情報幾何学と統計多様体の幾何学

Information geometry and geometry of statistical manifolds

松添博（MATSUZOE Hiroshi）
名古屋工業大学大学院工学研究科　しくみ領域

前半の内容は　竹内純一氏（九州大システム情報科学）
甘利俊一氏（理化学研究所）との共同研究

Example (Bernoulli Trial)

$$\Omega = \{0, 1\}$$

$x = 1$: success event $x = 0$: failure event

η : sccess probability ($1 - \eta$: failure probability)

$$p(x; \eta) = \eta^x (1 - \eta)^{1-x} \qquad \text{the Bernoulli distribution}$$

Suppose that η is unknown.
Let us infer the parameter η from experiments (traials).

trails: 500, success events: 298 $\Longrightarrow \eta = \dfrac{298}{500} \left(\approx \dfrac{3}{5} \right)$

the maximum likelihood estimation

trail: 1, success event: 1 $\Longrightarrow \eta = 1$ $\left(\text{We may answer } \dfrac{2}{3} \text{ (?)}, \dfrac{3}{4} \text{ (?)}. \right)$

Example (Bernoulli Trial)

$$\Omega = \{0, 1\}$$

$x = 1$: success event $\quad x = 0$: failure event

η : sccess probability ($1 - \eta$: failure probability)

$$p(x; \eta) = \eta^x (1 - \eta)^{1-x} \qquad \text{the Bernoulli distribution}$$

Suppose that η is unknown.
Let us infer the parameter η from experiments (traials).

$$\text{trails: } 500, \ \text{success events: } 298 \Longrightarrow \eta = \frac{298}{500} \left(\approx \frac{3}{5}\right)$$

the maximum likelihood estimation

$$\text{trail: } 1, \ \text{success event: } 1 \Longrightarrow \eta = 1 \left(\text{We may answer } \frac{2}{3} \ (?), \ \frac{3}{4} \ (?).\right)$$

Bayesian estimations

We would like to consider why the ratios $\frac{2}{3}$ or $\frac{3}{4}$ arise
form the viewpoint of differential geometry.

1 Geometry for Statistical Models

$(\Omega, \mathcal{B}, P)$: a probability space
Ξ : an open domain of R^n (a parameter space)

Definition 1.1
S is a statistical model or a parametric model on Ω
$\overset{\text{def}}{\Longleftrightarrow}$ S is a set of probability densities with parameter $\xi \in \Xi$ such that

$$S = \left\{ p(x;\xi) \,\middle|\, \int_\Omega p(x;\xi)dx = 1,\ p(x;\xi) > 0,\ \xi \in \Xi \subset R^n \right\},$$

where $P(A) = \int_A p(x;\xi)dx,\ (A \in \mathcal{B})$.

Example 1.2 (Normal distributions) $\xi = (\mu, \sigma) \in \Xi = R^2_+$
μ: mean $(-\infty < \mu < \infty)$, σ: standard deviation $(0 < \sigma < \infty)$.

$$S = \left\{ p(x;\mu,\sigma) \,\middle|\, p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-u)^2}{2\sigma^2}\right] \right\}$$

We assume S is a smooth manifold with local coordinate system Ξ.

$g = (g_{ij})$ is the Fisher information matrix of S

$$\overset{\text{def}}{\Longleftrightarrow} \qquad g_{ij}(\xi) := \int_\Omega \frac{\partial}{\partial \xi^i} \log p(x;\xi) \frac{\partial}{\partial \xi^j} \log p(x;\xi) p(x;\xi) dx$$

$$= \int_\Omega \partial_i p_\xi \left(\frac{\partial_j p_\xi}{p_\xi} \right) dx = E_\xi[\partial_i l_\xi \partial_j l_\xi]$$

For simplicity, we used following notations:

$$E_\xi[f] = \int_\Omega f(x) p(x;\xi) dx, \qquad \text{(the expectation of } f(x) \text{ w.r.t. } p(x;\xi)\text{)},$$

$$l_\xi = l(x;\xi) = \log p(x;\xi) \qquad \text{(the information of } p(x;\xi)\text{)},$$

$$\partial_i = \frac{\partial}{\partial \xi^i}.$$

We assume that g is positive definite and $g_{ij}(\xi)$ is finite for all i, j, ξ.
$\Longrightarrow$ We can define a Riemannian metric on S.
(the Fisher metric on S)

$g = (g_{ij})$ is the Fisher information matrix of S

$$\overset{\text{def}}{\Longleftrightarrow} \qquad g_{ij}(\xi) := \int_\Omega \frac{\partial}{\partial \xi^i} \log p(x;\xi) \frac{\partial}{\partial \xi^j} \log p(x;\xi) p(x;\xi) dx$$

$$= \int_\Omega \partial_i p_\xi \left(\frac{\partial_j p_\xi}{p_\xi} \right) dx = E_\xi[\partial_i l_\xi \partial_j l_\xi]$$

Proposition 1.3

The following conditions are equivalent.

(1) g is positive definite.

(2) $\{\partial_1 p_\xi, \ldots, \partial_n p_\xi\}$ are linearly independent.

