SUT Journal of Mathematics
Vol. 35, No. 1 (1999), 113–137

Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations

Shintaro Yoshizawa and Kunio Tanabe

(Received April 16, 1999)

Abstract. Amari showed that the geometry of a family of probability distributions is characterized by a dual differential geometry determined by a couple of affine connections and a divergence associated with a couple of dual potential functions. In this paper, a 2-parameter class of dual differential geometries is constructed on the manifold of the family of multivariate Gaussian distributions with nonzero means, as well as a new class of divergences. This class of geometry includes the Riemannian geometry studied by Skovgaard and the geometry associated with the Kullback-Leibler information. The specific dually flat charts for the latter geometry is given in conjunction with a detailed analysis of the associated connections. In order to facilitate the various calculations of differential geometric quantities in the analysis of our geometries, we introduce a new coordinate free differential calculus of a function of a symmetric matrix argument, based on a specific bilinear form defined on the domain of the function and its dual space. This calculus enables us to obtain a parallel formalism of the Legendre transformation in convex analysis even for a function of a matrix argument.

AMS 1991 *Mathematics Subject Classification.* 15A48, 26B25, 31C12, 53C05, 62B10, 62H99.

Key words and phrases. dual affine connections, Legendre transformation, matrix differential, matrix convex function, Dunford-Taylor integral, (β, γ)-divergence, Kullback-Leibler information.

§1. Introduction

Efron [8] introduced the concept of *statistical curvature* of a one-parameter family of distributions for investigating statistical characteristics of the family. Dawid [8, p.1231] interpreted it in the language of differential geometry to

introduce three kinds of affine connections for use in statistics. Amari [2] defined a one-parameter class of affine connections, which includes the three connections as special cases. He noted the importance of a pair of *dual affine connections* (∇, ∇^*) for understanding the differential geometry of a family of probability distributions. If a Riemannian manifold (M, g) is *flat* with respect to a pair of torsion-free dual affine connections, then there exists a pair of dual coordinate systems which is associated with *dual potential functions* via a *Legendre transformation* [2,p.80]. The *dual potential functions* lead to a *divergence* which forms a measure of discrepancy between two distributions. The similar geometry was also studied, for example, by Calabi [6], Chen-Yau [7] and Shima [15] from a view point of affine differential geometry.

Amari [2] investigated the dual differential geometry of a family $\left\{N(\mu, \sigma^2)\right\}$ of the univariate Gaussian distributions and gave closed form expressions for the *dually flat coordinates* and *dual potential functions*. Ohara-Suda-Amari [13] studied the same geometry of a family $\{N(0, \Sigma)\}$ of the multivariate Gaussian distributions with *zero means* and obtained the dual potential functions

$$\psi_0 = \frac{1}{2}\log(\det \Sigma) + \frac{n}{2}\log(2\pi), \tag{1}$$

$$\phi_0 = -\frac{1}{2}\log(\det \Sigma) - \frac{n}{2}\log(2\pi e). \tag{2}$$

We are concerned in this paper with the family $\{N(\mu, \Sigma)\}$ of all the multivariate Gaussian distributions whose *means are not necessary zero*. Contrary to the general expectation [2,pp.84-88], the dual differential geometry conceivable on this family of Gaussian distributions is not unique, because there are arbitrariness in modeling the interplay of μ- and Σ-coordinates. In fact, we construct, in Section 3, a 2-parameter class of dual differential geometries on this family of distributions, each of which is identified by a pair of charts, a bilinear form defined on the charts, a pair of potential functions,

$$\psi = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log(\det \Sigma) + \frac{n}{2}\log(2\pi), \tag{3}$$

and

$$\phi_{\beta,\gamma} = \frac{1}{2}(1 - \frac{\beta}{\gamma})\mu^T\Sigma^{-1}\mu - \frac{1}{2}\log(\det \Sigma) - \frac{n}{2}\log(2\pi e^{\gamma^{-1}}). \tag{4}$$

and a corresponding divergence. In particular, we define a new set of divergences which includes the Kullback-Leibler information. Both potential functions contain additional term, $\mu^T\Sigma^{-1}\mu$, which determines a dual differential geometry specifying the interplay of μ- and Σ- coordinates. The functions ϕ_0 and $\phi_{\beta,\gamma}$, ($0 \leq \beta,\ 0 < \gamma,\ 0 \leq \frac{\beta}{\gamma} \leq 1$), are convex in μ and Σ and the

function ψ_0 is concave in Σ, but the function ψ is neither convex nor concave with respect to (μ, Σ). This 2-parameter class of geometry includes both the Riemannian geometry studied by Skovgaard [16] and the geometry associated with the Kullback-Leibler information quantity which we deal with in Section 4. While the divergence induced by the former geometry is a metric function, the divergence induced by the latter geometry is a pseudo-metric function. Every geometry in this class induces a relative geometry on the subfamily of multivariate Gaussian distributions *with zero means* . All the geometries thus induced, however, are essentially identical to the unique geometry on the subfamily treated by Ohara-Suda-Amari [13].

One might suspect that it would be rather trivial from the arguments of Amari [2] and Fujiwara-Amari [9] to obtain the dual differential geometry on the distributions with *nonzero means. However, on top of the above mentioned lack of uniqueness of the geometry on* $\{N(\mu, \Sigma)\}$ *(See Section 3), a direct application of Amari [2,pp.80-88] and Fujiwara-Amari [9,p.318] leads to the divergence which doesn't conform to the Kullback-Leibler information* because of the definition of their bilinear form $\sum_{i \leq j} \Theta_{ij} H_{ij}$ defined for two symmtric matrices $\Theta = \{\Theta_{ij}\}, H = \{H_{ij}\}$. In order to give a *dually flat structure which is associated with the Kullback-Leibler information* , we must choose an appropriate bilinear form and define a pair of the dual potential functions by

$$\psi = \frac{1}{2}\mu^T \Sigma^{-1} \mu + \frac{1}{2}\log(\det \Sigma) + \frac{n}{2}\log(2\pi),$$

and

$$\phi_{1,1} = -\frac{1}{2}\log(\det \Sigma) - \frac{n}{2}\log(2\pi e). \tag{5}$$

In this specific case, our dual potential function $\phi_{1,1}$ reduces to the same potential function (2) which does not contain the component, $\mu^T \Sigma^{-1} \mu$.

In Section 2, we develop a new differential calculus of a function with a symmetric matrix argument for facilitating the analysis of the dual differential geometry. In Section 3, we construct a 2-parameter class of dual differential geometries of the Gaussian distributions $\{N(\mu, \Sigma)\}$ with non zero means and derive the associated set of new divergences between two probability distributions. In Section 4, we single out one differential geometry which is associated with the *Kullback-Leibler information* and analyze it as a typical geometry in the 2-parameter class.

We use the following notations.

R^n: Vector space of all the n dimensional column vectors

I: Identity matrix of appropriate size

X^T: Transpose of the matrix X

$Tr(X)$: Trace of the matrix X

$M(n,R)$: Vector space of all the n by n real matrices

$\mathfrak{S}_n$: Vector space of all the n by n real symmetric matrices

$\mathfrak{S}_n^+(\mathfrak{S}_n^-)$: Cone of all the positive (negative) definite matrices in $\mathfrak{S}_n$

$N(\mu,\Sigma)$: The Gaussian distribution with mean $\mu \in R^n$ and covariance Σ

$p(x;\mu,\Sigma)$: Density function of $N(\mu,\Sigma)$ with respect to the Lebesgue measure

$\mathfrak{N}$: Set $\{N(\mu,\Sigma)\}$ of all the multivariate Gaussian distributions

§2. A Differential Calculus of a Function with a Symmetric Matrix Argument

In order to manipulate the differential geometric quantities in the following sections, we need to differentiate functions with respect to a *symmetric* matrix argument. There have been extensive literatures [1,12,13,14] on various differential calculi of functions with a positive definite matrix argument. For example, Rao [14] has the formula

$$\frac{\partial}{\partial X_s}\log\det(X_s) = 2X_s^{-1} - \mathrm{diag}(X_s^{-1}) \tag{6}$$

for the derivative, where X_s is a symmetric and positive definite matrix and $\mathrm{diag}(X_s^{-1})$ is the diagonal matrix of diagonal elements of X_s^{-1}. Anderson-Olkin [1] has the formula

$$\{d\det(X_s)\}_{ij} = (2-\delta_{ij})X_{ij}^c dX_{ij} \tag{7}$$

for the differential, where X_s is a symmetric and positive definite matrix, $X_s \equiv \{X_{ij}\}$ and X_{ij}^c is the cofactor of X_{ij} and δ_{ij} is Kronecker's delta. These calculi which are essentially obtained through ad hoc examination of individual matrix elements are inappropriate for us to construct a dual differential geometric structure on the family of multivariate Gaussian distributions with nonzero means. Their calculi, in fact, are closely associated with the inner product $\sum_{i\le j}\Theta_{ij}H_{ij}$, which is not suitable for accommodating the space of positive definite matrices in our analysis. In order to faciliate the construction of the dual differential geometry which is consistent with the Kullback-Leibler information, we adopt the same bilinear form $<\Theta, H> \equiv Tr(\Theta H^T)$ for both symmetric and general matrices.

Definition 2.1. Let $\mathfrak{D}$ and $\mathfrak{D}^*$ be subspaces of $R^n \times M(n,R)$. We define a bilinear form $<,>$ on $\mathfrak{D}$ and $\mathfrak{D}^*$ by

$$<\tilde{X},\tilde{Y}> \equiv Tr(xy^T + XY^T). \tag{8}$$

where $\tilde{X} = (x,X) \in \mathfrak{D}$ and $\tilde{Y} = (y,Y) \in \mathfrak{D}^*$.

When the matrices are symmetric, this bilinear form amounts to the sum of a weighted sum of products of independent matrix elements and the usual sum of the products of vector elements. For use in statistics, Rao [14] defined two different kinds of differentiations, which include the formulas (6) and

$$\frac{\partial}{\partial X}\log(\det X) = X^{-1}, \tag{9}$$

respectively for the function with symmetric matrix argument and with general matrix argument. We would like to emphasize that there is no need to make this distinction and we should use essentially *the same differetial calculus* shown below for both cases. Taking advantage of the induced bilinear form on the subspaces of symmetric matrices, we introduce the notion of differentials and derivatives with respect to a symmetric matrix infinitesimal increment without recourse to the usual elementwise differentiations.

Definition 2.2 (differentials). Let $f(\tilde{X})$ be a real valued function defined on a smoothly imbeded submanifold $\mathfrak{O}$ of $R^n \times M(n,R)$. When, for any $\delta\tilde{X}$ such that $\tilde{X} + \delta\tilde{X} \in \mathfrak{O}$, the difference

$$\delta f = f(x + \delta x, X + \delta X) - f(x, X)$$

has an asymptotic expansion of the form

$$\delta f = \delta^{(1)} f + o(\|\delta\tilde{X}\|), \tag{10}$$

where $\delta^{(1)} f$ is a linear function of $\delta\tilde{X}$ and $\|\delta\tilde{X}\|$ is the norm defined by the bilinear form (8), then $\delta^{(1)} f$ is called the first order differential of f with respect to the manifold $\mathfrak{O}$.

