Some invariants of statistical manifolds

M. Sitaramayya[1]
School of Mathematics and Statistics,
University of Hyderabad, Hyderabad
email: malledi_sr@yahoo.com

K. Moinuddin
Department of Mathematics,
MANU University, Gachibowli, Hyderabad

Abstract

First we reformulate Gauss Theorema Egregium in terms of Inertial frames and give formula for Gaussian curvature K. Then we study the geometry of exponential families and more generally of statistical manifolds using differential forms. We define the first and second fundamental forms of a statistical manifold and give analogues of Gauss invariant K for statistical manifolds. Statistical manifolds are realized as injective immersions into an affine space. In appendix Amari's minimum divergence theorem was derived via differential forms and existence of inertial coordinate systems on an embedded surface S in $\mathbb{R}^3$ was proved after giving some historical remarks on them.

PART - 1
Gauss invariant (Modern approach)

1 Introduction

Many mathematical objects or concepts can be understood by finding various characterizations of them or by studying several invariants associated with them. For example, Gauss while studying the geometry of surfaces in $\mathbb{R}^3$ noticed in 1853 a concept, nowadays called as Gaussian curvature, that behaves extraordinarily which he stated and proved as 'Theorema Egregium'. We will have a relook of this remarkable theorem. We give Gauss's original formula and then simplify his method by a reduction using Inertial coordinate system and bring out the intrinsic nature of this invariant.

Then we attempt to understand some similar invariant concepts for statistical manifolds. As is well known in statistics, the simplest type of statistical manifolds are exponential families of probability distributions and then transformation models of probability measures under group actions. These models enjoy rich geometry and amenable to techniques of affine and differential geometries.

Here we focus on the geometry of exponential families which enjoy a natural affine geometry. We give various ways of understanding them by giving 4 characterizations of them, defining on the way the concept of first and 2nd fundamental forms associated to a statistical manifold P and also the concept of generalized statistical curvature.

---

[1] Invited talk for the Bharat Ganita Parishad Annual Conference at the Department of Mathematics & Astronomy, University of Lucknow during Nov. 18-19, 2017, delivered on 19th Novemeber 2017.

In [5] we studied the most generalized geometry called (F, G)-geometry on a statistical manifold P and investigated several properties of it. In [6] we investigated the geometry of dually flat structures and dual structures on statistical manifolds using differential geometric techniques and gave several examples and its invariant properties using convex analysis.

The exponential families and curved exponential families are rich in geometry, in particular they are naturally equipped with affine structure and hence have a vanishing second fundamental form which concept characterizes them. We investigate some invariants intrinsically associated with them. We attempt to study the geometry of statistical manifolds via differential forms approach throughout. In particular we interpret the Fisher-Rao Riemannian metric as a differential 1-form having values as 1-forms. This sophisticated way of looking brings out the essence of statistical divergences as an integrated version of Fisher-Rao metric g on statistical manifold P. The divergence minimizing theorem of Amari appears in differential form version and it is capable of generalizing to spacial fibrations with a local approximation by an estimator. Amari's minimum divergence theorem was proved rigorously using differential forms in Appendix Part A. In Part-B the existence of inertial coordinate systems was proved in two ways geometrically and analytically and some historical and philosophical remarks on them were given.

Acknowledgements:

The first author thanks Prof. Raj Kishore Singh for his continuous encouragement and inspiration in writing down our ideas in this research article. This paper was written while the first author is visiting the department of Mathematics of MANU University, Hyderabad and he thanks Prof. S.N. Hassan, Head of the Department of Mathematics for the excellent facilities provided and hospitality extended during the time of writing this article.

2 Gauss Theorema Egregium (Modern approach)

We setup the modern differential geometric approach for an embedded surface S in 3-space and give expression for its Gaussian curvature K.

2.1 First fundamental form and its geometric information:

Let S be a surface embedded in 3-space $\mathbb{R}^3$ with a coordinate patch
$\sigma : \cup^{\text{open}} \subset \mathbb{R}^2_{u,v} \to S \subset \mathbb{R}^3$. If $\gamma(t) = \sigma((u(t), v(t))$ is a curve of S in σ, then the tangent vector at $\gamma(t)$ of γ is given by

$$\dot{\gamma} = \sigma_u \dot{u} + \sigma_v \dot{v} \tag{2.1}$$

(where $\cdot = \frac{d}{dt}$, σ_u =partial derivative w.r.t. u). Then arc length of γ from $\gamma(t_o)$ to $\gamma(t)$ is given by

$$s = \int_{t_o}^{t} \|\dot{\gamma}(u)\| du \quad \text{where} \quad \|\dot{\gamma}(u)\|^2 = \dot{\gamma}(u) \cdot \dot{\gamma}(u) \tag{2.2}$$

$$= \int_{t_o}^{t} \left(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2 \right) dt \quad \text{where} \quad E = \sigma_u \cdot \sigma_u,\ F = \sigma_u \cdot \sigma_v,\ G = \sigma_v \cdot \sigma_v \tag{2.3}$$

Definition 2.1. *The quadratic differential* $\Pi_1 = Edu^2 + 2Fdudv + Gdv^2$ *with* E, F, G *as in (3) is called the first fundamental form of* σ *in* S.

Remark 2.1. *Define the infinitesimal arc length ds by $ds^2 = Edu^2 + 2Fdudv + Gdv^2$. Then $s = \int \sqrt{ds^2}$. Hence the first fundamental form Π_1 measures the length of curves in S intrinsically.*

Definition 2.2. *An isometry f between two surfaces S_1 and S_2 is a diffeomorphism carrying curves of S_1 to curves of S_2 of same length.*
In view of above remark we have.

Theorem 2.1. *(a) A diffeomorphism $f : S_1 \to S_2$ is an isometry if and only if for any surface patch σ of S_1, the patches σ and $f o \sigma$ of S_1 and S_2 respectively have the same first fundamental form Π_1..*
Theorem Egregium 4(b) (Gauss): *If $f : S_1 \to S_2$ is an isometry of surfaces then they have the same Gaussian curvature K, i.e. K is invariant under isometries.*

2.2 Area of a region and second fundamental form Π_2 of surface S

Let (U, σ) be a patch in S and let R be a region in $\cup^{\text{open}} \subset \mathbb{R}^2_{uv}$. Let $\sigma(R)$ be its image in S. Then the area of $\sigma(R)$ in S is $\mathcal{A}_\sigma(R)$ given by

$$\int_R \int \|\sigma_u \times \sigma_v\| du dv$$

and

$$\|\sigma_u \times \sigma_v\| = (\det Q)^{1/2} = (EG - F^2)^{1/2}$$

where $Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$ is the first fundamental matrix of Π_1 of patch σ in S.
This area $\mathcal{A}_\sigma(R)$ is invariant under reparametrization of the surface patch σ.

Let σ be a patch on S in $\mathbb{R}^3$ with standard unit normal N. For two nearby points $P = \sigma(u, v)$ and $Q = \sigma(u + du, v + dv)$ of S, as we move from P to Q the surface moves away from its tangent plane at $\sigma(u, v)$ by a distance $(\sigma(u + du, v + dv) - \sigma(u, v)) \cdot N$. By using Taylor expansion of $\sigma(u + du, v + dv)$ we get this deviation of S from its tangent plane is approximately given by

$$\frac{1}{2}\left(edu^2 + 2f dudv + gdv^2\right) \quad \text{where}$$

$e = \sigma_{uu} \cdot N$, $f = \sigma_{uv} \cdot N$ and $g = \sigma_{vv} \cdot N$

Definition 2.3. *The quadratic differential $\Pi_2 = edu^2 + 2f dudv + gdv^2$ is called the second fundamental form of σ in S.*

Remark 2.2. *1) The form Π_2 has information about the curvature of σ on S.*
2) The second fundamental matrix $L = \begin{pmatrix} e & f \\ f & g \end{pmatrix}$ where $e(u, v), f(u, v), g(u, v)$ are functions on S and the first fundamental matrix $Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$ determine the Gaussian curvature K of S.
3) Consider the equation $\det(L - \lambda Q) = 0$. Let λ_1, λ_2 be its roots. Then the quantities $H = \frac{1}{2}(\lambda_1 + \lambda_2)$

and $K = \lambda_1\lambda_2$ are called the mean curvature and the Gaussian curvature of S respectively. In terms of fundamental coefficient functions $E(u,v), F(u,v), G(u,v)$; $e(u,v), f(u,v), g(u,v)$ they are $K = \frac{eg-f^2}{EG-F^2}$ and $H = \frac{eG-2fF+gE}{2(EG-F^2)}$. The Gaussian curvature K is invariant under reparametrizations.

2.3 Infinitesimal interpretation of Guassian curvature K of S

First for a curve γ, its curvature is the rate of change of direction of the tangent vector of γ per unit length. Similarly, the direction of the tangent plane to a surface patch $\sigma : U \to \mathbb{R}^3$ is measured by its standard unit normal $\hat{N}$, suggesting the curvature of σ is measured by the rate of change of N per unit area.
More precisely, first note that $\hat{N}$ is a point of the unit sphere $S^2 = \{v \in \mathbb{R}^3 | \|v\| = 1\}$ in $\mathbb{R}^3$ with its base point varying on S. Define the Gauss map: $S \to S^2$ where S is the image of σ sending $\sigma(u,v) \in S$ to $\hat{N}(u,v)$ of S^2.

If $R \subset U$ is a region, the amount by which $\hat{N}$ varies over the part $\sigma(R)$ of S is measured by the area of the part $\hat{N}(R)$ of the sphere. Then the rate of change of $\hat{N}$ per unit area is approximately

$$\frac{\text{area } \hat{N}(R)}{\text{area } \sigma(R)} = \mathcal{A}_{\hat{N}}(R)/\mathcal{A}_{\sigma}(R). \tag{$*$}$$

Geometrically as region R shrinks to a point P, the limit of this ratio gives value of $|K_P|$.

