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

0 簡単な問題設定
I 統計モデル
II 統計的推論
III ベイズ推論
IV 統計多様体と等積構造
V 不可積分系のダイバージェンス

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)

\[ \Omega = \{0, 1\} \]
\[ x = 1 : \text{success event} \quad x = 0 : \text{failure event} \]
\[ \eta : \text{success probability} \quad (1 - \eta : \text{failure probability}) \]
\[ p(x; \eta) = \eta^x(1 - \eta)^{1-x} \quad \text{the Bernoulli distribution} \]

Suppose that \( \eta \) is unknown.
Let us infer the parameter \( \eta \) from experiments (trials).

trails: 500, success events: 298 \( \implies \eta = \frac{298}{500} \approx \frac{3}{5} \)
the maximum likelihood estimation

trail: 1, success event: 1 \( \implies \eta = 1 \) (We may answer \( \frac{2}{3} \) (?), \( \frac{3}{4} \) (?).)
Example (Bernoulli Trial)

\[ \Omega = \{0, 1\} \]

- \( x = 1 \): success event
- \( x = 0 \): failure event

\( \eta \): success probability (\( 1 - \eta \): failure probability)

\[ p(x; \eta) = \eta^x (1 - \eta)^{1-x} \] the Bernoulli distribution

Suppose that \( \eta \) is unknown.
Let us infer the parameter \( \eta \) from experiments (trials).

- trails: 500, success events: 298 \( \Rightarrow \eta = \frac{298}{500} \approx \frac{3}{5} \) the maximum likelihood estimation
- trail: 1, success event: 1 \( \Rightarrow \eta = 1 \) (We may answer \( \frac{2}{3} \) (?), \( \frac{3}{4} \) (?).)

Bayesian estimations

We would like to consider why the ratios \( \frac{2}{3} \) or \( \frac{3}{4} \) arise from the viewpoint of differential geometry.
1 Geometry for Statistical Models

\((\Omega, \beta, P)\) : a probability space
\(\Xi\) : an open domain of \(R^n\) (a parameter space)

**Definition 1.1**

\(S\) is a statistical model or a parametric model on \(\Omega\)
\(\iff S\) is a set of probability densities with parameter \(\xi \in \Xi\) such that
\[
S = \left\{ p(x; \xi) \left| \int_{\Omega} p(x; \xi) \, dx = 1, \ p(x; \xi) > 0, \ \xi \in \Xi \subset R^n \right\},
\]
where \(P(A) = \int_A p(x; \xi) \, dx, \ (A \in \beta)\).

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

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

We assume \(S\) is a smooth manifold with local coordinate system \(\Xi\).
\[ g = (g_{ij}) \text{ is the Fisher information matrix of } S \]

\[ g_{ij}(\xi) \overset{\text{def}}{=} \int_{\Omega} \frac{\partial}{\partial \xi^i} \log p(x; \xi) \frac{\partial}{\partial \xi^j} \log p(x; \xi)p(x; \xi) \, dx \]

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

For simplicity, we used the following notations:

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

\[ l_\xi = l(x; \xi) = \log p(x; \xi) \quad \text{(the information of } p(x; \xi)) \],

\[ \partial_i = \frac{\partial}{\partial \xi^i}. \]

We assume that \( g \) is positive definite and \( g_{ij}(\xi) \) is finite for all \( i, j, \xi \).

\[ \implies \text{We can define a Riemannian metric on } S. \]

\[ \text{(the Fisher metric on } S) \]
\( g = (g_{ij}) \) is the **Fisher information matrix** of \( S \)

\[
g_{ij}(\xi) := \int_{\Omega} \frac{\partial}{\partial \xi_i} \log p(x; \xi) \frac{\partial}{\partial \xi_j} \log p(x; \xi) p(x; \xi) dx
\]

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

**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.

\( \partial_i p_\xi \quad \overset{\text{def}}{=} \quad \text{mixture representation,} \)