Definition 2.3 (partial derivatives). We define *partial derivatives* $\partial_x f$ and $\partial_X f$ of f with respect to the manifold $\mathfrak{O}$ by

$$\delta^{(1)} f = < (\partial_x f, \partial_X f), (\delta x, \delta X) >,$$

where $\partial_x f \in R^n$ and $\partial_X f \in M(n,R)$.

Note that our derivatives depend heavily on the choice of bilinear form given in (8). We can derive Rao's derivatives such as (6) by replacing $Tr(XY^T)$ with $\sum_{i \leqq j} X_{ij} Y_{ij}$ in our definition.

Definition 2.4 (symmetric partial derivative). When infinitesimal increment δX is a symmetric matrix, $\partial_X f$ is taken to be a symmetric matrix. In this case, the derivative is denoted by $\partial_X^s f$, i.e.,

$$\partial_X^s f \equiv \frac{1}{2}[(\partial_X f) + (\partial_X f)^T].$$

Proposition 2.1. *If f is a function of a symmetric matrix X, then*

$$\partial_X^s f = \partial_X f. \tag{11}$$

In this case, the differential of the function f is given by

$$\delta^{(1)} f \equiv < \partial_X^s f, \delta X > . \tag{12}$$

Proposition 2.2. *Let x be a vector and let X be a non-singular matrix. Then we have*

(i) $\delta^{(1)}(\det X) = (\det X) Tr(X^{-1}\delta X)$,

(ii) $\delta^{(1)} \log(\det X) = Tr(X^{-1}\delta X)$,

(iii) $\delta^{(1)}(x^T X^{-1} x) = Tr(X^{-1}\delta x x^T + X^{-1} x \delta x^T) - Tr(X^{-1} x x^T X^{-1}\delta X)$,

(iv) $\delta^{(1)}[Tr(AX)] = Tr[A(\delta X)]$.

Proof. We give the proofs of each case.

(i) $\delta(\det X) = \det(X + \delta X) - \det X = (\det X)\det(I_n + X^{-1}\delta X) - \det X$
$= (\det X) Tr(X^{-1}\delta X) + O(\|\delta X\|^2)$.

(ii) $\delta \log(\det X) = \log[\det(I_n + X^{-1}\delta X)]$
$= \log[1 + Tr(X^{-1}\delta X) + O(\|\delta X\|^2)]$
$= Tr(X^{-1}\delta X) + O(\|\delta X\|^2)$.

(iii) $\delta(x^T X^{-1} x) = (x + \delta x)(X + \delta X)^{-1}(x + \delta x)^T - x^T X^{-1} x$
$= Tr(X^{-1} dx x^T + X^{-1} x dx^T) - Tr(X^{-1} x x^T X^{-1} dX) + O(\|\delta \tilde{X}\|^2)$.

(iv) This follows from the linearity of Trace map.

□

Corollary 2.3. *Let X be a non-singular matrix. Then we have*

(i) $\partial_X(\det X) = (\det X)(X^{-1})^T$,

(ii) $\partial_X[\log(\det X)] = (X^{-1})^T$,

(iii) $\partial_X(x^T X^{-1} x) = -(X^{-1})^T x x^T (X^{-1})^T$,

(iv) $\partial_X Tr(AX) = A^T$.

Proposition 2.4. *Let X be a non-singular symmetric matrix. Then we have*

(i) $\partial_X^s(\det X) = (\det X) X^{-1}$,

(ii) $\partial_X^s[\log(\det X)] = X^{-1}$,

(iii) $\partial_X^s(x^T X^{-1} x) = -X^{-1} x x^T X^{-1}$,

(iv) $\partial_X^s Tr(AX) = A$. (A : *a fixed symmetric matrix*).

Higher order differentials are defined recursively.

Definition 2.5. The k-th order differential $\delta^{(k)}f$ of f is defined by

$$\delta f(\tilde{X}) - \sum_{i=1}^{k-1} \frac{1}{i!}\delta^{(i)}f = \frac{1}{k!}\delta^{(k)}f + o(\|\delta\tilde{X}\|^k).$$

for $k = 2, 3, ...$

We note that the relation $\delta^{(k)}f = \delta^{(1)}(\delta^{(k-1)}f)$ holds for certain functions treated in Section 4. Due to the new formalism of differentials and derivatives defined in this section, we can avoid tedious elementwise computation of tensors in handling various differential geometric quantities in our analysis through our coordinate free arguments. Compare the treatment shown in the following sections with the arguments seen in Anderson-Olkin [1], Mitchell [12], Ohara-Suda-Amari [13], Rao [14], Skovgaard [16] for example.

§3. Two Parameter Class of Dual Differential Geometry

We are concerned with the family $\mathfrak{N}$ of multivariate Gaussian distributions whose density function is given by

(13)

$$\begin{aligned} &p(x;\mu,\Sigma) \\ &\quad = \frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \\ &\quad = \exp\{-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu - \frac{1}{2}\log(\det\Sigma) - \frac{1}{2}\mu^T\Sigma^{-1}\mu - \frac{n}{2}\log(2\pi)\}, \end{aligned}$$

where $x, \mu \in R^n$ and Σ is a symmetric positive definite matrix.

Following Bandorff-Nielsen [4], Bandorff-Nielsen and Blæsild [5] and Amari [2], we define a potential function ψ defined on the manifold $\mathfrak{N}$ by the cumulant transformation of (13). The function ψ is represented in terms of a chart $\Xi = (\mu, \Sigma)$ as

$$\psi = \frac{1}{2}Tr(\Sigma^{-1}\mu\mu^T) + \frac{1}{2}\log(\det\Sigma) + \frac{n}{2}\log(2\pi). \tag{14}$$

Contrary to the general belief ([2,pp.80-88], [9,p.318]) that the $\alpha = \pm 1$ connections in the information geometry lead to the Kullback-Leibler information, we construct in this section various dual differential geometries whose $\alpha = \pm 1$ connections don't necessarily lead to the Kullback-Leibler information. Since

$\log(\det \Sigma)$ is a concave function of Σ and $Tr(\Sigma^{-1}\mu\mu^T)$ is a convex function of μ and Σ, the function $\psi(\mu, \Sigma)$ is neither convex nor concave. There could be many possible charts on $\mathfrak{N}$, with respect to which the function ψ becomes a convex function. For each choice of such chart, we can construct a dual chart and the associated dual potential function which are related by the *Legendre transformation* (See Amari [2]).

In this section, by choosing a special class of charts

$$\text{(15)} \quad \mathfrak{I}_{\tilde{\Theta}_{\beta,\gamma}}:$$
$$\mathfrak{N} \ni N(\mu, \Sigma) \mapsto \tilde{\Theta}_{\beta,\gamma} \equiv (\theta_{\beta,\gamma}, \Theta_{\beta,\gamma}) \equiv (\Sigma^{-\frac{\beta+1}{2}}\mu, (2\Sigma)^{-\gamma}) \in \mathfrak{D}_{\beta,\gamma} \equiv R^n \times \mathfrak{S}_n^+$$

for the potential function ψ, we construct a 2-parameter class of dual differential geometries which includes both the dualistic geometry ($(\beta, \gamma) = (1, 1)$) associated with Kullback-Leibler information and the Riemaniann geometry ($(\beta, \gamma) = (0, \frac{1}{2})$) proposed by Skovgaard [16] as special cases, where the fractional power of a symmetric positive definite matrix is defined by the *Dunford-Taylor integral* (See Kato [10]) as

$$\begin{aligned} \Theta^{-s} &= \frac{1}{2\pi\sqrt{-1}} \int_\Gamma \lambda^{-s}(\lambda I - \Theta)^{-1} d\lambda \\ &= \frac{1}{2\pi\sqrt{-1}} \int_{-\infty}^{0} (e^{-\pi s\sqrt{-1}} - e^{\pi s\sqrt{-1}})(-\lambda)^{-s}(\lambda I - \Theta)^{-1} d\lambda \\ &= \frac{\sin(\pi s)}{\pi} \int_0^\infty \lambda^{-s}(\lambda I + \Theta)^{-1} d\lambda, \quad (\ 0 < s < 1\ ), \end{aligned} \tag{16}$$

where the integration path Γ runs from $-\infty$ to $-\infty$ in the resolvent set of Θ, making a turn around the origin in the positive direction. The values of λ^{-s} should be chosen in such a way that $\lambda^{-s} > 0$ at the point where Γ meets the positive real axis. The second equality of the equation (16) is obtained by reducing the path Γ to the union of the upper and lower edges of the negative real axis.

We represent the potential function ψ with respect to each charts indexed by (β, γ). From the relations $\Sigma = \frac{1}{2}\Theta_{\beta,\gamma}^{-\frac{1}{\gamma}}$ and $\mu = 2^{-\frac{\beta+1}{2}}\Theta_{\beta,\gamma}^{-\frac{\beta+1}{2\gamma}}\theta_{\beta,\gamma}$ we have the representation,

$$\begin{aligned} \tilde{\Psi}_{\beta,\gamma} &\equiv \psi(\mathfrak{I}^{-1}_{\tilde{\Theta}_{\beta,\gamma}}(\tilde{\Theta}_{\beta,\gamma})) \\ &= 2^{-(\beta+1)} Tr(\Theta_{\beta,\gamma}^{-\frac{\beta}{\gamma}}\theta_{\beta,\gamma}\theta_{\beta,\gamma}^T) - \frac{1}{2\gamma}\log(\det\Theta_{\beta,\gamma}) + \frac{n}{2}\log(\pi). \end{aligned} \tag{17}$$

Proposition 3.1. *Let $0 < \beta$, $0 < \gamma$ and let $\frac{\beta}{\gamma} < 1$. The potential function $\tilde{\Psi}_{\beta,\gamma}$ is a convex function with respect to $(\theta_{\beta,\gamma}, \Theta_{\beta,\gamma})$.*

For proving Proposition 3.1, we employ the following Lemma due to Lieb [11].

Lemma 3.2 (Lieb). *The function from* $\mathfrak{S}_n^+ \times M(n,R)$ *to the non-negative reals* $R^+ \cup \{0\}$ *is defined by*

$$\mathfrak{S}_n^+ \times M(n,R) \ni (X,Y) \mapsto Tr[X^{-p}Y^T X^{-q}Y] \in R^+ \cup \{0\} \tag{18}$$

is jointly convex in (X,Y) *whenever* $0 \leq p, \quad 0 \leq q$ *and* $p+q \leq 1$.

Proof of Proposition 3.1. If we put $X = \Theta_{\beta,\gamma} \in \mathfrak{S}_n^+, \quad Y^T = [\theta_{\beta,\gamma}\ O] \in M(n,R)$, where O is an n by $(n-1)$ zero matrix, and $p = \frac{\beta}{\gamma}$, $q = 0$ in Lemma 3.2, then we find that the function $Tr(\Theta_{\beta,\gamma}^{-\frac{\beta}{\gamma}}\theta_{\beta,\gamma}\theta_{\beta,\gamma}^T)$ is a convex function in $(\theta_{\beta,\gamma}, \Theta_{\beta,\gamma})$. The function $-\frac{1}{2\gamma}\log(\det\Theta_{\beta,\gamma})$ is obviously convex in $\Theta_{\beta,\gamma}$. Hence, we obtain the result.

