Introduction to Information Geometry

– based on the book “Methods of Information Geometry ”written by Shun-Ichi Amari and Hiroshi Nagaoka

Yunshu Liu

2012-02-17

Outline

1. Introduction to differential geometry
   - Manifold and Submanifold
   - Tangent vector, Tangent space and Vector field
   - Riemannian metric and Affine connection
   - Flatness and autoparallel
2. Geometric structure of statistical models and statistical inference
   - The Fisher metric and α-connection
   - Exponential family
   - Divergence and Geometric statistical inference

Part I

Introduction to differential geometry

Basic concepts in differential geometry

Basic concepts

Manifold and Submanifold

Tangent vector, Tangent space and Vector field

Riemannian metric and Affine connection

Flatness and autoparallel

Manifold

Manifold S

Manifold: a set with a coordinate system, a one-to-one mapping from S to $\mathbb{R}^n$, supposed to be "locally" looks like an open subset of $\mathbb{R}^n$"

Elements of the set(points): points in $\mathbb{R}^n$, probability distribution, linear system.

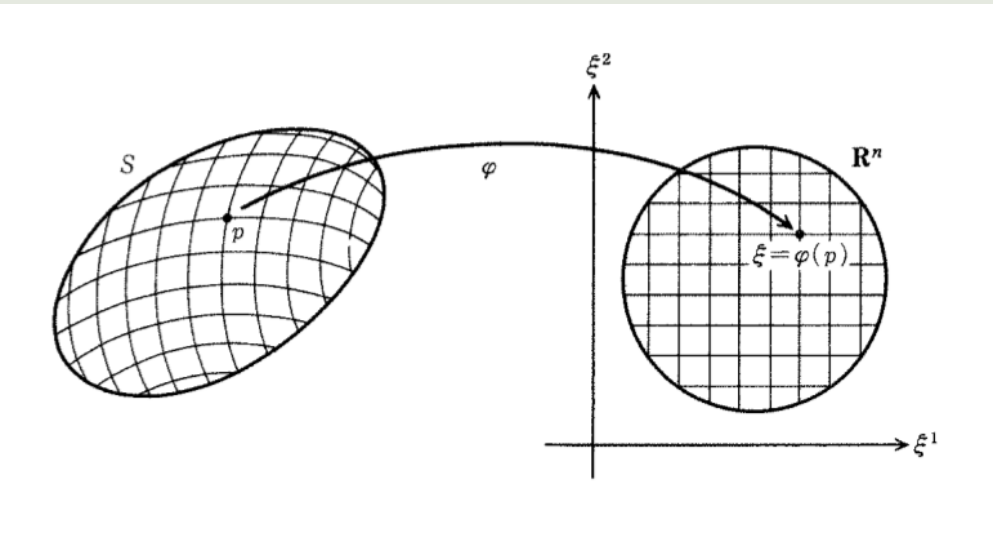


Figure : A coordinate system ξ for a manifold S

Manifold

Manifold S

Definition: Let S be a set, if there exists a set of coordinate systems $\mathcal{A}$ for S which satisfies the condition (1) and (2) below, we call S an n-dimensional C^∞ differentiable manifold.

- (1) Each element φ of $\mathcal{A}$ is a one-to-one mapping from S to some open subset of $\mathbb{R}^n$.
- (2) For all $\varphi \in \mathcal{A}$, given any one-to-one mapping ψ from S to R^n, the following hold:
  $\psi \in \mathcal{A} \Leftrightarrow \psi \circ \varphi^{-1}$ is a C^∞ diffeomorphism.

Here, by a C^∞ diffeomorphism we mean that $\psi \circ \varphi^{-1}$ and its inverse $\varphi \circ \psi^{-1}$ are both C^∞(infinitely many times differentiable).

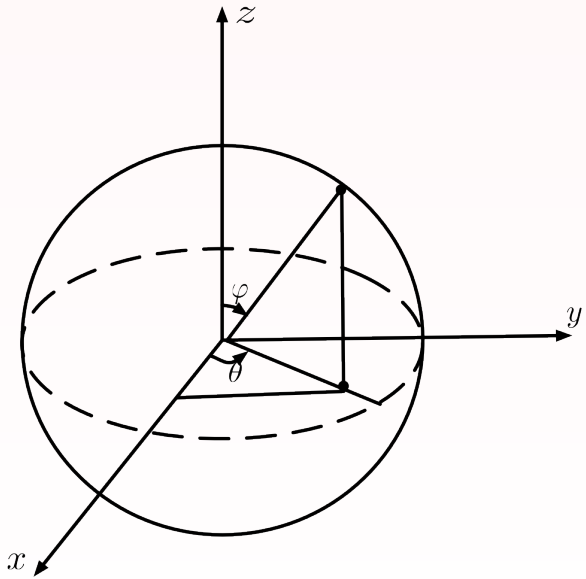
Examples of Manifold

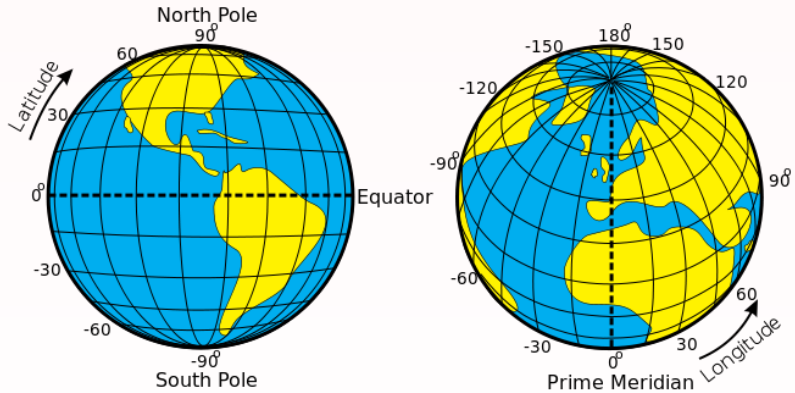
Examples of one-dimensional manifold

- A straight line: a manifold in $\mathbb{R}^1$, even if it is given in $\mathbb{R}^k$ for $\forall\, k \geqslant 2$.
  Any open subset of a straight line: one-dimensional manifold
  A closed subset of a straight line: not a manifold
- A circle: locally the circle looks like a line
  Any open subset of a circle: one-dimensional manifold
  A closed subset of a circle: not a manifold.

Examples of Manifold: surface of a sphere

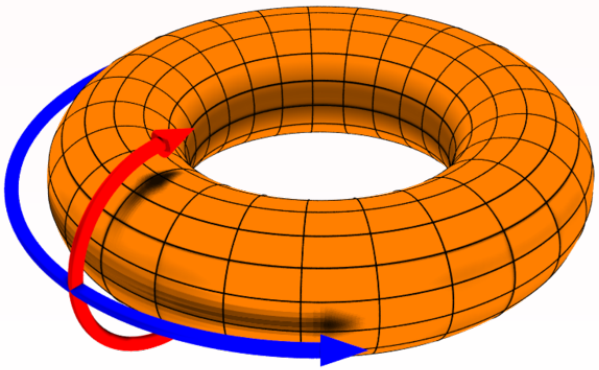
Surface of a sphere in $\mathbb{R}^3$, defined by $S = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 = 1\}$, locally it can be parameterized by using two coordinates, for example, we can use latitude and longitude as the coordinates.
nD sphere(n-1 sphere): $S = \{(x_1, x_2, ..., x_n) \in \mathbb{R}^n | x_1^2 + x_2^2 + ... + x_n^2 = 1\}$.

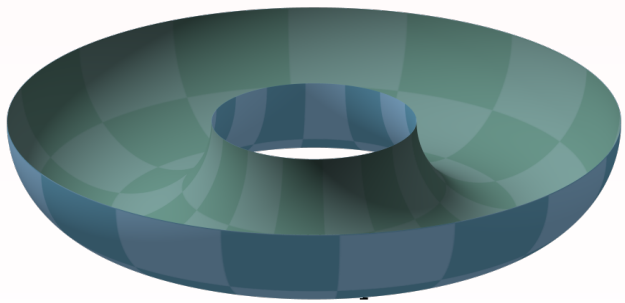




Examples of Manifold: surface of a torus

The torus in $\mathbb{R}^3$(surface of a doughnut):
$x(u,v) = ((a + b\cos u)cos\ v, (a + b\cos u)sin\ v, b\sin\ u),\ 0 \leqslant u, v < 2\pi.$
where a is the distance from the center of the tube to the center of the torus, and b is the radius of the tube. A torus is a closed surface defined as product of two circles: $T^2 = S^1 \times S^1$.
n-torus: T^n is defined as a product of n circles: $T^n = S^1 \times S^1 \times \cdots \times S^1$.



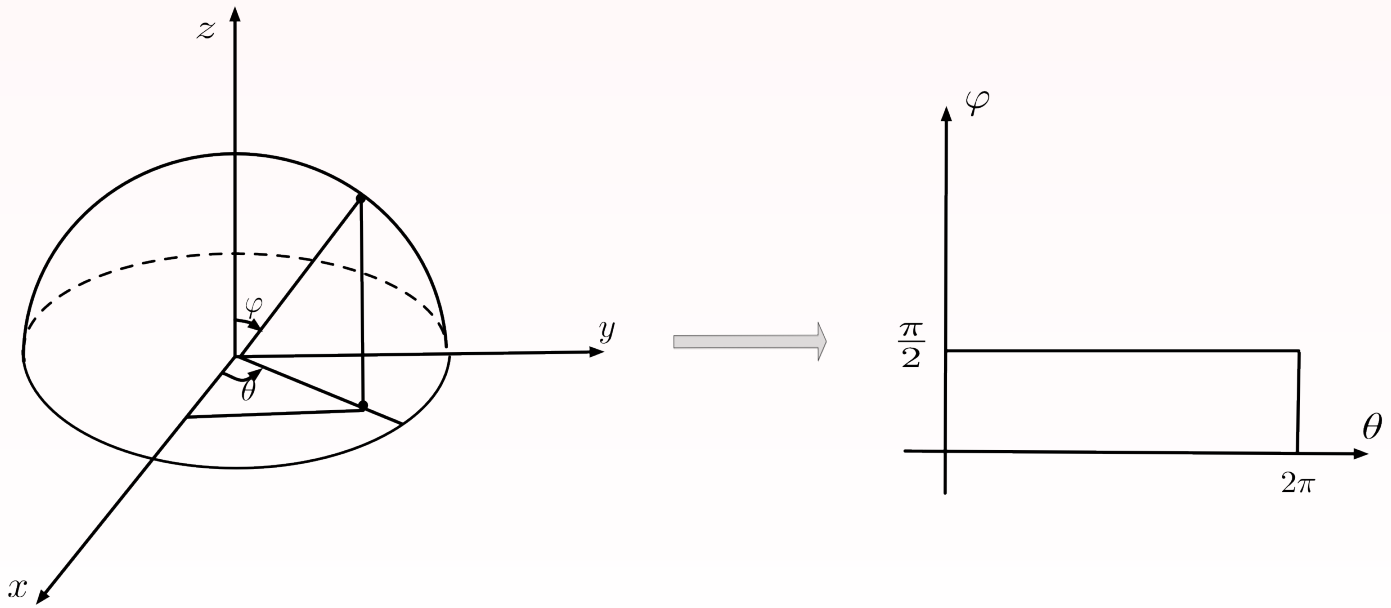


Coordinate systems for manifold

Parametrization of unit hemisphere

Parametrization: map of unit hemisphere into $\mathbb{R}^2$
(1) by latitude and longigude;

$$x(\varphi, \theta) = (sin\ \varphi\ cos\ \theta, sin\ \varphi\ sin\ \theta, cos\ \varphi), \quad 0 < \varphi < \pi/2, 0 \leqslant \theta < 2\pi \quad (1)$$