\( \partial_i l_\xi = \left( \frac{\partial_i p_\xi}{p_\xi} \right) \quad \overset{\text{def}}{=} \quad \text{exponential representation.} \)
For $\alpha \in R$, we define the $\alpha$-connection $\nabla^{(\alpha)}$ by the following formula:

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

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

We can check that $\nabla^{(\alpha)} (\forall \alpha \in R)$ is torsion-free and $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metric. 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 $\alpha \in R$, we define the $\alpha$-connection $\nabla^{(\alpha)}$ by the following formula:

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

$$g(\nabla_{\partial_i}^{(\alpha)} \partial_j, \partial_k) = \Gamma_{i,j,k}^{(\alpha)}$$

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

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

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

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

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

$T_\xi(X, Y, Z) := E_\xi[(X l_\xi)(Y l_\xi)(Z l_\xi)]$

the skewness or the cubic form.

(3) $(\nabla_X^{(\alpha)} g)(Y, Z) = (\nabla_Y^{(\alpha)} g)(X, Z) = \alpha T(X, Y, Z)$
A statistical model $S$ is an exponential family

\[ S = \{ p(x; \theta) \mid p(x; \theta) = \exp[C(x) + \theta^i F_i(x) - \psi(\theta)] \} , \]

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 $\Omega$

$\psi$ : a function on the parameter space $\Theta$

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

\[\text{Proposition 1.4}\]

For an exponential family,

1. $\nabla^{(1)}$ is flat
2. $[\theta^i]$ is an affine coordinate, i.e., $\Gamma^{(1)}_{ij,k} \equiv 0$

Proof:

\[ \Gamma^{(\alpha)}_{ij,k}(\theta) = E_{\theta} \left[ \left( \partial_i \partial_j l_\theta + \frac{1 - \alpha}{2} \partial_i l_\theta \partial_j l_\theta \right) (\partial_k l_\theta) \right] \]
A statistical model $S$ is an **exponential family**

\[
S = \{ p(x; \theta) \mid p(x; \theta) = \exp[C(x) + \theta^i F_i(x) - \psi(\theta)] \},
\]

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 $\Omega$

$\psi$ : a function on the parameter space $\Theta$

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 $S$

\[
M \text{ is a submanifold of } S \text{ such that } M = \{ p(x; \theta(u)) \mid p(x; \theta(u)) \in S, u \in U \subset \mathbb{R}^m \}\]
Normal distributions

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

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

The Fisher metric is

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

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

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

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

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

\( \eta_1 = E[x] = \mu, \ \eta_2 = E[x^2] = \sigma^2 + \mu^2. \)

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

\[ \Omega = \{ x_0, x_1, \cdots, x_n \}, \quad \text{dim } S = n \]

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

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

(an \ n\text{-dimensional simplex})

The Fisher metric is

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

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

\[ S \text{ is a space of constant positive curvature } \frac{1}{4}. \]
Finite sample space

\[ \Omega = \{ x_0, x_1, \cdots, x_n \}, \quad \text{dim } S = n \]

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

\[ \Xi = \left\{ \{ \eta_1, \cdots, \eta_n \} \mid \eta_i > 0 \text{ (\forall i)}, \sum_{j=1}^{n} \eta_j < 1 \right\} 
\text{ (an } n\text{-dimensional simplex)}
\]

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

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

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

\[ p(x; \xi) = \eta^x (1 - \eta)^{1-x} = \exp \left[ \log \eta^x (1 - \eta)^{1-x} \right] \]

\[ = \exp \left[ x\theta - \psi(\theta) \right]. \]

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

\[ g(\eta) = \frac{1}{\eta(1 - \eta)} \]
0 簡単な問題設定

I 統計モデル

II 統計的推論

III ベイズ推論

IV 統計多様体と等積構造

V 不可積分系のダイバージェンス
2 Statistical inference for curved exponential families

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

Given \( x^N = (x_1, \cdots, x_N) \), a function \( L \) on \( U \) can be defined by
\[
L(u) = p(x_1; u) \cdots p(x_N; u)
= \prod_{i=1}^{n} p(x_i; u)
= p(x^N; u)
\]

We call \( L \) a likelihood function.

