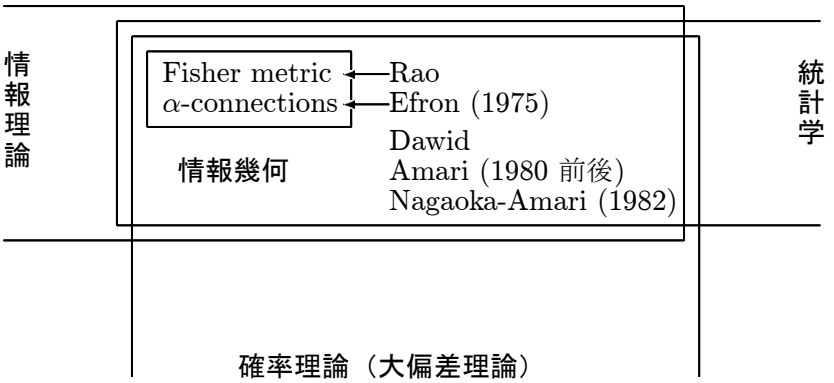
情報幾何の基礎概念

長岡 浩司（電通大）
ノート：野田知宣（OCAMI）

§ 0.

先ず情報幾何と今回の講義の概略を述べる。情報幾何という言葉は厳密な定義がある訳ではなく、人によって狭く捉えらえたり広く捉えらえたり、あるいは捉える場所も異なる。しかしながら確率分布、あるいは確率構造の一つ一つを点とするような空間を考え、その上に微分幾何的構造をのせて解析することは共通している。このような観点に立っても入る構造には色々ある。その中で今回は一番基本的且つ重要と思われる Fisher 計量（と云われる Riemann 計量）と α-接続（と云われる affine 接続）、これらは確率分布を要素とする多様体上にのる、の話をしたい。このような話が歴史的にどのように出て来たかと云えば、そもそもは統計学からであり、統計学の中で Fisher 情報行列（Fisher 情報量）がおそらく２０世紀前半に Fisher によって考案され、統計学的な推定理論において基本的であることが解った。一方少し統計学から離れて考えてみると Fisher 情報量は幾何学で云う Riemann 計量であることが判った。文献上で最初に登場するのは Rao（統計学の巨匠）の 1945 年の論文であり、この中で『 Fisher 情報量を Riemann 計量として考察する事は重要ではないか』との suggestion が与えられた。この辺りから Fisher 計量の幾何的考察が始まった。

一方計量的でない接続の考察は Efron（統計学）の 1975 年の論文に端を発する。彼は統計的推定理論の漸近理論（データ数が非常に大きいときにどのような良い推定が可能か、どのような限界があるかなどを調べる分野）において確率分布族が平坦であるか曲がっているかという事が重要である事を述べた。ところが彼の導入した曲がり具合を測る尺度は普通の意味での（Riemann 幾何的な）曲率ではなかった。このとき既に Fisher 計量は知られていたので、これに対する埋め込み曲率とも思われたがそうではなかった。Dawid は Efron の論文に対する discussion という part で『これは何か新しい接続を導入しているに違いない』と指摘した。これにより非計量的接続の導入に意味があることが分かってきた。これらをきちんと定義し組織的に一般理論を展開、そしてその有効性を示したのが Amari で 1980 年前後の事である。更に非計量的接続などをめぐる世界の統一を目的に研究がなされ duality が得られた（Nagaoka-Amari 1982）。α-接続の α は任意の実数を取り得る接続の集まりであるが、α-接続と $(-\alpha)$-接続とは非常に dual な関係にある。また α は或る意味で確率分布を何乗かしたところ、そのままでは積分して 1 であるが、何乗かして積分すると 1 でなくなるところ（精確には $\frac{1-\alpha}{2}$ 乗）に変換するとそこでの自然な接続を考えている事になる。これらの応用も見付かり、またこれにより統計学以外にも情報幾何に似たものが作れる事もわかった。



このように情報幾何の大部分は統計学の中に出来ている。一方、統計学と密接に関係する分野として情報理論がある。これは 1948 年に Shannon により提唱され、統計学とは別の問題意識を有している。Fisher 計量、α-接続はこれらとも密接に係る。またより proper な確率論とくに大偏差理論の話もこれら Fisher 計量などに関係している。これらの理論は互いに密接に関係しているが、確率論・大偏差理論は統計物理との関係が特に重要で、これにより物理と関係してくる。これら多くの分野の関係する部分に 1 つの幾何的な世界がある事を指摘する事は情報幾何において重要と思われる。

情報幾何の応用は大きく二つある。一つはパラメータ推定論、もう一つは相対エントロピーに係る話である。これらは、Čencov の定理によって幾何的構造は情報幾何構造しかないにも係らず、異なる世界のように見える。量子版を考えた場合これらは別の幾何構造になるから概念的に一致する必然性はないのであろう。また情報幾何が統計や確率論で有用となる理由は少なくとも二つの要因がある。一つは大偏差との関係であり、もう一つは推定理論の幾何学である。多くの場合これらは余り区別されないが、これら二つを紹介する。最後に無限次元の場合を見る。

今回の講義は数学の研究者、若しくは勉強している人で微分幾何についてはある程度知っている人たちを対象とする。

§ 1. 統計多様体と指数型分布族

統計多様体とは、ここでは確率分布（確率密度函数、事象が離散的な集合の場合は確率函数）を要素とするような多様体のことをいう。微分幾何ではより抽象的な或る構造を持った多様体を統計多様体というが[1]、ここでは確率分布を要素とするような具体的なもの（以下 [1] の例 1， 2 参照）を考える。

[1] ∇g が対称となるような affine 接続 ∇、（擬）Riemann 計量 g を備えた可微分多様体 (M, ∇, g) を**統計多様体**と云う。

[1] 測度空間 $(\Omega, \mathcal{F}, \mu)$ に対し

$$\mathcal{P} = \mathcal{P}(\Omega) = \mathcal{P}(\Omega, \mathcal{F}, \mu) := \{p \mid p : \Omega \to \underbrace{\mathbb{R}^+}_{\| \\ (0,\infty)}, \ \int_\Omega p d\mu = 1\}$$

とおく。いま

$$M = \{p_\theta \mid \theta = (\theta^1, \cdots, \theta^n) \in \Theta\} \subset \mathcal{P}, \quad \Theta : \text{open} \subset \mathbb{R}^n$$

が与えられていて

$$\theta \mapsto p_\theta$$

が 1 対 1 かつ十分に滑らかだとする。このとき M は $\theta = [\theta^i]$ を座標系とする多様体と見做すことができる。このような M を**統計多様体**（statistical manifold）と呼ぶ。

これは厳密な意味で数学的定義ではない（“十分滑らか”など）。しかしこれから挙げる例を念頭においておけば以下の話には充分である。また “多様体” といったが、これは一つの座標系で全体が覆われているので多様体の大域的性質には（あまり）関心がないと思って頂きたい。基本的には局所理論である。

例 1. $\Omega = \mathbb{R}$、μ: Lebesgue、

$$p_\theta(\omega) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\omega-\mu)^2}{2\sigma^2}} \quad : \text{正規分布（Gaussian (Normal)-distributions)},$$

に対し $\theta = (\mu, \sigma^2)$ とおけば 2 次元の多様体と見做せる。これは統計多様体の代表例である。

例 2. $\Omega = \{0, 1, 2, \ldots, n\}$（任意の有限集合）

$$\begin{aligned} \mathcal{P} = \mathcal{P}(\Omega) &= \{p \mid p : \Omega \to \mathbb{R}^+, \ \sum_\omega p(\omega) = 1\} \\ &= \{p_\theta \mid \theta = (\theta^1, \ldots, \theta^n) \in \Theta\}, \end{aligned}$$

ここで

$$\begin{aligned} &\theta^i = p(i), \ i \in \{1, 2, \ldots, n\}, \\ &\Theta = \{(\theta^i) \in \mathbb{R}^n \mid {}^\forall i, \ \theta^i > 0 \text{ かつ } \sum_{i=1}^n \theta^i < 1\}. \end{aligned}$$

即ち $\mathcal{P}$ は統計多様体（$|\Omega| = n + 1$ であるが、$\sum p(\omega) = 1$ から自由度は n. 即ち座標系は n 個指定すればよい。ここでは $\omega = 1, \ldots, n$ を入れた値を座標にしている。$\sum p(\omega) = 1$ から $p(0)$ は自動的に定まり、$\mathbb{R}^n$ の開集合となる）。

注意１. $|\Omega| < \infty$ の場合、全体集合 $\mathcal{P}$ が多様体なので統計多様体は全体の部分多様体、即ち
M $(\subset \mathcal{P})$ は統計多様体 $\Leftrightarrow$ M は $\mathcal{P}$ の部分多様体。

注意２. $|\Omega| = \infty$（可算、非可算ともに）の場合も実は $\mathcal{P}$ を無限次元 Banach 多様体とみなすことができる（Pistone-Sempi, 1995)。これについては § 8 参照。

[2] いま述べた２つの統計多様体の例には或る特別な構造が入る。それを述べよう。
統計多様体 $M = \{p_\theta\} \subset \mathcal{P}(\Omega)$ に対し

M は **指数型分布族**（exponential family）
$\overset{\text{def}}{\Longleftrightarrow} {}^{\exists}C : \Omega \to \mathbb{R}$, ${}^{\exists}F_i : \Omega \to \mathbb{R}$ $(i \in \{1, \dots, n\})$,
${}^{\exists}\psi : \Theta \to \mathbb{R}$,
${}^{\forall}\omega$, ${}^{\forall}\theta$, $p_\theta(\omega) = \exp\left[C(\omega) + \sum_{i=1}^{n} \theta^i F_i(\omega) - \psi(\theta)\right]$.