(3) $\{\partial_1 l_\xi, \ldots, \partial_n l_\xi\}$ are linearly independent.

$$\partial_i p_\xi \quad \overset{\text{def}}{\Longleftrightarrow} \quad \text{mixture representation,}$$

$$\partial_i l_\xi = \left(\frac{\partial_i p_\xi}{p_\xi} \right) \quad \overset{\text{def}}{\Longleftrightarrow} \quad \text{exponential representation.}$$

For $\alpha \in R$, we define the α-connection $\nabla^{(\alpha)}$ by the following formula:

$$\Gamma^{(\alpha)}_{ij,k}(\xi) = E_\xi\left[\left(\partial_i\partial_j l_\xi + \frac{1-\alpha}{2}\partial_i l_\xi \partial_j l_\xi\right)(\partial_k l_\xi)\right]$$
$$g(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(\alpha)}_{ij,k}$$

We can check that $\nabla^{(\alpha)}$ $({}^\forall\alpha \in R)$ is torsion-free and $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metricis. On the other hand,

$\nabla^{(1)}$: the exponential connection
$\nabla^{(-1)}$: the mixture connection

Exponential connections and mixture connections are very useful in geometric theory of statistical inferences.

For $\alpha \in \boldsymbol{R}$, we define the α-connection $\nabla^{(\alpha)}$ by the following formula:

$$\Gamma^{(\alpha)}_{ij,k}(\xi) = E_\xi\left[\left(\partial_i\partial_j l_\xi + \frac{1-\alpha}{2}\partial_i l_\xi \partial_j l_\xi\right)(\partial_k l_\xi)\right]$$

$$g(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(\alpha)}_{ij,k}$$

We can check that $\nabla^{(\alpha)}$ $({}^\forall\alpha \in \boldsymbol{R})$ is torsion-free and $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metricis. On the other hand,

$\nabla^{(1)}$: the exponential connection
$\nabla^{(-1)}$: the mixture connection

(1) $Xg(Y,Z) = g(\nabla^{(\alpha)}_X Y, Z) + g(Y, \nabla^{(-\alpha)}_X Z)$

$\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are called dual (or conjugate) with respect to g

(2) $g(\nabla^{(\alpha)}_X Y, Z) = g(\nabla^{(0)}_X Y, Z) - \frac{\alpha}{2}T(X,Y,Z)$

$T_\xi(X,Y,Z) := E_\xi[(Xl_\xi)(Yl_\xi)(Zl_\xi)]$

the skewness or the cubic form.

(3) $(\nabla^{(\alpha)}_X g)(Y,Z) = (\nabla^{(\alpha)}_Y g)(X,Z) = \alpha T(X,Y,Z)$

A statistical model S is an exponential family

$\overset{\text{def}}{\Longleftrightarrow}$

$$S = \left\{ p(x;\theta) \,\middle|\, p(x;\theta) = \exp[C(x) + \theta^i F_i(x) - \psi(\theta)] \right\},$$

where $\theta^i F_i(x) = \sum_{i=1}^{n} \theta^i F_i(x)$ (Einstein's convention) and

$C, F_1, \cdots, F_n$: random variables on Ω

ψ : a function on the parameter space Θ

The coordinate system $[\theta^i]$ is called the natural parameters.

Proposition 1.4
For an exponential family,
(1) $\nabla^{(1)}$ is flat
(2) $[\theta^i]$ is an affine coordinate, i.e., $\Gamma_{ij}^{(1)\,k} \equiv 0$

Proof:

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = E_\theta\left[\left(\partial_i \partial_j l_\theta + \frac{1-\alpha}{2}\partial_i l_\theta \partial_j l_\theta\right)(\partial_k l_\theta)\right]$$

A statistical model S is an exponential family

$\overset{\text{def}}{\Longleftrightarrow}$

$$S = \left\{ p(x;\theta) \,\middle|\, p(x;\theta) = \exp[C(x) + \theta^i F_i(x) - \psi(\theta)] \right\},$$

where $\theta^i F_i(x) = \sum_{i=1}^{n} \theta^i F_i(x)$ (Einstein's convention) and

$C, F_1, \cdots, F_n$: random variables on Ω

ψ : a function on the parameter space Θ

The coordinate system $[\theta^i]$ is called the natural parameters.

For simplicity, assume that $C = 0$.

Definition 1.5

M is a curved exponential family of S

$\overset{\text{def}}{\Longleftrightarrow}$ M is a submanifold of S such that

$$M = \{p(x;\theta(u)) \,| p(x;\theta(u)) \in S \; u \in U \subset R^m\}$$

Normal distributions

$\Omega = R,\ n = 2,\ \xi = (\mu, \sigma) \in R^2_+$ (the upper half plane).

$$S = \left\{ p(x;\mu,\sigma) \;\middle|\; p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-u)^2}{2\sigma^2} \right] \right\}$$

The Fisher metirc is

$$(g_{ij}) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \left(S \text{ is a space of constant negative curvature } -\frac{1}{2} \right).$$

$\nabla^{(1)}$ and $\nabla^{(-1)}$ are flat affine connections. In addition,

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

$$\Longrightarrow p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-u)^2}{2\sigma^2} \right] = \exp\left[x\theta^1 + (x)^2\theta^2 - \psi(\theta) \right]$$

$\{\theta^1, \theta^2\}$: natural parameters. ($\nabla^{(1)}$-geodesic coordinate system)

$$\eta_1 = E[x] = \mu, \qquad \eta_2 = E\left[x^2\right] = \sigma^2 + \mu^2.$$

$\{\eta_1, \eta_2\}$: moment parameters. ($\nabla^{(-1)}$-geodesic coordinate system)

Finite sample space

$$\Omega = \{x_0, x_1, \cdots, x_n\}, \ \dim S = n$$

$$p(x_i;\eta) = \begin{cases} \eta_i & (1 \le i \le n) \\ 1 - \sum_{j=1}^n \eta_j & (i = 0) \end{cases}$$

$$\Xi = \left\{ \{\eta_1, \cdots, \eta_n\} \,\middle|\, \eta_i > 0 \ ({}^\forall i), \ \sum_{j=1}^n \eta_j < 1 \right\}$$

(an n-dimensional simplex)

The Fisher metirc is

$$(g_{ij}) = \frac{1}{\eta_0} \begin{pmatrix} 1 + \frac{\eta_0}{\eta_1} & 1 & \cdots & 1 \\ 1 & 1 + \frac{\eta_0}{\eta_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & \cdots & 1 + \frac{\eta_0}{\eta_n} \end{pmatrix},$$

where $\eta_0 = 1 - \sum_{j=1}^n \eta_j$.