We say that a statistic is the maximum likelihood estimator if it maximizes the likelihood function:
\[
\hat{u} = \arg \max_{u \in U} L(u), \quad \left( L(\hat{u}) = \max_{u \in U} L(u) \right)
\]
Suppose that \( p(x; \theta), p(x; \theta') \in S \).

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

\[
KL(p(\theta) \mid \mid p(\theta')) = \int_{\Omega} \log \frac{p(\theta)}{p(\theta')} p(\theta) \, dx.
\]

\[KL(p(\hat{\eta}) \mid \mid p(u)) = \phi(\hat{\eta}) - \frac{1}{N} \log L(u).
\]

The maximum likelihood estimation \( \hat{u} \) is the point in \( M \) which minimizes the divergence from \( p(\hat{\eta}) \).
KL-divergence (statistically)

The Kullback-Leibler divergence

\[ KL(p(\theta)||p(\theta')) = \int_{\Omega} \log \frac{p(\theta)}{p(\theta')} p(\theta) dx \]
\[ = \int_{\Omega} (\log p(\theta) - \log p(\theta')) p(\theta) dx \]

The Kullback-Leibler divergence measures 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.

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

\( l_\theta : \) the tangent hyperplane of \( \phi \) at \( \theta \)

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

The Kullback-Leibler divergence measures the difference of the height 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 implies the KL-divergence is contained in the class of Bragman divergences (or canonical divergences).

---

**KL-divergence (geometrically)**

Suppose that $M$ is an exponential family.

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

$l_\theta$: the tangent hyperplane of $\phi$ at $\theta$

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

The Kullback-Leibler divergence measures the difference of the height between $l_\theta(\theta')$ and $\phi(\theta')$. 
0 簡単な問題設定
I 統計モデル
II 統計的推論
III ベイズ推論
IV 統計多様体と等積構造
V 不可積分系のダイバージェンス
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 $M$.

$\tilde{\rho}^{(0)} \overset{\text{def}}{=} \frac{(\det|g_{ab}|)^{1/2}}{\int_U (\det|g_{ab}|)^{1/2}du}du \quad g :$ 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 $f_\rho[x^N](x)$ to $M$ with respect to the Kullback-Leibler divergence:

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

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

\[ \Omega = \{0, 1\} \]

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

\[ \eta : \text{an expectation parameter} \]

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

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

<table>
<thead>
<tr>
<th>priors</th>
<th>( d\theta )</th>
<th>Jeffreys</th>
<th>( d\eta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>density ( \rho(\eta) ) w.r.t. ( d\eta )</td>
<td>( \frac{d\theta}{d\eta} = \frac{1}{\eta(1 - \eta)} )</td>
<td>( \frac{1}{\sqrt{\eta(1 - \eta)}} )</td>
<td>1</td>
</tr>
</tbody>
</table>

where \( d\theta \) and \( d\eta \) are uniform priors with respect to \( \theta \) and \( \eta \), respectively.
**α-parallel priors**

Recall the Bayes formula:

\[ \rho'(u|x) = \frac{p(x; u) \rho(u)}{\int_{\mathcal{U}} p(x; u) \rho(u) du} \]

The integral is carried out on the parameter space

\[ \Rightarrow \text{A prior distribution can be regarded as a volume element on } M. \]

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

**Proposition 3.1** \[ \nabla^{(0)} \tilde{\omega}^0 = 0 \]

**Definition 3.2**
\[ \tilde{\omega}^{(\alpha)} \text{ is an } \alpha-(\text{parallel}) \text{ prior} \iff \nabla^{(\alpha)} \tilde{\omega}^{(\alpha)} = 0 \]

For an exponential family
\[ d\theta \leftrightarrow 1-(\text{parallel}) \text{ prior} \quad d\eta \leftrightarrow -1-(\text{parallel}) \text{ prior} \]
Example (Bernoulli Trial)

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

\( \eta \) : an expectation parameter

\[ \theta = \log \frac{\eta}{1 - \eta} \] a natural parameter

\[ g(\eta) = \frac{1}{\eta(1 - \eta)} \] the Fisher information with respect to \( \eta \)

