The Annals of Statistics
1982, Vol. 10, No. 2, 357–385

DIFFERENTIAL GEOMETRY OF CURVED EXPONENTIAL FAMILIES—CURVATURES AND INFORMATION LOSS

SHUN-ICHI AMARI

University of Tokyo

The differential-geometrical framework is given for analyzing statistical problems related to multi-parameter families of distributions. The dualistic structures of the exponential families and curved exponential families are elucidated from the geometrical viewpoint. The duality connected by the Legendre transformation is thus extended to include two kinds of affine connections and two kinds of curvatures.

The second-order information loss is calculated for Fisher-efficient estimators, and is decomposed into the sum of two non-negative terms. One is related to the exponential curvature of the statistical model and the other is related to the mixture curvature of the estimator. Only the latter term depends on the estimator, and vanishes for the maximum-likelihood estimator. A set of statistics which recover the second-order information loss are given. The second-order efficiency also is obtained. The differential geometry of the function space of distributions is discussed.

1. Introduction. A statistical model specifies a family of distributions which are usually described by a set of parameters, thus constituting a parameter space. A parameter space has some natural geometrical structures due to the properties of the distributions. It is not only important but useful to take these structures into account when one treats statistical problems. Rao (1945) is one of the first who made differential-geometrical considerations on parameter spaces by introducing a Riemannian metric in terms of the Fisher information matrix.

Since then many researchers have tried to construct a geometrical theory in statistics. The author of the present paper remarked in 1959 that the two-dimensional parameter space of the family of one-dimensional normal distributions is a space of negative constant curvature. Although this work remains unpublished, it was followed and extended by Yoshizawa (1971), Takiyama (1974) and Sato et al. (1979). Ozeki (1977) noticed the Riemannian structure of the auto-regression models of time-series. There must have been many Riemannian-geometrical investigations in statistics. We can mention among others the works of Atkinson and Mitchell (1981), Ingarden et al. (1979, (1981) as well as Amari (1968). See also Reeds (1975).

The non-Riemannian point of view was introduced by Chentsov (1972) by defining a one-parameter family of affine connections in the space of statistical distributions. He considered the category of the spaces of distributions on a finite number of atoms with Markov morphisms. He proved that (a constant multiple of) the Fisher information metric is the only invariant metric and that the one-parameter family of affine connections are the only invariant connections in the category (Chentsov, 1972). He elucidated the geometric and dualistic structures of the space of the exponential family (Chentsov, 1966, 1972).

It was through Efron (1975) that an idea was opened up by introducing the "statistical curvature". Efron proved that the second-order information loss and the

Received March 1980; revised March 1981.

AMS 1970 *subject classification.* 62E20, 62B10.

Key words and phrases. Statistical curvatures, statistical affine connections, information loss, recovery of information, duality, second-order efficiency, Kullback-Leibler distance, asymptotic estimation theory.

second-order variance of the MLE are related to the statistical curvature of a curve representing a one-parameter family (one-dimensional space) of distributions. The concept of connections is indispensible for the curvature. Dawid (1975) suggested that the exponential connection is implicitly used in Efron's approach, and introduced other possible connections in the function space of distributions. Further fruitful results are anticipated by these papers. Madsen (1979) succeeded in extending the result of Efron (1975) to the multi-parameter case. She defined the (exponential) curvature of a submanifold embedded in an exponential family, and proved that the covariance of an unbiased efficient estimator is decomposed into the sum of three non-negative terms, of which one is the curvature term.

The present paper intends to give a differential-geometrical framework for analyzing statistical problems by the use of the one-parameter family of affine connections (α-connections). The second-order information loss and the second-order covariance of a general Fisher efficient estimator will be given in terms of the exponential ($\alpha = 1$) curvature of the statistical model and the mixture ($\alpha = -1$) curvature of the ancillary subspace associated with the estimator.

First, we study the geometrical structures of parameter spaces of distributions by defining the metric and a one-parameter family of affine connections. In order to do this, we shall develop further the idea of Dawid (1975). We then treat the structures of the exponential families and the curved exponential families, which can be embedded in the exponential families as subspaces (Efron, 1975, 1978). It will then be found that these families have multifold dualistic structures: the duality connected by the Legendre transformation (Chentsov, 1972; Barndorff-Nielsen, 1978), the duality between two kinds of connections (Chentsov, 1972), and the duality between two kinds of curvatures. All of these dualities are intimately related to yield a geometric edifice. We shall finally proceed to clarify the second-order information loss of Fisher-efficient estimators, and to decompose it into the sum of two non-negative terms. One is related to the exponential curvature of the statistical model and the other is related to the mixture curvature of the estimator. It is shown that the latter vanishes for the maximum likelihood estimator, proving that this estimator has the minimum loss. This is the generalization and extension of Efron's result (1975). A set of statistics which recover the second-order loss will be derived immediately. The second-order covariance of the estimator is also given in terms of geometrical quantities (cf. Madsen, 1979).

Our results are not necessarily completely new, and statisticians are aware of the second-order loss and the second- or third-order efficiency (K. Takeuchi, private communication). See also papers published by Fisher (1925), Rao (1962), Pfanzagl (1973), Akahira and Takeuchi (1979) and others. However, the geometrical approach seems to offer many suggestions to other theoretical and practical statistical problems concerning, for example, the role of ancillary statistics, robust estimators, fitness of statistical models, etc. The present approach will provide a primary basis for exploring such problems.

In the present paper we adopt a rather classical and intuitive way of description of differential geometry, referring mainly to Schouten (1954) (see also Eisenhart, 1927), but in a slightly modified manner. This is in order to provide an easier introduction for readers not familiar with differential geometry. We can of course rewrite the theory in terms of the modern approach using fibre-bundles, etc.

In the Appendix we will give a rough sketch of the differential geometry of the function space of distributions (cf. Dawid, 1977). To this end, we introduce a one-parameter family of quasi-distances in the function space of distributions. These distances are closely related to the α-connections, so that we can elucidate the role and meaning of the α-connections.

2. Geometry of parameter spaces.

2.1 *Parameter space, tangent space and metric.* Let us consider a family S of distributions of (vector) random variables $\mathbf{x}$, such that a distribution is specified by a set of n real parameters $\boldsymbol{\theta} = (\theta^1, \theta^2, \cdots, \theta^n)$. Then, we can construct an n-dimensional space

S^n of distributions with a coordinate system $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)$. Let $p(\mathbf{x}, \boldsymbol{\theta})$ denote the probability density function of $\mathbf{x}$ specified by $\boldsymbol{\theta}$. We assume the regularity conditions that the density functions $p(\mathbf{x}, \theta)$ exist on some carrier measure of the space of $\mathbf{x}$, that they are smooth with respect to θ, that

$$p(\mathbf{x}, \boldsymbol{\theta}) > 0$$

for all $\mathbf{x}$ in the common domain X of $\mathbf{x}$, and that $p(\mathbf{x}, \boldsymbol{\theta})$ is defined for $\boldsymbol{\theta}$ belonging to an open set Θ homeomorphic to an n-dimensional Euclidean space R^n. Let

$$\ell(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}, \boldsymbol{\theta}). \tag{2.1}$$

Then, we may regard every point $\boldsymbol{\theta}$ of S^n as carrying a function $\ell(\mathbf{x}, \boldsymbol{\theta})$ of $\mathbf{x}$.

Let $T_{\boldsymbol{\theta}}$ be the tangent space of S^n at $\boldsymbol{\theta}$, which is, roughly speaking, identified with a linearized version of a small neighborhood of $\boldsymbol{\theta}$ in S^n. Let $\mathbf{e}_i$ $(i = 1, 2, \ldots, n)$ be the natural basis of $T_{\boldsymbol{\theta}}$ associated with the coordinate system. Then, an infinitesimally small line-element $d\boldsymbol{\theta}$ stemming from the point $\boldsymbol{\theta} = (\theta^1, \theta^2, \cdots, \theta^n)$ to à neighboring point $\boldsymbol{\theta} + d\boldsymbol{\theta} = (\theta^1 + d\theta^1, \cdots, \theta^n + d\theta^n)$, is identified with a vector

$$d\boldsymbol{\theta} = \textstyle\sum_{i=1}^n d\theta^i \mathbf{e}_i \tag{2.2}$$

in the tangent space $T_{\boldsymbol{\theta}}$.

Since each point $\boldsymbol{\theta}$ of S^n carries a function $\ell(\mathbf{x}, \boldsymbol{\theta})$ of $\mathbf{x}$, it is natural to regard $\mathbf{e}_i(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ as representing the function

$$\mathbf{e}_i(\boldsymbol{\theta}) = \partial_i \ell(\mathbf{x}, \boldsymbol{\theta}),$$

where ∂_i is the abbreviation for the partial derivative with regard to θ^i, $\partial_i = \partial/\partial\theta^i$. We have

$$\textstyle\sum d\theta^i \mathbf{e}_i = \sum d\theta^i \partial_i \ell(\mathbf{x}, \boldsymbol{\theta}) = \ell(\mathbf{x}, \boldsymbol{\theta} + d\boldsymbol{\theta}) - \ell(\mathbf{x}, \boldsymbol{\theta}),$$

which represents an infinitesimal change in $\ell(\mathbf{x}, \boldsymbol{\theta})$ entailed by the change $d\boldsymbol{\theta}$. Hence, $T_{\boldsymbol{\theta}}$ is spanned by n functions $\partial_i \ell(\mathbf{x}, \boldsymbol{\theta})$. We now define the inner product "$\cdot$" in each $T_{\boldsymbol{\theta}}$ by defining n^2 quantities

$$g_{ij}(\boldsymbol{\theta}) = \mathbf{e}_i \cdot \mathbf{e}_j, \qquad i, j = 1, \cdots, n. \tag{2.3}$$

We define g_{ij} by the inner product of two derivatives $\partial_i \ell(\mathbf{x}, \boldsymbol{\theta})$ and $\partial_j \ell(\mathbf{x}, \boldsymbol{\theta})$ in the function space with respect to the weight $p(x, \theta)$.

DEFINITION 1.

$$g_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\partial_i \ell(\mathbf{x}, \boldsymbol{\theta}) \partial_j \ell(\mathbf{x}, \boldsymbol{\theta})\}, \tag{2.4}$$

where $E_{\boldsymbol{\theta}}$ denotes the expectation with respect to the distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

In order that $g_{ij}(\boldsymbol{\theta})$ is well defined and that g_{ij} is positive definite (or $\mathbf{e}_i$'s are linearly independent), we need some regularity conditions. These are all fulfilled in the case that S^n is a smooth submanifold in a full, regular, minimally represented exponential family (Barndorff-Nielsen, 1978). Under these conditions, the n^2 quantities $g_{ij}(\boldsymbol{\theta})$, $i, j = 1, \cdots, n$, constitute the metric tensor of S^n, by which the square of the length ds of an infinitesimal line-element $d\boldsymbol{\theta}$ is given with the quadratic form

$$ds^2 = d\boldsymbol{\theta} \cdot d\boldsymbol{\theta} = \textstyle\sum_{i,j,} g_{ij}(\boldsymbol{\theta})\, d\theta^i\, d\theta^j. \tag{2.5}$$

We hereafter assume Einstein's summation convention, in which the summation is automatically taken without the symbol $\sum$ for such indices as appear twice in one term, once as a subscript and once as a superscript, as for i and j in the above formula. Hence the summation symbol $\sum$ can be neglected on the right-hand side, $g_{ij}\, d\theta^i\, d\theta^j$ denoting the same quantity without the symbol $\sum$ as the one with it.

The tensor $g_{ij}(\boldsymbol{\theta})$, which has n^2 components, $i, j = 1, \cdots, n$, is called the Riemannian metric tensor. As is well known, in this case it is nothing other than the Fisher information matrix, whose role is clear from the following.

CRAMÉR-RAO THEOREM. *The covariance matrix of any unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ cannot be smaller than the inverse* (g^{ij}) *of the Fisher information matrix* (g_{ji}),

$$E_{\boldsymbol{\theta}}\{(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j)\} - g^{ij} \geq 0, \tag{2.6}$$

where the sign $\geq$ *is used in the sense that the matrix on the left-hand side of* (2.6) *is positive semi-definite.*

From the above theorem, we see that the distance ds in (2.5) represents a degree of distinction between the two distributions $p(\mathbf{x}, \boldsymbol{\theta})$ and $p(\mathbf{x}, \boldsymbol{\theta} + d\boldsymbol{\theta})$. The distance s between two distant points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is defined in a Riemannian space as the minimum value of the integral of ds along all the curves connecting the two points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The curve C that gives the minimum value is called the geodesic with respect to the metric g_{ij}.

2.2 *Affine connections.* Let $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ be the two tangent spaces at two neighboring points $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$, respectively. It is sometimes necessary to compare vectors in $T_{\boldsymbol{\theta}}$ and in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$. However, $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ are two different vector spaces so that there is no means of comparing them without introducing some correspondence between $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$. To this end, we introduce an affine connection, by which two tangent spaces $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ become feasible for comparison.

Let us take a natural basis vector $\mathbf{e}_i(\boldsymbol{\theta} + d\boldsymbol{\theta})$ of $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$, and consider which vector in $T_{\boldsymbol{\theta}}$ is in correspondence with it. The vector $\mathbf{e}_i(\boldsymbol{\theta})$ may not be the one, because the space S^n is "curved," or, if not, the coordinate system $\{\theta^i\}$ may be curvilinear. Let us presume that $\mathbf{e}_i(\boldsymbol{\theta} + d\boldsymbol{\theta})$ in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ corresponds to a vector $\mathbf{e}_i(\boldsymbol{\theta}) + \delta\mathbf{e}_i$ in $T_{\boldsymbol{\theta}}$. Because of the continuity of $\mathbf{e}_i$, $\delta\mathbf{e}_i$ should be small, depending linearly on $d\boldsymbol{\theta}$. Hence, when we represent $\delta\mathbf{e}_i$ as a linear combination of the basis $\{\mathbf{e}_k\}$ of $T_{\boldsymbol{\theta}}$, the coefficients of the combination are small quantities which are linear in $d\boldsymbol{\theta}$. Hence, we can write

$$\delta\mathbf{e}_i = d\theta^j\Gamma^k_{ji}\mathbf{e}_k(\boldsymbol{\theta}), \tag{2.7}$$

where the summations with respect to j and k are automatically assumed because of the Einstein summation convention. The quantity having n^3 components $\Gamma^k_{ji}(\boldsymbol{\theta})(i, j, k = 1, \cdots, n)$ defines an affine connection in S^n. An affine correspondence is established between two neighboring tangent spaces $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ by an affine connection. A natural basis $\mathbf{e}_i(\boldsymbol{\theta} + d\boldsymbol{\theta})$ in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ is mapped into a vector $\mathbf{e}_i(\boldsymbol{\theta}) + d\theta^j\Gamma^k_{ji}(\boldsymbol{\theta})\mathbf{e}_k(\boldsymbol{\theta})$ by this correspondence.