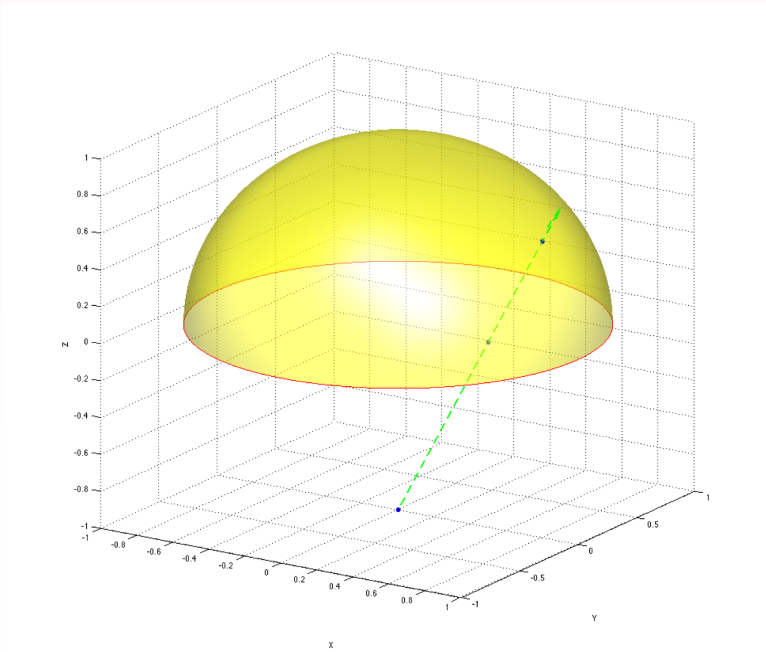


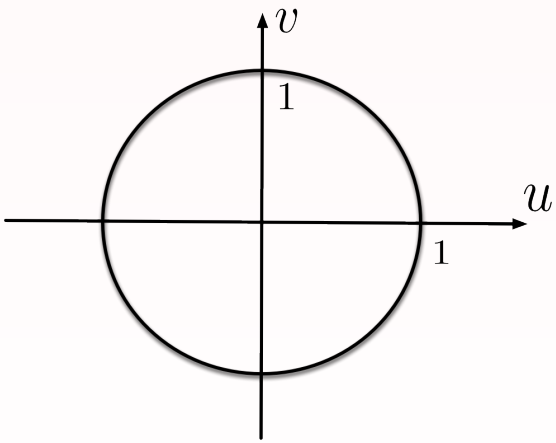
Coordinate systems for manifold

Parametrization of unit hemisphere

Parametrization: map of unit hemisphere into $\mathbb{R}^2$
(2) by stereographic projections.

$$x(u,v) = \left(\frac{2u}{1+u^2+v^2}, \frac{2v}{1+u^2+v^2}, \frac{1-u^2-v^2}{1+u^2+v^2}\right) \quad where\ u^2+v^2 \leqslant 1 \quad (2)$$



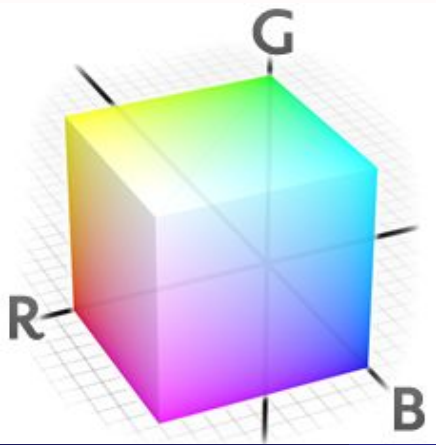


Examples of Manifold: colors

Parametrization of color models

3 channel color models: RGB, CMYK, LAB, HSV and so on.

- The RGB color model: an additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors.
- The Lab color model: three coordinates of Lab represent the lightness of the color(L), its position between red and green(a) and its position between yellow and blue(b).



Coordinate systems for manifold

Parametrization of color models

Parametrization: map of color into $\mathbb{R}^3$

Examples:

Submanifolds

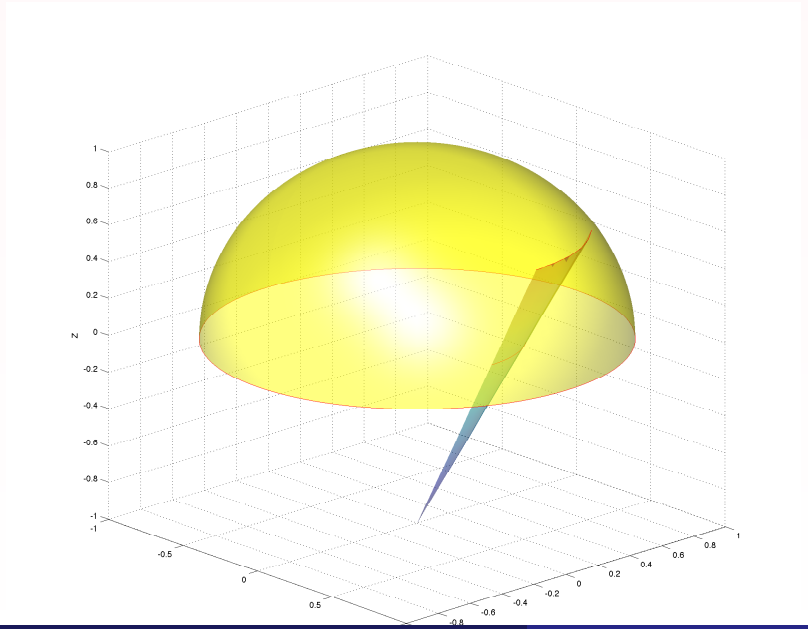
Submanifolds

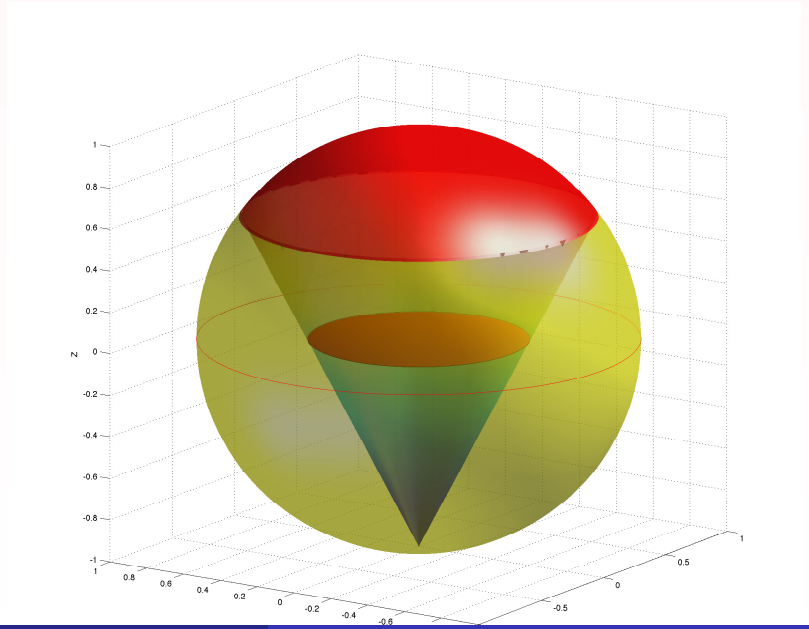
Definition: a submanifold M of a manifold S is a subset of S which itself has the structure of a manifold

An open subset of n-dimensional manifold forms an n-dimensional submanifold.

One way to construct m(<n)dimensional manifold: fix n-m coordinates.

Examples:

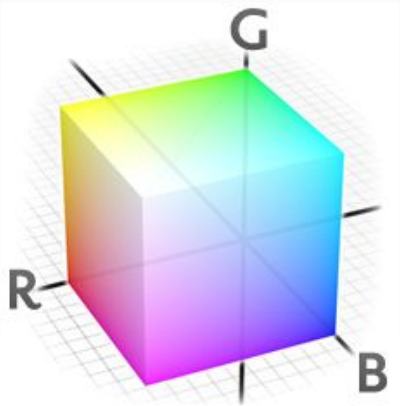




Submanifolds

Examples: color models

3-dimensional submanifold: any open subset;
2-dimensional submanifold: fix one coordinate;
1-dimensional submanifold: fix two coordinates.
Note: In Lab color model, we set a and b to 0 and change L from 0 to 100(from black to white), then we get a 1-dimensional submanifold.



Basic concepts in differential geometry

Basic concepts

Manifold and Submanifold

Tangent vector, Tangent space and Vector field

Riemannian metric and Affine connection

Flatness and autoparallel

Curves and Tangent vector of Curves

Curves

Curve γ: $I \to S$ from some interval $I(\subset \mathbb{R})$ to S.
Examples: curve on sphere, set of probability distribution, set of linear systems.
Using coordinate system $\{\xi^i\}$ to express the point $\gamma(t)$ on the curve(where t $\in$ I): $\gamma^i(t) = \xi^i(\gamma(t))$, then we get $\bar{\gamma}(t) = [\gamma^1(t), \cdots, \gamma^n(t)]$.

C^∞ Curves

C^∞: infinitely many times differentiable(sufficiently smooth).
If $\bar{\gamma}(t)$ is C^∞ for $t \in$ I, we call γ a C^∞ on manifold S.