□

Let $\mathfrak{D}^*_{\beta,\gamma} \equiv \{(\eta_{\beta,\gamma}, H_{\beta,\gamma}) \in R^n \times \mathfrak{S}_n^- \mid \partial_{\theta_{\beta,\gamma}}\tilde{\Psi}_{\beta,\gamma} = \eta_{\beta,\gamma}, \partial^s_{\Theta_{\beta,\gamma}}\tilde{\Psi}_{\beta,\gamma} = H_{\beta,\gamma}\}$. Since the dual potential functions $\tilde{\Psi}_{\beta,\gamma}$ and $\tilde{\Phi}_{\beta,\gamma}$ are convex on $\mathfrak{D}_{\beta,\gamma}$ and $\mathfrak{D}^*_{\beta,\gamma}$, respectively, we can define the Legendre transformation between the spaces $\mathfrak{D}_{\beta,\gamma}$ and $\mathfrak{D}^*_{\beta,\gamma}$ as follows.

$$\mathfrak{L}_{\beta,\gamma} \quad : \quad \mathfrak{D}_{\beta,\gamma} \ni \tilde{\Theta}_{\beta,\gamma} = (\theta_{\beta,\gamma}, \Theta_{\beta,\gamma}) \mapsto (\eta_{\beta,\gamma}, H_{\beta,\gamma}) = \tilde{H}_{\beta,\gamma} \in \mathfrak{D}^*_{\beta,\gamma}.$$

By using this Legendre transformation, we derive the dual map $\mathfrak{I}_{\tilde{H}_{\beta,\gamma}}$ for each member of the 2-paramater chart.

Lemma 3.3. *Choosing an* $\mathfrak{I}_{\tilde{\Theta}_{\beta,\gamma}}$ *map,*

$$\begin{aligned} \mathfrak{I}_{\tilde{\Theta}_{\beta,\gamma}} \ &: \ \mathfrak{N} \longrightarrow \mathfrak{D}_{\beta,\gamma} = R^n \times \mathfrak{S}_n^+ \\ &N(\mu,\Sigma) \longmapsto (\Sigma^{-\frac{\beta+1}{2}}\mu, (2\Sigma)^{-\gamma}) \equiv (\theta_{\beta,\gamma}, \Theta_{\beta,\gamma}) \equiv \tilde{\Theta}_{\beta,\gamma}, \end{aligned} \tag{19}$$

we have the dual map $\mathfrak{I}_{\tilde{H}_{\beta,\gamma}}$,

$$\begin{aligned} \mathfrak{I}_{\tilde{H}_{\beta,\gamma}} \ &: \ \mathfrak{N} \longrightarrow \mathfrak{D}^*{}_{\beta,\gamma} \subset R^n \times \mathfrak{S}_n^- \\ &N(\mu,\Sigma) \longrightarrow (\eta_{\beta,\gamma}, H_{\beta,\gamma}) \equiv \tilde{H}_{\beta,\gamma}, \end{aligned} \tag{20}$$

where

$$\begin{aligned} \eta_{\beta,\gamma} &\equiv \Sigma^{\frac{\beta-1}{2}}\mu, \\ H_{\beta,\gamma} &\equiv -\frac{2^{\gamma-1}}{\gamma}\Sigma^{\gamma} - \frac{2^{-(\beta+1)}\sin(\frac{\beta\pi}{\gamma})}{\pi} \\ &\quad \times \int_0^{\infty} \lambda^{-\frac{\beta}{\gamma}}[(2\Sigma)^{-\gamma} + \lambda I]^{-1}\Sigma^{-\frac{\beta+1}{2}}\mu\mu^T\Sigma^{-\frac{\beta+1}{2}}[(2\Sigma)^{-\gamma} + \lambda I]^{-1}d\lambda \end{aligned}$$

for $0 < \beta$, $0 < \gamma$ *and* $\frac{\beta}{\gamma} < 1$.

Proof. In order to derive the dual chart, we calculate the differential of potential function.

$$\delta\tilde{\Psi}_{\beta,\gamma} \equiv \tilde{\Psi}_{\beta,\gamma}(\theta_{\beta,\gamma}+\delta\theta_{\beta,\gamma}, \Theta_{\beta,\gamma}+\delta\Theta_{\beta,\gamma}) - \tilde{\Psi}_{\beta,\gamma}(\theta_{\beta,\gamma}, \Theta_{\beta,\gamma}).$$

From the equation (17) we have

$$\begin{aligned}
&\delta\tilde{\Psi}_{\beta,\gamma} \\
&= 2^{-(\beta+1)} Tr[-\frac{\sin(\frac{\beta\pi}{\gamma})}{\pi}\int_0^\infty \lambda^{-\frac{\beta}{\gamma}}(\Theta_{\beta,\gamma}+\lambda I)^{-1}\delta\Theta_{\beta,\gamma}(\Theta_{\beta,\gamma}+\lambda I)^{-1}d\lambda\theta_{\beta,\gamma}\theta_{\beta,\gamma}^T \\
&\quad + \Theta_{\beta,\gamma}^{-\frac{\beta}{\gamma}}(\delta\theta_{\beta,\gamma}\theta_{\beta,\gamma}^T + \theta_{\beta,\gamma}\delta\theta_{\beta,\gamma}^T)] - \frac{1}{2\gamma}\delta_{\tilde{\Theta}_{\beta,\gamma}}[\log(\det\Theta_{\beta,\gamma})] + O(\|\delta\tilde{\Theta}_{\beta,\gamma}\|^2).
\end{aligned}$$

By the formula $\log\det(X+\delta X) = \log\det(X) + Tr(X^{-1}\delta X) + O(\|\delta X\|^2)$, we have

$$\begin{aligned}
\delta\tilde{\Psi}_{\beta,\gamma} &= Tr[2^{-\beta}\Theta_{\beta,\gamma}^{-\frac{\beta}{\gamma}}\theta_{\beta,\gamma}\delta\theta_{\beta,\gamma}^T \\
&\quad + \{-\frac{\sin(\frac{\beta\pi}{\gamma})}{\pi}\int_0^\infty \lambda^{-\frac{\beta}{\gamma}}(\Theta_{\beta,\gamma}+\lambda I)^{-1}\theta_{\beta,\gamma}{\theta_{\beta,\gamma}}^T(\Theta_{\beta,\gamma}+\lambda)^{-1}d\lambda \\
&\quad - \frac{1}{2\gamma}\Theta_{\beta,\gamma}^{-1}\}\delta\Theta_{\beta,\gamma}^T] + O(\|\delta\tilde{\Theta}_{\beta,\gamma}\|^2).
\end{aligned}$$

Therefore, we obtain

(21)

$$\begin{aligned}
\tilde{H}_{\beta,\gamma} &\equiv (\eta_{\beta,\gamma}, H_{\beta,\gamma}) \\
&= (2^{-\beta}\Theta_{\beta,\gamma}^{-\frac{\beta}{\gamma}}\theta_{\beta,\gamma} \quad , -\frac{1}{2\gamma}\Theta_{\beta,\gamma}^{-1} - 2^{-(\beta+1)}\frac{\sin(\frac{\beta\pi}{\gamma})}{\pi} \\
&\quad \times \int_0^\infty \lambda^{-\frac{\beta}{\gamma}}(\Theta_{\beta,\gamma}+\lambda I)^{-1}\theta_{\beta,\gamma}{\theta_{\beta,\gamma}}^T(\Theta_{\beta,\gamma}+\lambda I)^{-1}d\lambda) \in R^n \times \mathfrak{S}_n^-.
\end{aligned}$$

□

In order to construct the dual potential function defined on the space $\mathfrak{D}_{\beta,\gamma}$ with respect to $\tilde{\Psi}_{\beta,\gamma}$, we prepare the following two Lemmas.

Lemma 3.4. *Let* $\tilde{\Theta} \equiv (\theta, \Theta) \in R^n \times \mathfrak{S}_n^+$ *and* $h(\tilde{\Theta}) = Tr(\Theta^{-\frac{\beta}{\gamma}}\theta\theta^T)$ *with* $0 < \gamma,\ 0 \le \beta$ *and* $\frac{\beta}{\gamma} \le 1$, *Then we have*

$$\begin{aligned}
\partial_\theta h &= 2\Theta^{-\frac{\beta}{\gamma}}\theta \in R^n, \\
\partial_\Theta^s h &= \frac{1}{2\pi\sqrt{-1}}\int_{\tilde{\Gamma}} \lambda^{-\frac{\beta}{\gamma}}(\lambda I - \Theta)^{-1}\theta\theta^T(\lambda I - \Theta)^{-1}d\lambda \in \mathfrak{S}_n.
\end{aligned}$$

Proof. We consider the differential of the function h and we find its derivative by the definition 2.3.

$$\begin{aligned}\delta h &\equiv h(\theta+\delta\theta,\Theta+\delta\Theta)-h(\theta,\Theta)\\ &= Tr[\frac{1}{2\pi\sqrt{-1}}\int_{\tilde{\Gamma}}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-\Theta-\delta\Theta)^{-1}d\lambda(\theta+\delta\theta)(\theta+\delta\theta)^T]\\ &\qquad -Tr[\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-\Theta)^{-1}d\lambda\theta\theta^T]\end{aligned}$$

Since $[(\lambda I-\Theta)-\delta\Theta]^{-1}=(\lambda I-\Theta)^{-1}+(\lambda I-\Theta)^{-1}\delta\Theta(\lambda I-\Theta)^{-1}+O(\|\delta\Theta\|^2)$, we have

$$\begin{aligned}\delta h = Tr[&\frac{1}{2\pi\sqrt{-1}}\int_{\tilde{\Gamma}}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-\Theta)^{-1}\delta\Theta(\lambda I-\Theta)^{-1}d\lambda\theta\theta^T]\\ &+Tr[\frac{1}{2\pi\sqrt{-1}}\int_{\tilde{\Gamma}}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-\Theta)^{-1}d\lambda(\theta\delta\theta^T+\delta\theta\theta^T)]+O(\|\delta\tilde{\Theta}\|^2).\end{aligned}$$

Hence, we obtain the result.