Let us consider a vector field $\mathbf{X}(\boldsymbol{\theta})$, and let $\mathbf{X}(\boldsymbol{\theta} + d\boldsymbol{\theta})$ and $\mathbf{X}(\boldsymbol{\theta})$ be two vectors in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}}$, respectively. We can define their "true difference" with the help of the affine connection. By representing $\mathbf{X}(\boldsymbol{\theta} + d\boldsymbol{\theta})$ and $\mathbf{X}(\boldsymbol{\theta})$ in terms of the respective natural bases, we have

$$\mathbf{X}(\boldsymbol{\theta} + d\boldsymbol{\theta}) = X^i(\boldsymbol{\theta} + d\boldsymbol{\theta})\mathbf{e}_i(\boldsymbol{\theta} + d\boldsymbol{\theta}), \qquad \mathbf{X}(\boldsymbol{\theta}) = X^i(\boldsymbol{\theta})\mathbf{e}_i(\boldsymbol{\theta}).$$

We have by expansion

$$X^i(\boldsymbol{\theta} + d\boldsymbol{\theta}) = X^i(\boldsymbol{\theta}) + dX^i,$$

where

$$dX^i = (\partial_j X^i)\, d\theta^j$$

represents the apparent change in the components of the two vectors. The affine connection maps $\mathbf{X}(\boldsymbol{\theta} + d\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ to

$$X^i(\boldsymbol{\theta} + d\boldsymbol{\theta})(\mathbf{e}_i + \delta\mathbf{e}_i) = (X^i + dX^i)(\mathbf{e}_i + \Gamma^k_{ji}\, d\theta^j\mathbf{e}_k) = (X^i + dX^i + \Gamma^i_{jk}X^k\, d\theta^j)\mathbf{e}_i$$

of $T_{\boldsymbol{\theta}}$. Hence the "true" or "intrinsic" difference is represented by the infintesimal vector $D\mathbf{X}$, whose components are given by

$$DX^i = dX^i + \Gamma^i_{jk}X^k\, d\theta^j, \tag{2.8}$$

where the second term arises due to the difference between the natural bases at $T_{\boldsymbol{\theta}}$ and

$T_{\theta+d\theta}$. When $DX^i = 0$, or the apparent change in the components satisfies

$$dX^i = -\Gamma^i_{jk} X^k \, d\theta^j,$$

then the two vectors $\mathbf{X}(\boldsymbol{\theta} + d\boldsymbol{\theta})$ and $\mathbf{X}(\boldsymbol{\theta})$ are regarded as essentially the same, and they are said to be parallel shifts of each other.

The rate

$$\frac{DX^i}{d\theta^j} = \frac{\partial X^i(\boldsymbol{\theta})}{\partial \theta^j} + \Gamma^i_{jk} X^k(\boldsymbol{\theta}) \tag{2.9}$$

represents the "true change" in the vector field $\mathbf{X}(\boldsymbol{\theta})$, and is called the covariant derivative of the vector field. It is a tensor having n^2 components, and is denoted shortly by $\nabla_j X^i$ (Schouten, 1954, page 122).

Let $\boldsymbol{\theta}(t)$ be a smooth curve in S^n. It is called a path or a geodesic with respect to the affine connection, when the tangent vectors $\dot{\boldsymbol{\theta}}(t) = d\boldsymbol{\theta}/\mathrm{d}t$ are parallel along the curve, i.e., the intrinsic change in direction vanishes

$$\frac{D\dot{\boldsymbol{\theta}}(t)}{dt} = 0 \tag{2.10}$$

along the curve, or $\boldsymbol{\theta}(t)$ satisfies the equation

$$\frac{d^2\theta^i(t)}{dt^2} + \Gamma^i_{jk} \dot{\theta}^j \dot{\theta}^k = 0. \tag{2.11}$$

The geodesic is a natural extension of the straight line in a Euclidean space. If $\Gamma^i_{jk}(\boldsymbol{\theta})$ vanishes identically, the geodesic equation is linear in t: $\theta^i(t) = a^i t + b^i$.

Mathematically speaking, the parameters of an affine connection $\Gamma^i_{jk}(\boldsymbol{\theta})$ can be defined arbitrarily (assuming adequate smoothness). In our statistical problem, the connection should be defined such that it represents the structure of distributions. Chentsov (1972) defined a one-parameter family of affine connections and proved that they are the only connections invariant in the categories of distributions on finite sample space. We now show "heuristics" in defining affine connections, following the ideas of Dawid (1975). This leads to the same one-parameter family of affine connections as introduced by Chentsov.

Multiplying both sides of (2.7) by $\mathbf{e}_m$ and taking the inner product, we have

$$\mathbf{e}_m \cdot \delta \mathbf{e}_i = d\theta^j \, \Gamma_{jim}, \tag{2.12}$$

where

$$\Gamma_{jim} = \Gamma^k_{ji} g_{km} \tag{2.13}$$

is the covariant (i.e., lower indices) expression for the affine connection, and

$$\Gamma^k_{ji} = \Gamma_{jim} g^{mk} \tag{2.14}$$

holds, where g^{mk} is the inverse (or the contravariant expression) of the metric tensor g_{km}. Since $\mathbf{e}_i(\boldsymbol{\theta})$ is represented by $\partial_i \ell(\mathbf{x}, \boldsymbol{\theta})$, a formal expansion yields

$$\mathbf{e}_i(\boldsymbol{\theta} + d\boldsymbol{\theta}) \sim \partial_i \ell(\mathbf{x}, \boldsymbol{\theta} + d\boldsymbol{\theta}) = \partial_i \ell(\mathbf{x}, \boldsymbol{\theta}) + \partial_i \partial_j \ell(\mathbf{x}, \boldsymbol{\theta}) \, d\theta^j.$$

Hence, if the additional term $\partial_i \partial_j \ell(\mathbf{x}, \boldsymbol{\theta}) \, d\theta^j$ is a linear combination of $\partial_i \ell(\mathbf{x}, \boldsymbol{\theta})$ or $\mathbf{e}_i(\boldsymbol{\theta})$, it is included in $T_{\boldsymbol{\theta}}$ and hence $\delta \mathbf{e}_i$ could be defined by the function $\partial_i \partial_j \ell(\mathbf{x}, \boldsymbol{\theta}) \, d\theta^j$ of $\mathbf{x}$. However, by virtue of

$$E_{\boldsymbol{\theta}}\{\partial_i \ell(\mathbf{x}, \boldsymbol{\theta})\} = 0, \tag{2.15}$$

any expression $\ell(\mathbf{x})$ of a vector in $T_{\boldsymbol{\theta}}$ should satisfy

$$E_{\boldsymbol{\theta}}\{\ell(\mathbf{x})\} = 0. \tag{2.16}$$

We get by simple calculations

$$E_{\boldsymbol{\theta}}\{\partial_i \partial_j \ell(\mathbf{x}, \boldsymbol{\theta})\} = -E_{\boldsymbol{\theta}}\{\partial_i \ell \partial_j \ell\} = -g_{ij}(\boldsymbol{\theta}),$$

and hence $\partial_i\partial_j\ell$is not included in T_θ. We modify it in two ways: One is

$$\overset{1}{\delta}_i(\mathbf{x}, \theta) = \partial_i\partial_j\ell(\mathbf{x}, \theta)\, d\theta^j + g_{ij}(\theta)\, d\theta^j$$

and the other is

$$\overset{2}{\delta}_i(\mathbf{x}, \theta) = \partial_i\partial_j\ell(\mathbf{x}, \theta)\, d\theta^j + \partial_i\ell(\mathbf{x}, \theta)\partial_j\ell(\mathbf{x}, \theta)\, d\theta^j.$$

Both of them satisfy the requirement (2.16). Combining them linearly, we have

$$\overset{\alpha}{\delta}_i\ell(\mathbf{x}, \theta) = \frac{1+\alpha}{2}\overset{1}{\delta}_i(\mathbf{x}, \theta) + \frac{1-\alpha}{2}\overset{2}{\delta}_i(\mathbf{x}, \theta), \tag{2.17}$$

where α is a parameter. If we regard the projection of $\delta_i(\mathbf{x}, \theta)$ to T_θ as the expression of the increment $\delta\mathbf{e}_i$, which defines the affine connection, then we have from (2.12) that

$$d\theta^j\, \overset{\alpha}{\Gamma}_{jim} = \mathbf{e}_m \cdot \delta\mathbf{e}_i = E_\theta\{\partial_m\ell(\mathbf{x}, \theta)\overset{\alpha}{\delta}_i(\mathbf{x}, \theta)\}, \tag{2.18}$$

where the affine connection depends on the parameter α. Here, we have replaced the inner product "$\cdot$" by E_θ in the function expressions. By substituting (2.17) in (2.18) and by calculating it, we are led to

DEFINITION 2.

$$\overset{\alpha}{\Gamma}_{jim}(\theta) = E_\theta\{\partial_j\partial_i\ell(\mathbf{x}, \theta)\partial_m\ell(\mathbf{x}, \theta)\} + \frac{1-\alpha}{2} E_\theta\{\partial_j\ell(\mathbf{x}, \theta)\partial_i\ell(\mathbf{x}, \theta)\partial_m\ell(\mathbf{x}, \theta)\}. \tag{2.19}$$

We have thus defined a one-parameter family of affine connections, and$\overset{\alpha}{\Gamma}_{ijk}$ is called the α-connection.

The roles of $\alpha = 1, -1$ and 0 connections were studied by Chentsov (1972) and also remarked upon by Dawid (1975). First, we treat an exponential family of distributions

$$p(\mathbf{x}, \theta) = \exp\{c(\mathbf{x}) + \theta^i x_i - \psi(\theta)\}, \tag{2.20}$$

specified by the natural parameters or the natural coordinate system $\theta = (\theta^i)$. (Other parametrizations are also possible for specifying the distributions.) We have, in this special case,

$$\partial_i\ell(\mathbf{x}, \theta) = x_i - \partial_i\psi(\theta), \qquad \partial_i\partial_j\ell(\mathbf{x}, \theta) = -\partial_i\partial_j\psi(\theta).$$

Hence, from (2.4) and (2.19), taking account of (2.15), we have

$$g_{ij}(\theta) = \partial_i\partial_j\psi(\theta),$$

$$\overset{\alpha}{\Gamma}_{ijk}(\theta) = \frac{1-\alpha}{2} E_\theta\{\partial_i\ell\partial_j\ell\partial_k\ell\}. \tag{2.21}$$

Since $\overset{\alpha}{\Gamma}_{ijk}(\theta)$ vanishes identically for $\alpha = 1$, any exponential family of distributions constitutes an uncurved space when the $\alpha = 1$ connection is adopted. The natural parameters play the role of the Cartesian coordinate system. The 1-connection is therefore called the exponential connection and is denoted by $\overset{e}{\Gamma}_{ijk}$.

Next, we consider a family of distributions given by a mixture of $n + 1$ prescribed linearly independent distributions$p_0(\mathbf{x})$, $\mathrm{p}_1(\mathbf{x}), \cdots, p_m(\mathbf{x})$,

$$p(\mathbf{x}, \theta) = \theta^i p_i(\mathbf{x}) + (1 - \textstyle\sum_{\iota=1}^n \theta^i)p_0(\mathbf{x}),$$

where $0 < \theta^\iota < 1$. In this case, we have

$$\partial_i\ell(\mathbf{x}, \theta) = \frac{1}{p(\mathbf{x}, \theta)}\{p_i(\mathbf{x}) - p_0(\mathbf{x})\},$$

$$\partial_i\partial_j\ell(\mathbf{x}, \theta) = -\partial_i\ell(\mathbf{x}, \theta)\partial_j\ell(\mathbf{x}, \theta),$$

so that

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \frac{1+\alpha}{2} E_{\theta}\{\partial_i \ell \partial_j \ell \partial_k \ell\}, \tag{2.22}$$

which vanishes identically for $\alpha = -1$. Hence, a family of mixture distributions constitutes an uncurved space with the -1-connection. This connection is called the mixture connection and is denoted by $\overset{m}{\Gamma}_{ijk}$.

When $\alpha = 0$, calculations yield

$$\overset{0}{\Gamma}_{ijk} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}), \tag{2.23}$$

whose right-hand side is a quantity called the Christoffel symbol of the first kind and usually denoted by $[ij;k]$, (Shouten, 1954, page 132). This connection is derived from the metric tensor g_{ij}. It is the natural connection compatible with the metric tensor, and the induced parallelism is known as the Levi-Civita parallelism. The space S^n is Riemannian in this case. The connection $\overset{0}{\Gamma}_{ijk}$ may be called the information connection, since it is derived from the Fisher information.

It should be noted that the length of a vector generally changes by a parallel shift, because the connection is defined independently of the metric. The angle between two vectors is also not invariant under a parallel shift. This is because the covariant differentiation of the metric tensor does not vanish in general. Such a connection is said to be nonmetric. Let

$$T_{ijk} = E_{\theta}(\partial_i \ell \partial_j \ell \partial_k \ell). \tag{2.24}$$

Then, we have

$$\overset{\alpha}{\nabla}_i g_{jk} = \alpha T_{ijk}, \tag{2.25}$$

where $\overset{\alpha}{\nabla}_i$ denotes the covariant derivative with respect to the α connection $\overset{\alpha}{\Gamma}_{ijk}$. Hence, the covariant differentiation of the metric vanishes for $\alpha = 0$ connection. The length of a vector is invariant under a parallel shift by the information connection.

By the use of the important tensor T_{ijk}, the connections are represented as

$$\overset{\alpha}{\Gamma}_{ijk} = [ij:k] - \frac{\alpha}{2} T_{ijk}. \tag{2.26}$$

2.3. *Geometric quantities and coordinate transformations.* We have so far defined the geometric structures of a general distribution space in terms of a specific coordinate system $\{\theta^i\}$. We may use another coordinate system $\{\eta^{i'}\}$ to describe the geometry, where the one-to-one relation of the coordinate transformation

$$\theta^i = \theta^i(\eta^1, \eta^2, \cdots, \eta^n),$$
$$\eta^{i'} = \eta^{i'}(\theta^1, \theta^2, \cdots, \theta^n), \qquad i, i' = 1, \cdots, n,$$

holds. Geometric structures are described by quantities, which consist of a number of components, e.g., the metric is represented by a quantity g_{ij} having n^2 components and the affine connection is represented by a quantity Γ_{ijk} having n^3 components.