注意. $\psi(\theta) = \log \int \exp[C(\omega) + \sum_i \theta^i F_i] d\mu(\omega)$、すなわち ψ は $\int p_\theta(\omega) = 1$ となる為のもの。

例１.

$$
\begin{aligned}
p_\theta(\omega) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\omega-\mu)^2}{2\sigma^2}} \\
&= \exp\left[-\frac{(\omega-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right] \\
&= \exp\left[\underbrace{\left(-\frac{1}{2\sigma^2}\right)}_{\| \ \theta^1}\underbrace{\omega^2}_{\| \ F_1(\omega)} + \underbrace{\left(\frac{\mu}{\sigma^2}\right)}_{\| \ \theta^2}\underbrace{\omega}_{\| \ F_2(\omega)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi}\sigma\right)}_{\| \ \psi(\theta)}\right]
\end{aligned}
$$

であり指数型分布族（$C(\omega) = 0$）。

例２. $\Omega = \{0, 1, \dots, n\}$ の場合、$\mathcal{P}(\Omega) \ni p$ に対し

$$
\begin{aligned}
\log p(\omega) &= \sum_{i=1}^{n} \log p(i)\delta_i(\omega) + \log p(0)\underbrace{\delta_0(\omega)}_{\| \ 1 - \sum_{i=1}^{n}\delta_i(\omega)} \\
&= \sum_{i=1}^{n} \underbrace{\log\frac{p(i)}{p(0)}}_{\| \ \theta^i}\underbrace{\delta_i(\omega)}_{\| \ F_i(\omega)} - \underbrace{(-\log p(0))}_{\| \ \psi(\theta) = \log(1 + \sum_{i=1}^{n} e^{\theta^i})}
\end{aligned}
$$

とすると指数型分布族であることがわかる。ここで δ_j は Kronecker's delta、即ち

$$\delta_j(\omega) = \begin{cases} 1 & \text{if } j = \omega, \\ 0 & \text{otherwise.} \end{cases}$$

正の確率分布全体は重要な集合であるが、それは指数型分布族を成している。$|\Omega| < \infty$ の場合には全ての統計多様体は或る大きな指数型分布族に含まれている、即ち部分多様体と見做せる（実は無限次元の場合にもそのような見方が出来る）。情報幾何では指数型分布族は特に重要な意味を持つ。

注意. $[\theta^i]$ を指数型分布族 $M = \{p_\theta\}$ の**自然座標系**（natural coordinate system）と呼ぶ（これには affine 変換の自由度がある）。

§ 2. Fisher 計量

[1] 統計多様体 $M = \{p_\theta \mid \theta = [\theta^i] \in \Theta\}$ に対し

$$g_{ij}(\theta) := E_\theta[\partial_i \ell_\theta \partial_j \ell_\theta],$$

これらを成分にもつ行列を $G(\theta) := (g_{ij}(\theta)) \in \mathbb{R}^{n\times n}$ とおく。$G(\theta)$ を M の（座標系 $[\theta^i]$ に関する）（点 p_θ における）**Fisher 情報行列**（Fisher information matrix）と呼ぶ。但し

$$\begin{cases} E_\theta[F] = \displaystyle\int F(\omega) p_\theta(\omega) d\mu : F \text{ の期待値}, \\ \partial_i = \dfrac{\partial}{\partial \theta^i}, \\ \ell_\theta = \log p_\theta. \end{cases}$$

定義から $G(\theta)$ の性質として次が判る：

- $G(\theta) \geqq 0$（半正定値）、
- g_{ij} は 2 階共変テンソル（g とおく）の成分。

いま $G(\theta) > 0$ を仮定する[2]。これにより g は Riemann 計量と見做せる。これを**Fisher 計量**（Fisher metric）と呼ぶ。この計量は或る不変性で特徴付けられる（§ 2 [4] 参照）。確率分布が要素である事を考慮に入れて考えると自然な計量はこれしかない（と云ってよいほど唯一無二）。log をとって微分する有難味が後々解ってくるであろう。

例 1. 正規分布の場合、

$$G(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

但し $\theta = (\mu, \sigma^2)$.

[2] 多くの場合満たされる。例えば $|\Omega| < \infty$ で M が $\mathcal{P}$ の部分多様体の場合など。

例２. $\mathcal{P}(\{0,1,\dots,n\})$ の場合、

$$g_{ij}(\theta)=\frac{\delta_{ij}}{\theta^i}+\frac{1}{1-\sum_i\theta^i},$$

但し $\theta^i=p(i)$ $(1\leqq i\leqq n)$.

[2] $g_{ij}=-E_\theta[\partial_i\partial_j\ell_\theta]$（これは重要な性質。これを定義とする事も可能)。

$\because)\ ^\forall\theta,\ 1=\int p_\theta(\omega)d\mu$ より

$$0=\partial_i\int p_\theta(\omega)d\mu=\int\partial_i p_\theta d\mu=\int(\partial_i\ell_\theta)p_\theta d\mu=E_\theta[\partial_i\ell_\theta].$$

これを微分して $\partial_j p_\theta=(\partial_j\ell_\theta)p_\theta$ から

$$0=\int(\partial_i\partial_j\ell_\theta)p_\theta d\mu+\int(\partial_i\ell_\theta)\partial_j p_\theta\ d\mu=E_\theta[\partial_i\partial_j\ell_\theta]+g_{ij}(\theta).$$

[3]

$$\begin{aligned}g_{ij}(\theta)=&4\int\partial_i\sqrt{p_\theta}\partial_j\sqrt{p_\theta}d\mu\\=&4\sum_\omega\partial_i\sqrt{p_\theta(\omega)}\partial_j\sqrt{p_\theta(\omega)}\quad(\text{if }|\Omega|<\infty)\end{aligned}$$

ユークリッド空間 $\mathbb{R}^\Omega$ 内の
半径 2 の球面の計量
$(\because\ \sum_\omega(2\sqrt{p_\theta(\omega)})^2=2^2)$.

即ち $2\sqrt{p_\theta}$ を座標とする点を $\mathbb{R}^\Omega$ にとっていくと半径 2 の球面になる。この変換により確率分布の集合が球面の形で、Fisher 計量はこの球面に自然に誘導される計量となっている事が判る。

[4] Fisher 計量の不変性
∘ データの（１対１）変換 $\Phi:\Omega\to\Omega'$ によって $M=\{p_\theta\}\subset\mathcal{P}(\Omega)$ が $M'=\{p'_\theta\}\subset\mathcal{P}(\Omega')$ に写されたとする。Φ が１対１ならば

$$G(\theta)=G'(\theta).$$

これは定義に従って確かめれば良い。離散の場合は ω の順番が換わるだけである。連続の場合には変換行列（に相当するもの）の Jacobian が出てくる。この Jacobian 込みで計算を行う。データ変換は θ に依らない変換なので $\log p+$ (θ に依らない項) となり、微分すると第２項は消える（log をとって微分した有難味がここに一つ)。
∘ dominating measure μ の変換：密度函数を μ から ν に替えると、$\frac{d\mu}{d\nu}$ が掛かるが、

これは θ に依らないから log をとって微分すると上と同様に消える。
∘ 十分統計量に関する不変性については § 3 [3] 参照。

データを変えても基本的に統計的状況が変わらないなら Fisher 計量は保存される。

§ 3. α-接続

[1] ここで出てくる接続は affine 接続に限るが、affine 接続を初等的に、また丁寧に書いてある本は案外少ない。また数学的には難しくないが、標準的 Riemann 幾何の教科書には載っていない事実も使うので、その辺りを先ず整理しておく。

(1) affine 接続
$$\updownarrow$$
共変微分 ∇（$:\mathfrak{X}\times\mathfrak{X}\to\mathfrak{X}:(X,Y)\mapsto\nabla_X Y$）
$$\updownarrow [\theta^i] : \text{given}$$
接続係数 $\{\Gamma_{ij}^k\}$（i.e. $\nabla_{\partial_i}\partial_j=\sum_k\Gamma_{ij}^k\partial_k$ により定まる n^3 個の（局所）函数）
$$\updownarrow \text{計量 } g : \text{given}$$
$\{\Gamma_{ij,k}\}$（$\Gamma_{ij,k}=\sum_h\Gamma_{ij}^h g_{hk}=g(\nabla_{\partial_i}\partial_j,\partial_k)$ ）
これらの内、どれを指定しても良い。

(2) 座標系 $[\theta^i]$ が ∇ に関し **affine 座標系**（affine coordinate system w.r.t. ∇）
$\iff {}^\forall i,j,k,\ \Gamma_{ij}^k=0 \quad (\Leftrightarrow {}^\forall i,j,k,\ \Gamma_{ij,k}=0)$
$\iff {}^\forall i,\ \nabla\partial_i=0 \quad (\partial_i \text{ は } \nabla\text{-平行})$