Tangent vector of Curves

A tangent vector is a vector that is tangent to a curve or surface at a given point.
When S is an open subset of $\mathbb{R}^n$, the range of γ is contained within a single linear space, hence we consider the standard derivative:

$$\dot{\gamma}(a) = \lim_{h \to 0} \frac{\gamma(a+h) - \gamma(a)}{h} \tag{3}$$

In general, however, this is not true, ex: the range of γ in a color model
Thus we use a more general "derivative" instead:

$$\dot{\gamma}(a) = \sum_{i=1}^{n} \dot{\gamma}^i(a) (\frac{\partial}{\partial \xi^i})_p \tag{4}$$

where $\gamma^i(t) = \xi^i \circ \gamma(t)$, $\dot{\gamma}^i(a) = \frac{d}{dt}\gamma^i(t)|_{t=a}$ and $(\frac{\partial}{\partial \xi^i})_p$ is an operator which maps $f \to (\frac{\partial f}{\partial \xi^i})_p$ for given function $f : S \to R$.

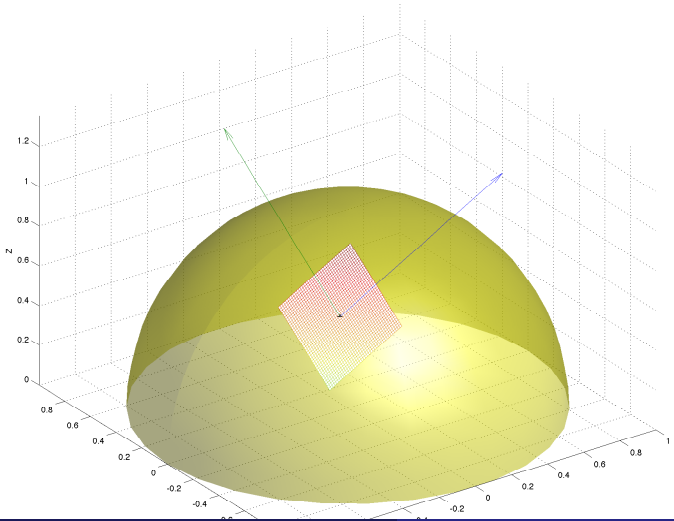
Tangent space

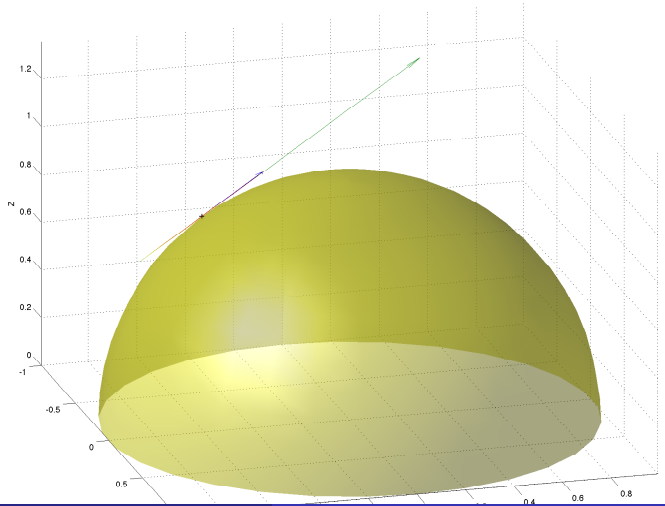
Tangent space

- Tangent space at p: a *hyperplane* T_p containing all the tangents of curves passing through the point $p \in S$. (dim $T_p(S)$ = dim S)

$$T_p(S) = \{\sum_{i=1}^{n} c^i (\frac{\partial}{\partial \xi i})_p | [c^1, \cdots, c^n] \in \mathbb{R}^n\}$$

Examples: hemisphere and color

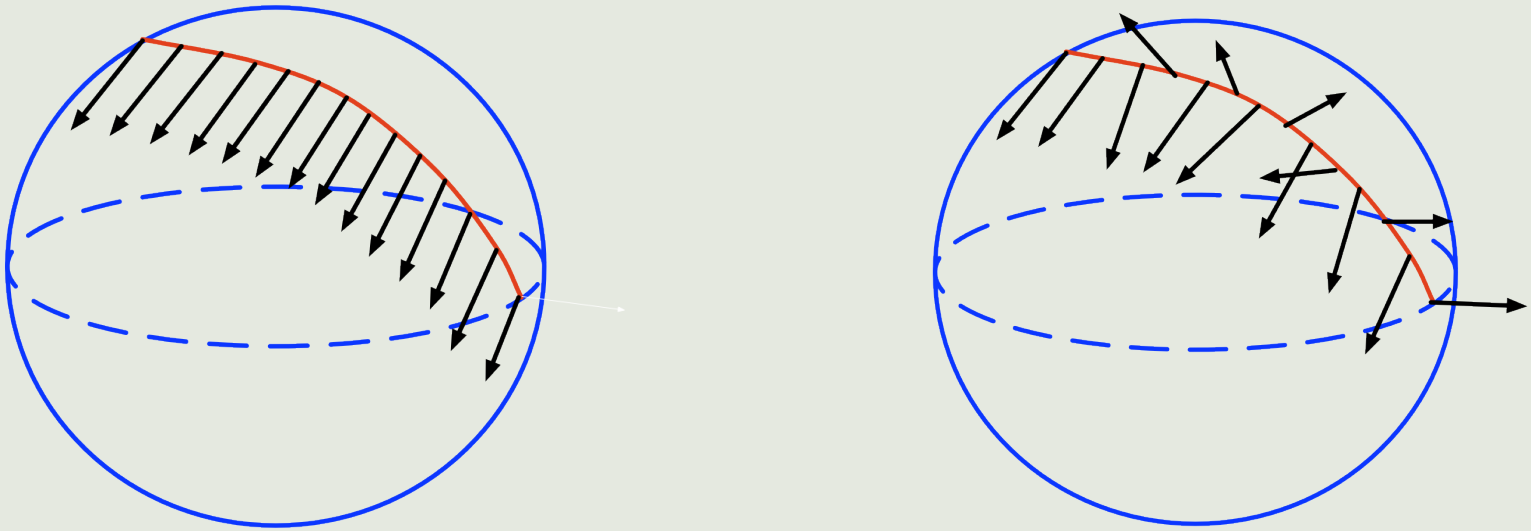




Vector fields

Vector fields

Vector fields: a map from each point in a manifold S to a tangent vector.
Consider a coordinate system $\{\xi_i\}$ for a n-dimensional manifold, clearly $\partial_i = \frac{\partial}{\partial \xi^i}$ are vector fields for $i = 1, \cdots, n$.



Basic concepts in differential geometry

Basic concepts

Manifold and Submanifold

Tangent vector, Tangent space and Vector field

Riemannian metric and Affine connection

Flatness and autoparallel

Riemannian Metrics

Riemannian Metrics: an inner product of two tangent vectors(D and $D^{'} \in T_p(S)$) which satisfy $\langle D, D^{'} \rangle_p \in \mathbb{R}$, and the following condition hold:

$$Linearity: \ \langle aD + bD^{'}, D^{''} \rangle_p = a\langle D, D^{''} \rangle_p + b\langle D^{'}, D^{''} \rangle_p$$
$$Symmetry: \ \langle D, D^{'} \rangle_p = \langle D^{'}, D \rangle_p$$
$$Positive - definiteness: \ If \ D \neq 0 \ then \ \langle D, D \rangle_p > 0$$

The components $\{g_{ij}\}$ of a Riemannian metric g $w.r.t.$ the coordinate system $\{\xi_i\}$ are defined by $g_{ij} = \langle \partial_i, \partial_j \rangle$, where $\partial_i = \frac{\partial}{\partial \xi_i}$.

Riemannian Metrics

Examples of inner product:

- For $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n)$, we can define inner product as $\langle X, Y \rangle_1 = X \cdot Y = \sum_{i=1}^{n} x_i y_i$, or $\langle X, Y \rangle_2 = YMX$, where M is any symmetry positive-definite matrix.
- For random variables X and Y, the expected value of their product: $\langle X, Y \rangle = E(XY)$
- For square real matrix, $\langle A, B \rangle = tr(AB^T)$

Riemannian Metrics

For unit sphere:

$$x(\varphi,\theta) = (sin\ \varphi\ cos\ \theta, sin\ \varphi\ sin\ \theta, cos\ \varphi), \quad 0 < \varphi < \pi, 0 \leqslant \theta < 2\pi \tag{5}$$

we have:

$$\begin{aligned}
\partial_\varphi &= (cos\ \varphi\ cos\ \theta, cos\ \varphi\ sin\ \theta, -sin\ \varphi) \\
\partial_\theta &= (-sin\ \varphi\ sin\ \theta, sin\ \varphi\ cos\ \theta, 0) \\
g_{11} &= \langle \partial_\varphi, \partial_\varphi \rangle, g_{22} = \langle \partial_\theta, \partial_\theta \rangle \\
g_{12} &= g_{21} = \langle \partial_\varphi, \partial_\theta \rangle = \langle \partial_\theta, \partial_\varphi \rangle
\end{aligned}$$

If we define $\langle X, Y \rangle = YMX = 2\sum_{i=1}^{n} x_i y_i$, where $M = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, then

$$(g_{i,j}) = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2sin^2\varphi \end{pmatrix} \tag{6}$$

Affine connection

Parallel translation along curves

Let γ: $[a, b] \to S$ be a curve in S, $X(t)$ be a vector field mapping each point $\gamma(t)$ to a tangent vector, if for all $t \in [a, b]$ and the corresponding infinitesimal dt, the corresponding tangent vectors are linearly related, that is to say there exist a linear mapping $\Pi_{p,p'}$, such that $X(t + dt) = \Pi_{p,p'}\ (X(t))$ for $t \in [a, b]$, we say X is parallel along γ, and call Π_γ the parallel translation along γ.
Linear mapping: additivity and scalar multiplication.