Although the components of a geometric quantity are described with reference to a specific coordinate system, they represent the geometric structures which are independent of coordinate systems. For example, the components g_{ij} of the metric and X^i of a vector $\mathbf{X}$ depend numerically on the coordinate system, but the length of $\mathbf{X}$

$$|\mathbf{X}|^2 = g_{ij}X^iX^j$$

is invariant.

In order to ascertain this invariance, we should know how the components of a quantity are transformed under the coordinate transformation. Let $\bar{B}_j^{i'}$ be the Jacobian matrix of the transformation, i.e.

$$\bar{B}_j^{i'} = \frac{\partial \eta^{i'}}{\partial \theta^j}$$

and let $B_{j'}^i$ be its inverse, i.e.

$$B_{j'}^i = \frac{\partial \theta^i}{\partial \eta^{j'}}.$$

A tensor, S_{ij}^k for example, is a quantity whose components are transformed to $\bar{S}_{i'j'}^{k'}$ in the $\eta^{i'}$-coordinate,

$$\bar{S}_{i'j'}^{k'} = B_{i'}^l B_{j'}^m \bar{B}_n^{k'} S_{lm}^n. \tag{2.27}$$

A vector X^i is a tensor having only one upper (i.e., contravariant) index, and is transformed to

$$\bar{X}^{i'} = \bar{B}_j^{i'} X^j.$$

The metric tensor g_{ij} has two lower (i.e., covariant) indices, and is transformed to

$$\bar{g}_{i'j'} = B_{i'}^k B_{j'}^l g_{kl}.$$

From these transformation rules, one can see that the length $|\mathbf{X}|$ of a vector is absolutely invariant in whatever coordinate system it is presented,

$$g_{ij} X^i X^i = \bar{g}_{i'j'} \bar{X}^{i'} \bar{X}^{j'}.$$

A tensorial equation, for example, $S_{ij}^k = 0$, is invariant for any coordinate system, because a 0-tensor is mapped to a 0-tensor by any coordinate transformation.

We can lower or raise any indices of a tensor by multiplying the metric tensor g_{ij} or its inverse g^{ji}, for example,

$$S_{ijk} = S_{ij}^m g_{mk}, \qquad S_{jk}^i = S_{mjk} g^{ki}.$$

These are different expressions of the same quantity. The inverse matrix g^{ij} itself is an alternate version of g_{ji} with upper indices, because of

$$g^{ij} = g^{ik} g^{jm} g_{km}.$$

There are geometric quantities whose components are not transformed like tensors. The affine connection is such a quantity and its components are represented in the $\eta^{i'}$-coordinate system as (Shouten, 1954, page 124)

$$\Gamma_{i'j'k'} = B_{i'}^l B_{j'}^m B_{k'}^n \Gamma_{lmn} + B_{k'}^l (\partial_{i'} B_{j'}^m) g_{lm}. \tag{2.28}$$

Hence, even if $\Gamma_{ijk}(\boldsymbol{\theta}) = 0$ holds for all $\boldsymbol{\theta}$ in a specific coordinate system, it is not necessarily equal to 0 in another coordinate system.

In the following, we analyze the geometric structures of distribution spaces with reference to some specific coordinate systems. However, essentially, we study the invariant structures. Hence, we should take care of the rules of transformations of geometric quantities to ascertain the invariance. We sometimes use a restricted equation

$$\Gamma_{ijk} \overset{*}{=} 0 \quad \text{or} \quad \Gamma_{ijk} \overset{*}{=} \alpha T_{ijk},$$

which is not a tensorial equation (because Γ_{ijk} is not a tensor). It holds only for the specific coordinate system we have currently adopted for the ease of analysis. We attach "*" above the equality sign in order to emphasize this specific character of the equation. Of course, these equations represent invariant geometrical structures, but they take different forms in different coordinate systems, and we can derive their exact forms from the rules of transformations of such quantities.

Refer to textbooks of differential geometry for more rigorous and detailed accounts. We mainly refer to Chapters III and V of Schouten (1954).

3. Geometry of exponential families.

3.1. *Geometric quantities in natural coordinates.* An exponential family can be written in the form (2.20) by choosing natural parameters θ^i where x_i are functions of the original sample variables, constituting a set of sufficient statistics. The natural coordinate system θ^i is determined uniquely to within affine transformations. Since the characteristic function

$$\phi(\boldsymbol{\xi}, \boldsymbol{\theta}) = E_\theta[\exp\{i\boldsymbol{\xi}\cdot\mathbf{x}\}], \qquad i^2 = -1,$$

of $p(\mathbf{x}, \boldsymbol{\theta})$ is easily calculated as

$$\log \phi(\boldsymbol{\xi}, \boldsymbol{\theta}) = \psi(\boldsymbol{\theta} + i\boldsymbol{\xi}) - \psi(\boldsymbol{\theta}),$$

the function $\psi(\boldsymbol{\theta})$ includes sufficient information for the distributions. The geometric quantities are calculated as follows. From

$$\partial_i \ell = x_i - \partial_i\psi, \qquad \partial_i\partial_j\ell = -\partial_i\partial_j\psi, \qquad \partial_i\partial_j\partial_k\ell = -\partial_i\partial_j\partial_k\psi,$$

we easily derive

$$E_\theta(x_i) = \partial_i\psi(\boldsymbol{\theta}), \tag{3.1}$$

$$g_{ij}(\boldsymbol{\theta}) \overset{*}{=} \partial_i\partial_j\psi(\boldsymbol{\theta}). \tag{3.2}$$

Since g_{ij} is positive-definite, $\psi(\boldsymbol{\theta})$ is a convex function. From

$$E_\theta\{\partial_i\partial_j\partial_k\ell(\mathbf{x}, \boldsymbol{\theta})\} \overset{*}{=} -E_\theta\{\partial_i\ell\partial_j\ell\partial_k\ell\},$$

we obtain

$$T_{ijk}(\boldsymbol{\theta}) \overset{*}{=} \partial_i\partial_j\partial_k\psi(\boldsymbol{\theta}). \tag{3.3}$$

From

$$E_\theta\{\partial_i\partial_j\ell\partial_k\ell\} = E_\theta\{-g_{ij}(\boldsymbol{\theta})\partial_k\ell\} = 0,$$

we have

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) \overset{*}{=} \frac{1-\alpha}{2} T_{ijk}(\boldsymbol{\theta}). \tag{3.4}$$

The left-hand side is not a tensor but the right-hand side is a tensor. Hence, this equality holds only for the natural coordinate system; cf. (2.26), which holds for any coordinate system.

The Riemann-Christoffel curvature tensor (Schouten, 1954, page 138) of this space is calculated as

$$\overset{\alpha}{R}_{ijkl}(\boldsymbol{\theta}) = \frac{1-\alpha^2}{2} T_{km[i}T_{j]ln}g^{mn}, \tag{3.5}$$

where $[ij]$ denotes the alteration with respect to indices i and j, i.e., $T_{km[i}T_{n]ln} = (T_{kmi}T_{jln} - T_{kmj}T_{iln})/2$. We see that $\overset{\alpha}{R}_{ijkl}$ vanishes for $\alpha = \pm 1$, i.e., for the exponential and mixture connections, giving spaces of distant parallelism. We give a simple example next.

EXAMPLE 3.1. *Normal distribution.* The family of one-dimensional normal distribution is specified by two parameters, the mean μ and the variance σ^2, and the density function is

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

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

Hence, by introducing the natural parameters $\boldsymbol{\theta} = (\theta^1, \theta^2)$,

$$\theta^1 = \frac{\mu}{\sigma^2}, \qquad \theta^2 = -\frac{1}{2\sigma^2},$$

and by introducing a vector $x = (x_1, x_2)$ defined by

$$x_1 = x, \qquad x_2 = x^2,$$

we have

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp\{\theta^i x_i - \psi(\boldsymbol{\theta})\},$$

where

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

We have

$$E(x_1) = \partial_1\psi = \mu(\boldsymbol{\theta}), \qquad E(x_2) = \partial_2\psi = \mu^2(\boldsymbol{\theta}) + \sigma^2(\boldsymbol{\theta}).$$

The metric tensor given by $\partial_i\partial_i\psi$ has the following matrix form,

$$[g_{ij}(\boldsymbol{\theta})] = \begin{bmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{bmatrix}.$$

The components of the tensor T_{ijk} are given by

$$T_{111} = 0, \qquad T_{112} = 2\sigma^4, \qquad T_{122} = 8\mu\sigma^4, \qquad T_{222} = 24\mu^2\sigma^4 + 8\sigma^6.$$

(Since T_{ijk} is symmetric, other components are not necessary.) The affine connection $\overset{\alpha}{\Gamma}_{ijk}$ is given by (3.4). the Riemann-Christoffel curvature tensor has only one independent component in the case of $n = 2$, and is

$$\overset{\alpha}{R}_{1212} = (1 - \alpha^2)\sigma^6.$$

The scalar curvature κ is defined by

$$\kappa = \frac{1}{n(n-1)} \overset{\alpha}{R}_{ijkl} g^{il} g^{jk},$$

and we have in this case

$$\kappa = 0 \quad \text{for} \quad \alpha = \pm 1,$$

$$\kappa = -\frac{1}{2} \quad \text{for} \quad \alpha = 0.$$

Hence, the Riemannian space ($\alpha = 0$) is a space of constant negative curvature, well known in non-Euclidean geometry. This result was first remarked by Amari.

3.2. *Dual space.* Let us consider the sample space $\tilde{S}^n$ whose point is specified by $\mathbf{x} = (x_1, \cdots, x_n)$. Given a distribution specified by $\boldsymbol{\theta}$, the sample expectation $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ is given by

$$\eta_i = E_{\boldsymbol{\theta}}(x_i) = \partial_i\psi(\boldsymbol{\theta}). \tag{3.6}$$

Since $\psi(\boldsymbol{\theta})$ is a convex function, there is a one-to-one correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ connected by the Legendre transformation in the case of a full, regular, canonical exponential family with a minimal representation (Chentsov, 1972; Barndorff-Nielsen, 1978). Defining

$$\phi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\{\eta_i\theta^i - \psi(\boldsymbol{\theta})\}, \tag{3.7}$$

we have the inverse transformation

$$\theta^i = \partial^i \phi(\boldsymbol{\eta}), \tag{3.8}$$

where ∂^i stands for

$$\partial^\iota = \frac{\partial}{\partial \eta_\iota}.$$

Moreover,

$$\psi(\boldsymbol{\theta}) = \max\{\eta_i \theta^i - \phi(\boldsymbol{\eta})\} \tag{3.9}$$

holds, and the two functions are connected by the identity

$$\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\eta}) = \theta^i \eta_i. \tag{3.10}$$

The two spaces S^n and $\tilde{S}^n$ are thus dually connected. Since $\boldsymbol{\eta}$ can also be used to specify the distribution, we can analyze the geometric structure of the distribution space in the dual space $\tilde{S}^n$.

By the one-to-one correspondence between S^n and $\tilde{S}^n$, the geometric structures such as the metric and affine connections can be transplanted in $\tilde{S}^n$. One may consider $\boldsymbol{\eta}$ as a coordinate system of S^n by identifying the two spaces S^n and $\tilde{S}^n$. However, it should be noted that $\boldsymbol{\eta}$ has a lower index, so that the roles of the upper (contravariant) and lower (covariant) indices are interchanged in this coordinate system $\boldsymbol{\eta}$ or in the dual space $\tilde{S}^n$.

The Jacobian matrix of the transformation is

$$\frac{\partial \eta_i}{\partial \theta^j} = \partial_j \partial_i \psi(\boldsymbol{\theta}) = g_{ji}, \tag{3.11}$$

and hence for the inverse transformation, it is

$$\frac{\partial \theta^i}{\partial \eta_j} = \partial^i \partial^j \phi(\boldsymbol{\eta}) = g^{ij}. \tag{3.12}$$

Any tensor, S_{ij}^k for example, is represented by

$$S_k^{ij} = S_{lm}^n g^{li} g^{nj} g_{nk}$$

in the dual space, i.e., only by lowering and raising the corresponding upper and lower indices, respectively. Obviously, the metric tensor in $\tilde{S}^n$ is g^{ij}, and the infinitesimal line-element $d\eta_i$ is related to $d\theta^j$ thus:

$$d\eta_i = g_{ij} d\theta^j, \qquad d\theta^j = g^{ji} d\eta_i,$$

and its length satisfies

$$ds^2 = g^{ij} d\eta_i d\eta_j.$$

The dual space, or the dual system, has an important meaning. It gives the coordinate system in which the Cramer-Rao bound is attained. Let

$$\hat{\eta}_i = x_i$$

define an estimator for the parameters $\boldsymbol{\eta}$. We see that it is unbiased

$$E(\hat{\eta}_i) = \eta_i,$$

and it attains the bound

$$E\{(\hat{\eta}_i - \eta_i)(\hat{\eta}_j - \eta_j)\} = g_{ij}.$$

EXAMPLE 3.1. (continued). In the case of the normal distribution, we have

$$\phi(\boldsymbol{\eta}) = -\frac{1}{2}\log(\eta_2 - \eta_1^2) - \frac{1}{2}\log 2\pi e.$$

The one-to-one correspondence is given by

$$\theta^1 = \frac{\eta_1}{\eta_2 - \eta_1^2}, \qquad \theta^2 = -\frac{1}{2(\eta_2 - \eta_1^2)},$$

and

$$\eta_1 = -\frac{\theta^1}{2\theta^2}, \qquad \eta_2 = \left(\frac{\theta^1}{2\theta^2}\right)^2 - \frac{1}{2\theta^2}.$$

The dual variables η are related to μ and σ^2 by

$$\eta_1 = \mu, \qquad \eta_2 = \mu^2 + \sigma^2.$$

The Fisher information, or the metric tensor, becomes g^{ij} in the dual space, which is the inverse of g_{ji}, and is given by

$$g^{ij} = \frac{1}{\sigma^4}\begin{bmatrix} \sigma^2 + 2\mu^2 & -\mu \\ -\mu & \dfrac{1}{2} \end{bmatrix}.$$

3.3. *Duality*. The duality between the two spaces has been studied by Chentsov (1972), Barndorff-Nielsen (1978) and Efron (1978). It can be extended to include the geometrical structures such as the $\alpha - (-\alpha)$ duality or the exponential connection and mixture connection duality; see also Chentsov (1972).