(3) ∇ は**平坦**（flat）
$\overset{\text{def}}{\iff} {}^\exists[\theta^i] : \nabla\text{-affine}$[3]
$$\iff \begin{cases}\text{torsion}=0\\ \text{curvature}=0\end{cases}$$

(4) M の affine 接続 ∇、M の部分多様体 N に対し、一般には

$$^\forall X,Y\in\mathfrak{X}(N),\quad \nabla_X Y\in\mathfrak{X}(N) \tag{3.1}$$

とはならない。(3.1) が成り立つとき、N は ∇ に関して M の中で**自己平行**（autoparallel、a.p.）であると言う。このとき $\nabla|_N$ は N 上の affine 接続となる。

---

[3] これは affine 接続特有の定義。affine 接続に限っても flat には 2 種類ある：平行移動が曲線に依らない事のみを要請するか、affine coord. sys. の存在まで要請するか。後者の方がより強い性質。これは affine 接続と linear 接続を区別する一つのポイント。ここでは後者を採用する。

また (3.1) が成り立たなくても、M に Riemann 計量 g が与えられているときは g に関する射影 π を用いて

$$\nabla'_X Y = \pi(\nabla_X Y)$$

で N 上の affine 接続 ∇' が定義できる（より一般に N への射影で充分)。

[2] 統計多様体 $M = \{p_\theta \mid \theta \in \Theta\} \subset \mathcal{P}(\Omega)$ に対し affine 接続 $\nabla^{(\alpha)}$（$\alpha \in \mathbb{R}$）を次のように定める：

$$\begin{aligned} g(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k) &= \Gamma^{(\alpha)}_{ij,k} \\ &:= E_\theta[\partial_i\partial_j\ell_\theta\partial_k\ell_\theta] + \frac{1-\alpha}{2}E_\theta[\partial_i\ell_\theta\partial_j\ell_\theta\partial_k\ell_\theta]. \end{aligned}$$

$\nabla^{(\alpha)}$ を M 上の **α-接続**（α-connection）と呼ぶ。但し g は Fisher 計量。これは座標系に依らない affine 接続を定めている（affine connection + tensor の形)。更に $\nabla^{(\alpha)}$ は torsion-free である（$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(\alpha)}_{ji,k}$)。

[3]

$$\ell^{(\alpha)}_\theta(\omega) := \begin{cases} \dfrac{2}{1-\alpha}p_\theta(\omega)^{\frac{1-\alpha}{2}} & (\alpha \neq 1) \\ \log p_\theta(\omega) & (\alpha = 1) \end{cases}$$

とおくと Fisher 計量、α-接続はそれぞれ

$$\begin{cases} g_{ij} = \displaystyle\int \partial_i\ell^{(\alpha)}_\theta(\omega)\partial_j\ell^{(-\alpha)}_\theta(\omega)d\mu \\ \Gamma^{(\alpha)}_{ij,k} = \displaystyle\int \partial_i\partial_j\ell^{(\alpha)}_\theta(\omega)\partial_k\ell^{(-\alpha)}_\theta d\mu \end{cases} \tag{3.2}$$

と表される（証明は単純計算)。この表示では [2] での定義と異なり $d\mu$ が θ に依らない測度となっている。これにより見通しが良くなる。この式は α-接続と $(-\alpha)$-接続の duality を示すときに使う（§ 5 [2] 参照)。g_{ij} の表示で一見 ∂_i と ∂_j の対称性が失われているように見えるが、ちゃんと対称になっている。また、接続とは 2 階微分をどのように行うかを決めるものであるが、それは $\partial_i\partial_j\ell^{(\alpha)}_\theta$ の部分に現れており、これから α-接続とは p_θ を $\ell^{(\alpha)}_\theta$ に変換しそこで自然に微分している（接空間への射影の項 $\partial_k\ell^{(-\alpha)}_\theta$ 付きで）事が解る。即ち $\nabla^{(\alpha)}$ は $\ell^{(\alpha)}_\theta$ の成す関数空間の自然な affine 構造から導かれる接続を M へ射影することによって得られる[4]。

[4] $\nabla^{(0)}$ は Fisher 計量 g に関する Levi-Civita 接続（Riemann 接続）になる。
($\because\ \partial_i g_{jk} = \Gamma^{(0)}_{ij,k} + \Gamma^{(0)}_{ik,j}$).

[5] $\alpha = 0$ は Fisher 計量の議論において自然に出てくる事は判ったが、それ以外で接続を考えて意味のある議論ではほぼ $\alpha = \pm 1$ の場合に限られる。これらには名前

[4]確率密度は L^1 であるが、何乗かすると L^p に属する（$p = \frac{2}{1-\alpha}$)。そして L^p の自然な affine 構造が入っている（が色々問題もある)。

が付いており $\alpha = 1$ のとき e-接続、$\alpha = -1$ のとき m-接続と云う。これらの由来について述べよう。

指数型分布族

$$p_\theta(\omega) = \exp[C(\omega) + \sum_i \theta^i F_i(\omega) - \psi(\theta)]$$

において自然座標系 $[\theta^i]$ に関する $\nabla^{(1)}$ の係数は

$$\Gamma^{(1)}_{ij,k} = E_\theta[\partial_i\partial_j\ell\partial_k\ell] = -\partial_i\partial_j\psi E_\theta[\partial_k\ell] = 0$$

$$\underbrace{}_{\|\ -\partial_i\partial_j\psi(\theta)} \qquad \underbrace{}_{\|\ 0\ (\because\ \S2\ [2])}$$

よって $\nabla^{(1)}$ は $[\theta^i]$ を affine 座標系とする平坦接続になる。$\nabla^{(1)}$ を**指数型接続**（exponential connection、e-connection）と呼び、$\nabla^{(1)} = \nabla^{(e)}$ と表す（Efron の “e” と云われた事もあり）。

[6] 混合型分布族[5]（mixture family）

$$p_\theta(\omega) = \sum_{i=1}^n \theta^i p_i(\omega) + (1 - \sum_{i=1}^n \theta^i)p_0(\omega)$$

において（$p_\theta(\omega) > 0$ となる範囲で θ を動かす） $[\theta^i]$ に関する $\nabla^{(-1)}$ の係数は (3.2) から

$$\Gamma^{(-1)}_{ij,k} = \int \underbrace{\partial_i\partial_j\ell^{(-1)}}_{\|\ \partial_i\partial_j p_\theta = 0} \partial_k\ell^{(1)} d\mu = 0$$

となる。よって $\nabla^{(-1)}$ は $[\theta^i]$ を affine 座標系とする平坦接続になる。$\nabla^{(-1)}$ を**混合型接続**（mixture connection、m-connection）と呼び、$\nabla^{(-1)} = \nabla^{(m)}$ と表す。情報幾何において指数型分布族はよく現れる。それに付随して混合型分布族もよく現れるが、ここでの形としては稀である。確率分布とは積分して 1 という条件を満たす函数であるが、この条件は線型（affine）拘束条件であるから、積分して 1 となる函数全体は全函数の中で余次元 1 の affine 部分空間を成す（$|\Omega| = \infty$ の場合は位相など難しくなる）。この平坦な空間の affine 部分空間で表されるものを**混合型分布族**と思えば良い。即ち、一般に統計多様体 $M \subset \mathcal{P}(\Omega)$ が $\mathbb{R}^\Omega = \{F \mid F : \Omega \to \mathbb{R}\}$ の中の affine 部分空間 V によって

$$M = \mathcal{P}(\Omega) \cap V$$

[5] $\theta^i > 0, (1 - \sum \theta^i) > 0$ の場合、$p_1 \sim p_n$ と p_0 の $n+1$ 個の分布の混合形、このように複数の分布から別の分布を作る事を混合を取るなどと云う。

と表されるとき M を**混合型分布族**（mixture family）と呼ぶ。これは幾つかの確率変数が与えられていて、その期待値が或る指定された値になるという条件を満たす確率分布族の集まりとしてよく現れる。

例. $M = \{p \in \mathcal{P} \mid E_p[F_i] = c_i,\ {}^\forall i \in \{1, \dots, k\}\}$ は混合型分布族、但し $F_i : \Omega \to \mathbb{R}$ と $c_i \in \mathbb{R}$ は given.