Remark 2.3. *This infinitesimal version of K as in (*) or $K = \frac{eg-f^2}{EF-G^2}$ involves the surface normal $\hat{N}$s movement in the ambient space and is extrinsic in nature. But Gauss observed K can be measured within the surface S itself by expressing the expression $eg - f^2$ in terms of E, F and G. This made Gauss to call his result 'thma Egregium'.*

2.4 Proof of Gauss Theorem:

Take a smooth orthonormal basis $\{e', e''\}$ of the tangent plane at each point of σ i.e., $e'(u,v), e''(u,v)$ are smooth functions of u and v. Then $B = \{e', e'', \hat{N}\}$ is an orthonormal basis of $\mathbb{R}^3$ where $\hat{N}$ is the standard unit normal of σ, with $\hat{N} = e' \times e''$. Express the partial derivatives of e', e'' w.r.t. u and v in terms of basis B as

$$e'_u = \alpha e'' + \lambda' N \qquad e'_v = \beta e'' + \mu' N \tag{2.1}$$

and

$$e''_u = -\alpha e' + \lambda'' N \qquad e''_v = -\beta e' + \mu'' N \tag{2.2}$$

with $\alpha, \beta, \lambda', \mu', \lambda'', \mu''$ scalars depending on u, v then we have
Lemma (Technical): We have the relations:

$$e'_u \cdot e''_v - e''_u \cdot e'_v = \lambda'\mu'' - \lambda''\mu' \tag{2.3}$$

$$= \alpha_v - \beta_u \tag{2.4}$$

$$= \frac{eg - f^2}{(EG - F^2)^{1/2}} \tag{2.5}$$

Using this lemma proof of Gauss theorem: we have

$$K = \frac{eg - f^2}{EG - F^2} = \frac{\alpha_v - \beta_u}{(EG - F^2)^{1/2}} \quad \text{(by (4) and (5))} \tag{2.6}$$

Hence it sufficies to show for a suitable choice of (e', e'') the scalars α and β depend only on E, F, G. We construct this by Gram-Schmidt process applied to the basis $\{\sigma_u, \sigma_v\}$ of the tangent plane. Once we do this then the Gaussian curvature K is intrinsically determined by the first fundamental form Π_1 of S and an isometry preserves Π_1. Then Gauss Theorem follows.

$$\text{First define } e' = \frac{\sigma_u}{\|\sigma_u\|} = \epsilon\sigma_u \quad \text{where } \epsilon = E^{-1/2} \tag{2.7}$$

Let $e'' = \gamma\sigma_u + \delta\sigma_v$ for scalars γ, δ, s.t. $e'' \perp e'$ and $\|e''\| = 1$. From these conditions we get

$$\delta = \frac{E^{1/2}}{(EG - F^2)^{1/2}},\ \gamma = \frac{FE^{-1/2}}{(EG - F^2)^{1/2}}\ ,\ \epsilon = E^{-1/2} \tag{2.8}$$

Thus

$$e' = \epsilon\sigma_u,\ e'' = \gamma\sigma_u + \delta\sigma_v \tag{2.9}$$

with above ϵ, γ, δ depending only on E, F, G.
Then we can compute α and β:

$$\begin{aligned} \alpha =& e'_u \cdot e'' = (\epsilon_u\sigma_u + \epsilon\sigma_{uu}) \cdot (\gamma\sigma_u + \delta\sigma_u) \\ =& \frac{1}{2}\epsilon\gamma E_u + \epsilon\delta\left(F_u - \frac{1}{2}E_v\right) \text{(after some calculation)} \end{aligned} \tag{2.10}$$

which indeed depend only on E, F, G.
Finally

$$\begin{aligned} \beta = e'_v \cdot e'' =& (\epsilon_v\sigma_u + \epsilon\sigma_{uv}) \cdot (\gamma\sigma_u + \delta\sigma_v) \\ =& \frac{1}{2}\epsilon\gamma E_v + \frac{1}{2}\epsilon\delta G_u \qquad \text{(after some calculation)} \end{aligned} \tag{2.11}$$

which depends only on E, F, G.
$\therefore$ The Gaussian curvature K of a surface is intrinsically given by Π_1 and hence is invariant under isometry of surfaces. q.e.d.

Remark 2.4. *The Gaussian curvature K is an intrinsic invariant of the surface S.*
Finally substituting the actual values of γ, δ, ϵ into the formulae for α and β and using $K = \frac{\alpha_v - \beta_u}{(EG-F^2)^{1/2}}$ we get the explicit formula for the Gaussian curvature K as follows

Corollary 2.1. *The Gaussian curvature K of S in terms of functions E, F and G is given by the formula*

$$K = \frac{\begin{vmatrix} -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu} & \frac{1}{2}E_u & F_u - \frac{1}{2}F_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{vmatrix} - \begin{vmatrix} 0 & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{vmatrix}}{(EG - F^2)^{1/2}} \tag{2.12}$$

Remark 2.5. *Special cases:*

$$\text{(2.13)} \qquad a) \quad F = 0,\ K = -\frac{1}{2\sqrt{EG}}\left\{\frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right) + \frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right)\right\}$$

$$\text{(2.14)} \qquad b) \quad E = 1 \text{ and } F = 0,\ K = \frac{-1}{\sqrt{G}}\frac{\partial^2\sqrt{G}}{\partial u^2}$$

3 A new reduction to Gauss Theorem via Inertial frames

3.1 General setup for oriented hypersurfaces in $\mathbb{R}^n$

Let Y be an oriented hypersurface in $\mathbb{R}^n$. that is, at each point x of Y we have chosen a unique (positive) unit normal vector N_x depending smoothly on x. This defines a well defined Gauss map $\nu : Y \to S^{n-1}$ sending x to $\nu(x) = N_x$, S^{n-1}: unit sphere in $\mathbb{R}^n$ and its differential $d\nu_x : T_x(Y) \to T_{\nu(x)}(S^{n-1}) \leftrightarrow T_x(Y)$ is a linear map, called the Weingarten map $W_x = d\nu_x$

Definition 3.1. *The 2nd fundamental form Π_2 of $Y^{n-1} \subset \mathbb{R}^n$ is the bilinear form on T_xY defined by*

$$\text{(3.1)} \qquad \Pi_2(v, w) = (W_x v, w)$$

where $(,)$ is the inner product of $\mathbb{R}^n$, $\forall\, v, w \in T_xY$. Then W_x is symmetric in v, w i.e., $(W_x u, v) = (u, W_x v)$, $\forall\, u, v$ and hence W_x is digonalizable with real eigenvalues.

Let $B = \{X_1, X_2, X_3, \cdots, X_{n-1}\}$ be a basis of T_xY.

$$\text{(3.2)} \qquad N_i = \sum_j W_{ji} X_j \quad (i = 1, 2, \cdots, n-1)$$

where $X_j, N_i \in \mathbb{R}^n$. Then (2) in matrix form is

$$\text{(3.3)} \qquad (N_1, N_2, \cdots, N_{n-1}) = (X_1, X_2, \cdots, X_{n-1})W$$

(N_is and X_js are column vectors of size n)

Multiplying (3) on the left by the matrix $(X_1, X_2, \cdots, X_{n-1})^T$ we get the matrix equation for $(n-1) \times (n-1)$ matrices

$$\text{(3.4)} \qquad L = QW$$

where $L = (L_{ij})$ with $L_{ij} = (X_i, N_j)$ and $Q = (Q_{ij})$ with $Q_{ij} = (X_i, X_j)$.

Definition 3.2. *We call $Q = (Q_{ij})$ the matrix of the first fundamental form Π_1 w.r.t. chosen coordinates $(y_i)_{i=1}^n$ on Y around x with $y_n = 0$ and basis B.*

L is called the matrix of the 2nd fundamental form Π_2 around x.

Taking dterminants of (4) we get

$$\text{(3.5)} \qquad \det W = \frac{\det L}{detQ}$$

Thus we proved.

Proposition 3.1. *The determinant* det W *of the Weiengarten map* W *which is a geometric property of the embedded hyper surface* Y *in* $\mathbb{R}^n$ *can be expressed as the quotient of two local expressions namely as the ratio of determinants of second and first fundamental matrices* L *and* Q.

Now specialising for $n-1=2$ *we have an embedded surface* S *in* $\mathbb{R}^3$ *and (5) becomes the Gauss formula for Gaussian curvature*

$$K = \det W = \frac{\det L}{\det Q} \tag{3.6}$$

3.2 Intrinsic geometry associated with first fundamental form Π_1 of Y in $\mathbb{R}^n$.

The first fundamental form Π_1 of hypersurface Y in $\mathbb{R}^n$ (or surface S in $\mathbb{R}^3$) encodes the intrinsic geometry of Y (or S) in terms of local coordinates and it gives the Euclidean geometry of the tangent space (plane) in terms of the basis $X_1, X_2, \cdots, X_{n-1}$, $(n-1=2)$.
For example, if we describe a curve $t \to \gamma(t)$ in Y in terms of the coordinates $y_1, y_2, \cdots, y_{n-1}$ by giving the functions $t \to y^j(t)$, $j = 1, 2, \cdots, n-1$ then by chain rule we have

$$\gamma'(t) = \sum_{j=1}^{n-1} X_j(y(t)) \frac{dy^j(t)}{dt} \quad \text{where } \; y(t) = (y^1(t), y^2(t), \cdots, y^{n-1}(t))$$

So the Euclidean squared length of the tangent vector $\gamma'(t)$ is

$$\|\gamma'(t)\|^2 = (\gamma'(t), \gamma'(t)) = \sum_{i,j=1}^{n-1} Q_{ij}(y(t)) \frac{dy^i(t)}{dt} \frac{dy^j(t)}{dt} \tag{3.7}$$

which is an intrinsic formulae for curve length in standard form.

Coming back to the embedded surface case S in $\mathbb{R}^3$ take $y_1 = u$, $y_2 = v$, then $X_1 = X_u$ partial derivative w.r.t. u, $X_2 = X_v$ and take $\cdot$ for scalar product of $\mathbb{R}^3$. Then clearly

$$Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

where $E = X_u \cdot X_u$, $F = X_u \cdot X_v$, $G = X_v \cdot X_v$ (called first fundamental coefficients) and so $\det Q = EG - F^2$.
Similarly as motivated in section 2 set $e = N \cdot X_{uu}$, $f = N \cdot X_{uv}$ and $g = N \cdot X_{vv}$ (called 2nd fundamental coeffients), and $L = \begin{pmatrix} e & f \\ f & g \end{pmatrix}$. Then $\det L = eg - f^2$.
So (6) becomes in this special case

$$K = \frac{eg - f^2}{EG - F^2} \tag{3.8}$$

which gives a local expression for the Gaussian curvature of S in $\mathbb{R}^3$.

3.3 Effect of change of coordinates of S on Π_1

Suppose we change coordinates from (u, v) to (u', v') on S via a coordinate patch $X(u, v)$ where $u = u(u', v')$, $v = v(u', v')$. Then we get

$$\begin{pmatrix} X_{u'} \\ X_{v'} \end{pmatrix}^T = \begin{pmatrix} X_u \\ X_v \end{pmatrix}^T J \quad \text{where } J = \begin{pmatrix} \frac{\partial u}{\partial u'}, & \frac{\partial u}{\partial v'} \\ \frac{\partial v}{\partial u'}, & \frac{\partial v}{\partial v'} \end{pmatrix}$$

is a matrix-valued function of u' and v', called the Jacobian and so

$$Q' = (X_{u'}, X_{v'})^T (X_{u'}, X_{v'}) =^t J(X_u, X_v)^T (X_u, X_v) J =^t JQJ \tag{3.9}$$

$$Q' = {}^tJQJ \tag{9a}$$

which is the transformation rule for Π_1 from (u, v) to (u', v') coordinate system change on S. Here t_J is the transpose of the Jacobian J.

Now let P be a point on S. By linear algebra we can always find a matrix K s.t. ${}^tKQ(u_p, v_p)K = I_2$. Take coordinates of P as $(0, 0)$. Then make a linear change of coordinates so that $J(0, 0)$ is K. Thus ${}^tKQ(0, 0)K = I_2$
Find coordinates such that $J(0, 0) = I$ and then $Q(0, 0) = I_2$ in this coordinate system.
In fact one can do more, namely, we can choose coordinates such that

$$Q(0) = I, \ \frac{\partial Q}{\partial u}(0, 0) = \frac{\partial Q}{\partial v}(0, 0) = 0 \tag{3.10}$$

holds.
We call such a coordinate system at $P \in S$ with coordinates (0,0) an inertial coordinate system of S at P.
(we prove (10) in Appendix, Part B)

Remark 3.1. *Note that the collection of all inertial coordinate systems based at P is intrinsically associated to the metric, since their definition depends only on the properties of Q in the coordinate system, just like the set of all orthonormal frames of S at p is intrinsic object depending on the metric.*

Theorem 3.1. *If $\{u, v\}$ is an Inertial coordinate system of an embedded surface S based at P then the Gaussian curvature is given by*

$$K(P) = F_{uv} - \frac{1}{2}G_{uu} - \frac{1}{2}E_{vv} \tag{3.11}$$

where the R.H.S. is evaluated at (0,0).