As we have shown, the dualistic correspondence of any tensor can be obtained by lowering the upper indices and by raising the lower indices of its components. Hence, the Fisher information or the metric tensor has the following dual correspondence in the two spaces

$$g_{ij} \overset{*}{=} \partial_i \partial_j \psi, \qquad g^{ij} \overset{*}{=} \partial^i \partial^j \phi.$$

Similarly, the tensor T_{ijk} has the following expressions

$$T_{ijk} \overset{*}{=} \partial_i \partial_j \partial_k \psi, \qquad T^{ijk} \overset{*}{=} -\partial^i \partial^j \partial^k \phi,$$

since we have

$$\begin{aligned} T^{ijk} &= g^{il} g^{jm} g^{kn} T_{lmn} = g^{il} g^{jm} g^{kn} \partial_l g_{mn} \\ &= -g^{jm} g_{mn} g^{il} (\partial_l g^{kn}) = -\partial^i \partial^j \partial^k \phi. \end{aligned}$$

Since $\overset{\alpha}{\Gamma}_{ijk}$ is not a tensor, the corresponding α-connection $\overset{\alpha}{\Gamma}{}^{ijk}$ in the dual space cannot be derived by raising the indices. We can calculate the $\overset{\alpha}{\Gamma}_{ijk}$ from the rule (2.27) of the transformation of the affine connection. Thus, the α-connection has the following dual expressions

$$\overset{\alpha}{\Gamma}_{ijk} \overset{*}{=} \frac{1-\alpha}{2} T_{ijk}, \qquad \overset{\alpha}{\Gamma}{}^{ijk} \overset{*}{=} -\frac{1+\alpha}{2} T^{ijk}.$$

It is interesting that the exponential connection $(\alpha = 1)$ $\overset{e}{\Gamma}_{ijk}$ vanishes identically in the normal coordinate system, while the mixture connection $(\alpha = -1)$ $\overset{m}{\Gamma}{}^{ijk}$ vanishes identically in the dual space. Hence the dual space (the dual coordinate system) is straight or Cartesian for the mixture connection, and the mixture connection plays the same role as the exponential connection does in the primal space. Thus, the duality is extended between the exponential and mixture connection, or the α- and $(-\alpha)$-connections (cf. Chentsov, 1972).

THEOREM 1. *The exponential connection vanishes identically in the normal coordinate system, while the mixture connection vanishes in the dual coordinate system. They are thus in a dual correspondence.*

We can study the dual of the Kullback-Leibler information (Chentsov, 1972; Efron, 1978). The Kullback-Leibler information (Kullback, 1959) defined by

$$K(1:2) = E_{\theta 1}\{\ell(\mathbf{x}, \boldsymbol{\theta}_1) - \ell(\mathbf{x}, \boldsymbol{\theta}_2)\} \tag{3.13}$$

is regarded as a quasi-distance from point $\boldsymbol{\theta}_1$ to point $\boldsymbol{\theta}_2$. It is not symmetric and does not satisfy the triangular inequality, but when $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1$ are located sufficiently close to each other, we have, by putting $d\boldsymbol{\theta} = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$ and neglecting the higher order terms,

$$K(1:2) = \frac{1}{2} g_{ij} d\theta^i \, d\theta^j.$$

For the exponential family, it is written as

$$K(1:2) = (\theta_1^i - \theta_2^i)\eta_{1i} - \{\psi(\boldsymbol{\theta}_1) - \psi(\boldsymbol{\theta}_2)\}, \tag{3.14}$$

where $\eta_1 = \eta_1(\boldsymbol{\theta}_1)$. We can define the dual $\tilde{K}(1:2)$ of $K(1:2)$ by interchanging formally the role of $\boldsymbol{\theta}$ and η as

$$\tilde{K}(1:2) = (\eta_{1i} - \eta_{2i})\theta_1^i - \{\phi(\eta_1) - \phi(\eta_2)\}. \tag{3.15}$$

By simple calculations, we can prove (Chentsov, 1972; Efron, 1978).

THEOREM 2.

$$\tilde{K}(1:2) = K(2:1). \tag{3.16}$$

This shows that $K(2:1)$ is the dual counterpart of $K(1:2)$.

4. Geometry of curved exponential families.

4.1. *Curved exponential families.* A family of distributions is called a curved exponential family when it can be embedded smoothly in the space of an exponential family. Let the curved exponential family S^m be specified by m parameters $\mathbf{u} = (u^1, u^2, \cdots, u^m)$, $p(\mathbf{x}, \mathbf{u})$ denoting the probability density function of S^m. Let $p(\mathbf{x}, \boldsymbol{\theta})$ be the probability density function of the enveloping space S^n, where $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)$ is the natural coordinates. Then, we have

$$\theta^i = \theta^i(\mathbf{u}), \tag{4.1}$$

such that

$$p(\mathbf{x}, \mathbf{u}) = p(\mathbf{x}, \boldsymbol{\theta}(\mathbf{u})).$$

This is the parametric representation of S^m, which forms an m-dimensional submanifold in S^n.

In order to calculate the geometric quantities of S^m, we need

$$\partial_a \ell(\mathbf{x}, \mathbf{u}) = \partial_a \log p(\mathbf{x}, \mathbf{u}), \tag{4.2}$$

where

$$\partial_a = \frac{\partial}{\partial u^a}, \qquad a = 1, 2, \cdots, m,$$

and hereafter we denote indices concerning the coordinate system $\mathbf{u} = (u^a)$ by the letters a, b, c, etc. From (4.2), we have

$$\partial_a \ell(\mathbf{x}, \mathbf{u}) = B_a^i(\mathbf{u}) \partial_i \ell(\mathbf{x}, \boldsymbol{\theta}(\mathbf{u})), \tag{4.3}$$

where

$$B_a^i(\mathbf{u}) = \frac{\partial \theta^i}{\partial u^a}. \tag{4.4}$$

THEOREM 3. *The metric g_{ab} and the α-connection $\overset{\alpha}{\Gamma}_{abc}$ of S^m is related to those in S^n by*

$$g_{ab}(\mathbf{u}) = B^i_a B^j_b g_{ij}(\boldsymbol{\theta}(\mathbf{u})), \tag{4.5}$$

$$\overset{\alpha}{\Gamma}_{abc}(\mathbf{u}) = (\partial_a B^i_b) B^j_c g_{ij} + B^i_a B^j_b B^k_c \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}(\mathbf{u})). \tag{4.6}$$

PROOF. From the definition (2.4) of the metric, we have

$$g_{ab} = E_{\mathbf{u}}[\partial_a \ell \partial_b \ell] = B^i_a B^j_b E_{\theta(\mathbf{u})}[\partial_i \ell \partial_j \ell] = B^i_a B^j_b g_{ij}.$$

From

$$\partial_a \partial_b \ell = \partial_a (B^j_b \partial_j \ell) = (\partial_a B^j_b) \partial_j \ell + B^i_a B^j_b \partial_i \partial_j \ell$$

and the definition (2.19) of $\overset{\alpha}{\Gamma}_{abc}$, we obtain (4.6) in a similar manner.

4.2. *Curvatures of S^m in S^n.* Let us consider the tangent space T_u of S^m at $\mathbf{u}$. It is a subspace of the tangent space T_θ at $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{u})$, and is spanned by m tangent vectors of S^m, $\mathbf{B}_1 = (B^i_1)$, $\mathbf{B}_2 = (B^i_2)$, $\cdots$, $\mathbf{B}_m = (B^i_m)$. Let

$$B_{aj}(\mathbf{u}) = g_{ji} B^i_a, \tag{4.7}$$

be the covariant expression of the tangent vectors. Since the vector n^i normal to S^m in S^n is defined by the equations

$$B_{aj} n^j = 0, \qquad a = 1, \cdots, m.$$

B_{aj} may be regarded as the vectors which implicitly define the directions normal to S^m with respect to the information metric.

The curvature of a subspace is defined by a quantity representing the intrinsic change in the tangent or normal directions of the subspace (Schouten, 1954, page 256). The intrinsic change is measured invariantly by the covariant differentiation by the use of the α-connection. The curvature depends on which connection we are using. We first consider the rate of change in the tangent direction from $B^i_b(\mathbf{u})$ at $\mathbf{u}$ to $B^i_b(\mathbf{u} + d\mathbf{u})$ at $\mathbf{u} + d\mathbf{u}$,

$$\lim_{d\mathbf{u} \to 0} \frac{\mathbf{B}_b(\mathbf{u} + d\mathbf{u}) - \mathbf{B}_b(\mathbf{u})}{du^a},$$

where the subtraction is carried into effect by mapping the vector $\mathbf{B}_b(\mathbf{u} + \mathbf{du})$ in $T_{\theta(\mathbf{u}+d\mathbf{u})}$ to $T_{\theta(\mathbf{u})}$ by the α-connection. This quantity can be formally written as

$$\overset{\alpha}{H}{}^i_{ab}(\mathbf{u}) = B^j_a(\mathbf{u}) \overset{\alpha}{\nabla}_j B^i_b(\mathbf{u}) = \partial_a B^i_b(\mathbf{u}) + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a(\mathbf{u}) B^k_b(\mathbf{u}) \tag{4.8}$$

and is called the α-curvature, where $\overset{\alpha}{\nabla}_j$ denotes the covariant derivative with respect to the α-connection.

Similarly, by treating the change in the normal directions, we have another α-curvature

$$\overset{\alpha}{L}_{abi} = B^j_a \overset{\alpha}{\nabla}_j B_{bi} = \partial_a B_{bi} - \overset{\alpha}{\Gamma}{}^j_{ki} B^k_a B_{bj}. \tag{4.9}$$

Since the normal and tangent directions are orthogonal to each other, both definitions usually coincide. However, our α-connections are in general non-metric, and the orthogonality is not preserved by the parallel shift of vectors. Hence, they do not coincide. We have indeed

$$\overset{\alpha}{L}_{abi} = B^j_a \overset{\alpha}{\nabla}_j (B^k_b g_{ki}) = (B^j_a \overset{\alpha}{\nabla}_j B^k_b) g_{ki} + B^j_a B^k_b \overset{\alpha}{\nabla}_j g_{ki} = \overset{\alpha}{H}{}^k_{ab} g_{ki} + \alpha T_{jik} B^j_a B^k_b, \tag{4.10}$$

which is different from the covariant form $\overset{\alpha}{H}_{abi} = \overset{\alpha}{H}{}^k_{ab} g_{ki}$ of $\overset{\alpha}{H}{}^i_{ab}$. They coincide only in the case of the information connection ($\alpha = 0$) which is metric.

We can treat the curved exponential family also in the dual space, where it is embedded as

$$\eta = \eta(\mathrm{u}), \tag{4.11}$$

and the curvatures are similarly obtained. The following theorem shows that the α-$(-\alpha)$ duality is accompanied by the H-L dualistic correspondence.

THEOREM 4.

$$\overset{\alpha}{H}{}^i_{ab} = \overset{-\alpha}{L}{}^i_{ab}. \tag{4.12}$$

PROOF. By direct calculations, we have

$$\overset{-\alpha}{L}{}^i_{ab} = g^{ij}\overset{-\alpha}{L}_{abj} = g^{ij}(\partial_a B_{bj} - \overset{-\alpha}{\Gamma}{}^l_{kj}B^k_aB_{bl}) = g^{ij}_a(B^k_b g_{kj}) - \frac{1+\alpha}{2}g^{ij}T^l_{kj}B^k_aB_{bl}$$

$$= \partial_a B^i_b + g^{ij}B_bB^l_a\partial_l g_{kj} - \frac{1+\alpha}{2}T^i_{ab} = \partial_a B^i_b + \frac{1-\alpha}{2}T^i_{ab} = \overset{\alpha}{H}{}^i_{ab},$$

where we put

$$T^i_{ab} = B^j_aB^k_bT^i_{jk}. \tag{4.13}$$

We now need to know the rule of transformation of the curvature $\overset{\alpha}{H}{}^i_{ab}$. This is a quantity having two lower indices a, b and one upper index i. It behaves like a vector in S^n with respect to the index i. Hence, for fixed a and b, it is a vector. When the coordinate system is changed from $\mathbf{u} = (u^a)$ to $\mathrm{v} = (\mathrm{v}^{a'})$ in S^m, the components change in the following manner,

$$\overset{\alpha}{H}{}^i_{a'b'} = B^a_{a'}B^b_{b'}\overset{\alpha}{H}{}^i_{ab} + B^i_b\partial_{a'}B^b_{b'}, \tag{4.14}$$

where

$$B^a_{a'} = \frac{\partial u^a}{\partial u^{a'}}.$$

Hence, this quantity is not a tensor. However, it gives a tensor, when it is projected to the normal subspace of $T_\mathbf{u}$ in T_θ, i.e., its normal components with respect to i form a tensor, because the second non-tensorial term of the right-hand side of (4.14) has tangent components only and they vanish when projected to the normal subspace. Let N^i_j be the projection operator of a vector in T_θ to the subspace normal to the subspace $T_\mathbf{u}$ of S^m. Then, the normal components of $\overset{\alpha}{H}{}^i_{ab}$ are written as $N^i_j\overset{\alpha}{H}{}^j_{ab}$, and they form a tensor representing the intrinsic curvature of S^m. We have introduced the curvature tensors $N^i_j\overset{\alpha}{H}{}^j_{ab}$ and $N^i_j\overset{\alpha}{L}{}^j_{ab}$ and in a little different way from Schouten (1954, page 256), but the results are the same. It also should be remarked that the projection of $\overset{\alpha}{H}{}^i_{ab}$ to the subspace $T_\mathbf{u}$ is nothing but the α-connection of S^m, as

$$\overset{\alpha}{\Gamma}_{abc} = \overset{\alpha}{H}{}^i_{ab}B^j_cg_{ij}. \tag{4.15}$$

EXAMPLE 4.1. Let z be a normal random variable $z \sim N(1, a^2)$ where a is a constant and let

$$x = uz,$$

where u is an unknown parameter. Then, x is also a normal random variable

$$x \sim N(u, u^2a^2)$$

specified by one parameter u. The distributions constitute a one-dimensional curved exponential family S^1, which are embedded in the S^2 of the normal distributions by

$$\mu = u, \qquad \sigma^2 = a^2u^2,$$

or in terms of the natural coordinates by

$$\theta^1 = \frac{1}{a^2u}, \qquad \theta^2 = -\frac{1}{2a^2u^2}.$$

Therefore, S^1 is a parabola

$$\theta^2 = -\frac{a^2}{2}(\theta^1)^2$$

in S^2. The tangent vector is

$$B_1^i = \frac{d\theta^i}{du} = \frac{1}{a^2u^3}[-u, 1].$$

The metric g_{ab} of S^1 has only one component g_{11} in this case, and is

$$g_{11} = B_1^iB_1^jg_{ij} = \left(\frac{2a^2+1}{a^2}\right)\frac{1}{u^2}.$$

The α-curvature of S^1 is a vector in S^2 given by

$$\overset{\alpha}{H}{}_{ab}^i = \overset{\alpha}{H}{}_{11}^i = \frac{1}{2a^4u^4}\{[-2u, 1-2a^2] + \alpha[2u(1+2a^2), -(4a^2+1)]\}.$$

The normal component of $\overset{\alpha}{\mathrm{H}}{}_{ab}^i$ is given by

$$N_j^i\overset{\alpha}{H}{}_{11}^j = \frac{1+2a^2-\alpha}{2a^4(2a^2+1)u^4}[-2u(1+a^2), 1].$$

4.3. *The maximum-likelihood estimator and its dual.* Let $\bar{\mathbf{x}}$ be the observed data from an unknown distribution belonging to the curved exponential family S^m. When N independent observations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ are made, we take the sufficient statistics

$$\bar{\mathbf{x}} = \frac{1}{N}\textstyle\sum_{i=1}^N \mathbf{x}_i. \tag{4.16}$$

In this case, the likelihood function $\ell(\mathbf{x}_1, \cdots, \mathbf{x}_N; \mathbf{u})$ becomes

$$\ell(\mathbf{x}_1, \cdots, \mathbf{x}_N; \mathbf{u}) = N\ell(\bar{\mathbf{x}}, \mathbf{u}), \tag{4.17}$$

so that we only have to make a slight modification by multiplying the metric, the affine connection, and the tensor T_{ijk} by N, to adopt Ng_{ij}, $N\overset{\alpha}{\Gamma}{}_{ijk}$ and NT_{ijk}, respectively.