<table>
<thead>
<tr>
<th>priors</th>
<th>( d\theta )</th>
<th>Jeffreys</th>
<th>( d\eta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>density ( \rho(\eta) ) w.r.t. ( d\eta )</td>
<td>( \frac{d\theta}{d\eta} = \frac{1}{\eta(1 - \eta)} )</td>
<td>( \frac{1}{\sqrt{\eta(1 - \eta)}} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Experiment \( N = 1 \), success event \( k = 1 \)

<table>
<thead>
<tr>
<th>the projected Bayes estimator</th>
<th>( d\theta )</th>
<th>Jeffreys</th>
<th>( d\eta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>( \frac{3}{4} )</td>
<td>( \frac{2}{3} )</td>
</tr>
</tbody>
</table>

General case

| the projected Bayes estimator | \( \frac{k}{N} \) | \( \frac{k + \frac{1}{2}}{N + 1} \) | \( \frac{k + 1}{N + 2} \) |
0 簡単な問題設定

I 統計モデル

II 統計的推論

III ベイズ推論

IV 統計多様体と等積構造

V 不可積分系のダイバージェンス
4 Statistical manifolds and equiaffine structures

\((M, g)\) : a Riemannian manifold
\(\nabla\) : 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 \((M, \nabla, g)\) a statistical manifold
\(\overset{\text{def}}{\iff} \nabla g\) is totally symmetric.

Definition 4.2
\(\nabla^*\): dual (or conjugate) connection of \(\nabla\) 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) \((M, \nabla, g)\) : statistical manifold \(\iff\) \((M, \nabla^*, g)\) : statistical manifold
\((M, \nabla^*, g)\) : 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

\[
g(\nabla_X Y, Z) := g(\nabla_X^{(0)} Y, Z) - \frac{1}{2} C(X, Y, Z),
\]

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

\(\implies\) (1) \(\nabla\) and \(\nabla^*\) are torsion-free dual affine connections.

(2) \(\nabla g\) and \(\nabla^* g\) are totally symmetric.

Proposition 4.4

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

(1) \(\nabla\) is torsion-free.

(2) \(\nabla^*\) 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

\[ g(\nabla_X Y, Z) := g(\nabla_X^{(0)} Y, Z) - \frac{1}{2}C(X, Y, Z), \]

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

\[ \Rightarrow (1) \ \nabla \text{ and } \nabla^* \text{ are torsion-free dual affine connections.} \]

\[ (2) \ \nabla g \text{ and } \nabla^* g \text{ are totally symmetric.} \]

Definition 4.1 (Kurose)

We call the triplet \((M, \nabla, g)\) a statistical manifold

\[ \def \nabla g \text{ is totally symmetric.} \]

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

\( (\Omega, \beta, P) \) : a probability space
\[ \Xi : \text{an open domain of } \mathbb{R}^n \quad (\text{a parameter space}) \]

\( S \) is a statistical model or a parametric model on \( \Omega \)
\[ \text{def} \quad S \text{ is a set of probability densities with parameter } \xi \in \Xi \text{ such that} \]
\[ S = \left\{ p(x; \xi) \left| \int_{\Omega} p(x; \xi) \, dx = 1, p(x; \xi) > 0, \xi \in \Xi \subset \mathbb{R}^n \right\} , \]
where \( P(A) = \int_A p(x; \xi) \, dx, \quad (A \in \beta) \).

\( g = (g_{ij}) \) is the Fisher information matrix of \( S \)
\[ \text{def} \quad g_{ij}(\xi) := \int_{\Omega} \frac{\partial}{\partial \xi_i} \log p(x; \xi) \frac{\partial}{\partial \xi_j} \log p(x; \xi) p(x; \xi) \, dx \quad (= E_\xi[\partial_i l_\xi \partial_j l_\xi]) \]

We assume that \( g \) is positive definite and \( g_{ij}(\xi) \) is finite for all \( i, j, \xi \).
\[ \implies \text{We can define a Riemannian metric on } S. \quad (\text{the Fisher metric}) \]
For $\alpha \in R$, the $\alpha$-connection $\nabla^{(\alpha)}$ on $S$