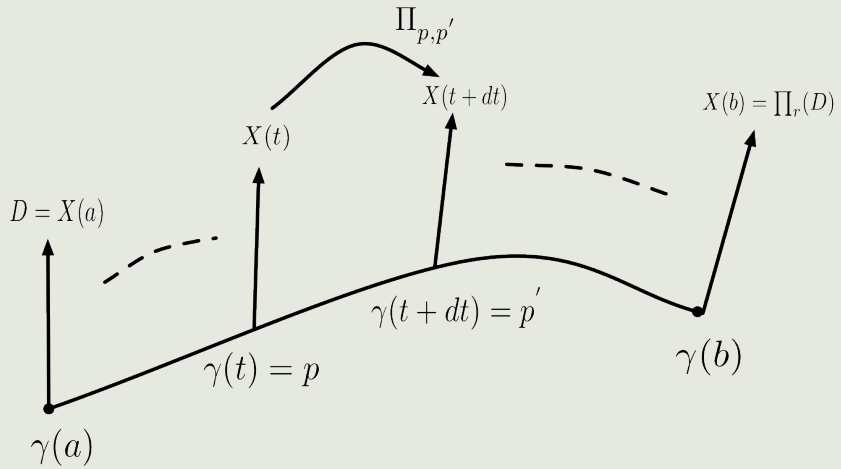


Figure : Translation of a tangent vector along a curve

Affine connection

Affine connection: relationships between tangent space at different points.
Recall:

- Natural basis of the coordinate system $[\xi^i]$: $(\partial_i)_p = (\frac{\partial}{\partial \xi^i})_p$: an operator which maps $f \to (\frac{\partial f}{\partial \xi^i})_p$ for given function $f : S \to R$ at p.
- Tangent space:

$$T_p(S) = \{\sum_{i=1}^{n} c^i (\frac{\partial}{\partial \xi^i})_p | [c^1, \cdots, c^n] \in \mathbb{R}^n\}$$

Tangent vector(elements in Tangent space) can be represented as linear combinations of ∂_i.

$$\text{Tangent space } T_p \to \text{Tangent vector } X_p \to \text{Natural basis } (\partial_i)_p = (\frac{\partial}{\partial \xi^i})_p$$

Affine connection

If the difference between the coordinates of p and $p^{'}$ are very small, that we can ignore the second-order infinitesimals $(d\xi^i)(d\xi^j)$, where $d\xi^i = \xi^i(p^{'}) - \xi^i(p)$, then we can express difference between $\Pi_{p,p'}((\partial_j)_p)$ and $((\partial_j)_{p'})$ as a linear combination of $\{d\xi^1, \cdots, d\xi^n\}$:

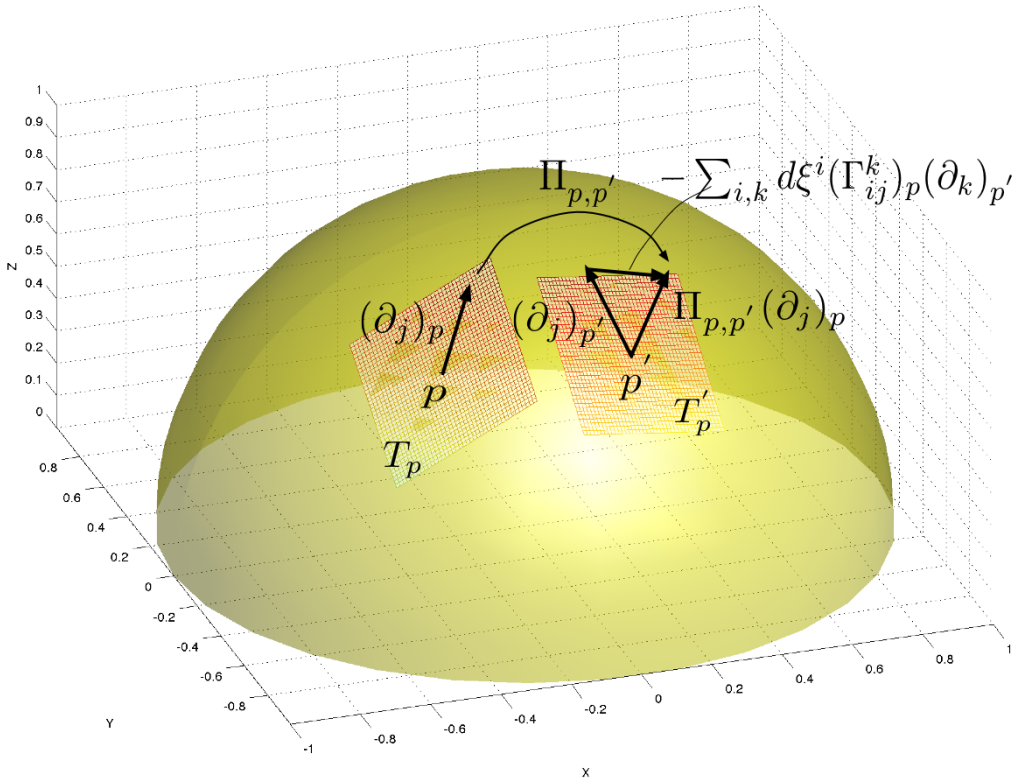
$$\Pi_{p,p'}((\partial_j)_p) = (\partial_j)_{p'} - \sum_{i,k}(d\xi^i(\Gamma_{ij}^k)_p(\partial_k)_{p'}) \tag{7}$$

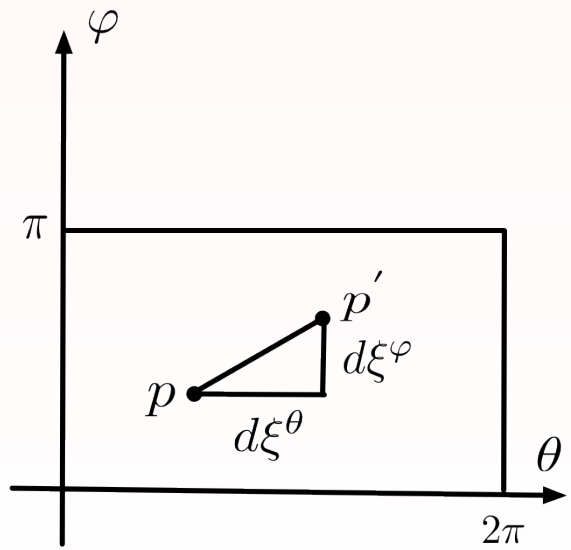
where $\{(\Gamma_{ij}^k)_p; i,j,k = 1, \cdots, n\}$ are n^3 numbers which depend on the point p. From $X(t) = \sum_{i=1}^n X^i(t)(\partial_i)_p$ and $X(t+dt) = \sum_{i=1}^n (X^i(t+dt)(\partial_i)_{p'})$, we have

$$\Pi_{p,p'}(X(t)) = \sum_{i,j,k}(\{X^k(t) - dt\dot{\gamma}^i(t)X^j(t)(\Gamma_{ij}^k)_p\}(\partial_k)_{p'}) \tag{8}$$

Affine connection

$$\Pi_{p,p'}((\partial_j)_p) = (\partial_j)_{p'} - \sum_{i,k}(d\xi^i(\Gamma^k_{ij})_p(\partial_k)_{p'})$$





Connection coefficients(Christoffel's symbols): $(\Gamma_{ij}^k)_p$

Given a connection on the manifold S, the value of $(\Gamma_{ij}^k)_p$ are different for different coordinate systems, it shows how tangent vectors changes on a manifold, thus shows how basis vectors changes.
In

$$\Pi_{p,p'}((\partial_j)_p) = (\partial_j)_{p'} - \sum_{i,k}(d\xi^i(\Gamma_{ij}^k)_p(\partial_k)_{p'})$$

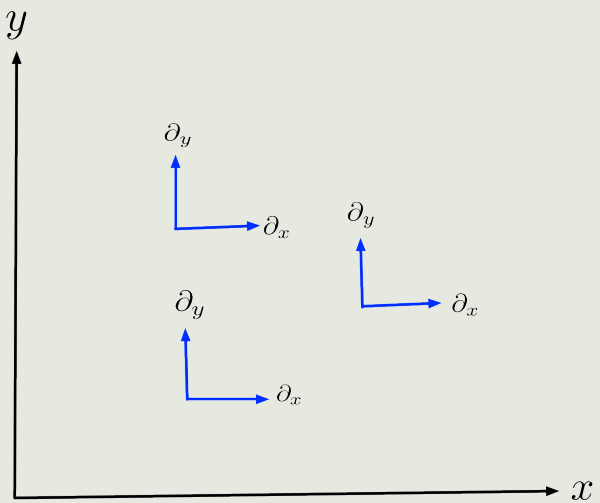
if we let $\Gamma_{ij}^k = 0$ for $i,j,k = x,y$, we will have

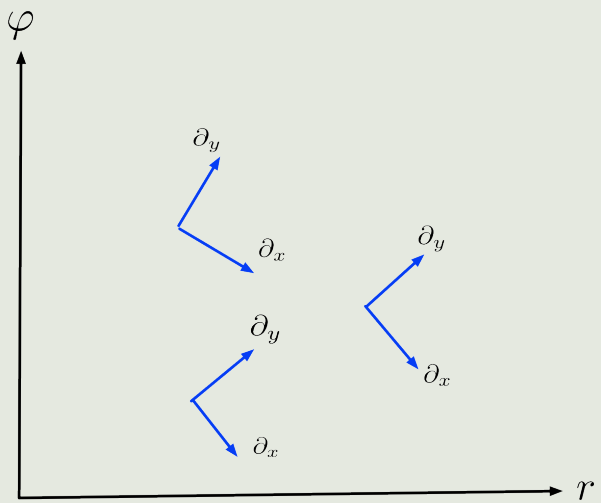
$$\Pi_{p,p'}((\partial_j)_p) = (\partial_j)_{p'}$$

Connection coefficients(Christoffel's symbols): $(\Gamma_{ij}^{k})_p$

Given a connection on a manifold S, $\Gamma_{i,j}^{k}$ depend on coordinate system. Define a connection which makes $\Gamma_{i,j}^{k}$ to be zero in one coordinate system, we will get non-zero connection coefficients in some other coordinate systems.

Example: If it is desired to let the connection coefficients for Cartesian Coordinates of a 2D flat plane to be zero, $\Gamma_{ij}^{k} = 0$ for $i, j, k = x, y$, we can calculate the connection coefficients for Polar Coordinates: $\Gamma_{r\varphi}^{\varphi} = \Gamma_{\varphi r}^{\varphi} = \frac{1}{r}$, $\Gamma_{\varphi\varphi}^{r} = -r$, and $\Gamma_{ij}^{k} = 0$ for all others.

