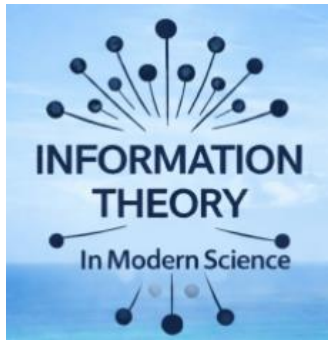
Geometric Information Theory

~ Hub to information sciences ~

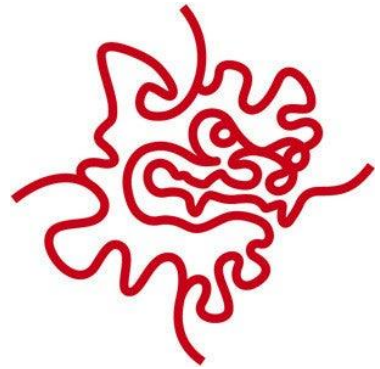
Part I: Dualistic framework of information geometry

Shannon dual geometry



Frank Nielsen

Sony Computer Science Laboratories, Inc.



Sony CSL

July 6th, 2026

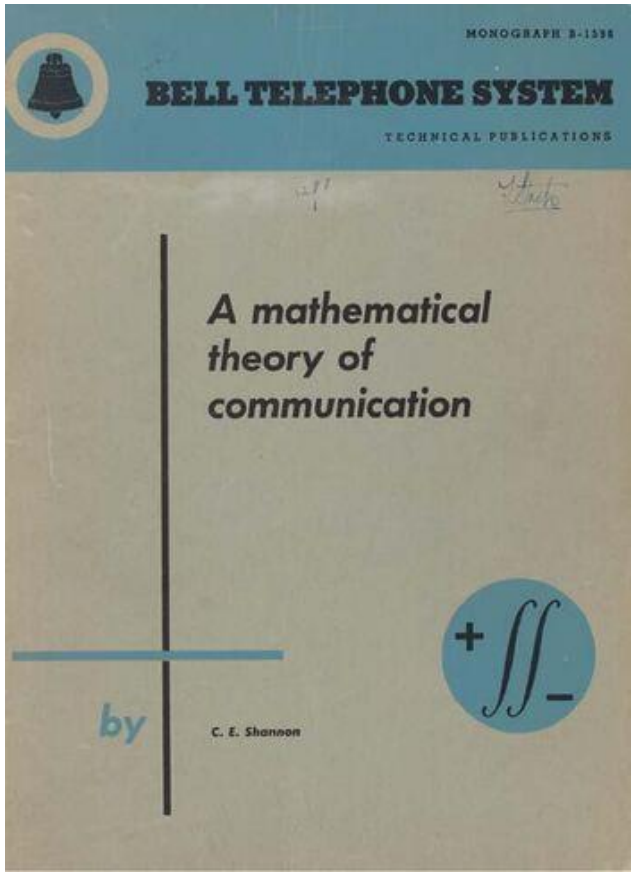
Outline: Geometric Information Theory (GIT)

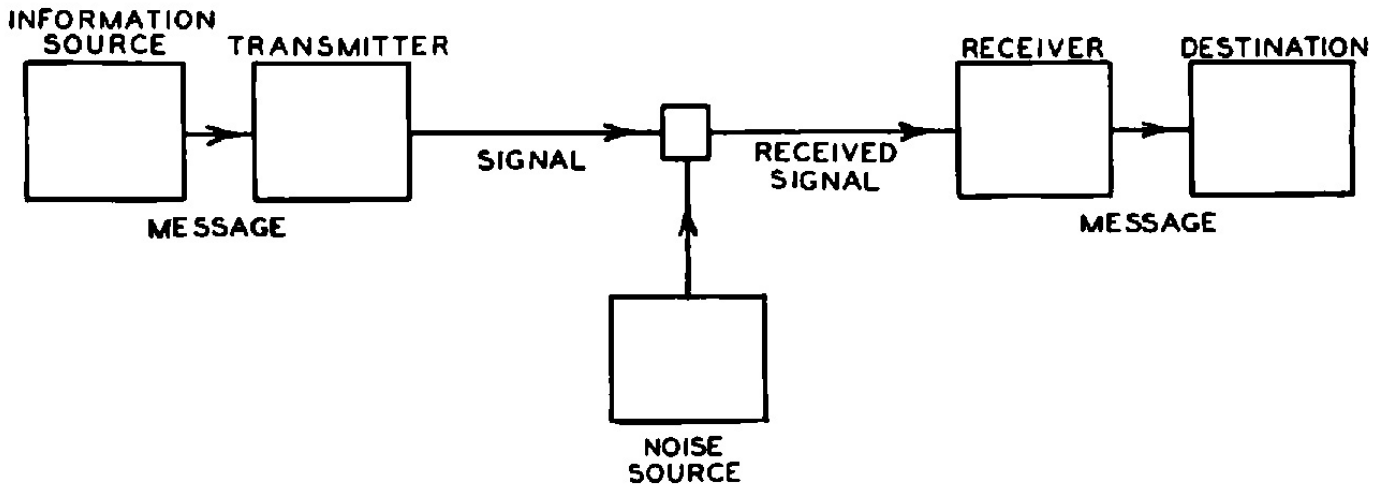
"There is Nothing More **Practical** Than A Good **Theory**"
Kurt Lewin

- Rich interplay of
Shannon information theory
with differential (information) geometry

- Applications to information sciences **(HUB)**

Breakthrough #1: Birth of information theory (1948)

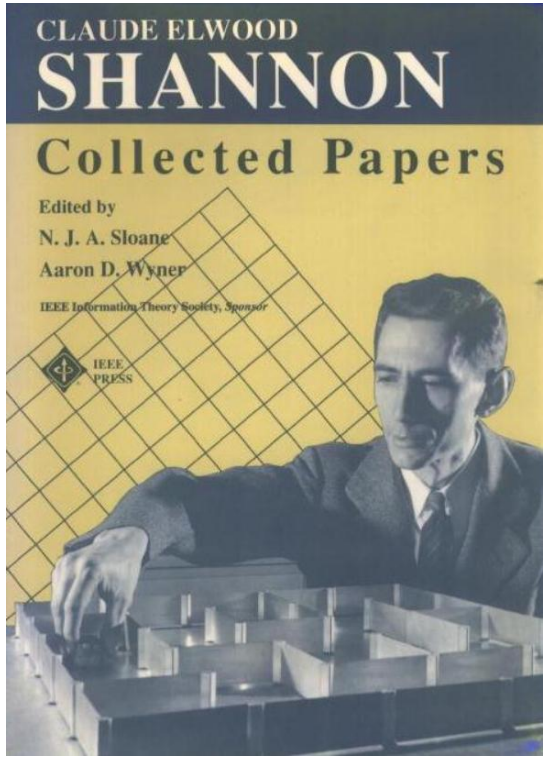


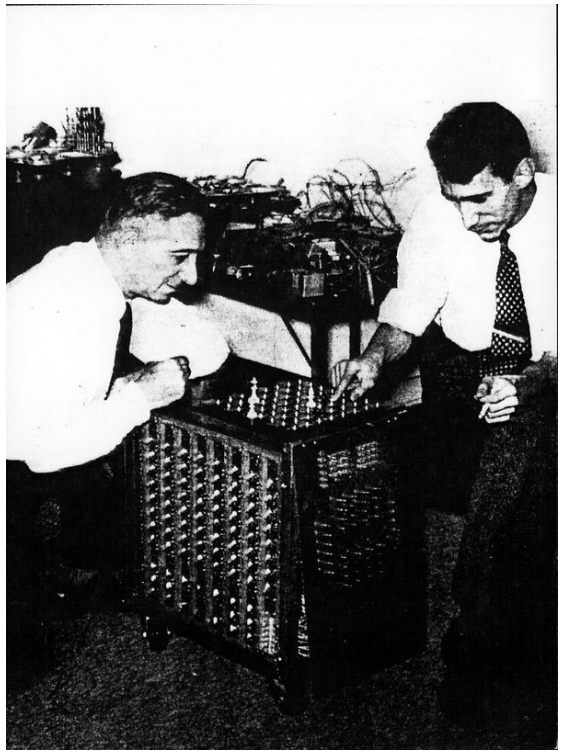


- **Modeling** of a message transmission problem
- **Revolutionize** telecommunication industry
- **Axiomatization** yielding **Shannon entropy** (H or S):
  … quantify uncertainty of random variables

We shall call $H = -\Sigma p_i \log p_i$ the entropy of the set of probabilities

Claude Shannon: A curious and playful mind





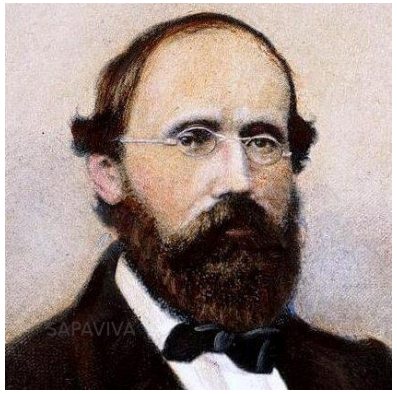
- Shannon's **source coding theorem** (Lossless compression)
- Shannon's **noisy channel coding theorem** (Reliable comm. for transmission rate<channel capacity)
- Shannon's **perfect secrecy theorem**
- Nyquist–Shannon **sampling theorem**
- Shannon–McMillan theorem: **Asymptotic Equipartition Property** (AEP)
- Etc.

https://neilsloane.com/doc/shannon.html

Breakthrough #2:

20th century revolution: Curved 4D Spacetime

Gauss
1827

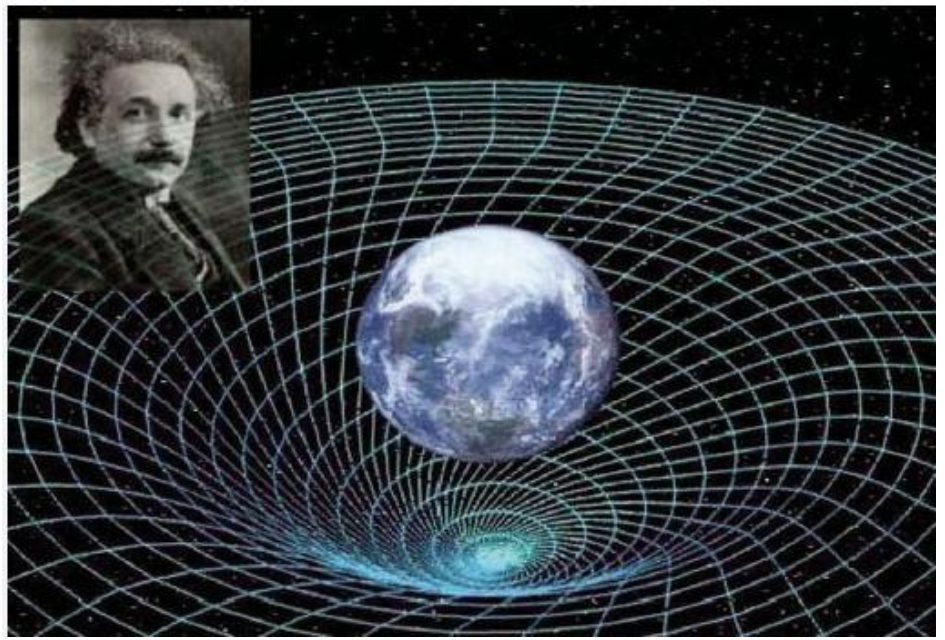


Riemann 1854

$$ds^2 = \sum_{i,j} g_{ij}\, dx^i\, dx^j$$

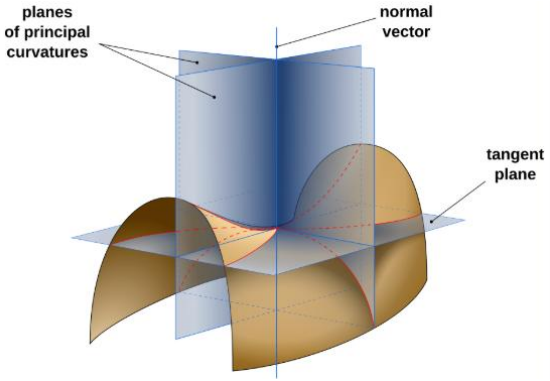
Length element

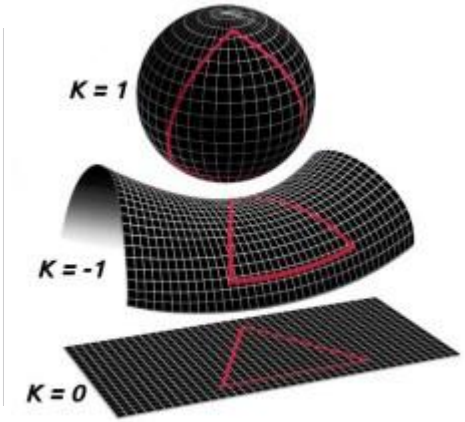
A first ``Killer'' application

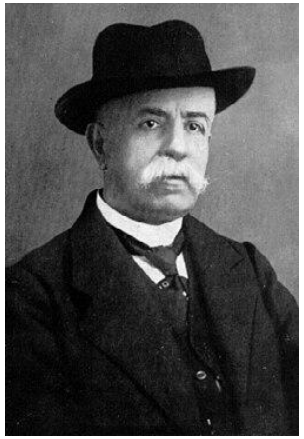


General Relativity 1915

Intrinsic curvature







Ricci-Curbastro

Tensor calculus

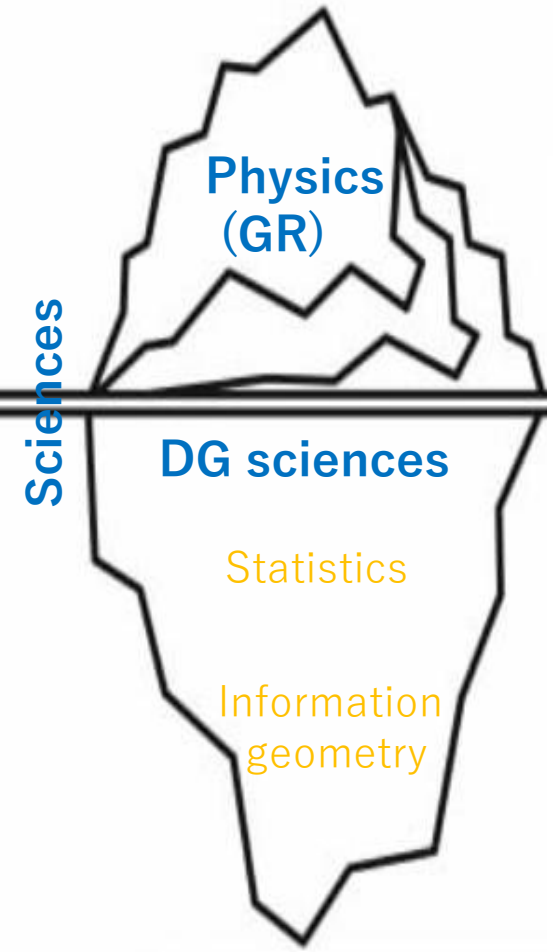
"Space-time tells matter how to move matter tells space-time how to curve."

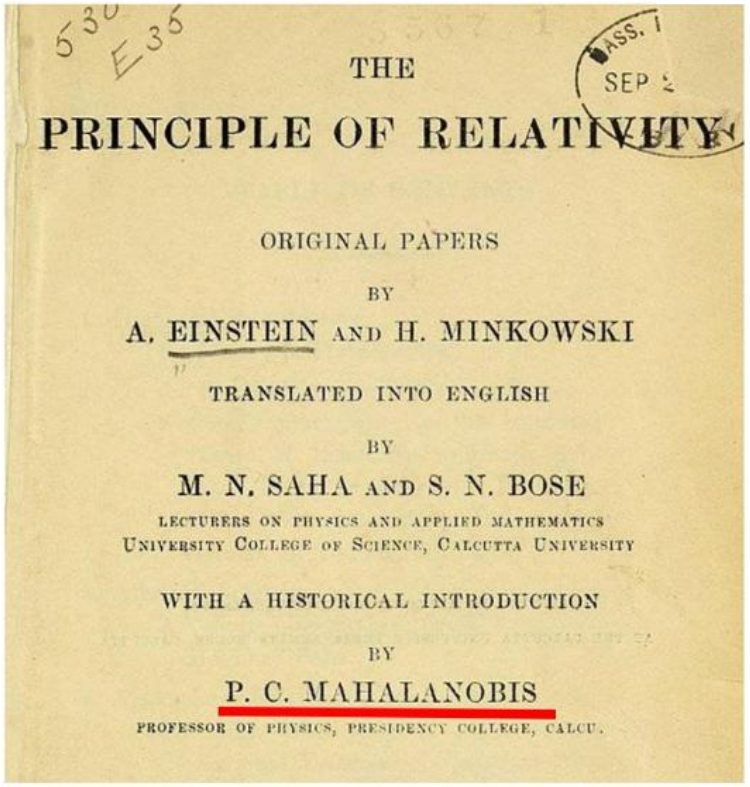
4D pseudo-Riemannian dynamic geometry

Einstein's iceberg: Differential geometry in Sciences

DG popularized by GR became a **scientific curiosity**…

$$\Delta[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] = \sqrt{(\mu_1-\mu_2)^\top \Sigma^{-1} (\mu_1-\mu_2)}$$





THE

PRINCIPLE OF RELATIVITY

ORIGINAL PAPERS

BY

A. EINSTEIN AND H. MINKOWSKI

TRANSLATED INTO ENGLISH

BY

M. N. SAHA AND S. N. BOSE

LECTURERS ON PHYSICS AND APPLIED MATHEMATICS
UNIVERSITY COLLEGE OF SCIENCE, CALCUTTA UNIVERSITY

WITH A HISTORICAL INTRODUCTION

BY

P. C. MAHALANOBIS

PROFESSOR OF PHYSICS, PRESIDENCY COLLEGE, CALCU.

This we may call a generalized statistical field in which the values of $\alpha_{\mu\nu}$'s will vary from point to point in the field. In such a field we may still continue to use

$$P \,.\, D_1^2 = a_{\mu\nu} \,.\, (da)^\mu \,.\, (da)^\nu \quad .. \quad .. \quad .. \quad (5{\cdot}0)$$

as the expression for the line element.

ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

1936

Equation (2·5) can then be written in the form

$$P \,.\, \Delta^2 = \alpha_{\mu\nu} \,.\, (d\alpha)^\mu \,.\, (d\alpha)^\nu \quad .. \quad .. \quad .. \quad (2{\cdot}7)$$

Comparing with the formula for ds^2

$$ds^2 = g_{\mu\nu} \,.\, (dx)^\mu \,.\, (dx)^\nu \quad .. \quad .. \quad .. \quad (2{\cdot}71)$$

we notice that $P \,.\, \Delta^2$ in statistics is the exact analogue of ds^2 in the restricted theory of relativity.

This merely implies that a consistent geometrical representation is possible in both cases. It is possible, however, to use this formal equivalence to establish an exact correspondence between results in the two subjects.

3. We see therefore that a statistical field in which the dispersion is same everywhere (values of $\alpha_{\mu\nu}$'s same at all points of the field and independent of mean values) corresponds to the physical field in the restricted theory of relativity ($g_{\mu\nu}$'s same everywhere and independent of co-ordinate values). In fact $\alpha_{\mu\nu}$'s play the same part in statistics as $g_{\mu\nu}$'s in the theory of relativity, and all the results involving ds^2 can be formally obtained from the results for a statistical field in which the dispersion is constant by putting

Fisher information matrix $I_X(\theta)$

- parametric family of laws indexed by D parameters
- **Fisher Information Matrix** (FIM) = covariance matrix of the **score**

$$X \sim p_\theta(x)$$

$$s_X(\theta) = \nabla_\theta \log p_\theta(x)$$

$$I_X(\theta) = \mathrm{Cov}[s_\theta]$$

Regularity conditions

$$I_X(\theta) = E_\theta \left[\nabla \log p_\theta(x) (\nabla \log p_\theta(x))^\top \right]$$

$$I_X(\theta) = -E_{p_\theta} \left[\nabla^2 l_x(\theta) \right] = - \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} l_x(\theta) l_x(\theta) \right]_{ij}$$

- FIM is symmetric and positive semi-definite (could be undefined too!)
- Statistical model is said **regular** when positive-definite:

  ex.: Gaussian models are regular but Gaussian mixture models are not.

Cramér–Rao lower bound:

Covariance of **unbiased estimator** of iid random vector of n observations:

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta)^{-1}$$

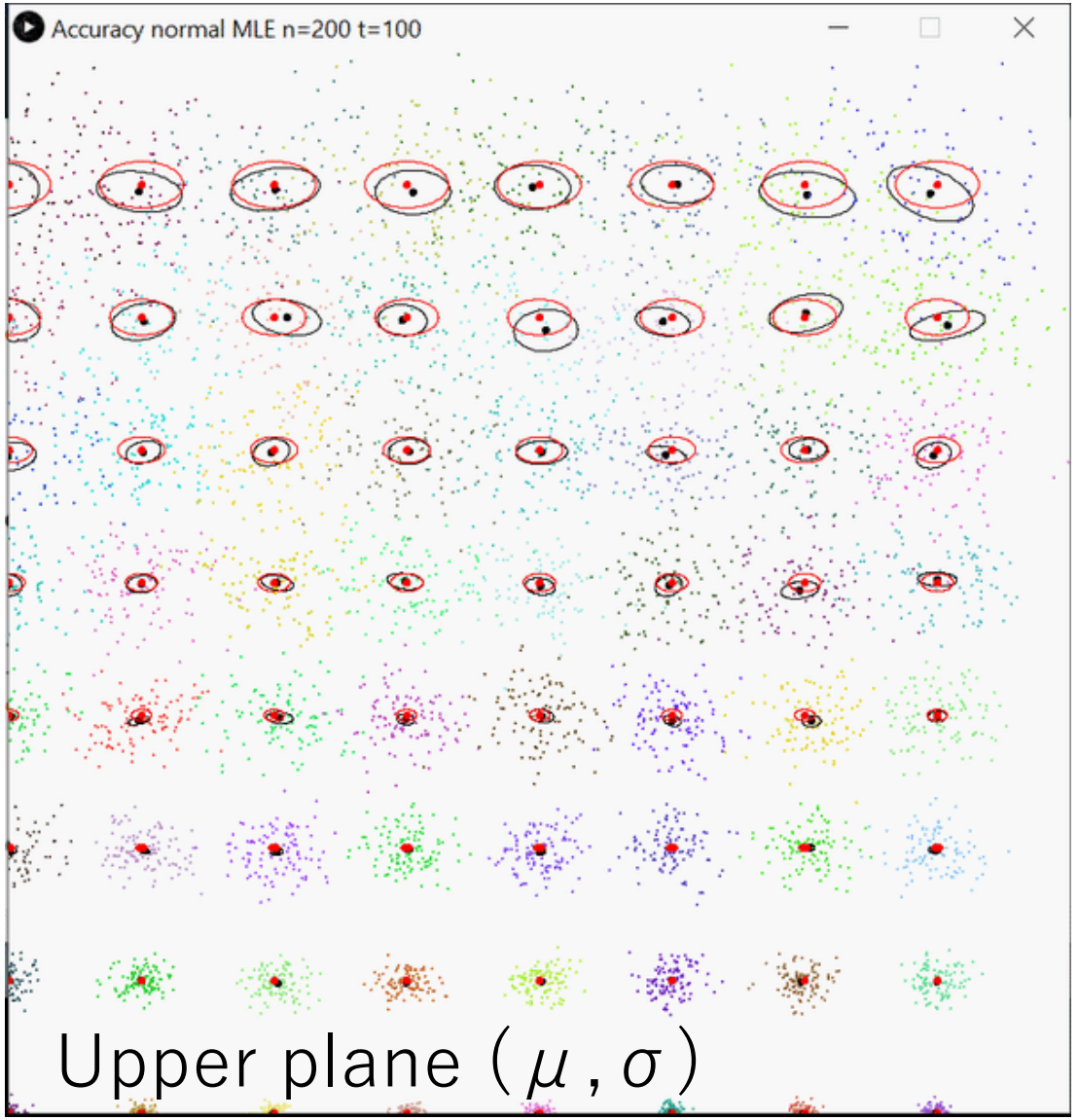
Estimator is **efficient** is bound attained

Bound attained only for **exponential families**

$$p_\theta(x) = \exp\left(\langle \theta, t(x) \rangle - F(\theta)\right)$$

EFs: normal, Beta, Poisson, Wishart, etc.
Non EFs: Gaussian mixtures, uniform, Cauchy, etc

CRLB improved by Chapman–Robbins bound



$$\mathrm{Cov}[\hat{\theta}_n] \quad \text{vs} \quad I(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

SPACES OF STATISTICAL PARAMETERS

By Harold Hotelling, Stanford University.

