ランダムマルコフ場モデルなどと呼ばれることがある．また，カルマンフィルタや隠れマルコフモデルなどもその一種とみなすことができる．

さて，グラフィカルモデルでは，局所的な関係が全体に影響を及ぼすため，ある確率変数に関する期待値を取るだけでも，確率変数全体に対する和を計算しなければならず指数的に大きな計算量が必要となることがある(注17)．

(注15) $\mathrm{E}_{\mathrm{emp}}[\,|\,x]$ は観測データに基づく経験分布での条件付き期待値を表わす．M 自身が観測データに依存したものになっているので，通常の統計的推定とはこの意味でも若干異なることに注意．

(注16) 厳密な説明は省くが，直感的には，確率分布全体の空間の中では，指数分布族のように確率分布の log の線形空間が e-平坦で，混合分布族のように確率分布そのものの線形空間が m-平坦な部分空間となる．3.5 でも同様の議論を使う．

(注17) 詳細は省略するが，無向グラフで表わしたときに，グラフ内にループがあるような場合に多くの計算量が必要となる．

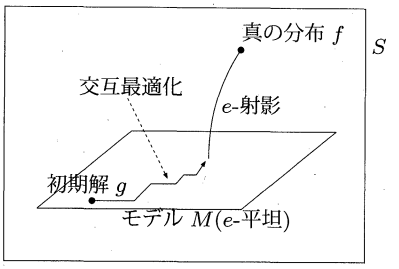

図9　ナイーブ平均場近似．変分ベイズ法では交互最適化によって局所最適解に収束させる．

そこで用いられるのが，平均場近似（あるいは変分ベイズ法）と呼ばれる近似法である[20]．ここではその中でも，最も単純なナイーブ平均場近似についてその幾何的な意味を説明する．

一般に，$f(x_1,\ldots,x_m)$ という確率分布が与えられたとき，各確率変数が独立ならば，変数ごとの計算にばらすことができるので都合がよい．そこで，独立な確率分布全体の空間 M を取り，もとの分布 f を M に射影する．

M の要素 $g(x_1,\ldots,x_m)$ はその周辺確率分布の積

$$g(x_1,\ldots,x_m) = g(x_1)\cdots g(x_m) \qquad (23)$$

で書ける．これは e-平坦な部分空間である．情報幾何の観点からは e-平坦な部分空間へは m-射影を取るのが自然であるが，m-射影を取るために必要なカルバックダイバージェンスはもとの分布 f に関する平均操作を必要とするため計算が容易でない．一方 e-射影は M の分布での平均操作なので，変数ごとにばらばらに行えばよく非常に都合がよい．

そこで，e-平坦な部分空間と m-射影という美しい組み合わせはあきらめて，e-射影を取るというのがナイーブ平均場近似の考え方である．e-射影なので，射影の一意性などは保証されないが，少ない計算量で最適化ができる．変分ベイズ法ではある初期解からスタートし，1 ステップで 1 つの変数だけに着目して射影する（交互最適化）ことによって局所最適解に収束させることが多い（**図9**）．

グラフィカルモデルを用いた現実的な問題（特に最近は符号化への応用が盛んである）では，ナイーブ平均場近似では近似が荒すぎるので，より複雑な近似手法が開発され，それらに関しても幾何的な理解が進みつつある[16),17),19)] (注18)．

4.　おわりに

本稿では確率的な学習モデルを幾何的に眺める方法について，特に平坦な空間への射影という観点から大まかに説明した．本稿で扱えなかった問題として，グラフィカルモデルにおけるマルコフ連鎖モンテカルロ（MCMC）法の幾何

(注18) 基本的に類似な手法だが，クラスタ変分法，TAP 平均場近似，ルーピービリーフプロパゲーション，CCCP 法などといったようにいろいろなバリエーションがある．

的解釈[27]や，確率分布のパラメータの次元縮小[2),13)]などがあり，やはり平坦な構造に着目している．一方，情報幾何は平坦でない場合についてもさまざまな研究がある．接続係数から計算される曲率や捩率と呼ばれる幾何的な量が学習モデルの性能解析や性能向上に重要な役割を果たす．

紙面の制約と筆者の力不足から，必ずしもやさしい解説になったかどうか自信がないが，少しでも情報幾何に興味をもっていただける方が増えれば幸いである．また最後に挙げた面白いトピックについても触れることができなかったが，多くの参考文献を挙げておいたので詳しくはそちらを参考にしていただきたい．

(2005 年 1 月 31 日受付)