Remark 3.2. *Note that formula (11) expresses the Gaussian curvature K in terms of local expression for the metric in an inertial coordinate system, this theorem implies Gauss Theorema Egregium.*

Proof of theorem: Step 1 reduction: First make a rotation and translation in $\mathbb{R}^3$ so that $X(P)$ is at the origin and $T_P(S)$ is the xy-plane in $\mathbb{R}^3$. Then $Q(0) = I$ implies that the vectors $X_u(0), X_v(0)$ form an orthonormal basis of the xy-plane.
So by a further rotation, if necessary, we may assume that X_u is the unit vector in the positive x-direction and X_v is the unit length vector in the positive y-direction. These euclidean motions have no effect on the curvature.
After these reductions our embedded surface S in $\mathbb{R}^3$ can be parametrically described by

$$X(u,v) = \begin{pmatrix} u + r(u,v) \\ v + s(u,v) \\ h(u,v) \end{pmatrix} \tag{3.12}$$

where r, s and h are functions which vanish together with their first derivatives at the origin in the uv-space. All this reductions we could do using only $Q(0) = I$ of the definition of inertial coordinate system.
Step 2 we claim that if the coordinate system is inertial then all the second partials of r and s also vanish at the origin.
In fact, we have

$$X_u = \begin{pmatrix} 1 + r_u \\ s_u \\ h_u \end{pmatrix} \; ; \; X_v = \begin{pmatrix} r_v \\ 1 + s_v \\ h_v \end{pmatrix}$$

so that

$$\begin{aligned} E = X_u \cdot X_u =& (1 + r_u)^2 + s_u^2 + h_u^2 && \text{(a)} \\ F = X_u \cdot X_v =& (1 + r_u) r_v + s_u (1 + s_v) + h_u h_v && \text{(b)} \\ \text{and} \quad G = X_v \cdot X_v =& r_v^2 + (1 + s_v)^2 + h_v^2 && \text{(c)} \end{aligned}$$

Since $Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$ and $\frac{\partial}{\partial u} Q(0) = 0$ can be computed and, $\frac{\partial}{\partial v} Q(0)$ can be computed

$$\begin{aligned} E_u =& 2(1 + r_u) r_{uv} + 2 s_{uu} s_u + 2 h_{uu} h_u \\ E_v =& 2(1 + r_u) r_{uv} + 2 s_u s_{uv} + 2 h_u h_{uv} \\ \text{and so} \quad E_u(0) =& 2 r_{uu}(0) \; ; \; E_v(0) = 2 r_{uv}(0) \end{aligned} \tag{3.13}$$

Similarly calculate

$$F_u(0) = r_{uv}(0) + s_{uu}(0); \quad F_v(0) = r_{vv}(0) + s_{uv}(0) \tag{3.14}$$

Similarly calculate

$$G_u(0) = 2 s_{uv}(0), G_v(0) = 2 s_{vv}(0) \tag{3.15}$$

By (13), (14) and (15), $\frac{\partial Q}{\partial u}(0) = 0$ and $\frac{\partial Q}{\partial v}(0) = 0$ give that all the 2nd partials of r and s vanish at (0,0).
Step 3 completion of proof of formula in (11).

Now $N(0)$ the surface normal vector is just the unit vector in the positive z-direction. Since $e = N \cdot X_{uu}$; $f = N \cdot X_{uv}$ and $g = N \cdot X_{vv}$ we get $e = h_{uu}$, $f = h_{uv}$ and $g = h_{vv}$.
Now $Q(0) = I_2 = \begin{pmatrix} E(0) & F(0) \\ F(0) & G(0) \end{pmatrix}$ gives $EG - F^2 = 1$ at the origin.
Hence the Gaussian curvature K at $P(0,0)$ is $\frac{eg-f^2}{EG-F^2}$ at the origin, which is equal to

$$h_{uu}h_{vv} - h_{uv}^2 \tag{3.16}$$

From formula (a),(b),(c) taking partials and evaluating at (0,0) gives

$$F_{uv} = r_{uvv} + s_{uuv} + h_{uu}h_{vv} + h_{uv}^2$$
$$E_{vv} = 2[r_{uvv} + h_{uv}^2] \text{ and } G_{uu} = 2[s_{uuv} + h_{uv}^2]$$

and hence $K(P) = h_{uu}h_{vv} - h_{uv}^2 = F_{uv} - \frac{1}{2}E_{vv} - \frac{1}{2}G_{uu}$ at the origin proving formula (11). q.e.d.
We close this section with some remarks.

Corollary 3.1. *The non-linear coordinate choice of Inertial coordiantes over Euclidean ones as in §2 which has the power of transforming goedesic curves to straight lines enormously simplified the formula of Gauss for curvature K (compare formula (12) of section 2 with formula (11) of section 3). (see Appendix for remarks on inertial coordinates)*

Remark 3.3. *(1) By formula (11) of §3 K is an intrinsic invariant of the surface S in $\mathbb{R}^3$.*
(2) As we worked with a tangential frame $\begin{pmatrix} X_u \\ X_v \end{pmatrix}$ or B for $T(S)$ we can as well work with the coframe $\begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix}$ for $T^*(S)$ and as in the work of Elie Cartan we get the Cartan equation:

$$K = \det W = (\omega_{31} \wedge \omega_{32})(E_1, E_2) \ \textit{where} \ (\theta_1 \wedge \theta_2)(E_1, E_2) = 1 \,.$$

and hence

$$\begin{aligned} \omega_{31} \wedge \omega_{32} &= K\theta_1 \wedge \theta_2 \\ \text{and} \qquad d\omega_{12} &= -\,K\theta_1 \wedge \theta_2 \end{aligned} \tag{3.17}$$

Formulas (17) show the Gaussian curvature K is an intrinsic invariant of S in $\mathbb{R}^3$ as starting with any orthonormal frame field (E_1, E_2) adopted to surface S and its corresponding dual coframe field $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ and above Cartan Calculus works giving (17) and the concept of an orthonormal frame field on a surface S depends only on the intrinsic geometry of S. (cf [4], [7], [8]).

PART -2:
Geometry of Statistical manifolds via differential forms

1 Geometry of exponential families

Parametric statistics is concerned with parametrized families of probability distributions $P = \{p(x, \theta) : \theta = (\theta^1, \theta^2, \cdots, \theta^n) \in U$ given subset of $\mathbb{R}^n\}$ defined over some sample space X or Ω.

The most common example is the normal family $N(\theta) = \{p(x, \theta) | x \in \mathbb{R}$ sample space, $\theta = (\mu, \sigma) \in \mathbb{R}^2$ with $\mu \in \mathbb{R}$ and $\sigma > 0\}$ where $p(x, \mu, \sigma) = \frac{1}{\sqrt{2\Pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ is a density function w.r.t. the Lebesgue measure dx on the sample space $\mathbb{R}$.

From a geometric point of view such parametric families P of probability distributions support a variety of geometries and these are entities independent of any particular parametrization and enjoy some 'shape concept' depending on a particular parametrization. The statistical properties of P may sometimes indicate significance of a special geometry on P to understand P better, for example affine geometry for the study of exponential families.

Definition 1.1. *A parametric family $P = \{p(x, \theta)\}$ where*

$$p(x, \theta) = \exp\left(\sum_{i=1}^{n} x^i \theta^i - K(\theta)\right) d\mu \tag{1.1}$$

where $\theta = (\theta^1, \theta^2, \cdot, \theta^n)$ are parameters; x^i random variables, K: normalizing function and μ a measure on a sample space or measure space is called an exponential family.

Remark 1.1. *We regard a parametrized family P of probability distributions as an analogue of a surface S with a coordinate system on S in statistics. Then $p(x, \theta)$ represents a point on S and its parameter θ is its coordinates*

The normal family $N(\mu, \sigma)$ is an exponential family in above sense as it can be parametrized as

$$p(x, \theta) = p(\theta^1, \theta^2, x) = \exp(x^2\theta^1 + x\theta^2 - K(\theta)) \tag{1.2}$$

where

$$\theta^1 = \frac{-1}{2\sigma^2}\ ,\ \theta^2 = \frac{\mu}{\sigma^2} \ \text{ and }\ K(\theta) = \frac{1}{2}\log\left(\frac{-\Pi}{\theta^1}\right) - \frac{(\theta^2)^2}{4\theta^1} \tag{1.3}$$

The parameter $\theta = (\theta^1, \theta^2)$ is called the canonical parameter in statistics and it varies over the open subset U of $\mathbb{R}^2$ given by $\theta^1 < 0$.

Proposition 1.1. *The exponential family P has an affine structure, namely its coordinate systems are affinely related.*

Proof: Let the density function $p(x, \theta) \in P$ be given by (1). Let us suppose we have another set of random variables, parameters and normalizing function $\{y^i(x), \varphi^i(\theta), J(\theta)\}$ for P so that

$$p(x, \theta) = \exp\left(\sum_{i=1}^{n} y^i(x)\varphi^i(\theta) - J(\theta) + f(x)\right) \tag{1.4}$$

with a new measure ν such that $d\mu = f(x)d\nu$ on the sample space X or Ω. Then from (1) and (4) we get

$$\sum_{i=1}^{n} x^i\theta^i - K(\theta) = \sum_{i=1}^{n} y^i(x)\varphi^i(\theta) - J(\theta) + f(x) \tag{1.5}$$

which on differentiation w.r.t. x^i gives

$$\theta^i = \sum_{j=1}^{n} \frac{\partial y^j}{\partial x^i}\varphi^j + \frac{\partial f}{\partial x^i} \quad \text{for } i = 1, 2, \cdots, r \tag{1.6}$$

choosing a point θ such that $\varphi^i(\theta) = 0\ \forall i$ so that $\xi^i = \frac{\partial f}{\partial x^i}$ must be a constant vector and $X_i^j = \frac{\partial y^j}{\partial x^i}$ must be a constant matrix giving that the two canonical parameter systems $\{\theta^i\}$ and $\{\varphi^j\}$ are related by an affine transformation

$$\theta^i = \sum_{j=1}^{r} X_j^i\varphi^j + \xi^i \tag{1.7}$$

Thus on the exponential family P any two canonical coordinate systems are affinely related giving on affine structure to P.

Abstract Setup for realizing a parametric family P:

If Ω is a measure space, consider the set of families of non-negative measures which are absolutely continuous w.r.t. each other i.e. measures μ and ν are equivalent if they have the same sets of measure zero. Under this relation measures in any equivalence class form a family $\mathcal{M}$ which is naturally an affine space. That is, R_Ω is the vector space of measurable functions f on Ω and the translation by f is the operation of sending the measure $d\mu$ to $e^f d\mu$.

Definition 1.2. *An affine space is a pair (X, V) consisting of a set X and a vector space V acting as a set of transformations: $X \times V \to X$ (called translations) on X such that*

(i) $(p + v) + w = p + (v + w)$, $\forall\, p \in X$, $v, w \in V$

(ii) for $\forall$ two points $p, q \in X$, there exists a unique vector $v \in V$ such that $q = p + v$

A measure class $\mathcal{M}$ is an affine space under its log-likelihood structure $l : \mathcal{M} \to R_\Omega : e^f\mu \to f$.

Note that expressing measures as densities w.r.t. any one of the measures say μ of a measure class $\mathcal{M}$ is exactly the process of choosing μ as the origin of this affine space $\mathcal{M}$. Then define $l : \mathcal{M} \to R_\Omega$ by $l(pd\mu) = \log p$ where p is the density function, the log-likelihood function.