[Abstract]

Here $g_{\alpha\beta}$ is the mathematical expectation of

$$\frac{\partial^2 L}{\partial p_\alpha \, \partial p_\beta},$$

and is a covariant tensor of second order under transformations $p'_\alpha = \phi_\alpha(p_1, \ldots, p_k)$—though of course the second derivative is not itself a tensor.

This tensor property suggests that

$$g_{\alpha\beta}\, dp^\alpha \, dp^\beta$$

be taken as distance element in a space of coordinates $p^1, \ldots, p^k$. Indeed a considerable amount of differential geometry carries over immediately to give novel statistical conclusions. It should be said at once that these spaces are not flat, but are curved in a manner depending on the initial population distributions.

1930

0804.2996

Information and the Accuracy Attainable in the Estimation of Statistical Parameters

C Radhakrishna Rao

(Communicated by Mr. R C Bose–Received August 23, 1945

The population space

Let the distribution of a certain number of characters in a population be characterised by the probability differential

$$\phi(x, \theta_1, \ldots, \theta_q)dv. \tag{6.1}$$

The quantities $\theta_1, \theta_2, \ldots, \theta_q$ are called population parameters. Given the functional form in x's as in (6.1) which determines the type of the distribution function, we can generate different populations by varying $\theta_1, \theta_2, \ldots, \theta_q$. If these quantities are represented in a space of q dimensions, then a population may be identified by a point in this space which may be defined as the population space (P.S).

Let $\theta_1, \theta_2, \ldots, \theta_q$ and $\theta_1 + d\theta_1, \theta_2 + d\theta_2, \ldots, \theta_q + d\theta_q$ be two contiguous points in (P.S). At any assigned value of the characters of the populations corresponding to these contiguous points, the probability densities differ by

$$d\phi(\theta_1, \theta_2, \ldots, \theta_q) \tag{6.2}$$

retaining only first order differentials. It is a matter of importance to consider the relative discrepancy $d\phi/\phi$ rather than the actual discrepancy. The distribution of this quantity over the x's summarises the consequences of replacing $\theta_1, \theta_2, \ldots, \theta_q$ by $\theta_1 + d\theta_1, \ldots, \theta_q + d\theta_q$. The variance of this distribution or the expectation of the square of this relative discrepancy comes out as the positive definite quadrate differential form

$$ds^2 = \sum\sum g_{ij} d\theta_i d\theta_j, \tag{6.3}$$

where

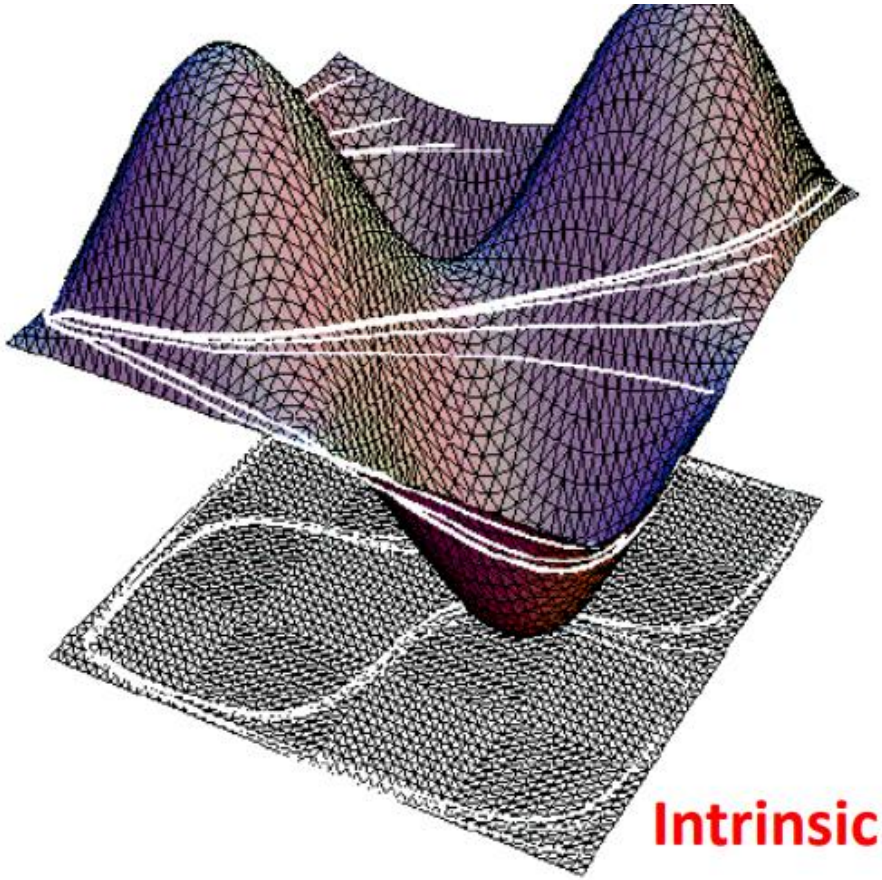
$$g_{ij} = E\left(\frac{1}{\phi}\frac{\partial \phi}{\partial \theta_i}\right)\left(\frac{1}{\phi}\frac{\partial \phi}{\partial \theta_j}\right). \tag{6.4}$$

1945

1301.3578

Intrinsic vs extrinsic Riemannian geometry

Isometric embedding:



Extrinsic geometry

Intrinsic geometry

Riemannian manifold of **dimension D** can always be thought as a **submanifold of Euclidean space** in **2D dimensions** by Whitney isometric embedding theorem.

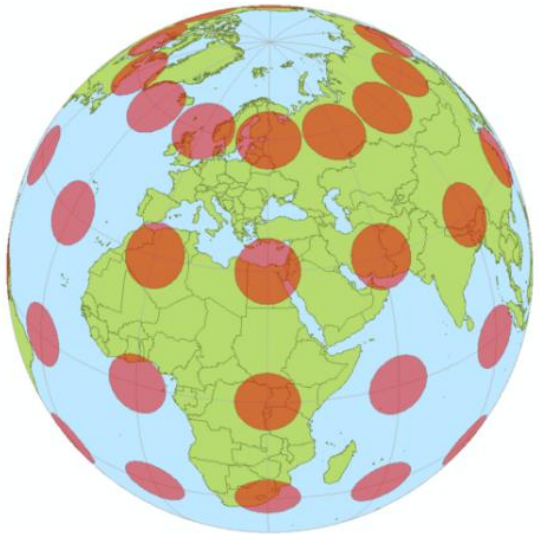
Riemannian geodesic distance

- Geodesics are *locally* **shortest path lengths** on a surface.
- Geodesics $\gamma(s)$ are parameterized by **arc length** s, not necessarily unique.

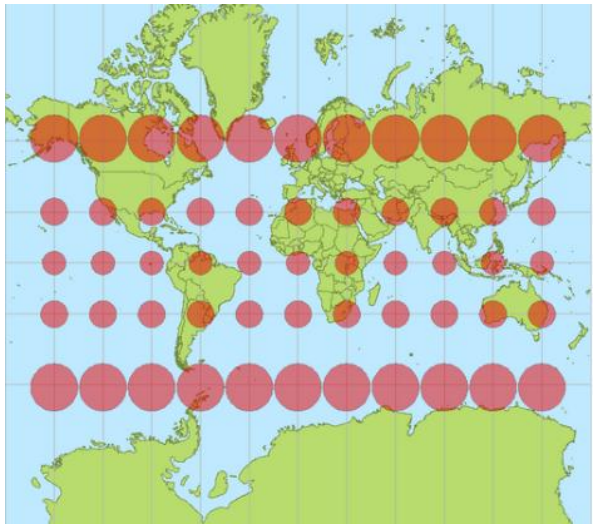
(soon a better proper definition)

Locally shortest path but not globally

Geodesics = local shortest paths



Extrinsic Euclidean view of S^2 in R^3



Intrinsic map/chart view with **Tissot indicatrix**

Fisher-Rao geodesic distance: First applications

Distance in tests of significance and classification

We apply the metric (7.1) to find the distance between two normal populations defined by (m_1, σ_1) and (m_2, σ_2) the distribution being of the type

$$\phi(x, m, \sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp -\frac{1}{2}\frac{(x-m)^2}{\sigma^2}. \tag{7.5}$$

The quantities g_{ij} defined above have the values

$$g_{11} = 1/\sigma^2, \quad g_{12} = 0, \quad g_{22} = 2/\sigma^2, \tag{7.6}$$

so that the element of length is obtained from

$$ds^2 = \frac{(dm)^2}{\sigma^2} + \frac{2}{\sigma^2}(d\sigma)^2. \tag{7.7}$$

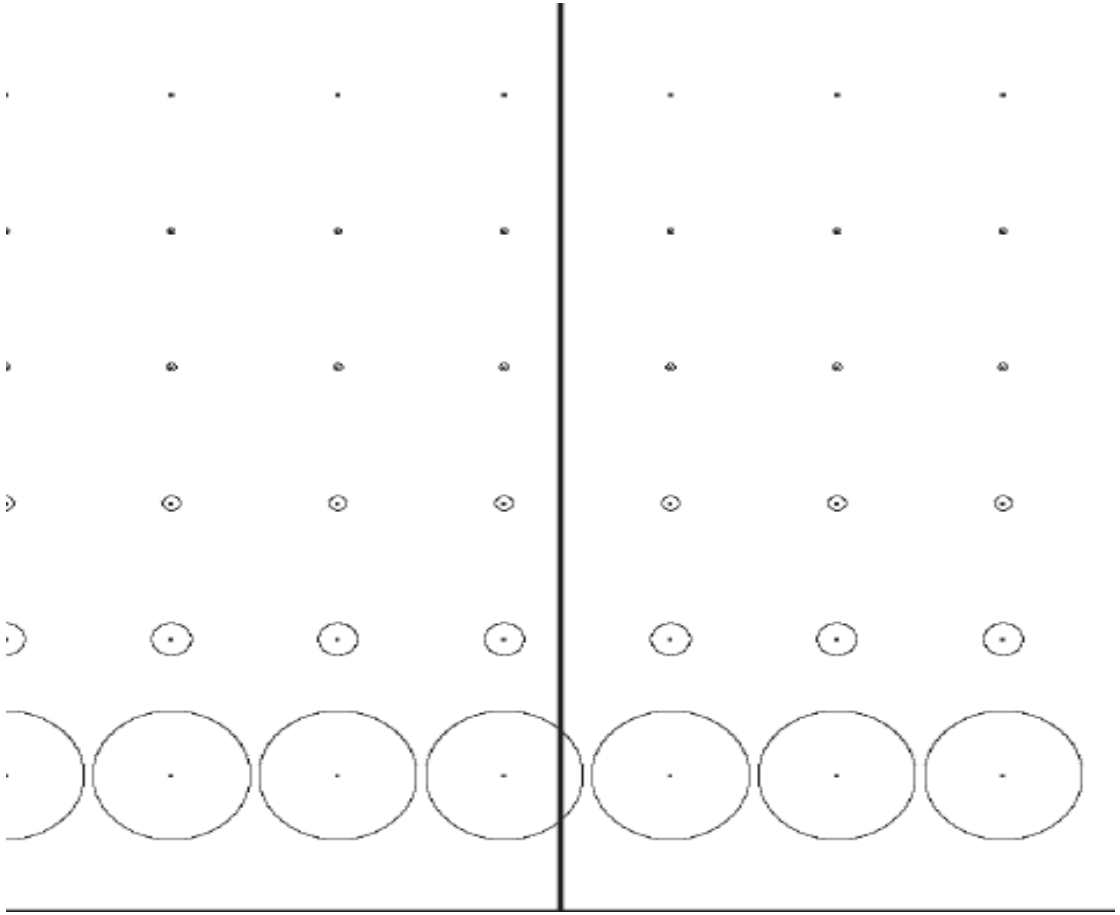
If $m_1 \neq m_2$ and $\sigma_1 \neq \sigma_2$ then the distance comes out as

$$D_{AB} = \sqrt{2} \log \frac{\tan \theta_1/2}{\tan \theta_2/2} \tag{7.8}$$

where

$$\theta_i = \sin^{-1} \sigma_i/\beta \text{ and } \beta^2 = \sigma_1^2 + [(m_1 - m_2)^2 + 2(\sigma_2^2 - \sigma_1^2)]^2/8(m_1 - m_2)^2. \tag{7.9}$$