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

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

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

\begin{enumerate}
\item $Xg(Y, Z) = g(\nabla^{(\alpha)}_{X} Y, Z) + g(Y, \nabla^{(-\alpha)}_{X} Z)$

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

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

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

\item $(\nabla^{(\alpha)}_{X} g)(Y, Z) = (\nabla^{(\alpha)}_{Y} g)(X, Z) = \alpha T(X, Y, Z)$
\end{enumerate}
\((M, g)\) : 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 \(\alpha\)-connection is defined by

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

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

\[
Xg(Y, Z) = g(\nabla^{(\alpha)}_XY, Z) + g(Y, \nabla^{(-\alpha)}_XZ).
\]

(2) \(\alpha \in R \implies (\nabla^{(\alpha)}_Xg)(Y, Z) = \alpha C(X, Y, Z)\)

(3) \((M, \nabla^{(\alpha)}, g)\) is a statistical manifold.

**Definition 4.6**

\((M, \nabla, g)\) : a statistical manifold, 
\(T\) : the Tchebychev form, and \(#T\) : the Tchebychev vector field

\[
\leftarrow^{\text{def}} \quad T(X) := \text{trace}_g\{(Y, Z) \mapsto C(X, Y, Z)\},
\]

\[
g(#T, X) := T(X)
\]
Definition 4.7 \( \{ \nabla, \omega \} \) is **(locally) equiaffine structure** on \( M \).

\[
\iff \nabla \omega = 0
\]

\( \nabla \) is called a **(locally) equiaffine connection**, 
\( \omega \) is called a **parallel volume element**.

Proposition 4.8
\((M, \nabla, g) : a \text{ statistical manifold}
\)
\( T : the \ Tchebychev \ form \)

Then \( \nabla \) **is an equiaffine connection** \iff \( dT = 0 \)

\( \nabla^{(\alpha)} \) **is equiaffine**

\[
\implies dT = 0 \implies \text{there exists a function } \phi \text{ on } M \text{ such that } T = d\phi.
\]

Hence \( g(#T, X) = X\phi \)

The Tchebychev vector field is a gradient vector field of some function \( \phi \) on \( M \).
Proposition 4.9

\((M, g, C)\) : a statistical manifold
\(\nabla^{(\alpha)},\nabla^{(-\alpha)}\) : affine connections 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
\(\hat{\mu}\) : 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 \(\alpha\)-parallel propr

\[
\Rightarrow u^c(\tilde{f}^{(\alpha)}[x^N]) = \hat{u}^c + \frac{1 - \alpha}{2N} \hat{C}_{abd} \hat{g}^{ab} \hat{g}^{cd} + o\left(\frac{1}{N}\right)
\]
\[
= \hat{u}^c + \frac{1 - \alpha}{2N} \#\hat{T}^c + o\left(\frac{1}{N}\right)
\]
0 簡単な問題設定

I 統計モデル

II 統計的推論

III ベイズ推論

IV 統計多様体と等積構造

V 不可積分系のダイバージェンス
5 Divergences for non-integrable systems

5.1 Dually flat spaces

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

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

\(\nabla^*\) : dual (or conjugate) connection of \(\nabla\) with respect to \(g\) by
\[ Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z). \]

\(\nabla\) is flat \(\iff\) \(\nabla^*\) is flat.

\((M, g, \nabla, \nabla^*)\) : dually flat space \(\overset{\text{def}}{\iff} \nabla, \nabla^*\) are flat affine connections.
An affine connection $\nabla$ is flat
$\implies$ there exists a local coordinate system on $M$ such that
$$\Gamma^k_{ij} \equiv 0.$$ 
We call such a coordinate system an affine coordinate system.