Let us define the data point $\bar{\boldsymbol{\eta}}$ in the dual space by

$$\bar{\boldsymbol{\eta}} = \bar{\mathbf{x}}, \tag{4.18}$$

or the data point $\bar{\boldsymbol{\theta}}$ in the primal space by

$$\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}(\bar{\boldsymbol{\eta}}). \tag{4.19}$$

The point $\bar{\boldsymbol{\eta}}$ or $\bar{\boldsymbol{\theta}}$ is by itself a sufficient statistic and represents a distribution in $\tilde{S}^n$ or in S^n.

The maximum likelihood estimator (MLE) $\hat{\mathbf{u}}$ is defined as the estimator which maximizes the likelihood $\ell(\bar{\mathbf{x}}, \mathbf{u})$. It is easy to see that the MLE $\hat{\mathbf{u}}$ represents the point in S^m that is located closest to the data point in the sense of the Kullback-Liebler distance,

$$\min_{\mathbf{u}\in S^m} K(\bar{\boldsymbol{\theta}} : \boldsymbol{\theta}(\mathbf{u})) = K(\bar{\boldsymbol{\theta}} : \boldsymbol{\theta}(\hat{\mathbf{u}})).$$

The equation determining $\hat{\mathbf{u}}$ is obtained by differentiation as

$$(4.20) \qquad \{\bar{\eta}_i - \eta_i(\hat{\mathbf{u}})\}B_a^i(\hat{\mathbf{u}}) = 0,$$

where $\eta_i(\mathbf{u})$ is the dual point of $\theta^i(\mathbf{u})$.

In a dual manner to the above, one can define an estimator $\tilde{\mathbf{u}}$, which is the point in S^m closest to the data point in the sense of the dual of the Kullback-Leibler distance,

$$\min_{\mathbf{u}\in S^m}\tilde{K}(\bar{\boldsymbol{\theta}}:\boldsymbol{\theta}(\mathbf{u})) = \min_{\mathbf{u}\in S^m}K(\boldsymbol{\theta}(\mathbf{u}):\bar{\boldsymbol{\theta}}) = K(\boldsymbol{\theta}(\tilde{\mathbf{u}}):\bar{\boldsymbol{\theta}}),$$

(Kullback, 1959; Efron, 1978). We call $\tilde{\mathbf{u}}$ the dual maximum likelihood estimator (DMLE). The equation for determining $\tilde{\mathbf{u}}$ is given by

$$(4.21) \qquad \{\bar{\theta}^i - \theta^i(\tilde{\mathbf{u}})\}B_{ai}(\tilde{\mathbf{u}}) = 0.$$

It should be noted that these estimators are defined independently of the specific coordinate systems $\mathbf{u}$.

In order to elucidate the properties of these estimators, we introduce a subset $\hat{A}(\mathbf{u})$ of S^n by

$$(4.22) \qquad \hat{A}(\mathbf{u}) = \{\boldsymbol{\theta} \mid \min_{\mathrm{v}\in S^m}K(\boldsymbol{\theta}:\boldsymbol{\theta}(\mathrm{v})) = K(\boldsymbol{\theta}:\boldsymbol{\theta}(\mathbf{u}))\},$$

that is, $\hat{A}(\mathbf{u})$ consists of those points $\boldsymbol{\theta}$ from which $\boldsymbol{\theta}(\mathbf{u})$ is the closest among all the points in the subspace S^m. When and only when the data point $\bar{\boldsymbol{\theta}}$ belongs to $\hat{A}(\mathbf{u})$ is $\mathbf{u}$ obtained by the MLE. We call $\hat{A}(\mathbf{u})$ the ancillary domain of $\mathbf{u}$ by maximum likelihood estimation.

Dually to the above, we introduce the ancillary domain $\hat{A}(\mathbf{u})$ of $\mathbf{u}$ with respect to the dual maximum likelihood estimation by

$$(4.23) \qquad \tilde{A}(\mathbf{u}) = \{\boldsymbol{\theta} \mid \min_{\mathrm{v}\in S^m}K(\boldsymbol{\theta}(\mathrm{v}):\boldsymbol{\theta}) = K(\boldsymbol{\theta}(\mathbf{u}):\boldsymbol{\theta})\}.$$

When and only when the data point belongs to $\tilde{A}(\mathbf{u})$ is $\mathbf{u}$ obtained by the DMLE. The two sets $\hat{A}(\mathbf{u})$ and $\tilde{A}(\mathbf{u})$ form $(n - m)$-dimensional submanifolds which intersect S^m at the point $\boldsymbol{\theta}(\mathbf{u})$ only.

Let us elucidate the geometric meaning of the MLE and DMLE. A geodesic with respect to the exponential connection (mixture connection) will be called an exponential geodesic (a mixture geodesic). More generally, a geodesic with respect to the α-connection will be called an α-geodesic.

THEOREM 5. $\hat{A}(\mathbf{u})$ *and* $\tilde{A}(\mathbf{u})$ *are composed respectively of the mixture and exponential geodesics which intersect* S^m *at point* $\boldsymbol{\theta}(\mathbf{u})$ *and are orthogonal to* S^m *at this point.*

PROOF. We first study $\tilde{A}(\mathbf{u})$. We see from (4.21) that a point $\boldsymbol{\theta}$ belongs to $\tilde{A}(\mathbf{u})$, only when it satisfies

$$(4.24) \qquad \{\theta^i - \theta^i(\mathbf{u})\}B_a^j(\mathbf{u})g_{ij}(\mathbf{u}) = 0,$$

where $g_{ij}(\mathbf{u})$ implies $g_{ij}(\boldsymbol{\theta}(\mathbf{u}))$. Let us consider an exponential geodesic $\boldsymbol{\theta}(t)$ which passes through $\boldsymbol{\theta}(\mathbf{u})$ and is orthogonal to S^m. Then, the equation of the geodesic is

$$\frac{D\dot{\boldsymbol{\theta}}(t)}{dt} = \frac{d^2\theta^i}{dt} + \Gamma_{jk}^i\dot{\theta}^j\dot{\theta}^k = \ddot{\theta}^i = 0$$

because of $\overset{e}{\Gamma}{}_{jk}^i = 0$ in the natural coordinates, with the initial conditions

$$\theta^i(0) = \theta^i(\mathbf{u}), \qquad \dot{\theta}^i(0)B_a^j(\mathbf{u})g_{ij}(\boldsymbol{\theta}(\mathbf{u})) = 0.$$

The solution is given by

$$(4.25) \qquad \theta^i(t) = \dot{\theta}^i(0)t + \theta^i(\mathbf{u}),$$

which satisfies (4.24). Conversely, every solution of (4.24) is written in the form of (4.25) with suitable $\dot{\theta}^i(0)$ orthogonal to S^m. Hence, we have proved the theorem for $\tilde{A}(\mathbf{u})$.

We use the dual space or the dual coordinates $\boldsymbol{\eta}$ to obtain $\hat{A}(\mathbf{u})$. A point $\boldsymbol{\eta}$ belongs to $\hat{A}(\mathbf{u})$ only when

$$\{\eta_i - \eta_i(\mathbf{u})\}B^i_a(\mathbf{u}) = 0. \tag{4.26}$$

Now consider a mixture goedesic $\boldsymbol{\eta}(t)$ which passes through $\boldsymbol{\eta}(\mathbf{u})$ and which is orthogonal to S^m at $\boldsymbol{\eta}(\mathbf{u})$. The equation of the mixture geodesic path is written as $\ddot{\eta}_i(t) = 0$, because the mixture connection vanishes in these coordinates. The initial conditions are

$$\eta_i(0) = \eta_i(\mathbf{u}), \qquad \dot{\eta}_i(0)B_{ja}(\mathbf{u})g^{ij} = \dot{\eta}_i B^i_a(\mathbf{u}) = 0,$$

because the tangent vectors are

$$\partial_a \eta_j(\mathbf{u}) = \partial_a \theta^i \partial_i \eta_j = g_{ij}B^i_a = B_{ja}.$$

Hence, the solution is

$$\eta_i(t) = \dot{\eta}_i(0)t + \eta_i(\mathbf{u}),$$

which satisfies (4.26) and vice versa, proving the theorem. Since the geodesic is a geometric concept defined invariantly, the theorem holds in any coordinate systems.

Extending this result, we can now define the α-estimator $\overset{\alpha}{\mathbf{u}}$. Let $\overset{\alpha}{A}(\mathbf{u})$ be the set composed of the α-geodesics that pass through $\boldsymbol{\theta}(\mathbf{u})$ and are orthogonal to S^m at this point. Then, the α-estimator $\overset{\alpha}{\mathbf{u}}$ is the estimator whose domain is $\overset{\alpha}{A}(\mathbf{u})$, that is, $\overset{\alpha}{\mathbf{u}} = \mathbf{u}$ when and only when the data point belongs to $\overset{\alpha}{A}(\mathbf{u})$. The MLE is the (-1)-estimator, and the DMLE is the 1-estimator. As will be shown in the following, all of these estimators are Fisher efficient.

EXAMPLE 4.1 (continued). We analyze the MLE and DMLE of Example 4.1 in the dual space. The equation (4.20) of the MLE is written as

$$\bar{\eta}_1\hat{u} - \bar{\eta}_2 + a^2\hat{u}^2 = 0.$$

Hence, $\hat{A}(u)$ consists of the straight line

$$\hat{A}(u) = \{\boldsymbol{\eta} \mid \eta_2 = u\eta_1 + a^2u^2\}$$

which passes through $\boldsymbol{\eta}(u) = [u, (1 + a^2)u^2]$ and is orthogonal to S^1. The MLE is given by

$$\hat{u} = \frac{1}{2a^2}\{-\bar{x}_1 + \sqrt{(\bar{x}_1^2 + 4a^2\bar{x}_2)}\}.$$

The equation of the DMLE is given in the primal space by

$$\bar{\theta}_1\tilde{u} + 2(1 + a^2)\bar{\theta}^2\tilde{u}^2 + 1 = 0,$$

which is

$$\bar{\eta}_2 = \bar{\eta}_1^2 - \tilde{u}\eta_1 + (1 + a^2)\tilde{u}^2$$

in the dual space. Hence, $\tilde{A}(u)$ is a parabola in the dual space,

$$\tilde{A}(u) = \{\boldsymbol{\eta} \mid \eta_2 = \eta_1^2 - u\eta_1 + (1 + a^2)u^2\},$$

and

$$\tilde{u} = \frac{1}{2(1 + a^2)}[\bar{x}_1 + \sqrt{\{4(1 + a^2)\bar{x}_2 - (3 + 4a^2)\bar{x}_1^2\}}].$$

We can obtain the $\overset{\alpha}{\mathbf{u}}$ by solving the equation of geodesics with respect to the α-connection.

5. Curvature and information loss.

5.1. *Consistent estimators.* We can study the asymptotic properties of consistent estimators in the framework of the geometry of curved exponential families. Let $\bar{\boldsymbol{\theta}}$ be the data point in S^n depending on N independent observations, and let

$$\bar{\mathbf{u}} = \bar{\mathbf{u}}(\bar{\boldsymbol{\theta}})$$

be a smooth estimator independent of N. Then, its ancillary domain $A(\mathbf{u})$ is defined by

$$A(\mathbf{u}) = \{\boldsymbol{\theta} \mid \mathbf{u} = \bar{\mathbf{u}}(\boldsymbol{\theta})\},$$

i.e., $\mathbf{u}$ is obtained by the estimation when and only when the data point $\bar{\boldsymbol{\theta}}$ is included in $A(\mathbf{u})$. The set $A(\mathbf{u})$ forms a smooth $(n - m)$-dimensional subspace in S^n.

The estimator $\bar{\mathbf{u}}$ is consistent if and only if $A(\mathbf{u})$ intersects S^m only at point $\boldsymbol{\theta}(\mathbf{u})$. This is proved as follows. By the law of large numbers, the data point $\bar{\mathbf{x}}$ tends to the average $\boldsymbol{\eta}(\mathbf{u})$ as N tends to ∞, where $\mathbf{u}$ is the true parameter of the distribution. The data point $\bar{\boldsymbol{\theta}}$ also tends to $\boldsymbol{\theta}(\mathbf{u})$. Hence, the estimator is consistent if and only if $\boldsymbol{\theta}(\mathbf{u})$ is included in $A(\mathbf{u})$.