□

Lemma 3.5. *Let* $(\theta,\Theta)\in R^n\times\mathfrak{S}_n^+$ *and let* $X=\Theta^{-\gamma},\ (0<\gamma)$, *and*

$$Y=\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-X)^{-1}\theta\theta^T(\lambda I-X)^{-1}d\lambda \text{ with } 0\le\beta \text{ and } \frac{\beta}{\gamma}\le 1.$$

Then we have the pairing between X *and* Y *as*

$$< X,Y >= -\frac{\beta}{\gamma}Tr(\Theta^\beta\theta\theta^T). \tag{22}$$

Proof. In this case, since the bilinear form $< X,Y >$ is $Tr(XY)$, we calculate the trace of matrix XY.

$$\begin{aligned}Tr(XY) &= Tr[X\cdot\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-X)^{-1}\theta\theta^T(\lambda I-X)^{-1}d\lambda]\\ &= \frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}}\theta^T(\lambda I-X)^{-1}X(\lambda I-X)^{-1}\theta d\lambda\\ &= \theta^T[\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}}(\lambda I-X)^{-1}X(\lambda I-X)^{-1}d\lambda]\theta\\ &= \theta^T[\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}+1}(\lambda I-X)^{-2}d\lambda]\theta-\theta^T\Theta^\beta\theta.\end{aligned}$$

Here, by employing the Cauchy's integral formula, we have

$$(-\frac{\beta}{\gamma}+1)\Theta^\beta=\frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\lambda^{-\frac{\beta}{\gamma}+1}(\lambda I-X)^{-2}d\lambda.$$

Hence, we obtain the result.

□

Corollary 3.6. *Let* $\tilde{\Theta}_{\beta,\gamma} \equiv (\theta_{\beta,\gamma}, \Theta_{\beta,\gamma})$ *and* $\tilde{H}_{\beta,\gamma} \equiv (\eta_{\beta,\gamma}, H_{\beta,\gamma})$ *be defined in Lemma 3.3. Then we have*

$$< \tilde{\Theta}_{\beta,\gamma}, \tilde{H}_{\beta,\gamma} >= (1 - \frac{\beta}{2\gamma})Tr(\Sigma^{-1}\mu\mu^T) - \frac{n}{2\gamma}.$$

From Lemma 3.4 and Lemma 3.5, we have the dual potential function $\tilde{\Phi}_{\beta,\gamma}$ corresponding to the potential function $\tilde{\Psi}_{\beta,\gamma}$.

Proposition 3.7. *Let* $\tilde{\Phi}_{\beta,\gamma} \equiv \phi_{\beta,\gamma}(\mathfrak{I}^{-1}_{\tilde{H}_{\beta,\gamma}}(\tilde{H}_{\beta,\gamma}))$, *where*

$$\phi_{\beta,\gamma} = \frac{1}{2}(1 - \frac{\beta}{\gamma})Tr(\Sigma^{-1}\mu\mu^T) - \frac{1}{2}\log(\det\Sigma) - \frac{n}{2}\log(2\pi e^{\gamma^{-1}}).$$

Then the potential function $\tilde{\Phi}_{\beta,\gamma}$ *coinsides with* $< \tilde{\Theta}_{\beta,\gamma}, \tilde{H}_{\beta,\gamma} > -\tilde{\Psi}_{\beta,\gamma}$, *i.e.,*

$$\tilde{\Phi}_{\beta,\gamma} =< \tilde{\Theta}_{\beta,\gamma}, \tilde{H}_{\beta,\gamma} > -\tilde{\Psi}_{\beta,\gamma}.$$

Proof. We only have to substitute the result of Corollary 3.6 into the relations

$$\tilde{\Phi}_{\beta,\gamma} \equiv< \tilde{\Theta}_{\beta,\gamma}, \tilde{H}_{\beta,\gamma} > -\tilde{\Psi}_{\beta,\gamma}.$$

□

We note that the relation between two potential functions $\tilde{\Phi}_{\beta,\gamma}$ and $\phi_{\beta,\gamma}$ is $\phi_{\beta,\gamma}(\cdot) = \tilde{\Phi}_{\beta,\gamma}(\mathfrak{I}_{\tilde{H}_{\beta,\gamma}}(\cdot))$.

Now we can construct a *general divergence* for each dual geometry in the class. Let $\tilde{\Theta}_2 \equiv \mathfrak{I}_{\tilde{\Theta}_{\beta,\gamma}}(N(\mu_2, \Sigma_2))$ and $\tilde{H}_1 \equiv \mathfrak{I}_{\tilde{H}_{\beta,\gamma}}(N(\mu_1, \Sigma_1))$. If we define a divergence by

$$Div_{\beta,\gamma}(N(\mu_2, \Sigma_2), N(\mu_1, \Sigma_1)) \equiv \tilde{\Psi}_{\beta,\gamma}(\tilde{\Theta}_2) + \tilde{\Phi}_{\beta,\gamma}(\tilde{H}_1) - < \tilde{\Theta}_2, \tilde{H}_1 >, \tag{23}$$

then we have

Proposition 3.8. *The divergence* $Div_{\beta,\gamma}$ *between the two distributions* $N(\mu_1, \Sigma_1)$ *and* $N(\mu_2, \Sigma_2)$ *is given by*

$$\begin{aligned}
&Div_{\beta,\gamma}(N(\mu_2, \Sigma_2), N(\mu_1, \Sigma_1)) \\
&\quad = \frac{1}{2}Tr(\Sigma_2^{-1}\mu_2\mu_2^T) + \frac{1}{2}\log\left(\frac{\det\Sigma_2}{\det\Sigma_1}\right) + \frac{1}{2}(1 - \frac{\beta}{\gamma})Tr(\Sigma_1^{-1}\mu_1\mu_1^T) - \frac{n}{2\gamma} \\
&\qquad - Tr(\Sigma_1^{\frac{\beta-1}{2}}\Sigma_2^{-\frac{\beta+1}{2}}\mu_2\mu_1^T) + \frac{1}{2\gamma}Tr(\Sigma_1^{\gamma}\Sigma_2^{-\gamma}) \\
&\qquad - 2^{-(\beta+1)}\mu_1^T\Sigma_1^{-\frac{\beta+1}{2}}(\frac{1}{2\pi\sqrt{-1}} \\
&\qquad\qquad \times \int_\Gamma \lambda^{-\frac{\beta}{\gamma}}[\lambda I - (2\Sigma_1)^{-\gamma}]^{-1}(2\Sigma_2)^{-\gamma}[\lambda I - (2\Sigma_1)^{-\gamma}]^{-1}d\lambda)\Sigma_1^{-\frac{\beta+1}{2}}\mu_1,
\end{aligned}$$

which satisfies the inequality,

$$Div_{\beta,\gamma}(N(\mu_2,\Sigma_2),N(\mu_1,\Sigma_1)) \geq 0,$$

where the equality holds if and only if $(\mu_1,\Sigma_1)=(\mu_2,\Sigma_2)$.

Proof. Substituting the pairing $<\tilde{\Theta}_2,\tilde{H}_1>$, the dual potential functions $\tilde{\Psi}_{\beta,\gamma}$ and $\tilde{\Phi}_{\beta,\gamma}$ into the definition (23) of divergence, we obtains the result. □

We will identify each geometry of the 2-parameter class by

$$(\mathfrak{N},((\mathfrak{I}_{\tilde{\Theta}_{\beta,\gamma}},\mathfrak{D}_{\beta,\gamma}),(\mathfrak{I}_{\tilde{H}_{\beta,\gamma}},\mathfrak{D}^*_{\beta,\gamma}),<,>),\tilde{\Psi}_{\beta,\gamma},\tilde{\Phi}_{\beta,\gamma},Div_{\beta,\gamma}).$$

The 2-parameter class of geometry includes a couple of important geometries. One is the Riemannian geometry studied by Skovgaard [16]. It is given by putting $(\beta,\gamma)=(0,\frac{1}{2})$, i.e.,

$$\begin{aligned}
&\mathfrak{I}_{\tilde{\Theta}_{0,\frac{1}{2}}}(N(\mu_1,\Sigma_1))=(\Sigma_1^{-\frac{1}{2}}\mu_1,(2\Sigma_1)^{-\frac{1}{2}}),\\
&\mathfrak{D}_{0,\frac{1}{2}}\equiv R^n\times\mathfrak{S}_n^+,\\
&\mathfrak{I}_{\tilde{H}_{0,\frac{1}{2}}}(N(\mu_2,\Sigma_2))=(\Sigma_2^{-\frac{1}{2}}\mu_2,-(2\Sigma_2)^{\frac{1}{2}}),\\
&\mathfrak{D}^*_{0,\frac{1}{2}}\equiv R^n\times\mathfrak{S}_n^-,\\
&<\tilde{\Theta}_1,\tilde{H}_2>=\mu_1^T\Sigma_1^{-\frac{1}{2}}\Sigma_2^{-\frac{1}{2}}\mu_2-Tr[(2\Sigma_1)^{-\frac{1}{2}}(2\Sigma_2)^{-\frac{1}{2}}],\\
&\psi=\frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1+\frac{1}{2}\log(\det\Sigma_1)+\frac{n}{2}\log(2\pi),\\
&\qquad(\text{i.e., }\tilde{\Psi}_{0,\frac{1}{2}}=\frac{1}{2}Tr(\theta_{0,\frac{1}{2}}\theta_{0,\frac{1}{2}}^T)-\log(\det\Theta_{0,\frac{1}{2}})+\frac{n}{2}\log(\pi)),\\
&\phi_{0,\frac{1}{2}}=\frac{1}{2}\mu_2^T\Sigma_2^{-1}\mu_2-\frac{1}{2}\log(\det\Sigma_2)-\frac{n}{2}\log(2\pi e^2),\\
&\qquad(\text{i.e., }\tilde{\Phi}_{0,\frac{1}{2}}=\frac{1}{2}Tr(\eta_{0,\frac{1}{2}}\eta_{0,\frac{1}{2}}^T)-\log[\det(-H_{0,\frac{1}{2}})]+\frac{n}{2}\log(\pi e^2)).\\
&Div_{0,\frac{1}{2}}=\frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1+\frac{1}{2}\log(\det\Sigma_1)\\
&\qquad\qquad+\frac{1}{2}\mu_2^T\Sigma_2^{-1}\mu_2-\frac{1}{2}\log(\det\Sigma_2)\\
&\qquad\qquad-\mu_1^T\Sigma_1^{-\frac{1}{2}}\Sigma_2^{-\frac{1}{2}}\mu_2+Tr[(2\Sigma_1)^{-\frac{1}{2}}(2\Sigma_2)^{\frac{1}{2}}]-n.
\end{aligned}$$

There isn't an interplay of the variables $\theta_{0,\frac{1}{2}}$ and $\Theta_{0,\frac{1}{2}}$.
Skovgaard [16] considered the same geometry from a point of view of the Levi-Civita connection associated with the Fisher information. The other geometry is the dual geometry which leads to the Kullback-Leibler information. We will discuss this topic in the next section.

§4. Geometry Associated with Kullback-Leibler Information

In this section we single out one ($(\beta, \gamma) = (1, 1)$) of the dual differential geometries discussed in the previous section, that leads to the Kullback-Leibler information, and analyze the associated dual connections (Theorem 4.9.).

For convenience of notation, we write $\mathfrak{D}_{1,1}$ and $\mathfrak{D}^*_{1,1}$ as $\mathfrak{D}$ and $\mathfrak{D}^*$ in this section. Let M be a Riemannian manifold with metric g. Two affine connections represented by covariant derivatives ∇ and ∇^* on M are said to be *dual* or *conjugate* [2,9] with respect to g if, for any vector fields X, Y and Z on M,

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z), \tag{24}$$

where $g(Y, Z)$ denotes the inner product of Y and Z with respect to the metric g. If the torsion and the Riemannian curvatures of M with respect to the connections ∇ and ∇^* vanish, (M, ∇, ∇^*) is said to be *dually flat*.