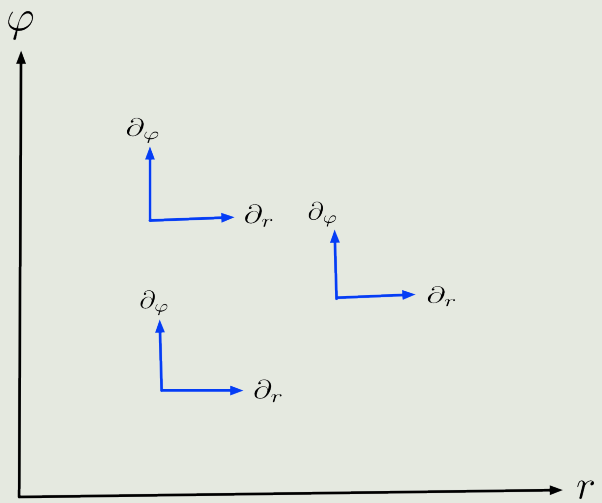


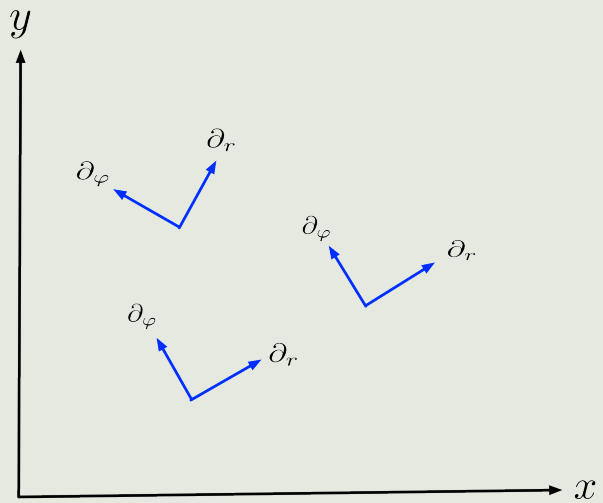


Connection coefficients(Christoffel's symbols): $(\Gamma_{ij}^k)_p$

Example(cont.):

Now if we want to let the connection coefficients for Polar Coordinates to be zero, $\Gamma_{ij}^k = 0$ for $i,j,k = r,\varphi$, we can calculate the connection coefficients for Polar Coordinates: $\Gamma_{xx}^x = \frac{-\sin^2\varphi\cos\varphi}{r}$, $\Gamma_{xx}^y = \frac{\sin\varphi(1+\cos^2\varphi)}{r}$, $\Gamma_{xy}^x = \Gamma_{yx}^x = \frac{-\sin^3\varphi}{r}$, $\Gamma_{xy}^y = \Gamma_{yx}^y = \frac{-\cos^3\varphi}{r}$, $\Gamma_{yy}^x = \frac{\cos\varphi(1+\sin^2\varphi)}{r}$, and $\Gamma_{yy}^y = \frac{-\sin\varphi\cos^2\varphi}{r}$.



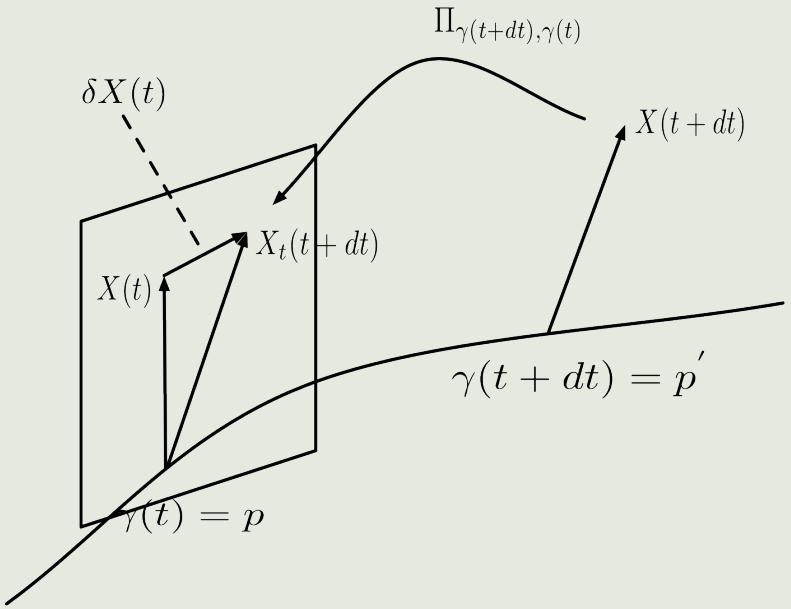


Affine connection

Covariant derivative along curves

Derivative: $\frac{dX(t)}{dt} = \lim_{dt\to 0} \frac{X(t+dt)-X(t)}{dt}$, what if $X(t)$ and $X(t+dt)$ lie in different tangent spaces? $X_t(t+dt) = \Pi_{\gamma(t+dt),\gamma(t)}(X(t+dt))$

$$\delta X(t) = X_t(t+dt) - X(t) = \Pi_{\gamma(t+dt),\gamma(t)}(X(t+dt)) - X(t)$$



Affine connection

Covariant derivative along curves

We call $\frac{\delta X(t)}{dt}$ the covariant derivative of $X(t)$:

$$\frac{\delta X(t)}{dt} = \lim_{dt\to 0} \frac{X_t(t+dt) - X(t)}{dt} = \frac{\Pi_{\gamma(t+dt),\gamma(t)}(X(t+dt)) - X(t)}{dt} \tag{9}$$

$$\Pi_{\gamma(t+dt)}(X(t+dt)) = \sum_{i,j,k}(\{X^k(t+dt) + dt\dot{\gamma}^i(t)X^j(t)(\Gamma^k_{ij})_{\gamma(t)}\}(\partial_k)_{\gamma(t)}) \tag{10}$$

$$\frac{\delta X(t)}{dt} = \sum_{i,j,k}(\{\dot{X}^k(t) + \dot{\gamma}^i(t)X^j(t)(\Gamma^k_{ij})_{\gamma(t)}\}(\partial_k)_{\gamma(t)}) \tag{11}$$

Affine connection

Covariant derivative of any two tangent vector

Covariant derivative of Y w.r.t. X, where $X = \sum_{i=1}^{n}(X^i\partial_i)$ and $Y = \sum_{i=1}^{n}(Y^i\partial_i)$:

$$\nabla_X Y = \sum_{i,j,k}(X^i\{\partial_i Y^k + Y^j\Gamma_{ij}^k\}\partial_k) \tag{12}$$

$$\nabla_{\partial_i}\partial_j = \sum_{k=1}^{n}\Gamma_{ij}^k\partial_k \tag{13}$$

Note: $(\nabla_X Y)_p = \nabla_{X_p} Y \in T_p(S)$

Examples of Affine connection

metric connection

Definition: If for all vector fields $X, Y, Z \in \mathcal{T}(S)$,

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle.$$

where $Z\langle X, Y\rangle$ denotes the derivative of the function $\langle X, Y\rangle$ along this vector field Z, we say that ∇ is a metric connection w.r.t. g.
Equivalent condition: for all basis $\partial_i, \partial_j, \partial_k \in \mathcal{T}(S)$,

$$\partial_k\langle \partial_i, \partial_j\rangle = \langle \nabla_{\partial_k}\partial_i, \partial_j\rangle + \langle \partial_i, \nabla_{\partial_k}\partial_i\rangle.$$

Property: parallel translation on a metric connection preserves inner products, which means parallel transport is an isometry.

$$\langle \Pi_\gamma(D_1), \Pi_\gamma(D_2)\rangle_q = \langle D_1, D_2\rangle_p.$$

Examples of Affine connection

Levi-Civita connection

- For a given connection, when $\Gamma_{ij}^k = \Gamma_{ji}^k$ hold for all i, j and k, we call it a symmetric connection or torsion-free connection.
  From $\nabla_{\partial_i}\partial_j = \sum_{k=1}^n \Gamma_{ij}^k \partial_k$, we know for a symmetric connection:
  $\nabla_{\partial_i}\partial_j = \nabla_{\partial_j}\partial_i$
- If a connection is both metric and symmetric, we call it the Riemannian connection or the Levi-Civita connection w.r.t. g.

Basic concepts in differential geometry

Basic concepts

Manifold and Submanifold

Tangent vector, Tangent space and Vector field

Riemannian metric and Affine connection

Flatness and autoparallel

Flatness

Affine coordinate system

Let $\{\xi^i\}$ be a coordinate system for S, we call $\{\xi^i\}$ an affine coordinate system for the connection ∇ if the n basis vector fields $\partial_i = \frac{\partial}{\partial \xi^i}$ are all parallel on S. Equivalent conditions for a coordinate system to be an affine coordinate system:

$$\nabla_{\partial_i}\partial_j = \sum_{k=1}^{n}(\Gamma_{ij}^k \partial_k) = 0 \quad \textit{for all } i \textit{ and } j \tag{14}$$

$$\Gamma_{ij}^k = 0 \quad \textit{for all } i, j \textit{ and } k \tag{15}$$

Flatness

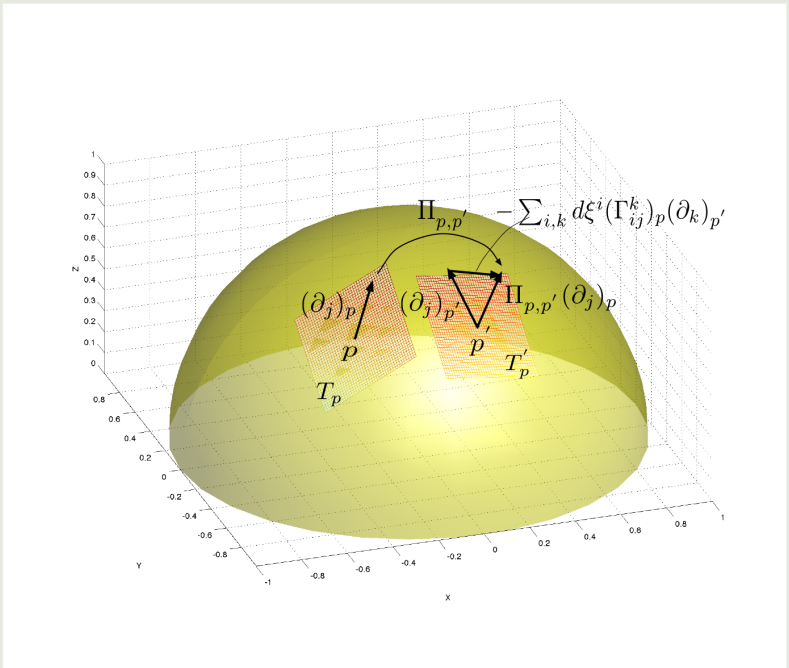
S is flat w.r.t the connection ∇: an affine coordinate system exist for the connection ∇.