Thus treating measures as densities w.r.t. a base measure μ and considering the log-likelihood of these densities amounts precisely to choosing an origin for $\mathcal{M}$ and identifying points of $\mathcal{M}$ with their translation vectors from the origin.

Now let $\mathcal{P}$ denote the space of all probability measures in $\mathcal{M}$. But $\mathcal{P}$ is not an affine subspace in $\mathcal{M}$. To get this we modify the equivalence relation on measures above, by considering nonnegative measures upto scale, i.e. $\mu \sim \nu$ if $\nu = e^f \mu$, f is a constant function. Under this equivalence relation, $\mathcal{M}$ becomes an affine space and $\mathcal{P}$ regarded as non-negative measures upto scale form an affine subspace of $\mathcal{M}$ i.e. the set corresponding to finite measures upto scale. Thus we claim the following:

Proposition 1.2. *A family P of measures upto scale which form a finite dimensional affine subspace is an exponential family.*

Proof: Recall that an affine subspace is a subset which is closed under translation by a subspace of translation vectors. If this subspace is finite dimensional and is spanned by the random variables $x^1, x^2, \cdots, x^r$ and if μ is one of the measures upto scale then the measures in such a family have the form

$$\mu + \theta^1 x^1 + \theta^2 x^2 + \cdots + \theta^r x^r = \exp(\theta^1 x^1 + \cdots + \theta^r x^r - K)\mu \tag{1.8}$$

where changing K changes the scale of the measure and hence such P is an exponential family by definition 1 above.

Remark 1.2. *The probability measures in this subspace correspond to those non-negative measures upto scale which are finite and thus the set $\mathcal{P}$ of all probability measures regarded as non-negative measures upto scale forms an affine subspace of $\mathcal{M}$ namely set of finite measures upto scale. Moreover the precise probability measure is obtained by choosing K to be $K(\theta)$ as determined by the formula*

$$K(\theta) = \log \int_\Omega \exp(\theta^1 x^1 + \cdots + \theta^r x^r)\mu \tag{1.9}$$

Thue we have

Theorem 1.1. *A parametric family $P \subset \mathcal{P} \subset \mathcal{M}$ is an exponential family if and only if P is a finite dimensional affine subspace of measures upto scale with their natural log-likelihood affine structure.*

Remark 1.3. *Note that to realize P in above proposition 7 in the explicit form of exponential family given by (8) and (9) we have to construct an affine coordinate system for this finite dimensional affine subspace P via a choice of origin and a choice of an ordered basis for the space of translations. That is, the canonical coordinates $\theta^1, \theta^2, \cdots, \theta^r$ are affine coordinate system for the affine geometry and hence any two such coordinate systems are affine related as we proved above in proposition 1.*

A Geometric realization of exponential family P

Let $P = \{p(\theta)\}$ be a parametric family of probability measures regarded as a 'surface' in $\mathcal{P} \subset \mathcal{M}$. Choose an origin and define the log-likelihood function $l(\theta) : \mathcal{M} \to R_\Omega$. Assume $l(\theta)$ is a differentiable function of θ and P has a well defined tangent space at each point $p \in P$. This is an affine subspace of $\mathcal{M}$ which is tangent to P at p and hence has the form $\{p + v = \exp(v)p | v \in V\}$ for some vector space V of random variables. We call V the tangent space of P at p denoted by $T_p(P)$. This vector space T_pP has basis consisting of the score vectors $l_i = \frac{\partial l}{\partial \theta^i}(p)$ $i = 1, 2, \cdots, r$.

Since $P \subset \mathcal{P}$, $\mathcal{P}$ is affine space of positive measures upto scale, extend P as a family $\tilde{P}$ of positive measures by $\tilde{P} = \{\exp(\lambda)p | \lambda \in \mathbb{R}, p \in P\}$. Also extend the local coordinates θ^i by

$\theta^i(e^\lambda p) = \theta^i(p)$ and define a new coordinate θ^o by $\theta^o(e^\lambda p) = \lambda$.
Let $\mathbb{R} \cdot 1$ denote the vector space of constant random variables. Consider the quotient vector space $\mathbb{R}_\Omega/\mathbb{R} \cdot 1$ which is the space of random variables upto addition of constants. Then we have.

Proposition 1.3. *P is an affine subspace of $\mathcal{P}$ generated by a vector space $V \subset \mathbb{R}_\Omega | \mathbb{R} \cdot 1$ iff $\tilde{P}$ is an affine subspace of $\mathcal{M}$ generated by $\tilde{V} = \{f \in \mathbb{R}_\Omega | f + R \cdot 1 \in V\}$. The RHS of this proposition means: choose origin μ for $\tilde{P}$ then every other measure in $\tilde{P}$ has the form $e^{\tilde{l}}\mu$ where $\tilde{l}$ ranges over the $f.d.$ subspace $\tilde{V}$ of R_Ω. Extend l function as $\tilde{l}(e^\lambda p) = l(p) + \lambda$ and let $\tilde{l}_i$ denote the derivatives of $\tilde{l}(i = 0, 1, 2, \cdots, r)$. Note $\tilde{l}_o$ is the constant $r.v.1$. Since all the values of $\tilde{l}$ are in the vector space $\tilde{V}$ so all the values of $\tilde{l}_o, \tilde{l}_1, \cdots, \tilde{l}_r$ lie in $\tilde{V}$ and they form basis for $\tilde{V}$. When $\tilde{P}$ is affine, the partial derivatives of the score vectors $\tilde{l}_i$ namely $\tilde{l}_{ij}$ must also lie in the space of translations $\tilde{V}$. Thus $\frac{\partial^2 \tilde{l}}{\partial\theta^i\partial\theta^j} = \tilde{l}_{ij}$. can be written as*

$$\tilde{l}_{ij}(\theta) = \sum_{k=0}^{r} \gamma_{ij}^k(\theta)\tilde{l}_k(\theta) \tag{1.10}$$

equivalently, the 2nd derivatives of l i.e. l_{ij} are in the span of the scores l_i and the constant $r.v.1$. Thus we get

Proposition 1.4. *(a) The derivatives of the scores lie in the span of the scores at each point is the characteristic property of affine subspaces (11)*

(b) In particular $P \subset \mathcal{P} \subset \mathcal{M}$ is an exponential family of probability measures iff (11) holds. We illustrate this result in (b) for the normal family.

$p(x, \mu, \sigma) = \frac{1}{\sqrt{2\Pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$ and then $l(p, (x, \mu, \sigma)) = \frac{-(x-\mu)^2}{2\sigma^2} - \log(\sqrt{2\Pi}\sigma)$ and scores are $l_1 = \frac{x-\mu}{\sigma^2}$, $m_2 = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$ and their non-zero 2nd partials are $l_{11} = \frac{-1}{\sigma^2}$, $l_{12} = l_{21} = \frac{-2(x-\mu)}{\sigma^3}$ and $l_{22} = \frac{-3(x-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2}$. Then we have the relations:
$l_{11}(\mu, \sigma) = \frac{-1}{\sigma^2} \cdot 1$; $l_{12}(\mu, \sigma) = \frac{-2}{\sigma} l_1(\mu, \sigma)$ and $l_{22}(\mu, \sigma) = \frac{-3}{\sigma} l_2(\mu, \sigma) - \frac{2}{\sigma^2} \cdot 1$.
These conditions hold for $l_{oj}(\mu, \sigma)$ trivally as they are all zero.

A differential geometric realization of exponential family: We give a geometric criterion to decide whether a given parameter family P admits an exponential parametrization.
Setup as above; $P \subset \mathcal{P} \subset \mathcal{M}$ and $\tilde{P}$ extension and then by above proposition, it suffices to decide whether the 2nd derivatives l_{ij} are in $T_p\tilde{P}$ for each $i, j = 1, 2, \cdots, r$. Let E_p denote the expectation of functions w.r.t measure p. Define an inner product on the subspace of p-square integrable functions f i.e.

$$E_p(f^2) < \infty \qquad \text{by} \qquad < f, g >_p = E_p(fg) \tag{1.12}$$

Definition 1.3. *The $r \times r$ matrix*

$$g_{ij}(p) = E_p(l_i l_j) \tag{1.13}$$

is called the Fisher-information matrix.

Remark 1.4. *If we restrict this inner product to T_pP which has basis $\{l_i\}_{i=1}^r$, it defines inner product on the tangent space T_pP. The Fisher Information matrix is the matrix representation of this inner product w.r.t the basis of scores $\{l_i\}_{i=1}^r$.*

More generally we have the above inner product $<,>$ on the space of all random variables R_Ω given by (12). Using $<,>$ on R_Ω we have

$$R_\Omega = T_p\tilde{P} + N_p \tag{1.14}$$

where N_p is the orthogonal complement of $T_p\tilde{P}$ in R_Ω w.r.t. $<,>$ called the normal space of P to $T_p\tilde{P}$. Then $\forall\ f \in R_\Omega$ has its tangential and normal component given by

$$\sum_{m,n} g^{mn} E_p(f l_m) l_n \text{ and } f - \sum_{m,n} g^{mn} E_p(f l_m) l_n - E_p(f) \tag{1.15}$$

where $g^{m,n}$ is the inverse of the Fisher matrix and the normal component of f is denoted by $\Pi_p(f)$ with $\Pi_p : R_\Omega \to N_p$ projection map.
Indeed the later vector $\pi_p(f)$ is in N_p as $< \pi_p(f), 1 >= 0$ and $< \pi_p(f), l_i >= 0 \cdot \forall\ i$. Hence we get

Proposition 1.5. *A parameter family $P \subset \mathcal{P} \subset \mathcal{M}$ is an exponential family iff its 2nd fundamental form $\alpha = (\alpha_{ij})$ vanishes.*
Proof: *By (11) P is exponential family iff the functions l_{ij} are always tangential to $\tilde{P}$ i.e. iff the normal component of each l_{ij} vanishes. That is to say,*

$$\alpha_{ij} \overset{def}{=} \pi_p(l_{ij}) = l_{ij} - \sum_{m,n} g^{mn} E_p(l_{ij} l_m) l_n - E(l_{ij}) = 0 \tag{1.16}$$

Remark 1.5. *1) For this definition the motivation is the 2nd fundamental form matrix L for a hypersurface $Y \subset \mathbb{R}^n$ given in part I and it measures the flatness of a surface.*
2) Geometric implication of vanishing of $\alpha = (\alpha_{ij})$: Regard the scores $\tilde{l}_i$ as tangent vectors to $\tilde{P}$ and as we move around in $\tilde{P}$ these tangent vectors change and the $\tilde{l}_{ij}$ is measuring this change. The changes are due to (i) the lines in P in which the parameters are constant are bending around and (ii) the surface $\tilde{P}$ itself is bending around in $\mathcal{M}$ and the tangential components and normal components of $\tilde{l}_{ij}$'s measure precisely these two types of bending respectively. So the vanishing of the second fundamental form α of P implies the surface $\tilde{P}$ is not bending in $\mathcal{M}$ i.e., the structure of $\tilde{P}$ is rigid and hence the affine structure of $\tilde{P}$ as an affine subspace of $\mathcal{M}$ is characterized by the vanishing of α.

Definition 1.4. *The quantity α_{ij} or $\alpha = (\alpha_{ij})_{r\times r}$ matrix is called the second fundamental form of the parametric family P*

Definition 1.5. *We can define a scalar function γ on the parameter family P depending on the 2nd fundamental form $\alpha = (\alpha_{ij})$ by*

$$\gamma(p) = \sum_{ijkl} g^{ij}(p) g^{kl}(p) E_p(\alpha_{ik}(p)\alpha_{jl}(p)) \tag{1.17}$$

called the generalizaed statistical curvature of the family P, where g^{ij} is the inverse of the Fisher information matrix.