While we have shown, in Lemma 3.3, general form of the dual charts, we can now specify the *dually flat charts* for the dual differential geometry associated with the Kullback-Leibler information. We define the canonical map $\mathfrak{I}_{\tilde{\Theta}}$ and the moment map $\mathfrak{I}_{\tilde{H}}$ as follows:

$$\begin{aligned} \mathfrak{I}_{\tilde{\Theta}} \quad : \quad & \mathfrak{N} \longrightarrow R^n \times \mathfrak{S}^+_n \\ & N(\mu, \Sigma) \longmapsto (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) \equiv (\theta, \Theta) \equiv \tilde{\Theta}, \end{aligned} \tag{25}$$

and

$$\begin{aligned} \mathfrak{I}_{\tilde{H}} \quad : \quad & \mathfrak{N} \hookrightarrow R^n \times \mathfrak{S}^-_n \\ & N(\mu, \Sigma) \longmapsto (\mu, -(\Sigma + \mu\mu^T)) \equiv (\eta, H) \equiv \tilde{H}. \end{aligned} \tag{26}$$

The map $\mathfrak{I}_{\tilde{\Theta}}$ is a bijection and the map $\mathfrak{I}_{\tilde{H}}$ is an into-injection. Note that the sign of our map $\mathfrak{I}_{\tilde{\Theta}}$ is different from that of Bandorff-Nielsen [4] and Amari [2]. We have also the following relations.

$$\mu = \frac{1}{2}\Theta^{-1}\theta = \eta, \qquad \Sigma = \frac{1}{2}\Theta^{-1} = -H - \eta\eta^T, \tag{27}$$

$$\Theta = -\frac{1}{2}(H + \eta\eta^T)^{-1}, \qquad \theta = -(H + \eta\eta^T)^{-1}\eta. \tag{28}$$

If we employ the canonical map, then we have

Proposition 4.1. *Let* $\tilde{\psi}(\tilde{\Theta}) \equiv \psi(\mathfrak{I}^{-1}_{\tilde{\Theta}}(\tilde{\Theta}))$. *The potential function* ψ *can be written by*

$$\tilde{\psi}(\tilde{\Theta}) = \frac{1}{4}Tr(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log(\det\Theta) + \frac{n}{2}\log(\pi) \tag{29}$$

which is strictly convex in $\tilde{\Theta} = (\theta, \Theta)$.

Proof. Let $\Theta = \Theta_0 + \delta\Theta_0$, where $\Theta_0 \in \mathfrak{S}_n^+$, $O \neq \delta\Theta_0 \in \mathfrak{S}_n$, and let $\theta = \theta_0 + \delta\theta_0$, where $\theta_0,\ 0 \neq \delta\theta_0 \in R^n$. Since

$$\begin{aligned}\delta_{\tilde{\Theta}}\tilde{\psi} &= \frac{1}{4}Tr\left[\delta\Theta^{-1}\theta\theta^T + \Theta^{-1}\delta(\theta\theta^T)\right] - \frac{1}{2}\delta[\log(\det\Theta)] \\ &= \frac{1}{4}Tr\left[-\Theta^{-1}\delta\Theta_0\Theta^{-1}\theta\theta^T + \Theta^{-1}(\delta\theta_0\theta^T + \theta\delta\theta_0^T)\right] - \frac{1}{2}Tr(\Theta^{-1}\delta\Theta_0),\end{aligned}$$

we have

$$\begin{aligned}&\delta^{(2)}_{\tilde{\Theta}}\tilde{\psi} \\ &= Tr\left[\Theta^{-1}\delta\Theta_0\Theta^{-1}\delta\Theta_0\Theta^{-1}\theta\theta^T - \Theta^{-1}\delta\Theta_0\Theta^{-1}(\delta\theta_0\theta^T + \theta\delta\theta_0^T) + \Theta^{-1}\delta\theta_0\delta\theta_0^T\right] \\ &\qquad - \frac{1}{2}Tr(\Theta^{-1}\delta\Theta_0\Theta^{-1}\delta\Theta_0) \\ &= \left\{(\delta\Theta_0\Theta_0^{-1}\theta - \delta\theta_0)^T\Theta^{-1}(\delta\Theta_0\Theta_0^{-1}\theta - \delta\theta_0) + Tr(\Theta^{-1}\delta\Theta_0\Theta^{-1}\delta\Theta_0)\right\}\end{aligned}$$

Hence, we obtain $\delta^{(2)}_{\tilde{\Theta}}\tilde{\psi} > 0$ when $\delta\Theta_0 \neq 0$ and $\delta\theta_0 \neq 0$.

□

Let $\mathfrak{D} \equiv \mathfrak{I}_{\tilde{\Theta}}(\mathfrak{N})$. The function $\tilde{\psi}$ is defined on $\mathfrak{D}$. Consulting Amari's book [2], we define the dual potential function $\tilde{\phi}$ by

$$\tilde{\phi} = Tr(\theta\eta^T + \Theta H^T) - \tilde{\psi} + C \tag{30}$$

such that

$$\tilde{\phi} + \tilde{\psi} - < \tilde{\Theta}, \tilde{H} >= 0 \quad \text{at} \quad \mathfrak{I}^{-1}_{\tilde{\Theta}}(\tilde{\Theta}) = \mathfrak{I}^{-1}_{\tilde{H}}(\tilde{H}).$$

The constant term C of the equation (30) doesn't appear in [2,9], but this term plays an important role in constructing the *divergence*. Under the above condition, we have

Proposition 4.2. *The dual potential function*

$$\begin{aligned}\tilde{\phi}(\tilde{H}) &= -\frac{1}{2}\log[\det(-H - \eta\eta^T)] - \frac{n}{2}\log(2\pi e) \\ &= -\frac{1}{2}\log\det(I + H^{-1}\eta\eta^T) - \frac{1}{2}\log\det(-H) - \frac{n}{2}\log(2\pi e) \\ &= -\frac{1}{2}\log(1 + \eta^T H^{-1}\eta) - \frac{1}{2}\log\det(-H) - \frac{n}{2}\log(2\pi e)\end{aligned}$$

is defined on $\mathfrak{D}^* = \left\{\ \tilde{H} = (\eta, H) \in R^n \times \mathfrak{S}_n^- \mid 1 + \eta^T H^{-1}\eta > 0\ \right\}$.

Proof. Substituting the formula (28) into the equation (30), we have

$$\begin{aligned}\tilde{\phi} &= \eta^T[-(H + \eta\eta^T)^{-1}\eta] + Tr[-\frac{1}{2}(H + \eta\eta^T)^{-1}H] \\ &\quad - \frac{1}{4}Tr[-2(H + \eta\eta^T)\{-(H + \eta\eta^T)^{-1}\eta\}\{-\eta^T(H + \eta\eta^T)^{-1}\}] \\ &\quad + \frac{1}{2}\log\det[-\frac{1}{2}(H + \eta\eta^T)^{-1}] - \frac{n}{2}\log(\pi) + C.\end{aligned}$$

We must determine the constant term C of the equation (30), which satisfies

$$-\frac{n}{2}-\frac{n}{2}\log 2-\frac{n}{2}\log(\pi)+C+\frac{n}{2}\log(\pi)=-\frac{n}{2}.$$

Hence, we have the constant $C\equiv\frac{n}{2}\log(2)$.

□

We note that the space $\mathfrak{D}^*$ is not a cone, but we can consider the domain $\mathfrak{D}^*$ as a cone when the mean μ of $\{N(\mu,\Sigma)\}$ is zero.

If we define the divergence Div as

$$Div=\psi+\phi-<\mathfrak{I}_{\tilde{\Theta}}^{-1}(\tilde{\Theta}),\mathfrak{I}_{\tilde{H}}^{-1}(\tilde{H})>, \tag{31}$$

then the divergence Div satisfies

$$(i)\quad Div(\mathfrak{I}_{\tilde{\Theta}}^{-1}(\tilde{\Theta}),\mathfrak{I}_{\tilde{H}}^{-1}(\tilde{H}))\geqq 0,\quad \text{for all } \tilde{\Theta}\in\mathfrak{D},\ \tilde{H}\in\mathfrak{D}^*,$$
$$(ii)\quad Div(\mathfrak{I}_{\tilde{\Theta}}^{-1}(\tilde{\Theta}),\mathfrak{I}_{\tilde{H}}^{-1}(\tilde{H}))=0\quad \text{if and only if}\quad \mathfrak{I}_{\tilde{\Theta}}^{-1}(\tilde{\Theta})=\mathfrak{I}_{\tilde{H}}^{-1}(\tilde{H}).$$

Proposition 4.3. *The potential function $\tilde{\phi}(\eta,H)$ defined on $\mathfrak{D}^*=\left\{\ (\eta,H)\in R^n\times\mathfrak{S}_n^-\mid 1+\eta^TH^{-1}\eta>0\ \right\}$ is a convex function in $\tilde{H}=(\eta,H)$.*

Proof. Let $H=H_0+\delta H_0$ and $\eta=\eta_0+\delta\eta_0$, where $(\delta\eta_0,\delta H_0)\in(R^n\setminus 0)\times(\mathfrak{S}_n\setminus O)$ and $(\eta,H)\in R^n\times\mathfrak{S}_n^-$ such that $-H-\eta\eta^T>O$.

Since $\delta_{\tilde{H}}^{(1)}\tilde{\phi}=-\frac{1}{2}Tr\left\{(-H-\eta\eta^T)^{-1}\delta_{\tilde{H}}(-H-\eta\eta^T)\right\}$,
we have

$$\begin{aligned}&\delta_{\tilde{H}}^{(2)}\tilde{\phi}(\tilde{H})\\&=Tr[(-H-\eta\eta^T)^{-1}(\delta H_0-\delta\eta_0\eta^T-\eta\delta\eta_0^T)\\&\qquad\times(-H-\eta\eta^T)^{-1}(\delta H_0-\delta\eta_0\eta^T-\eta\delta\eta_0^T)]+Tr\left[(-H-\eta\eta^T)^{-1}\delta\eta_0\delta\eta_0^T\right].\end{aligned}$$

Hence, we have the result.

□

It is interesting to consider the domain $\mathfrak{D}^*$ with the metric $g^*\equiv\delta_{\tilde{H}}^{(2)}(\tilde{H})$ as a Riemannian space. We note the following proposition.