Flatness

Examples:

$$\nabla_{\partial_i}\partial_j = \sum_{k=1}^{n}(\Gamma_{ij}^{k}\partial_k) = 0 \quad \textit{for all } i \textit{ and } j$$

$$\Gamma_{ij}^{k} = 0 \quad \textit{for all } i, j \textit{ and } k$$



Flatness

Curvature R and torsion T of a connection

$$R(\partial_i, \partial_j)\partial_k = \sum_l (R^l_{ijk}\partial_l) \text{ and } T(\partial_i, \partial_j) = \sum_k (T^k_{ij}\partial_k) \tag{16}$$

where R^l_{ijk} and T^k_{ij} can be computed in the following way:

$$R^l_{ijk} = \partial_i \Gamma^l_{jk} - \partial_j \Gamma^l_{ik} + \Gamma^l_{ih}\Gamma^h_{jk} - \Gamma^l_{jh}\Gamma^h_{ik} \tag{17}$$

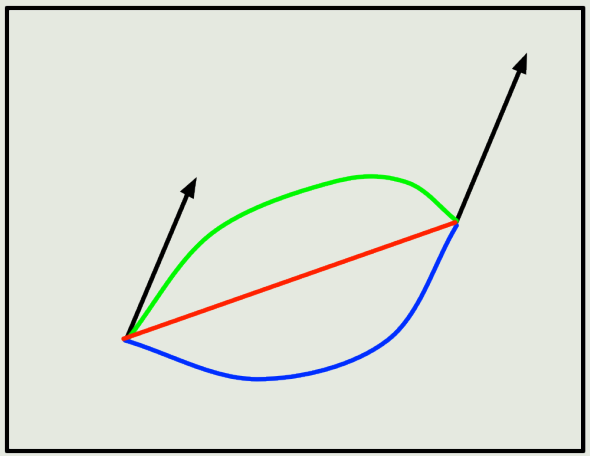
$$T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji} \tag{18}$$

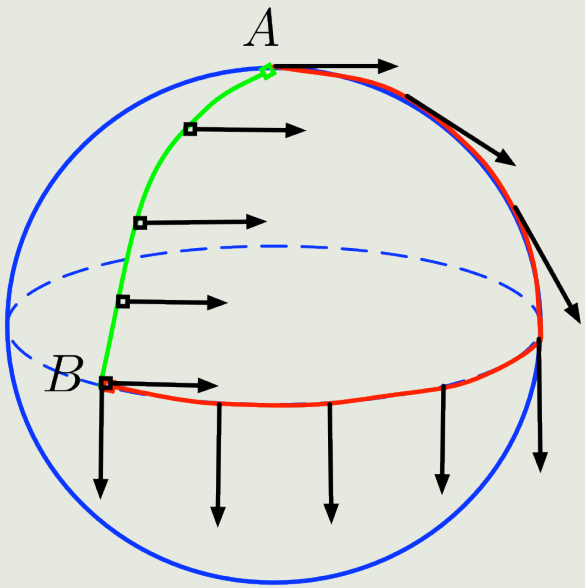
If a connection is flat, then T=R=0;
If T= 0, $\Gamma^k_{ij} = 0$ *for all* i, j *and* k, we get the symmetry connnection.

Flatness

Curvature

Curvature $R = 0$ iff parallel translation does not depend on curve choice.
Curvature is independent of coordinate system, under Riemannian connection, we can calculate:
Curvature of 2 dimensional plane: $R = 0$;
Curvature of 3 dimensional sphere: $R = \frac{2}{r^2}$.





Autoparallel submanifold

Equivalent condition for a submanifold M of S to be autoparallel

$$\nabla_X Y \in \mathcal{T}(M) \text{ for } \forall X, Y \in \mathcal{T}(M) \tag{19}$$

$$\nabla_{\partial_a} \partial_b \in \mathcal{T}(M) \text{ for all } a \text{ and } b \tag{20}$$

$$\nabla_{\partial_a} \partial_b = \sum_c (\Gamma_{ab}^c \partial_c) \tag{21}$$

where $\partial_a = \frac{\partial}{\partial u^a}$ and $\partial_b = \frac{\partial}{\partial u^b}$ are the basis for submanifold M w.r.t. coordinate system $\{u^i\}$.

Examples of autoparallel submanifold:
Open subsets of manifold S are autoparallel;
A curve with the properity that all the tangent vector are parallel

Autoparallel submanifold

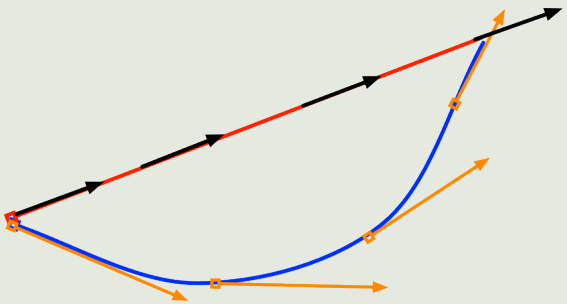
Geodesics

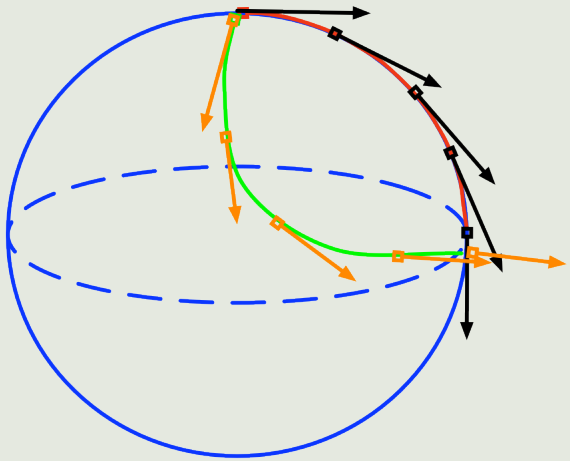
Geodesics(autoparallel curves): A curve with tangent vector transported by parallel translation.

Examples under Riemannian connection:

2 dimensional flat plane: straight line

3 dimensional sphere: great circle





Autoparallel submanifold

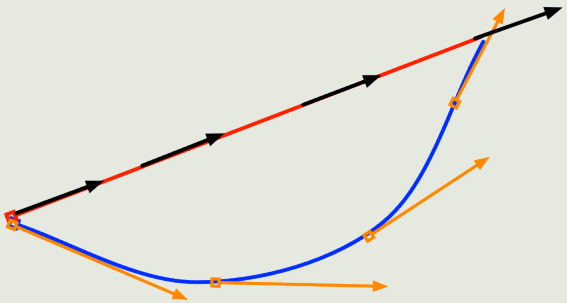
Geodesics

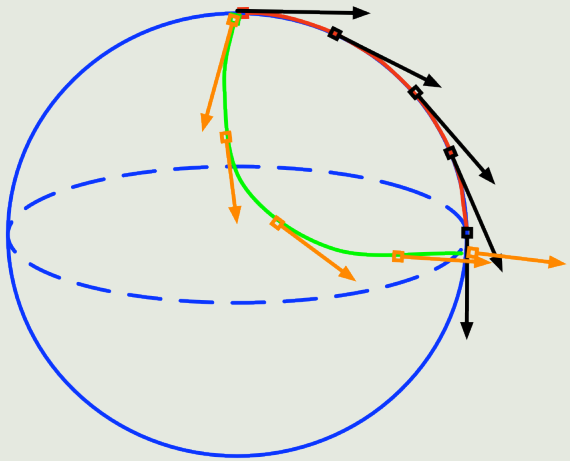
The geodesics with respect to the Riemannian connection are known to coincide with the shortest curve joining two points.

Shortest curve: curve with the shortest length.

Length of a curve $\gamma : [a, b] \to S$:

$$\|\gamma\| = \int_a^b \|\frac{d\gamma}{dt}\| dt = \int_a^b \sqrt{g_{ij}\dot{\gamma}^i\dot{\gamma}^j} dt \tag{22}$$





Part II

Geometric structure of statistical models and statistical inference

Motivation

Motivation

Consider the set of probability distributions as a manifold.

Analysis the relationship between the geometric structure of the manifold and statistical estimation.

Introduce concepts like metric, affine connection on statistical models and studying quantities such as distance, the tangent space (which provides linear approximations), geodesics and the curvature of a manifold.

Statistical models

Statistical models

$$\mathcal{P}(\mathcal{X}) = \{p : \mathcal{X} \to \mathbb{R} \mid p(x) > 0 \ (\forall x \in \mathcal{X}), \int p(x)dx = 1\} \tag{23}$$

Example Normal Distribution:

$$\mathcal{X} = \mathbb{R}, n = 2, \xi = [\mu, \sigma], \Xi = \{[\mu, \sigma] | -\infty < \mu < \infty, 0 < \sigma < \infty\}$$

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

Geometric structure of statistical models and statistical inference

Basic concepts

The Fisher metric and α-connection

Exponential family

Divergence and Geometric statistical inference

The Fisher information matrix

Fisher information matrix $G(\xi) = [g_{i,j}(\xi)]$, and

$$g_{i,j}(\xi) = E_\xi[\partial_i \ell_\xi \partial_j \ell_\xi] = \int \partial_i \ell(x;\xi)\partial_j \ell(x;\xi) p(x;\xi) dx$$

where $\ell_\xi = \ell(x;\xi) = log\ p(x;\xi)$ and E_ξ denotes the expectation w.r.t. the distribution p_ξ.

Motivation:
Sufficient statistic and Cramér-Rao bound

The Fisher information matrix

Sufficient statistic

Sufficient statistic: for $Y = F(X)$, given the distribution $p(x;\xi)$ of X, we have $p(x;\xi) = q(F(x);\xi)r(x;\xi)$, if $r(x;\xi)$ does not depend on ξ for all x, we say that F is a sufficient statistic for the model S. Then we can write $p(x;\xi) = q(y;\xi)r(x)$.

A sufficient statistic is a function whose value contains all the information needed to compute any estimate of the parameter (e.g. a maximum likelihood estimate).

Fisher information matrix and sufficient statistic

Let $G(\xi)$ be the Fisher information matrix of $S = p(x;\xi)$, and $G_F(\xi)$ be the Fisher information matrix of the induced model $S_F = q(y;\xi)$, then we have $G_F(\xi) \leqslant G(\xi)$ in the sense that $\Delta G(\xi) = G_F(\xi) - G(\xi)$ is positive semidefinite. $\Delta G(\xi) = 0$ iff. F is a sufficient statistic for S.

Cramér-Rao inequality

Cramér-Rao inequality

The variance of any unbiased estimator is at least as high as the inverse of the Fisher information.

Unbiased estimator $\hat{\xi}$: $E_\xi[\hat{\xi}(X)] = \xi$

The variance-covariance matrix $V_\xi[\hat{\xi}] = [v_\xi^{ij}]$ where

$$v_\xi^{ij} = E_\xi[(\hat{\xi}^i(X) - \xi^i)(\hat{\xi}^j(X) - \xi^j)]$$

Thus Cramér-Rao inequality state that $V_\xi[\hat{\xi}] \geqslant G(\xi)^{-1}$, and an unbiased estimator $\hat{\xi}$ satisfying $V_\xi[\hat{\xi}] = G(\xi)^{-1}$ is called an efficient estimator.