Fisher information metric of **normals** is **deformed Poincaré hyperbolic metric**



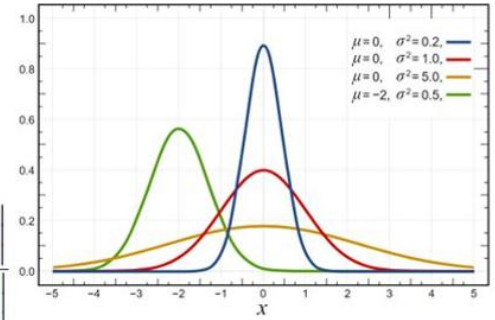
Poincaré Riemannian upper plane:
Intrinsic vs extrinsic view (partial in 3D, pseudosphere)

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \lambda = (\mu, \sigma) \in \mathbb{H} \right\}$$

$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

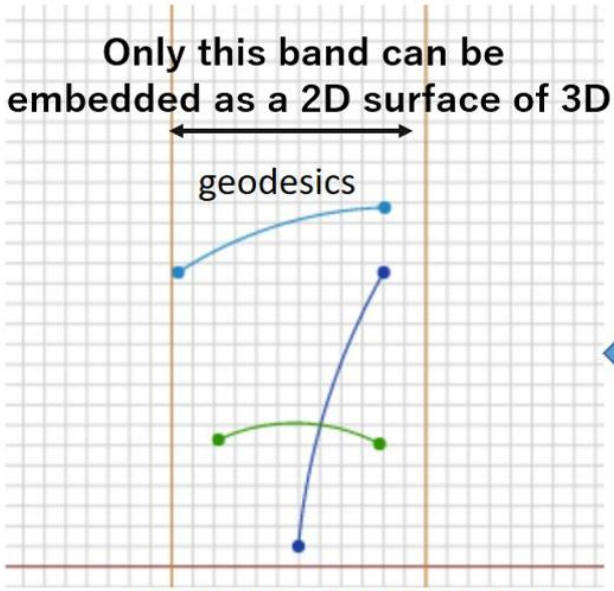
Fisher-Rao geodesic distance:

$$D_{\mathrm{Rao}}\left[p_{\mu_1,\sigma_1}, p_{\mu_2,\sigma_2}\right] = \sqrt{2}\ln \frac{\left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right) \right\| + \left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right) \right\|}{\left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right) \right\| - \left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right) \right\|}$$



$$ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$

Constant curvature -1/2
(= hyperbolic manifold)

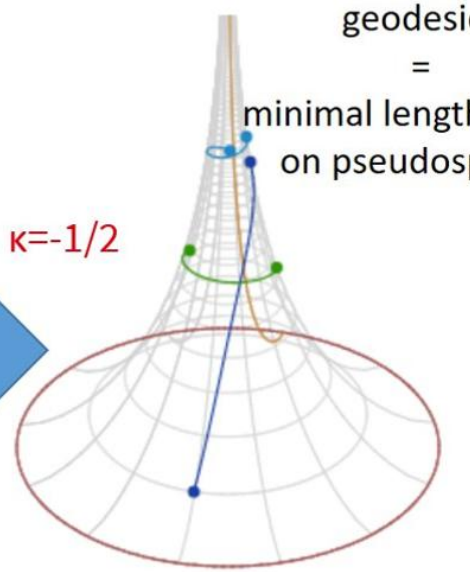


Constant Gaussian
negative curvature

Isometric embedding
(partial/periodic)

κ=-1/2

geodesics
=
minimal length curves
on pseudosphere



Intrinsic view
(in GR, space expands!)

Extrinsic view
Hilbert theorem (1901): partial embedding

Invariance of Fisher-Rao distance by reparam.

If we parameterize Gaussians by (μ, σ^2) or $(\mu, \log(\sigma))$ instead of (μ, σ), it should *not change* the distance nor the interpolating paths called **geodesics**

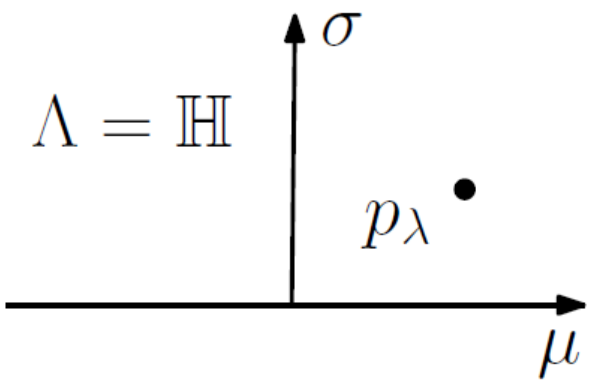
Same family of Gaussians but different parameter spaces

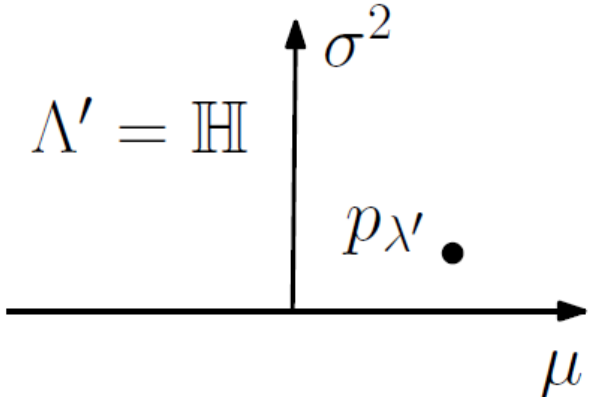
$$D[p_{\lambda_1}, p_{\lambda_2}] = D[p_{\lambda'_1}, p_{\lambda'_2}] = D[p_{\lambda''_1}, p_{\lambda''_2}]$$

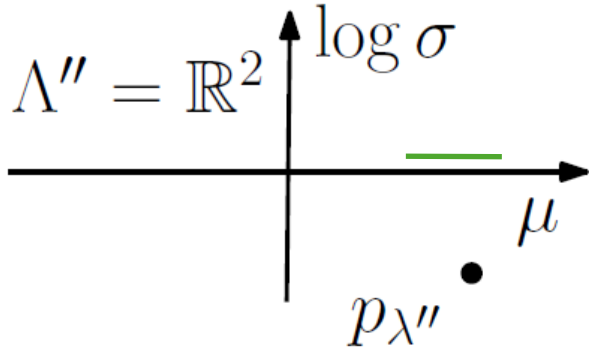
$$\gamma_{p_{\lambda_1}, p_{\lambda_2}}(t) = \gamma_{p_{\lambda'_1}, p_{\lambda'_2}}(t) = \gamma_{p_{\lambda''_1}, p_{\lambda''_2}}(t), \forall t \in [0,1]$$

Fisher matrix is covariant
but length element is invariant

arc length parameterization







$$\mathcal{P} = \left\{ p_{\lambda''}(x) = \frac{1}{\sqrt{2\pi}\exp(\lambda''_2)} \exp\left(-\frac{(x-\lambda''_1)^2}{2\exp(2\lambda''_2)}\right), \quad \lambda'' = (\mu, \log\sigma) \in \mathbb{R}^2 \right\}$$

Covariance of FI Matrix and invariance of Fisher metric

Consider two different parameterizations of a statistical model:

$$\mathcal{P} = \{p_\theta \ : \ \theta \in \Theta\} = \{p_\eta \ : \ \eta \in H\}$$

Covariance transformation of the FIM under reparameterization

$$\eta(\theta) \Leftrightarrow \theta(\eta) \qquad I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial \theta_i}{\partial \eta_j}\right]$$

Invariance of length element…

$$\mathrm{d}s^2(\theta, \mathrm{d}\theta) = \mathrm{d}\theta^\top I_\theta(\theta)\mathrm{d}\theta = \mathrm{d}s^2(\eta, \mathrm{d}\eta) = \mathrm{d}\eta^\top I_\eta(\eta)\mathrm{d}\eta \qquad \mathrm{d}s_\theta = \mathrm{d}s_\eta$$

…implies **Invariance** of Fisher-Rao element

$$\rho_{\mathrm{Rao}}(p_{\eta 1}, p_{\eta 2}) = \rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})$$

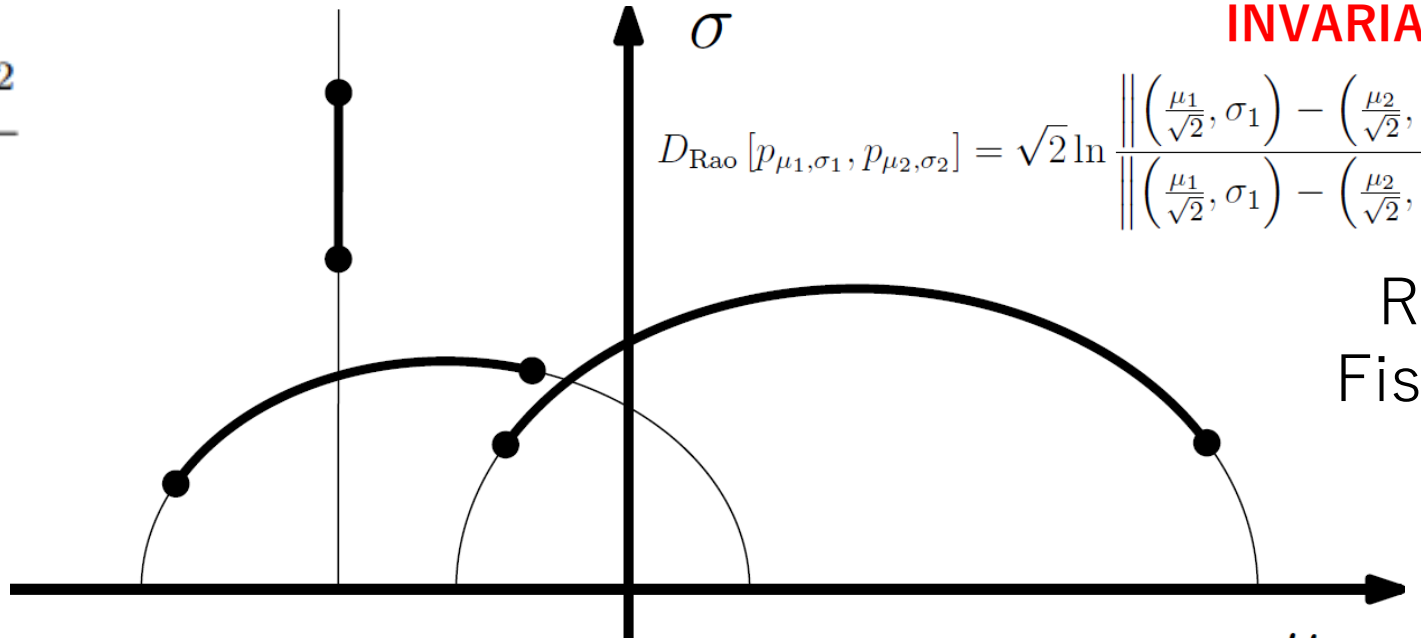
Fisher-Rao geometry: univariate normal distributions

