Differential Analytic aspects of statistical models

M. Sitaramayya[*1], K.S.S. Moosath[2] and S.N. Hasan[3]

[1] *School of Maths & Statistics, University of Hyderabad, Hyderabad, India*
[2] *Dept. of Maths., IIST, Valiamalai,Thiruvananthapuram,Kerala, India*
[3] *Dept. of Maths.,MANU University,Gachibowli, Hyderabad, India*

malladi_sr@yahoo.com

Abstract

Statistical models carry mathematical structures new to conventional mathematics such as Dual structures or Amari-Chentsov (AC) structures and in particular dually flat structures or Bregman (or exponential) families. These dual flat Riemannian manifolds are natural generalizations of Euclidean spaces and enjoy properties like pythagorean theorem and orthogonal projection theorem in general form. They are universal models of Riemannian spaces. The topological aspects of dual flat spaces and obstructions for a Riemannian space to be dually flat and Hessian are analyzed. In fact dual spaces are curved submanifolds of dual flat spaces. Convex dual data living on dual flat spaces is brought out via) convex analysis of Legendre. The problem when a Riemannian metric becomes a Hessian metric is analyzed via) AC structures (M, G, T) and some open problems are suggested.

1 Introduction

Here we analyze the structure living on statistical models namely, Dual structures or A-C structures; dually flat structures which are new in mathematical structures. We study these models as differentiable manifolds with a special Riemannian metric called the Fisher-Rao metric g. We study how they generalize the Euclidean spaces. Since Euclidean distance is an integrated version of Euclidean metric we do the similar analysis in the reverse way to recover the metric from the distance. We do this in statistical models via the concept of divergences and recover the F-R metric from the divergence. The space of discrete probability distributions S_N on a finite set and the space of exponential families of probability distributions $\mathcal{E}$ of a contnuous random variable and the family $\mathfrak{F}$ of all probability distributions on an infinite sample space are good examples of statistical structures and serve as universal models.

We analyse the space of Bregman divergences $\mathbb{B}$ and the space $\mathcal{E}$ of all exponential families of probability distributions and study the correspondence between them.

Then we study the space of dually flat Riemannian spaces and the space of dual structures and its relation to space of general divergences (non flat ones). Finally we analyze when a Riemannian metric g on M becomes a Hessian metric and the resulting topological obstructions.

2 Statistical models as (differentiable) manifolds

In this section we give classical examples from statistics of differentiable manifolds. **Example 2.1:** Manifold of probability distributions

(a) Gaussian or normal distributions of a continuous random variable ($r.v.$) Recall the probability density function of a Gaussian random variable x is given by

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.1}$$

where μ is the mean and σ^2 is the variance. Hence we can regard the set $\mathcal{N}$ of all the normal distributions of $x \in \mathbb{R}$ as a 2-dimensional differential manifold where a point in $\mathcal{N}$ denotes a probability density function and

$$\xi = (\mu,\sigma), \qquad \sigma > 0 \tag{2.2}$$

is a coordinate system which is topologically equivalent to $H^\dagger$ upper half plane of Euclidean space $\mathbb{R}^2$. This manifold $\mathcal{N}$ is covered by a single coordinate system. There are also other coordinate systems such as $\zeta = (m_1, m_2)$ (called the moment coordinate system) where

$$m_1 = E[x] = \mu \qquad \text{and} \qquad m_2 = E[x^2] = \mu^2 + \sigma^2 \tag{2.3}$$

and E is the expectation of a random variable.
Also $\underline{\theta} = (\theta_1, \theta_2)$ given by

$$\theta_1 = \frac{\mu}{\sigma^2} \quad \text{and} \quad \theta_2 = \frac{-1}{2\sigma^2} \tag{2.4}$$

gives another coordinate system on $\mathcal{N}$ of natural parameters.

(b) Discrete $r.v.$ case.

Let $x$be a discrete $r.v$ taking values on $X = \{0, 1, 2, \cdots, N\}$. A probability distribution $p(x)$ is specified by $N+1$ probabilities

$$p_i = \text{prob.}(x = i), \qquad i = 0, 1, 2, \cdots, N \tag{2.5}$$

so that $p(x)$ is represented by a probability vector $\underline{p} = (p_o, p_1, \cdots, p_N)$ satisfying

$$\sum_{i=0}^{N} p_i = 1, \qquad p_i > 0 \tag{2.6}$$

Then the set S_N of all probability distributions $\underline{p}$ forms an N-dimensional manifold with its coordinate system

$$\xi = (p_1, p_2, \cdots, p_N) \tag{2.7}$$

and p_o is a functions of these coordinates as

$$p_o = 1 - \sum_{i=1}^{N} \xi_i \tag{2.8}$$

This S_N is an N-dimensional (open) simplex called the probability simplex. S_2 is the interior of a triangle and S_3 is the interior of a tetrahedron etc
δ-interpretation of S_N: Define $N+1$ random variables $\delta_i(x), i = 0, 1, 2, \cdots, N$ by

$$\begin{aligned} \delta_i(x) &= 1 \quad x = i \\ &= 0 \quad x \neq i \end{aligned} \tag{2.9}$$

Then a probability distribution of x is denoted by

$$p(x;\xi) = \sum_{i=1}^{N} \xi_i \delta_i(x) + p_o(\xi)\delta_o(x) \tag{2.10}$$

in terms of coordinates $\underline{\xi}$.
On S_N we have another coordinate system $\underline{\theta}$ given by

$$\theta_i = \log \frac{p_i}{p_o}, \qquad i = 1, 2, \cdots, N \tag{2.11}$$

(c) Regular statistical model as a differential manifold.

Let $\underline{x}$ be a random variable which may take discrete, scalar or vector continuous values. Then a statistical model is a family of probability distributions $M = \{p(x;\xi)\}$ specified by a *vector* parameter ξ. When it satisfies certain regularity conditions (see [20]) M is called a regular statistical model. Such a model M is a (differential) manifold where $\underline{\xi}$ gives a coordinate system. Example (a) and (b) above are regular statistical models. Note that information geometry may be defined as the study of invariant geometrical structures of regular statistical model.

(d) Manifold of positive measures.

Let x be a variable taking values in a set $N = \{1, 2, \cdots, n\}$.

we assign a positive measure or weight m_i to each element i, $i = 1, 2, \cdots, n$. Then

$$\xi = (m_1, m_2, \cdots, m_n), \qquad m_i > 0 \tag{2.12}$$

defines a distribution of measures over N. The set $\mathcal{M}$ of all such measures sits in the first quadrant $\mathbb{R}^n_+$ of an n-dimensional Euclidean space $\mathbb{R}^n$ and is an n-dimensional manifold. the sum $m = \sum_{i=1}^{n} m_i$ is called the total mass of $\underline{m} = (m_1, m_2, \cdots, m_n)$. Such measure vector $\underline{m}$ with total mass equal to 1 is a probability distribution belonging to the simplex S_{n-1}. Hence S_{n-1} is included in $\mathbb{R}^n_+$ as submanifold.

(e) manifold of positive definite matrices.

Let A be a $n \times n$ matrix over $\mathbb{R}$. They form n^2-dimensional manifold $\mathcal{M}$. When A is a symmetric and positive definite they form a $n(n+1)/2$-dimensional manifold P which is a submanifold of $\mathcal{M}$ with their upper right elements including the diagonal giving a coordinate system.

(f) Neural manifold.

A neural network is composed of a large number of neurons connected with each other, where the dynamics of information processing takes place. A network is specified by connection weights w_{ji} connecting neuron i with neuron j. the set of all such networks forms a manifold where $W = (w_{ji})$ is a coordinate system.

(g) Exponential family of probability distributions.
Define a probability density function by

$$p(x,\theta) = \exp\left[\sum_{i=1}^{n} \theta_i x_i + k(x) - \psi(\theta)\right] \tag{2.13}$$

of vector random variable $\underline{x}$ specified by vector parameter $\underline{\theta}$ and $k(x)$ is a function of $\underline{x}$ satisfying

$$\int p(x,\theta)dx = 1 \tag{2.14}$$

Then $\psi(\theta)$ is given by

$$\psi(\theta) = \log \int \exp\left[\sum_{i=1}^{n} \theta_i x_i + k(x)\right] dx \tag{2.15}$$

Let $\mathcal{M} = \{p(\underline{x}, \underline{\theta})\}$. Then $\mathcal{M}$ is regarded as a manifold and $\underline{\theta}$ is its coordinate system, called an exponential family of probability distributions.

(h) Infinite dimensional Manifold of probability distribution (formal approach).
Let x be a continuous random variable. Consider the set $\mathfrak{F}$ of all probability density functions $p(x)$ which is an infinite dimensional function space which can be regarded as an infinite dimensional manifold enjoying geometry similar to those of S_N probability simplex. (we discuss its geometry later).

3 Divergence and Fisher-Rao metric on a Statistical model

Let $S = \{p(x, \xi = \theta)\}$ be a n-dimensional statistical model. Let $l_\theta(x) = \log p_\theta(x) = \log p(x,\theta)$ be log-density function. Then the Fisher information matrix $(g_{ij}(\theta)) = G$ is given by

$$g_{ij}(\theta) = E_\theta[\partial_i l_\theta \partial_j l_\theta] = \int_\chi \partial_i l_\theta(x) \partial_j l_\theta(x) p(x,\theta) dx \qquad \partial_i = \frac{\partial}{\partial \theta^i} \tag{3.1}$$

which is a symmetric positive definite matrix and hence defines an inner product on the tangent space $T_\theta(S)$ at θ depending smoothly on $\theta \in \Theta$ open $\subset \mathbb{R}^n$ and hence defines a Riemannian metric on the parameter space Θ called the Fisher-Rao metric of S. (cf.[20] for details).

The general philosophy is the geometry of S will be studied thru several invariant tensors like G, T etc., we realize these tensors by a generalized concept of "metric-like distance" called "divergences" by a "differentiation process".

We define the concept of divergence analytically first, (geometrically later) and derive a special class of divergences defined by a general convex function ψ on a statistical model M.

