速習情報幾何

Arithmer Cloud Div. Kimiaki Kinugasa

2018/10/25

本スライドは、当社と国立大学法人東京大学との間の共同研究に基づき、2018年11月20日、同大学大学院数理科学研究科・理学部数学科において開催された離散数理モデリングセミナーに甘利俊一博士を招請するにあたり、事前の知識習得のために行った社内講義資料を公開用に一部修正したものです。

Self-Introduction

- Kimiaki Kinugasa
  - Academic Background
    - The University of Tokyo
    - Bachelor of Engineering
  - Former Job
    - Pal Software Service, Inc.
      - Programmer
  - Current Job
    - Cloud Architect
    - Cloud Developer
    - Team Manager

目次

- ガウス分布の幾何構造
- 指数型分布族
- 双対平坦空間

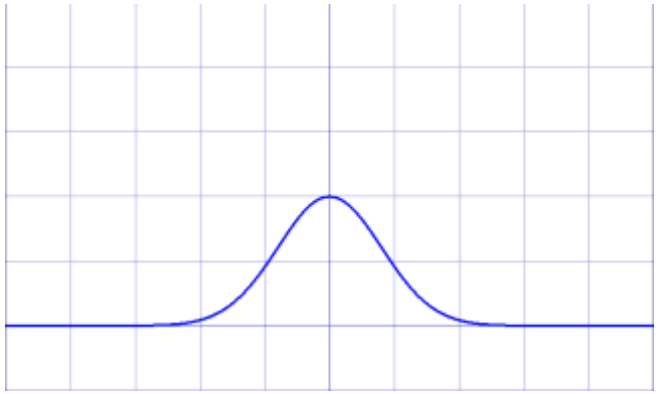
目次

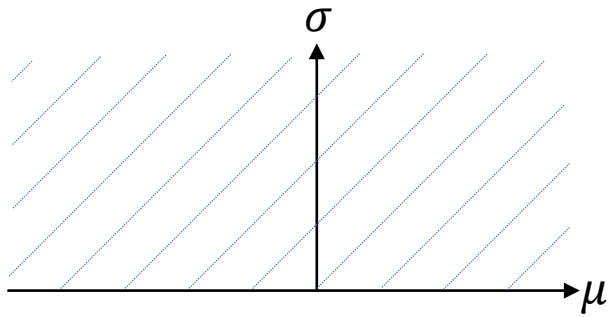
- ガウス分布の幾何構造
- 指数型分布族
- 双対平坦空間

ガウス分布の幾何構造

$$\mu \in \mathbb{R}, \sigma > 0, x \in \mathbb{R}$$