Let us introduce in the submanifold $A(\mathbf{u})$ a local coordinate system $\mathbf{v} = (v^\kappa)$, $\kappa = 1, 2, \cdots, n - m$, such that the point $\boldsymbol{\theta}(\mathbf{u}) \in A(\mathbf{u})$ has the coordinates $\mathbf{v} = 0$. We use letters κ, λ, μ, etc., to represent the indices with respect to this coordinate system of $A(\mathbf{u})$. The n coordinates $(\mathbf{u}, \mathbf{v})$ together form a local coordinate system of S^n in a neighborhood of $\boldsymbol{\theta}(\mathbf{u})$. Let

$$\theta^i = \theta^i(\mathbf{u}, \mathbf{v}) \tag{5.1}$$

be the $\boldsymbol{\theta}$-coordinates of a point $(\mathbf{u}, \mathbf{v})$, i.e., a point having coordinate $\mathbf{v}$ in $A(\mathbf{u})$. Thus, the points satisfying $\mathbf{v} = 0$ constitute S^m, and the points satisfying $\mathbf{u} = \mathbf{u}_0$ constitute $A(\mathbf{u}_0)$.

The tangent space of S^m at $\boldsymbol{\theta}(\mathbf{u})$ is spanned by m vectors $B_1^i(\mathbf{u}), \cdots, B_m^i(\mathbf{u})$, where

$$B_a^i(\mathbf{u}) = \partial_a \theta^i(\mathbf{u}, 0). \tag{5.2}$$

The tangent space of $A(\mathbf{u})$ at $\boldsymbol{\theta}(\mathbf{u})$ is spanned by $n - m$ vectors

$$B_\kappa^i(\mathbf{u}, 0) = \partial_\kappa \theta^i(\mathbf{u}, 0), \qquad \kappa = 1, 2, \cdots, n - m, \tag{5.3}$$

where

$$\partial_\kappa = \frac{\partial}{\partial v^\kappa}.$$

These n vectors span the tangent space T_θ of S^n at $\boldsymbol{\theta}(\mathbf{u})$.

The α-curvature of S^m is written as

$$\overset{\alpha}{H}{}_{ab}^{i}(\mathbf{u}) = \partial_a B_b^i(\mathbf{u}) + \overset{\alpha}{\Gamma}{}_{ab}^{i}(\mathbf{u}), \tag{5.4}$$

where we put

$$\overset{\alpha}{\Gamma}{}_{ab}^{i}(\mathbf{u}) = \overset{\alpha}{\Gamma}{}_{jk}^{i}(\mathbf{u}) B_a^j(\mathbf{u}) B_b^k(\mathbf{u}) \tag{5.5}$$

and $\Gamma_{jk}^i(\mathbf{u}) = \Gamma_{jk}^i(\boldsymbol{\theta}(\mathbf{u}))$. The α-curvature of the subspace $A(\mathbf{u})$ at $\boldsymbol{\theta}(\mathbf{u})$ is defined in a similar manner,

$$\overset{\alpha}{H}{}_{\kappa\lambda}^{i}(\mathbf{u}) = B_\kappa^j(\mathbf{u}, 0) \overset{\alpha}{\nabla}_j B_\lambda^i(\mathbf{u}, 0) = \partial_\kappa B_\lambda^i(\mathbf{u}, 0) + \overset{\alpha}{\Gamma}{}_{\kappa\lambda}^{i}(\mathbf{u}), \tag{5.6}$$

where

$$\overset{\alpha}{\Gamma}{}_{\kappa\lambda}^{i}(\mathbf{u}) = \overset{\alpha}{\Gamma}{}_{jk}^{i}(\mathbf{u}) B_\kappa^j(\mathbf{u}) B_\lambda^k(\mathbf{u}). \tag{5.7}$$

and $B_\kappa^i(\mathbf{u})$ is the abbreviation for $B_\kappa^j(\mathbf{u}, 0)$.

We may study as an example, the β-estimator $\overset{\beta}{\mathbf{u}}$, whose domain is $\overset{\beta}{A}(\mathbf{u})$. Let

$n^i_1, n^i_2, \cdots, n^i_{n-m}$ be a set of independent solutions of

$$B^i_a(\mathbf{u}) n^j g_{ij}(\mathbf{u}) = 0, \qquad a = 1, \cdots, m.$$

Then, they span the directions perpendicular to S^m. Let $\boldsymbol{\theta}^i(t)$ be a β-geodesic, passing through $\boldsymbol{\theta}(\mathbf{u})$ and perpendicular to S^m. Then, it satisfies

$$\boldsymbol{\theta}(0) = \boldsymbol{\theta}(\mathbf{u}), \qquad \dot{\boldsymbol{\theta}}(0) = a^\kappa n^i_\kappa$$

for some $\mathbf{a} = (a^1, \cdots, a^{n-m})$. Moreover, we have

$$\ddot{\theta}^i(0) = -\overset{\beta}{\Gamma}{}^i_{jk}(\mathbf{u}) \dot{\theta}^j(0) \dot{\theta}^k(0)$$

from the equation of the β-geodesic. Expanding $\theta^i(t)$, we have

$$\begin{aligned} \theta^i(t) &= \theta^i(0) + \dot{\theta}^i(0) t + ½\, \ddot{\theta}^i(0) t^2 + O(t^3) \\ &= \theta^i(\mathbf{u}) + t a^\kappa n^i_\kappa - ½\, \overset{\beta}{\Gamma}{}^i_{jk}(\mathbf{u}) n^j_\kappa n^k_\lambda a^\kappa a^\lambda t^2 + O(t^3), \end{aligned}$$

where $O(t^3)$ is the terms of order t^3. Since $\overset{\beta}{A}(\mathbf{u})$ is composed of these geodesics, $v^\kappa = t a^\kappa$ can be used as a coordinate system of $\overset{\beta}{A}(\mathbf{u})$ in a neighborhood of $\boldsymbol{\theta}(u)$, and we have

$$\theta^i(\mathbf{u}, \mathbf{v}) = \theta^i(\mathbf{u}) + v^\kappa n^i_\kappa - ½\, \overset{\beta}{\Gamma}{}^i_{jk} n^j_k n^k_\lambda v^\kappa v^\lambda + O(|\mathbf{v}|^3).$$

Obviously,

$$\begin{aligned} &B^i_\kappa(\mathbf{u}, 0) = n^i_\kappa, \\ &\partial_\kappa B^i_\lambda(\mathbf{u}, 0) = -\overset{\beta}{\Gamma}{}^i_{jk} B^j_\kappa(\mathbf{u}, 0) B^k_\lambda(\mathbf{u}, 0) = -\overset{\beta}{\Gamma}{}^i_{\kappa\lambda}(\mathbf{u}). \end{aligned}$$

Hence, by substituting this in (5.6), the α-curvature of $\overset{\beta}{A}(\mathbf{u}$ of the β-estimator $\overset{\beta}{\mathbf{u}}$ is given by

$$\overset{\alpha}{H}{}^i_{\kappa\lambda} = \overset{\alpha}{\Gamma}{}^i_{\kappa\lambda} - \overset{\beta}{\Gamma}{}^i_{\kappa\lambda}. \tag{5.8}$$

Especially, the α-curvature of $\overset{\alpha}{A}(\mathbf{u})$ vanishes.

5.2 *First-order efficiency.* When the true parameter of the distribution is $\mathbf{u}$, the distribution of the data $\bar{\mathbf{x}}$ is asymptotically normal,

$$\bar{\mathbf{x}} \sim N(\boldsymbol{\eta}(\mathbf{u}), \frac{1}{N} g_{ij}(\mathbf{u})).$$

Hence, asymptotically the data point $\bar{\boldsymbol{\theta}}$ is also normally distributed

$$\bar{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}(\mathbf{u}), \frac{1}{N} g^{ij}(\mathbf{u})).$$

We can represent the data point $\bar{\boldsymbol{\theta}}$ in terms of the $(\mathbf{u}, \mathbf{v})$-coordinates of a consistent estimator. Let $\mathbf{u}$ be the true parameters. Since $\bar{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}(\mathbf{u})$ for large N, we have

$$\bar{\theta}^i = \theta^i(\mathbf{u} + \bar{\mathbf{u}}, \bar{\mathbf{v}}),$$

where $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are small terms, and the estimator takes the value $\mathbf{u} + \bar{\mathbf{u}}$. We have by expansion

$$\bar{\theta}^i = \theta^i(\mathbf{u}) + B^i_a(\mathbf{u}) \bar{u}^a + B^i_\kappa(\mathbf{u}) \bar{v}^\kappa, \tag{5.9}$$

where the higher order terms are neglected. Now we shift the coordinates $\mathbf{u}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ without losing generality such that

$$\mathbf{u} = 0, \qquad \boldsymbol{\theta}(0) = 0, \qquad \boldsymbol{\eta}(0) = 0.$$

Here, the true parameter is $\mathbf{u} = 0$. Then, (5.9) becomes

$$\bar{\theta}^i = B^i_a \bar{u}^a + B^i_\kappa \bar{v}^\kappa, \tag{5.10}$$

where B^i_a and B^i_κ are, respectively, $B^i_a(0)$ and $B^i_\kappa(0)$. We can solve this linearized equation, and obtain

$$\bar{u}^a = D^a_i \bar{\theta}^i, \qquad \bar{v}^\kappa = D^\kappa_i \bar{\theta}^\iota, \tag{5.11}$$

where the $n \times n$ matrix

$$D = \begin{bmatrix} D^a_i \\ D^\kappa_i \end{bmatrix}$$

is the inverse of the $n \times n$ matrix

$$B = [B^i_a \quad B^i_\kappa].$$

By introducing an n-vector $\mathbf{w} = [\bar{\mathbf{u}}, \bar{\mathbf{v}}]^T$, we have

$$\mathbf{w} = D\bar{\theta}, \quad \bar{\theta} = B\mathbf{w}$$

in the matrix notation, where T denotes the transposition.

Since $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are asymptotically jointly normal, having asymptotically zero mean, we calculate their variance-covariance matrix as

$$E(\bar{u}^a \bar{u}^b) = \frac{1}{N} \bar{g}^{ab}, \qquad E(\bar{u}^a \bar{v}^\kappa) = \frac{1}{N} \bar{g}^{a\kappa}, \qquad E(\bar{v}^\kappa \bar{v}^\lambda) = \frac{1}{N} \bar{g}^{\kappa\lambda}. \tag{5.12}$$

Let

$$\frac{1}{N} \bar{V} = E(\mathbf{w}\mathbf{w}^T) = \frac{1}{N} \begin{bmatrix} \bar{g}^{ab} & \bar{g}^{a\kappa} \\ \bar{g}^{\lambda b} & \bar{g}^{\lambda\kappa} \end{bmatrix}$$

in the partitioned form, and let $G = (g_{ij})$. Then,

$$\frac{1}{N} \bar{V} = DE[\bar{\theta}\bar{\theta}^T]D^T = \frac{1}{N} DG^{-1}D^T$$

and hence we have

$$\bar{V}^{-1} = B^T G B.$$

By putting

$$g_{ab} = B^i_a B^j_b g_{ij}, \qquad g_{a\kappa} = B^i_a B^j_\kappa g_{ij}, \qquad g_{\kappa\lambda} = B^i_\kappa B^j_\lambda g_{ij}, \tag{5.13}$$

we have

$$\bar{V} = \begin{bmatrix} g_{ab} & g_{a\lambda} \\ g_{\kappa b} & g_{\kappa\lambda} \end{bmatrix}^{-1}$$

in the partitioned form.

By the partitioned calculus of matrices, we have

$$\bar{g}^{ab} = (g_{ba} - g^{\kappa\lambda} g_{\kappa b} g_{\lambda a})^{-1}, \tag{5.14}$$

where $g^{\kappa\lambda}$ is the inverse of $(n - m) \times (n - m)$ matrix $g_{\lambda\kappa}$. Since $\bar{g}^{ab}/N$ is the covariance matrix of the estimator $\bar{\mathbf{u}}$, we have the following theorem.

THEOREM 6. *A consistent estimator $\bar{\mathbf{u}}$ is Fisher or first-order efficient, if and only if $A(\mathbf{u})$ is orthogonal to S^m at the intersecting point $\theta(\mathbf{u})$. The loss of information Δg_{ab} due to summarizing the data into the estimator $\bar{\mathbf{u}}$ is given asymptotically by*

$$\frac{1}{N} \Delta g_{ab} = g^{\kappa\lambda} g_{\lambda a} g_{\kappa b}. \tag{5.15}$$

PROOF. Since $g_{\lambda a}$ is the inner product of two tangent vectors B^{i}_{a} of S^m and B^{j}_{λ} of $A(\mathbf{u})$, all the $g_{\lambda a}$ vanish when and only when $A(\mathbf{u})$ is orthogonal to S^m. In this case,

$$\frac{1}{N}\bar{g}^{ab} = \frac{1}{N}g^{ab}$$

holds, where g^{ab} is the inverse of the Fisher information matrix g_{ab} of S^m. Hence, the estimator $\bar{\mathbf{u}}$ asymptotically attains the Cramer-Rao bound, so that it is Fisher or first-order efficient. When $A(\mathbf{u})$ is not orthogonal, the inverse of the variance $\bar{g}^{ab}/N$ of the estimator $\bar{\mathbf{u}}$ becomes smaller than the Fisher information Ng_{ab} by $Ng^{\kappa\lambda}g_{\lambda b}g_{\kappa a}$, as is shown by (5.14). Since $\bar{\mathbf{u}}$ is asymptotically normal, it carries the Fisher information $N(\bar{g}^{ab})^{-1}$. Therefore, the loss of information is given by (5.15).

When $A(\mathbf{u})$ is orthogonal to S^m, we have

$$\bar{g}^{ab} = g^{ab}, \qquad \bar{g}^{\kappa\lambda} = g^{\kappa\lambda}. \tag{5.16}$$

Moreover, D^a_i and D^κ_i are obtained as follows:

$$D^a_i = g^{ab}B^j_b g_{ji}, \qquad D^\kappa_i = g^{\kappa\lambda}B^j_\lambda g_{ji}. \tag{5.17}$$

It can also be proved that

$$\bar{u}^a = D^a_i\bar{\theta}^i \quad \text{and} \quad \bar{v}^\kappa = D^\kappa_i\bar{\theta}^i$$

are the orthogonal projections of $\bar{\boldsymbol{\theta}}$ to S^m and $A(\mathbf{u})$, respectively. More precisely,

$$T^j_i = B^j_a D^a_i, \qquad N^j_i = B^j_\kappa D^\kappa_i,$$

are the projection matrices to the tangent space of S^m and to the tangent space of $A(\mathbf{u})$, respectively.