[7]

∘ M：指数型分布族、$N \subset M$：部分多様体のとき

N が M において e-自己平行 $\Leftrightarrow$ N が指数型分布族

（M が指数型分布族なので自然座標系で書けている。N が e-自己平行なら M の自然座標系に関して affine 部分空間を成す。その affine 部分空間の具体的表示を使って N の分布を書き直すと N 自身が指数型分布族である事が判る。逆はもう少し注意深く行う必要がある。英語版（Reference [2]）には書いてある）。

∘ M：混合型分布族、$N \subset M$ のとき

N が M で m-自己平行 $\Leftrightarrow$ N が混合型分布族。

注意１. $M = \mathcal{P}(\Omega)$ $(|\Omega| < \infty)$ の場合[6]

$$\begin{cases} \text{指数型分布族} \ \Leftrightarrow\ \text{e-自己平行、} \\ \text{混合型分布族} \ \Leftrightarrow\ \text{m-自己平行。} \end{cases}$$

注意２. この話の α-version がある。α-接続への拡張が面白いかは別にして、そもそも非自明な結果が余り多くない。これは非自明なものの一つで、指数型分布族、混合型分布族の α-版として α-family が考えられ、$|\Omega| < \infty$ の場合は $\mathcal{P}(\Omega)$ 自体は任意の α に対して α-family になる。このとき α-autoparallel がどういう形になるかはちゃんと判っている（英語版（Reference [2]）には載っている）。

§ 4. 不変性と単調性 （配布資料 [1] 参照）

ここでの内容は情報幾何の応用というよりは基礎付けである。しかしながら応用に関係する事もある。不変性、単調性は全て確率分布を別の確率分布に変換する操作と係った概念である。何かしらの確率系があった場合にその結果を観測し、その結果に何か情報処理をして別のものにする。但し元の確率構造に関しては何も知らないとする。このとき元の確率構造が変われば変換した後の確率構造も変わる。このような状況において単調性とは、計量に関する性質であるが、操作を行うと計量は等しいか減るのどちらかであり、決して増える事はないというものである。不変性とは確率構造の変換が可逆であれば計量は不変に保たれるというものである。不変性については接続についても定義できて、特に α-接続は不変に保たれる。逆にこ

[6] これは指数型分布族であり、また混合型分布族の自明な場合でもある

のような条件を課すと Fisher 計量と α-接続しかない事も判る（Čencov の定理[7]）。これらについて概観する。先ずどのような変換を考えるかについてから始める。ω を x に写像で変換するか、あるいは（より一般に）確率的に変換することを考える。以下 $|\Omega| < \infty$ を仮定する。

[1] 有限集合 Ω、$\mathfrak{X}$（$|\mathfrak{X}| < \infty$ は仮定）に対し

$$\begin{aligned} Q :&\Omega \times \mathfrak{X} \to \mathbb{R} \\ &(\omega, x) \mapsto Q(x|\omega) \geqq 0 \text{ s.t. } {}^{\forall}\omega,\ \sum_x Q(x|\omega) = 1 \end{aligned}$$

を満たす Q を Ω から $\mathfrak{X}$ への**通信路**（channel）と呼ぶ[8]。特に写像 $F : \Omega \to \mathfrak{X}$ から

$$Q_F(x|\omega) = \begin{cases} 1 & \text{if } x = F(\omega) \\ 0 & \text{otherwise} \end{cases}$$

により定まる Q_F を **deterministic channel** と呼ぶ。これはデータに関する変換である。

通信路 Q に対し

$$\begin{aligned} \Phi_Q : \mathcal{P}(\Omega) &\to \mathcal{P}(\mathfrak{X}) \\ p \mapsto \Phi_Q(p) : x &\mapsto \sum_\omega Q(x|\omega)p(\omega) \end{aligned}$$

と定める（$\Phi_Q(\mathcal{P}(\Omega)) \subset \mathcal{P}(\mathfrak{X})$ は仮定する）：

$$p \sim \omega \ \longrightarrow \ \underset{\text{channel}}{\boxed{Q}} \ \longrightarrow \ x \sim \Phi_Q(p)$$

（確率分布 p に従ってデータ ω が出てくる。これを通信路 Q に通して出てきたデータ x の従う確率分布が $\Phi_Q(p)$）。このような Φ_Q を（ここでは Čencov に敬意を表して）**マルコフ写像**（Markov map）と呼ぶ。特に deterministic channel Q_F の場合これは確率分布 p から F に関する分布を導く操作に対応している。

---

[7] Chentsov とも表される彼は著書（Reference [3]）において α-接続（に相当するもの）を最初に導入した。当時 Fisher 計量と不変性は既に知られていた。彼は逆に不変性で特徴付けられることを示した。この著書に α-接続の話は少ししか出てこない。当時はロシア語で書かれ西側には殆ど伝わらなかった。Efron、Amari の発見は独立である。接続があると平均の概念が（測地線の中間として）定まり、幾らか議論はしているが、曲率や（データ数の多くなったときの）漸近理論と関係はさせていない。著書は全てカテゴリーの言葉で書かれており解読は大変である。

[8] これは情報理論の通信への応用を考えて出てきた言葉であるが、今では通信と関係ない分野でも用いられる。元々は communication channel であったものが channel となったのであるが、日本語では “路” と云う表現は余り広まっておらず通信路と云われる。また推移確率と呼んでも良い。

[2] Fisher 計量の単調性

$$\begin{array}{ccc}\mathcal{P}(\Omega) \supset M = \{p_\theta\} & \overset{\text{Fisher}}{\longrightarrow} & G = [g_{ij}] \\ \Phi_Q \downarrow & & \\ \mathcal{P}(\mathfrak{X}) \supset M' = \{p'_\theta = \Phi_Q(p_\theta)\} & \overset{\text{Fisher}}{\longrightarrow} & G' = [g'_{ij}]\end{array}$$

このとき $\boxed{G(\theta) \geq G'(\theta)}$ ($^\forall\theta$) が成り立つ（i.e. （左辺）−（右辺）が半正定値）。Fisher 情報量とはデータが未知パラメータ θ に関して持っている情報量であり、θ に依存しない変換（操作）によって θ に関する情報量は減る事はあっても増える事はない。

証明は（逆向きの）条件付確率を使えば簡単。ポイントは

$(\star)\quad p_\theta(\omega)Q(x|\omega) = p'_\theta(x)Q'_\theta(\omega|x)$

となる $Q'_\theta(\omega|x)$ を使う。幾何的には Φ を接空間の対応にした場合

接ベクトルのノルムは減る事はあっても増える事はないという性質。配布資料 [1] 参照。

[3] $M \subset \mathcal{P}(\Omega)$、$\Phi = \Phi_Q : \mathcal{P}(\Omega) \to \mathcal{P}(\mathfrak{X})$ に対し

Φ は M 上で**可逆**（M-rev.）$\overset{\text{def}}{\iff} {}^\exists\Psi(= \Phi_R) : \mathcal{P}(\mathfrak{X}) \overset{\text{Markov}}{\longrightarrow} \mathcal{P}(\Omega)$

s.t. $^\forall p \in M,\ \Psi(\Phi(p)) = p.$

と定める。可逆のとき単調性から不変性が従う：

○ Fisher 計量の不変性：

$M, M' = \Phi(M)$ の Fisher 計量を G, G' とおくと

Φ は M 上で可逆 $\Rightarrow {}^\forall\theta,\ G(\theta) = G'(\theta).$

（実は $\Leftarrow$ も成り立つ）

○ $Q = Q_F$（deterministic channel）の場合：

Φ_Q が M 上で可逆 $\Leftrightarrow$ F は M の十分統計量（sufficient for M）。

余り統計では意識されていないが、可逆性と十分統計量は同等。上の $(\star)$ で θ に依らない Q'_θ を作るとちゃんと可逆になっているという事。

○ 特に $F : \Omega \to \mathfrak{X}$ が 1 対 1（単射）ならば Φ_Q は可逆。さらに $F : \Omega \overset{1:1}{\longrightarrow} \Omega$ の場合は (M, g) に関する対称性を導く（配布資料 [1] 参照）。

例. 正規分布 $N(\mu, \sigma^2)$

$0 \neq a,\ b \in \mathbb{R}$ に対し変換

$$\begin{array}{c}\mathbb{R} \to \mathbb{R} \\ \omega \mapsto a\omega + b\end{array}$$

を考える。ω が正規分布に従えば $a\omega+b$ も正規分布に従う。またこの変換は 1 対 1。これより不変性は

$$\begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{a^2} \end{bmatrix} G_{(\mu',\sigma'^2)} \begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{a^2} \end{bmatrix} = G_{(\mu,\sigma^2)}$$

（但し $\mu'=a\mu+b,\ \sigma'^2=a^2\sigma^2$）が全ての μ,σ と a,b に対して成立する。逆にこれらからどれくらい決定できるかと云うと、この条件を満たす計量 G は

$$G_{(\mu,\sigma^2)} = \begin{bmatrix} \frac{C_1}{\sigma^2} & 0 \\ 0 & \frac{C_2}{\sigma^4} \end{bmatrix}, \qquad C_1>0, C_2>0$$

の形に限られる事が判る。実際の Fisher 計量は

$$G = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

である。$C_1=2C_2$ ならば不変性で完全に特徴付けられる事になるが、これは導けない。即ち、一つの統計多様体の対称性だけでは計量や接続を特徴付けるのは無理である。

[4] 接続 $\nabla^{(\alpha)}$ も計量 g と同様の不変性を満たす。Markov 変換で接続係数がどう変わるかを見ればよい。可逆な場合には接続係数が変わらない。また幾何的には共変微分が $\nabla^{(\alpha)\prime}\Phi_*=\Phi_*\nabla^{(\alpha)}$ を満たす。しかしながら、計量とは違い単調性に相当するものはない。計量と同様に 1 つの多様体上の対称性からの特徴付けは難しい（配布資料 [1] 参照）。

[5] Čencov の定理：配布資料 [1] 参照。

§ 5. 双対接続

[1] 一般に多様体 M 上の Riemannn 計量 g、affine 接続 ∇,∇^* に対し

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y), \qquad {}^\forall X,Y,Z: \text{ vector fields on } M$$