α-connection

α-connection

Let $S = \{p_\xi\}$ be an n-dimensional model, and consider the function $\Gamma^{(\alpha)}_{ij,k}$ which maps each point ξ to the following value:

$$(\Gamma^{(\alpha)}_{ij,k})_\xi = E_\xi[(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi)(\partial_k\ell_\xi)] \tag{24}$$

where α is an arbitrary real number. We defined an affine connection $\nabla^{(\alpha)}$ which satisfy:

$$\left\langle \nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k \right\rangle = \Gamma^{(\alpha)}_{ij,k} \tag{25}$$

where $g = \langle , \rangle$ is the Fisher metric. We call $\nabla^{(\alpha)}$ the α-connection

α-connection

Properties of α-connection

- α-connection is a symmetric connection
- Relationship between α-connection and β-connection:

$$\Gamma_{ij,k}^{(\beta)} = \Gamma_{ij,k}^{(\alpha)} + \frac{\alpha - \beta}{2} E[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi]$$

- The 0-connection is the Riemannian connection with respect to the Fisher metric.
- $$\Gamma_{ij,k}^{(\beta)} = \Gamma_{ij,k}^{(0)} + \frac{-\beta}{2} E[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi]$$
- $$\nabla^{(\alpha)} = \frac{1+\alpha}{2} \nabla^{(1)} + \frac{1-\alpha}{2} \nabla^{(-1)}$$

Geometric structure of statistical models and statistical inference

Basic concepts

The Fisher metric and α-connection

Exponential family

Divergence and Geometric statistical inference

Exponential family

Exponential family

$$p(x;\theta) = exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)]$$

$[\theta^i]$ are called the natural parameters(coordinates), and ψ is the potential function for $[\theta^i]$, which can be calculated as

$$\psi(\theta) = log \int exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x)]dx$$

The exponential families include many of the most common distributions, including the normal, exponential, gamma, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, and so on.

Exponential family

Exponential family

Examples: Normal Distribution

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{26}$$

where $C(x) = 0, F_1(x) = x, F_2(x) = x^2$, and $\theta^1 = \frac{\mu}{\sigma^2}, \theta^2 = -\frac{1}{2\sigma^2}$ are the natural parameters, the potential function is :

$$\psi = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2} log(-\frac{\pi}{\theta^2}) = \frac{\mu^2}{2\sigma^2} + log(\sqrt{2\pi}\sigma) \tag{27}$$

Mixture family

Mixture family

$$p(x;\theta) = C(x) + \sum_{i=1}^{n} \theta^i F_i(x)$$

In this case we say that S is a mixture family and $[\theta^i]$ are called the mixture parameters.

e-connection and m-connection

The natural parameters of exponential family form a 1-affine coordinate system($\Gamma^{(1)}_{ij,k} = 0$), which means the connection is 1-flat, we call the connection $\nabla^{(1)}$ the e-connection, and call exponential family e-flat.
The mixture parameters of mixture family form a (-1)-affine coordinate system($\Gamma^{(-1)}_{ij,k} = 0$), which means the connection is (-1)-flat, and we call the connection $\nabla^{(-1)}$ the m-connection and call mixture family m-flat.

Dual connection

Dual connection

Definition: Let S be a manifold on which there is given a Riemannian metric g and two affine connection ∇ and ∇^*. If for all vector fields $X, Y, Z \in \mathcal{T}(S)$,

$$Z < X, Y >=< \nabla_Z X, Y > + < X, \nabla_Z^* Y > \tag{28}$$

hold, we say that ∇ and ∇^* are duals of each other w.r.t. g and call one the dual connection of the other.
Additional, we call the triple (g, ∇, ∇^*) a dualistic structure on S.

Dual connection

Properties

- For any statistical model, the α-connection and the $(-\alpha)$-connection are dual with respect to the Fisher metric.
- $$\langle \Pi_\gamma(D_1), \Pi^*_\gamma(D_2) \rangle_q = \langle D_1, D_2 \rangle_p.$$

  where Π_γ and Π^*_γ are parallel translation along γ w.r.t. ∇ and ∇^*.
- $$R = 0 \Leftrightarrow R^* = 0$$

  where R and R^* are the curvature tensors of ∇ and ∇^*.

Dually flat spaces and dual coordinate system

Dually flat spaces

Let (g, ∇, ∇^*) be a dualistic structure on a manifold S, then we have $R = 0 \Leftrightarrow R^* = 0$, and if the connnection ∇ and ∇^* are both symmetric($T = T^* = 0$), then we see that ∇-flatness and ∇^*-flatness are equivalent.

We call (S, g, ∇, ∇^*) a dually flat space if both duals ∇ and ∇^* are flat.

Examples: Since α-connections and $-\alpha$-connections are dual w.r.t. Fisher metric and α-connections are symmetry, we have for any statistical model S and for any real number α

$$S \text{ is } \alpha - \text{flat} \Leftrightarrow S \text{ is } (-\alpha) - \text{flat} \tag{29}$$

Dually flat spaces and dual coordinate system

Dual coordinate system

For a particular ∇-affine coordinate system $[\theta^i]$, if we choose a corresponding ∇^*-affine coordinate system $[\eta_j]$ such that

$$g = \langle \partial_i, \partial^j \rangle = \delta_i^j$$

where $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial^j = \frac{\partial}{\partial \eta_j}$.
Then we say the two coordinate systems mutually dual w.r.t. metric g, and call one the dual coordinate system of the other.

Existence of dual coordinate system

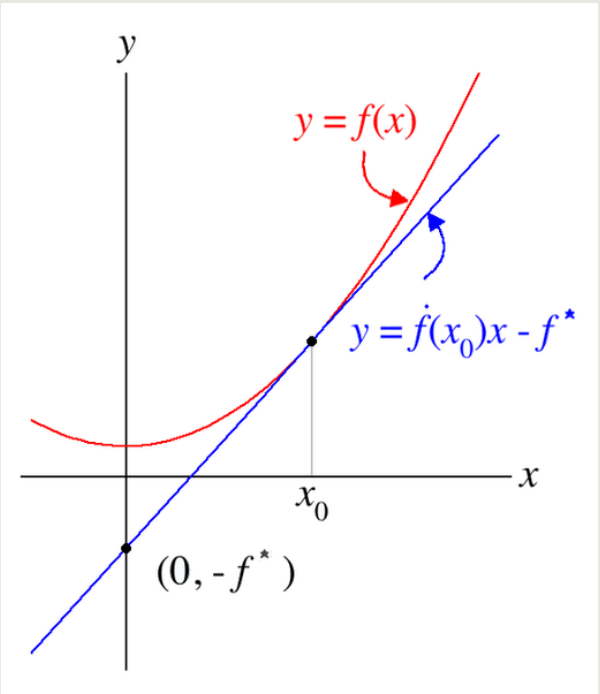
A pair of dual coordinate system exist if and only if (S, g, ∇, ∇^*) is a dually flat space.

Legendre transformations

Consider mutually dual coordinate system $[\theta^i]$ and $[\eta_i]$ with functions $\psi : S \to \mathbb{R}$ and $\varphi : S \to \mathbb{R}$ satisfy the following equations:

$$\partial_i \psi = \eta_i$$
$$\partial^i \varphi = \theta^i$$
$$g_{i,j} = \partial_i \eta_j = \partial_j \eta_i = \partial_i \partial_j \psi$$
$$\varphi(\eta) = max_\theta \{\theta^i \eta_i - \psi(\theta)\}$$
$$\psi(\theta) = max_\eta \{\theta^i \eta_i - \varphi(\eta)\}$$

Legendre transformations



Geometric interpretation for $f^*(p) = max_x(px - f(x))$:
A convex function $f(x)$ is shown in red, and the tangent line at point $(x_0, f(x_0))$ is shown in blue. The tangent line intersects the vertical axis at $(0, -f^*)$ and f^* is the value of the Legendre transform $f^*(p_0)$, where $p_0 = \dot{f}(x_0)$. Note that for any other point on the red curve, a line drawn through that point with the same slope as the blue line will have a y-intercept above the point $(0, -f^*)$, showing that is indeed a maximum.

Examples of Legendre transformations

Examples:

- The Legendre transform of $f(x) = \frac{1}{p}|x|^p$ (where $1 < p < \infty$) is $f^*(x^*) = \frac{1}{q}|x^*|^q$ (where $1 < q < \infty$),
- The Legendre transform of $f(x) = e^x$ is $f^*(x^*) = x^* ln\, x^* - x^*$ (where $x^* > 0$),
- The Legendre transform of $f(x) = \frac{1}{2}x^T Ax$ is $f^*(x^*) = \frac{1}{2}x^{*T}A^{-1}x^*$,
- The Legendre transform of $f(x) = |x|$ is $f^*(x^*) = 0$ if $x^* \leqslant 1$, and $f^*(x^*) = \infty$ if $x^* > 1$.

The natural parameter and dual parameter of Exponential family

- For distribution $p(x;\theta) = exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)]$, $[\theta^i]$ are called the natural parameters
- If we define $\eta_i = E_\theta[F_i] = \int F_i(x)p(x;\theta)dx$, we can verify $[\eta_i]$ is a (-1)-affine coordinate system dual to $[\theta^i]$, we call this $[\eta_i]$ the expectation parameters or the dual parameters.

The natural parameter and dual parameter of Exponential family

Recall: Normal Distribution

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{30}$$

where $C(x) = 0, F_1(x) = x, F_2(x) = x^2$, and $\theta^1 = \frac{\mu}{\sigma^2}, \theta^2 = -\frac{1}{2\sigma^2}$ are the natural parameters, the potential function is :

$$\psi = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2} log(-\frac{\pi}{\theta^2}) = \frac{\mu^2}{2\sigma^2} + log(\sqrt{2\pi}\sigma) \tag{31}$$

The dual parameter are calculated as $\eta_1 = \frac{\partial \psi}{\partial \theta^1} = \mu = -\frac{\theta^1}{2\theta^2}$,
$\eta_2 = \frac{\partial \psi}{\partial \theta^2} = \mu^2 + \sigma^2 = \frac{(\theta^1)^2 - 2\theta^2}{4(\theta^2)^2}$, It has potential function:

$$\varphi = -\frac{1}{2}(1 + log(-\frac{\pi}{\theta^2})) = -\frac{1}{2}(1 + log(2\pi)) + 2log\sigma) \tag{32}$$

Geometric structure of statistical models and statistical inference

Basic concepts

The Fisher metric and α-connection

Exponential family

Divergence and Geometric statistical inference

Divergences

Let S be a manifold and suppose that we are given a smooth function $D = D(\cdot\|\cdot) : S \times S \to \mathbb{R}$ satisfying for any $p, q \in S$:

$$D(p\|q) \geqslant 0 \textit{ with equality iff }. p = q) \tag{33}$$

Then we introduce a distance-like measure of the separation between two points.