Definition 3.1 Let $M = \{p(x, \xi)\}$ be a parametrized statistical model. Let $P, Q \in M$ with coordinates ξ_P and ξ_Q respectively. We define a differentiable function $D : M \times M \to \mathbb{R}$ by

$$D(P, Q) = D(\xi_P, \xi_Q) \tag{3.2}$$

and is called a divergence if

1. $D(P, Q) \geq 0$.
2. $D(P, Q) = 0 \Leftrightarrow P = Q$
3. When P, Q points are sufficiently close with coordinates ξ_P and $\xi_Q = \xi_P + d\xi$

$$D(\xi_P, \xi_p + d\xi) = \frac{1}{2} \sum g_{ij}(\xi_p) d\xi_i d\xi_j + \bigcirc(|d\xi|^3) \tag{3.3}$$

and $G = (g_{ij})$ is a positive definite matrix depending smoothly on ξ_P.

Remark: Divergence measures the separation between two points of the model, not symmetric $D(P, Q) \neq D(Q, P)$ and triangular inequality also fails in general.

If P and Q are sufficiently close we define the square of the infinitesimal distance ds between them using (18) by

$$ds^2 = 2D[\xi, \xi + d\xi] = \sum g_{ij}(\xi) d\xi_i d\xi_j \tag{3.4}$$

Since $G = (g_{ij}(\xi))$ is $p.d$ at each point ξ of M (19) shows that divergence D provides M with a Riemannian metric.

3.2 Examples:

(a) $M = (\mathbb{R}^n, <, >)$ Euclidean space. Then define

$$D[P, Q] = \frac{1}{2} \sum_{i=1}^{n} (\xi_P^i - \xi_Q^i)^2 \tag{3.5}$$

and then $G - Id_{n \times n}$ matrix and

$$ds^2 = \sum (d\xi_i)^2 \tag{3.6}$$

(Euclidean divergence)

(b) Kullback-Leibler Divergence: Let $p(x), q(x)$ be two probability distributions of a $r.v.x$ in a manifold M of probability distributions. Define

$$D_{KL}[p(x), q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{3.7}$$

when x is a discrete $r.v.$ replace integral by summation.
KL-divergence for positive measures on $\mathbb{R}^n_+$ is defined by

$$D_{KL}[\underline{m}_1, \underline{m}_2] = \sum m_{1i} \log \frac{m_{1i}}{m_{2i}} - \sum m_{1i} + \sum m_{2i} \tag{3.8}$$

which for probability measures reduces to (22) (extended KL divergence).

(c) Divergences on $p.d$ matrices $\mathbb{P}$: Define for $P, Q \in \mathbb{P}$

$$(i) \quad D[P, Q] = tr(P \log P - P \log Q - P + Q) \tag{3.9}$$
(Von-Neumann divergence)

$$(ii) \quad D[P, Q] = tr[PQ^{-1}] - \log[\det(PQ^{-1})] - n \tag{3.10}$$
(KL-divergence for multivariate normal distribution)

$$(iii) \quad D[P, Q] = \frac{4}{1-\alpha^2} tr \left[(-P^{\frac{1-\alpha}{2}} Q^{\frac{1+\alpha}{2}} + \frac{1-\alpha}{2} P + \frac{1+\alpha}{2} Q) \right], \tag{3.11}$$
$\alpha \in \mathbb{R}$ (α-divergence)

Now we specialize to a class $\mathbb{B}$ of divergences on a statistical model, M called Bregman divergences defined by a convex function ψ on M.

Definition 3.3. A nonlinear differentiable function $\psi(\xi)$ of coordinate ξ on M is called convex if

$$\lambda\psi(\xi_1) + (1-\lambda)\psi(\xi_2) \geq \psi[\lambda\xi_1 + (1-\lambda)\xi_2] \tag{3.12}$$

holds for each $\xi_1, \xi_2 \in M$ and scalar λ in $0 \leq \lambda \leq 1$.
Then a function ψ is convex iff its Heissian

$$H(\xi) = \left(\frac{\partial^2}{\partial\xi_i \partial\xi_j} \psi(\xi) \right) \tag{3.13}$$

is positive definite.
Example (1):

$$\psi(\xi) = \frac{1}{2} \sum \xi_i^2 \tag{3.14}$$

is a convex function on Euclidean space.
(2) On the probability simplex S_n,

$$\varphi(p) = -H(p) = \sum p_i \log p_i \tag{3.15}$$

(negative entropy) is a convex function.
(3) For the exponential family statistical model, the cumulant generating function

$$\psi(\theta) = \log \int \exp\left[\sum \theta_i x_i + k(x)\right] dx \tag{3.16}$$

can be shown to be a convex function namely $\nabla\psi(\theta) = E(\underline{x})$ and $\nabla\nabla\psi(\theta) = Hess(\psi) = Var(\underline{x})$ cov. matrix of $\underline{x}$ which is positive definite.

3.4. Bregman divergence of a convex function ψ on M.

Consider the graph of $z = \psi(\xi)$. Let $\xi_o \in M$. Then the tangent hyper plane H to M at ξ_o is given by

$$z = \psi(\xi_o) + \nabla\psi(\xi_o) \cdot (\xi - \xi_o) \tag{3.17}$$

Define the divergence $D_\psi[\xi, \xi_o]$ =the vertical height of $\psi(\xi)$ from the tangent hyper plane by

$$D_\psi(\xi, \xi_o) = \psi(\xi) - \psi(\xi_o) - \nabla\psi(\xi_o) \cdot (\xi - \xi_o) \tag{3.18}$$

on M called the Bregmann divergence defined by the convex function ψ.
Examples (1) $\psi(\xi) = \frac{1}{2}\sum \xi_i^2$. Then $D[\xi, \xi_o] = \frac{1}{2}|\underline{\xi} - \underline{\xi}_o|^2$ (Euclidean divergence).
(2) $\psi(\xi) = -\sum_{i=1}^n \log \xi_i$ convex function on $\mathbb{R}_+^n$. Then

$$D[\xi, \xi'] = \sum (\xi_i \log \frac{\xi_i}{\xi_i'} - \xi_i + \xi_i')$$

(3) For the exponential family $M = \{p(\underline{x}, \underline{\theta})\}$ the Bregmann divergence is defined by

$$D_\psi[\underline{\theta}, \underline{\theta}'] = \psi(\theta) - \psi(\theta') - \nabla\psi(\theta') \cdot (\theta - \theta') \tag{3.19}$$

$$= D_{KL}[p(x, \theta'), p(x, \theta)] = \int p(x, \theta') \log \frac{p(x, \theta')}{p(x, \theta)}\, dx \tag{3.20}$$

3.5 Convex analysis associated with convex function ψ on M. Denote by

$$\xi^* = \nabla\psi(\xi) \tag{3.21}$$

the gradient vector of ψ at ξ which is the normal vector $\underline{n}$ on the supporting Hyper plane H at ξ and hence the coordinate transformation from ξ to ξ^* by (36) is 1-1 and differentiable. Hence a point ξ on M can be represented by the coordinate ξ^* is called the Legendre transform of ξ by (36) and we say the two coordinate system (ξ) and (ξ^*) are coupled by (36) the Legendre transformation. The Legendre transformation defines a dualistic structure on M as follows: Define a new function of ξ^* by

$$\psi^*(\xi^*) = \xi \cdot \xi^* - \psi(\xi) \tag{3.22}$$

where $\xi, \xi^* = \sum \xi_i \xi_i^*$ and ξ is a function of ξ^* say

$$\xi = f(\xi^*) \tag{3.23}$$

and (38) defines the inverse function of (36). Differentiating (37) w.r.t. ξ^* we get

$$\nabla\psi^*(\xi^*) = \xi + \frac{\partial\xi}{\partial\xi^*}\xi^* - \nabla\psi(\xi)\frac{\partial\xi}{\partial\xi^*} = \xi \quad \text{by (36)}$$

Thus we have the dualistic structure given by $\xi^* = \nabla\psi(\xi)$ and $\xi = \nabla\psi^*(\xi^*)$ where ψ^* is defined by (37).
ψ^* is called the Legendre dual of ψ and ψ^* is also a convex function of ξ^* because $G(\xi) = \nabla\nabla\psi(\xi)$ is p.d. Hessian as ψ is convex and $G^*(\xi^*) = \nabla\nabla\psi^*(\xi^*) = \frac{\partial\xi}{\partial\xi^*} = G^{-1}(\xi) = \frac{\partial\xi^*}{\partial\xi}$ which is p.d. ψ and ψ^* define their Bregmann divergences D_ψ and D_{ψ^*} by

$$D_\psi(\xi, \xi') = \psi(\xi) - \psi(\xi') - \nabla\psi(\xi') \cdot (\xi - \xi')$$

and

$$D_{\psi^*}(\xi^*, \xi^{*\prime}) = \psi^*(\xi^*) - \psi^*(\xi^{*\prime}) - \nabla\psi^*(\xi^{*\prime}) \cdot (\xi^* - \xi^{*\prime}) \tag{3.24}$$

and they are related by

$$D_{\psi^*}[\xi^*, \xi^{*\prime}] = D_\psi[\xi', \xi] \tag{3.25}$$

(same upto order of points of M).
Remark: Using Legendre duals of ξ and ψ one can derive

$$D_\psi[P, Q] = \psi(\xi_P) + \psi^*(\xi^*_Q) - \xi_P \cdot \xi^*_Q$$

Examples of convex duals:

1) $\psi(\xi) = \frac{1}{2}|\xi|^2$ and its dual is $\psi^*(\xi^*) = \frac{1}{2}|\xi^*|^2$ self-dual convex function

2) For $\psi(\xi) = -\sum_{i=1}^{n} \log \xi_i$, its convex dual is $\psi^*(\xi^*) = -\sum[1 + \log(-\xi_i)]$ (3.26)

3) For $\psi(\xi) = \sum \xi_i \log \xi_i$, its convex dual is $\psi^*(\xi^*) = \sum \exp[\xi_i^* - 1]$ (3.27)

4) For the $\psi(\theta)$ of exponential family, its convex dual is

$$\psi^*(\xi^*) = \xi^* \cdot \xi - \psi(\xi) = \text{negative entropy} = \int p(x,\theta) \log p(x,\theta) dx \tag{3.28}$$

and $D_{\psi^*}[\theta^*, \theta^{*\prime}] = D_{KL}[p(x,\theta), p(x,\theta')]$ (3.29)
where $\theta = \nabla\psi^*(\theta^*)$ and $\theta' = \nabla\psi^*(\theta^{*\prime})$.