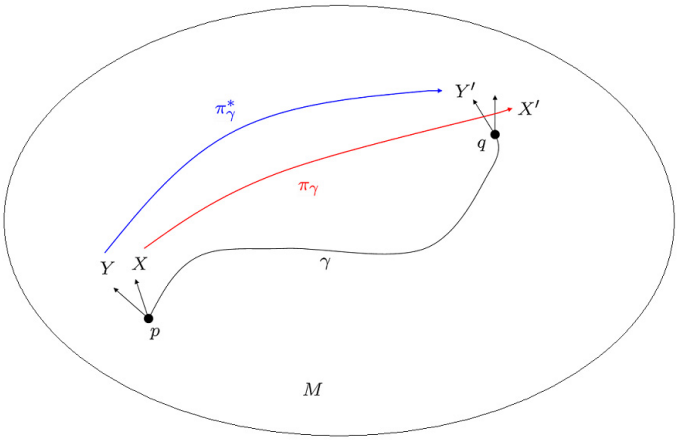
が成り立つとき ∇ と ∇^* は g に関して**双対的**（dual）であると云い、∇^* を ∇ の g に関する**双対計量**と呼ぶ。このとき $(\nabla^*)^*=\nabla$ と $\partial_i g_{jk} = \Gamma_{ij,k} + \Gamma^*_{ik,j}$ が成立。

双対性の意味：affine 接続から曲線に沿った平行移動が定義される。γ を 2 点 p,q を結ぶ任意の曲線とし、∇,∇^* に関する平行移動をそれぞれ

$$\pi_\gamma : T_p \overset{\nabla\text{-平行}}{\longrightarrow} T_q$$

$$\pi^*_\gamma : T_p \overset{\nabla^*\text{-平行}}{\longrightarrow} T_q$$

とすると $g(X,Y)=g(X',Y')$ が成立。証明は略す。計量接続の場合を少し拡張するだけである。



注意. ∇ は自己双対（$\nabla = \nabla^*$）$\Leftrightarrow$ ∇ は g を保存（metric connection）。

[2] $\nabla^{(\alpha)}$ と $\nabla^{(-\alpha)}$ は Fisher 計量 g に関して双対的。（$\because$ § 3 [3] の $\ell_\theta^{(\alpha)}$ を用いた表現から明らか。）

[3] 一般に互いに双対的な ∇, ∇^* に関して

∇ は curvature free $\Leftrightarrow$ ∇^* は curvature free.

（曲率テンソルが零だというのは、局所的には、平行移動が曲線に依らず始点と終点のみで定まる事であるが、一方が曲線に依らなければ $g(X,Y) = g(X',Y')$ から他方も曲線に依らない。また ∇, ∇^* の曲率テンソルを具体的に書いても確かめられる。）

特に α-接続は torsion-free なので

$\nabla^{(\alpha)}$ は平坦 $\Leftrightarrow$ $\nabla^{(-\alpha)}$ は平坦。

例えば

$$\begin{cases} \text{指数型分布族上の } \nabla^{(m)} \text{ は平坦、} \\ \text{混合型分布族上の } \nabla^{(e)} \text{ も平坦。} \end{cases}$$

[4] ∇ と ∇^* がともに平坦のとき (M, g, ∇, ∇^*) を**双対平坦空間**（dually flat space）と呼ぶ。これに対し以下が成り立つ。

(1)

$$\begin{cases} [\theta^i] \text{ を } \nabla \text{ の affine 座標系, } \partial_i := \frac{\partial}{\partial \theta^i} \text{ を自然基底} \\ [\eta_i] \text{ を } \nabla^* \text{ の affine 座標系, } \partial^i := \frac{\partial}{\partial \eta_i} \text{ を自然基底} \end{cases}$$

とすると ∂_i は ∇-平行、∂^i は ∇^*- 平行なので、双対性より

$$g(\partial_i, \partial^j) \equiv \text{cont. on } M.$$

affine 座標系は affine 変換の自由度があるから、特に $g(\partial_i, \partial^j) = \delta_i{}^j$ を満たすように $[\theta^i]$, $[\eta_i]$ をとることができる。このとき $[\theta^i]$ と $[\eta_j]$ は g に関して **dual** であるという。

例. 指数型分布族において

自然座標系 $[\theta^i]$: e-affine（e-接続について affine な座標系）

$\updownarrow$ dual

期待値座標系[9] $[\eta_i]$: m-affine（m-接続について affine な座標系），

ここで $\eta_i := E_\theta[F_i]$ であり、F_i は $p_\theta = \exp(C + \sum_i \theta^i F_i - \psi)$ の F_i.

証明は計算で $g(\partial_i, \partial^j) = \delta_i{}^j$ を確かめればよい。

(2)

$$\begin{cases} \partial_i \eta_j = g_{ij} = g(\partial_i, \partial_j) : \eta_j \text{ の } \theta^i \text{ に関する変換行列} \\ \qquad\qquad\qquad \updownarrow \text{ 逆行列} \\ \partial^i \theta^j = g^{ij} = g(\partial^i, \partial^j) : \theta^i \text{ の } \eta_i \text{ に関する変換行列} \end{cases}$$

これより座標変換行列が計量行列である事が判る。

(3) $\partial_i \eta_j = g_{ij} = g_{ji} = \partial_j \eta_i$ から $\eta_j d\theta^i$ は積分出来る（exact）。Poincaré の補題を用いると次が判る：$^\exists\psi : M \to \mathbb{R}$, $^\exists\varphi : M \to \mathbb{R}$ s.t.

$$\text{(5.1)} \qquad \begin{cases} \eta_j = \partial_j \psi, \ \theta^j = \partial^i \varphi, \\ \psi + \varphi = \sum_i \theta^i \eta_i, \end{cases}$$

（但し φ, ψ は局所函数）。

例. 指数型分布族では

$$\begin{cases} \psi = \log \int \exp[C(\omega) + \sum_i \theta^i F_i(\omega)] d\mu, \\ \varphi = \int p_\theta \log p_\theta d\mu - E_\theta[C]. \end{cases}$$

φ は統計力学の free energy の一種であり、$C = 0$ ならばマイナス・エントロピー。この φ, ψ は一意には定まらず、本質的に affine 変換に相当する自由度が残る。これらを使って次の [5] で canonical divergence を定義するがこれは不定性を有しない。また指数型分布族上でこれは相対エントロピーになる。

(4)

$$\begin{cases} \psi \text{ は } \theta \text{ の函数として凸} \quad (\because \text{ (2) と (5.1) から } \partial_i \partial_j \psi = g_{ij}), \\ \varphi \text{ は } \eta \text{ の函数として凸} \quad (\because \text{ (2) と (5.1) から } \partial^i \partial^j \varphi = g^{ij}), \end{cases}$$

であり

[9] expectation coordinates

$$\begin{cases} \varphi(\eta) = \max\limits_{\theta}\{\sum\limits_i \theta^i\eta_i - \psi(\theta)\}, \\ \psi(\theta) = \max\limits_{\eta}\{\sum\limits_i \theta^i\eta_i - \varphi(\eta)\} \end{cases}$$

が成り立つ。これを **Legendre 変換**という（但し θ は自然座標として、η は期待値座標として意味のある範囲を動くものとする。境界に行くと色々変な事も起きる）。

$\because$) $\eta_i(\theta) = \partial_i\psi(\theta)$ より

$$\begin{cases} \frac{\partial}{\partial\theta^i}(\sum_j \theta^j\eta_j - \psi(\theta))|_{\eta=\eta(\theta)} = 0, \\ \theta \mapsto (\sum_j \theta^j\eta_j - \psi(\theta)) \text{ は凹（上に凸）。} \end{cases}$$

$\therefore$ $\max\limits_{\theta}\{\sum\limits_i \theta^i\eta_i - \psi(\theta)\}$

$= \sum\limits_i \theta^i(\eta)\eta_i - \psi(\theta(\eta))$ ($\because$ $\theta = \theta(\eta) \Leftrightarrow \eta = \eta(\theta)$: 同一点の 2 つの座標の値)

$= \varphi(\eta)$ ($\because$ (5.1)).