$\left(S \text{ is a space of constant positive curvature } \frac{1}{4} \right)$.

Finite sample space

$$\Omega = \{x_0, x_1, \cdots, x_n\}, \ \dim S = n$$

$$p(x_i; \eta) = \begin{cases} \eta_i & (1 \le i \le n) \\ 1 - \sum_{j=1}^n \eta_j & (i = 0) \end{cases}$$

$$\Xi = \left\{ \{\eta_1, \cdots, \eta_n\} \,\middle|\, \eta_i > 0 \ ({}^\forall i), \ \sum_{j=1}^n \eta_j < 1 \right\}$$

(an n-dimensional simplex)

$\{\theta^1, \cdots, \theta^n\}$: natural parameters. ($\nabla^{(1)}$-geodesic coordinate system)

where $$\theta^i = \log \frac{\eta_i}{1 - \sum_{j=1}^n \eta_j} = \log \frac{p(x_i)}{p(x_0)}.$$

$\{\eta_1, \cdots, \eta_n\}$: moment parameters. ($\nabla^{(-1)}$-geodesic coordinate system)

Bernoulli distributions

$\Omega = \{0, 1\}, n = 1, \xi = \eta.$

$$C(x) = 0, \quad F(x) = x, \quad \theta = \log \frac{\eta}{1-\eta},$$
$$\psi(\theta) = -\log(1-\eta) \;=\; \log(1+e^{\theta})$$

Then we obtain

$$\begin{aligned} p(x;\xi) \;&=\; \eta^{x}(1-\eta)^{1-x} \;=\; \exp\left[\log \eta^{x}(1-\eta)^{1-x}\right] \\ &=\; \exp\left[x\theta - \psi(\theta)\right]. \end{aligned}$$

This implies that Bernoulli distributions are an exponential family.
The expectation parameter is:

$$E[x] \;=\; 1 \cdot \eta + 0 \cdot (1-\eta) \;=\; \eta$$

The Fisher metric is

$$g(\eta) = \frac{1}{\eta(1-\eta)}$$

2 Statistical inference for curved exponential families

S : an exponential family
M : a curved exponential family embedded into S
$x_1, \cdots, x_N$: N independent observations of the random variable x distributed to $p(x; u) \in M$

Given $x^N = (x_1, \cdots, x_N)$, a function L on U can be defined by

$$\begin{aligned} L(u) &= p(x_1; u) \cdots p(x_N; u) \\ &= \prod_{i=1}^{n} p(x_i; u) \\ &= p(x^N; u) \end{aligned}$$

We call L a likelihood function.

We say that a statistic is the maximum likelihood estimator if it maximizes the likelihood function:

$$\hat{u} = \arg\max_{u \in U} L(u), \qquad \left(L(\hat{u}) = \max_{u \in U} L(u) \right)$$

Suppose that $p(x;\theta), p(x;\theta') \in S$.

KL : the Kullback-Leibler divergence
(or the relative entropy) of S

$\overset{\text{def}}{\Longleftrightarrow}$ KL is a function on $S \times S$ such that

$$KL(p(\theta)||p(\theta')) = \int_\Omega \log \frac{p(\theta)}{p(\theta')} p(\theta) dx.$$

$$\begin{aligned}
\bar{x} &= \frac{1}{N}\sum x_i \qquad \text{(the sample mean of } x^N\text{)} \\
\hat{\eta}_i &= \frac{1}{N}\sum_{j=1}^{N} F_i(x_j) \qquad \text{(the sample mean of the random variable } F_i.\text{)} \\
\phi(\theta) &= E_\theta[\log p(\theta)] \qquad (-\phi(\theta) \text{ is the entropy of } p(\theta))
\end{aligned}$$

Then the Kullback-Leibler divergence is given by

$$KL(p(\hat{\eta})||p(u)) = \phi(\hat{\eta}) - \frac{1}{N}\log L(u).$$

The mixmum likelihood estimation $\hat{u}$ is the point in M which minimizes the divergence from $p(\hat{\eta})$.

KL-divergence (statistically)

the Kullback-Leibler divergence

$$\begin{aligned} KL(p(\theta)||p(\theta')) &= \int_\Omega \log \frac{p(\theta)}{p(\theta')} p(\theta) dx \\ &= \int_\Omega (\log p(\theta) - \log p(\theta'))\, p(\theta) dx \end{aligned}$$

The Kullback-Leibler divergence mesures the difference of the mean of informations from $\log p(\theta)$ to $\log p(\theta')$.