4 Geometry of M with a convex function ψ

Let θ be a coordinate system in which $\psi(\theta)$ is convex. Take θ to be an affine coordinate system and M is locally flat in θ-coordinate system called primal flatness. Similarly $\theta^* = \nabla\psi(\theta)$ is another type of coordinate system in which $\psi^*(\theta^*)$ is convex. Then M has a dual

affine structure defined by $\psi^*(\theta^*)$ and M is flat in this θ^*-coordinate system called the dual flatness. Thus M is a dually flat manifold with θ-flatness (primal flat) and θ^*-flatness (dual flatness) and the two flat coordinates θ and θ' are connected by Legendre transformation.

4.1. Tangent space, basis vectors and Riemannian metric of M.

Recall

$$ds^2 = 2D_\psi[\theta, \theta + d\theta] = \sum g_{ij} d\theta^i d\theta^j \tag{4.1}$$

and

$$g_{ij}(\theta) = \frac{\partial^2}{\partial\theta^i \partial\theta^j}\psi(\theta) \tag{4.2}$$

Let $\{e_i,\ i = 1, 2, n\}$ be a set of tangent vectors along the coordinate curves of θ. Then the span of $\{e_i\}$ is the tangent space of M at each point. θ is affine coordinate system implies $\{e_i\}$ basis everywhere. Hence $\forall\ A \in T(M)$ is

$$A = \sum A^i e_i \tag{4.3}$$

similarly for θ^*-affine coordinate system, $\{e^{*i}\}$ basis vectors of dual affine coordinate curves with

$$A = \sum A_i e^{*i}\,. \tag{4.4}$$

At line element level we have

$$d\theta = \sum d\theta^i e_i \tag{4.5}$$

and

$$d\theta^* = \sum d\theta^*_i e^{*i} \tag{4.6}$$

$$ds^2 =< d\theta, d\theta >= g_{ij} d\theta^i d\theta^j =< e_i, e_j > d\theta^i d\theta^j \quad \text{(by (49))} \tag{4.7}$$

Hence $(M, G = (g_{ij}))$ is a Riemannian manifold.

Remark:

1. When $\psi(\xi) = \frac{1}{2}\sum \xi_i^2$, $g_{ij} = \delta_{ij}$ and hence Euclidean space is a special Riemannian manifold and a manifold induced from a convex function ψ is non-Euclidean in general.
2. The Riemannian metric can also be represented in the dual affine coordinate θ^*. Then

$$ds^2 =< d\theta^*, d\theta^* >= g^{*ij} d\theta^*_i d\theta^*_j \tag{4.8}$$

where $g^{*ij} =< e^{*i}, e^{*j} >$. Then $d\theta^* = Gd\theta, d\theta = G^{-1}d\theta^*$ where $G = G^{*-1}$. So the two Riemannian metric tensors are mutually inverse and $e^{*i} = g^{ij}e_j$ and $e_i = g_{ij}e^{*j}$ and so $< e_i, e^{*j} >= \delta_i^j$ by $G = G^{*-1}$.
Hence $\{e_j\}$ and $\{e^{*j}\}$ are mutually dual or complimentarily orthogonal.
Then $A^i =< A, e^{*i} >$, $A_i =< A, e_i >$ and $A_i = g_{ij}A^j$ or $A^i = g^{*ij}A_j$.
Thus a convex function ψ on M defines a dually flat Riemannian structure.
Note that for a vector A, $|A|^2 = A^iA_i$ changes under parallel transport of θ to θ'. If $< A, B >= g_{ij}A^iB^j = g^{*ij}A_iB_j = A_iB^i$
If $< A_iB >= 0$ we say A and B are orthogonal at θ. But orthogonality is not preserved under parallel transport of A, B from θ and θ' (at θ').
Note if A is transported in parallel and B is transported in dual parallel then orthogonality of A, B at θ holds orthogonality at θ' also.
Then dually coupled parallel transports preserve orthogonality and accordingly generalized Pythagoras theorem holds in dually flat Riemannian manifolds and projection theorem also holds [4] (cf. Appendix §10).

3. Summary of our local analysis:
We have seen on a statistical model M admitting a convex function $\psi(x)$ gives a class of special divergence called Bregman divergences. Using convex analysis we get Legendre duals of θ as θ^* and of ψ as ψ^* and these ψ, ψ^*, θ and θ^* give rise to primal flat affine structure by ψ and dual flat affine structure by ψ^* and ψ gives rise to a Riemannian structure $G = (g_{ij}(\theta))$ and ψ^* with G^* and $G = G^{*-1}$. In fact G gives rise to an invariant 2-tensor G on M. We have

Theorem 4.1 Let (M, ψ) be a statistical model with a convex function ψ. Then ψ gives rise to a Bregman divergence which inturn makes (M, G) a dually (affine) flat Riemannian manifold and there are dual affine coordinate systems on such (M, G).

Remarks 4.2 (a) If we consider the class $Div(M)$ of all divergences on M then $Div(M) \supsetneqq \text{Breg}(M)$. We have studied in detail the set $Div(M)$ in our previous paper [20] and a general divergence D on M gives rise to a Riemannian structure on (M, g_D) with a pair of dual affine connections which are in general non-flat and there are no affine coordinate systems on such M. (cf. §7 of [20]).
(b) Recall the Nash embedding theorem that each Riemannian manifold (M, g) can be isometrically embedded in a Euclidean space. But a dually flat manifold arising from a Bregman divergence can be regarded as a statistical geometrical generalization of a Euclidean space such M reduces to Euclidean structure when the two dual flat affine structures become self-dual.
Hence we can expect theorems like Pythagorean theorem and orthogonal projection theorem of Euclidean space hold in dually flat manifolds in a general form (see Appendix in §10). By Nash theorem every Riemannian manifold can be regarded as a curved submanifold of a Euclidean space.
Hence a general non-flat manifold (M, g_D) arising out of a general divergence can be regarded as a curved submanifold of dually flat manifolds.
We study the geometry (global and local) of dually flat Riemannian manifolds in the next section. In a sense the set of all dually flat manifolds is a universal model.

Theorem 4.3 (a) A dually flat manifold arises from a Bregman divergence of a convex function ψ.
(b) Conversely given a dually flat manifold M there exists a convex function ψ and the corresponding Bregman divergence gives the original dually flat geometric structures.
These Bregman divergences are related to exponential families (later).

5 Geometry of exponential family of probability distributions

An important class of statistical models are the exponential families of probability distributions which includes the well known families of probability distributions of statistics such as discrete probability distributions S_n, normal distributions, multinomial distributions, gamma distributions etc. These are typical models universal in nature and arise from a convex function $\psi(x)$ which is called the cumulant generating function so that they come from Bregman divergence and conversely. Thus exponential families are dually flat Riemannian manifolds.

5.1. Exponential family of probability distributions.

The standard family $\mathcal{E}$ of exponential probability distributions is given by the probability function

$$p(\underline{x}, \underline{\theta}) = \exp\{\theta^i h_i(x) + k(x) - \psi(\theta)\} \tag{5.1}$$

where x is a random variable, $\underline{\theta} = (\theta^1, \cdots, \theta^n)$ is an n-dimensional vector parameter specifying the distribution, $h_i(x)$ are n functions of x which are linearly independent, $k(x)$ is a function of x and ψ is the normalizing factor function. We can put (53) in the standard form

$$p(x, \underline{\theta})dx = \exp(\underline{\theta} \cdot \underline{x} - \psi(\theta)\}d\mu(x) \tag{5.2}$$

where $\underline{x} - (h_1(x), \cdots, h_n(x))$ with new sample space X and measure $d\mu(x) = \exp(k(x))dx$. Then $\mathcal{E} = \{p(\underline{x}, \underline{\theta})\}$ is a n-dimensional statistical model with $\underline{\theta}$ as its coordinate system. Then $\int p(\underline{x}, \underline{\theta})dx = 1$ gives

$$\psi(\theta) = \log \int \exp(\underline{\theta} \cdot \underline{x})d\mu(x) \tag{5.3}$$

and we saw earlier $\psi(\theta)$ is a convex function of θ and hence $\mathcal{E}$ gets a dually flat Riemannian structure by the standard procedure explained before thru the data

$$\underline{\theta}, \theta^* = \eta, \psi(\underline{\theta}), \psi^*(\theta^*) = \varphi(\eta) \tag{5.4}$$

$\underline{\theta}$ is called the natural parameter or canonical parameter of $\mathcal{E}$ and $\underline{\eta} = E[\underline{x}] = \int \underline{x}p(x, \theta)d\mu(x)$ is called the expectation parameter of $\mathcal{E}$. $\underline{\theta}$ and $\underline{\eta}$ are the two dually flat affine coordinates on $\mathcal{E}$ and the Riemannian metric

$$g_{ij}(\theta) = \frac{\partial^2 \psi(\underline{\theta})}{\partial \theta^i \partial \theta^j} \tag{5.5}$$

The dual convex function $\varphi(\eta)$ is the negative entropy given by

$$\varphi(\eta) = \int p(x,\theta)\log p(x,\theta)dx \tag{5.6}$$

where θ is a function of η thru $\eta = \nabla\psi(\theta)$ and

$$\begin{aligned} D_\psi[\theta',\theta] =& \psi(\theta') - \psi(\theta) - \eta.(\theta'-\theta) \\ =& \int p(x,\theta)\log\frac{p(x,\theta)}{p(x,\theta')}d\mu(x) = D_{KL}[\theta,\theta'] \end{aligned} \tag{5.7}$$

the Riemannian metric of the exponential family is the Fisher-Rao information metric of $\mathcal{E}$ defined by

$$\begin{aligned} g_{ij} =& E[\partial_i \log p(x,\theta)\partial_j \log p(x,\theta)] \\ =& E[(x_i-\eta_i)(x_j-\eta_j)] = \nabla\nabla\psi(\theta) = \frac{\partial^2\psi(\theta)}{\partial\theta^i\partial\theta^j} \\ & (cf.[1]) \end{aligned} \tag{5.8}$$

5.2 Examples

(a) Gaussian distributions with means μ and variance σ^2 has the probability density function

$$p(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{5.9}$$

Take a new *r.v.* $\underline{x} = (x_1, x_2) = (x, x^2)$ and new parameters $\theta^1 = \frac{\mu}{\sigma^2}$, $\theta^2 = -\frac{1}{2\sigma^2}$ and write (61) as

$$p(\underline{x},\underline{\theta}) = \exp[(\underline{\theta}\cdot\underline{x}) - \psi(\theta)] \tag{5.10}$$

with the convex function $\psi(\theta)$ given by

$$\psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma) = -\frac{(\theta^1)^2}{4\theta^2} - \frac{1}{2}\log(-\theta^2) + \frac{1}{2}\log\pi \tag{5.11}$$

The new measure is $d\mu(x) = \delta(x_2 - x_1^2)dx$ and the dual affine coordinate η are given by

$$\eta = E(x) = \int xp(x,\theta)d\mu(x) \tag{5.12}$$

and $\eta = (\eta_1, \eta_2)$ with $\eta_1 = \mu$ and $\eta_2 = \mu^2 + \sigma^2$.