Divergence, semimetrics and metrics

- A distance satisfying positive-definiteness, symmetry and triangle inequality is called a *metric*;
- A distance satisfying positive-definiteness and symmetry is called *semimetrics*;
- A distance satisfying only positive-definiteness is called a *divergence*.

Kullback-Leibler divergence

Discrete random variables p and q:

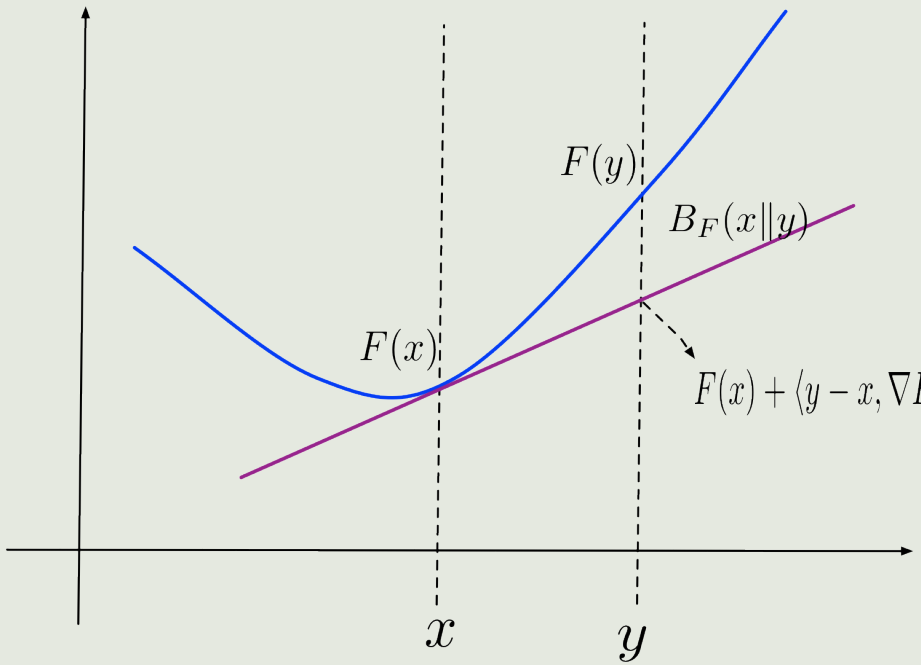
$$D_{KL}(p\|q) = \sum_i p(x) log \frac{p(x)}{q(x)} \tag{34}$$

Continuous random variables p and q:

$$D_{KL}(p\|q) = \int p(x) log \frac{p(x)}{q(x)} dx \tag{35}$$

Generally , we use Kullback-Leibler divergence to measurethe difference between two probability distributions p and q. KL measures the expected number of extra bits required to code samples from p when using a code based on q, rather than using a code based on p. Typically p represents the ”true” distribution of data, observations, or a precisely calculated theoretical distribution. The measure q typically represents a theory, model, description, or approximation of p.

Bregman divergence



Bregman divergence associated with F for points $p, q \in \Delta$ is :
$B_F(x\|y) =$
$F(y) - F(x) - \langle (y-x), \nabla F(x) \rangle$,
where F(x) is a convex function defined on a closed convex set Δ.

Examples:
$F(x) = \|x\|^2$, then $B_F(x\|y) = \|x-y\|^2$.
More generally, if $F(x) = \frac{1}{2}x^T Ax$, then $B_F(x\|y) = \frac{1}{2}(x-y)^T A(x-y)$.
KL divergence:if $F = \sum_i x\, logx - \sum x$, we get KullbackLeibler divergence.

Canonical divergence

Canonical divergence(a divergence for dually flat space)

Let (S, g, ∇, ∇^*) be a dually flat space, and $\{[\theta^i], [\eta_j]\}$ be mutually dual affine coordinate systems with potentials $\{\psi, \varphi\}$, then the canonical divergence($(g, \nabla) - \textit{divergence}$) is defined as:

$$D(p\|q) = \psi(p) + \varphi(q) - \theta^i(p)\eta_j(q) \tag{36}$$

Canonical divergence

Properties:

- Relation between $(g, \nabla) - \textit{divergence}$ and $(g, \nabla^*) - \textit{divergence}$: $D^*(p\|q) = D(q\|p)$
- If M is a autoparallel submanifold w.r.t. either ∇ or ∇^*, then the (g_M, ∇_M)-divergence $D_M = D|_{M\times M}$ is given by $D_M(p\|q) = D(p\|q)$
- If ∇ is a Riemannian connection($\nabla = \nabla^*$) which is flat on S, there exist a coordinate system which is self-dual($\theta^i = \eta_i$), then $\varphi = \psi = \frac{1}{2}\sum_i (\theta^i)^2$, then the canonical divergence is

$$D(p\|q) = \frac{1}{2}\{d(p,q)\}^2$$

where $d(p,q) = \sqrt{\sum_i \{\theta^i(p) - \theta^i(q)\}^2}$

Canonical divergence

Triangular relation

Let $\{[\theta^i], [\eta_i]\}$ be mutually dual affine coordinate systems of a dually flat space (S, g, ∇, ∇^*), and let D be a divergence on S. Then a necessary and sufficient condition for D to be the (g, ∇)-divergence is that for all $p, q, r \in S$ the following triangular relation holds:

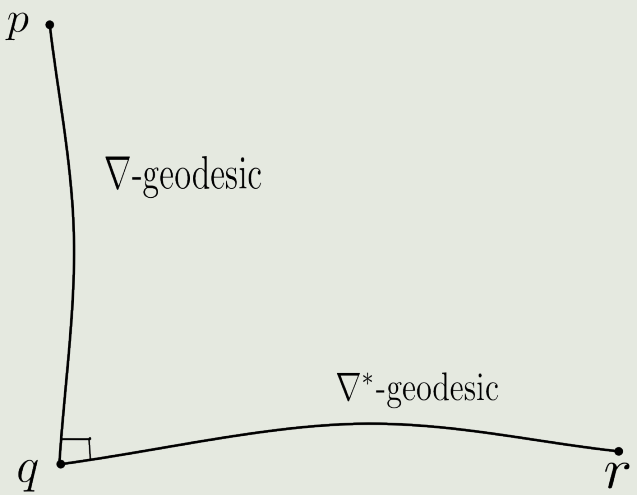
$$D(p\|q) + D(q\|r) - D(p\|r) = \{\theta^i(p) - \theta^i(q)\}\{\eta^i(p) - \eta^i(q)\} \tag{37}$$

Canonical divergence

Pythagorean relation

Let p, q, and r be three points in S. Let γ_1 be the ∇-geodesic connecting p and q, and let γ_2 be the ∇^*-geodesic connecting q and r. If at the intersection q the curve γ_1 and γ_2 are orthogonal(with respect to the inner product g), then we have the following Pythagorean relation.

$$D(p\|r) = D(p\|q) + D(q\|r) \tag{38}$$



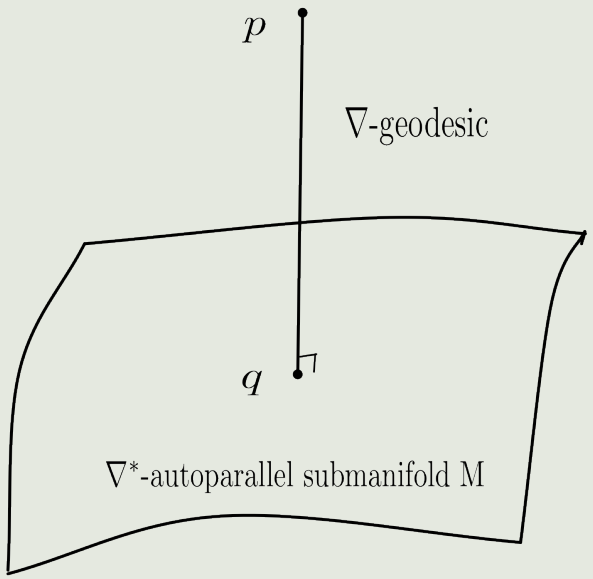
Canonical divergence

Projection theorem

Let p be a point in S and let M be a submanifold of S which is ∇^*-autoparallel. Then a necessary and sufficient condition for a point q in M to satisfy

$$D(p\|q) = min_{r\in M} D(p\|r) \tag{39}$$

is for the ∇-geodesic connecting p and q to be orthogonal to M at q.



Canonical divergence

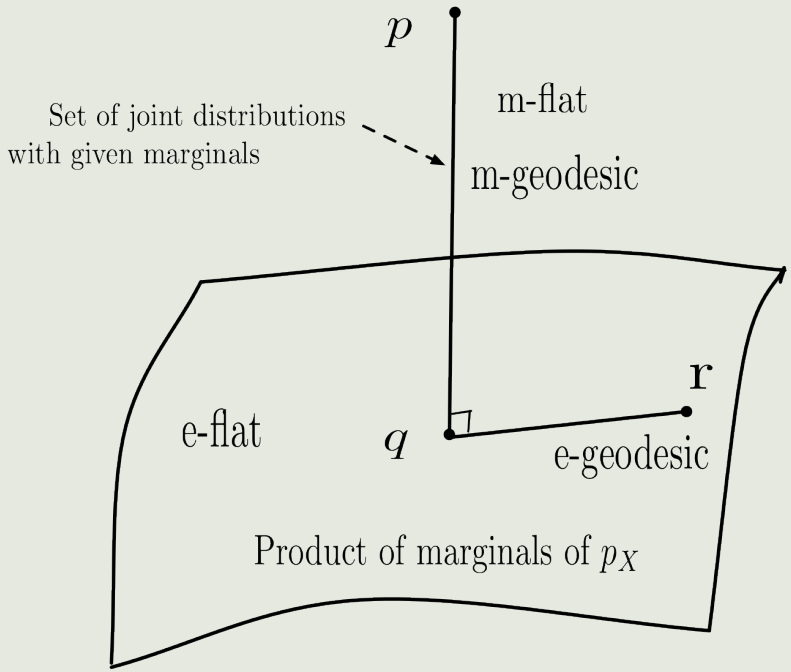
Examples

From the definition of exponential family and mixture family, the product of exponential family are still exponential family, the sum of mixture family are still mixture family.

- e-flat submanifold: set of all product distributions:
  $E_0 = \{p_X | p_X(x_1, \cdots, x_N) = \prod_{i=1}^{N} p_{X_i}(x_i)\}$
- m-flat submanifold: set of joint distributions with given marginals:
  $M_0 = \{p_X | \sum_{X \setminus i} p_X(x) = q_i(x_i) \quad \forall i \in \{1, \cdots, N\}\}$

Canonical divergence

Examples



Thanks!

Thanks!

Question?