Manifold of probability distribution

Yunshu Liu

2012-04-20

Outline

1. Examples of manifold: Transmit two bits
2. Important Submanifolds
3. Information projection

Probability distribution viewed as manifold

Outline

Examples of manifold: Transmit two bits
Important Submanifolds
Information projection

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.

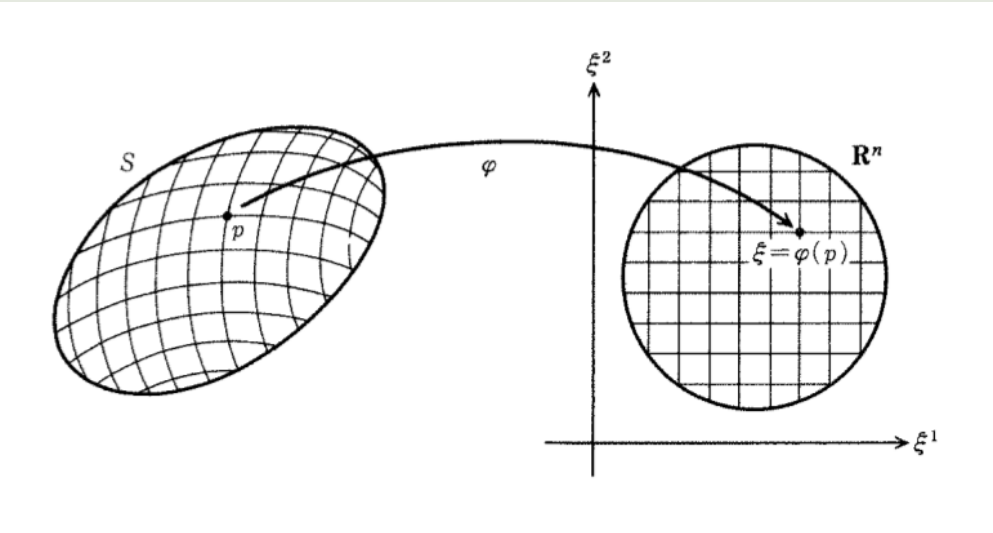


Figure: A coordinate system ξ for a manifold S

Coordinate systems for manifold

Parametrization of color models

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

Examples:

Let $x = \{x_1, x_2\}^T$, $x_i \in \{-1, 1\}$, consider the family of all the probability distributions over x:

$$S = \{p(x) \mid p(x) > 0, \sum_{x \in C} p(x) = 1\} \tag{1}$$

Outcomes of x:

$$C = \{C_0 = \{-1, -1\}, C_1 = \{-1, 1\}, C_2 = \{1, -1\}, C_3 = \{1, 1\}\} \tag{2}$$

Parameterization of S:

$$p(x = C_i) = \eta^i \; for \; i = 1, 2, 3, \; and \; p(x = C_0) = 1 - \sum_{j=1}^{3} \eta^j \tag{3}$$

where $\{\eta^1, \eta^2, \eta^3\}$ can be considered as a coordinate system of the manifold S.

A new coordinate $\{\theta_i\}$, which is defined as

$$\theta_i = ln \frac{\eta^i}{1 - \sum_{j=1}^{3} \eta^j} for\ i = 1, 2, 3. \tag{4}$$

Introduce new random variables:

$$\delta_{C_i}(x) = \begin{cases} 1 & \text{if x = } C_i \\ 0 & \text{otherwise} \end{cases}$$

and let

$$\psi = -ln\ (1 - \sum_{j=1}^{3} \eta^j) = -ln\ p(x = C_0) \tag{5}$$

Then S is an exponential family with natural parameter $\{\theta_i\}$:

$$p(x, \theta) = exp(\sum_{i=1}^{3} \theta_i \delta_{C_i}(x) - \psi(\theta)) \tag{6}$$