(b) Discrete distributions of a *r.v.x* over $X = \{0, 1, 2, \cdots, n\}$ form a probability simplex S_n and $p \in S_n$ is represented by

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x) \tag{5.13}$$

Introduce new *r.v.* $x_i = h_i(x) = \delta_i(x)$ $i = 1, 2, \cdots, n$ and new parameters $\theta^i = \log \frac{p_i}{p_o}$. Then (65) can be written

$$p(x, \theta) = \exp\left\{\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)\right\} \tag{5.14}$$

where

$$\psi(\theta) = -\log p_o = \log\left\{1 + \sum_{i=1}^{n} \exp(\theta^i)\right\} \tag{5.15}$$

(use $\log p(x, \theta)$'s expression)
The dual affine coordinates η are

$$\eta_i = E[h_i(x)] = p_i \qquad i = 1, 2, \cdots, n \tag{5.16}$$

and

$$\varphi(\eta) = \sum \eta_i \log \eta_i + (1 - \sum \eta_1) \log(1 - \sum \eta_i) \tag{5.17}$$

(c) Mixture family of probability distributions.
We compose a new probability distribution out of given $n+1$ probability distribution $q_o(x), q_1(x), q_2(x), \cdots, q_n(x)$ which are linearly independent, by

$$p(x, \eta) = \sum_{i=0}^{n} \eta_i q_i(x) \tag{5.18}$$

where $\sum_{i=0}^{n} \eta_i = 1, \quad \eta_i > 0 \ \forall \ i$
$M = \{p(x, \eta)\}$ is a statistical model called the mixture family where $\eta = [n_1, n_2, \cdots, n_n]$ is a coordinate system with $\eta_o = 1 - \sum_{i=1}^{n} \eta_i$.
Note a discrete distribution $p(x) = \sum p_i \delta_i(x) \in S_n$ can be written as a mixture family with $q_i(x) = \delta_i(x)$ $i = 0, 1, 2, \cdots, n$ and $\eta_i = p_i$ for $i = 0, 1, 2, \cdots, n$ and η is a dual affine coordinate system of the exponential family S_n.

Remark: if a general mixture family is given by (70) which is not an exponential family. Then also the negative entropy given by $\varphi(\eta) = \int p(x, \eta) \log p(x, \eta) dx$ is a convex function of η and hence defines a dually flat structure to $M = \{p(x, \eta)\}$ having η as the dual affine

coordinate system. then $\theta = \nabla\varphi(\eta)$ defines the primal affine structure dually coupled with η but θ is not the natural parameter of an experimental family. and

$$D_\varphi[\eta, \eta'] = \text{the } KL \text{ div } = \int p(x,\eta) \log p(x,\eta) \Big/ p(x,\eta')dx$$

Remark: On an exponential family we have θ-primal affine flat structure called e-flat and e-geodesic. Similarly the dual η-coordinate system defies an affine flat structure giving η-flatness or m-flatness and m-geodesics. Thus $\mathcal{E}$ has a dually flat Riemannian structures see [4] and [1].

6 Infinite Dimensional Manifold of Probability Distributions (a formal approach).

Recall that S_n the probability simplex of discrete probability distributions is an exponential family and a mixture family at the same time. S_n can be regarded as a super manifold in which every statistical model of a discrete random variable is embedded as a submanifold. When x is a continuous random variable formally we consider the set $\mathfrak{F}$ of all probability density functions $p(x)$ as a manifold and assume it is an experimental family and also a mixture family simultaneously. It is a supermanifold including all statistical models of a continuous $r.v$ as submanifolds.

Let $p(x)$ be a probability density function of a real $r.v.x \in \mathbb{R}$ which is mutually absolutely continuous w.r.t. the Lebesgue measure dx or the Gaussian measure

$$d\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] dx$$

Consid4er the function space

$$F = \{p(x) | p(x) > 0, \quad \int p(x)dx = 1\} \tag{6.1}$$

For $p_1(x), p_2(x) \in F$, the expermental family connecting them is

$$P_{\exp}(x,t) = \exp\{(1-t)\log p_1(x) + t \log p_2(x) - \psi(t)\} \tag{6.2}$$

provided it exists in F and the mixture family connecting them is

$$P_{\text{mix}}(x,t) = (1-t)p_1(x) + tp_2(x) \tag{6.3}$$

assuming it belongs to F,. Thus F can be regarded as an exponential family and a mixture family of probability distributions.

6.1 General idea: Discretize the real line $\mathbb{R}$ into $n+1$ intervals $I_o, I_1, \cdots, I_n$. Then the discretized version of $p(x)$ is given in the discrete probability distribution $\underline{p} = (p_o, p_1, \cdots, p_n)$ where

$$p_i = \int_{I_i} p(x)dx, \quad i = 0, 1, 2, \cdots, n \tag{6.4}$$

we get a map: $F \to S_n\ p(x) \to \underline{p}$ and if $p_i \to 0$ on each I_i as $n \to \infty$ we get $F = \lim_{n\to\infty} S_n$.

Remark 6.2: In a more general set up of generalized parametric measure models and sufficient statistics this idea was done with mathematical rigor ([6],[10]).

Similar to discrete case $S_n p(x) = \sum p_i \delta_i(x)$. introduce a family of random variables $\delta(s-x)$, $s \in \mathbb{R}$ real point (corresponds to index i in $\delta_i(x)$ of S_n). Then

$$p(x) = \int p(s)\delta(x-s)ds \tag{6.5}$$

which shows F is a mixture family given by $\{\delta(s-x)\}$. Similarly

$$p(x) = \exp\{\int \theta(s)\delta(s-x)dx - \psi\} \tag{6.6}$$

where

$$\theta(s) = \log p(x) + \psi \tag{6.7}$$

and ψ is a functional of $\theta(x)$ formally given by

$$\psi(\theta(s)) = \log\left[\int \exp[\theta(s)]ds\right] \tag{6.8}$$

Hence F is an exponential family where $\theta(s) = \log p(s) + \psi$ is the θ-affine coordinates and $n(s) = p(s)$ is the dual affine coordinate η. The dual convex function is

$$\varphi[\eta(s)]] = \int \eta(s)\log\eta(s)ds \tag{6.9}$$

Note the dual coordinates are

$$\eta(s) = E_\varphi[\delta(s-x)] = p(s) \tag{6.10}$$

and we have

$$\eta(s) = \nabla\psi[\theta(s)] \tag{6.11}$$

where ∇ is the Frechet-derivative w.r.t $\theta(s)$. the $e$geodesic connecting $p(x)$ and $q(x)$ is given by (72) and m-geodesic by (73).
The KL-divergence is

$$D_{KL}[p(x), q(x)] = \int p(x)\log\frac{p(x)}{q(x)}dx \tag{6.12}$$

which is the Bregmann divergence derived from $\psi(\theta)$ and hence F becomes a dually flat Riemannian manifold with the Riemannian metric for two nearby points $p(x)$ and $p(x) + \delta p(x)$ is given by

$$ds^2 = \int [\frac{\delta p(x)}{p(x)}]^2 dx \tag{6.13}$$

Thus we proved.

Theorem 6.3 the infinite-dimensional function space Fhas a dually flat Riemannian structure and the Riemannian metric in θ-coordinates is give η by

$$g(s,t) = \nabla\nabla\psi = p(x)\delta(s-t) \tag{6.14}$$

Remark: For a manifold structure on F where F is a Banach space refer to the works of G. Pistone and his collaborators [11],[18],[12],[13],[14],[15],[16],[17].

6.4 Bregman divergence and Exponential family

We saw that an exponential family induces a Bregman divergence $D_\psi[\theta, \theta']$. Now we want to answer the converse problem: Given a Bregman divergence $D_\psi[\theta, \theta']$ does there exist a corresponding exponential family $p(x, \theta)$?
We answer this now:
Start with a $r.v.x$. It specifies a point $\eta' = x$ in the η-coordinates of a dually flat manifold given by ψ. Let θ' be its θ-coordinate. Then the ψ-divergence from θ to θ' (θ' is the θ-coordinate of the point $\eta' = x$) is

$$D_\psi[\theta, \theta'(x)] = \psi(\theta) + \varphi(x) - \theta \cdot x \tag{6.15}$$

Then define the probability density function $p(x, \theta)$ by

$$p(x,\theta) = \exp[-D_\psi[\theta,\theta'] + \varphi(x)] = \exp\{\underline{\theta} \cdot \underline{x} - \psi(\theta)\} \tag{6.16}$$

where θ' is determined from x as the θ-coordinates of $\eta' = x$. Thus we proved:

Theorem 6.5 Given a Bregman divergence there exists an exponential family $D_\psi[\theta, \theta']$ defined by (86).
In other words, given a convex function $\psi(\theta)$ find a measure $d\mu(x)$ such that (85) holds or equivalently

$$\exp\{\psi(\theta)\} = \int \exp(\underline{\theta} \cdot \underline{x}) d\mu(x) \tag{6.17}$$

holds which is the inverse of the Laplace transform.