5.3. *Curvature and second-order information loss.* We now calculate the loss of information due to summarizing the data by a first-order efficient estimator $\bar{\mathbf{u}}$. The amount of information which an estimator $\bar{\mathbf{u}}$ carries per sample is written as

$$\bar{g}_{ab}(\mathbf{u}) = \frac{1}{N}E_{\mathbf{u}}[\partial_a\bar{\ell}(\bar{\mathbf{u}}, \mathbf{u})\partial_b\bar{\ell}(\bar{\mathbf{u}}, \mathbf{u})],$$

where $\bar{\ell}(\bar{\mathbf{u}}, \mathbf{u})$ is the logarithm of the probability density function of $\bar{\mathbf{u}}$ with the true parameters $\mathbf{u}$. When the estimator is first-order efficient, $\bar{g}_{ab}(\mathbf{u})$ coincides asymptotically with $g_{ab}(\mathbf{u})$ which is the amount of information carried by one original sample. The total loss of information is hence given by

$$\Delta g_{ab}(\mathbf{u}) = N\{g_{ab}(\mathbf{u}) - \bar{g}_{ab}(\mathbf{u})\}$$

and is of order $O(1)$. The term of this order is called the second-order information loss, and has been calculated by many statisticians (e.g., Rao, 1962; Hosoya, 1979; see also Akahira and Takeuchi, 1974). We give its geometrical presentation, extending the result of Efron (1975).

We use the following lemma.

LEMMA. *The loss of information by taking a statistic* $\mathbf{T}$ *is given by*

$$\Delta g_{ab}(\mathbf{u}) = E_{\mathbf{u}}[\mathrm{Cov}\{\partial_a\ell(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{u})\,|\,\mathbf{T}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \mathbf{T}\}] \tag{5.18}$$

where $\mathrm{Cov}\{\cdot\,|\,\cdot\}$ *is the conditional covariance and E is expectation with respect to the distribution of* $\mathbf{T}$.

By the use of this lemma, we have the following main theorem, which states that the second-order loss of information is decomposed into the sum of two non-negative curvature terms. One term is the square of the exponential curvature tensor of the model S^m,

$$\overset{e}{H}_{ac\kappa} = \overset{e}{H}{}^{i}_{ac}B^j_\kappa g_{ij}. \tag{5.19}$$

Since this denotes the normal components of $\overset{e}{H}{}^{i}_{ac}$, it is a tensor representing the intrinsic curvature of S^m. The other term is the square of the mixture curvature tensor of the estimator,

$$\overset{m}{H}_{\kappa\lambda a} = \overset{m}{H}{}^{i}_{\kappa\lambda} B^{j}_{a} g_{ji}. \tag{5.20}$$

This also denotes the normal components of the curvature of $A(\mathbf{u})$, and hence is a tensor representing the intrinsic curvature of $A(\mathbf{u})$. Since the α-estimator $\overset{\alpha}{\mathbf{u}}$ has the mixture curvature

$$\overset{m}{H}_{\kappa\lambda a} = \frac{1+\alpha}{2} T_{\kappa\lambda a} = \frac{1+\alpha}{2} T_{ijk} B^{i}_{\kappa} B^{j}_{\lambda} B^{k}_{a},$$

the curvature term vanishes for $\alpha = -1$, i.e., for the MLE. The MLE thus has the least second-order information loss.

THEOREM 7. *The second-order information loss of a first-order efficient estimator is given by*

$$\Delta g_{ab} = \overset{e}{H}_{ac\kappa} \overset{e}{H}_{bd\lambda} g^{\kappa\lambda} g^{cd} + \frac{1}{2} \overset{m}{H}_{\nu\mu a} \overset{m}{H}_{\kappa\lambda b} g^{\nu\kappa} g^{\mu\lambda}, \tag{5.21}$$

which becomes

$$\Delta g_{ab} = \overset{e}{H}_{ac\kappa} \overset{e}{H}_{bd\lambda} g^{\kappa\lambda} g^{cd} + \frac{(1+\alpha)^2}{8} T_{\nu\mu a} T_{\kappa\lambda b} g^{\nu\kappa} g^{\mu\lambda}$$

for the α-estimator $\overset{\alpha}{\mathbf{u}}$.

PROOF. We have

$$\partial_a \ell(\bar{\mathbf{x}}, 0) = B^{i}_{a} \bar{x}_i, \tag{5.22}$$

where we assume that $\mathbf{u} = 0$ is the true parameter and $\boldsymbol{\theta}(0) = \boldsymbol{\eta}(0) = 0$, and all the quantities such as B^{i}_{a} are evaluated at $\mathbf{u} = 0$. We represent $\partial_a \ell$ in terms of $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$, neglecting the terms of $O(|\bar{\mathbf{x}}|^3)$. We have by expansion

$$\bar{x}_i = \partial_i \psi(\bar{\boldsymbol{\theta}}) = g_{ij} \bar{\theta}^j + \tfrac{1}{2} T_{ijk} \bar{\theta}^j \bar{\theta}^k, \tag{5.23}$$

and

$$\bar{\theta}^i = B^{i}_{a} \bar{u}^a + B^{i}_{\kappa} \bar{v}^\kappa + \tfrac{1}{2} \partial_a B^{i}_{b} \bar{u}^a \bar{u}^b + \tfrac{1}{2} \partial_\kappa B^{i}_{\lambda} \bar{v}^\kappa \bar{v}^\lambda + \partial_a B^{i}_{\kappa} \bar{u}^a \bar{v}^\kappa. \tag{5.24}$$

By substituting (5.23) and (5.24) in (5.22), we have

$$\partial_a \ell = g_{ab} \bar{u}^b + \tfrac{1}{2} B^{i}_{a} (T_{bci} + g_{ij} \partial_b B^{j}_{c}) \bar{u}^b \bar{u}^c + \tfrac{1}{2} B^{i}_{a} (T_{\kappa\lambda i} + g_{ji} \partial_\kappa B^{j}_{\lambda}) \bar{v}^\kappa \bar{v}^\lambda + B^{i}_{a} (g_{ij} \partial_b B^{j}_{\kappa} + T_{\kappa b i}) \bar{u}^b \bar{v}^\kappa,$$

where the orthogonality condition

$$B^{i}_{a}(\mathbf{u}) B^{j}_{\kappa}(\mathbf{u}) g_{ij}(\mathbf{u}) = 0 \tag{5.25}$$

is taken into account and $T_{\kappa b i} = B^{k}_{\kappa} B^{j}_{b} T_{kji}$, etc. We have from (4.6)

$$B^{i}_{a} (T_{bci} + g_{ji} \partial_b B^{j}_{c}) = \overset{m}{\Gamma}_{bca},$$

and from (5.6)

$$B^{i}_{a} (T_{\kappa\lambda i} + g_{ji} \partial_\kappa B^{j}_{\lambda}) = \overset{m}{H}_{\kappa\lambda a}.$$

By differentiating the orthogonality condition (5.25) with respect to u^b, we have

$$\overset{e}{H}_{ba\kappa} = (\partial_b B^{i}_{a}) B^{j}_{\kappa} g_{ji} = -B^{i}_{a} (T_{\kappa b i} + g_{ij} \partial_b B^{j}_{\kappa}).$$

Hence, we obtain the important relation

$$\partial_a \ell = g_{ab}\bar{u}^b + \tfrac{1}{2} \overset{m}{\Gamma}_{bca}\bar{u}^b\bar{u}^c + \tfrac{1}{2} \overset{m}{H}_{\kappa\lambda a}\bar{v}^\kappa\bar{v}^\lambda - \overset{e}{H}_{ab\kappa}\bar{u}^b\bar{v}^\kappa. \tag{5.26}$$

Next, we calculate the conditional covariance of $\partial_a \ell$. Since $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ are asymptotically normal, we have

$$E(\bar{v}^\kappa\bar{v}^\lambda) = \frac{1}{N} g^{\kappa\lambda},$$

$$E(\bar{v}^\kappa\bar{v}^\lambda\bar{v}^\nu\bar{v}^\mu) = \frac{1}{N^2}(g^{\kappa\lambda}g^{\nu\mu} + g^{\kappa\nu}g^{\lambda\mu} + g^{\kappa\mu}g^{\lambda\nu}),$$

$$E(\bar{u}^a\bar{u}^b) = \frac{1}{N} g^{ab},$$

and the expectations of $\bar{u}^a\bar{v}^\kappa$ as well as of $\bar{v}^\kappa\bar{v}^\lambda\bar{v}^\nu$ vanish to within the higher order terms. By substituting these in (5.18), and taking (4.17) into account, we have (5.21), which proves the theorem.

EXAMPLE 5.1. The second-order information loss of the α-estimator in Example 4.1 is given by

$$\Delta g_{ab} = \frac{1}{8(2a^2+1)^2u^2}\{16a^2 + (1+\alpha)^2(4a^2-3)^2\},$$

where the first term is due to the exponential curvature of the model and the second term is due to the mixture curvature of the α-estimator. The latter vanishes for $\alpha = -1$ or MLE.

5.4. *Recovery of information loss.* We have shown the second order information loss of an efficient estimator $\bar{\mathbf{u}}$. We need some additional statistics to recover the second order loss. Such statistics are derived directly from (5.26). Let

$$A_{ab} = \overset{e}{H}{}^{\kappa}_{ab}(\bar{\mathbf{u}})B^i_\kappa(\bar{\mathbf{u}})\bar{x}_i, \qquad B_a = \tfrac{1}{2}\overset{m}{H}{}^{\kappa\lambda}_{a}(\bar{\mathbf{u}})B^i_\kappa(\bar{\mathbf{u}})B^j_\lambda(\bar{\mathbf{u}})\bar{x}_i\bar{x}_j$$

be two statistic tensors. Then, we have from (5.26)

$$\partial_a \ell(\bar{\mathbf{x}}, \mathbf{u}) = g_{ab}\bar{u}^b + \tfrac{1}{2}\overset{m}{\Gamma}_{bca}\bar{u}^b\bar{u}^c + B_a - A_{ab}\bar{u}^b + O(|\bar{\mathbf{x}}|^3). \tag{5.27}$$

Hence, the expectation of the conditional covariance of $\partial_a \ell(\bar{\mathbf{x}}, \mathbf{u})$ on the condition $\bar{\mathbf{u}}$, A_{ab} and B_a is of the order $1/N^3$. This proves the following theorem.

THEOREM 8. *The second-order loss of information for a first-order efficient estimator* $\bar{\mathbf{u}}$ *is recovered by the statistics* A_{ab} *and* B_a. *That is,*

$$\lim_{N\to\infty} \Delta g_{ab} = 0 \tag{5.28}$$

for the statistics $\bar{\mathbf{u}}$, A_{ab} *and* B_a.

The tensor A_{ab} is related to the exponential curvature of S^m. The vector B_a is related to the mixture curvature of $A(\mathbf{u})$, and vanishes for MLE. These terms are asymptotically ancillary. Since A_{ab} and B_a together have $m(m+3)/2$ components, information is summarized in $\bar{\mathbf{u}}$, A_{ab} and B_a only when $m(m+3)/2 < n - m$. Otherwise, $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ having n components carry sufficient information.

It is another important and interesting problem to see how to use these statistics to get better inference (see e.g., Efron, 1978; Efron and Hinkley, 1978; Hosoya, 1979).

5.5. *The second-order covariance of efficient estimator.* Efron (1975) pointed out that the second order term of the covariance of a first-order efficient estimator includes the square of the statistical curvature. The result was extended to the multi-parameter case by

Madsen (1979), where she used the multi-dimensional exponential curvature tensor. Here we show the result in our framework, which will be treated elsewhere in more detail.

THEOREM 9. *The covariance of an efficient estimator* $\hat{\mathbf{u}}$ *is given by*

$$\begin{aligned}\operatorname{Cov}(\hat{u}^a, \hat{u}^b) = \frac{1}{N} g^{ab} + \frac{1}{2N^2} (\overset{m}{\Gamma}{}^a_{cd}\overset{m}{\Gamma}{}^b_{ef} g^{ce} g^{df} + \overset{m}{H}{}^a_{\kappa\lambda}\overset{m}{H}{}^b_{\nu\mu} g^{\kappa\nu} g^{\lambda\mu} \\ + 2\overset{e}{H}{}^a_{c\kappa}\overset{e}{H}{}^b_{d\lambda} g^{cd} g^{\kappa\lambda} - g^{ab}\partial_c b^b - g^{bc}\partial_c b^a) + O(N^{-3}),\end{aligned} \tag{5.29}$$

where

$$\frac{b^a}{N} = E(\hat{u}^a) = -\frac{1}{2}(\overset{m}{\Gamma}{}^a_{bc} g^{bc} + \overset{m}{H}{}^a_{\kappa\lambda} g^{\kappa\lambda})$$

is the bias of the estimator, which can be eliminated by modifying $\hat{u}^a$ *into*

$$\hat{u}^{*a} = \hat{\mathrm{u}}^a - \frac{b^a(\hat{\mathbf{u}})}{N}.$$

It should be noted that the second-order squared error of $\hat{\mathbf{u}}^*$ can be decomposed into the following positive semidefinite terms,

$$\frac{1}{N^2}\left(\frac{1}{2}\cdot\overset{m}{\Gamma}{}^a_{cd}\overset{m}{\Gamma}{}_{ef} g^{ce} g^{df} + \Delta g_{cd} g^{ca} g^{db}\right).$$

The first term is the square of the mixture connection, which depends on the parametrization of S^m. This sometimes called the naming curvature. There exists a parametrization by which this term vanishes identically, if and only if the Riemann-Christoffel mixture curvature tensor $\overset{m}{R}_{abcd}$ vanishes. The term Δg_{cd} is indeed the second-order information loss, which is the sum of the squares of the exponential curvature of the subspace S^m and of the mixture curvature of the associated ancillary subspace $A(\hat{\mathbf{u}})$. The latter vanishes for the MLE.

Conclusion. We have given a differential-geometrical framework for describing statistical problems by introducing the Fisher information metric and the one-parameter family of α-connections. It is shown that the α-curvatures play a fundamental ·ole in the asymptotic theory of estimation. The second-order information loss and covariance of an efficient estimator are given in terms of geometrical quantities. The meaning of the α-connection is elucidated further by introducing the α-quasi-distances in the function space of statistical distributions (see the Appendix).

There remain many important problems to be analyzed in the geometrical framework. They are, for example, higher-order efficiency, conditional estimation and ancillary statistics, asymptotic theory of testing hypotheses, robust estimation, and fitness of statistical models (cf. Akaike, 1974). Moreover, our geometrical theory should be extended to the function space. We have already obtained some results, which will be published in forthcoming papers (Amari, 1982, Amari and Kumon, 1981, Kumon and Amari, 1981a, Kumon and Amari, 1981b).

Appendix. *α-distance and α-connection in the function space of distributions.*

A1. *Function space of distributions.* Let us consider the set of all the smooth density functions $p(\mathbf{x})$ of a random variable $\mathbf{x} \in X$ with respect to some carrier measure P on X, with $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in X$. Let us put

$$\ell(\mathbf{x}) = \log p(\mathbf{x}). \tag{A.1}$$

Let S be the space consisting of all such densities.