Properties of Exponential family S

Exponential family

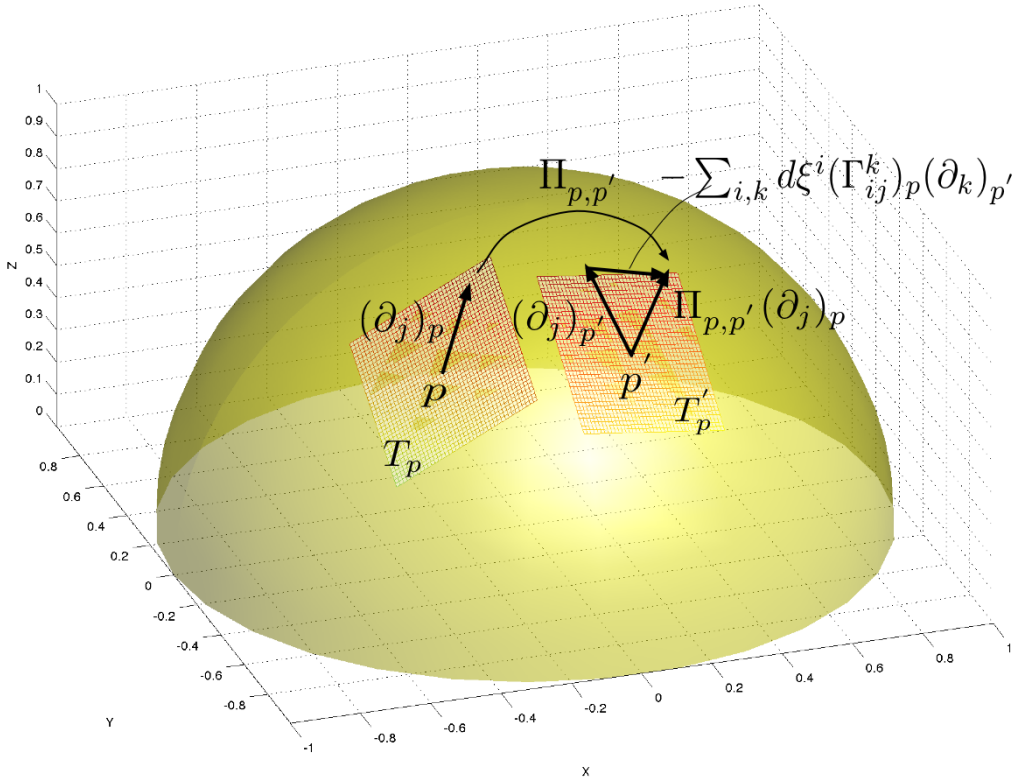
$$p(x;\theta) = exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)]$$

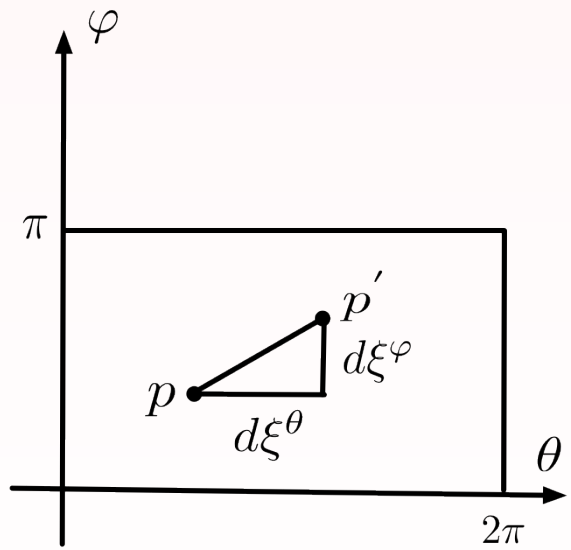
- An exponential family is e-flat and its nature parameters θ form an e-affine coordinate system
- An exponential family is m-flat and its dual parameters η form an m-affine coordinate system

Notes: e-flat means S is flat w.r.t. 1-connection(α = 1), m-flat means S is flat w.r.t. (-1)-connection(α = -1)

Affine connection: relationships between tangent space at different points

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





α-connection

α-connection

For $S = \{p_\xi\}$, consider the function $\Gamma^{(\alpha)}_{ij,k}$ which maps each point ξ to the following value:

$$(\Gamma^{(\alpha)}_{ij,k})_\xi = E_\xi[(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi)(\partial_k\ell_\xi)] \tag{7}$$

where α is an arbitrary real number, $\ell_\xi = \ell(x;\xi) = log\ p(x;\xi)$ and E_ξ denotes the expectation w.r.t. the distribution p_ξ.
We defined an affine connection $\nabla^{(\alpha)}$ which satisfy:

$$\left\langle \nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k \right\rangle = \left\langle \sum_h \Gamma^{h(\alpha)}_{i,j}\partial_h, \partial_k \right\rangle = \sum_h \Gamma^{h(\alpha)}_{i,j} g_{hk} = \Gamma^{(\alpha)}_{ij,k} \tag{8}$$

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

Affine connection

Affine connection ∇: $\mathcal{T}(S) \times \mathcal{T}(S) \to \mathcal{T}(S)$

All the affine connection on a manifold forms an affine space:
For any affine connection ∇ and ∇', and for any real number $\alpha \in \mathbb{R}$, the affine connection $\alpha\nabla + (1-\alpha)\nabla'$ defines another affine connection.
Example: for α-connection:

$$\nabla^{(\alpha)} = \alpha\nabla^{(1)} + (1-\alpha)\nabla^{(0)} \tag{9}$$

Affine connection

Riemannian connection($\alpha = 0$): symmetric and metric

Symmetric connection: $\Gamma_{ij}^{k} = \Gamma_{ji}^{k}$ hold for all i, j and k
Metric connction:

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

Dual connection ∇ and ∇^*

- 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 ∇^*.