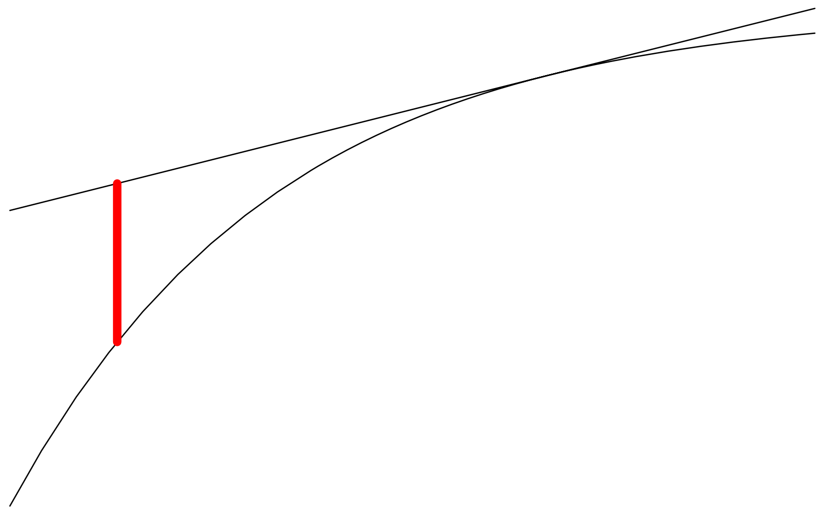
KL-divergence (geometrically)

Suppose that M is an exponential family.

$$\begin{aligned} \phi(\theta) &= E_\theta[\log p(\theta)] \quad (-\phi(\theta) \text{ is the entropy of } p(\theta)) \\ l_\theta &: \text{ the tangent hyperplane of } \phi \text{ at } \theta \end{aligned}$$

$$KL(p(\theta)||p(\theta')) = l_\theta(\theta') - \phi(\theta')$$

The Kullback-Leibler divergence mesures the difference of the hight between $l_\theta(\theta')$ and $\phi(\theta')$.



The KL-divergence $KL(p||q)$ is equivalent to the difference between $\phi(\theta')$ and $l_\theta(\theta')$.

This imples the KL-divergence is contained in the class of Bragman divergences (or canonical divergences).

KL-divergence (geometrically)

Suppose that M is an exponential family.

$$\phi(\theta) = E_\theta[\log p(\theta)] \quad (-\phi(\theta) \text{ is the entropy of } p(\theta))$$

$$l_\theta \quad : \quad \text{the tangent hyperplane of } \phi \text{ at } \theta$$

$$KL(p(\theta)||p(\theta')) = l_\theta(\theta') - \phi(\theta')$$

The Kullback-Leibler divergence mesures the difference of the hight between $l_\theta(\theta')$ and $\phi(\theta')$.

3 Bayesian inference of curved exponential families

S : an exponential family
M : a curved exponential family embedded into S
$p(x;\theta(u))$: the model distribution which generates data
$\rho(u)du$: a prior distribution

e.g. $\tilde{\rho}^{(0)}$: the Jeffreys prior of M.

$$\overset{\text{def}}{\Longleftrightarrow} \quad \tilde{\rho}^{(0)} = \frac{(\det|g_{ab}|)^{1/2}}{\int_U (\det|g_{ab}|)^{1/2} du} du \qquad g : \text{the Fisher metric of } M.$$

We define the posterior distribution by

$$\rho'(u|x) = \frac{p(x;u)\rho(u)}{\int_U p(x;u)\rho(u)du}.$$

x^N : N observations obtained from $p(x;\theta(u))$.
We define the Bayesian mixture distribution by

$$f_\rho[x^N](x) = \int_U p(x;u)\rho'(u|x^N)du$$

Let us consider the projection from $\boldsymbol{f_\rho[x^N](x)}$ to $\boldsymbol{M}$ with respect to the Kullback-Leibler divergence:

$$\boldsymbol{u\left(\tilde{f}_\rho[x^N]\right) = \arg\min_{u \in U} KL\left(f_\rho[x^N] || p(x^N; u)\right).}$$

$\boldsymbol{u\left(\tilde{f}_\rho[x^N]\right)}$ **: the projected Bayesian estimation.**

Example (Bernoulli Trial)

$$\Omega = \{0, 1\}$$

$$p(x;\eta) = \eta^x (1-\eta)^{1-x}$$

$$\eta \ : \ \text{an expectation parameter}$$

$$\theta = \log \frac{\eta}{1-\eta} \quad \text{a natural parameter}$$

$$g(\eta) = \frac{1}{\eta(1-\eta)} \quad \text{the Fisher information with respect to } \eta$$

priors	$d\theta$	Jeffreys	$d\eta$
density $\rho(\eta)$ w.r.t. $d\eta$	$\frac{d\theta}{d\eta} = \frac{1}{\eta(1-\eta)}$	$\frac{1}{\sqrt{\eta(1-\eta)}}$	1

where $d\theta$ and $d\eta$ are uniform priors with respect to θ and η, respectively.

α-parallel priors

Recall the Bayes formula:

$$\rho'(u|x) = \frac{p(x;u)\rho(u)}{\int_U p(x;u)\rho(u)du}$$

The integral is carried out on the parameter space
$\Longrightarrow$ A prior distribution can be regarded as a volume element on M.

M : a statistical model
g : the Fisher metric on M
$\nabla^{(0)}$: the Levi-Civita connection with respect to g
$\tilde{\omega}^0$: the Jeffreys prior distribution

Proposition 3.1 $\nabla^{(0)}\tilde{\omega}^0 = 0$

Definition 3.2
$\tilde{\omega}^{(\alpha)}$ is an α-(parallel) prior $\overset{\text{def}}{\Longleftrightarrow}$ $\nabla^{(\alpha)}\tilde{\omega}^{(\alpha)} = 0$

For an exponential family
$d\theta \leftrightarrow$ 1-parallel prior $d\eta \leftrightarrow -1$-parallel prior

Example (Bernoulli Trial)

$$\Omega = \{0,1\}, \quad p(x;\eta) = \eta^x(1-\eta)^{1-x}$$

$$\eta \ : \ \text{an expectation parameter}$$

$$\theta = \log\frac{\eta}{1-\eta} \quad \text{a natural parameter}$$

$$g(\eta) = \frac{1}{\eta(1-\eta)} \quad \text{the Fisher information with respect to } \eta$$

priors	$d\theta$	Jeffreys	$d\eta$
density $\rho(\eta)$ w.r.t. $d\eta$	$\frac{d\theta}{d\eta} = \frac{1}{\eta(1-\eta)}$	$\frac{1}{\sqrt{\eta(1-\eta)}}$	1