もう一つも同様に示せる。

[5] 以上の状況（双対平坦空間）において M 上の 2 変数函数を

$$D : M \times M \to \mathbb{R}, \quad (p, q) \mapsto D(p\|q) := \varphi(p) + \psi(q) - \sum_i \eta_i(p)\theta^i(q) \tag{5.2}$$

とおくと $^\forall p, q, r \in M$ に対し

$$D(p\|q) + D(q\|r) - D(p\|r) = \sum_i \{\eta_i(p) - \eta_i(q)\}\{\theta^i(r) - \theta^i(q)\} \tag{5.3}$$

が成り立つ。また

$$\begin{cases} D(p\|q) \geqq 0 \quad (^\forall p, q \in M) \\ \text{等号} \Leftrightarrow p = q \end{cases}$$

が成り立つ（$\because$ 凸性と (5.1) より）。逆に非負値関数 $D : M \times M \to \mathbb{R}$ が (5.3) を満たせば必ず (5.2) の形に表せる。この D を (M, g, ∇, ∇^*) の（∇^* に関する）**canonical divergence** とよぶ（注 : ∇, ∇^* の順番に依る）。

注意. ∇ に関する canonical divergence は $(p, q) \mapsto D(q\|p)$ になる。

例. 指数型分布族では $\nabla^{(m)}$-divergence（m-divergence、相対エントロピー、KL divergence）は

$$\boxed{D(p\|q) = \int p \log \frac{p}{q} d\mu}$$

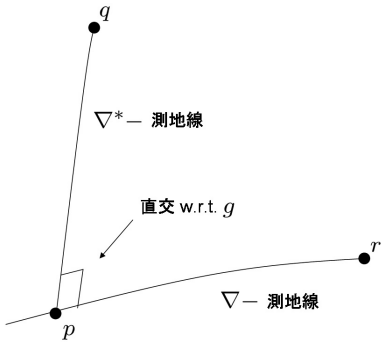
と表される（証明は (5.2) または (5.3) と非負性を確かめる）。

[6] 拡張ピタゴラス

$D(p\|q)$ は p と q の距離の自乗のようなものである。実際

$$D(p\|r) = D(p\|q) + D(q\|r)$$

が成立：



注：

測地線 = 自己平行曲線（：接ベクトルが接続に関し平行）

∇^*-測地線 = $[\eta_i]$ についての直線

∇-測地線 = $[\theta^i]$ についての直線

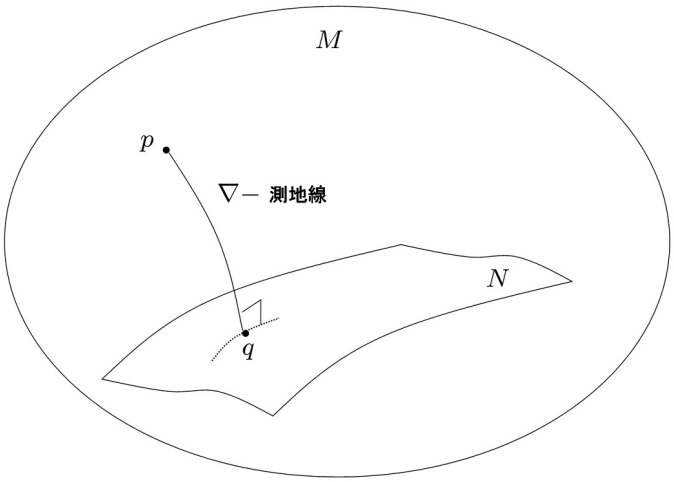
[7] ∇-射影

(M, g, ∇) と M の部分多様体 N、$p \in M$ と $q \in N$ に対し

q が p の N への **∇-射影**（∇-projection）

$\Longleftrightarrow$ q と p を結ぶ ∇-測地線が q において N と直交

と定める。



双対平坦空間 (M, g, ∇, ∇^*)、∇^*-divergence D に対して

- $q \in N$ が p の N への ∇-射影 $\Leftrightarrow$ q が $D(\cdot\|p)|_N$ の停留点
- $q \in N$ が p の N への ∇^*-射影 $\Leftrightarrow$ q が $D(p\|\cdot)|_N$ の停留点

$\because$) $D(q\|p) = \varphi(q) + \psi(p) - \sum_i \eta_i(q)\theta^i(p)$

に対し p を fix して q について微分すると（$\tilde{\partial}$ で表す）

$$\tilde{\partial}^i D(q\|p) = \underbrace{\partial^i \varphi(q)}_{\| \ \theta^i(q)} - \theta^i(p)$$

より明らか（N の座標系を導入する必要あり。準備が必要なので略）。

○ 指数型分布族では $\nabla = \nabla^{(e)}$, $\nabla^* = \nabla^{(m)}$ で m-射影、e-射影の話になる。
○ N が ∇^*-自己平行ならば ∇-射影 q は p に対し一意に定まり

$$D(q\|p) = \min_{r \in N}(r\|p)$$

になる（拡張ピタゴラスより）。
○ N が ∇-自己平行でも同様の主張が成り立つ。

例１．m-射影が一番良く出てくるのは N のどこかに真の分布があり（当然どこかは不明で）何も方法のないときの最尤推定である。M を指数型分布族とし

データ $\underbrace{\omega_1, \omega_2, \ldots, \omega_N}_{\downarrow}$

$$\frac{1}{N}\sum_{t=1}^{N} F_i(\omega_t) = \eta_i(\hat{\theta})$$

とすると $\hat{\theta}$ は M での最尤推定

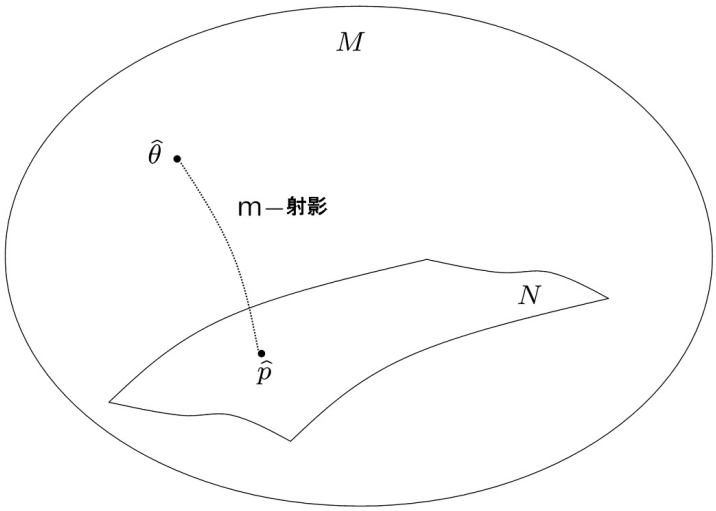
$\downarrow$ N への m-射影

$\hat{p}$

$\hat{p}$ は N での最尤推定（尤度方程式の解）[10]。

m-射影はデータが与えられ M での最尤を求めたら、その点と N の点との divergence（相対エントロピー）を考え、その意味で一番近いものとして出てくる。最尤推定は符号を変えると尤度 + 定数と見做せるので最小化の操作が尤度最大となる。

[10]指数型分布族の中の点には３つの捉え方がある：確率分布、自然座標系での座標値、η-座標系での座標値。これらは文字の違い：$\theta, \eta, \hat{\theta}, \hat{\eta}$ などで判読せよ。



例２. e-射影は大偏差（large deviation）で現れる。次節参照。

§ 6. 指数型分布族における大偏差問題

i.i.d（independent and identically distributed; 独立に同じ分布に従う）の場合の大偏差を学ぶと必ず二つの定理に出会う。一つは Sanov の定理（経験分布が或る領域に入る確率のレート函数で相対エントロピーが現れる）、もう一つは Cramér の定理（実確率変数に関する大偏差で積率母関数（moment generating function）、ψ が出てくる形でレート函数が与えられている）。有限次元の指数型分布族ではこれらは同じ定理であり、同じものを別の見方をしている。これを理解する事がここでの目的である。

[1] 指数型分布族 $S = \{p_\theta\}$ で

$$p_\theta(\omega) = \exp\left[C(\omega) + \sum_{i=1}^{d} \theta^i F_i(\omega) - \psi(\theta)\right]$$

となっているものが与えられているとする。このとき次の 1 対 1 対応がある：

$$\begin{array}{ccc} \mathcal{P}(\Omega) \supset S & & \mathbb{R}^d \\ \cup & & \cup \\ p_\theta & \longleftrightarrow & \theta = \theta(\eta) \\ \| & \swarrow & \updownarrow \\ p_\eta & & \\ & & \eta = [E_\theta[F_i]] = \eta(\theta) \\ & & \cap \\ & & \mathbb{R}^d \end{array}$$

いま $S \supset W$ と $\mathbb{R}^d \supset V$ が

$$\begin{array}{ccc} W & & V \\ \cup & & \cup \\ p & \longleftrightarrow & \eta \end{array}$$

によって互いに対応しているとする。ただし、W, V は d 次元の閉領域[11]であるとする。このとき任意の $q \in S$ に対し

$$\gamma_N := \mathrm{Prob}\left\{ \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{F}(\omega_t) \in V \mid \underbrace{(\omega_1, \dots, \omega_N)}_{\| \ \omega^N} \overset{i.i.d.}{\sim} q \right\} \qquad (\text{where } \boldsymbol{F} = (F_1, \dots, F_d))$$

$$= \mathrm{Prob}\{\hat{p}_{\omega^N} \in W \mid \omega^N \overset{i.i.d.}{\sim} q\},$$

但し[12]

$$\begin{aligned} \hat{p}_{\omega^N} &= p_{\hat{\theta}(\omega^N)} = p_{\eta = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{F}(\omega_t)} \\ &= \arg\max_{\theta} p_\theta(\omega_1) \cdot \dots \cdot p_\theta(\omega_N) \\ &\quad (\text{最尤推定}) \end{aligned}$$

とおくと[13]

$$\boxed{\lim_{N\to\infty} \frac{1}{N} \log \gamma_N = -\min_{p \in W} D(p\|q) = -\min_{\eta \in V} D(p_\eta \| q)}$$

が成り立つ（一般化された Sanov の定理)。
ここで $q = p_{\theta_0}$, $\xi^i = \theta - \theta_0^i$, $\psi_q(\xi) = \psi(\theta) - \psi(\theta_0)$ と変換すると

$$S = \{p_\xi\}, \quad \text{where } p_\xi(\omega) = q(\omega) \exp[\sum_i \xi^i F_i(\omega) - \psi_q(\xi)]$$

$$\psi_q(\xi) = \log \int q(\omega) \exp[\sum_i \xi^i F_i(\omega)] d\mu$$

$$= \log E_q[e^{\sum_i \xi^i F_i}]$$

---

[11]即ち V の内部の閉包が V 自身に等しい、i.e., $\overline{V^\circ} = V$.