参 考 文 献

1) 赤穂昭太郎：EM アルゴリズムの幾何学，情報処理，**37**–1, 43/51 (1996)
2) S. Akaho: The e-PCA and m-PCA: dimension reduction by information geometry, Proc. of Int. Joint Conf. on Neural Networks (IJCNN) (2004)
3) S. Amari: Differential Geometrical Methods in Statistics, Springer Lecture Notes in Statistics, **28** (1985)
4) S. Amari: Differential geometry of a parametric family of invertible linear-systems—Riemannian metric, dual affine connections and divergence, Mathematical Systems Theory, **20**, 53/82 (1987)
5) 甘利俊一，ほか：特集 情報幾何，数理科学，No.303 (1988)
6) 甘利俊一：情報幾何への招待，特集 どこへでも顔を出す微分幾何，数理科学，No.318, 25/29 (1989)
7) 甘利俊一：情報幾何学，応用数理，**2**–1, 37/56 (1992)
8) 甘利俊一，長岡浩司：情報幾何の方法，岩波講座 応用数学 6 [対象 12], 岩波書店 (1993)
9) 甘利俊一，ほか：特集 情報空間 その応用の広がり，数理科学，No.366 (1993)
10) S. Amari: Information Geometry of the EM and em Algorithms for Neural Networks, Neural Networks, **8**–9, 1379/1408 (1995)
11) 甘利俊一：統計学と情報幾何，特集 知としての統計学，数理科学，No.389, 69/75 (1995)
12) O. Barndorff-Nielsen : Parametric Statistical Models and Likelihood, Lecture Notes in Statistics, **50** (1988)
13) M. Collins, S. Dasgupta and R.E. Schapire: A Generalization of Principal Component Analysis to the Exponential Family, Advances in Neural Information Processing Systems, **14** (2002)
14) 江口真透：統計的パタン識別の情報幾何—*U* ブースト学習アルゴリズム，特集 統計科学の最前線，数理科学，No.489, 53/59 (2004)
15) 江口真透：情報幾何と統計的パタン認識，数学，**55**, 岩波書店 (2004)
16) 池田思朗，田中利幸，甘利俊一：ターボ復号の情報幾何，電子情報通信学会論文誌，**J85-D-II**–5, 758/765 (2002)
17) S. Ikeda, T. Tanaka and S. Amari: Information geometry of turbo and low-density parity-check codes, IEEE Trans. on Information Theory, **50**–6, 1097/1114 (2004)
18) 伊理正夫，韓太舜：ベクトルとテンソル第 II 部 テンソル解析入門，シリーズ新しい応用の数学，1-II, 教育出版 (1973)
19) S. Ikeda, T. Tanaka and S. Amari: Stochastic reasoning, free energy, and information geometry, Neural Computation, **16**–9, 1779/1810 (2004)
20) 樺島祥介，上田修功：平均場近似・EM 法・変分ベイズ法，汪，田栗，手塚，樺島，上田：計算統計 I，統計科学のフロンティア 11, 岩波書店 (2003)
21) G. Lebanon and J. Lafferty: Boosting and maximum likelihood for exponential models, Technical Report CMU-CS-01-144, School of Computer Science, Carnegie Mellon University (2001)
22) 村田 昇：推定量を組み合わせる，バギングとブースティング，麻生，津田，村田：パターン認識と学習の統計学，統計科学のフロンティア 6, 岩波書店 (2003)
23) N. Murata, S. Eguchi, T. Takenouchi and T. Kanamori: Information Geometry of U-Boost and Bregman Divergence, Neural Computation, **16**–7, 1437/1481 (2004)
24) M. K. Murray and J. W. Rice: Differential Geometry and Statistics, Monographs on Statistics and Applied Probability, **48**, Chapman & Hall (1993)
25) 小原敦美：線形状態フィードバックシステムの幾何学的構造，計測と制御，**32**–6, 486/494 (1993)
26) M. Opper and D. Saad (eds.): Advanced Mean Field Methods, Theory and Practice, MIT Press (2001)
27) K. Takabatake: Information Geometry of Gibbs Sampler, Proc. of WSEAS Int. Conf. on Neural Networks and Applications (NNA) (2004)
28) 竹内啓，広津千尋，公文雅之，甘利俊一：統計学の基礎 II, 統計科学のフロンティア 2, 岩波書店 (2004)
29) K. Tsuda, S. Akaho and K. Asai: The em Algorithm for Kernel Matrix Completion with Auxiliary Data, J. of Machine Learning Research, **4**, 67/81 (2003)
30) 渡辺美智子，山口和範（編）：EM アルゴリズムと不完全データの諸問題，多賀出版 (2000)

[著 者 紹 介]

赤穂 昭太郎（あかほ しょうたろう）君

1988 年東京大学工学部計数工学科卒業．90 年東京大学大学院工学系研究科修士課程修了．同年，電子技術総合研究所に入所．2001 年より産業技術総合研究所 脳神経情報研究部門 情報数理研究グループ．博士（工学）．統計的学習理論に関する研究に従事．日本神経回路学会，電子情報通信学会などの会員．