Remark 6.6. There is a deeper result due to A. Banerjee [7] which gives a 1-1 correspondence between the set of regular Bregman divergences $\mathbb{B}$ and the set of regular exponential families $\mathcal{E}$. Thus $\mathbb{B}$ and $\mathcal{E}$ are super models for dually flat Riemanian manifolds.
Note that all discrete probability distributions on a finite set is a curved submanifold of probabilities simplex S_n and hence S_n is a universal model in the discrete case.
Similarly each dually flat manifold is a curved submanifold of $\mathbb{B}$ or $\mathcal{E}$.

7 Study of dually flat manifolds

We have studied the geometry of the class of Bregman divergences out of convex functions ψ on statistical models. However we have a bigger class of general divergences Div. on M which induces a dualistic structure on M which is non-flat.

Definition 7.1. Two affine connections Γ, Γ^* on M are dual w.r.t. the Riemannian metric g if locally

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma^*_{ikj} \tag{7.1}$$

or globally

$$Z < X, Y >=< \nabla_Z X, Y > + < X, \nabla^*_Z Y > \tag{7.2}$$

for any three vector fields X, Y, Z on M and ∇, ∇^* are covariant derivatives of Γ, Γ^* respectively..
Then define a symmetric 3-tensor T on M by

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk} \tag{7.3}$$

and define

$$\Gamma^o_{ijk} = \frac{1}{2}(\Gamma_{ijk} + \Gamma^*_{ijk}) \tag{7.4}$$

which is the Levi-Civita connection of g and so

$$\Gamma_{ijk} = \Gamma^o_{ijk} - \frac{1}{2}T_{ijk} \tag{7.5}$$

$$\Gamma^*_{ijk} = \Gamma^o_{ijk} + \frac{1}{2}T_{ijk} \tag{7.6}$$

Infact

$$\nabla_i g_{jk} = T_{ijk} \tag{7.7}$$

Remark 7.2: Amari called T as a cubic tensor.
A Statistical model (M, G, T) equipped with symmetric $p.d$ 2-tensor G and a symmetric 3-tensor T is called a statistical manifold and $\{G, T\}$ is called a Amari-Chentsov structure.

7.3 A general divergence $D[\xi, \xi']$ induces on M the Amari-Chentsov structure $\{(g^D_{ij}), \{T^D_{ijk}\}\}$ on M by higher order differentiation of $D[\xi, \xi']$ as follows: Define

$$D_i = \frac{\partial}{\partial \xi^i} D[\xi, \xi']\Big|_{\xi'=\xi} \tag{7.8}$$

$$D_{ij} = \frac{\partial}{\partial \xi^i}\,\frac{\partial}{\partial \xi^j} D[\xi, \xi']\Big|_{g'=\xi} \tag{7.9}$$

$$D_{,j} = \frac{\partial}{\partial \xi'^j} D[\xi, \xi']\Big|_{\xi=\xi'} \tag{7.10}$$

and

$$D_{ij,k} = \frac{\partial^2}{\partial\xi^i\partial\xi^j}\frac{\partial}{\partial\xi' k}D[\xi : \xi']_{\xi'=\xi} \tag{7.11}$$

etc. Using these define

$$g^D_{ijh} = -\, D_{i,j} \tag{7.12}$$

$$\Gamma^D_{ijk} = -\, D_{ij;k} \tag{7.13}$$

$$\Gamma^{D^*}_{ijk} = -\, D_{k,ij} \tag{7.14}$$

Then Γ^D and Γ^{D*} are affine connections and define

$$T^D_{ijk} = \Gamma^{D*}_{ijk} - \Gamma^D_{ijk} \tag{7.15}$$

defines third order symmetric tensor. Then we get a dual structure on M induced by divergence D.

7.4 Recall the f-divergence on the manifold of probability distributions defined by

$$D_f(p(x), q(x)) = \int p(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

(f is a suitable function). Then we can calculate the Amari-Chentsov tensors as

$$g^f_{ij} = g_{ij} \quad \text{(Firher-Rao metric)} \tag{7.16}$$

and

$$T^f_{ijk} = \alpha T_{Ijk} \tag{7.17}$$

with

$$T_{ijk} = E[\partial_i l(x,\xi)\partial_j l(x,\xi)\partial_k l(x,\xi)] \tag{7.18}$$

and

$$\alpha = 2f'''(1) + 3\,.$$

(simply differentiate $D_f[\xi,\xi'] = \int p(x,\xi) f\left(\frac{p(x,\xi')}{p(x,\xi)}\right) dx$ w.r.t. ξ, ξ' and put $\xi = \xi'$ as required.

7.5 Amari's α-geometry.

Consider the Amari-Chentsov tensor $\{G, \alpha T\}$ for $\forall\, \alpha \in \mathbb{R}$ in

$$\Gamma^\alpha_{ijk} = \Gamma^0_{ijk} - \frac{\alpha}{2}T_{ijk}\ , \quad \Gamma^{-\alpha}_{ijk} = \Gamma^0_{ijk} + \frac{\alpha T_{ijk}}{2}$$

we get dual structures $\{\Gamma^{\alpha}, \Gamma^{-\alpha}\}$ on M w.r.t. G and this α-geometry of M is invariant.

Remark 7.6 Thus the two invariant tensors G and T generate the 1-parameter family of α-geometries on M by a formal process. However there is a more rigorous way of understanding these α-geometry of Amari as an embedding geometry see [20]. We can justify geometrically the dually flatness of M as in

Proposition 7.7: When the Riemannian Christoffel curvature R vanishes w.r.t. one affine connection on a dually flat manifold (M, g, Γ, Γ^*), the RC-curvature R^* w.r.t the dual connection vanishes and vice versa.

Proof: When $R = 0$ parallel transport operator π leaves every tangent vector A invariant i.e. $A = \pi(A)$. We always have

$$< A, B >=< \pi A, \pi^* B >=< A, \pi^* B > \tag{7.19}$$

which implies $B = \pi^* B$ and hence $R^* = 0$.

Thus a manifold is always dually flat when it is flat w,.r.t. one connection. When M is dually flat there exists an affine coordinate θ for which $\Gamma_{ijk}(\theta) = 0$.
Similarly there exists a dual affine coordinate system η for which $\Gamma^{*ijk}(\eta) = 0$ and hence in these dual coordinate systems each coordinate curve is a geodesic with their respective directions given by $\{e_i\}$ and $\{e^{*i}\}$ respectively. Then the Jacobians of these coordinate transformation among $\underline{\theta} = (\theta^i)$ and $\eta = (\eta_j)$ satisfy

$$g_{ij} = \frac{\partial \eta_i}{\partial \theta^j} \qquad \text{and} \qquad g^{ij} = \frac{\partial \theta^i}{\partial \eta_j}$$

and hence the two bases $\{e_i\}$ and (e^{*i}) satisfy

$$e_i = g_{ij} e^{*j} \qquad \text{and} \qquad e^{*i} = g^{ij} e_i$$

Thus we proved.

Theorem 7.8 In a dually flat manifold there exist affine coordinate systems $\underline{\theta}$ and $\underline{\eta}$ the dual affine coordinate systems such that their tangent vectors are reciprocally orthogonal $< e_i, e^{*j} >=< \partial_i, \partial^{*j} >= \delta_i^j$ and orthogonality is preserved under mixed parallel transportations i.e. $< e_i, e^{*j} >= \delta_i^j$ and $< \pi e_i, \pi^* e^{*j} >=< e_i, e^j >= \delta_i^j$.

7.9 Canonical divergence in a dually flat manifold.

Recall a general divergence induces on M a dual structure which is nonflat in general and a dually flat structure is induced from a Bregman divergence. On a dual manifold there are no dually flat affine coordinate system $\underline{\theta} = (\theta^i)$ and $\underline{\eta} = (\eta_j)$ as described above. Several divergences give the same dual structure.

Proposition 7.10: (dual convex scheme $\{\theta, \eta, \psi, (\theta)$ and $\varphi(\eta)\}$ on a dually flat manifold). Let M be dually flat manifold (Riemannian). Then there are a pair of dual affine coordinate systems $\underline{\theta}$ and $\underline{\eta}$ and a pair of Legendre-dual convex functions $\psi(\theta)$ and $\varphi(\eta)$ satisfying

$$\psi(\theta) + \varphi(\eta) - \theta^i \eta_i = 0 \tag{7.20}$$

such that the Riemannian metric is given by

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta); \quad g^{ij}(\eta) = \partial^i \partial^j \varphi(\eta) \tag{7.21}$$

and the Amari-Cubic tensor by

$$T_{ijk}(\theta) = \partial_i \partial_j \partial_k \psi(\theta)$$

and

$$T^{ijk}(\eta) = \partial^i \partial^j \partial^k \varphi(\eta) \tag{7.22}$$

This was proved earlier. Using this we get.

Theorem 7.11 When M is a dually flat manifold there exists a Legendre pair of convex functions $\psi(\theta)$ and $\varphi(\eta)$ and a canonical divergence given by the Bregman divergence

$$D[\theta, \theta'] = \psi(\theta) + \varphi(\eta') - \theta \cdot \eta' \tag{7.23}$$

They are uniquely determined upto affine transformation. Conversely the canonical divergence gives the original dually flat Riemannian structure. (pf. trivial from 7.10).

Remark 7.12 Note that the KL-divergence is the canonical divergence of an exponential family of probability distributions which is invariant and dually flat.

Remark 7.13 On a general dual manifold one can define a canonical divergence using an averaging process along primal geodesic $\xi(t)$ and dual geodesic $\xi^*(t)$ connecting given two points ξ_p and ξ_q using experimental map: $T(M) \to M$. (cf. Ay-Amari [3]).

Theorem 7.14 (Ay and Amari) The geometric structure derived from the canonical divergence coincides with the original geometry. When M is dually flat it gives the canonical divergence of Bregman type. When M is Riemannian ($T = 0$) it gives half of squared Riemannian distance.

We close this section with a remark on Invariant geometry.

Remark 7.15: Originally it is Fisher and Rao that showed their Riemannian metric g is invariant and unique. Later in 1972 Chentsov showed the two tensors G and T are invariant on the probability simplex S_N for discrete case and also their uniqueness. Then in 1982 Amari and Nagaoka formulated invariant α-geometries on any statistical model M in general from these two tensors G and T (cf. [4]).