[12]前頁の図式から $\frac{1}{N}\sum \boldsymbol{F}(\omega_t) = \eta \ (\in V)$ から θ が定まり、更に S の分布 p が定まる。この分布 p は与えられたデータ $\boldsymbol{F}$ に対する S での最尤推定となる。η では閉領域 V が p では閉領域 W となるのでこれらの確率は当然等しい。

[13]$\hat{p}$ は $|\Omega| < \infty$ で $S = \mathcal{P}(\Omega)$ の場合、経験分布になる。本節末を参照。

と表される。このとき

$$\begin{aligned} D(p_\eta\|q) =&\varphi_q(p_\eta) + \underbrace{\psi_q(q)}_{\|\,0} - \sum_i \eta_i \cdot \underbrace{\xi^i(q)}_{\|\,0} \\ =&\sum_i \eta_i\xi^i(\eta) - \psi_q(\xi(\eta)) \qquad (\because\ \varphi+\psi = \sum_i \eta_i\theta^i) \\ =&\max_{\xi\in\mathbb{R}^d}(\sum \eta_i\xi^i - \psi_q(\xi)) \end{aligned}$$

$$\therefore\ \lim_{N\to\infty}\frac{1}{N}\log\gamma_N = -\min_{\eta\in V}\max_{\xi}(\sum_i \eta_i\xi^i - \psi_q(\xi))$$

（Cramér の定理）.

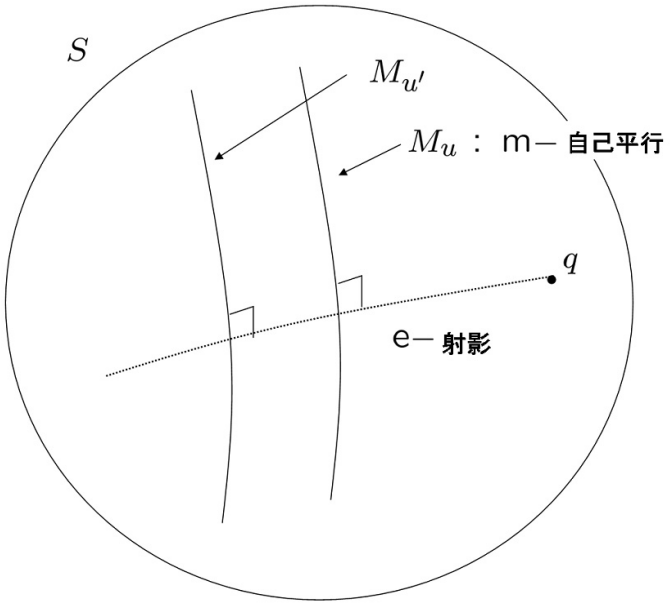
[2] 確率変数 $H_1,\dots,H_{d'} \in \mathrm{span}_{\mathbb{R}}\{F_1,\dots,F_d,1\}$（$d'<d$）と $\mathbb{R}^{d'}$ の領域 U に対し大偏差を考えると

$$\lim_{N\to\infty}\frac{1}{N}\log \mathrm{Prob}\{\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{H}(\omega_t)\in U \mid \omega^N \overset{i.i.d.}{\sim} q\} \qquad (\boldsymbol{H}=(H_1,\dots,H_{d'}))$$

$$= -\min_{p\in W} D(p\|q) \qquad \text{但し } W := \{p\in S \mid E_p[\boldsymbol{H}]\in U\}$$

$$= -\min_{u\in U}\min_{p\in M_u} D(p\|q) \qquad \text{但し } M_u := \{p\in S \mid E_p[\boldsymbol{H}]=u\}$$

$$= -\min_{u\in U} D(p'_u\|q) \qquad \text{但し } \begin{cases} p'_u(\omega) = q(\omega)\exp[\sum_{i=1}^{d'}\theta^i H_i(\omega) - \psi'_q(\theta)] \\ u = E_{p'_u}[\boldsymbol{H}] \qquad (\theta \overset{1:1}{\longleftrightarrow} u \text{ は前提。}) \end{cases}$$



外側の空間が $\mathcal{P}$ 全体とは限らないので M_u は混合型分布族とは限らない。あくまで S の中で m-自己平行。このとき e-射影で直交化する。M_u を動かしても $M_{u'}$ はちゃんと直交している。

これにより次元の高いところの大偏差から次元の低いところの大偏差が導かれる。

注意. $S = \mathcal{P}(\Omega)$（$|\Omega| < \infty$）の場合

$$\hat{p}_{\omega^N} = \frac{1}{N}\sum_{t=1}^{N} \delta_{\omega_t} \quad \text{（経験分布、type）.}$$

ここで δ_{ω_t} は ω_t のところだけ確率 1 をもつ（Kronecker's delta のような）もの。有限集合上の経験分布は情報理論でよく出てきて、type に関する大偏差が Sanov の定理。上記の議論は例えば Sanov の定理から 1 次元の Cramér の定理を導くのに使われる。

§ 7. 推定（配布資料その 2 参照）

推定の話は射影だとよく云われる。要は近似と同じと。しかしそれは本当であろうか。外側に指数型分布族があってそこでの最尤推定が利用出来る状況だとそこから全てが導ける（十分統計量になっている）。外側に S、その中にモデル M がある場合には外側の最尤推定からの写像（射影）だけ考えれば良い。では外側の最尤推定は何かという話になると普通それは幾何ではなくなる。しかし量子推定を考える場合、量子推定には射影はなかなか出てこない。それは十分統計量に相当するものがないからであるが、推定という問題は定式化出来、計量も出てくる。

射影の事をよく知っていると量子推定も理解出来ると思われがちであるが、量子相対エントロピーと一向に結びつかない。寧ろ量子相対エントロピーも量子推定も古典版では一つの（同一）幾何構造になっているだけであり、概念的には必ずしも同一ではない。改めて古典的推定を見直すと『どこで幾何が現れるか?』となる。推定を幾何的に捉える事がここでの目的である。

ここで一番云いたい事は配布資料その 2 [5] の Theorem である。簡単にこれを説明する。これは確率変数のバラつき具合の尺度をどのように定義するかという内容である。確率変数は観測すると値はバラつく（だから確率変数）。その特別な場合としていつも同じ値を返す一定な確率変数もある。バラつき具合はその一定な確率変数からどれくらいズレているかを測れば良い。この測り方には二通りある。右辺 $V_p(F)$ は分散。即ち分布 p の下で確率変数 F がどれくらいバラつくかは $\langle (F - \langle F \rangle_p)^2 \rangle_p$ で測れる。これは一つの分布 p に着目して測る尺度である。

一方左辺は F の期待値が p を少し変えたときにどれくらい変わるかを見ている。もし F が確率変数として一定ならば、期待値は分布に依らない。分布を少し変えたときに期待値が大きく変わるとすると、それは p の近くで F が大きくバラついている事を意味する。幾何学的には p に F の期待値を対応させる事によって得られる多様体上の函数の微分（余接ベクトル）をとり、この余接ベクトルの或る計量に関するノルムを測る。

Fisher 計量はこれら二つの尺度が一致する計量である事を定理は主張している：

$$\boxed{\|(d\langle F\rangle)_p\|^2 = V_p[F]}$$

これが成立するから推定論では非常に重要と考えるのが自然であり、量子推定の場合もここを出発点として色々な結果が得られる。この立場から双対接続なども出てくるが、相対エントロピーとは結びつかない。

その他詳細は配布資料その２を参照の事。以下に配布資料その２ [4] の gradient の導入の補足の図式を与えておく：

u の e-表現 $u^{(e)} = F - \langle F\rangle_p$ に対し

$$\begin{array}{ccccc} (d\langle F\rangle)_p & \overset{g}{\longleftrightarrow} & (\mathrm{grad}\langle F\rangle)_p & \overset{\text{e-rep.}}{\longleftrightarrow} & (\mathrm{grad}\langle F\rangle)_p^{(e)} \\ \cap & & \cap & & \cap \\ T_p^* & & T_p & & T_p^{(e)} \subset \mathcal{F} \end{array}$$

§ 8. 無限次元

ここでは $|\Omega| = \infty$ の場合の $\mathcal{P}(\Omega, \mathcal{F}, \mu)$ の多様体構造を概観する。先ず Ω が無限集合になると生じる困難の部分から始める。

$$\mathcal{P}(\Omega, \mathcal{F}, \mu) = \{p \in L^1 \mid \underbrace{p(\omega) > 0 \text{ a.e. } \mu}_{\text{難しい原因}}, \ \int_\Omega p d\mu = 1\}.$$