Proposition 4.4. *The Legendre transformation and its inverse are given by the following equations.*

$$\partial_\theta\tilde{\psi}=\eta,\qquad \partial_\Theta^s\tilde{\psi}=H. \tag{32}$$

$$\partial_\eta\tilde{\phi}=\theta,\qquad \partial_H^s\tilde{\phi}=\Theta. \tag{33}$$

Proof. From Proposition 2.4, we obtain the equation (32).
Since

$$\begin{aligned}
\delta_{\tilde{H}}\tilde{\phi} &= \tilde{\phi}(\eta+\delta\eta, H+\delta H) - \tilde{\phi}(\eta, H) \\
&= -\frac{1}{2}\log[\det\{-H-\delta H-(\eta+\delta\eta)(\eta^T+\delta\eta^T)\}] + \frac{1}{2}\log[\det(-H-\eta\eta^T)] \\
&= -\frac{1}{2}\log[\det\{I_n + (-H-\eta\eta^T)^{-1}(-\delta H-\eta\delta\eta^T-\delta\eta\eta^T-\delta\eta\delta\eta^T)\}] \\
&= \frac{1}{2}Tr[(-H-\eta\eta^T)^{-1}(\delta H+\eta\delta\eta^T+\delta\eta\eta^T)] + O(\|\delta\tilde{H}\|^2),
\end{aligned}$$

we obtain the equation (33).

□

The equations (32) and (33) are parallel to the Amari-Nagaoka's ([2], p80). Note that this parallel formalism is only possible by our definition 2.3 of derivative.

In order to investigate the geometric structures between the dual spaces $\mathfrak{D}$ and $\mathfrak{D}^*$ in terms of the two metrics and the dual connections, we will prepare three Lemmas by using the higher order differential operators $\delta^{(n)}$ up to the third order. For later use we introduce the notation $O(n) \equiv O(\|\delta\tilde{X}\|^n)$.

Since the dual potential functions $\tilde{\psi}$ and $\tilde{\phi}$ are convex, we can employ the second order differentials for the potentials as Riemannian metrics. The metrics g on $\mathfrak{D}$ and g^* on $\mathfrak{D}^*$ are defined as

$$g \equiv \delta^{(2)}_{\tilde{\Theta}}\tilde{\psi}(\tilde{\Theta}), \qquad g^* \equiv \delta^{(2)}_{\tilde{H}}\tilde{\phi}(\tilde{H}). \tag{34}$$

Each tangent vector $\partial_{\tilde{\Theta}} \in T_{\tilde{\Theta}}\mathfrak{D}$ and $\partial_{\tilde{H}} \in T_{\tilde{H}}\mathfrak{D}^*$ correspond to the increment $\delta\tilde{\Theta}$ and $\delta\tilde{H}$, respectively. We define two kinds of inner product

$$\begin{aligned}
g((\partial_{\tilde{\Theta}})_1, (\partial_{\tilde{\Theta}})_2)\big|_{\tilde{\Theta}} \equiv \frac{1}{2}Tr[&\Theta^{-1}\delta\theta_2\delta\theta_1^T - \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\delta\theta_1^T - \Theta^{-1}\theta\delta\theta_2^T\Theta^{-1}\delta\Theta_1 \\
&+ \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1 + \Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1], \\
g^*((\partial_{\tilde{H}})_1, (\partial_{\tilde{H}})_2)\big|_{\tilde{H}} \equiv -\frac{1}{2}Tr[&2(H+\eta\eta^T)^{-1}\delta\eta_1\delta\eta_2^T \\
&- (H+\eta\eta^T)^{-1}(\delta H_1 + \eta\delta\eta_1^T + \delta\eta_1\eta^T)(H+\eta\eta^T)^{-1} \\
&\times (\delta H_2 + \delta\eta_2\eta^T + \eta\delta\eta_2^T)].
\end{aligned}$$

Then we have

Lemma 4.5. *Under the definition (34) of metrics, we have*

$$\begin{aligned}
g((\partial_{\tilde{\Theta}})_1, (\partial_{\tilde{\Theta}})_2)\big|_{\tilde{\Theta}} &= <(\partial_{\tilde{\Theta}})_1 < (\partial_{\tilde{\Theta}})_2\tilde{\psi}, \delta\tilde{\Theta}_2>, \delta\tilde{\Theta}_1> \big|_{\tilde{\Theta}} = \delta^{(1)}_{\tilde{\Theta}_1}(\delta^{(1)}_{\tilde{\Theta}_2}\tilde{\psi})\big|_{\tilde{\Theta}}, \\
g^*((\partial_{\tilde{H}})_1, (\partial_{\tilde{H}})_2)\big|_{\tilde{H}} &= <(\partial_{\tilde{H}})_1 < (\partial_{\tilde{H}})_2\tilde{\phi}, \delta\tilde{H}_2>, \delta\tilde{H}_1> \big|_{\tilde{H}} = \delta^{(1)}_{\tilde{H}_1}(\delta^{(1)}_{\tilde{H}_2}\tilde{\phi})\big|_{\tilde{H}}.
\end{aligned}$$

Proof. Since

$$\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi} = Tr[(\frac{1}{2}\Theta^{-1}\theta)\delta\theta_1^T - (\frac{1}{2}\Theta^{-1} + \frac{1}{4}\Theta^{-1}\theta\theta^T\Theta^{-1})\delta\Theta_1],$$

we have only to calculate the following quantity.

$$\begin{aligned}
&\delta_{\tilde{\Theta}_2}(\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi}) \\
&= Tr[\{\frac{1}{2}(\Theta+\delta\Theta_2)^{-1}(\theta+\delta\theta_2)\}\delta\theta_1^T - \frac{1}{2}\Theta^{-1}\theta)\delta\theta_1^T] \\
&\quad - Tr[\{\frac{1}{2}(\Theta+\delta\Theta_2)^{-1} \\
&\quad + \frac{1}{4}(\Theta+\delta\Theta_2)^{-1}(\theta+\delta\theta_2)(\theta+\delta\theta_2)^T(\Theta+\delta\Theta_2)^{-1}\}\delta\Theta_1 \\
&\quad + (\frac{1}{2}\Theta^{-1} + \frac{1}{4}\Theta^{-1}\theta\theta^T\Theta^{-1})\delta\Theta_1] + O(3) \\
&= Tr[\frac{1}{2}(\Theta^{-1} - \Theta^{-1}\delta\Theta_2\Theta^{-1} + O(2))(\theta\delta\theta_1^T + \delta\theta_2\delta\theta_1^T) - (\frac{1}{2}\Theta^{-1}\theta)\delta\theta_1^T] \\
&\quad - Tr[\frac{1}{2}(\Theta^{-1} - \Theta^{-1}\delta\Theta_2\Theta^{-1} + O(2))\delta\Theta_1 \\
&\quad + \frac{1}{4}(\Theta^{-1} - \Theta^{-1}\delta\Theta_2\Theta^{-1} + O(2))(\theta\theta^T + \theta\delta\theta_2^T + \delta\theta_2\theta^T + \delta\theta_2\delta\theta_2^T) \\
&\quad \times (\Theta^{-1} - \Theta^{-1}\delta\Theta_2\Theta^{-1} + O(2))\delta\Theta_1 \\
&\quad - \frac{1}{2}\Theta^{-1}\delta\Theta_1 - \frac{1}{4}\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1] + O(3) \\
&= \frac{1}{2}Tr[\Theta^{-1}\delta\theta_2\delta\theta_1^T - \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\delta\theta_1^T] \\
&\quad + \frac{1}{2}Tr[\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1 - \frac{1}{2}\Theta^{-1}\theta\delta\theta_2^T\Theta^{-1}\delta\Theta_1 - \frac{1}{2}\Theta^{-1}\delta\theta_2\theta^T\Theta^{-1}\delta\Theta_1 \\
&\quad + \frac{1}{2}\Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1 + \frac{1}{2}\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1] + O(3).
\end{aligned}$$

Hence we have

$$\begin{aligned}
\delta^{(1)}_{\tilde{\Theta}_2}(\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi}) = \frac{1}{2}Tr[&\Theta^{-1}\delta\theta_2\delta\theta_1^T - \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\delta\theta_1^T - \Theta^{-1}\theta\delta\theta_2^T\Theta^{-1}\delta\Theta_1 \\
&+ \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1 + \Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1].
\end{aligned}$$

In the same way, the second equation can be derived.

□

We next define the two connections associated with the two potential functions $\tilde{\psi}$ and $\tilde{\phi}$.

Lemma 4.6. *Let* $\tilde{\Theta} = (\theta, \Theta)$, $\tilde{H} = (\eta, H)$. *For any vectors* $(\partial_{\tilde{\Theta}})_2$, $(\partial_{\tilde{\Theta}})_3 \in T_{\tilde{\Theta}}\mathfrak{D}$ *and* $(\partial_{\tilde{H}})_2$, $(\partial_{\tilde{H}})_3 \in T_{\tilde{H}}\mathfrak{D}^*$ *if we identify two kinds of covariant deriva-*

tives as follows.

$$T_{\tilde{\Theta}}\mathfrak{D} \ni \nabla_{(\partial_{\tilde{\Theta}})_3}(\partial_{\tilde{\Theta}})_2 \leftrightarrow (x, X) \in R^n \times \mathfrak{S}_n,$$
$$T_{\tilde{H}}\mathfrak{D}^* \ni \nabla^*_{(\partial_{\tilde{H}})_3}(\partial_{\tilde{H}})_2 \leftrightarrow (y, Y) \in R^n \times \mathfrak{S}_n,$$

where

$$\begin{aligned}
(x, X) \equiv\ & (-\delta\Theta_3\Theta^{-1}\delta\theta_2 - \delta\Theta_2\Theta^{-1}\delta\theta_3, -\delta\Theta_2\Theta^{-1}\delta\Theta_3 - \delta\Theta_3\Theta^{-1}\delta\Theta_2),\\
y \equiv\ & -\delta H_2(H + \eta\eta^T)^{-1}\delta\eta_3 - (\delta\eta_2\eta^T + \eta\delta\eta_2^T)(H + \eta\eta^T)^{-1}\delta\eta_3\\
& - \delta H_3(H + \eta\eta^T)^{-1}\delta\eta_2 - (\delta\eta_3\eta^T + \eta\delta\eta_3^T)(H + \eta\eta^T)^{-1}\delta\eta_2,\\
Y \equiv\ & -\delta H_2(H + \eta\eta^T)^{-1}\delta H_3 - \delta H_3(H + \eta\eta^T)^{-1}\delta H_2\\
& - \delta H_2(H + \eta\eta^T)^{-1}\eta\delta\eta_3^T - \delta H_3(H + \eta\eta^T)^{-1}\eta\delta\eta_2\\
& - \delta\eta_2\eta^T(H + \eta\eta^T)\delta H_3 - \delta\eta_3\eta^T(H + \eta\eta^T)^{-1}\delta H_2\\
& - \delta\eta_2\eta^T(H + \eta\eta^T)^{-1}(\delta\eta_3\eta^T + \eta\delta\eta_3^T)\\
& - \delta\eta_3\eta^T(H + \eta\eta^T)^{-1}(\delta\eta_2\eta^T + \eta\delta\eta_2^T)\\
& - (\delta\eta_2\eta^T + \eta\delta\eta_2^T)(H + \eta\eta^T)^{-1}\eta\delta\eta_3^T\\
& - (\delta\eta_3\eta^T + \eta\delta\eta_3^T)(H + \eta\eta^T)^{-1}\eta\delta\eta_2^T\\
& + \delta\eta_2\delta\eta_3^T + \delta\eta_3\delta\eta_2^T,
\end{aligned}$$

then we have

$$g(\nabla_{(\partial_{\tilde{\Theta}})_3}(\partial_{\tilde{\Theta}})_2, (\partial_{\tilde{\Theta}})_1) = \delta^{(1)}_{\tilde{\Theta}_3}(\delta^{(1)}_{\tilde{\Theta}_2}(\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi})),$$
$$g^*(\nabla^*_{(\partial_{\tilde{H}})_3}(\partial_{\tilde{H}})_2, (\partial_{\tilde{H}})_1) = \delta^{(1)}_{\tilde{H}_3}(\delta^{(1)}_{\tilde{H}_2}(\delta^{(1)}_{\tilde{H}_1}\tilde{\phi})).$$