8 Amari-Chentsov structures and their specialization on M

Let (M, G, T) be a Amari-Chentsov (AC) manifold which is realizable by a general divergence D such that $g^D = G$ and $T^D = T$. So we have a many-to-one map from $Div(M) \to \{\text{A-C structures on } M\}$. By Ay-Amari Canonical divergence construction we can make this map 1-1 locally by this choice. Similarly this map is 1-1 at dually flat structures level and Bregman divergences level.

8.1: Bregman divergence, Hessian metric

We have seen earlier that $Div(M) \supset \mathbb{B} = \mathcal{E}$ and $\{\text{Dual structure on} M\} \supsetneqq \{\text{Dually flat Riemannian Structure on } M\}$. Let M be a dually flat Riemannian manifold. Then there exists convex function ψ, its convex dual φ and coordinate system $\underline{\theta} = (\theta^i)$ and its dual coordinate system (η) such that the AC tensors G and T are realized in terms of this data. In particular the Riemannian metric

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta) \tag{8.1}$$

Definition 8.2 A Riemannian metric g is called *Hessian metric* if there exist local coordinates such that g can be expressed as the Hessian of some convex potential function ψ as in (111).

Hessian metrics play important role in studies on optimization and convex programming, statistical manifolds and in string theory via special Kahler manifolds. In Hessian geometry there is a natural duality principle given by the Legendre-Fenchel transform. Recall that given any affine connection ∇ on a Riemannian manifold (M, g) we define its "g-dual connection ∇^*" by

$$g(\nabla_X, Y, Z) = g(Y, \nabla^*_X Z) \tag{8.2}$$

and Levi-civita connection of (M, g) is self-dual.

Definition 8.3: A g-dually flat structure is a pair of g-dual connections which are both flat.

So a metric locally admits a g-dually flat structure if and only if g is Hessian metric. We have shown this before analytically. In geometrical sense one simply needs to know that the geodesics of g-dually flat connection define local coordinate system w.r.t. which g is a Hessian and converse is also true [4], [19].

Amari [4] raised the question whether a Riemannian metric g on M can be realized as a Hessian?

If not, what are the obstructions?

Characterize this subclass.

Consider an AC manifold (M, G, T) put sufficient conditions on T so that G gives a Hessian metric. This is a kind of "flattening process" for M.

Remark: In a different context Weil [24] studied similar problem

$$g_{ij} = \frac{\partial^2 \varphi}{\partial z_1 \partial z_j} \quad \text{or} \frac{\partial^2 \varphi}{\partial z_i \partial \bar{z}_j} \tag{8.3}$$

in the context of smooth function φ (not convex) for understanding special Kahler manifolds.

Amari's problem being local in nature is equivalent to determining whether g is a Hessian metric. This is a deep problem. We outline some results on this.

Theorem 8.4 In dimensions $n \geq 3$, a generic Riemannian metric $(M^n g)$ does not admit a compatible dually flat structure, even locally.

Proof: If (M^n, g) admits a dually flat structure then, in some neighborhood of any point p, there exists local coordinates $x : M^n \to \mathbb{R}^n$ and a potential function φ such that in these coordinates the metric is given by

$$g_{ij} = \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \tag{8.4}$$

Thus the k-jet of g at p is determined by the $(k+2)-$jet of x and φ at p. If we fix some reference coordinates around p then the function x is defined by n real valued functions in n-variables and the coordinate φ is a single real-valued function of n variables: Hence the dimension of the space of $(k+2)$-jets of (x, φ) at p is equal to

$$\dim J_{k+2}(x, \varphi) := \sum_{i=0}^{k+2} (n+1) \dim(S^i T_p) = \sum_{i=0}^{k+2} (n+1) \binom{n+i-1}{i} \tag{8.5}$$

On these same reference coordinates the metric g is defined by $n(n+1)/2$ real valued functions, so the space of k-jets of metric at p has dimension given by

$$\dim J_k(g) = \sum_{i=0}^{k} \frac{n(n+1)}{2} \dim(S^i T_p) = \sum_{i=0}^{k} \frac{n(n+1)}{2} \binom{n+i-1}{i} \tag{8.6}$$

Thus

$$\dim J_k(g) - \dim J_{k+2}(x, \varphi) = (n+1)(a_{k,n} - b_{k,n}) \tag{8.7}$$

where

$$a_{kn} = \Big(\frac{n}{2} - 1\Big) \sum_{i=1}^{k} \binom{n+i-1}{i}$$

$$\text{and} \quad b_{k,n} = \binom{n+k}{k+1} + \binom{n+k+1}{k+2}$$

For fixed $n > 2$, $a_{k,n}$ grows as order k^n where as $b_{k,n}$ grows as order k^{n-1}. So for sufficiently large k

$$\dim J_k(g) > \dim J_{k+2}(x, \varphi)$$

Hence it follows that if $n > 2$, for sufficiently large k.

The generic k-Jet of a metric tensor does not admit any compatable dually flat structure no matter how one extends this k-Jet to a smooth metric

Remark: The above counting argument can be summarized by saying that the number of metric depends upon $\frac{1}{2}n(n+1)$ functions of n-variables whereas the data for a Hessian structure depends upon only $(n+1)$ functions of n-variables and our argument makes it more precise and answers negatively Amari's question for $n \geq 3$.
We have the following result of Shima [19].

Proposition 8.5: Let (M, g) be a Riemannian manifold. Let ∇ be the Levi-Civita connection of g. Suppose there exists tensor A on M such that $\tilde{\nabla} = \nabla + A$ is a g-dually flat connection. Then

(i) The tensor A_{ijk} lies in S^3T^* called the S^3-tensor of $\bar{\nabla}$.

(ii) The S^3-tensor determines the Riemannian curvature tensor as follows:

$$R_{ijkl} = -g^{ab}A_{ika}A_{jlb} + g^{ab}A_{ila}A_{jkb} \tag{8.8}$$

Definition 8.6 Define a quadratic equivariant map

$$\begin{aligned} \rho : S^3T^* \to & \Lambda^2T^* \otimes \Lambda^2T^* \\ \rho(A_{ijk}) = & - g^{ab}A_{ika}A_{jlb} + g^{ab}A_{ila}A_{jkb} \end{aligned} \tag{8.9}$$

Then we get a necessary condition for a Riemannian metric g to be a Hessian metric.

Corollary 8.7 In dimension of $M > 4$ the condition that R_g lies in the image of ρ gives a non-trivial necessary condition for a metric g to be Hessian metric

Proof: $\dim S^3T = \binom{n+2}{n-1} = \frac{1}{6}n(n+1)(n+2)$.
The dimension of the space of algebraic curvature tensor, $\mathbb{R}$ is $\dim \mathbb{R} = \frac{1}{12}n^2(n^2-1)$ (cf.[23]).

So $\dim \mathbb{R} - \dim S^3T = \frac{1}{12}n(n-4)(n+1)^2$ which is positive if $n > 4$.

Remark 8.8 For the case of $n = 4$, the space of possible curvature tensors for Hessian 4-manifold is 18-dimensional. In particular the curvature tensor must satisfy the identities

$$\alpha\left(R^b_{ija}R^a_{klb}\right) = 0 \tag{8.10}$$

$$\alpha(R_{iajb}R^b_{kcd}R^{dac}_l - 2R_{ajb}R^a_{kcd}R^{dbc}_l) = 0 \tag{8.11}$$

where α denotes anti-symmetrization of the i, j, k and l indices.

The identity (120) gives that the Pontryagin forms i.e. the closed differential forms given as polynomials in the curvature tensor that represent its class vanish on a Hessian manifold.

Hence this suggests that the Pontryagin classes are the topological obstructions for the existence of a Hessian metric on n-dimensional manifold for $n \geq 4$.

8.9 Examples: 1) all analytic 2-manifolds are Hessian
2) All products of Hessian manifolds are Hessian
3) All hyperbolic manifolds are Hessian (see [21],[22] for more examples)
4) In dimension 3 also all possible Riemannian curvature tensors occur as the curvature tensor of a Hessian metric (see [2]).

9 Some general Remarks

We have defined geometrically (a) dual manifold (b) dually flat manifold among the class of Riemannian manifolds.

Recall the definition of statistical manifold of Lauritzen [1987] [8] as a Riemannian manfold (M, g) with a 3-symmetric tensor T Amari's question: What is the relation among general dual manifold and statistical manifolds?
and Lauritzen's question: if any statistical manifold is a statistical model?
More precisely for a given statistical manifold (M, g, T) can we find a sample space Ω and probability distribution $p(x, w), x \in M$, $w \in \Omega$ such that $p(x, w)$ is a potential function for g and T?

Definition 9.1 A statistical model is a (possibly immersed) submanifold M in the space $\mathrm{Cap}(\Omega)$ of all probability measures on a sample space Ω. If $\dim \Omega > 1$ then $\mathrm{Cap}(\Omega)$ is an infinite dimensional space. We represent the measure $p(x, w)$ as $p(x, w)dw$ with $p(x, w)$ considered as an almost everywhere positive density function and dw is a specific Borel measure on Ω.

Then we put two conditions on $p(x, w)$ namely

$$p(x, w) > 0 \ \forall \ (x, w) \in M \times \Omega \tag{9.1a}$$

and

$$\int_\Omega p(x, w)dw = 1 \qquad \forall \ x \in M \tag{9.1b}$$

Let $p(x, w)$ be a probability potential on a statistical model M.