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

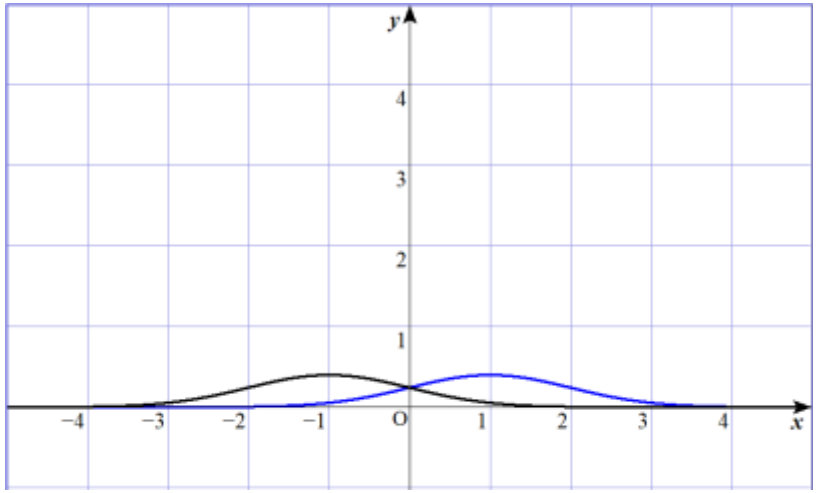




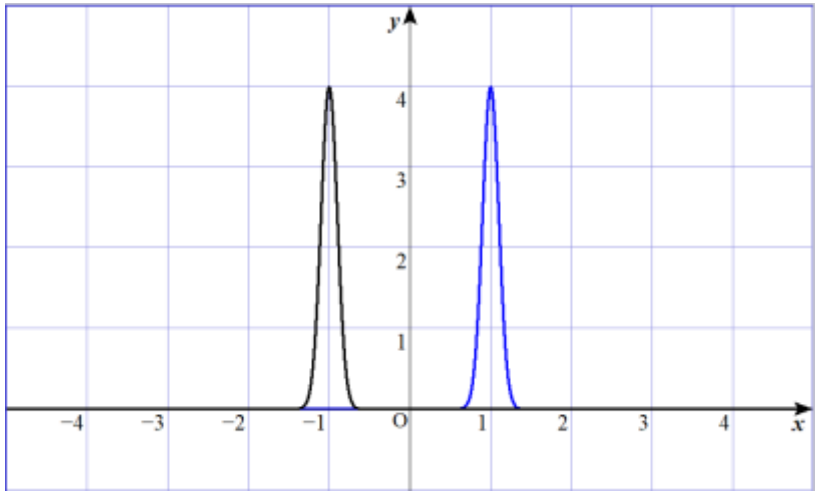
μ, σ空間の幾何構造について考える

μ, σ空間はユークリッド空間?

$(-1,1),\quad (1,1)$　の距離は2



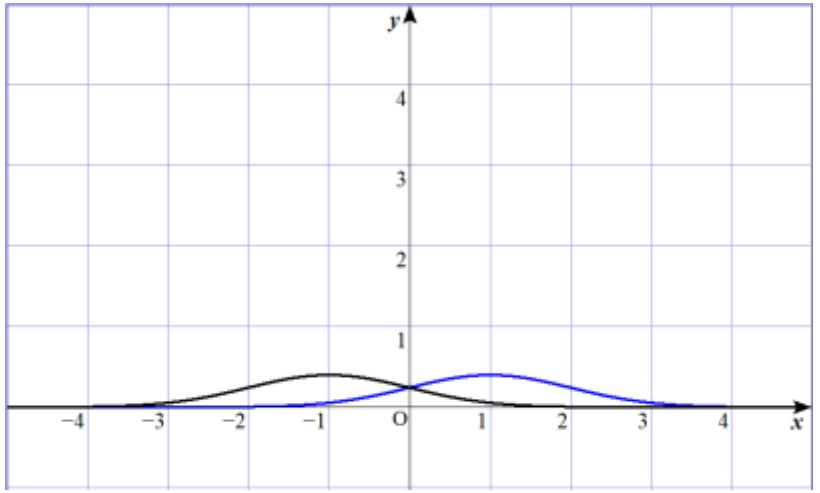
$(-1,0.1), (1,0.1)$ の距離も2



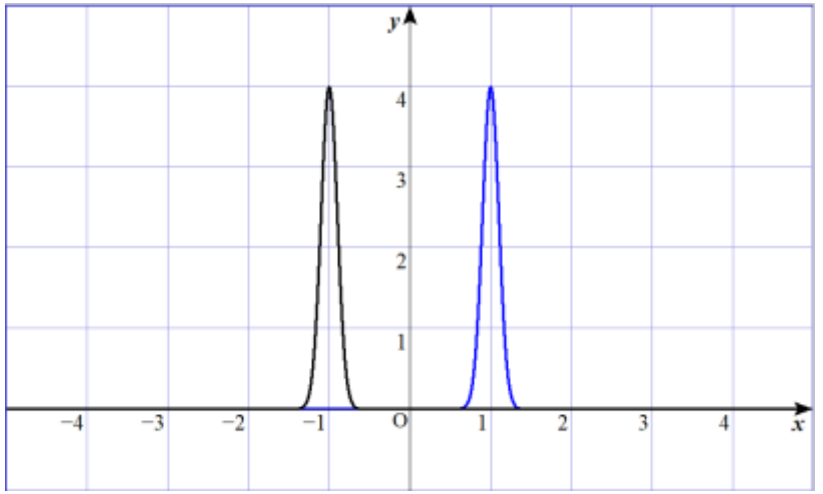
変な気がする・・・

μ, σ空間はユークリッド空間?

$(-1,1),\quad (1,1)$　の距離は2



$(-1,0.1), (1,0.1)$ の距離も2



変な気がする・・・

μ, σ空間が確率分布族の空間であることを無視している

μ, σ空間のリーマン計量

μ, σ空間の微小な2点間の距離は

$$ds^2 = (d\mu \quad d\sigma)\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}$$

と表せる

μ, σ空間をユークリッド空間とみなすと

$$ds^2 = (d\mu \quad d\sigma)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}$$

となる

リーマン計量gとして何を選ぶべきか?