INVARIANT

$$ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$

Fisher metric

stretched Poincaré half-plane

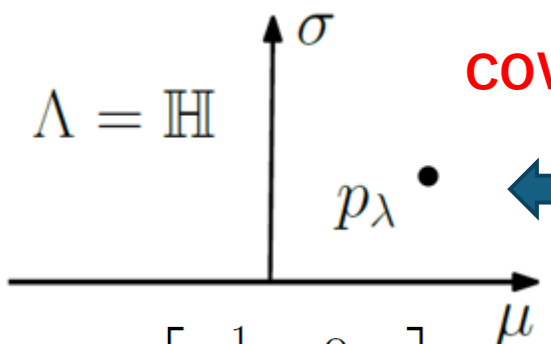


INVARIANT

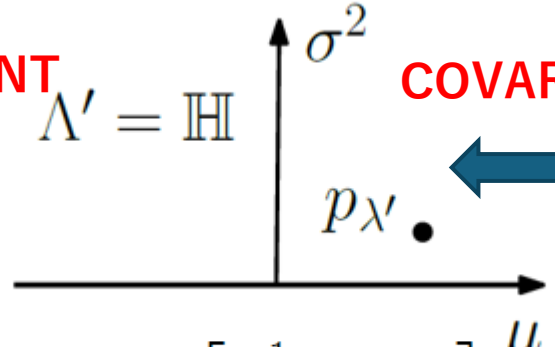
$$D_{\mathrm{Rao}}\left[p_{\mu_1,\sigma_1}, p_{\mu_2,\sigma_2}\right] = \sqrt{2}\ln \frac{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| + \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| - \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}$$

Rao's distance or Fisher-Rao distance

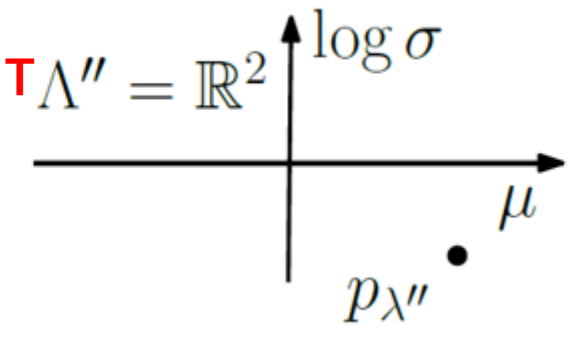
FIM and domains for various parameterizations



COVARIANT



COVARIANT



$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

$$I(\mu,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

$$I(\mu,\log\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2 \end{bmatrix}$$

In general, **location-scale families** yield a **hyperbolic** Fisher-Rao geometry

Classification using Fisher-Rao geodesic distance

Distance in tests of significance and classification

We apply the metric (7.1) to find the distance between two normal populations defined by (m_1, σ_1) and (m_2, σ_2) the distribution being of the type

$$\phi(x, m, \sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp -\frac{1}{2}\frac{(x-m)^2}{\sigma^2}. \tag{7.5}$$

The quantities g_{ij} defined above have the values

$$g_{11} = 1/\sigma^2, \quad g_{12} = 0, \quad g_{22} = 2/\sigma^2, \tag{7.6}$$

so that the element of length is obtained from

$$ds^2 = \frac{(dm)^2}{\sigma^2} + \frac{2}{\sigma^2}(d\sigma)^2. \tag{7.7}$$

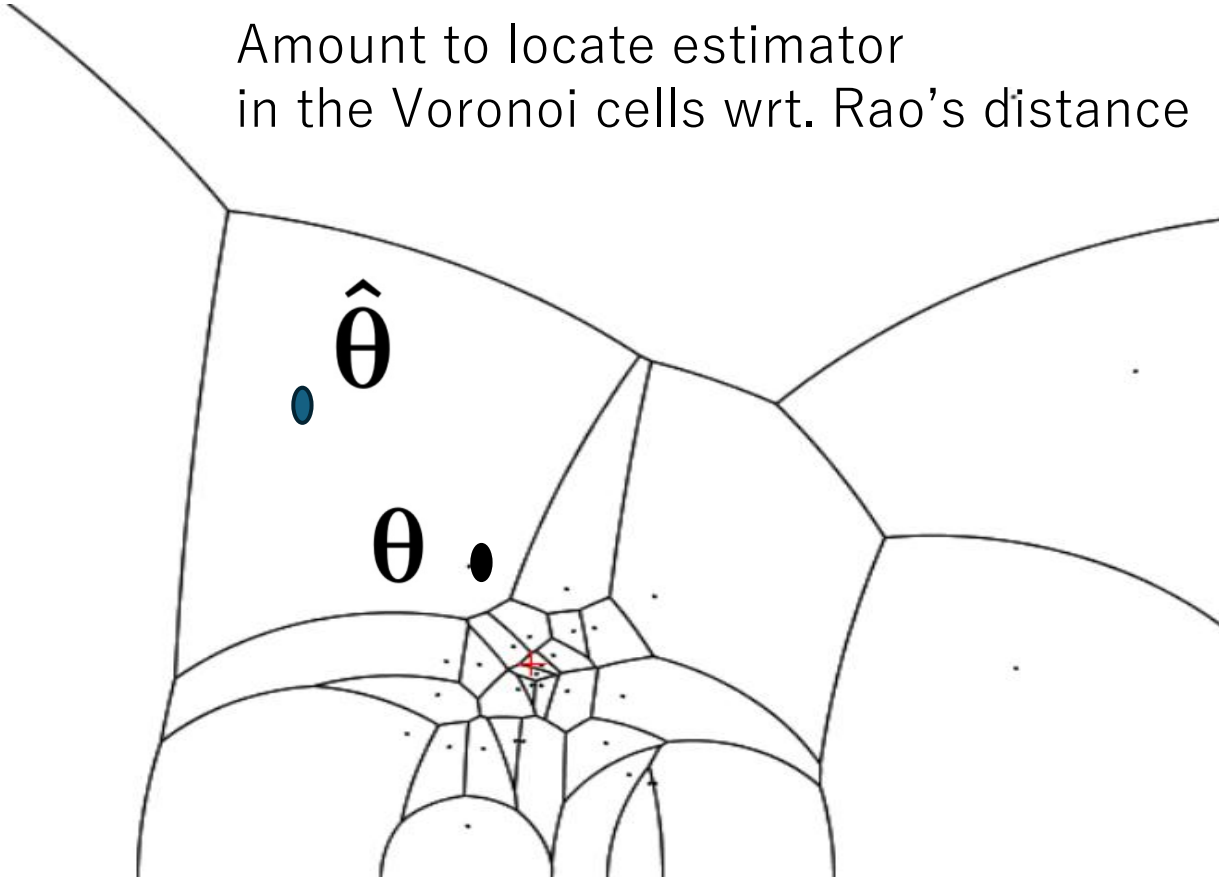
If $m_1 \neq m_2$ and $\sigma_1 \neq \sigma_2$ then the distance comes out as

$$D_{AB} = \sqrt{2} \log \frac{\tan \theta_1/2}{\tan \theta_2/2} \tag{7.8}$$

where

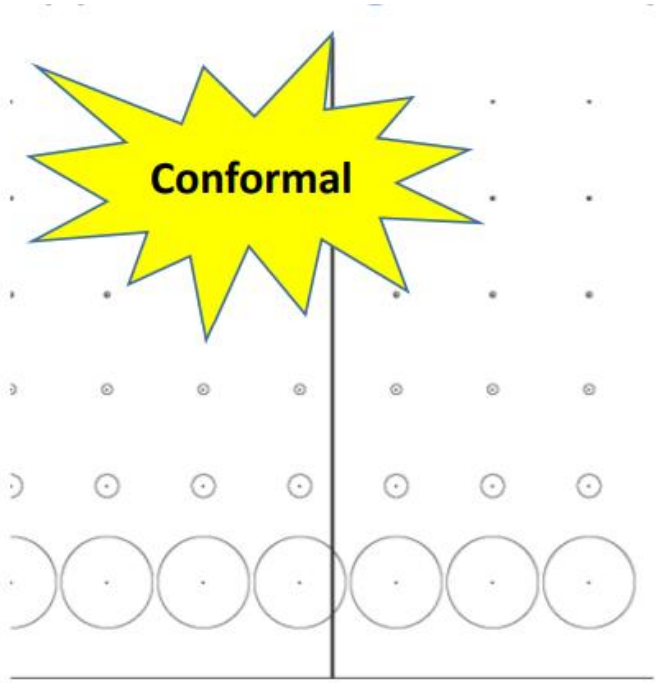
$$\theta_i = \sin^{-1} \sigma_i/\beta \text{ and } \beta^2 = \sigma_1^2 + [(m_1 - m_2)^2 + 2(\sigma_2^2 - \sigma_1^2)]^2/8(m_1 - m_2)^2. \tag{7.9}$$

Amount to locate estimator
in the Voronoi cells wrt. Rao's distance

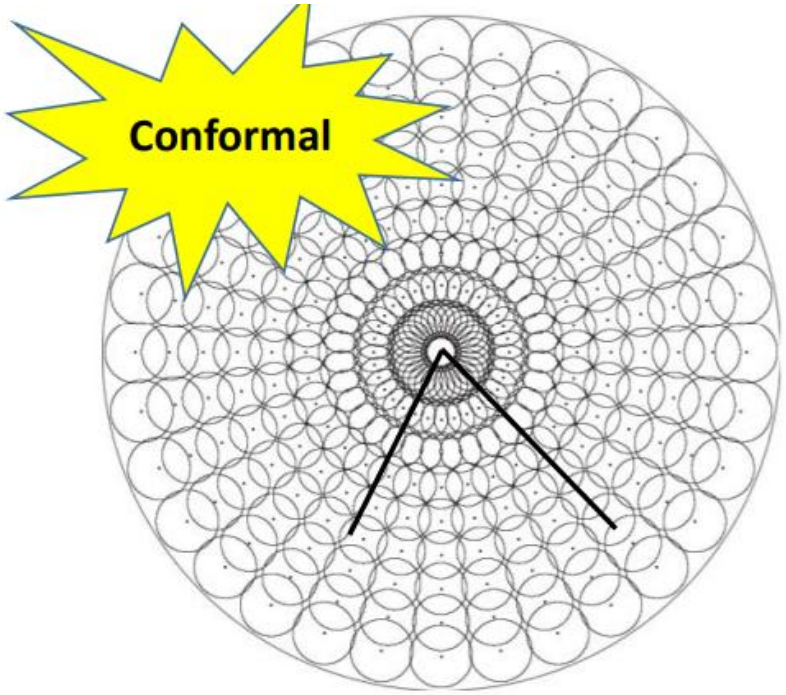


hyperbolic Voronoi diagram on the stretched Poincare hyperbolic metric on the upper plane,