Affine coordinate system and 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 parallel translation is independent of curve choices and the connection is symmetric.
Equivalent conditions for a coordinate system to be an affine coordinate system:

$$\nabla_{\partial_i}\partial_j = \sum_{k=1}^{n}(\Gamma_{ij}^k\partial_k) = 0 \quad \textit{for all } i \textit{ and } j \tag{10}$$

$$\Gamma_{ij}^k = 0 \quad \textit{for all } i,j \textit{ and } k \tag{11}$$

Flatness

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

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: A sufficient statistic is a function whose value contains all the information needed to compute any estimate of the parameter.

$$p(x;\xi) = q(y;\xi)r(x) \tag{12}$$

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.

$$V_\xi[\hat{\xi}] \geqslant G(\xi)^{-1} \tag{13}$$

where $E_\xi[\hat{\xi}] = \xi$, the variance-covariance matrix $V_\xi[\hat{\xi}] = [v_\xi^{ij}]$ and $v_\xi^{ij} = E_\xi[(\hat{\xi}^i(X) - \xi^i)(\hat{\xi}^j(X) - \xi^j)]$

An unbiased estimator $\hat{\xi}$ satisfying $V_\xi[\hat{\xi}] = G(\xi)^{-1}$ is called an **efficient estimator**(minimum variance among all the unbiased estimator).

Fact: The best unbiased estimator does not always achieve Cramér-Rao bound, a necessary and sufficient condition for a coordinate system ξ of a model S to have an efficient estimator is that S is an exponential family and ξ is a m-affine coordinate system.

Probability distribution viewed as manifold

Outline

Examples of manifold: Transmit two bits
Important Submanifolds
Information projection

Submanifolds

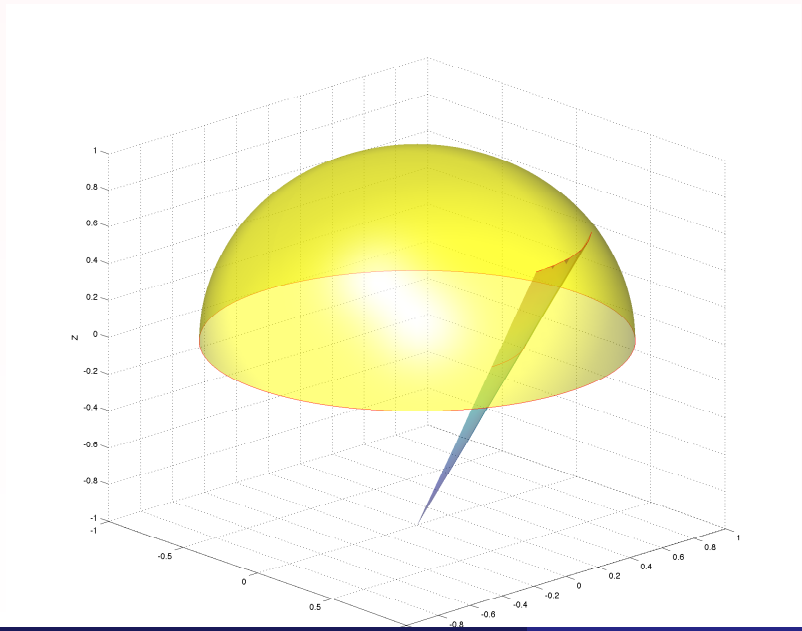
Submanifolds

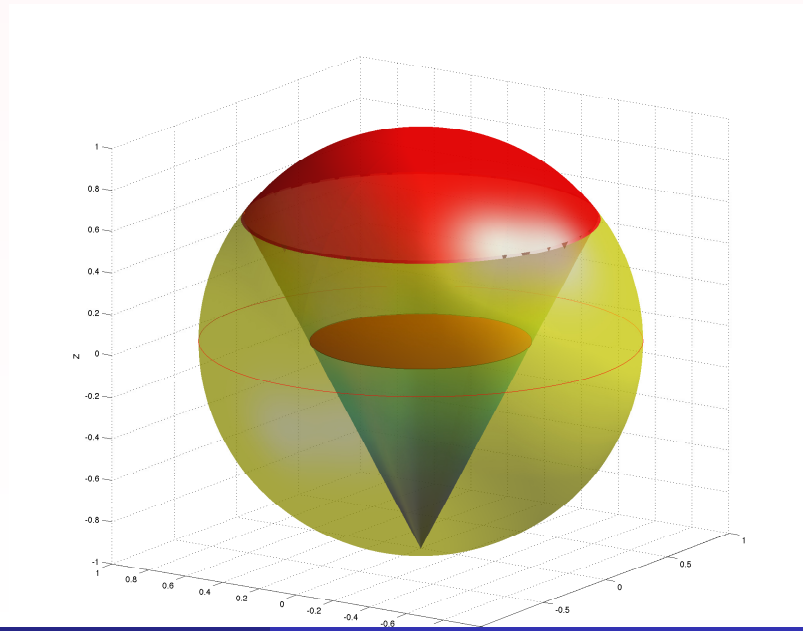
Definition: a submanifold M of a manifold S is a subset of S which itself has the structure of a manifold

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:





Two independent bits

Definition

The set of all product distributions, which is defined as

$$\mathcal{E} = \{p(x) \mid p(x) = p_{X_1}(x_1)p_{X_2}(x_2)\} \tag{14}$$

$p_{X_1}(x_1 = 1) = p_1, p_{X_2}(x_2 = 1) = p_2$
then $p_{X_1}(x_1 = -1) = 1 - p_1, p_{X_2}(x_2 = -1) = 1 - p_2$

Parameterization

Define $\lambda_i = \frac{1}{2}\, ln\, \frac{p_i}{1-p_i}$ for i = 1, 2, and $\psi(\lambda) = \sum_{i=1,2} ln\, (e^{\lambda_i} + e^{-\lambda_i})$,

$$p(x, \lambda) = e^{\lambda_1 x_1 - \psi(\lambda_1)} e^{\lambda_2 x_2 - \psi(\lambda_2)} = e^{\lambda x - \psi(\lambda)} \tag{15}$$

Thus $\mathcal{E}$ is an exponential family with nature parameter $\{\lambda_i\}$

Properties of Exponential family $\mathcal{E}$

Dual parameters: expectations

$$\gamma = \begin{pmatrix} \gamma^1 \\ \gamma^2 \end{pmatrix} = \begin{pmatrix} \partial_1 \psi(\lambda) \\ \partial_2 \psi(\lambda) \end{pmatrix} = \begin{pmatrix} 2p_1 - 1 \\ 2p_2 - 1 \end{pmatrix} = E_p(x)$$

Properties

- $\mathcal{E}$ is an e-autoparallel submanifold(plane in θ coordinate) of S, but not a m-autoparallel submanifold(not a plane in η coordinate):

$$\theta(p) = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \lambda(p)$$

- $\mathcal{E}$ is e-flat with e-affine coordinate system $\{\lambda\}$
- $\mathcal{E}$ is m-flat with m-affine coordinate system $\{\gamma\}$

Autoparallel submanifold

Condition for a submanifold M of S to be autoparallel

$$\nabla_X Y \in \mathcal{T}(M) \text{ for } \forall X, Y \in \mathcal{T}(M) \tag{16}$$

$$\nabla_{\partial_a} \partial_b \in \mathcal{T}(M) \text{ for all } a \text{ and } b \tag{17}$$

where M is a submanifold of S and X, Y are vector fields of M, $\partial_a = \frac{\partial}{\partial u^a}$ and $\partial_b = \frac{\partial}{\partial u^b}$ are the basis for submanifold M w.r.t. coordinate system $\{u^i\}$

Geodesics

Geodesics(1-dimensional autoparallel submanifolds): A curve with tangent vector transported by parallel translation.

Autoparallel submanifold

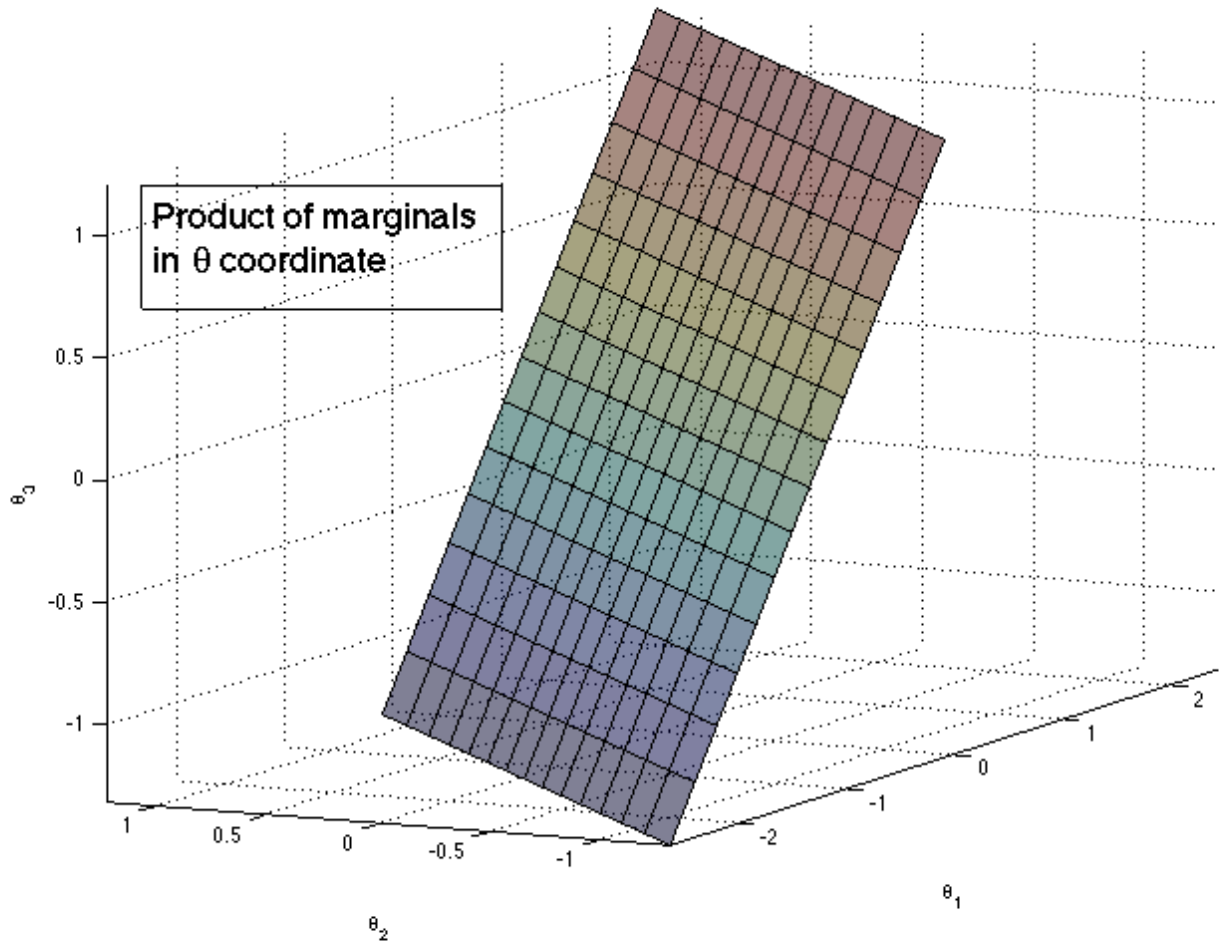
Condition for a submanifold M of S to be autoparallel