Proof. We first show the equation $g(\nabla_{(\partial_{\tilde{\Theta}})_3}(\partial_{\tilde{\Theta}})_2, (\partial_{\tilde{\Theta}})_1) = \delta^{(1)}_{\tilde{\Theta}_3}(\delta^{(1)}_{\tilde{\Theta}_2}(\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi}))$. From Lemma 4.5 and the facts

$$(i) \qquad \delta_{\tilde{\Theta}_3}[\frac{1}{2}Tr(\Theta^{-1}\delta\theta_2\delta\theta_1^T)] = -\frac{1}{2}Tr(\Theta^{-1}\delta\Theta_3\Theta^{-1}\delta\theta_2\delta\theta_1^T) + O(4),$$

$$\begin{aligned}
(ii) \qquad & \delta_{\tilde{\Theta}_3}[-\frac{1}{2}Tr(\Theta^{-1}\delta\Theta_j\Theta^{-1}\theta\delta\theta_i^T)]\\
= & -\frac{1}{2}Tr[\Theta^{-1}\delta\Theta_j\Theta^{-1}\delta\theta_3\delta\theta_i^T - \Theta^{-1}\delta\Theta_j\Theta^{-1}\delta\Theta_3\Theta^{-1}\theta\delta\theta_i^T\\
& - \Theta^{-1}\delta\Theta_3\Theta^{-1}\delta\Theta_j\Theta^{-1}\theta\delta\theta_i^T] + O(4), \qquad ((i,j) = (1,2), (2,1))
\end{aligned}$$

$$\begin{aligned}
(iii) \qquad & \delta_{\tilde{\Theta}_3}[\frac{1}{2}Tr(\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1)]\\
= & -Tr(\Theta^{-1}\delta\Theta_3\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1) + O(4),
\end{aligned}$$

$$(iv) \qquad \delta_{\tilde{\Theta}_3}[\frac{1}{2}Tr(\Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1)]$$

$$
\begin{aligned}
&= \frac{1}{2}Tr(-\Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1\Theta^{-1}\delta\Theta_3 \\
&\qquad - \Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_3\Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\Theta_1 \\
&\qquad - \Theta^{-1}\theta\theta^T\Theta^{-1}\delta\Theta_1\Theta^{-1}\delta\Theta_3\Theta^{-1}\delta\Theta_2 \\
&\quad + \Theta^{-1}\delta\Theta_2\Theta^{-1}\theta\delta\theta_3^T\Theta^{-1}\delta\Theta_1 + \Theta^{-1}\delta\Theta_2\Theta^{-1}\delta\theta_3\theta^T\Theta^{-1}\delta\Theta_1) + O(4),
\end{aligned}
$$

we obtain the equation

$$
\begin{aligned}
&\delta^{(1)}_{\tilde{\Theta}_3}(\delta^{(1)}_{\tilde{\Theta}_2}(\delta^{(1)}_{\tilde{\Theta}_1}\tilde{\psi})) \\
&\quad = Tr[\frac{1}{2}(\Theta^{-1}x - \Theta^{-1}X\Theta^{-1}\theta)\delta\theta_1^T \\
&\qquad + \frac{1}{2}(-\Theta^{-1}\theta x^T\Theta^{-1} + \Theta^{-1}X\Theta^{-1}\theta\theta^T\Theta^{-1} + \Theta^{-1}X\Theta^{-1})\delta\Theta_1^T] \\
&\quad = g(\nabla_{(\partial_{\tilde{\Theta}})_3}(\partial_{\tilde{\Theta}})_2, (\partial_{\tilde{\Theta}})_1)
\end{aligned}
$$

We omit the proof of the other equation because it is derived in the same way with tedious calculations.

□

We recall the Legendre transformation, i.e.,

$$
\mathfrak{L} \quad : \quad \mathfrak{D} \ni (\theta, \Theta) \mapsto (\eta, H) = (\frac{1}{2}\Theta^{-1}\theta, -\frac{1}{2}\Theta^{-1} - \frac{1}{4}\Theta^{-1}\theta\theta^T\Theta^{-1}) \in \mathfrak{D}^*.
$$

We show that the dual metrics g and g^* are transformed each other by the Legendre transformation, i.e.,

Lemma 4.7. *Let*

$$
\mathfrak{L}^* : T^*_{\tilde{H}}\mathfrak{D}^* \ni \delta\tilde{H} = (\delta\eta, \delta H) \mapsto \delta\tilde{\Theta} = (\delta\theta, \delta\Theta) \in T^*_{\tilde{\Theta}}\mathfrak{D}
$$

be defined by

$$
\begin{aligned}
\delta\theta &= (H + \eta\eta^T)^{-1}(\delta H + \delta\eta\eta^T + \eta\delta\eta^T)(H + \eta\eta^T)^{-1}\eta - (H + \eta\eta^T)^{-1}\delta\eta, \\
\delta\Theta &= \frac{1}{2}(H + \eta\eta^T)^{-1}(\delta H + \delta\eta\eta^T + \eta\delta\eta^T)(H + \eta\eta^T)^{-1}.
\end{aligned}
$$

Then we have

$$
\mathfrak{L}^*(g^*) = g.
$$

The transformation $\mathfrak{L}^*$ is derived from the Legendre transformation as follows. Differentiating the equation $\eta = \frac{1}{2}\Theta^{-1}\theta$, we have

$$
2\delta\Theta\eta + 2\Theta\delta\eta = \delta\theta. \tag{35}
$$

On the other hand, differentiating the equation $-(H+\eta\eta^T)=\frac{1}{2}\Theta^{-1}$, we have

$$\delta\Theta = 2\Theta(\delta H + \delta\eta^T + \eta\delta\eta^T)\Theta. \tag{36}$$

Substituting the equation (36) into the equation (35), we have the transformation $\mathfrak{L}^*$.

For the proof of Lemma 4.7., we should prepare the second order differential expansion of the potential function $\tilde{\psi}(\tilde{\Theta})$, which is based on the following lemma.

Lemma 4.8. *For any n by n matrix X and its eigenvalues λ_i ($i = 1,2,\ldots,n$), if $\max_{1\leqq i\leqq n}|\lambda_i|<1$ then*

$$\log\det(I+X) = Tr[\log(I+X)].$$

Now the second order differential expansion of the potential function $\tilde{\psi}$ is given as follows.

$$\begin{aligned}
\delta\tilde{\psi} &= \tilde{\psi}(\theta+\delta\theta,\Theta+\delta\Theta)-\tilde{\psi}(\theta,\Theta)\\
&= \frac{1}{4}Tr\left[\left\{\Theta(I+\Theta^{-1}\delta\Theta)\right\}^{-1}(\theta\theta^T+\theta\delta\theta^T+\delta\theta\theta^T+\delta\theta\delta\theta^T)\right]\\
&\quad -\frac{1}{2}\log\left[\det\left\{\Theta(I+\Theta^{-1}\delta\Theta)\right\}\right]-\frac{1}{4}Tr(\Theta^{-1}\theta\theta^T)+\frac{1}{2}\log(\det\Theta).
\end{aligned}$$

From $(I+\Theta^{-1}\delta\Theta)^{-1}=I-\Theta^{-1}\delta\Theta+(\Theta^{-1}\delta\Theta)^2+O(3)$ and Lemma 4.8, we have

$$\delta\tilde{\psi} = \delta^{(1)}\tilde{\psi}+\frac{1}{2!}\delta^{(2)}\tilde{\psi}+O(3),$$

where the differentials $\delta^{(i)}\tilde{\psi}$ $(i=1,2)$ are given as

$$\begin{aligned}
\delta^{(1)}\tilde{\psi} &= Tr[(\frac{1}{2}\Theta^{-1}\theta)\delta\theta^T-(\frac{1}{2}\Theta^{-1}+\frac{1}{4}\Theta^{-1}\theta\theta^T\Theta^{-1})\delta\Theta]\\
&= Tr[\eta\delta\theta^T+H\delta\Theta^T] =< \tilde{H},\delta\tilde{\Theta}>,\\
\delta^{(2)}\tilde{\psi} &= \frac{1}{2}Tr[\Theta^{-1}\delta\theta\delta\theta^T-2\Theta^{-1}\delta\Theta\Theta^{-1}\delta\theta\theta^T+(\Theta^{-1}\delta\Theta)^2\Theta^{-1}\theta\theta^T\\
&\qquad +(\Theta^{-1}\delta\Theta)^2].
\end{aligned}$$

We next consider the second order differential expansion of the dual potential function $\tilde{\phi}$.

$$\begin{aligned}
\delta\tilde{\phi} &= -\frac{1}{2}\log[\det\{-H-\delta H-(\eta+\delta\eta)(\eta+\delta\eta)^T]+\frac{1}{2}\log[\det(-H-\eta\eta^T)]\\
&= -\frac{1}{2}\log[\det\{I_n+(H+\eta\eta^T)(\delta H+\eta\delta\eta^T+\delta\eta\eta^T+\delta\eta\delta\eta^T)\}]
\end{aligned}$$

$$= -\frac{1}{2}Tr[(H+\eta\eta^T)^{-1}(\delta H+\eta\delta\eta^T+\delta\eta\eta^T+\delta\eta\delta\eta^T)$$
$$-\frac{1}{2}\{(H+\eta\eta^T)^{-1}(\delta H+\eta\delta\eta^T+\delta\eta\eta^T+\delta\eta\delta\eta^T)\}^2+O(3)].$$

Hence, we have

$$\delta^{(1)}\tilde{\phi} = -\frac{1}{2}Tr[(H+\eta\eta^T)^{-1}(\delta H+\eta\delta\eta^T+\delta\eta\eta^T)].$$
$$\delta^{(2)}\tilde{\phi} = Tr[-(H+\eta\eta^T)^{-1}\delta\eta\delta\eta^T+\{(H+\eta\eta^T)^{-1}\eta\delta\eta^T\}^2$$
$$+2(H+\eta\eta^T)^{-1}\delta H(H+\eta\eta^T)^{-1}\eta\delta\eta^T$$
$$+(H+\eta\eta^T)^{-1}\eta\delta\eta^T(H+\eta\eta^T)^{-1}\delta\eta\eta^T+\frac{1}{2}\{(H+\eta\eta^T)^{-1}\delta H\}^2].$$

Proof of Lemma 4.7. Decomposing the second order differential $\delta^{(2)}\tilde{\psi}$ into four parts, we substitute the increment of $\tilde{\Theta}$ into these parts.