Let us consider a smooth curve $p(\mathbf{x}, t)$ in S and put

$$\ell(\mathbf{x}, t) = \log p(\mathbf{x}, t).$$

Then,

$$\dot{\ell}(\mathbf{x}, t) = \frac{d}{dt}\ell(\mathbf{x}, t) \tag{A.2}$$

can be considered as a tangent vector of the curve at $p(\mathbf{x}, t)$. This satisfies

$$E_t[\dot{\ell}(\mathbf{x}, t)] = 0,$$

where E_t is the expectation at $p(\mathbf{x}, t)$.

The tangent space T_p of S at $p(\mathbf{x})$ thus consists of smooth random variables $a(\mathbf{x})$ with

$$E_p\{a(\mathbf{x})\} = 0, \tag{A.3}$$

$$T_p = \{a(\mathbf{x}) \mid E_p[a(\mathbf{x})] = 0\} \tag{A.4}$$

(cf. Dawid, 1975). To be mathematically rigorous, we need some subsidiary conditions, such as the boundedness of $a(\mathbf{x})p(\mathbf{x})$, the finiteness of $E_p[\{\dot{a}(\mathbf{x})\}^2]$, etc., which we do not discuss here. A tangent vector is a linear mapping from the set of smooth functionals F to the set of real numbers, satisfying the Leipnitz law. For $f \in F$, $\dot{\ell}(\mathbf{x}, t) \in T_{p_t}$, we have $\dot{\ell}f = (d/dt)f(p(\mathbf{x}, t))$. In general, we can write, for $a \in T_p$, $af = (Df)a$, where Df is the Fréchet derivative.

We can introduce the inner product in T_p by

$$a \cdot b = E_p\{a(\mathbf{x})b(\mathbf{x})\}. \tag{A.5}$$

This is the information metric, because

$$\|\dot{\ell}\|^2 = E_p(\dot{\ell}^2)$$

is the Fisher information for the one-parameter family of distributions $p(\mathbf{x}, t)$.

Let $a(\mathbf{x}, t)$ be a family of tangent vectors defined on the curve $p(\mathbf{x}, t)$. When $a(\mathbf{x}, t)$ is the solution of the equation

$$\dot{a} + \frac{1-\alpha}{2}a\dot{\ell} + \frac{1+\alpha}{2}E_t(a\dot{\ell}) = 0, \tag{A.6}$$

where α is a parameter, we call $a(\mathbf{x}, t)$ the parallel displacement of a with respect to the α-connection along the curve. This defines the α-connection in S. By substituting (A6) in $(d/dt)E_t\{a(\mathbf{x}, t)\}$, we have

$$\frac{d}{dt}E_t\{a(\mathbf{x}, t)\} = E_t(a\dot{\ell}) + E_t(\dot{a}) = 0,$$

which shows that the paralleled displacement $a(\mathbf{x}, t)$ indeed belongs to $T_{p(\mathbf{x},t)}$.

When the tangent vectors $\dot{\ell}$ of a curve are by themselves the parallel displacements along the curve with respect to the α-connection, the curve is called an α-geodesic. The equation of the α-geodesic is

$$\ddot{\ell} + \frac{1-\alpha}{2}\dot{\ell}^2 + \frac{1+\alpha}{2}i(t) = 0, \tag{A.7}$$

where

$$i(t) = E_t(\dot{\ell}^2).$$

The ($\alpha = 1$)-geodesic connects two distributions by a one-parameter exponential family, while the ($\alpha = -1$)-geodesic connects two distributions by a one-parameter mixture family.

A2. *Quasi-α-distance.* Let $k(u)$ be a smooth function satisfying

$$k(1) = 0, \qquad k''(u) > 0, \qquad k''(1) = 2. \tag{A.8}$$

Then, for two points $p(\mathbf{x}),\ q(\mathbf{x}) \in S$,

$$D_k(p, q) = E_p\left[k\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)\right] \tag{A.9}$$

satisfies

$$D_k(p, q) \geq 0,$$

where the equality holds when and only when $p = q$. We call $D_k(p, q)$ the quasi-k-distance from p to q. This does not satisfy the triangular inequality and is not symmetric in general. However, this is an extension of the Fisher information metric, because we have

$$D_k\{p(\mathbf{x}, 0), p(\mathbf{x}, t)\} = i(t)t^2 + O(t^3),$$

for small t.

Let us define the following one-parameter family of functions

$$k_\alpha(u) = \begin{cases} 2u \log u, & \alpha = 1, \\ \dfrac{8}{1-\alpha^2}(1 - u^{(1+\alpha)/2}), & \alpha \neq \pm 1, \\ -2 \log u, & \alpha = -1. \end{cases} \tag{A.10}$$

They satisfy the requirements in (A.8). (They satisfy the single equation $k''_\alpha(u) = 2u^{(\alpha - 3)/2}$.)

We call

$$D_\alpha(p, q) = D_{k_\alpha}(p, q) \tag{A.11}$$

the (quasi)-α-distance from p to q. The $(\alpha = -1)$-distance gives twice the Kullback-Leibler information, and the $(\alpha = 0)$- distance is four-times the Hellinger distance, which is connected to the true Riemannian distance (cf. Dawid, 1977). The α-distance $(\alpha \neq \pm 1)$ is related to the Chernoff distance (Chernoff, 1952).

The following duality holds for the α-distances.

$$D_\alpha(p, q) = D_{-\alpha}(q, p), \tag{A.12}$$

so that the α-distance from p to q is equal to the $(-\alpha)$- distance from q to p.

A3. *Ancillary subspaces for the α-connection.* Let us consider a statistical model S^n specified by n parameters $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)$, $S^n = \{p(\mathbf{x}, \boldsymbol{\theta})\}$. We assume that S^n forms an n-dimensional manifold in S, and the Fisher information $g_{ij}(\boldsymbol{\theta})$ is well-defined and positive-definite in it.

For a point $q \in S$, let $p(\mathbf{x}, \hat{\boldsymbol{\theta}})$ be the point which minimizes the α-distance from q to S^n, i.e.,

$$\min_{S^n} D_\alpha\{q, p(\mathbf{x}, \boldsymbol{\theta})\} = D_\alpha\{q, p(\mathbf{x}, \hat{\boldsymbol{\theta}})\}. \tag{A.13}$$

We call $p(\mathbf{x}, \hat{\boldsymbol{\theta}})$ the α-approximation of q by S^n. We assume that there exists a neighborhood U of S^n, in which every q has a unique α-approximation in S^n. Let M_α be the mapping from U to S^n, which gives the α-approximation.

Let us call the subset of S of which every point has the same α-approximation $p(\mathbf{x}, \boldsymbol{\theta})$, i.e.,

$$A_\alpha(\boldsymbol{\theta}) = M_\alpha^{-1}(\boldsymbol{\theta}) = \{q \mid M_\alpha q = p(\mathbf{x}, \boldsymbol{\theta}),\ q \in U\}, \tag{A.14}$$

the α-ancillary subspace of $p(\mathbf{x}, \boldsymbol{\theta}) \in S^n$. We then have the following theorem, which elucidates the meaning and role of the α-connections.

THEOREM. *The α-ancillary subspace of $p(\mathbf{x}, \boldsymbol{\theta}) \in S^n$ consists of all the points on the α-geodesics which pass through the point $p(\mathbf{x}, \boldsymbol{\theta})$ and are orthogonal to S^n at that point.*

Proof Outline. Let $q(\mathbf{x}, t)$ be an α-geodesic passing through a point $p(\mathbf{x}, \boldsymbol{\theta})$,

$$q(\mathbf{x}, 0) = p(\mathbf{x}, \boldsymbol{\theta}) \tag{A.15}$$

and orthogonal to S^n at that point,

$$E_0\{\partial_i \ell(\mathbf{x}, \boldsymbol{\theta}) \frac{d}{dt} \log q(\mathbf{x}, t)|_{t=0}\} = 0. \tag{A.16}$$

We first prove that

$$M_\alpha q(\mathbf{x}, t) = p(\mathbf{x}, \boldsymbol{\theta}).$$

To this end, we put

$$R_i(t) = \partial_i D_\alpha\{q(\mathbf{x}, t), p(\mathbf{x}, \boldsymbol{\theta})\}, \tag{A.17}$$

where $\partial_i = \partial/\partial\theta^i$. The conditions (A.15) and (A.16) imply

$$R_i(0) = 0 \quad \text{and} \quad \dot{R}_i(0) = 0, \tag{A.18}$$

respectively. Moreover, by differentiation of (A.17), we have, after a little complicated calculation, the equation

$$\ddot{R}_i(t) = -\frac{1-\alpha^2}{4} i(t) R(t). \tag{A.19}$$

By solving this with the above initial conditions, we have $R_i(t) = 0$, from which we can prove that the α-distance from $g(\mathbf{x}, t)$ to S^n is minimum at $p(\mathbf{x}, \boldsymbol{\theta})$.

Conversely, let $q(\mathbf{x})$ be a point which satisfies

$$M_\alpha q = p(x, \boldsymbol{\theta}).$$

Let $q(\mathbf{x}, t)$ be the α-geodesic connecting $q(\mathbf{x})$ and $p(\mathbf{x}, \boldsymbol{\theta})$, with $q(\mathbf{x}, 0) = p(\mathbf{x}, \boldsymbol{\theta})$ and $q(\mathbf{x}, \tau) = q(\mathbf{x})$. Then, $R_i(t)$ defined by (A.17) for this $q(\mathbf{x}, t)$ satisfies (A.19) with $R_i(0) = 0$, and $R_i(\tau) = 0$, from which we have $\dot{R}_i(0) = 0$ or (A.16). This implies that the α-geodesic is orthogonal to S^n, so that $q(\mathbf{x})$ is on an orthogonal α-geodesic.

Acknowledgement. The author would like to express his sincere gratitude to Professor Kazuo Kondo, who introduced him to the world of applied geometry and led him with penetrating spirits. He also thanks Professor S. Moriguti for his suggestion of the importance of geometrical considerations in statistical problems. The author is indebted to Professor K. Takeuchi and Dr. L. T. Madsen Skovgaard for many useful discussions. The author thanks one of the reviewers, who not only sent the author some personal translations of Chentsov's book, but also gave him many useful comments on the manuscript. He also expresses his thanks to Professor M. Akahira, Professor T. Kitagawa, Professor T. Okuno, Professor T. S. Han, Professor B. Efron and Professor A. P. Dawid for their encouragements. Thanks are also due to Mr. K. Hayami and Mr. M. Kumon for comments on the manuscript and to Mrs. T. Shintani for her cordiality in typing.

REFERENCES

Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency.* Lecture Notes in Statistics No. 7, Springer, Berlin.

Akaike, H. (1974). A new look at the statistical model indentification. *IEEE Trans. Automat. Control* **19** 716–723.

Amari, S. (1968). Theory of information spaces—a geometrical foundation of the analysis of communication systems. *RAAG Memoirs* **4** 373–418.

Amari, S. (1982). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* **69** 1–17.

Amari, S. and Kumon, M. (1981). Differential geometry of Edgeworth expansions in curved exponential family. Technical Report, METR 81-7, University of Tokyo.

ATKINSON, C. and MITCHELL, A. F. S. (1981). Rao's distance measure. *Sankhyā A* **43** (to appear).

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

CHENTSOV, N. N. (1966). A systematic theory of exponential families of probability distributions. *Theor. Probability Appl.* **11** 425–435.

CHENTSOV, N. N. (1972). *Statistical Decision Rules and Optimal Conclusions.* Nauka, Moscow (in Russian).

CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Ann. Math. Statist.* **23** 493–507.

DAWID, A. P. (1975). *Discussions to Efron's paper. Ann. Statist.* **3** 1231–1234.

DAWID, A. P. (1977). Further comments on a paper by Bradley Efron. *Ann. Statist.* **5** 1249.

EFRON, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency) (with Discussion). *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with Discussion) *Biometrika* **65** 457–487.

EISENHART, L. P. (1927). *Non-Riemannian Geometry.* Am. Coll. Publ. **8**.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **122** 700–725.

HOSOYA, Y. (1979). Conditionality and maximum-likelihood estimation. Paper presented at Int. Con. Statist. held in Tokyo.

INGARDEN, R. S., SATO, Y., SUGAWA, K. and KAWAGUCHI, M. (1979). Information thermodynamics and differential geometry. *Tensor, n.s.* **33** 347–353.

INGARDEN, R. S. (1981). Information geometry in function spaces of classical and quantum finite statistical systems. *Internat. J. Engrg. Sci.* **19** 1609–1633.

KULLBACK, S. L. (1959). *Information Theory and Statistics.* Wiley, New York.

KUMON, M. and AMARI, S. (1981a). Geometrical theory of higher-order asymptotically most powerful two-sided test. Technical Report, METR 81-8, University of Tokyo.

KUMON, M. and AMARI, S. (1981b). Geometry of interval estimation: A higher-order asymptotic theory. Technical Report, METR **81-9**, University of Tokyo.

MADSEN, L. T. (1979). The geometry of statistical model—a generalization of curvature. *Research. Report.* **79-1**, Statist. Res. Unit., Danish Medical Res. Council.

OZEKI, K. (1977). A Riemannian metric structure on autoregressive parameter space (in Japanese). Report S77-04, Acoustic Society of Japan.

PFANZAGL, J. (1973). Asymptotic expansions related to minimum contrast estimators. *Ann. Statist.* **1** 993–1026.

RAO, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta. Math. Soc.* **37** 81–91.

RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. B* **24** 46–72.

REEDS, J. (1975). Discussions to Efron's paper. *Ann. Statist.* **3** 1234–1238.

SATO, Y., SUGAWA, K. and KAWAGUCHI, M. (1979). The geometrical structure of the parameter space of two-dimensional normal distribution. *Rep. Math. Phys.* **16** 111–119.

SCHOUTEN, J. A. (1954). *Ricci-Calculus, 2nd ed.*, Springer, Berlin.

TAKIYAMA, R. (1974). On geometrical structures of parameter spaces of one-dimensional distributions (in Japanese). *Trans. Inst. Electr. Comm. Eng. Japan* **57-A** 67–69.

YOSHIZAWA, T. (1971). A geometrical interpretation of location and scale parameters. Memo TYH-2, Harvard Univ.

UNIVERSITY OF TOKYO
DEPARMENT OF MATHEMATICAL ENGINEERING AND
INSTRUMENTATION PHYSICS
BUNKYO-KU
TOKYO, JAPAN