Experiment $N = 1$, success event $k = 1$

	$d\theta$	Jeffreys	$d\eta$
the projected Bayes estimator	1	$\frac{3}{4}$	$\frac{2}{3}$

General case

the projected Bayes estimator	$\frac{k}{N}$	$\frac{k+\frac{1}{2}}{N+1}$	$\frac{k+1}{N+2}$

4 Statistical manifolds and equiaffine structures

(M, g) : a Riemannian manifold
∇ : a torsion-free affine connection on M
i.e. $T^{\nabla}(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y] \equiv 0$

Definition 4.1
We call the triplet (M, ∇, g) a statistical manifold
$\overset{\text{def}}{\Longleftrightarrow}$ ∇g is totally symmetric.

Definition 4.2
∇^*: dual (or conjugate) connection of ∇ with respect to g by

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

(1) $(\nabla^*)^* = \nabla$

(2) Set $\nabla^{(0)} = \frac{1}{2}(\nabla + \nabla^*) \implies \nabla^{(0)} g = 0$

(3) (M, ∇, g) : statistical manifold $\Longleftrightarrow$ (M, ∇^*, g) : statistical manifold
(M, ∇^*, g) : the dual statistical manifold

Proposition 4.3

(M, g) : *Riemannian manifold with the Levi-Civita connection* $\nabla^{(0)}$
C : *totally symmetric* $(0,3)$*-tensor field*

$$g(\nabla_X Y, Z) := g(\nabla^{(0)}_X Y, Z) - \frac{1}{2}C(X, Y, Z),$$
$$g(\nabla^*_X Y, Z) := g(\nabla^{(0)}_X Y, Z) + \frac{1}{2}C(X, Y, Z),$$

$\Longrightarrow$ (1) ∇ *and* ∇^* *are torsion-free dual affine connections.*
(2) ∇g *and* $\nabla^* g$ *are totally symmetric.*

Proposition 4.4

If we assume two conditions from the followings, then the others hold.

(1) ∇ is torsion-free.

(2) ∇^* is torsion-free.

(3) $C = \nabla g$ is totally symmetric.

(4) $\nabla^{(0)} = (\nabla + \nabla^*)/2$ is the Levi-Civita connection with respect to g.

Proposition 4.3
(M, g) : *Riemannian manifold with the Levi-Civita connection* $\nabla^{(0)}$
C : *totally symmetric* $(0, 3)$*-tensor field*

$$g(\nabla_X Y, Z) := g(\nabla^{(0)}_X Y, Z) - \frac{1}{2}C(X, Y, Z),$$
$$g(\nabla^*_X Y, Z) := g(\nabla^{(0)}_X Y, Z) + \frac{1}{2}C(X, Y, Z),$$

$\Longrightarrow$ (1) ∇ *and* ∇^* *are torsion-free dual affine connections.*
(2) ∇g *and* $\nabla^* g$ *are totally symmetric.*

Definition 4.1 (Kurose)
We call the triplet (M, ∇, g) a statistical manifold
$\overset{\text{def}}{\Longleftrightarrow}$ ∇g is totally symmetric.

Definition 4.5 (Lauritzen)
(M, g) : a Riemannian manifold
C : a totally symmetric $(0, 3)$-tensor field
We call the triplet (M, g, C) a statistical manifold.

Parametric statistical model

$(\Omega, \mathcal{B}, P)$: a probability space
Ξ : an open domain of R^n (a parameter space)

S is a statistical model or a parametric model on Ω
$\overset{\text{def}}{\Longleftrightarrow}$ S is a set of probability densities with parameter $\xi \in \Xi$ such that

$$S = \left\{ p(x;\xi) \,\middle|\, \int_\Omega p(x;\xi)dx = 1, p(x;\xi) > 0, \xi \in \Xi \subset R^n \right\},$$

where $P(A) = \int_A p(x;\xi)dx$, $(A \in \mathcal{B})$.

$g = (g_{ij})$ is the Fisher information matrix of S

$$\overset{\text{def}}{\Longleftrightarrow} \quad g_{ij}(\xi) := \int_\Omega \frac{\partial}{\partial \xi^i} \log p(x;\xi) \frac{\partial}{\partial \xi^j} \log p(x;\xi) p(x;\xi) dx \; (= E_\xi[\partial_i l_\xi \partial_j l_\xi])$$

We assume that g is positive definite and $g_{ij}(\xi)$ is finite for all i, j, ξ.
$\Longrightarrow$ We can define a Riemannian metric on S. (the Fisher metric)

For $\alpha \in R$, the α-connection $\nabla^{(\alpha)}$ on S

$$\overset{\text{def}}{\Longleftrightarrow} \quad \Gamma^{(\alpha)}_{ij,k}(\xi) = E_\xi\left[\left(\partial_i\partial_j l_\xi + \frac{1-\alpha}{2}\partial_i l_\xi \partial_j l_\xi\right)(\partial_k l_\xi)\right]$$

$$g(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(\alpha)}_{ij,k}$$

We can check that $\nabla^{(\alpha)}$ ($^\forall \alpha \in R$) is torsion-free and $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metrics. On the other hand,

$\nabla^{(1)}$: the exponential connection
$\nabla^{(-1)}$: the mixture connection

(1) $Xg(Y,Z) = g(\nabla^{(\alpha)}_X Y, Z) + g(Y, \nabla^{(-\alpha)}_X Z)$

$\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are called dual (or conjugate) with respect to g

(2) $g(\nabla^{(\alpha)}_X Y, Z) = g(\nabla^{(0)}_X Y, Z) - \dfrac{\alpha}{2}T(X,Y,Z)$

$T_\xi(X,Y,Z) := E_\xi[(Xl_\xi)(Yl_\xi)(Zl_\xi)]$: skewness, or cubic form.