Now consider a flat submanifold M of S, assume $\{\xi\}$ and $\{u\}$ are affine coordinate systems of S and M respectively, then a necessary and sufficient condition for M to be autoparallel is that there exist an $n \times m$ matrix A and an n-dimensional vector B such that

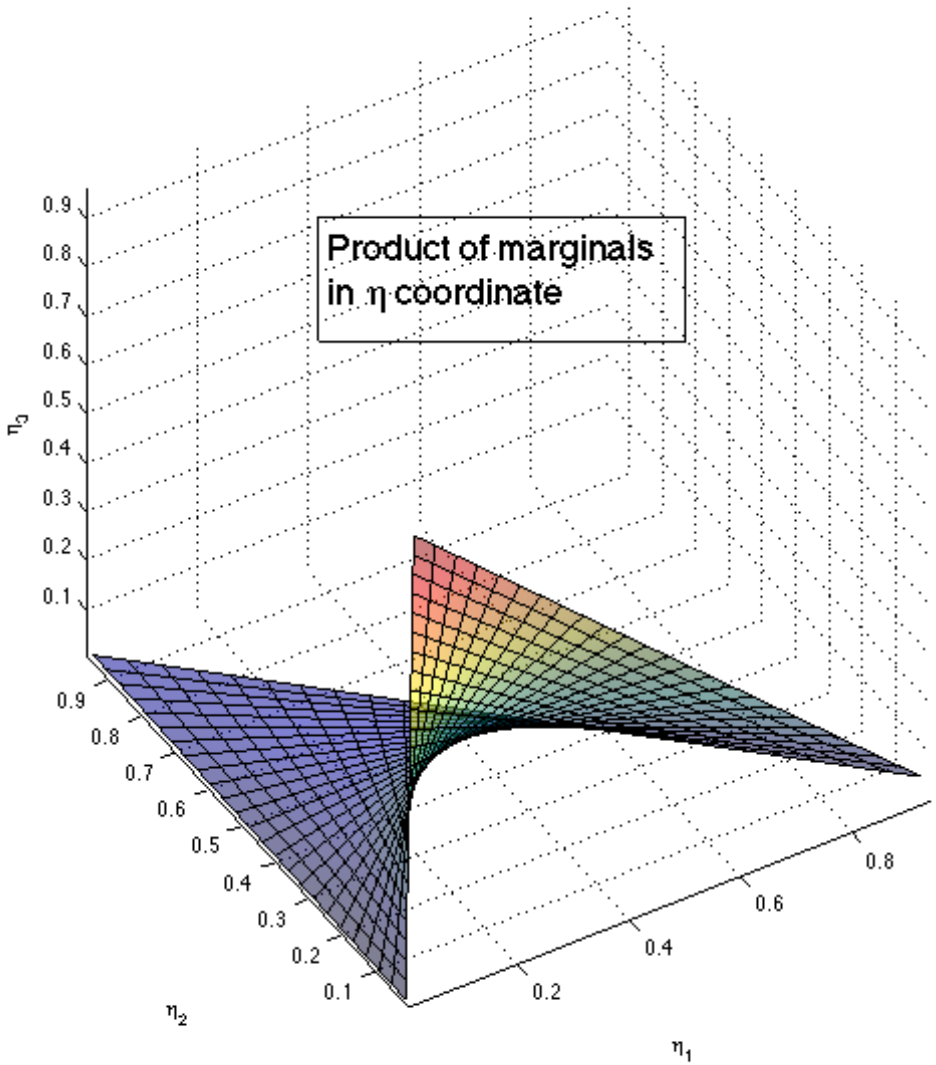
$$\xi(p) = Au(p) + B \quad (\forall p \in M) \tag{18}$$

A subspace of $\mathbb{R}^n$ which can be expressed as $\{Au + B | u \in \mathbb{R}^m\}$ is called an affine subspace of $\mathbb{R}^n$, when B = 0 it becomes a linear subspace.