$$(i)\quad \frac{1}{2}Tr(\Theta^{-1}\delta\theta\delta\theta^T)$$
$$= Tr[-(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)\eta\eta^T(H+\eta\eta)^{-1}(\delta H+\delta\eta^T+\eta\delta^T)$$
$$\times(H+\eta\eta^T)^{-1}$$
$$+2(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)\eta\delta\eta^T(H+\eta\eta^T)^{-1}$$
$$-\delta\eta\delta\eta^T(H+\eta\eta^T)^{-1}].$$
$$(ii)\quad -Tr[\Theta^{-1}\delta\Theta\Theta^{-1}\delta\theta\theta^T]$$
$$= 2Tr[(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}(\delta H+\delta\eta^T+\eta\delta\eta^T)$$
$$\times(H+\eta\eta^T)^{-1}\eta\eta^T(H+\eta\eta^T)^{-1}$$
$$-(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}\delta\eta\eta^T(H+\eta\eta^T)^{-1}].$$
$$(iii)\quad \frac{1}{2}Tr[(\Theta^{-1}\delta\Theta)^2\Theta^{-1}\theta\theta^T]$$
$$= -Tr[(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}(\delta H+\delta\eta\eta+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}$$
$$\times\eta\eta^T(H+\eta\eta^T)]$$
$$(iv)\quad \frac{1}{2}Tr(\Theta^{-1}\delta\Theta)^2=\frac{1}{2}Tr[(\delta H+\delta\eta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}]^2.$$

From previous quantities, we have

$$(37)\quad (\mathfrak{L}^*)^{-1}(g)$$
$$= -Tr[(H+\eta\eta^T)^{-1}\delta\eta\delta\eta^T]+\frac{1}{2}Tr[(\delta H+\delta\eta^T+\eta\delta\eta^T)(H+\eta\eta^T)^{-1}]^2.$$

Hence, we obtain the equation for the metrics.

□

From previous observations, we have one of two main theorems in this section.

Theorem 4.9. *Under the assumptions of Lemma 4.5.,Lemma 4.6., and Lemma 4.7.,*

(i) The connection ∇ *satisfies the equation* $Xg(Y,Z) = g(\nabla_X Y, Z)$ *for any vector fields* X, Y *and* Z *on* $\mathfrak{D}$.

(ii) The connection ∇^* *satisfies the equation* $Xg^*(Y,Z) = g^*(Y, \nabla^*_X Z)$ *for any vector fields* X, Y *and* Z *on* $\mathfrak{D}^*$.

(iii) Let $\mathfrak{X}, \mathfrak{Y}$ *and* $\mathfrak{Z}$ *be three vector fields on* $\mathfrak{N}$. *If we pull back the metric* g *and the connections* ∇ *and* ∇^* *by the maps* $\mathfrak{I}_{\tilde{\Theta}}$ *and* $\mathfrak{I}_{\tilde{H}}$, *i.e.,*

$$\hat{g} = \mathfrak{I}_{\tilde{\Theta}}{}^*(g), \quad \widehat{\nabla} = 2\mathfrak{I}_{\tilde{\Theta}}{}^*(\nabla), \quad \widehat{\nabla}^* = 2\mathfrak{I}_{\tilde{H}}{}^*(\nabla^*),$$

then the two connections $\widehat{\nabla}$ *and* $\widehat{\nabla}^*$ *become the dual connections with respect to the metric* $\hat{g}$ *on* $\mathfrak{N}$, *i.e.,*

$$\mathfrak{X}\hat{g}(\mathfrak{Y}, \mathfrak{Z}) = \hat{g}(\widehat{\nabla}_{\mathfrak{X}}\mathfrak{Y}, \mathfrak{Z}) + \hat{g}(\mathfrak{Y}, \widehat{\nabla}^*_{\mathfrak{X}}\mathfrak{Z}).$$

Proof. From Lemma 4.5 and Lemma 4.6, we find (i) and (ii). From Lemma 4.7., we find $\hat{g} = \mathfrak{I}_{\tilde{\Theta}}{}^*(g) = \mathfrak{I}_{\tilde{H}}{}^*(g^*)$. Adding the two equations in (i) and (ii), we have the result (iii).

□

Remark that Theorem 4.9 implies the theorem 1 in Ohara-Suda-Amari [13] as a special case. In fact, we can find that the connections defined in Lemma 4.6 conform to the dual connections in [13] by putting $\theta = \eta = 0 \in R^n$. While Ohara-Suda-Amari [13] treated the connections on the cone of symmetric positive definite matrices, the space $\mathfrak{D}^*$ treated in this paper is not a cone. Note also that Mitchell [12] obtained the coefficients of the α-connection ($\alpha = 0$, Skovgaard [16]) for the original parameter (μ, Σ).

The Kullback-Leibler information originated in information theory for measuring the divergence between two distributions. The following result is a consequence of above propositions.

Theorem 4.10. *Let* $\Xi_1 = (\mu_1, \Sigma_1) \in R^n \times \mathfrak{S}_n^+$, $\Xi_2 = (\mu_2, \Sigma_2) \in R^n \times \mathfrak{S}_n^+$ *and* $\tilde{\Theta}_2 = (\theta_2, \Theta_2) = (\Sigma_2^{-1}\mu_2, \frac{1}{2}\Sigma_2^{-1}) \in \mathfrak{D}$, $\tilde{H}_1 = (\eta_1, H_1) = (\mu_1, -(\Sigma_1 + \mu_1\mu_1^T)) \in \mathfrak{D}^*$. *We have the divergence Div defined on* $\mathfrak{N} \times \mathfrak{N}$

$$\begin{aligned} Div(N(\mu_2, \Sigma_2), N(\mu_1, \Sigma_1)) &\equiv Div(\mathfrak{I}_{\tilde{\Theta}}^{-1}(\tilde{\Theta}_2), \mathfrak{I}_{\tilde{H}}^{-1}(\tilde{H}_1)) \\ &\equiv \widetilde{Div}(\tilde{\Theta}_2, \tilde{H}_1) \equiv \tilde{\psi}(\tilde{\Theta}_2) + \tilde{\phi}(\tilde{H}_1) - < \tilde{\Theta}_2, \tilde{H}_1 > \\ &= \int_{R^n} p(x; \Xi_1) \log \frac{p(x; \Xi_1)}{p(x; \Xi_2)} dx, \end{aligned}$$

where $p(x;\Xi_i)$ is the density function of the Gaussian distribution with respect to the Lebesgue measure dx.

Proof. In fact,

$$\begin{aligned}
&\tilde{\psi}(\tilde{\Theta}_2)+\tilde{\phi}(\tilde{H}_1)- <\tilde{\Theta}_2,\tilde{H}_1> \\
&= \left[\frac{1}{4}Tr(\Theta_2^{-1}\theta_2\theta_2^T)-\frac{1}{2}\log(\det\Theta_2)\right]+\left[-\frac{1}{2}\log\left\{\det(-H_1-\eta_1\eta_1^T)\right\}\right] \\
&- \left[\theta_2^T\eta_1+Tr(\Theta_2 H_1^T)\right]-\frac{n}{2} \\
&= \left[\frac{1}{2}Tr(\Sigma_2^{-1}\mu_2\mu_2^T)+\frac{1}{2}\log(\det\Sigma_2)\right]+\left[-\frac{1}{2}\log(\det\Sigma_1)\right] \\
&- \left[Tr(\Sigma_2^{-1}\mu_2\mu_1^T)-\frac{1}{2}Tr\left\{\Sigma_2^{-1}(\Sigma_1+\mu_1\mu_1^T)\right\}\right]-\frac{n}{2} \\
&= \frac{1}{2}\log\left(\frac{\det\Sigma_2}{\det\Sigma_1}\right)+\frac{1}{2}Tr[\Sigma_1\Sigma_2^{-1}+\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T]-\frac{n}{2} \\
&= \int_{R^n} p(x;\Xi_1)\log\frac{p(x;\Xi_1)}{p(x;\Xi_2)}dx.
\end{aligned}$$

□

Remark 4.1. The definition of our *divergence* is different from that of *divergence* in [2,9] because of the definition (8).

Acknowledgment . The authors wish to thank the referees and Professor Kenro Furutani for various suggestions for improving the paper.

References

[1] T.W. Anderson and I. Olkin, *Maximum-Likelihood estimation of the parameters of a multivariate normal distribution*, Linear Algebra and Its Applications, 70, 141-171 (1985).

[2] S. Amari, *Differential geometrical method in statistics*, Lec. Notes in Statist. Vol.28, Springer Verlag, Berlin (1985).

[3] S. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Laurizen and C.R.Rao, *Differential geometry in statistical inferences*, IMS Lec. Notes Monograph Ser. Vol.10, Inst. Math. Statist., Hayward (1987).

[4] O.E. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley,Chichester (1978).

[5] O.E. Barndorff-Nielsen and P. Blæsild, *Exponential models with affine dual foliations*, Ann.Statist. 11, 735-769 (1983).

[6] E. Calabi, *Improper affine hyperspheres of convex type and a generalization of a theorem by K.Jörgens*, Michigan Math.J., Vol.5, 105-126 (1958).

[7] C. Chen and S. Yau, *On the regularity of the Monge-Ampere equation*, Comm.Pure Appl.Math. Vol XXX, 41-68 (1977).

[8] B. Efron, *Defining the curvature of a statistical problem (with Discussion)*, Ann. Statist. 6, 1189-1242 (1975).

[9] A. Fujiwara and S. Amari, *Gradient systems in view of information geometry*, Physica D80, No.3, 317-327 (1995).

[10] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Second Edition, (1976).

[11] E.H. Lieb, *Convex Trace Functions and the Wigner-Yanase-Dyson Conjecture*, Advances in Mathematics, 11, 267-288 (1973).

[12] A.F.S. Mitchell, *The informatiton matrix,skewness tensor and α-connections for the general multivariate elliptic distribution*, Ann.Inst.Statist.Math. Vol.41, No.2, 289-304 (1989).

[13] A. Ohara, N. Suda and S. Amari, *Dualistic Differential Geometry of Positive Definite Matrices and Its Applications to Related Problems*, Linear Algebra and Its Applications, 247, 31-53 (1996).

[14] C.R. Rao and M.B. Rao, *Matrix Algebra and Its Applications to Statistics and Econometrics*, World Scientific (1998).

[15] H. Shima, *Hessian manifolds of constant Hessian sectional curvature*, J.Math.Soc.Japan. Vol.47, No.4, 735-753 (1995).

[16] L.T. Skovgaard, *A Riemannian geometry of the multivariate normal model*, Scand.J.Statist., 11, 211-223 (1984).

Shintaro Yoshizawa
The Graduate University for Advanced Studies
4-6-7 Minami-Azabu Minato-ku Tokyo 106-8569 Japan.
E-mail address: yosizawa@ism.ac.jp

Kunio Tanabe
The institute of statistical mathematics
4-6-7 Minami-Azabu Minato-ku Tokyo 106-8569 Japan.
E-mail address: tanabe@ism.ac.jp