(3) $(\nabla^{(\alpha)}_X g)(Y,Z) = (\nabla^{(\alpha)}_Y g)(X,Z) = \alpha T(X,Y,Z)$

(M, g) : a Riemannian manifold
$\nabla^{(0)}$: the Levi-Civita connection with respect to g
C : a totally symmetric $(0, 3)$-tensor field on M
For fixed $\alpha \in R$, an α-connection is defined by

$$g(\nabla^{(\alpha)}_X Y, Z) := g(\nabla^{(0)}_X Y, Z) - \frac{\alpha}{2} C(X, Y, Z)$$

(1) $\nabla^{(\alpha)}, \nabla^{(-\alpha)}$ are mutullay dual torsion-free affine connections

$$Xg(Y, Z) = g(\nabla^{(\alpha)}_X Y, Z) + g(Y, \nabla^{(-\alpha)}_X Z).$$

(2) $\alpha \in R \implies (\nabla^{(\alpha)}_X g)(Y, Z) = \alpha C(X, Y, Z)$
(3) $(M, \nabla^{(\alpha)}, g)$ is a statistical manifold.

Definition 4.6
(M, ∇, g) :,a statistical manifold,
T : the Tchebychev form, and ${}^{\#}T$: the Tchebychev vector field

$$\overset{\text{def}}{\iff} \quad \begin{aligned} T(X) &:= \mathrm{trace}_g\{(Y, Z) \mapsto C(X, Y, Z)\}, \\ g({}^{\#}T, X) &:= T(X) \end{aligned}$$

M : an n-dimensional manifold
∇ : a torsion-free affine connection on M
ω : a volume element of M,

Definition 4.7 $\{\nabla, \omega\}$ is (locally) equiaffine structure on M.

$$\overset{\text{def}}{\Longleftrightarrow} \quad \nabla\omega = 0$$

∇ is called a (locally) equiaffine connection,
ω is called a parallel volume element.

Proposition 4.8
(M, ∇, g) : a statistical manifold
T : the Tchebychev form
Then $\quad \nabla$ *is an equiaffine connection* $\quad \Longleftrightarrow \quad dT = 0$

$\nabla^{(\alpha)}$ is equiaffine
$\Longrightarrow \quad dT = 0 \Longrightarrow$ there exists a function ϕ on M such that $T = d\phi$.
Hence $g(^{\#}T, X) = X\phi$
The Tchebychev vector field is a gradient vector field of some function ϕ on M.

Proposition 4.9

(M, g, C) : a statitical manifold

$\nabla^{(\alpha)}, \nabla^{(-\alpha)}$: affine connectoins determined by g, C

$T = d\phi$: the Tchebychev form on (M, g, C)

Then

$\{\nabla^{(\alpha)}, \omega\}$ is an equiaffine structure

$\Longleftrightarrow$ $\{\nabla^{(-\alpha)}, e^{-\alpha\phi}\omega\}$ is an equiaffine structure.

Theorem 4.10

$\hat{u}$: the maximum likelihood estimator (MLE)

$\hat{g}$: the Fisher metric with respect to MLE

$\hat{C}$: the skewness tensor with respect to MLE

$u(\tilde{f}^{(\alpha)}[x^N])$: the projected Bayesian estimator with α-paralell propr

$$\Longrightarrow \qquad u^c(\tilde{f}^{(\alpha)}[x^N]) = \hat{u}^c + \frac{1-\alpha}{2N}\hat{C}_{abd}\hat{g}^{ab}\hat{g}^{cd} + o\left(\frac{1}{N}\right)$$

$$= \hat{u}^c + \frac{1-\alpha}{2N}{}^{\#}\hat{T}^c + o\left(\frac{1}{N}\right)$$

5 Divergences for non-integrable systems

5.1 Dually flat spaces

(M, g) : a Riemannian manifold
∇ : a torsion-free affine connection on M

We call the triplet (M, ∇, g) a statistical manifold
$\overset{\text{def}}{\Longleftrightarrow}$ ∇g is totally symmetric.

∇^*: dual (or conjugate) connection of ∇ with respect to g by

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

∇ is flat $\Longleftrightarrow$ ∇^* is flat.

(M, g, ∇, ∇^*) : dually flat space $\overset{\text{def}}{\Longleftrightarrow}$ ∇, ∇^* are flat affine connections.

An affine connection ∇ is flat
$\Longrightarrow$ there exists a local coordinate system on M such that

$$\Gamma^{\nabla}_{ij}{}^{k} \equiv 0.$$

We call such a coordinate system an affine coordinate system.

Proposition 5.1
(M, g, ∇, ∇^*) : ***dually flat space***
$\{\theta^i\}$: ***∇-affine coordinate system***
$\Longrightarrow$ ***there exits an ∇^*-affine coordinate system $\{\eta_i\}$ such that***

$$g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\eta_j}\right) = \delta_i^j.$$

$\{\eta_i\}$: the dual coordinate system with respect to $\{\theta^i\}$.