Remark 1.6. *Efron [3] defined the statistical curvature γ for a 1-dimensional family $\{p(\theta)|\theta \in \mathbb{R}\}$. We have the following.*

Proposition 1.6. *A parametric family P is exponential iff its generalized statistical curvature γ vanishes.*

Proof: P is exponential $\Leftrightarrow \alpha = (\alpha_{ij})$ vanishes $\Leftrightarrow \gamma$ vanishes as the inverse $(g^{ij}(p))$ Fisher matrix is positive definite.
Putting together all the above parts we get

Theorem 1.2. *For a family $P \subset \mathcal{P} \subset \mathcal{M}$ of parametric measures to be an exponential one the following are equivalent.*

(a) A certain subspace $\tilde{P}$ of probability measures be affine (this is independent of parameters).

(b) The second derivatives of the log-likelihood are in the span of the scores and the constant r.v. functions.

(c) The second fundamental form $\alpha = (\alpha_{ij})$ of the family P vanishes.

(d) The generalized statistical curvature function of the family P vanishes.

Remark 1.7. *(1) Property in (a) characterising exponential family is independent of the parameters.*
(2) The property in (b) can be proved to be invariant under a change of parameters from $(\theta^i)_{i=1}^r$ to $(\chi^j)_{j=1}^r$.
(3) The vanishing of 2nd fundamental form of P is an invariant property independent of parametrization even though the fundamental form α itself depends, on the parametrization as we have

$$\alpha_{ab} = \alpha_{ij} \frac{\partial \theta^i}{\partial \chi^a} \frac{\partial \theta^j}{\partial \chi^b} \tag{1.18}$$

where (θ^i) and (χ^j) are two different parametrizations of P, $(\alpha_{ab}) = \alpha_\chi$, $(\alpha_{ij}) = \alpha_\theta$ second fundamental forms of P w.r.t (χ) and (θ)
(18) gives the change of parameter formula for the second fundamental form of P and α_θ vanishes iff α_χ vanishes.
(4) Similarly the Fisher information is not independent of parametrization but obeys the rule

$$g_{ab} = g_{ij} \frac{\partial \theta^i}{\partial \chi^a} \frac{\partial \theta^j}{\partial \chi^b} \tag{1.19}$$

where

$$g_{ij} = E_p \left(\frac{\partial^2 l}{\partial \theta^i \partial \theta^j} \right)$$

and

$$g_{ab} = E_p \left(\frac{\partial^2 l}{\partial \chi^a \partial \chi^b} \right)$$

and $l = \log$ likelihood function of P.
Using formulae (18) and (19) and the definition of the curvature function γ of P it can be checked that γ is invariant of the family P independent of parametrization. That is, γ is independent of the "shape" of P but is a property of the 'surface' P.

2 Statistical Manifolds and second fundamental form

Definition 2.1. *We say a subset Q of a manifold P is a submanifold if around each point q in Q we can find coordinates $\theta = (\theta^1, \theta^2, \cdots, \theta^n)$ on P such that the points of Q near q are precisely those for which $\theta^{r+1}, \cdots, \theta^n$ vanish and we say locally Q is inside P.*

Example 2.1. *If P is an exponential family then a submanifold Q of P is called a curved exponential family.*
For such Q we have a function $\theta : Q \to \mathbb{R}^n$ such that $q(x, \theta) = \exp\left(\sum_i x^i \theta^i(q) - K(q)\right)$ is a probability distribution on Q. Here in Q, θ is not bijection but it is on P.

Definition 2.2. *For a manifold Q, a smooth map $i : Q \to P$ is called an immersion if its tangent map or differential $di_q : T_q(M) \to T_{i(q)}P$ is an injective linear map at each point q in Q.*

Remark 2.1. *(1) An immersion makes Q locally a submanifold of P whereas an injective immersion makes Q as a submanifold of P.*
(2) We regard a submanifold Q is inside manifold P and the inclusion map $i : Q \to P$ is a smooth map and $di_q : T_q(Q) \to T_q(P)$ is isomorphism into at $\forall\, q \in Q$.
Our special case in statisties: Let P be an affine space modelled on a vector space V. Then we can identify T_pP with the vector space V and so the embedding of a submanifold Q in P has a tangent map $d_pi : T_pQ \to T_pP = V$ and identifies $T_p(Q)$ with a subspace of V.
More precisely, fix an origin p for P and define $l : P \to V$ by $q = p + l(q)$ and d_pl carries $\frac{\partial}{\partial\theta^i} \to \frac{\partial l}{\partial\theta^i}$ and is an isomorphism into.

Now in our statistical set up with $P \subset \mathcal{P} \subset \mathcal{M}$ where $\mathcal{M}$ is a measure class, $\mathcal{P}$ is the affine space of all probability measures and P is a parametric family of probability distributions with a sample or measure space Ω. We regard a statistical manifold P as a 'surface' in the affine space $\mathcal{P}$ or P is injectively immersed in $\mathcal{P}$. More precisely,

Definition 2.3. *(a) Let P be a subset of $\mathcal{P} \subset \mathcal{M}$. let $f : P \to R_\Omega$ be a function. We say f is smooth if $\forall\, x \in \Omega$, the real-valued function $p \to f(p)(x) : P \to \mathbb{R}$ is smooth (provided P has a differentiable structure).*
(b) We define a statistical manifold is a subset P of $\mathcal{P}$ the space of probability measures in a measure class $\mathcal{M}$ on a sample space Ω such that P is a manifold and satisfies:
(i) choosing an origin μ for $\mathcal{M}$, the log-likelihood function $l : P \to R_\Omega : p = e^f \mu \to \log(e^f) = f$ is a smooth function in the sense of (a) above.
(ii) for any point p and coordinates θ around p the random variables $\frac{\partial l}{\partial\theta^i}(p)$
$l = 1, 2, \cdots, r$ are linearly independent.

Remark 2.2. *(i) is essentially the requirement that inclusion map $i : P \to \mathcal{P}$ is a smooth map*
(ii) is the requirement that map i is an immersion.

Hence a statistical manifold is an injective immusion of P into the affine space $\mathcal{P}$ or its image in $\mathcal{P}$ and $l : P \to R_\Omega$ is smooth and its differential $d_pl : T_p(P) \to R_\Omega$ is a linear map called the score. In coordinates (θ^i) of P,

$$d_pl\left(\frac{\partial}{\partial\theta^i}\right) = \frac{\partial l}{\partial\theta^i}(p) : \Omega \to \mathbb{R}.$$

(ii) says $\left\{ \frac{\partial l}{\partial \theta^i}(p) \right\}$ *$i = 1, 2, \cdots, n$ is a basis of $T_p(P)$.*
In statistical literature the vector

$$\left(\frac{\partial l}{\partial \theta^1}, \frac{\partial l}{\partial \theta^2}, \cdots \frac{\partial l}{\partial \theta^r} \right)$$

is called a score vector on P or a frame of P.

Note that the score can be understood as an R_Ω-valued differential 1-form on P given by

$$d_p l = \sum_i \frac{\partial l}{\partial \theta^i}(p) d_p \theta^i \tag{2.1}$$

Thus we have proved

Theorem 2.1. *A statistical manifold P is a family of probability measures for which the log-likelihood function l is a differentiable map and the score $d_p l$ is an inclusion identifying $\frac{\partial}{\partial \theta^i}$ with $\frac{\partial l}{\partial \theta^i}$ (That is, $T_p(P) \hookrightarrow R_\Omega$)*

Example 2.2. *For the exponential family $P = \{p(x, \theta)\}$, its log-likelihood function is $l(p)(x) = \sum \theta^i x^i - K(p)$ and its score as differential form is*

$$d_p l = \sum_i x^i d_p \theta^i - d_p K \tag{2.2}$$

and the score fucntion $\frac{\partial l}{\partial \theta^i} : \Omega \to \mathbb{R}$ is an element of R_Ω and image of a basic tangent vector $\frac{\partial}{\partial \theta^i}$ under score is $\frac{\partial l}{\partial \theta^i} \in R_\Omega$, which is $x^i - \frac{\partial K}{\partial \theta^i}$, $i = 1, 2, \cdots, r$.
(b) For the curved exponential family Q in P given by

$$q(x, \theta) = \exp\left(\sum_i x^i \theta^i(q) - K(q) \right),$$

if χ is a local coordinate system on Q then we have its score

$$d_q l = \sum_j \left(\sum_i x^i \frac{\partial \theta^i}{\partial \chi^j} - \frac{\partial K}{\partial \chi^j} \right) d_q \chi^j \tag{2.3}$$

Second fundamental form of a statistical manifold P :
We give a coordinate-free definition of second fundamental form α of a statistical manifold P.

In our setup we have an inner product on the space of random variables R_Ω defined by $< f, g >_p = E_p(fg)$ and w.r.t. this $<,>$ we have decomposition $R_\Omega = T_p\tilde{P} \oplus N_p$. Let Π_p denote the projection from R_Ω onto N_p given explicitly by $\Pi_p f = f - \sum g^{mn} E_p(f l_m) l_n - E_p(f)$

Definition 2.4. *If X and Y are vector fields on P we define its second fundamental form α of P by*

$$\alpha(p)(X,Y) = \pi_p(X(Y(l)) \tag{2.4}$$

at each point $p \in P$. In coordinates (θ^i) on P we have

$$\begin{aligned} \alpha(p)(X,Y) =& \pi_p\left(X^iY^j\frac{\partial^2 l}{\partial\theta^i\partial\theta^j} + X^i\frac{\partial Y^j}{\partial\theta^i}\frac{\partial l}{\partial\theta^j}\right) \\ =& \pi_p\left(X^iY^j\frac{\partial^2 l}{\partial\theta^i\partial\theta^j}\right) \end{aligned} \tag{2.5}$$

as $\pi_p\left(\frac{\partial l}{\partial\theta^j}\right) = 0$
Thus $\alpha : T_pP \times T_pP \to N_p$ is a bilinear symmetric map for each $p \in P$ and its components w.r.t coordinate tangent vectors are given by

$$\alpha(p)\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right) = \pi_p\left(\frac{\partial^2 l}{\partial\theta^i\partial\theta^j}\right) = \alpha_{ij} \tag{2.6}$$

which was our earlier formula for exponential family P (cf. (16) above).

Remark 2.3. *In the general case of a curved exponential family Q in P higher order fundamental forms $\alpha^{(k)}$ of Q can be defined as we specify some specific subspace V of N^p valued bilinear form on T_pP into V.*
Just like the case of exponential family P, the vanishing of these $\alpha^{(k)}$s characterises a curved exponential family Q.

3 Statistical manifolds and differential forms

Statistical manifold P has a natural Riemannian metric g called the Fisher-Rao metric which is covariant under reparametrizations and invariant under random variable transformations. We call such quantities invariant. We discussed most general such geometries on statistical manifolds, studied their properties and their invariance in [5,6]. For example the Amari-Chentsov 3-tensor is invariant. Because of the F-R metric on P and as a Riemannian metric gives a 1-1 correspondence between vector fields and 1-forms, a statistical manifold P comes naturally equipped with a dual structure (P, ∇, ∇^*) and also dually flat structures on P [6].
More generally statistical manifolds have natural affine connections like exponential connection, mixture connection, a family of α-connections (α real) on P which are all invariant.