**Proposition 5.1**

$(M, g, \nabla, \nabla^*)$ : dually flat space

$\{\theta^i\}$ : $\nabla$-affine coordinate system

$\implies$ there exits an $\nabla^*$-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

\((M, g, \nabla, \nabla^*)\) : a dually flat space

\(\{\theta^i\} \): a \(\nabla\)-affine coordinate system

\(\{\eta_i\} \): the dual coordinate system of \(\{\theta^i\}\)

\[\Rightarrow\] there exists functions \(\psi, \phi\) on \(M\) such that

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

In addition, the following formulas hold

\[g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}, \quad g^{ij} = \frac{\partial^2 \phi}{\partial \eta_i \partial \eta_j}, \quad (2)\]

where

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

\((g^{ij}) \): the inverse matrix of \((g_{ij})\)

\(\psi\) is the \(\theta\)-potential function, and \(\phi\) is the \(\eta\)-potential function.

We say that the relation (1) the Legendre transformation.
Definition 5.3
\[ \rho : M \times M \rightarrow \mathbb{R} : \text{(canonical) divergence on } (M, g, \nabla, \nabla^*) \]
\[ \rho(p || q) := \psi(p) + \phi(q) - \sum_{i=1}^{n} \theta^i(p) \eta^i(q), \quad (p, q \in M). \]

Proposition 5.4
The definition of \( \rho \) is independent of choice of affine coordinate system on \( M \).

Example 5.5 (Euclidean space)
\( \mathbb{R}^m \) : Euclidean space
\( \langle \ , \ \rangle \) : the standard inner product
\( D \) : the standard flat affine connection.
\[ \implies (M, \langle \ , \ \rangle, D, D) \text{ is a dually flat space,} \]
\[ \rho(p || q) = \frac{1}{2}d(p, q)^2. \]
the Kullback-Leibler divergence

\[ KL(p(\theta)||p(\theta')) = \int_{\Omega} \log \frac{p(\theta)}{p(\theta')} p(\theta) \, dx \]

\[ = \int_{\Omega} \left( \log p(\theta) - \log p(\theta') \right) p(\theta) \, dx \]

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

\[ KL(p(\theta)||p(\theta')) = \int_{\Omega} \left( \sum_{i=1}^{n} \theta_i F_i(x) - \psi(\theta) - \sum_{i=1}^{n} \theta'_i F_i(x) + \psi(\theta') \right) p(\theta) \, dx \]

\[ = \psi(\theta') - \psi(\theta) + \sum_{i=1}^{n} \theta_i \eta_i - \sum_{i=1}^{n} \theta'_i \eta_i \]

\[ = \psi(\theta') + \phi(\theta) - \sum_{i=1}^{n} \theta'_i \eta_i \]

\[ = \rho(p(\theta')||p(\theta)) \]
\((M, \nabla, h)\) : 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 \(\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} = g_{ij}\)

\(\implies f: M \to R^{n+1}: \) 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 \(\{f, \xi\},\)

\[v = (-\eta_1, \ldots, -\eta_n, 1) \quad \eta_i = \frac{\partial \psi}{\partial \theta^i}\]

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

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

\[= -\sum \eta_i(q)\theta^i(p) + \psi(p) + \sum \eta_i(q)\theta^i(q) - \psi(q)\]

\[= \psi(p) + \phi(q) - \sum \eta_i(q)\theta^i(p) = \rho^C(p, q)\]

The geometric divergence coincides with the canonical divergence.
\((M, \nabla, h)\) : 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
\[\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} = g_{ij}\]

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

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

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

\[\iff \quad \langle v(p), \xi_p \rangle = 1, \quad \langle v(p), f^*_p X_p \rangle = 0\]

We define a function on \(M \times M\) by

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

\(\rho\) is called the geometric divergence on \(M\).
最尤法の復習

\( S \) は、指数族を表す
\( M \) は、\( S \) に埋め込まれた曲線指数族

与えられた \((x_1, \cdots, x_N)\)，の最尤関数 \( L \) は以下のよう定義されます。

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

\( \hat{u} \)：最大尤度推定量

実際に最尤推定量を求めることは、対数尤度方程式を解く。

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

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

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

すなわち

\( \hat{u} \)：ある推定量

\[
d(KL\hat{\eta})(X_u) = 0 \quad (\forall X_u \in T_uM)
\]
5.2 Divergences for non-integrable systems