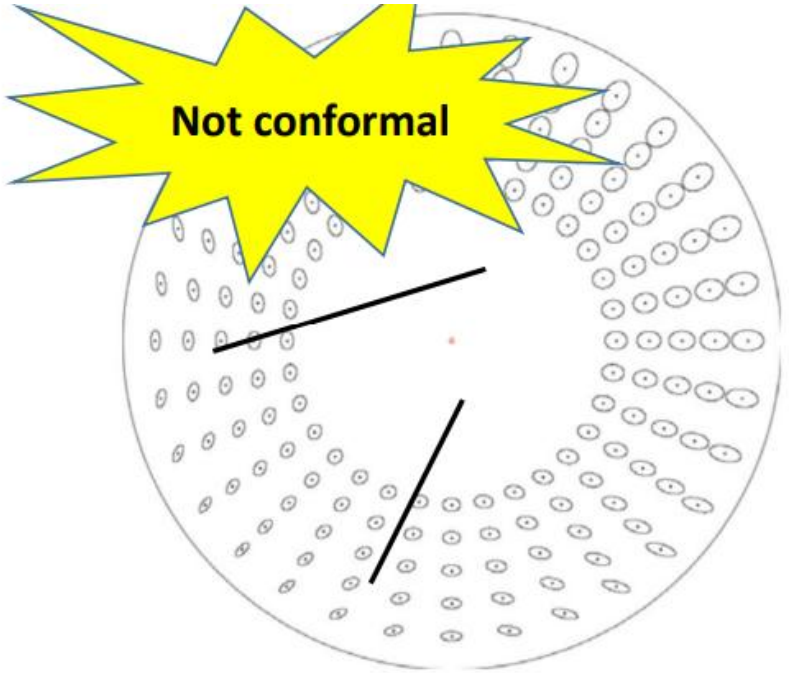
Tissot indicatrices and (non)conformal geometry



Upper Poincare plane
(conformal)



Poincare disk
(conformal)



Klein disk
(non-conformal)

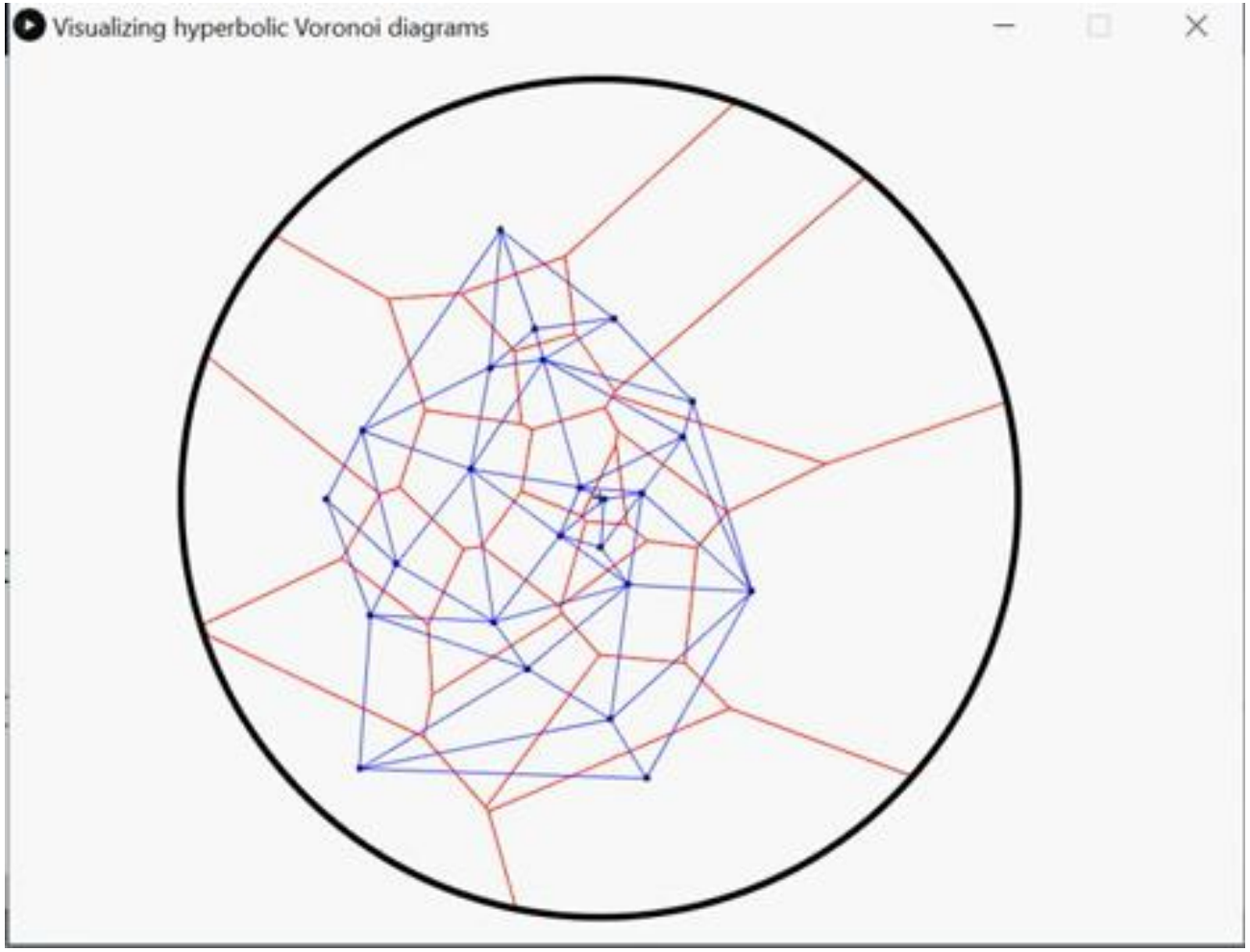
Metric tensor scaled by positive function: $\hat{g}_p = e^{f(p)} g$

Conformal: metric tensor a scalar-value function of the Euclidean metric tensor

In conformal geometry, we can measure angles without distortions

Fisher-Rao Voronoi diagrams of Gaussians

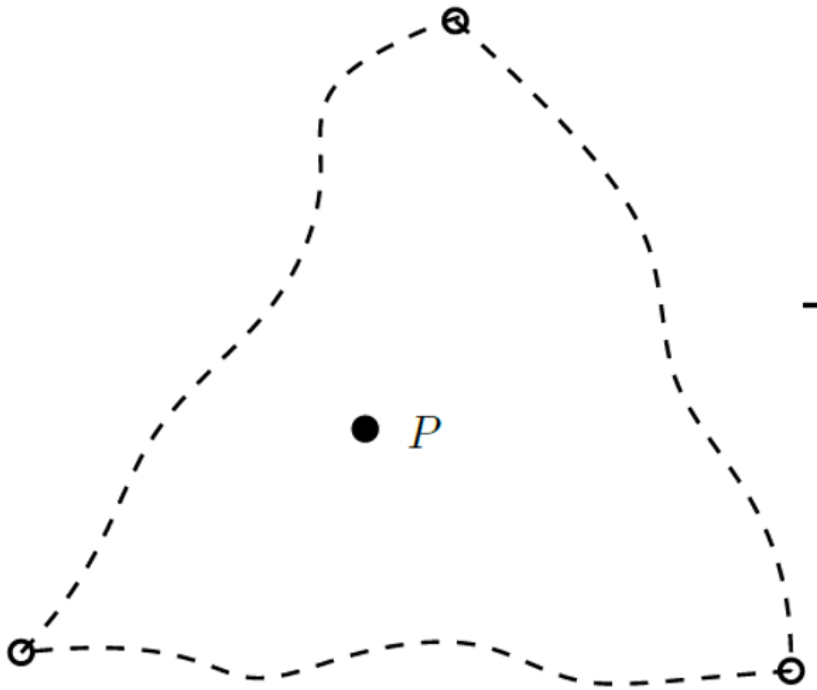
Compute fast+robust in **Klein model** and then convert in other models



0903.3287

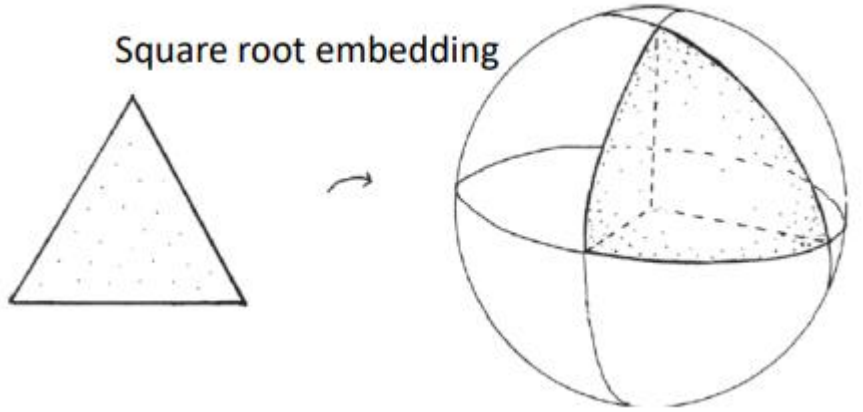
Embedding Fisher-Rao probability simplex manifold into Euclidean spherical orthant

Open simplex manifold Δ_2
Figure (mental image)



embedding

Square root embedding



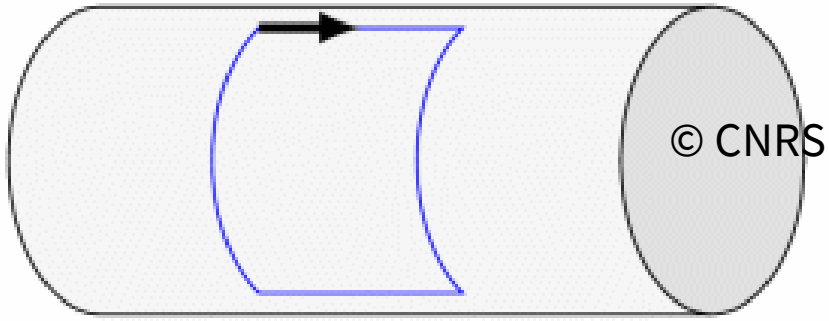
Fisher information metric:

$$g_{ij}(p) = \frac{\delta_{ij}}{\lambda_p^i} + \frac{1}{\lambda_p^0}.$$

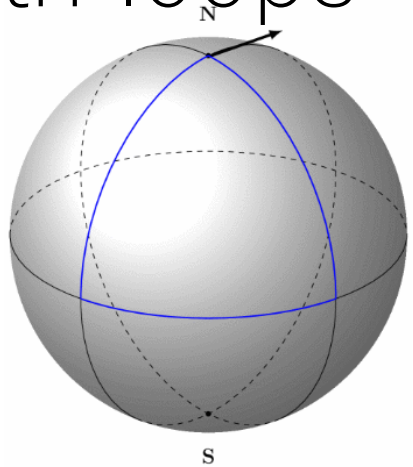
(Hotelling)-Fisher-Rao distance:

$$\rho_{\mathrm{FHR}}(p,q) = 2\arccos\left(\sum_{i=0}^{d}\sqrt{\lambda_p^i\lambda_q^i}\right)$$

Affine connection ∇ : Visualizing the curvature by the ∇-parallel transport along smooth loops



use embedding in R^3 and Euclidean connection



Élie Cartan

Cylinder is **flat**, 0 curvature Sphere has **positive constant** curvature

Geodesic equation is wrt. to affine connection (Christoffel symbols)

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,$$

2nd order differential equation
(geodesics = autoparallel curves)

Parameterization is **constant velocity**, i.e., no acceleration = t **modulo affine transformations**