μ, σ空間のリーマン計量

μ, σ空間の微小な2点間の距離は

$$\mathrm{d}s^2 = (\mathrm{d}\mu \quad \mathrm{d}\sigma) \begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} \begin{pmatrix} \mathrm{d}\mu \\ \mathrm{d}\sigma \end{pmatrix}$$

と表せる

μ, σ空間をユークリッド空間とみなすと

$$\mathrm{d}s^2 = (\mathrm{d}\mu \quad \mathrm{d}\sigma) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathrm{d}\mu \\ \mathrm{d}\sigma \end{pmatrix}$$

となる

リーマン計量gとして何を選ぶべきか?

結論
Fisher情報行列を使うべき

Fisher情報行列

定義:
確率密度関数$p(x;\theta)$に対して、そのFisher情報行列を

$$\mathrm{E}\left[\frac{\partial \log p}{\partial \theta}\left(\frac{\partial \log p}{\partial \theta}\right)^{\mathrm{T}}\right]$$

と定義する。

ガウス分布の場合、Fisher情報行列は

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

となる。

Fisher情報行列

定義:
確率密度関数$p(x;\theta)$に対して、そのFisher情報行列を

$$\mathrm{E}\left[\frac{\partial \log p}{\partial \theta}\left(\frac{\partial \log p}{\partial \theta}\right)^{\mathrm{T}}\right]$$

と定義する。

ガウス分布の場合、Fisher情報行列は

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

となる。

平面(ユークリッド空間)じゃない！

ガウス曲率

ガウス曲率: 空間の各点における曲がり具合を表現する量

2次元曲面の微小距離$\mathrm{d}s^2$が1次微分形式θ^1, θ^2を用いて

$$\mathrm{d}s^2 = \theta^1\theta^1 + \theta^2\theta^2$$

で与えられたるとき、

第1構造式

$$\mathrm{d}\theta^1 = \theta^2 \wedge \omega_2^1$$
$$\mathrm{d}\theta^2 = \theta^1 \wedge \omega_1^2$$

但し、 $\omega_2^1 = -\omega_1^2$

第2構造式

$$\mathrm{d}\omega_2^1 = K\theta^1 \wedge \theta^2$$

但し、 Kはガウス曲率

μ, σ空間のガウス曲率

$$ds^2 = (d\mu \quad d\sigma)\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}$$

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

より、

$$ds^2 = \frac{1}{\sigma^2}d\mu^2 + \frac{2}{\sigma^2}d\sigma^2$$

従って、

$$\theta^1 = \frac{1}{\sigma}d\mu, \theta^2 = \frac{\sqrt{2}}{\sigma}d\sigma$$

と表せる。

μ, σ空間のガウス曲率

$$ds^2 = (d\mu \quad d\sigma)\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}$$

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

より、

$$ds^2 = \frac{1}{\sigma^2}d\mu^2 + \frac{2}{\sigma^2}d\sigma^2$$

従って、

$$\theta^1 = \frac{1}{\sigma}d\mu, \theta^2 = \frac{\sqrt{2}}{\sigma}d\sigma$$

と表せる。第1構造式、第2構造式から計算すると、

$$\omega_2^1 = -\frac{1}{\sqrt{2}}\theta^1$$

$$K = -\frac{1}{2}$$

μ, σ空間のガウス曲率

$$ds^2 = (d\mu \quad d\sigma)\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}$$

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

より、

$$ds^2 = \frac{1}{\sigma^2}d\mu^2 + \frac{2}{\sigma^2}d\sigma^2$$

従って、

$$\theta^1 = \frac{1}{\sigma}d\mu, \theta^2 = \frac{\sqrt{2}}{\sigma}d\sigma$$

と表せる。第1構造式、第2構造式から計算すると、

$$\omega_2^1 = -\frac{1}{\sqrt{2}}\theta^1$$

$$K = -\frac{1}{2}$$

ガウス分布は負の定曲率空間になる！美しい！

ガウス分布間の距離

ガウス分布のリーマン計量は定まった

$$\begin{pmatrix} g_{\mu\mu} & g_{\mu\sigma} \\ g_{\mu\sigma} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