A Riemannian metric g on P as a bilinear form $g(v, w)$ on $T_p(P)$ can be interpreted by changing its arguments v and w one at a time as a 1-form with values as 1-forms.
Given a connection ∇ on P, η is a 1-form on P, then the rate of change of a 1-form w.r.t. ∇ gives a 1-form with values as 1-form i.e., $\nabla\eta(v) \in T_p^*P, \forall\, v \in T_p(P)$. If (θ^i) is a coordinate system on P, $\nabla\eta$ is completely determined by the values of $\nabla d\theta^i, i = 1, 2, \cdots, r$.
In general for a smooth function f on P, ∇df is a symmetric bilinear form. This suggests that

given a connection ∇ on P express the Riemannian metric g as rate of change of 1-form θ w.r.t ∇, i.e. solve the equation

$$\nabla\theta = g \qquad \text{for} \qquad \theta \tag{3.1}$$

If such θ exists, it has to be closed since g is symmetric and so θ is of the form $d\psi$ for some function ψ on P. Both θ and ψ can be made unique by specifying initial conditions: i.e.

$$\nabla\theta = g \tag{3.2}$$

and

$$\theta = d\psi \qquad \text{on} \qquad P \Leftrightarrow \nabla d\psi = g. \tag{3.3}$$

Let ψ_p be the solution of $\theta = d\psi$ for which $\psi_p(p) = 0$ and $d\psi_p = 0$. Then such solutions can be found iff ∇ (and hence ∇^*) is flat and ∇ and ∇^* are symmetric.
Thus we proved

Theorem 3.1. *Let P be a statistical manifold. Given a connection ∇ on P and a Riemannian metric g then g can be realized as the rate of change of a 1-form θ w.r.t. ∇ i.e. $\exists$ 1-form θ on P s.t. $\nabla\theta = g$ and θ is of the form $\theta = d\psi$, ψ a smooth function on P iff ∇ (and hence ∇^*) is flat and both ∇ and its dual ∇^* w.r.t. g are symmetric.*

Proof: We proved everything above except closedness of θ. $\nabla\theta = g$ means $\nabla\theta(v)(w) = g(v,w)\,\forall\, v, w \in T_p(P)$. Then

$$d\theta(v,w) = \nabla\theta(v)(w) - \nabla\theta(w)(v) = g(v,w) - g(w,v) = 0$$

since g is symmetric and so θ is a closed 1-form and hence by Poincaré lemma θ is exact locally and hence there eixsts a smooth function ψ s.t. $\theta = d\psi$. With suitable initial conditions on ψ and θ one can prove that they are unique. For existence of such 1-form θ, ∇ must be flat. □

Now we apply this theorem to the exponential family $P = \{p(\theta, x)\}$ which is a statistical manifold sitting as a submanifold of $\mathcal{P} \subset \mathcal{M}$, where $p(x,\theta) = \exp\left(\sum \theta^i x^i + k(x) - K(\theta)\right)$ be probability density function so that $\int p(x,\theta)dx = 1$ giving

$$K(\theta) = \log \int \exp\left(\sum \theta^i x^i + k(x)\right) dx \tag{3.4}$$

Let ∇ be the exponential connection given by the ($\alpha = 1$) embedding $p \to \log p$. Then there exists a smooth function ψ_P on P s.t.

$$\nabla d\psi = g \tag{3.5}$$

where g is the Fisher Rao metric. But we know from [6], differentiating (29) w.r.t. the affine coordinates (θ^i) of P $\frac{\partial}{\partial\theta^i}K(\theta) = \int x^i p(x,\theta)dx\,\forall\, i$
$\therefore$ The 1-form $\varphi = \sum \frac{\partial}{\partial\theta^i}K(\theta)d\theta^i$ is obtained and so

$$\nabla\varphi = \sum_{i,j} \frac{\partial^2 K}{\partial\theta^i\partial\theta^j} d\theta^i d\theta^j\,. \tag{3.6}$$

Evaluating this bilinear form on basic tangent vectors

$$\frac{\partial}{\partial\theta^i} \quad \text{and} \quad \frac{\partial}{\partial\theta^j} \quad \text{we get}$$

$$\nabla\varphi\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right) = g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right) \quad \text{and so}$$

$$\sum_{i,j} \frac{\partial^2 K}{\partial\theta^i\partial\theta^j} d\theta^i d\theta^j \left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right) = g_{ij} = -\int_\Omega \frac{\partial^2 l}{\partial\theta^i\partial\theta^j} p(x,\theta)dx = \frac{\partial^2 K}{\partial\theta^i\partial\theta^j} \tag{3.7}$$

Thus we have that for $\nabla\varphi = \nabla d\psi = g$, g:Fisher-Rao metric above computation showed the solution function ψ in the Poincaré Lemma is the K function in the definition of exponential family. As we proved above in (32)

$$\frac{\partial^2 K}{\partial\theta^i\partial\theta^j} = -\int \frac{\partial^2 l}{\partial\theta^i\partial\theta^j} p(x,\theta)dx = g_{ij}$$

$$\text{i.e.}\nabla dK = g \ \text{ i.e. } \ \nabla\varphi = g, \quad \varphi = dK \tag{3.8}$$

Thus the normalizing function K in the definition of exponential family P is naturally giving rise to a solution of (33) for the exponential family P with ∇ being the exponential connection (log p-embedded) and with Fisher-Rao Riemannian metric.
Thus we have

Theorem 3.2. *P is exponential family with exponential connection (which is flat) and g is Fisher-Rao metric. Then the cumulant generating function K on P and Kullback-Leibler information η on P for the dual connection i.e. mixture connection on P are intrinsic invariants of P depending on the geometric data of P only, not on the embedding of P in $\mathcal{P} \subset \mathcal{M}$.*

Proof: We proved this above as

$$E_p\left(\frac{-\partial^2 l}{\partial\theta^i\partial\theta^j}\right) = \frac{\partial^2 K}{\partial\theta^i\partial\theta^j}$$

which gives

$$-\int \frac{\partial^2 l}{\partial\theta^i\partial\theta^j} p(x,\theta)dx = \frac{\partial^2 K}{\partial\theta^i\partial\theta^j} \tag{3.9}$$

the L.H.S. of (34) is the Fisher information of P and the R.H.S. is a constant random variable on Ω.
$\therefore$ K can be measured on P by the Fisher-Rao Riemannian metric g, which is the analogue of first fundamental form for statistical manifolds. So it is an intrinsic invariant of P independent of embedding in $\mathcal{P} \subset \mathcal{M}$.
Similarly with identity embedding $p \to p$ ($\alpha = -1$ case) the solution function of $\theta = d\psi$ and $\nabla\theta = g$ gives $\psi_p(q)$ is the Kullback-Leibler information of $q(\theta)$ relative to $p(\theta)$ given by

$$\psi_p(q) = \int_\Omega p(x)\frac{\log p(x)}{q(x)}dx$$

p fixed, q varies as distribution $p = p(x,\theta)$, $q = p(x,\theta)$. Thus this $\psi_p(q)$ is intrinsically determined by the Fisher information. ☐

Remark 3.1. *(1) This result is a statistical analogue for exponential families of Gauss Theorem Egregium for surfaces $S \subset \mathbb{R}^3$ or $Y^{n-1} \subset \mathbb{R}^n$ with its Gaussian curvature K as an intrinsic invariant. The Fisher information is the analogue of first fundamental form and K is completely determined by this intrinsic first fundamental form.*
(2) The above result is a special case of general principle that the geometric second derivative of a statistical divergence function ψ on a statistical manifold P is related to Riemannian metrics on P. In this sense the above theorem is a geometric realization of F.R. metric g on P in terms of statistical divergences. [cf. 6].

We can define a statistical divergence in terms of differential forms as.

Definition 3.1. *The statistical divergence on a statistical manifold P determined by a given Riemannian metric g and a symmetric connection ∇ is a function $\psi(p, q)$ in pairs of points of P such that if ψ_p is the function defined by $\psi_p(q) = \psi(p, q)$ (i.e. holding p fixed) then we have*

$$\nabla d\psi_p = g \tag{3.10}$$

$$d\psi_p(v) = 0 \quad \forall\, v \in T_p P \tag{3.11}$$

$$\text{and} \quad \psi_p(p) = 0 \tag{3.12}$$

Remark 3.2. *In [6] we gave analytic definition of divergence ψ in terms of taylor expansion of ψ around p and constructed many such divergences using smooth convex functions.*

By our above discussion of the system $\nabla\theta = g$ and $\theta = d\psi$ of PD Eqns., a statistical divergence can be constructed by finding, for each point p, a solution θ_p of the equation $\nabla\theta = 0$ which satisfies $\theta_p(v) = 0$ for $\forall\, v \in T_p P$. This makes the solution θ of $\nabla\theta = g$ unique as the difference of two such solutions $\theta_1 - \theta_2 = \alpha$ is a 1-form such that $\nabla\alpha = 0$ and α vanishes on $T_p P$ and hence α is constant. By parallel translation its value is zero on $T_q P \;\forall\, q \in P$. Given θ_p there is a function ψ_p satisfying $d\psi_p = \theta_p$. this ψ_p is unique by the condition $\psi_p(p) = 0$.
Setting $\psi(p, q) = \psi_p(q)$ on $P \times P$ we get a statistical divergence.
We close this article with a few remarks.

Remark 3.3. *(1) By definition of ψ_p, it has a critical point at p with value zero at p i.e., $\psi_p(p) = 0$. Since its Hessian form $\nabla d\psi = g$ where g is positive definite, this critical point p gives a minimum value ψ_p can have. There are no other critical points and so ψ_p is a convex function with a unique minimum value zero occurring at p. Thus $\psi(p, q) \geq 0 \;\forall\, p, q \in P$ and $\psi(p, q) = 0$ iff $p = q$.*
(2) The condition that $\nabla\theta = g$ should have a solution, at each point $p \in P$, which is zero on $T_p P$ forces the connection ∇ to be flat. Thus a statistical divergence ψ determined by a Riemannian metric g and a connection ∇ on a statistical manifold exists iff ∇ (and hence its dual ∇^*) are flat and both ∇ and ∇^* are symmetric connections.
(3) Amari's divergence minimizing, theorem via dual geodesics can be geometrically derived in the setup of differential forms on statistical manifolds [Amari [2], [6] projection theorem]. (cf. Appendix, Part A for details)
We have

Theorem 3.3. *(Amari):The point q on a submanifold S of a statistical manifold P which has minimum divergence from point p of P not in S, must have the dual geodesic from p to q cutting S orthogonally at q.*
(see Appendix for details)

(4) The above theorem has generalization by replacing the geodesic by ancilliary submanifolds N *with a dual connection which is transversal to* S *in* P *giving* P *a fiber-bundle like structure (see [1] Theorem 5.6 on asymptotic estimation).*
(5) For discrete random variable case the exponential family behaves like a universal model in the sense that any statistical model having discrete $r.v.$ *can be regarded as a submanifold of an exponential family i.e., realized as curved exponential family.*
(6) For continuous $r.v.$ *case: there are many statistical models with continuous* $r.v$ *which are not subfamilies of exponential family but most of them are curved exponential families. In the case of a truly non-exponential statistical model we use its local approximation by using a larger exponential family. This gives rise to bundle-like structure with exponential connection on the base and mixture connection on the fibers and the fibers meet base transversally. Amari (cf [1] asymptotics and estimators).*
(7) For exponential family P*, the cummulant generating function* K *is so much an intrinsic part of* P *geometrically that, just like the score vector* $\left(\frac{\partial l}{\partial \theta^1}(p), \cdots, \frac{\partial l}{\partial \theta^r}(p)\right)$ *is a basis for* $T_p(P) \subset R_\Omega$*, the free energy vector* $\left(\frac{\partial K}{\partial \theta^1}(p), \frac{\partial K}{\partial \theta^2}(p), \cdots, \frac{\partial K}{\partial \theta^r}(p)\right)$ *is precisely the coordinates of the preimage of the point p under the inverse of the maximum likelyhood estimator map in the sample space* Ω*, i.e. under the map* $MLE : \Omega \to P$*.*
(8) In a sequel we investigate the general asymptotic estimators on statistical manifold P *and MLE's asymptotics using differential forms approach giving the deep geometry of amari's fibration.*

References

[1] S. Amari: Differential geometrical Methods in Statistics. Lecture Notes in Statistics 28, 1985.