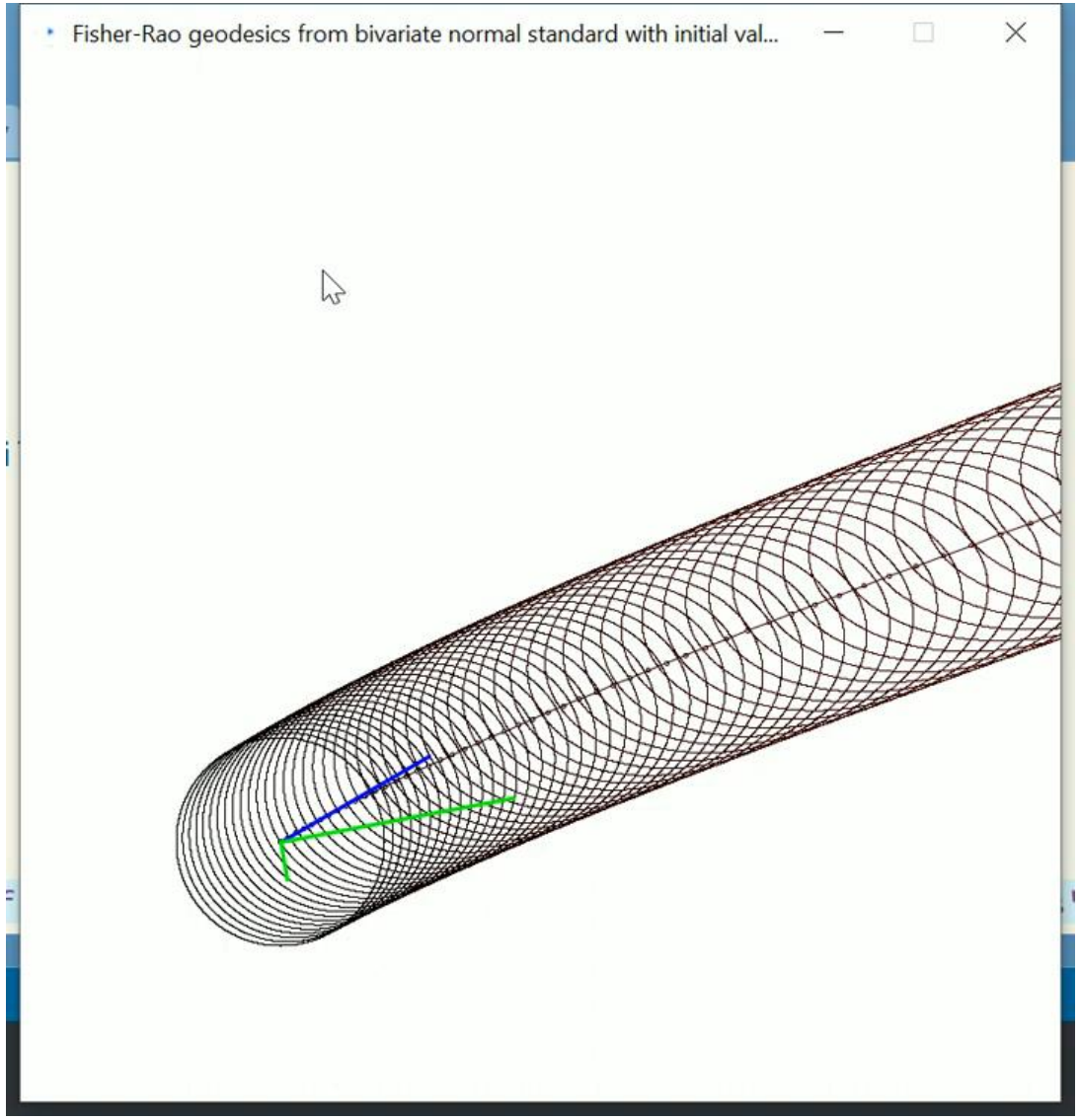
In Riemannian geometry, we derive the **Levi-Civita connection** from the **metric g**

$$\Gamma_{\mu\nu}^{\lambda} = \frac{1}{2} g^{\lambda\rho}\left(\partial_\mu g_{\nu\rho} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu} \right)$$

Parameterization is **unique** by arc length

Fisher-Rao distance requires (1) solving geodesics + (2) integrating length elements

Fisher-Rao geodesics: 2D Gaussians, initial value



$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,$$

Geodesic equation wrt. Levi-Civita connection

$$\begin{cases} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} & = & 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^{\top} - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} & = & 0. \end{cases}$$

Solve geodesic equation either with

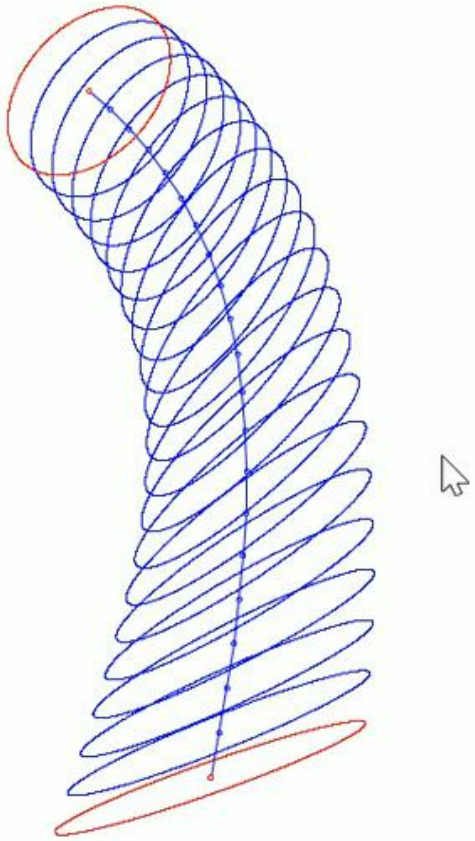
- **initial conditions**
  = **starting point** + **tangent vector**

- **boundary conditions**
  = **starting** + **ending points**

Blue vector is initial tangent vector for μ_0

Green vectors are the 2 eigenvectors of the initial tangent vector for Σ_0, symmetric matrix

Fisher-Rao geodesics 2D Gaussians: boundary conditions



Geodesics are defined wrt. to a connection

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,$$

$$\begin{cases} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} &= 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^{\top} - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} &= 0. \end{cases}$$

Red ellipsoids are the boundary conditions:
That is bivariate normal distributions
(μ_0, Σ_0) and (μ_1, Σ_1)

Multivariate normal Fisher-Rao **geodesics in closed-form**
But ***no known formula for Fisher-Rao distance*** for MVNs!

A simple approximation method for the Fisher–Rao distance between multivariate normal distributions, Entropy 25.4 (2023)

The Annals of Probability
1975, Vol. 3, No. 1, 146–158

KL projections, KL Pythagoras' thm.
KL parallelogram law, etc.

I-DIVERGENCE GEOMETRY OF PROBABILITY DISTRIBUTIONS AND MINIMIZATION PROBLEMS

BY I. CSISZÁR

1975

Mathematical Institute of the Hungarian Academy of Sciences

Some geometric properties of PD's are established, Kullback's *I*-divergence playing the role of squared Euclidean distance. The minimum discrimination information problem is viewed as that of projecting a PD onto a convex set of PD's and useful existence theorems for and characterizations of the minimizing PD are arrived at. A natural generalization of known iterative algorithms converging to the minimizing PD in special situations is given; even for those special cases, our convergence proof is more generally valid than those previously published. As corollaries of independent interest, generalizations of known results on the existence of PD's or nonnegative matrices of a certain form are obtained. The Lagrange multiplier technique is not used.

Dual information geometry

and the special case of

Bregman/Hessian/Shannon manifolds

Amari & Nagaoka: pair of affine e/m connections

- Historically, built the e-connection (exponential, α=1) and m-connection (mixture, α=-1) for statistical models

Log-likelihood $\ell(p_\xi)(x) = \ln p_\xi(x)$

e-connection $\Gamma^{(1)}_{ij,k}(\xi) = g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = E_\xi[(\partial_i\partial_j\ell)\,(\partial_k\ell)].$

m-connection $g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(-1)}_{ij,k} = E_\xi[(\partial_i\partial_j\ell + \partial_i\ell\,\partial_j\ell)\,(\partial_k\ell)]$

Meaning averaging Christoffel symbols yield Levi-Civita Fisher g connection

Dual connections with respect to the Fisher information (Riemannian) metric

$$\frac{\nabla + \nabla^*}{2} = {}^g\nabla$$

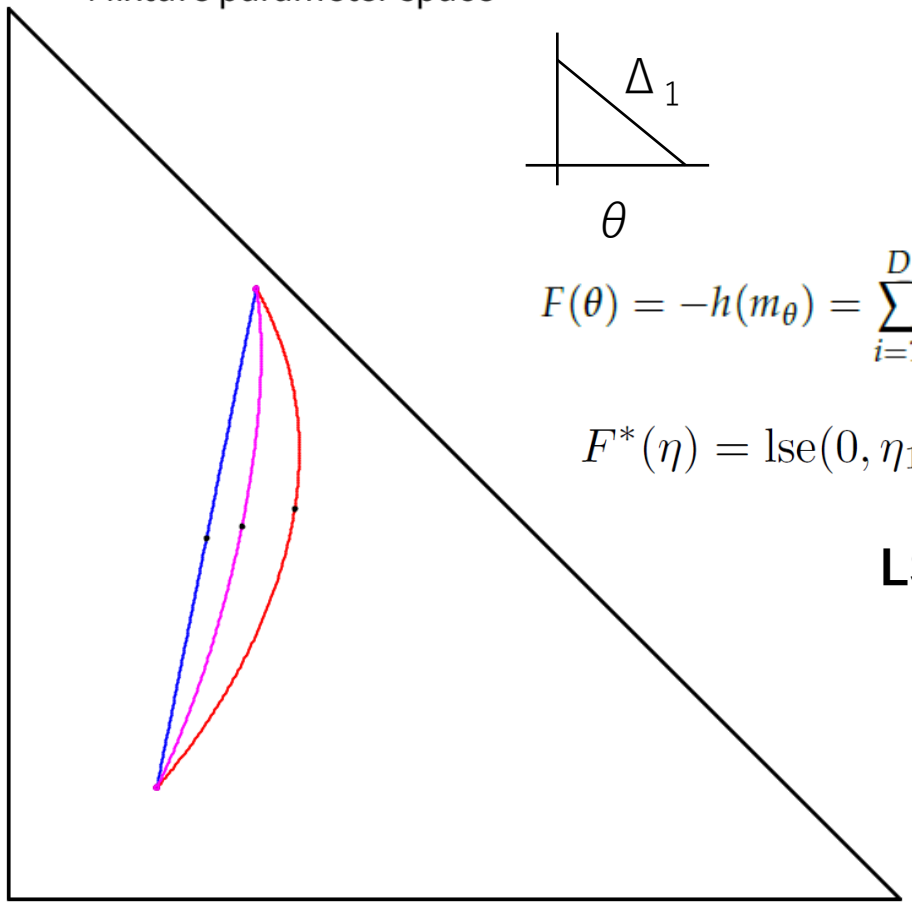
A connection ∇ is **flat** if there exists a coordinate system θ such that all Christoffel symbols vanish: $\Gamma(\theta)=0$.