これを使えば、2点間の最短距離は以下のように定義できる。

$$d\big((\mu_1,\sigma_1),(\mu_2,\sigma_2)\big) = \min_{(\mu(t),\sigma(t))\in C} \int_0^1 \sqrt{\frac{1}{\sigma^2(t)}\left(\frac{\mathrm{d}\mu}{\mathrm{d}t}(t)\right)^2 + \frac{2}{\sigma^2(t)}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t)\right)^2}\,\mathrm{d}t$$

但し$C = \{\big(\mu(0),\sigma(0)\big) = (\mu_1,\sigma_1), \big(\mu(1),\sigma(1)\big) = (\mu_2,\sigma_2)$を満たす曲線の集合$\}$

最短距離は求めるのが大変。もっと便利なものは?

Kullback–Leiblerダイバージェンス

定義:
確率密度関数p, qのKLダイバージェンスを

$$D(p||q) = \int p \log \frac{p}{q} \mathrm{d}x$$

と定義する。

- **二つの確率分布間の分離度を測るもの**
- **最尤推定に使われたりする**
- $D(p||q) \geq 0$
- $D(p||q) = 0 \Leftrightarrow p = q$
- $D(p||q) \neq D(q||p)$

ガウス分布間のKLダイバージェンス

$$
\begin{aligned}
D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big) &= \int p(\mu_1,\sigma_1)\log\frac{p(\mu_1,\sigma_1)}{p(\mu_2,\sigma_2)}\,\mathrm{d}x \\
&= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2+\mu_1^2-2\mu_1\mu_2+\mu_2^2}{2\sigma_2}
\end{aligned}
$$

ガウス分布間のKLダイバージェンス

$$D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big) = \int p(\mu_1,\sigma_1)\log\frac{p(\mu_1,\sigma_1)}{p(\mu_2,\sigma_2)}\mathrm{d}x$$
$$= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2+\mu_1^2-2\mu_1\mu_2+\mu_2^2}{2\sigma_2}$$

KLダイバージェンスは微分するとリーマン計量が出てくる

$$g_{\mu\mu} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\mu_1\partial\mu_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = \frac{1}{\sigma^2}$$

$$g_{\mu\sigma} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\mu_1\partial\sigma_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = 0$$

$$g_{\sigma\sigma} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\sigma_1\partial\sigma_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = \frac{2}{\sigma^2}$$

ガウス分布間のKLダイバージェンス

$$D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big) = \int p(\mu_1,\sigma_1)\log\frac{p(\mu_1,\sigma_1)}{p(\mu_2,\sigma_2)}\mathrm{d}x$$
$$= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2+\mu_1^2-2\mu_1\mu_2+\mu_2^2}{2\sigma_2}$$

KLダイバージェンスは微分するとリーマン計量が出てくる

$$g_{\mu\mu} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\mu_1\partial\mu_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = \frac{1}{\sigma^2}$$

$$g_{\mu\sigma} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\mu_1\partial\sigma_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = 0$$

$$g_{\sigma\sigma} = \frac{\partial^2 D\big(p(\mu_1,\sigma_1)||p(\mu_2,\sigma_2)\big)}{\partial\sigma_1\partial\sigma_1}\bigg|_{\mu_1=\mu_2=\mu,\sigma_1=\sigma_2=\sigma} = \frac{2}{\sigma^2}$$

KLダイバージェンスは、リーマン計量の情報を含んでいる

目次

- ガウス分布の幾何構造
- 指数型分布族
- 双対平坦空間

指数型分布族

定義:
n次元モデルの確率密度関数が

$$p(x;\theta) = C(x)\exp\left(\sum_{i=1}^{n}\theta^i F_i(x) - \psi(\theta)\right) = C(x)\exp\left(\theta^i F_i(x) - \psi(\theta)\right)$$

と表せるとき、このモデルを指数型分布族という。

ガウス分布は2次元の指数型分布族

$$C(x) = 1, F_1(x) = x, F_2(x) = x^2, \theta^1 = \frac{\mu}{\sigma^2}, \theta^2 = -\frac{1}{2\sigma^2}$$

$$\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta^2}\right)$$

指数型分布族の性質

指数型分布族の期待値

$$\mathrm{E}[F_i(x)] = \frac{\partial \psi}{\partial \theta^i}(\theta)$$

指数型分布族のFisher情報行列は

$$\mathrm{E}\left[\frac{\partial \log p}{\partial \theta}\left(\frac{\partial \log p}{\partial \theta}\right)^{\mathrm{T}}\right] = \frac{\partial^2 \psi}{\partial \theta^2}(\theta)$$