Proposition 5.2
(M, g, ∇, ∇^*) : a dually flat space
$\{\theta^i\}$: a ∇-affine coordinate system
$\{\eta_i\}$: the dual coordinate system of $\{\theta^i\}$
$\Longrightarrow$ there exists functions ψ, ϕ on M such that

$$\frac{\partial \psi}{\partial \theta^i} = \eta_i, \quad \frac{\partial \phi}{\partial \eta_i} = \theta^i, \quad \psi(p) + \phi(p) - \sum_{i=1}^{m} \theta^i(p)\eta_i(p) = 0. \tag{1}$$

In addition, the following formulas hold

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}, \quad g^{ij} = \frac{\partial^2 \phi}{\partial \eta_i \partial \eta_j}, \tag{2}$$

where
(g_{ij}) : the component matrix of a Riemannian metric g,
(g^{ij}) : the inverse matrix of (g_{ij})

ψ is the θ-potential function, and ϕ is the η-potential function.
We say that the relation (1) the Legendre transformation.

Definition 5.3
$\rho : M \times M \to R$: (canonical) divergence on (M, g, ∇, ∇^*)

$$\overset{\text{def}}{\Longleftrightarrow} \qquad \rho(p||q) \ := \ \psi(p) + \phi(q) - \sum_{i=1}^{n} \theta^i(p)\eta_i(q), \quad (p, q \in M).$$

Proposition 5.4
The definition of ρ is independent of choice of affine coordinate system on M.

Example 5.5 (Euclidean space)
R^m : Euclidean space
$\langle \ , \ \rangle$: the standard inner product
D : the standard flat affine connection.
$\Longrightarrow \qquad (M, \langle \ , \ \rangle, D, D)$ is a dually flat space,

$$\rho(p||q) = \frac{1}{2} d(p, q)^2.$$

KL-divergence

the Kullback-Leibler divergence

$$\begin{aligned} KL(p(\theta)||p(\theta')) &= \int_\Omega \log \frac{p(\theta)}{p(\theta')} p(\theta) dx \\ &= \int_\Omega \left(\log p(\theta) - \log p(\theta')\right) p(\theta) dx \end{aligned}$$

Suppose that S is an exponential family and $p(\theta), p(\theta') \in S$.

$$\begin{aligned} KL(p(\theta)||p(\theta')) &= \int_\Omega \left(\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta) - \sum_{i=1}^n {\theta'}^i F_i(x) + \psi(\theta') \right) p(\theta) dx \\ &= \psi(\theta') - \psi(\theta) + \sum_{i=1}^n \theta^i \eta_i - \sum_{i=1}^n {\theta'}^i \eta_i \\ &= \psi(\theta') + \phi(\theta) - \sum_{i=1}^n {\theta'}^i \eta_i \\ &= \rho(p(\theta')||p(\theta)) \end{aligned}$$

(M, ∇, h) : a simply connected flat statistical manifold.
$(\implies (M, h, \nabla, \nabla^*)$ is a dually flat space.)

$\implies \exists\psi$: a function on M (potential function) such that $\dfrac{\partial^2 \psi}{\partial\theta^i \partial\theta^j} = g_{ij}$

$\implies f : M \to R^{n+1}$: an immersion ($\{f, \xi\}$ a graph immersion)

$$f : \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^n \end{pmatrix} \mapsto \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^n \\ \psi(\theta) \end{pmatrix}, \quad \xi = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$v : M \to R_{n+1}$: the conormal map of $\{f, \xi\}$,

$$v = (-\eta_1, \dots, -\eta_n, 1) \qquad \eta_i = \frac{\partial \psi}{\partial \theta^i}$$

From $\phi(q) = \sum \eta_i(q)\theta^i(q) - \psi(q)$, we define the geometric divergence by

$$\begin{aligned} \rho(p, q) &= \langle v(q), f(p) - f(q) \rangle \\ &= -\sum \eta_i(q)\theta^i(p) + \psi(p) + \sum \eta_i(q)\theta^i(q) - \psi(q) \\ &= \psi(p) + \phi(q) - \sum \eta_i(q)\theta^i(p) \ = \ \rho^C(p, q) \end{aligned}$$

The geometric divergence coincides with the canonical divergence.

(M, ∇, h) : a simply connected flat statistical manifold.
($\Longrightarrow$ (M, h, ∇, ∇^*) is a dually flat space.)

$\Longrightarrow$ $\exists\psi$: a function on M (potential function) such that $\dfrac{\partial^2\psi}{\partial\theta^i\partial\theta^j} = g_{ij}$

$\Longrightarrow$ $f : M \to R^{n+1}$: an immersion ($\{f, \xi\}$ a graph immersion)

$$f : \begin{pmatrix}\theta^1\\ \vdots \\ \theta^n\end{pmatrix} \mapsto \begin{pmatrix}\theta^1\\ \vdots \\ \theta^n \\ \psi(\theta)\end{pmatrix}, \quad \xi = \begin{pmatrix}0\\ \vdots \\ 0 \\ 1\end{pmatrix}$$

$v : M \to R_{n+1}$ is the conormal map of $\{f, \xi\}$

$$\overset{\text{def}}{\Longleftrightarrow} \quad \begin{aligned}\langle v(p), \xi_p\rangle &= 1,\\ \langle v(p), f_* X_p\rangle &= 0\end{aligned}$$

We define a function on $M \times M$ by

$$\rho(p, q) = \langle v(q), f(p) - f(q)\rangle$$

ρ is called the geometric divergence on M.