[2] S. Amari and H. Nagaoka Methods in information geometry Amer. Math. and Oxford Univ. Press 2000.

[3] B. Efron: Defining the curvature of a statistical problem., annals. of Statistics, 3 (1975) p. 1189-1242.

[4] S. Kobayashi, K. Nomizu: Foundations of Differential Geometry Interscience 1963 Vol. 1.

[5] M. Sitaramayya, K.S.S. Moosath, K.V. Harsha: Generalized geometric structures on statistical manifolds Ganita vol. 65, (2016) p 19-44.

[6] M. Sitaramayya, K.S.S. Moosath, S.N. Hasan, Differential Analytic aspects of Statistical models GANITA vol. 67 (2017) p. 1-31.

[7] M. Spivak, A comprehensive Introduction to Differential Geometry Publish or pensh, Boston vol. I, 1970.

[8] S. Sternberg: Introduction to Differential geometry. Prentice Hall 1970.

Appendix

PART-A: Amari's theorem in differential forms set up

In this part we give a rigorous proof of Amari's theorem using differnetial forms.
Let P be a statistical manifold and S be a submanifold. Then a point $q \in S$ will be a minimum value point for a function f given on S if the differnetial df is zero on

(3.1) $Tq(S)$ i.e. $df(v) = 0 \quad \forall\, v \in T_q(S)$.

Definition 3.2. *The gradient vector field ∇f is defined by*

(3.2) $df(v) = g(\nabla f, v)$

where g is some Riemannian metric on P.
Then above condition df vanishes on $T_q(S)$ is equivalent to $\nabla f(q)$ the gradient vector of f at q is orthogonal to $T_q(S)$. (3)

Now we give a geometric interpretation of condition (3).
Let X be any vector field in P. By an integral variation of X we mean a variation $\gamma(t)$ whose tangent vector at each point p is given by $X(p)$. That is, γ must satisfy

(3.4) $$\gamma'(t) = X(\gamma(t)) \qquad \text{for every } t.$$

Let $\{\varphi^1, \varphi^2, \cdots, \varphi^n\}$ be a coordinate system on P. Then applying the differentials $d\varphi^1, d\varphi^2, \cdots, d\varphi^n$ to equation (4) we have

(3.5) $$\frac{d}{dt}\varphi^i \circ \gamma(t) = d\varphi^i(\gamma(t)) = d\varphi^i(X(\gamma(t)), \qquad i = 1, 2, \cdots, n$$

Denote $\varphi^i \circ \gamma(t)$ by $\varphi^i(t)$ and write $X = \sum_{i=1}^n X^i \left(\frac{\partial}{\partial \varphi^i}\right)$ so that $d\varphi^i(X) = X^i$. Then (5) can be put in coordinate form as

(3.6) $$\dot{\varphi}^i = \bar{X}^i(\varphi^i, \varphi^2, \cdots, \cdots, \varphi^n), \qquad i = 1, 2, \cdots, n$$

where $\bar{X}^i$ is the coordinate form of the function X^i. (6) is a system of ordinary differential equations and by the standard existence and uniqueness that solutions exist over small time intervals and are unique. Thus for any smooth vector field X, through every point p of the manifold P there passes a unique integral variation of X (Frobenius theorem). In particular, given a function f, the integral variations of the gradient vector field $X = \nabla f$ trace out a family of curves in an increasing direction of f. In fact, if γ is an integral variation of ∇f then

$$(f \circ \gamma)'(t) = df(\gamma'(t)) = g(\nabla f(\gamma(t)), \qquad \nabla f(\gamma(t))) \geq 0$$

The curves traced out by these integral variations of ∇f are called the paths of steepest descent of f oriented in the direction of decreasing values of f. Since the tangents to these curves are in the direction of ∇f at each point, in order that q be a minimum value point of f over the submanifold S, the curve of steepest descent through $\underline{q}$ must be orthogonal to S at q.

Now we discuss the global situation:
Suppose that, as in the case of statistical divergences, the minimum value of the function f over the entire manifold P occurs at a unique point p. Then along its curves of steepest descent, the values of f are strictly decreasing and hence the values must approach the minimum value and be confined to a compact regiion of P containing the unique minimum point p.
Thus p must be a limit of curves of steepest descent. In other words, all curves of steepest descent radiate out from p. Hence we can say that for a statistical divergence $\psi(p,q)$, the curves of steepest descent of the function ψ_p all radiate into p and if q is the point on any submanifold S at which $\psi(p,q)$ attains its minimum value, then the curve of steepest descent from p through q must cut orthogonally through S.
This is the geometric meaning of (3).

Next we want to identify the curves of steepest descent for the function ψ_p with geodesics of $\hat{\nabla}$, the dual connection of ∇, emanating from p.(7)
Let $\theta = d\psi_p$ be the differential 1-form and $X_\theta = \nabla\psi_p$ be the dual of θ. Then the integral variations of $\nabla\psi_p$ trace out curves of steepest descent for ψ_p.
By the definition of dual connection $\hat{\nabla}$ of ∇ we have
$\nabla\theta(Y)(Z) = g(\hat{\nabla}X_\theta(Y), Z)$ for vector fields Y and Z.
Hence for any vector field Y, since θ is a solution of the equation $\nabla\theta = g$ we have $\hat{\nabla}X_\theta(Y) = Y$
Since $\hat{\nabla}(X_\theta) = X_\theta$ we have $\hat{\nabla}(fX_\theta) = df(X_\theta)(X_\theta) + fX_\theta$ (since we have the formula $\hat{\nabla}(fX) = df(X) + f\hat{\nabla}X$.
Now the equation $df(X_\theta) + f = 0$ is a first order-partial differential equation which can be solved by the method of characteristics giving rise to a family of ordinary differential equations along the integral variations of X_θ giving the solution function f. With this choice of f we conclude that

$$\begin{aligned}\hat{\nabla}(fX_\theta)(fX_0) &= f(\hat{\nabla}(fX_\theta)(X_\theta))\\ &= f(df(X_\theta)+f)X_\theta = 0\,.\end{aligned}$$

If γ is any integral variation of fX_θ then by definition $\hat{\nabla}\gamma'(\gamma') = 0$. Hence the integral variations of fX_θ which is a rescaled gradient vector field of ψ_p, are geodesics of $\hat{\nabla}$ which proves (7) above.

Now it remains only to observe that the integral variations of any multiple fX of a vector field X trace out the same curves as the integral variations of X itself (8).
In fact, for any variation $\gamma(t)$ we have by chain rule that $\gamma(F(t))$ has tangent vector $F'(t)\gamma'(F(t))$ and hence $\gamma(F(t))$ and $\gamma(t)$ trace the same cvurve. If $\gamma(t)$ is integral variation of X then $\gamma(F(t))$ will be an integral variation of fX provided F satisfies

$$\begin{aligned}F'(t)\gamma'(F(t)) &= f(\gamma(F(t)))X(\gamma(F(t)))\\ &= f(\gamma(F(t)))\gamma'(F(t)) \qquad (3.9)\end{aligned}$$

That is, we must choose F to satisfy the ordinary differential equation

$$F'(t) = f(\gamma(F(t))) \qquad (3.10)$$

which can always be solved so that the integral variations of fX always have the form $\gamma(F(t))$ where $\gamma(t)$ is an integral variation of X.

Thus the integral variations of fX and X trace out the same curves in the general case which proves (8).

In particular, the integral variations of the gradient vector field of a statistical divergence $\psi(p,\cdot) = \psi_p$ i.e. of $\nabla\psi_p$ trace out geodesics of dual connection $\hat{\nabla}$, since there is a rescaling of $\nabla\psi_p$ whose integral variations are $\hat{\nabla}$-geodesics.
Thus we proved.

Theorem 3.4. *(Amari): The point q on the submanifold S of a statistical family P which has minimum divergence from a point p not in S, must have the $\hat{\nabla}$-geodesic from p to q cutting S orthogonally.*

Remark 3.4. *This theorem admits a deeper generalization giving best estimation for such q giving rise to bundle-like fibrations with applications in statistical inference and estimation which will be discussed in a sequel.*

PART-B: Some Remarks on Inertial Coordinate systems and Gauss Lemma

Inertial coordinate systems or Inertial frames have a long history since the days of Galileo and many great Mathematicians like Galileo, Newton, Gauss, Einstein etc., thought over this concept considerably. We give some historical remarks and philosophical thoughts on them below and then attempt to rigourously derive them in the name of Geodesic coordinate system or simply Gauss Lemma. For better geometric understanding we do it in polar coordinate patch first and then *uv*-coordinate patch on the surface S.

We saw in Part I that the intrinsic geometry of a surface gives rise to the notion of arc length of a curve on the surface (cf. formula (7) of §3) and this arc length is expressible in terms of the first fundamental form. Then we can look for curves on the surface which have the property that they minimize the arc length between any two points on the curve. These special curves exist on S and are called geodesics. For example on a sphere these are the great circles through north pole. In fact, if we view the sphere from above through the north pole, the geodesics i.e. circles of longitude look like straight lines and these straight lines are perpendicular to the circles of latitude. This process is projecting of the north pole to the centre. This is an illustration of Gauss Lemma (cf. lemma 44 below). It should be mentioned this result was invented by great Islamic scientist Ahmad-al-Biruni of Uzbekistan around 1010.

Some Historical remarks on inertial frames:

Galileo's law of inertia says that particles subject to no forces move along "straightlines" with constant velocity. For this to make sense we need to know that there is a family of "inertial frames" or "non-moving frames" relative to which the notion of a "straight line" is meaningful. That is, there must be a concept of "absolute space".

Newton used Theology to solve the problem of absolute space. According to Newton, absolute space is "God's Sensorium" and "attributes of God", possibly meaning by Newton that it is a neighborhood of infinity and projections from that point at infinity.

Finally Einstein solved the problem of absolute space. According to Einstein, it is the geometry of space-time that determines the motion of particles subject to gravitational forces. The distribution of matter-energy determines the geometry of space-time and ponderbale matter moves along geodesics. In an Inertial coordinate system geodesics look like straight lines. (cf. Corollary 47 below).

Gauss Lemma:

Let P be a point on the surface S in $\mathbb{R}^3$ and let v be a unit tangent vector to S at P. Let $\gamma^\theta(r)$ be the unit speed geodesic on S passing through P when $r = 0$ and with tangent vector v at P. Let $\sigma(r, \theta) = \gamma^\theta(r)$ be a smooth surface patch on S defined for $0 < r < \epsilon$ and $\theta \in [\alpha, 2\pi + \alpha]$, $\epsilon > 0$, $\alpha \in \mathbb{R}$. σ is called a geodesic polar coordinate patch of S.