Fisher情報行列は定義より非負定値行列
従って、 $\psi(\theta)$は凸関数

指数型分布族の性質

指数型分布族の期待値

$$\mathrm{E}[F_i(x)] = \frac{\partial \psi}{\partial \theta^i}(\theta)$$

指数型分布族のFisher情報行列は

$$\mathrm{E}\left[\frac{\partial \log p}{\partial \theta}\left(\frac{\partial \log p}{\partial \theta}\right)^{\mathrm{T}}\right] = \frac{\partial^2 \psi}{\partial \theta^2}(\theta)$$

Fisher情報行列は定義より非負定値行列
従って、 $\psi(\theta)$は凸関数

今後、Fisher情報行列を計量として使いたいので、
$\frac{\partial^2 \psi}{\partial \theta^2}(\theta)$ は常に正定値であると仮定する。

目次

- ガウス分布の幾何構造
- 指数型分布族
- 双対平坦空間

指数型分布族の双対座標系

指数型分布族

$$p(x;\theta) = C(x)\exp\left(\theta^i F_i(x) - \psi(\theta)\right)$$

に対して、 θを 自然パラメータまたは正準パラメータという。

$F_i(x)$ の期待値もパラメータと考えることができる。

$$\eta_i = \mathrm{E}[F_i(x)] = \frac{\partial \psi}{\partial \theta^i}(\theta)$$

これを期待値パラメータという。

θとηは以下の関係性がある。(Legendre変換)

$$\phi(\eta) \coloneqq \max_{\theta}\left(\theta^i \eta_i - \psi(\theta)\right)$$

$$\eta_i = \frac{\partial \psi}{\partial \theta^i}(\theta), \theta^i = \frac{\partial \phi}{\partial \eta_i}(\eta)$$

$$\psi(\theta) = \max_{\eta}\left(\theta^i \eta_i - \phi(\eta)\right)$$

$\phi(\eta)$も凸関数になる。

指数型分布族のKLダイバージェンス

$$
\begin{aligned}
D\big(p(\theta)||p(\theta')\big) &= \psi(\theta') - \psi(\theta) - \mathrm{E}_\theta[F_i(x)]\big(\theta'^i - \theta^i\big) \\
&= \psi(\theta') - \psi(\theta) - \frac{\partial \psi}{\partial \theta^i}(\theta)\big(\theta'^i - \theta^i\big)
\end{aligned}
$$

これは、凸関数$\psi(\theta)$ に関するBregman ダイバージェンス

$$
D_\psi(\theta'||\theta) \coloneqq \psi(\theta') - \psi(\theta) - \frac{\partial \psi}{\partial \theta^i}(\theta)\big(\theta'^i - \theta^i\big)
$$

に一致する。

$\phi(\eta)$に関するBregman ダイバージェンス

KLダイバージェンスはϕのBregman ダイバージェンスでも表せる

$$
\begin{aligned}
D\big(p(\theta)||p(\theta')\big) &= D_{\psi}(\theta'||\theta) \\
&= \psi(\theta') - \psi(\theta) - \frac{\partial \psi}{\partial \theta^i}(\theta)\big(\theta'^i - \theta^i\big) \\
&= \theta'^i \eta'_i - \phi(\eta') - \big(\theta^i \eta_i - \phi(\eta)\big) - \eta_i\big(\theta'^i - \theta^i\big) \\
&= \phi(\eta) - \phi(\eta') - \theta'^i(\eta_i - \eta'_i) \\
&= \phi(\eta) - \phi(\eta') - \frac{\partial \phi}{\partial \eta'_i}(\eta')(\eta_i - \eta'_i) \\
&= D_{\phi}(\eta||\eta')
\end{aligned}
$$