Then the Fisher-Rao metric is defined on M by

$$g_{ij} = \int_\Omega \partial_i \log p(x, w)\partial_j \log p(x, w)p(x, w)dw = E_x(\partial_i \log p \partial_j \log p) \tag{9.2}$$

and Chentsov tensor T on M is given by

$$T(X, Y, Z) = E_x(\partial_X \log p(x, w)\partial_Y \log p(x, w)\partial_Z \log p(x, w)) \tag{9.3}$$

$$\begin{aligned} &= \int_\Omega \frac{1}{p^2}(\partial_X p(x, w)\partial_Y p(x, w)\partial_Z p(x, w))dw \\ &=8 \int_\Omega \frac{1}{\sqrt{p}}\partial_X \sqrt{p}\partial_Y \sqrt{p}\partial_Z \sqrt{p} \end{aligned}$$

Definition 9.4: We say that a function $p(x, w)$ on $M \times \Omega$ is a probability potential for the metric g if $p(x, w)$ satisfies (9,1a) and (9.1b) as well as g is derived from $p(x, w)$ via (9.2) and if moreover T is derived from $p(x, w)$ by (9.3a,b,c) we say that $p(x, w)$ is probability potential for (M, g, T). In this case the statistical manifold (M, g, T) is a statistical model associated with the sample space Ω. We have seen that for a general dual manifold its structure arises from a general divergence D and it induces $(g^D, \Gamma^D, \Gamma^{D*})$ structure and T^D and hence is a statistical manifold in the sense of Lauritzen. For dually flat ones also this works and is a statistical manifold.
However there is a more concrete realization by Banerjee's theorem that every dually flat M can be realized (1-1 correspondence) as an exponential family of probability distributions (cf. [7]).

Theorem 9.5 (Lé) Any smooth (at least C^1) statistical manifold (M^n, g, T) can be immersed into a statistical manifold $[Cap_+^N, g^F, T^{A-C}]$ for some finite N, In other words every smooth statistical manifold can be realized as a statistical submanifold of probability simplex S_N for sufficiently large N and hence is a statistical model.
Corollary 9.6: A general dual manifold can be realized as a statistical model.

Remark 9.7: In a much general set up (less differential analytic set up) the delicate problems of topology on the manifold F of infinite dimensional probability distributions as Function space were treated in [6],[10],[9].

Remark 9.8: Finally we close this article with an open problem.

(a) Given a Riemannian manifold (M, G). Put geometric Tensorial data so that G becomes Hessian metric or dually flat. In other words find obstructions (curvature conditions) so that (M, G) is flattenable? If $\dim M = n = 2$ this is always true. For $n \geq 3$ there are obstructions. For some attempts on this see Amari-Armstrong [2] and Ay-Tuschmann [5] (see Appendix in §10 for global topological properties).

(b) Whether the vanishing of Pontryagin classes or equivalent curvature conditions (120) and (121) are sufficient for existence of dually flat structure or Hessian structure?

References

[1] S. Amari: Differential geometry of curved exponential families curvature and information loss Ann. of Stat. 10 (1982) p357-385.

[2] ——— and J. Armstrong: Curvature of Hessian manifolds, Differential Geometry and its Applications, 33 (2014) 1-12.

[3] ——— and N. Ay A novel approach to canonical divergences Entropy 17 (2015) 8191-8129.

[4] ——— and H. Nagaoka Methods of information geometry, AMS monograph vol. 191 (2000).

[5] N. Ay and W. Tuschmann: Dually flat manifold and global information geometry: open Sys. & Inf. Dyn. 9 (2002) p195-200.

[6] N. Ay, J. Jost, H.V. Le, L. Schwachhöfer: Information geometry and sufficient statistics, Probability Theory and related topics, 162 (2015) 327-364.

[7] A. Banerjee, S. Merugu, I.S. Dhillon, J. Glrosh clustering with Bregman Divergences, Jl. of Machine Learning Research 6 (2005) 1705-1749.

[8] S. Lauritzen: Statistical manifolds in differential geometry Institute of Math. Statistics, Monograph 10, Hayward, California 1987, 163-216.

[9] H.V. Le Statistical manifolds are statistical models Jl. of Geom. 84 (2005) 83-93.

[10] ——,——, and ——: Parametrized statistical models: arXiv. 25th October 2015, p29 (authors as in [6]).

[11] G. Pistone and C. Sempi, An infinite dimensional structure on the space of probability measures equivalent to a given one. Ann. of Stat. 5 (1995) 1543-1561.

[12] ——, A. Cena: Exponential Statistical models, Ann. Inst. Stat. Math. 59 (2007) p27-56.

[13] G. Pistone: Non parametric Information geometry, SCC of Inf. Lecture notes in computer science 8085 (2013) Springer, Heidelberg.

[14] ——, P. Cibiliso: connections on non-parametric statistical manifolds by orlicz space geometry, Inf. Diml. Analysis, Quantum probability and related topics 1 (1998) 325-347.

[15] ——, Malago, Malteucci: Natural gradient, fitness modeling and model selection, A unifying prospective, IEEE congress on Evolutionary computation 2013, p 486-493.

[16] ——, Malago combinational optimization with inf. geom., Newton method, Entropy 16, 2014, 4260-4289.

[17] ——: Examples of the application of non parametric information theory to statistical physics, Entropy 15, 2013, p 4042-4065.

[18] ——, M.P. Rogantin: Exponential statistical manifold, mean parameters, orthogonality and space transformation, Bernoulli 5, 1999, p 721-760.

[19] H. Shima: The geometry of Hessian structures, Vol. 1, World Scientific 2007.

[20] M. Sitaramayya, K.S.S. Moosath, K.V. Harsha: Generalized geometric structures on statistical manifolds Jl. of BGP, Lucknow Univ. 2015, GANITHA, Vol. 65 (2015) p 19-45..

[21] —— : Differential geometric study of a class of Hypersurfaces in a complex Torus, Ramanujan Inst. Seminar: Hyperbolic complex analysis proceedings (1977) p 205-222.

[22] —— : Hyperbolic complex manifolds, same proceedings as in [21] pages 53-79.

[23] —— : Kahler curvature Tensors, Trans. Amer. Math. Soc. 163 (1973) p 341-353.

[24] A. Weil: Varieté de Kahlerienne, Herman 1962.

APPENDIX

10 Global geometry and topology of dually flat manifolds

We study the topological obstructions and classification of dually flat manifolds and also discuss the geodesic property enjoyed by them.

Definition 10.1: A dually flat manifold is defined as a smooth Riemannian manifold (M, g) equipped with a pair of flat torsion- free affine connections ∇ and ∇^* which are dual to each other with respect to metric g i.e. for all vector fields X, Y, Z on M we have

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

We assume one of these two connections ∇ or ∇^* is complete on M i.e. the domain of all geodesics on M is the whole real line $\mathbb{R}$.

Example 10.2: (a) Let (M, g) be an exponential family of probability distributions equipped with the Fisher-Rao metric and $\nabla = \nabla^{(e)}$ is the exponential connection which is Fisher-Rao dual to the canonical mixture connection $\nabla^* = \nabla^{(m)}$ on M. Then the exponential connection $\nabla^{(e)}$ is complete and $\nabla^{(m)}$ is not complete.

(b) Let (M, g) be the set of positive density operators on a finite dimensional Hilbert space equipped with the Bogoliubov inner product and $\nabla = \nabla^{(e)}$ is the exponential connection and $\nabla^* = \nabla^{(m)}$ the mixture connection on M. Then $\nabla^{(e)}$ is complete though $\nabla^{(m)}$ is not.

10.3 Hick's ([1]) setup:

Let G be a connected Lie group and $\mathrm{Aut}(G)$ denote the group of continuous automorphisms of G. Then $GX\mathrm{Aut}(G)$ is a Lie group with group multiplication

$$(\lambda, k) \cdot (\mu, l) = (\lambda \cdot k(\mu), kol) \tag{1}$$

The $GX\ \mathrm{Aut}(G)$ acts effectively on G by

$$(\lambda, k) \cdot g = \lambda \cdot k(g) \tag{2}$$

On the Lie group G, there exists a canonical complete flat affine connection ∇ which is characterized by the fact that it is the one for which all left invariant vector fields on G are parallel. Then $GX\ \mathrm{Aut}(G)$ is isomorphic to the group of affine transformations of ∇, that is, isomorphic to the group of connection - preserving diffeomorphisms of $G \to G$ as follows:

If $\varphi : G \to G$ is a diffeomorphism then φ is connection - preserving if and only if $\varphi_*(g) = g$ where g=Lie algebra of G. thus the group of connection - preserving diffeomorphisms of G contains all left translations of G and all automorphisms of G. Conversely, if φ is a connection - preserving diffeomorphism of G,

Set $\lambda = \varphi(id)$, so that $\lambda^{-1} \circ \varphi$ leaves g invariant.

Thus $\lambda^{-1} \circ \varphi$ is an automorphism of G and $\varphi = \lambda \cdot (\lambda^{-1} \cdot \varphi)$.

Hence the group of connection-preserving diffeomorphisms of G is isomorphic to $GX\mathrm{Aut}(G)$.

Remark 10.4: The torsion tensor of the complete flat connection ∇ is invariant under parallel translation and if Γ is a subgroup of $GX\ \mathrm{Aut}(G)$ which acts freely and properly discontinuously on G then the quotient space G/Γ is a manifold whose fundamental group is isomorphic to Γ and carries an induced flat and complete connection ∇^Γ whose torsion tensor is invariant under parallel transport with respect to ∇^Γ. **Theorem 10.5:**

(Hicks-Ay-Tushmann) Let (M, ∇) be a manifold with a flat connection where torsion tensor is invariant under parallel translation. If ∇ is complete then there exists a connection-preserving diffeomorphism $\Phi : (M, \nabla) \to (G/\Gamma, \nabla^{\Gamma})$ where G is a connected and simply connected Lie group;

$\Gamma \simeq \pi_1(M)$ is a subgroup of the affine group $GX\,\mathrm{Aut}(G)$ acting freely and properly-discontinuously on G and ∇^{Γ} denotes the connection induced from the canonical connection ∇ on G for which all left invariant vector fields are parallel.

Proof: Let M satisfy the assumptions of the theorem. Consider the universal covering $\tilde{M}$ of M and Γ denote the group of deck-transformations of $\tilde{M}$ acting freely and properly discontinuously on $\tilde{M}$. Since the projection: $\tilde{M} \to \tilde{M}/\Gamma$ is a local diffeomorphism, $\tilde{M}$ carries a complete flat connection whose torsion tensor is invariant under parallel translation and which induces the given connection on M. Since $\tilde{M}$ is simply connected by [Hicks theorem 5 [2]],

$\tilde{M}$ is connection - preservingly diffeomorphic to a simply connected Lie group equipped with its canonical connection. Following [1] the Lie group G is determined as follows. The flatness of the connection on $\tilde{M}$ and the invariance of its torsion tensor under parallel transport imply that the vector fields on $\tilde{M}$ which are invariant under the parallel transport defined by this connection, form a finite-dimensioned Lie algebra which in turn uniquely determines a simply connected and connected Lie group G.