$\mathcal{E}$ in θ coordinate



$\mathcal{E}$ in η coordinate



1-dimensional submanifold

set of joint distributions with given marginals

Note

$$\begin{cases} p(x = C_0) + p(x = C_1) = p_{X_1}(x_1 = -1) \\ p(x = C_2) + p(x = C_3) = p_{X_1}(x_1 = 1) \end{cases}$$

and

$$\begin{cases} p(x = C_1) + p(x = C_3) = p_{X_2}(x_2 = -1) \\ p(x = C_0) + p(x = C_2) = p_{X_2}(x_2 = 1) \end{cases}$$

If given $p_{X_1}(x_1)$ or $p_{X_2}(x_2)$, we only need two variable to represent any point in S, thus it forms a 2-dimensional submanifold of S. Further more, if $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ are both given, we have a 1-dimensional submanifold of S.

1-dimensional submanifold

Definition and Property

$$\mathcal{M} = \{p_X | \sum_{X \backslash i} p_X(x) = q_i(x_i) \quad \forall i \in \{1, \cdots, N\}\} \tag{19}$$

. $\mathcal{M}$ is a m-autoparallel submanifold of S, which is both e-flat and m-flat.

Parameterization

Denote $\mathcal{M}$ the set of joint distributions with all the marginals given: $p_{X_1}(x_1 = 1) = p_1$ and $p_{X_2}(x_2 = 1) = p_2$, let $k = p(x = C_3)$ be the coordinate system for $\mathcal{M}$, we have

$$\begin{cases} p(x = C_0) = 1 + k - p_1 - p_2 \\ p(x = C_1) = p_2 - k \\ p(x = C_2) = p_1 - k \\ p(x = C_3) = k \end{cases}$$

Probability distribution viewed as manifold

Outline

Examples of manifold: Transmit two bits
Important Submanifolds
Information projection

Divergences

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 $p, q \in S$:

$$D(p\|q) \geqslant 0 \text{ with equality iff. } p = q) \tag{20}$$

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

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($(g, \nabla) - divergence$) is defined as:

$$D(p\|q) = \psi(p) + \varphi(q) - \theta^i(p)\eta_j(q) \tag{21}$$

From divergence to metric and connection

Riemannian metric:

$$g_{i,j}^{(D)}(p) = \langle \partial_i, \partial_j \rangle^{(D)} = -D[\partial_i \| \partial_j] = -\partial_i \partial_j' D[p\|q]_{|q=p} \tag{22}$$

Affine connection $\nabla^{(D)}$ with coefficients $\Gamma_{ij,k}^{(D)}$:

$$\Gamma_{ij,k}^{(D)}(p) = \langle \nabla_{\partial_i}^{(D)} \partial_j, \partial_k \rangle^{(D)} = -D(\partial_i \partial_j \| \partial_k) = -\partial_i \partial_j \partial_k' D[p\|q]_{|q=p} \tag{23}$$

where $\partial_i = \frac{\partial}{\partial p_i}$ and $\partial_i' = \frac{\partial}{\partial q_i}$.

Dual connection ∇ and ∇^*:

$$D^*(p\|q) = D(q\|p) \tag{24}$$

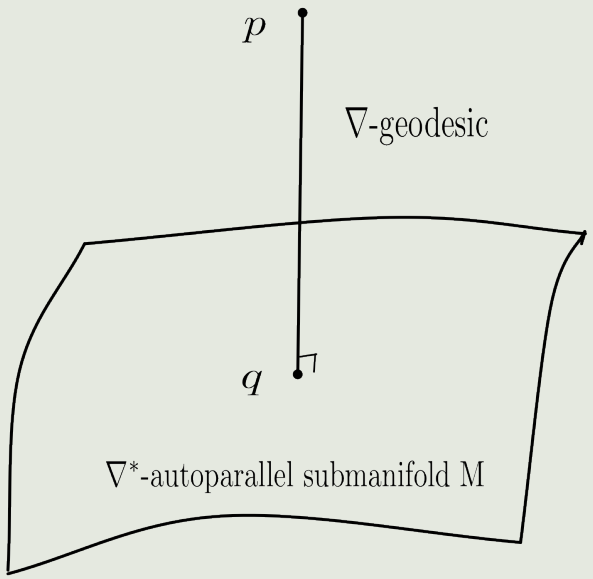
Projection theorem

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

$$D(p\|q) = min_{r\in M} D(p\|r) \tag{25}$$

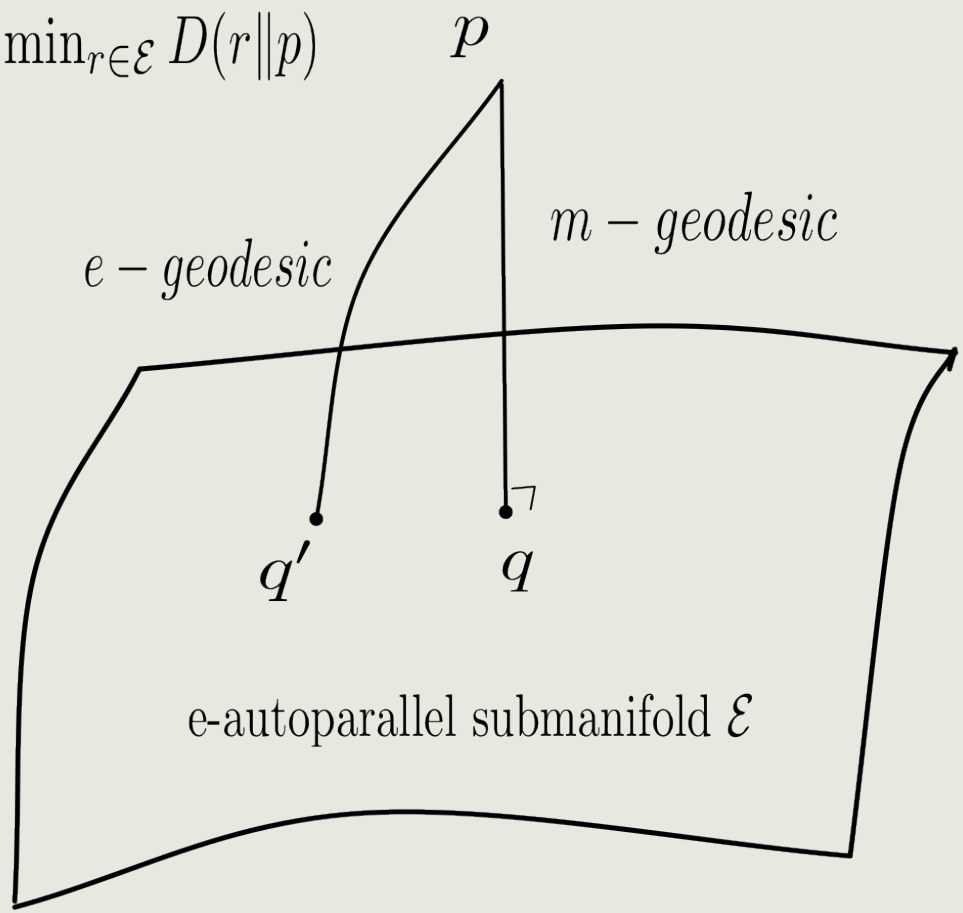
is for the ∇-geodesic connecting p and q to be orthogonal to M at q.