θ座標系とη座標系の直線

指数型分布族に対して、双対な2つの座標系が得られた。
それぞれの座標系における直線は以下の式を満たす。

θ座標系の直線(m-測地線)$\theta(t)$

$$\frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} = 0$$

η座標系の直線(e-測地線) $\eta(t)$

$$\frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} = 0$$

θ座標系の直線はη座標系でどう表せるか？
η座標系の直線はθ座標系でどう表せるか？

$$\frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} = 0 \Leftrightarrow \frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} + \underbrace{\frac{\partial^2\psi}{\partial\theta^i\partial\theta^l}\frac{\partial^3\phi}{\partial\eta_j\partial\eta_k\partial\eta_l}}_{\Gamma_i^{\mathrm{m}jk}}\frac{\mathrm{d}\eta_j}{\mathrm{d}t}\frac{\mathrm{d}\eta_k}{\mathrm{d}t} = 0$$

$$\frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} = 0 \Leftrightarrow \frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} + \underbrace{\frac{\partial^2\phi}{\partial\eta_i\partial\eta_l}\frac{\partial^3\psi}{\partial\theta^j\partial\theta^k\partial\theta^l}}_{\Gamma_{jk}^{\mathrm{e}i}}\frac{\mathrm{d}\theta^j}{\mathrm{d}t}\frac{\mathrm{d}\theta^k}{\mathrm{d}t} = 0$$

指数型分布族の幾何構造

指数型分布族

$$p(x;\theta) = C(x)\exp\left(\theta^i F_i(x) - \psi(\theta)\right)$$

のなす空間は次のよう構造をもつ

θ座標系で表すと,
リーマン計量

$$g_{ij} = \frac{\partial^2 \psi}{\partial\theta^i \partial\theta^j}$$

2種類の直線(測地線)

$$\frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} = 0, \frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} + \Gamma_i^{\mathrm{e}jk}\frac{\mathrm{d}\theta^j}{\mathrm{d}t}\frac{\mathrm{d}\theta^k}{\mathrm{d}t} = 0$$

η座標系で表すと
リーマン計量

$$g^{ij} = \frac{\partial^2 \phi}{\partial\eta_i \partial\eta_j}$$

2種類の直線(測地線)

$$\frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} + \Gamma_i^{\mathrm{m}jk}\frac{\mathrm{d}\eta_j}{\mathrm{d}t}\frac{\mathrm{d}\eta_k}{\mathrm{d}t} = 0, \frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} = 0$$

このような空間を双対平坦空間という。
(厳密な定義はしていない)

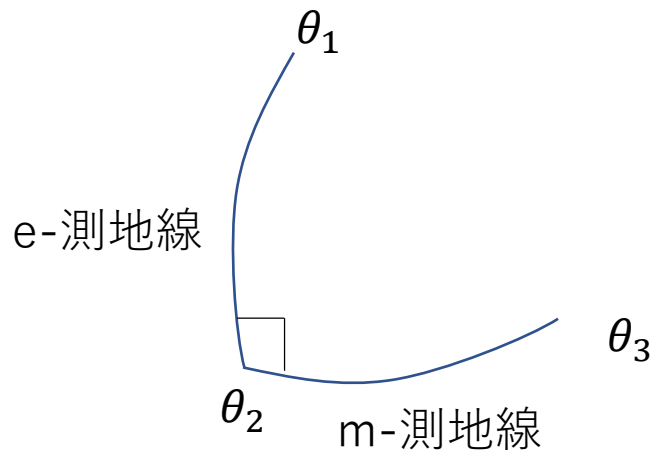
双対平坦空間上のピタゴラスの定理

指数型分布族

$$p(x;\theta) = C(x)\exp\left(\theta^i F_i(x) - \psi(\theta)\right)$$

の異なる3点$\theta_1 = \left(\theta_1^i\right), \theta_2 = \left(\theta_2^i\right), \theta_3 = \left(\theta_3^i\right)$に対して

θ_1, θ_2を結ぶe-測地線とθ_2, θ_3結ぶm-測地線はθ_2で直交する
$\Leftrightarrow D(\theta_1||\theta_3) = D(\theta_1||\theta_2) + D(\theta_2||\theta_3)$



参考文献

- Amari, S. Information Geometry and Its Applications (Applied Mathematical Sciences); Springer, 2016.
- 甘利俊一; 長岡浩司; 情報幾何の方法; 岩波書店, 1993.
- 藤原彰夫 情報幾何学の基礎; 牧野書店, 2015.
- Amari, S. Information geometry of the EM and em algorithms for neural networks; Neural Networks 1995, 8, 1379-1408