Since each deck transformation is connection-preserving, Γ is isomorphic to a subgroup of $GX\,\mathrm{Ant}(G)$ which acts freely and properly discontinuously on G. q.e.d.

From above Theorem 10.5 we get the following structure theorem for dually flat manifolds.

Theorem 10.6: Let (M, g, ∇, ∇^*) be a dually flat manifold of dimension m. Assume connection ∇ is complete. then there exists a connection-preserving diffeomorphism $\Phi : (M, \nabla) \to (\mathbb{R}^m/\Gamma, \nabla^{\Gamma})$ where $\Gamma \simeq \pi_1(M)$ is a subgroup of the group $\mathbb{R}^m XGl(m, \mathbb{R})$ of affine motions of $\mathbb{R}^m$ which acts freely and properly discontinuously on $\mathbb{R}^m$ and where ∇^{Γ} denotes the connection on $\mathbb{R}^m/\Gamma$ which is induced by the canonical flat affine connection on $\mathbb{R}^m$.

Proof: If M satisfies the assumptions of the theorem the torsion-freeness of ∇ implies that the structural constants of the Lie algebra which appears in the proof of Theorem 10.5 are all trivial[1,2] so that the Lie group in question here is simply a flat Euclidean space $\mathbb{R}^m$. q.e.d.

Corollary 10.7: Let (M, G, ∇, ∇^*) be a dually flat manifold of dimension m. Assume ∇ is complete. Then the universal covering $\tilde{M}$ of M is diffeomorphic to the Euclidean space $\mathbb{R}^m$ and $\pi_1(M)$ is isomorphic to a subspace of affine motions of $\mathbb{R}^m$ acting freely and properly discontinuously ion $\mathbb{R}^m$.

Proof: This directly follows from Theorem 10.6.

Corollary 10.8: Let (M, g, ∇, ∇^*) be a dually flat manifold of dimension m. Assume ∇ is complete. then the higher homotopy groups $\pi_k(M)$ of M vanish for all $2 \leq k \leq m$.

Proof: This follows from Corollary 10.7 and the fact that for a topological space whose universal covering space is contractible all higher homotopy groups vanish.

10.9: Geodesic property of dually flat manifolds.

For a complete affine connection ∇ on a manifold M it is in general false that any two

points in M can be joined by a ∇-geodesic., This is not true even if M is compact. But dually flat manifolds enjoy this property.

Theorem 10.10: (geodesic property): Let (M, g, ∇, ∇^*) be a dually flat manifold. Assume ∇ is complete. Then any two points of M can be joined by a ∇-geodesic.

Proof: Let dimension $M = m$ then by Theorem 10.5 there exists a connection-preserving diffeomorphism $\Phi : (M, \nabla) \to (\mathbb{R}^m/\Gamma, \nabla^\Gamma)$ where ∇^Γ denotes the connection on $\mathbb{R}^m/\Gamma$ which is induced by the canonical flat affine connection on $\mathbb{R}^m$. Let p, q be any two points of $M = \mathbb{R}^m/\Gamma$. Then choose corresponding points $P, Q \in \mathbb{R}^m$ which project down to p and q respectively.

Now P and Q can be joined by a geodesic in $\mathbb{R}^m$ whose projection is a $\nabla = \nabla^\Gamma$ geodesic in M. q.e.d.

Theorem 10.11: Let (M, g, ∇, ∇^*) be a dually flat manifold. Assume M is compact. Then the fundamental group $\pi_1(M)$ of M has infinite order.

Proof: Let M be a compact manifold with a flat and torsion free affine connection ∇. Let $\tilde{M}$ be the universal covering of M. Since $\tilde{M}$ is a simply connected and since the holonomy group of the induced connection on $\tilde{M}$ is trivial [2], $\tilde{M}$ admits a flat Riemannian metric $\tilde{g}$ whose geodesics project onto ∇-geodesics of M is complete. However we know that a complete and simply connected flat Riemannian manifold is isometric to some Euclidean space $\mathbb{R}^k$. This is a contradiction. q.e.d.

Corollary 10.12: Compact Riemannian manifolds with trivial or finite fundamental groups do never admit dually flat structures.

Proof: this follows directly from 10.11.

10.13: Some Remarks:

1. Theorem 10.11 shows that there exist general tropological obstruction to the existence of dually flat structures on a Riemannian manifold (M, g) and these obstructions are independent of the metric.

2. Theorem 10.5 gives a complete topological classification of dually flat manifolds under the assumption of one of the connections is complete.

3. Corollaries 10.6 and 10.7 provide further and strong topological obstructions to the existence of such dually flat structures. Infact any such manifold must be "aspherical" [3] and possesses a fundamental group of a very restricted type.

10.14: Generalized Pythagorean Theorem and Projection theorem for dually flat manifolds

Euclidean spaces are self-dual flat spaces with Pythagoras theorem holding in them for the Euclidean distance. A dually flat manifold is a generalization of the Euclidean space and hence we expect a Pythagorean type theorem in some generalized sense.

Let P, Q, R be three distinct points in a dually flat manifold M which form a triangle. We call it an orthogonal triangle when the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R at Q.

We have seen earlier such a manifold M arises out of a convex function ψ on M via) the Bregman divergence D_ψ [Theorem 4.3]. Now we state.

Theorem 10.15: (Generalize Pythagorean Theorem): If triangle PQR is orthogonal such that the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and

R at Q of a dually flat manifold M then the following generalized Pythagorean relation holds:

$$D_\psi(P,R) = D_\psi(P,Q) + D_\psi(Q,R) \qquad (A)$$

Proof: Recall the relation

$$D_\psi(P,Q) = \psi(\theta_p) + \psi^*(\theta^*_Q) - \theta_p \cdot \theta^*_Q \qquad (1)$$

and so

$$D_\psi(P,Q) + D_\psi(Q,R) - D_\psi(P,R) = (\theta^*_P - \theta^*_Q)\cdot(\theta_Q - \theta_R) \qquad (2)$$

The dual geodesic connecting P and Q is

$$\theta^*_{PQ}(t) = (1-t)\theta^*_P + t\theta^*_Q \qquad (3)$$

in parametric form and its tangent vector is given by

$$\dot\theta^*_{PQ}(t) = \theta^*_Q - \theta^*_P \qquad (4)$$

Dually, the geodesic connecting Q and R is

$$\theta_{QR}(t) = (1-t)\theta_Q + t\theta_R \qquad (5)$$

and its tangent vector is

$$\dot\theta_{QR}(t) = \theta_R - \theta_Q \qquad (6)$$

and since the two tangent vectors are orthogonal we have

$$(\theta^*_P - \theta^*_Q)\cdot(\theta_Q - \theta_R) = 0 \qquad (7)$$

and hence the right hand side of (2) is zero proving the relation (A). q.e.d.

Corollary 10.16: (dual relation) If triangle PQR is orthogonal such that the geodesic connecting P and Q is orthogonal to the dual geodesic connecting Q and R then

$$D_{\psi^*}(P,R) = D_{\psi^*}(P,Q) + D_{\psi^*}(Q,R) \qquad (B)$$

This follows as the Divergence is asymmetric.

Remark 10.17: Taking $\psi(\xi) = \frac{1}{2}\sum \xi_i^2$ we get $D[\xi,\xi_o] = \frac{1}{2}|\xi - \xi_o|^2$ and the affine coordinate system is exactly same as the dual affine coordinate system because the affine structure is self dual and the geodesic between two points coincides with its dual geodesic and hence the above two relations reduce to the classical Pythagoras theorem in a Euclidean space.

Projection Theorem: Consider a point P in a dually flat manifold M and let S be a smooth submanifold of M.

Definition 10.18: Define the divergence from a point P to submanifold S by

$$D_\psi(P:S) = \min D_\psi(P,R) \qquad \text{over } R \text{ in } S \qquad (8)$$

we study the problem of finding the point in S that is closest to P in the sense of divergence. This gives an approximation of P by using a point inside S.

We say a Curve $\theta(t)$ is orthogonal to S when its tangent vector $\dot{\theta}(t)$ is orthogonal to any tangent vectors of S at the point of intersection $\hat{P}_S$.

Definition 10.19: Such a point of intersection $\hat{P}_S$ is called the geodesic projection of P to S when the geodesic connecting P and $\hat{P}_S \in S$ is orthogonal to S. Dually, $\hat{P}_S^*$ is the dual geodesic projection of P to S when the dual geodesic connecting P and $\hat{P}_S^* \in S$ is orthogonal to S.

We now have the projection theorem stated as follows.

Theorem 10.20: (projection theorem): Given $P \in M$ and a submanifold $S \subset M$, the point $\hat{P}_S^*$ that minimizes the divergence $D_\psi(P, R), R \in S$ is the dual geodesic projection of P to S. The point $\hat{P}_S$ that minimizes the dual divergence $D_{\psi^*}(P, R), R \in S$, is the geodesic projection of P to S.

Proof: Let $\hat{P}_S^*$ be the dual geodesic projection of P to S. Consider a point $Q \in S$ which is infinitesimally close to $\hat{P}_S^*$. Then the three points $P, \hat{P}_S^*$ and Q form an orthogonal triangle because the small line element connecting $\hat{P}_S^*$ and Q is orthogonal to the dual geodesic connecting P and $\hat{P}_S^*$. Hence by the Pythagorean relation we get

$$D_\psi(P, Q) = D_\psi(P, \hat{P}_S^*) + D_\psi(\hat{P}_S^*, Q) \tag{9}$$

for any neighboring Q. This shows $\hat{P}_S^*$ is a critical point of $D_\psi(P, Q), Q \in S$. q.e.d.

Remark: If S is flat submanifold of a dually flat manifold M, the dual projection point $\hat{P}_S^*$ of P to S is unique and minimizes the divergence and its dual version also holds.

References

[1] N. Hicks: a theorem on affine connection Ill. Jl. of Math. 3 (1959) p 342-

[2] S. Kobayashi and K. Nomizu: Diffl. Geometry Vol. I.

[3] E. Spanier: algebraic topology, Springer 1996.