Ω が有限集合の場合、$\mathcal{P}(\Omega)$ が指数型分布族となる大前提は $\mathcal{P}(\Omega)$ が多様体となっている事である。例えば $|\Omega| = n$ としたとき、p は n 次元ベクトルと思って良い。このとき条件 $p(\omega) > 0$ はそのベクトルが第一象限に入っているとなる。第一象限は（$\mathbb{R}^n$ の）開集合なので問題はない。また $\int p d\mu = 1$[14] は次元が 1 減るだけである（逆に $\int p d\mu = 1$ を満たす p の中で $p > 0$ を満たすものは開集合を成すので、affine 部分空間の中の開集合として自然に多様体の構造が入るとしても良い）。$|\Omega|$ が大きくなるにつれ条件 $p(\omega) > 0$ は厳しい条件となってくる。第一象限は次元が高くなるにつれ“狭く”なる。$|\Omega| \to \infty$ で第一象限の次元は失われ細い集合となる。

$|\Omega| < \infty$ の場合、$\{f \mid \int f d\mu = 0\}$ が接空間となる。接空間では正値性は出てこない。第一象限は（全空間と等しいだけの）充分な次元を持つので、一点の近傍上では何の制約にもならない。しかし $|\Omega| = \infty$ の場合には $p(\omega) > 0$ から接空間は

[14] $|\Omega| < \infty$ の場合、この条件は $\sum p(\omega) = 1$ であるが、ここでは積分形で表しておく。以下も同様。

$\{f \mid \int f d\mu = 0\}$ とは云えない：即ち

$$T_p^{(m)}(\mathcal{P}) = \{\partial p \mid \partial \in T_p(\mathcal{P})\}$$
$$\stackrel{?}{=} \{f \mid \int f d\mu = 0\}.$$

これはまた、基点 p に対し f が接ベクトルなら $p+tf$ は $\mathcal{P}$ の内部に留まっているか？ と云い換えられる：

$$p \in \mathcal{P},\ \int f d\mu = 0 \stackrel{?}{\Longrightarrow} {}^{\exists}\varepsilon > 0, {}^{\forall}t \in (-\varepsilon, \varepsilon),\ p + tf \in \mathcal{P}.$$

$|\Omega| < \infty$ ならこれは成り立つ。またこれが成立すれば $p_t = p + tf \subset \mathcal{P}$ を微分して $\frac{dp_t}{dt} = f$ と接ベクトルが出てくる。$|\Omega| = \infty$ の場合ではこれは成立しない（Ω がコンパクトなら inf が存在し、それより小さい範囲で動かせば可能であり、関係する仕事もある。しかし統計としてこれでは弱い）。

さて $p \in L^1$ なので L^1 から入れるのが当然であり、基本的に多様体構造は混合型分布族から入れるべきで、指数型分布族は混合型の双対として現れると思われていた。しかし Pistone-Sempi は指数型分布族から $\mathcal{P}(\Omega)$ に多様体構造が入る事を示した。それを次に述べる。

$|\Omega| < \infty$ の場合、$\mathcal{P}(\Omega)$ が指数型分布族である事は既に見たが、指数型分布族の表現方法は沢山ある。自然パラメータは affine 座標系なので affine 変換の自由度がある訳である。いま $p \in \mathcal{P}(\Omega)$ を fix する。このとき

$$\mathcal{P}(\Omega) = \{e_p(u) \mid u \in T_p^{(e)}\}$$

という表現もある。ここで

$$\begin{cases} T_p^{(e)} = \{u \in \mathbb{R}^\Omega \mid E_p[u] = 0\}, \\ e_p(u) = \frac{1}{Z_p(u)} p e^u = \exp[\log p + u - \log Z_p(u)], \\ Z_p(u) := \sum_\omega p(\omega) e^{u(\omega)} = E_p[e^u]. \end{cases}$$

この表現を使って無限次元のときに多様体構造を入れる。

Orlicz space :

$\Phi : \mathbb{R} \to [0, \infty]$ が Young function

$\stackrel{\text{def}}{\Longleftrightarrow}$

$$\begin{cases} \circ\ \Phi \text{ is convex}, \\ \circ\ {}^{\forall}x \in \mathbb{R},\ \Phi(-x) = \Phi(x), \\ \circ\ \Phi(0) = 0, \\ \circ\ \lim_{x\to\infty} \Phi(x) = \infty, \end{cases}$$

と定義する。Young function Φ と $(\Omega, \mathcal{F}, pd\mu)$ に対し[15]

$$L^{\Phi}(pd\mu) \;=\; L^{\Phi}(p) \;:= \{u \mid u : \Omega \overset{\text{measurable}}{\longrightarrow} \mathbb{R} \text{ かつ } \|u\|_{p,\Phi} < \infty\}$$

を **Orlicz space** w.r.t. $(\Phi, pd\mu)$ と云う。ここで

$$\|u\|_{p,\Phi} = \inf\{k > 0 \mid \int \Phi(\frac{|u|}{k}) pd\mu \leqq 1\} \quad : \text{Luxembourg norm}$$

(Orlicz norm もあるが定義はこちらの方が簡単。norm としては同値)。このとき $(L^{\Phi}(p), \|\cdot\|_{p,\Phi})$ は Banach 空間になる。

例. $\Phi(x) = |x|^r \;\Rightarrow\; L^{\Phi}(p) = L^r(p)$ (: r 乗可積分函数全体)。

以下では

$$\Phi(x) = \cosh x - 1$$

とおく (当然 Young function)。このとき

$$L^{\Phi}(p) = \{u \in L^1(p) \mid {}^{\exists}\varepsilon > 0,\ {}^{\forall}t \in (-\varepsilon, \varepsilon),\ E_p[e^{tu}] < \infty\}$$

となる (Reference : Rao & Ren [4] 参照)。即ち積率母関数が 0 の近傍で存在する。

これがモデル空間である。次に座標近傍を導入する。以下 $p \in \mathcal{P}(\Omega, \mathcal{F}, \mu)$ (=: $\mathcal{P}$) を fix する。

$$\begin{aligned} B_p :=& \{u \in L^{\Phi}(p) \mid E_p[u] = 0\} \quad (= T_p^{(e)}(\mathcal{P})), \\ e_p : B_p &\to \mathcal{P}, \\ u &\mapsto \frac{1}{Z_p(u)} p e^u, \end{aligned}$$

ここで $Z_p(u) := E_p[e^u]$ であるが、$Z_p(u) < \infty$ が怪しい。そこでいま

$$\mathcal{V}_p := \{u \in B_p \mid \|u\|_{p,\Phi} < 1\}$$

とおくと ${}^{\forall}u \in \mathcal{V}_p \;\Rightarrow\; Z_p(u) < \infty$.

$\mathcal{U}_p := \{e_p(u) \mid u \in \mathcal{V}_p\}$ $(\subset \mathcal{P})$ とおく。すると $\{(\mathcal{U}_p, e_p^{-1})\}_{p \in \mathcal{P}}$ は $\mathcal{P}$ に C^{∞}-級多様体構造を定義する (Pistone-Sempi, 1995)。

$\mathcal{P}$ の位相は積率母関数を使って収束を定義する (詳細は略す)。$\mathcal{P}$ は連結ではなく無数の連結成分から成る。連結成分は

$$\begin{aligned} &\{e_p(u) \mid u \in \{u \mid Z_p(u) < \infty\}^{\circ}\} \\ =&\{q \mid q \text{ と } p \text{ を結ぶ } 1 \text{ 次元指数型分布族が (連続に) 作れる }\} \end{aligned}$$

となる。幾つか注意点を挙げると

$\circ$ $T_p(\mathcal{P}) \cong B_p$ と見做せる ($B_p = T_p^{(e)}(\mathcal{P})$)。

[15] ここで $pd\mu$ は probability measure.

∘ $B_p \subset L^2(p)$ なので Fisher 計量 $g(u,v) = E_p[uv]$ が定義される[16]。
∘ g は連続だが完備でない（Hilbert 空間にならない）。
∘ e-接続は OK（上手くいくよう作ってある）。
∘ m-接続は苦しい。$\mathcal{P}$ の中だけで閉じた接続は作れていない。しかし m-測地線は定義出来る：
$p, q \in C$：一つの連結成分 $\Rightarrow {}^\forall\varepsilon \in [0,1],\ \varepsilon p + (1-\varepsilon)q \in C$
（共変微分は外にはみ出るが、零だけは問題ないので測地線は上手くいく）。

References

[1] 甘利俊一, 長岡浩司, 情報幾何の方法, 岩波講座応用数学, 東京:岩波書店, 1993

[2] S. Amari and H. Nagaoka, Methods of information geometry, Translated from the 1993 Japanese original by Daishi Harada. Translations of Mathematical Monographs, 191. American Mathematical Society, Providence, RI; Oxford University Press, Oxford, 2000. x+206 pp. ISBN 0-8218-0531-2

[3] N.N. Chentsov, Statistical decision rules and optimal inference, Translation from the Russian edited by Lev J. Leifman. Translations of Mathematical Monographs, 53. American Mathematical Society, Providence, R.I., 1982. viii+499 pp. ISBN: 0-8218-4502-0

[4] M.M. Rao and Z.D. Ren, Theory of Orlicz spaces, Monographs and Textbooks in Pure and Applied Mathematics, 146. Marcel Dekker, Inc., New York, 1991, xii+449 pp. ISBN: 0-8247-8478-2

[16] e-表現の L^2-内積が Fisher 計量。