\(\omega : \Gamma(TM) \to R^{n+1}: a \ R^{n+1}\)-valued 1-form

\(\xi : M \to R^{n+1}: a \ R^{n+1}\)-valued function

\[\text{Definition 5.6}\]
{\(\omega, \xi}\)} is called an affine distribution
\(\iff\) For an arbitrary point \(p \in M\),

\[
\begin{align*}
(1) \quad R^{n+1} &= \text{Image } \omega_p \oplus R\{\xi_x\} \\
(2) \quad \text{Image } (d\omega)_p &\subset \text{Image } \omega_p
\end{align*}
\]

\[X\omega(Y) = \omega(\nabla_X Y) + h(X, Y)\xi, \quad X\xi = -\omega(SX) + \tau(X)\xi.\]

\(\omega: \text{non-degenerate } \iff h: \text{non-degenerate}\)

\{\(\omega, \xi\): equiaffine \(\iff\) \(\tau = 0\)

\[\text{Remark 5.7} \quad \text{Image } (d\omega)_p \subset \text{Image } \omega_p \iff h: \text{symmetric.}\]
**SLD Fisher metrics**

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

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

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

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

For $P \in \mathcal{S}$, $X \in \mathcal{A}_0$, define $\omega_P(\tilde{X})$ ($\in \text{Herm}(d)$) and $\xi$ by 

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

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

The induced quantities are given by 

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

$$
\nabla_{\tilde{X}} \tilde{Y} = h_P(\tilde{X}, \tilde{Y})P - \frac{1}{2}(X\omega_P(\tilde{Y}) + \omega_P(\tilde{Y})X).
$$
\{\omega, \xi\} : nondegenerate, equiaffine

\(v : M \to R_{n+1}\) is the conormal map of \(\{\omega, \xi\}\)

\(\exists \overset{\text{def}}{\iff} \langle v(p), \xi_p \rangle = 1,\)
\(\langle v(p), \omega(X_p) \rangle = 0\)

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

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

\(\rho\) is called the geometric pre-divergence on \(M\).
<table>
<thead>
<tr>
<th>Information geometry</th>
<th>Differential geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>$\nabla$-affine coordinates</td>
</tr>
<tr>
<td>natural parameters</td>
<td>\</td>
</tr>
<tr>
<td>exponential arc</td>
<td>$\nabla$-geodesic</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$\nabla^*$-affine coordinates</td>
</tr>
<tr>
<td>mixture parameters</td>
<td></td>
</tr>
<tr>
<td>expectation parameters</td>
<td></td>
</tr>
<tr>
<td>$g$ or $h$</td>
<td>Fisher metric</td>
</tr>
<tr>
<td>expectation parameters</td>
<td>Riemannian metric</td>
</tr>
<tr>
<td>$T$ or $C$</td>
<td>skewness</td>
</tr>
<tr>
<td>skewness</td>
<td>cubic form</td>
</tr>
<tr>
<td>$\psi$</td>
<td>cumulant generating function</td>
</tr>
<tr>
<td>free energy</td>
<td>affine hypersurface</td>
</tr>
<tr>
<td>$\phi$</td>
<td>entropy</td>
</tr>
<tr>
<td>entropy</td>
<td>dual map</td>
</tr>
<tr>
<td>Legendre transformation</td>
<td>dual transformation</td>
</tr>
<tr>
<td>$D, \rho, \cdots$</td>
<td>Kullback-Leibler divergence</td>
</tr>
<tr>
<td>relative entropy</td>
<td>geometric divergence</td>
</tr>
<tr>
<td>relative entropy</td>
<td>affine support function</td>
</tr>
<tr>
<td>$\rho, \omega, \cdots$</td>
<td>prior distribution</td>
</tr>
<tr>
<td>prior distribution</td>
<td>volume form</td>
</tr>
<tr>
<td>$T, \cdots$</td>
<td>bias-correction</td>
</tr>
<tr>
<td>bias-correction</td>
<td>Tchebychev vector</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>
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 ↔ 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