最尤法の復習

S ： an exponential family

M ： a curved exponential family embedded into S, $\quad U \subset M$

Given $(x_1, \cdots, x_N)$, the likelihood function L is defined by

$$L(u) = p(x_1; u) \cdots p(x_N; u) = \prod_{i=1}^{n} p(x_i; u)$$

$\hat{u}$: the maximum likelihood estimator $\overset{\text{def}}{\Longleftrightarrow}$ $\hat{u} = \arg\max_{u \in U} L(u)$

実際に最尤推定量を求めるためには，対数尤度方程式を解く．

$$\frac{\partial}{\partial u^1} \log L(u) = 0, \quad \ldots, \quad \frac{\partial}{\partial u^m} \log L(u) = 0$$

$$\hat{\eta}_i = \frac{1}{N} \sum_{j=1}^{N} F_i(x_j) \quad (F_i \text{ の標本平均}), \quad \phi(\theta) = E_\theta[\log p(\theta)]$$

$$KL(p(\hat{\eta}) || p(u)) = \phi(\hat{\eta}) - \frac{1}{N} \log L(u)$$

すなわち $\quad \hat{u}$: **ある推定量** $\Longleftrightarrow$ $d(KL_{\hat{\eta}})(X_u) = 0 \ ({}^\forall X_u \in T_u M)$

5.2 Divergences for non-integrable systems

$\omega : \Gamma(TM) \to \mathbf{R}^{n+1}$: a $\mathbf{R}^{n+1}$-valued 1-form
$\xi : M \to \mathbf{R}^{n+1}$: a $\mathbf{R}^{n+1}$-valued function

Definition 5.6
$\{\omega, \xi\}$ is called an affine distribution
$\overset{\text{def}}{\Longleftrightarrow}$ For an arbitrary point $p \in M$,

$$\begin{aligned} &(1) \quad \mathbf{R}^{n+1} = \text{Image } \omega_p \oplus \mathbf{R}\{\xi_x\} \\ &(2) \quad \text{Image } (d\omega)_p \subset \text{Image } \omega_p \end{aligned}$$

$$\begin{aligned} X\omega(Y) &= \omega(\nabla_X Y) + h(X, Y)\xi, \\ X\xi &= -\omega(SX) + \tau(X)\xi. \end{aligned}$$

ω : non-degenerate $\overset{\text{def}}{\Longleftrightarrow}$ h : non-degenerate
$\{\omega, \xi\}$: equiaffine $\overset{\text{def}}{\Longleftrightarrow}$ $\tau = 0$

Remark 5.7 Image $(d\omega)_p \subset$ Image ω_p $\Longleftrightarrow$ h*: symmetric.*

SLD Fisher metrics

Herm(d) : the set of all Hermitian matrices of degree d.
$\mathcal{S}$: a space of quantum states

$$\mathcal{S} = \{P \in \mathrm{Herm}(d) \mid P > 0,\ \mathrm{trace}P = 1\}$$

$T_P\mathcal{S} \cong \mathcal{A}_0 \qquad \mathcal{A}_0 = \{X \in \mathrm{Herm}(d) \mid \mathrm{trace}X = 0\}$

We denote by $\widetilde{X}$ the corresponding vector field of X.

For $P \in \mathcal{S}$, $X \in \mathcal{A}_0$, define $\omega_P(\widetilde{X})$ ($\in \mathrm{Herm}(d)$) and ξ by

$$X = \frac{1}{2}(P\omega_P(\widetilde{X}) + \omega_P(\widetilde{X})P), \qquad \xi = -I_d$$

Then $\{\omega, \xi\}$ is an equiaffine distribuion.
($\omega_P(\widetilde{X})$ is the symmetric logarithmic derivative of X !)

The induced quantities are given by

$$h_P(\widetilde{X}, \widetilde{Y}) = \frac{1}{2}\mathrm{trace}\left(P(\omega_P(\widetilde{X})\omega_P(\widetilde{Y}) + \omega_P(\widetilde{Y})\omega_P(\widetilde{X}))\right),$$

$$\nabla_{\widetilde{X}}\widetilde{Y} = h_P(\widetilde{X}, \widetilde{Y})P - \frac{1}{2}(X\omega_P(\widetilde{Y}) + \omega_P(\widetilde{Y})X).$$

$\{\omega, \xi\}$: nondegenerate, equiaffine

$v : M \longrightarrow R_{n+1}$ is the conormal map of $\{\omega, \xi\}$

$$\overset{\text{def}}{\Longleftrightarrow} \quad \begin{aligned} \langle v(p), \xi_p \rangle &= 1, \\ \langle v(p), \omega(X_p) \rangle &= 0 \end{aligned}$$

We define a function on $\Gamma(TM) \times M$ by

$$\rho(X, q) = \langle v(q), \omega(X) \rangle.$$

ρ is called the geometric pre-divergence on M.

	Information geometry	Differential geometry
θ	natural papramelers	∇-affine coordinates
	exponential arc	∇-geodesic
η	mixture papramelers expectation parameters	∇^*-affine coordinates
g or h	Fisher metric	Riemannian metric affine fundamental form
T or C	skewness	cubic form
ψ	cumulant generating function free energy	affine hypersurface
ϕ	entropy	dual map
	Legendre transformation	dual transformation
$D, \rho, \cdots$	Kullback-Leibler divergence relative entropy	geometric divergence affine support function
$\rho, \omega, \cdots$	prior distribution	volume form
$T, \cdots$	bias-correction	Tchebychev vector
	$\vdots$	$\vdots$

Final remarks

パラメータ空間に多様体構造を入れる話を考えた.（情報幾何学）

- Geometry for dually flat spaces
  - AdaBoost, U-Boost, . . .
  - Modern information theory (LDPC codes, etc.)
  - Linear programming problems
- Bayesian statistics
  - prior distribution $\Longleftrightarrow$ volume form
- Infinite dimensional case (Orlicz space geometry)
  - affine immersion into an functional space
- Quantum version of information geometry
  - statistical manifold admitting torsion
- **標本空間に多様体構造を入れる話は，情報幾何学とよばれていない**
  - kernel method