Lemma 3.1. *If $0 < R < \epsilon$ then*

$$\int_0^R \left\| \frac{d\gamma^\theta}{dr} \right\|^2 dr = R\,. \tag{3.11}$$

Differentiating (11) w.r.t θ gives

$$\sigma_r \cdot \sigma_\theta = 0 \quad \textit{(Gauss formula)} \tag{3.12}$$

Geometrically (12) means that the parameter curve $r = R$ called the geodesic circle with centre P and radius R is perpendicular to the radial lines from the centre P that is, to the geodesics passing through P and also that the first fundamental form of σ is $dr^2 + G(r,\theta)d\theta^2$ for some smooth function $G(r, \theta)$.

Proof. Since $\gamma^\theta(r)$ is unit speed,

$$\sigma_r \cdot \sigma_r = 1, \quad \text{so} \quad \int_0^R \sigma_r \cdot \sigma_r dr = R \tag{3.13}$$

Differentiating (13) w.r.t θ gives

$$\int_0^R \sigma_r \cdot \sigma_{r\theta} dr = 0 \tag{3.14}$$

and then integrating by parts we get

$$\sigma_\theta \cdot \sigma_r \Big|_{r=0}^{r=R} - \int_0^R \sigma_\theta \cdot \sigma_{rr} dr = 0 \tag{3.15}$$

Now $\sigma(0, \theta) = P$ for all θ in $[0, 2\pi]$ and so $\sigma_\theta = 0$ when $r = 0$. So we must show that the integral in (15) vanishes. But $\sigma_{rr} = \ddot{\gamma}^\theta$ where $\cdot$ denotes the derivative w.r.t. r of the geodesic γ^θ and so σ_{rr} is parallel to the unit normal $\hat{N}$ of σ. Since $\sigma_\theta \cdot \hat{N} = 0$, it follows that $\sigma_\theta \cdot \sigma_{rr} = 0$.
Hence from (15) we get $\sigma_r \cdot \sigma_\theta = 0$ when $r = R$ which proves (12).
The first fundamental form, since $\sigma_r \cdot \sigma_\theta = 0$ and $\sigma_r \cdot \sigma_r = 1$ $\sigma_\theta \cdot \sigma_\theta$ is arbitrary, is of the form.
$\pi_1(r, \theta) = dr^2 + G(r, \theta)d\theta^2$, for some smooth ($E = 1, F = 0, G = \sigma_\theta \cdot \sigma_\theta$) function on S. □

Construction of general Geodesic coordinates on the surface S: The existence of geodesics on a surface S allows us to construct a very special coordinate system for S as follows.
Let P be a point on S and γ with parameter v be a unit speed geodesic on S with $\gamma(0) = P$. For any value of v, there is a unique geodesic of unit speed $\tilde{\gamma}^v$ with parameter u such that $\tilde{\gamma}^v(0) = \gamma(v)$ and s.t. $\tilde{\gamma}^v$ is perpendicular to γ at $\gamma(v)$. We define

$$\sigma(u, v) = \tilde{\gamma}^v(u) \tag{3.16}$$

Thus we get a σ-patch on S.

Proposition 3.1. *With notation as in above paragraph, there exists an open subset* $U \subset \mathbb{R}^2_{uv}$ *containing* $(0,0)$ *such that* $\sigma : U \to \mathbb{R}^3$ *is a smooth surface patch for* S. *Moreover the first fundamental form of* σ *is of the form* $du^2 + G(u,v)dv^2$ *where* G *is a smooth function on* U *such that* $G(0,v) = 1$ *and* $G_u(0,v) = 0$ *whenever* $(0,v) \in U$.

Proof. By inverse function theorem there exists a suitable open set U in $\mathbb{R}^2_{uv}$ on which σ is a coordinate patch for surface S.
Note that for any value of v,

$$\sigma_u(0,v) = \frac{d}{du}\tilde{\gamma}^v(u)\Big|_{u=0} \text{ and } \sigma_v(0.v) = \frac{d}{dv}\tilde{\gamma}^v(0) = \frac{d}{dv}\gamma(v) \tag{3.17}$$

and these are perpendicular unit vectors by construction of atlas on S given in above paragraph (cf. (16)). If $\sigma(u.v)$ is parametrically represented of S in $\mathbb{R}^3$ by $(f(u,v), g(u,v), h(u,v))$ then its Jacobian matrix $((f_u, f_v), (g_u, g_v), (h_u, h_v))^T$, as column matrix of order 3×2 has rank 2 at origin $(0,0) \in U$ say $\begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix}$ at $(0,0)$ is invertible matrix.(18).
Hence by Inverse function theorem there exists an open subset U of $\mathbb{R}^2$ such that the map given by $F(u,v) = (f(u,v), g(u,v))$ is a bijection: $U \to F(U) \subset \mathbb{R}^2$ and its inverse: $F(U) \to U$ is also smooth. By smoothness the matrix in (18) is invertible at each point (u,v) of U which is true at $(0,0)$ first of U. So σ_u and σ_v are linearly independent at (u,v) in U. Hence $\sigma : U \to \mathbb{R}^3$ is a genuine coordinate patch proving the first part of the proposition.

For the first fundamental form π_1 of σ, note that $E = \sigma_u \cdot \sigma_u = \left\| \frac{d}{du}\tilde{\gamma}^v(u) \right\|^2 = 1$ as $\tilde{\gamma}^v$ is of unit speed.
Recall the characterization of geodesics on S of $\mathbb{R}^3$:

Result (19): A curve γ on a surface S in $\mathbb{R}^3$ is a geodesic iff for any part $\gamma(t) = \sigma(u(t), v(t))$ of γ contained in a surface patch σ of S satisfies the geodesic equations:
(a) $\frac{d}{dt}(E\dot{u} + F\dot{v}) = \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2)$
(b) $\frac{d}{dt}(F\dot{u} + G\dot{v}) = \frac{1}{2}(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2)$ where $\pi_1 = Edu^2 + 2Fdudv + Gdv^2$ of S.

Now we apply geodesic equation (b) to the curve $\tilde{\gamma}^v$ and here the unit speed parameter is u and v is constant. We get $F_u = 0$. But when $u = 0$, we already have seen σ_u and σ_v are perpendicular, so $F = 0$ and hence by smoothness of F, $F = 0$ everywhere. Hence the first fundamental form of σ is of the form

$$du^2 + G(u,v)dv^2 \text{ as } E = 1 \text{ and } F = 0\,. \tag{3.20}$$

we have

$$G(0,v) = \|\sigma_v(0,v)\|^2 = \left\| \frac{d}{dv}\gamma \right\|^2 = 1 \tag{3.21}$$

as γ is of unit speed.
Finally applying geodesic equation (a) to the geodesic γ for which $u = 0$ and v is unit speed parameter, we get $G_u(0,v) = 0$ □

Corollary 3.1. *By inter changing the roles of parameters u and v in above proposition with curves $\gamma(u)$ and $\tilde{\gamma}^u(v)$ there exists a coordinate patch $\tilde{\sigma}$ on S whose first fundamental form $\tilde{\pi}_1$ is of the form $E(u,v)du^2 + dv^2$ with $E(u,0) = 1$ and $E_v(u,0) = 0$ as $F_v = 0$ and hence $F = 0$.*

Corollary 3.2. *By simultaneously doing these two processes we get a coordinate system σ on S whose first fundamental form π_1 is of the form $du^2 + dv^2$ with $E = 1, F = 0, G = 1$ on U and so*

$$Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ at } (0,0) \text{ and } Q_u = \begin{pmatrix} E_u & F_u \\ F_u & G_u \end{pmatrix}$$

and so $Q_u(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
Also $Q_v = \begin{pmatrix} E_v & F_v \\ F_v & G_v \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ *at* $(0,0)$.
Thus we proved.

Theorem 3.5. *Let S be a surface in $\mathbb{R}^3$. Then there exists an inertial coordinate system (u,v) with a coordinate patch σ on S whose fundamental form π_1 is of the form $du^2 + dv^2$ and its first fundamental matrix $Q = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$ satisfies $Q(0) = I$, $\frac{\partial Q}{\partial u}(0,0) = 0$ and $\frac{\partial Q}{\partial v}(0,0) = 0$ as matrices (cf. formula (10) of part-I).*
We used this theorem in the proof of Theorem 15 in part-I. We conclude this appendix by giving a direct proof of Theorem 48 using properties of Jacobian matrices.
Theorem 48(a): *On a surface S in $\mathbb{R}^3$ there exist locally at a point P on S an inertial coordinate system $\{u,v\}$ such that its first fundamental matrix Q_σ of a surface patch σ satisfies*

$$Q_\sigma(0) = I, \ \frac{\partial Q}{\partial u}(0) = 0 \text{ and } \frac{\partial Q}{\partial v}(0) = 0$$

as matrices and with P as origin (0,0).

Proof. From §3.3 of Part-I we can find a coordinate system $\{u,v\}$ s.t. $Q(0,0) = I_2$. Now we look for a change of coordinates with $J(0,0) = I_2$ to $u(u',v')$ and $v(u',v')$ so that $Q'(0,0) = I_2$ in $u'v'$-coordinate system.
Now differentiating the relation

$$Q'(u',v') =^t J(u',v')Q(u,v)J(u',v') \text{ (formula(9) of part-I)}$$

w.r.t. u' and evaluating at $(0,0)$ and using $J(0,0) = I_2$, $Q(0,0) = I$ and $Q'(0,0) = I$ we get

$$\frac{\partial}{\partial u'}(J +^t J)(0,0) = -\frac{\partial Q}{\partial u}(0,0) \tag{3.22}$$

Similarly differentiating (9) above w.r.t v' and evaluating at (0,0) we get

$$\frac{\partial}{\partial v'}(J +^t J)(0,0) = -\frac{\partial Q}{\partial v}(0,0) \tag{3.23}$$

But computing the left hand sides of (22) and (23) we get the matrices

$$\begin{pmatrix} \dfrac{2\partial^2 u}{(\partial u')^2} & \dfrac{\partial^2 u}{\partial u' \partial v'} + \dfrac{\partial^2 v}{(\partial u')^2} \\ \dfrac{\partial^2 u}{\partial u' \partial v'} + \dfrac{\partial^2 v}{(\partial u')^2} & 2\dfrac{\partial^2 v^{(*)}}{\partial u' \partial v'} \end{pmatrix} \tag{3.24}$$

and

$$\begin{pmatrix} \dfrac{2\partial^2 u^{(+)}}{\partial u' \partial v'} & \dfrac{\partial^2 u}{(\partial v')^2} + \dfrac{\partial^2 v}{(\partial u')^2 \partial v'} \\ \dfrac{\partial^2 u}{(\partial v')^2} + \dfrac{\partial^2 v}{\partial u' \partial v'} & \dfrac{2\partial^2 v}{(\partial v')^2} \end{pmatrix} \tag{3.25}$$

evaluated at (0,0) respectively.

Note that these two matrices are symmetric matrices and having determined the (∗) term and (+) term in these two matrices all other 2nd order derivatives are also determined and because of $J(0,0) = I$ these matrices (24) and (25) reduce to 2×2 zero matrices and hence from (22) and (23) we get, $Q_u(0,0) = 0$, and $Q_v(0,0) = 0$ and hence there exists an inertial coordinate system $\{u, v\}$ around (0,0) in the surface S. □

Remark 3.5. *In Theorems 48 and 49(a) we gave two proofs for relation (10) of Part-I guaranteeing the existence of Inertial coordinate systems.*