e-projection and m-projection

m-projection $q = \arg\min_{r \in \mathcal{E}} D(p\|r)$

e-projection $q' = \arg\min_{r \in \mathcal{E}} D(r\|p)$

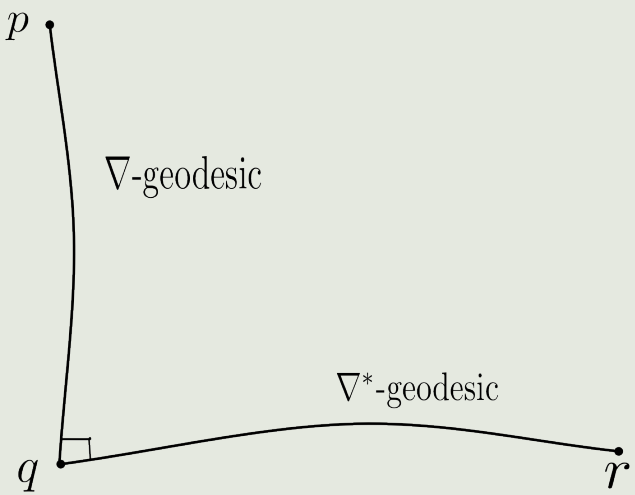


Pythagorean relation

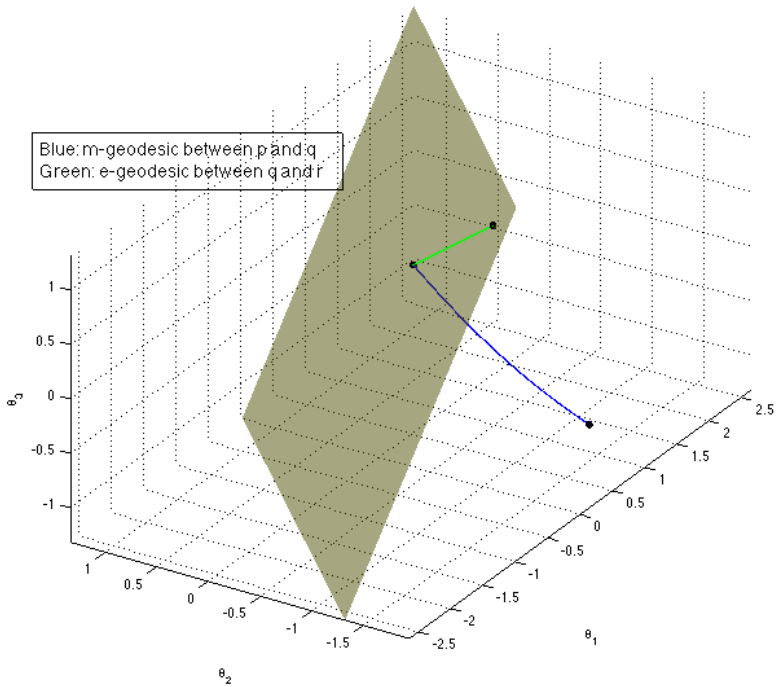
Pythagorean relation

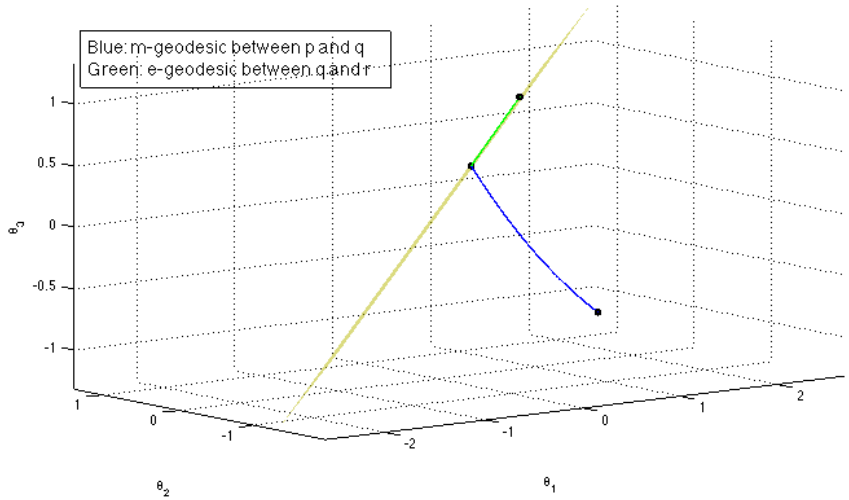
Let p, q, and r be three points in S. Let γ_1 be the ∇-geodesic connecting p and q, and let γ_2 be the ∇^*-geodesic connecting q and r. If at the intersection q the curve γ_1 and γ_2 are orthogonal(with respect to the inner product g), then we have the following Pythagorean relation.

$$D(p\|r) = D(p\|q) + D(q\|r) \tag{26}$$

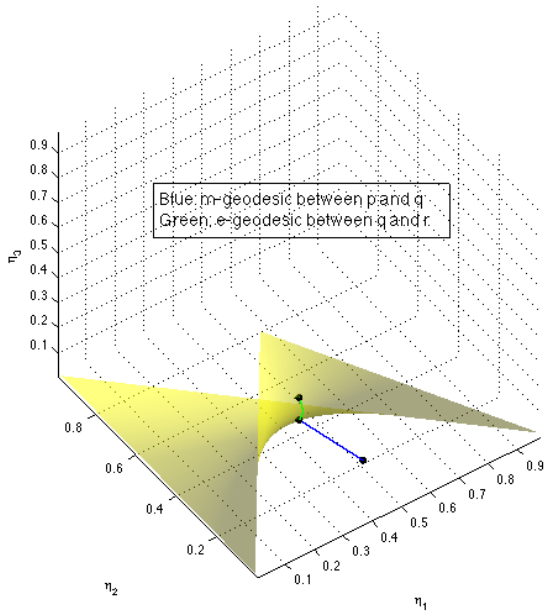


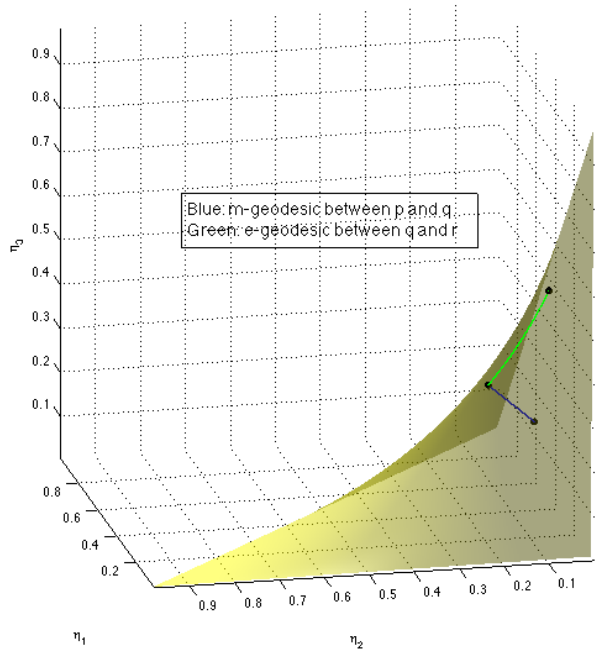
Projection in θ coordinate





Projection in η coordinate





Thanks!

Thanks!