∇-geodesic solves trivially as **line segments**

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p}\Gamma^k_{ij}\frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0$$

Dual geodesics and Fisher-Rao geodesics on the categorical distribution manifold

Mixture parameter space

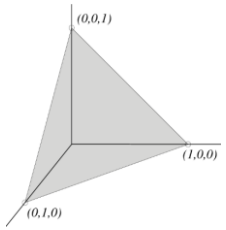


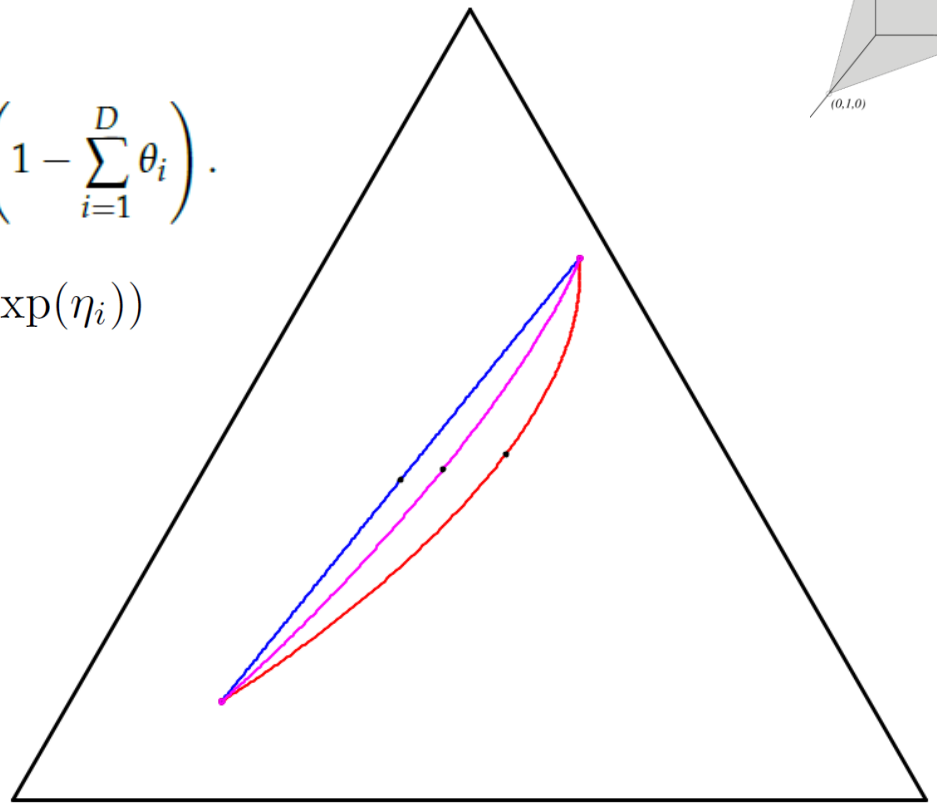
$$F(\theta) = -h(m_\theta) = \sum_{i=1}^{D} \theta_i \log \theta_i + \left(1 - \sum_{i=1}^{D} \theta_i\right) \log \left(1 - \sum_{i=1}^{D} \theta_i\right).$$

$$F^*(\eta) = \mathrm{lse}(0, \eta_1, \ldots, \eta_D) = \log(1 + \sum_{i=1}^{n} \exp(\eta_i))$$

LSE_0^+ is strictly convex

Probability simplex/Categorical manifold





Coordinate chart

Exponential ∇-geodesic
Mixture ∇^*-geodesic
Fisher-Rao ∇^g-geodesic (Levi-Civita)

Embedded manifold

Example: dually flat space of multivariate normals

(M,g, ∇ , ∇*)

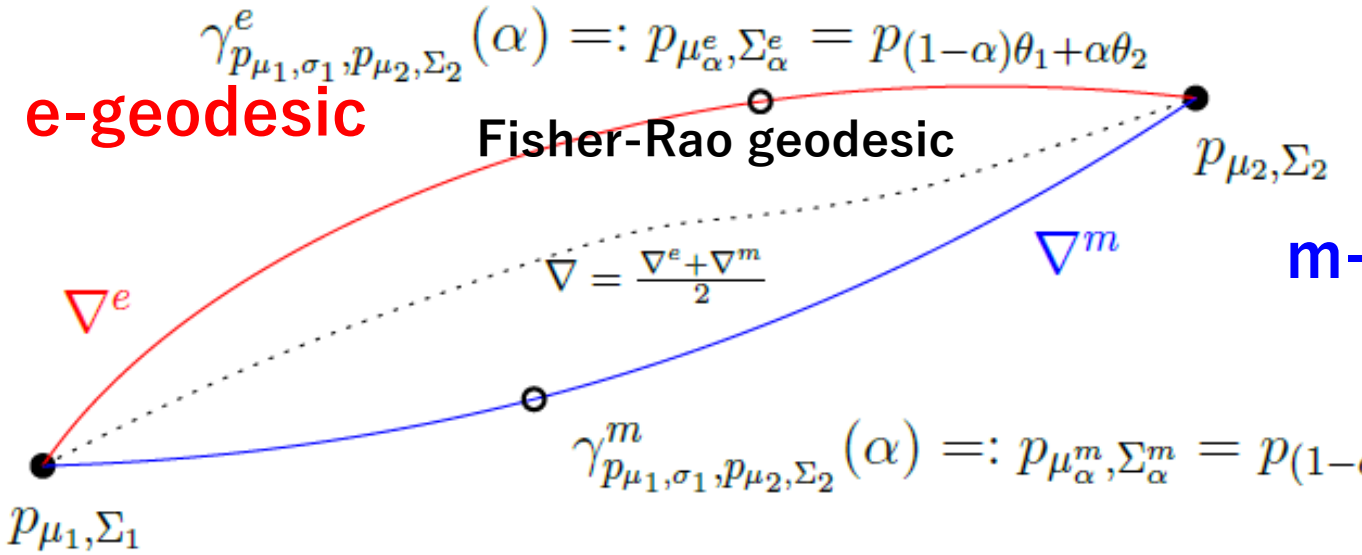
$$\begin{aligned}\mu_\alpha^e &= \Sigma_\alpha^e\left((1-\alpha)\Sigma_1^{-1}\mu_1+\alpha\Sigma_2^{-1}\mu_2\right)\\ \Sigma_\alpha^e &= \left((1-\alpha)\Sigma_1^{-1}+\alpha\Sigma_2^{-1}\right)^{-1}\end{aligned}$$

Cumulant function is convex:

$$F_\theta(\theta)=\frac{1}{2}\left(d\log\pi-\log|\theta_M|+\frac{1}{2}\theta_v^\top\theta_M^{-1}\theta_v\right)$$

with respect to natural parameters:

$$\theta=(\Sigma^{-1}\mu,\tfrac{1}{2}\Sigma^{-1}) \qquad \theta=(\theta_v,\theta_M)=\left(\Sigma^{-1}\mu,\frac{1}{2}\Sigma^{-1}\right)$$



But not convex wrt (μ Σ) parameters

m-geodesic :not mixture of Gaussians

$$p_{\eta_t}=\frac{1}{Z(\eta_t)}\exp(\langle\nabla F^*(\eta_t),t(x)\rangle)$$

$$\eta=(\mu,-\Sigma-\mu\mu^\top)$$

$$\begin{aligned}\mu_\alpha^m &= (1-\alpha)\mu_1+\alpha\mu_2=:\bar{\mu}_\alpha\\ \Sigma_\alpha^m &= (1-\alpha)\Sigma_1+\alpha\Sigma_2+(1-\alpha)\mu_1\mu_1^\top+\alpha\mu_2\mu_2^\top-\bar{\mu}_\alpha\bar{\mu}_\alpha^\top\end{aligned}$$

$$F_\eta^*(\eta)=-\frac{1}{2}\left(\log(1+\eta_v^\top\eta_M^{-1}\eta_v)+\log|-\eta_M|+d(1+\log 2\pi)\right)$$

$$\begin{cases}\eta_v(\theta)=\frac{1}{2}\theta_M^{-1}\theta_v\\ \eta_M(\theta)=-\frac{1}{2}\theta_M^{-1}-\frac{1}{4}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^\top\end{cases}\Leftrightarrow\begin{cases}\theta_v(\eta)=-(\eta_M+\eta_v\eta_v^\top)^{-1}\eta_v\\ \theta_M(\eta)=-\frac{1}{2}(\eta_M+\eta_v\eta_v^\top)^{-1}\end{cases}$$

Convex duality via Legendre-Fenchel transform

- Legendre-Fenchel transform of a convex function F:

$$F^*(\eta) = \sup_{\theta \in \Theta} \{ < \theta, \eta > - F(\theta) \}$$

- Consider "***nice***" convex functions = **Legendre-type functions** $(\Theta, F(\theta))$:

(i) Θ open, and (ii) $\lim_{\theta \to \partial\Theta} \| \nabla F(\theta) \| = \infty$

Then we get:

❶ **reciprocal gradient maps** $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$, $\nabla F^* = (\nabla F)^{-1}$

❷ conjugation yields $(H, F^*(\eta))$ of Legendre type

❸ biconjugation is an **involution**: $(H, F^*(\eta))^* = (H^* = \Theta, F^{**} = F(\theta))$

- Convex conjugate: $F^*(\eta) = < \nabla F^{-1}(\eta), \eta > - F(\nabla F^{-1}(\eta))$ since $\eta = \nabla F(\theta)$

Dual structures of information geometry

(M, g, ∇, ∇^*) such that $\frac{\nabla + \nabla^*}{2} = {}^g\nabla$

Meaning averaging
Christoffel symbols
yield Levi-Civita connection

Not necessarily the e/m connections

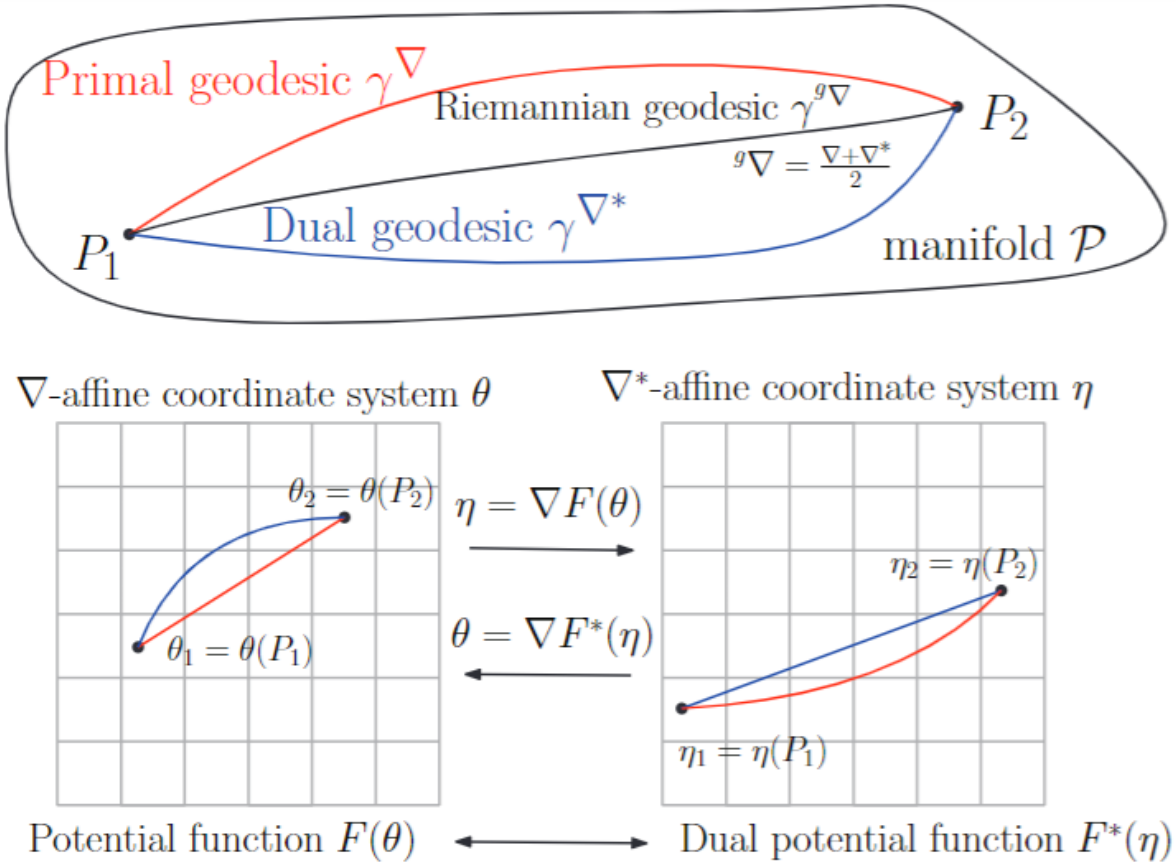
The metric g is **Hessian** if there exists a coordinate system such that g is written as the **Hessian of a potential function** then by Legendre-Fenchel convex duality there exists a dual potential function written as the Hessian with respect to the dual coordinate system

Fenchel-Young inequality $F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2$ yields a dissimilarity measure called **Fenchel-Young divergence**:

$$Y_{F,F^*}(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 = Y_{F^*,F}(\eta_2, \theta_1)$$

Dual geometry of Bregman manifolds: Convex conjugates (F, F*) yield dual flat connections

$(M, F \rightarrow g(\theta) = \nabla^2 F(\theta), F \rightarrow \nabla, F^* \rightarrow \nabla^*)$



- A connection ∇ is **flat** if there exists a coordinate system θ such that all Christoffel symbols vanish: $\Gamma(\theta) = 0$.

- **∇-geodesic** solves as **line segments**

$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0$$

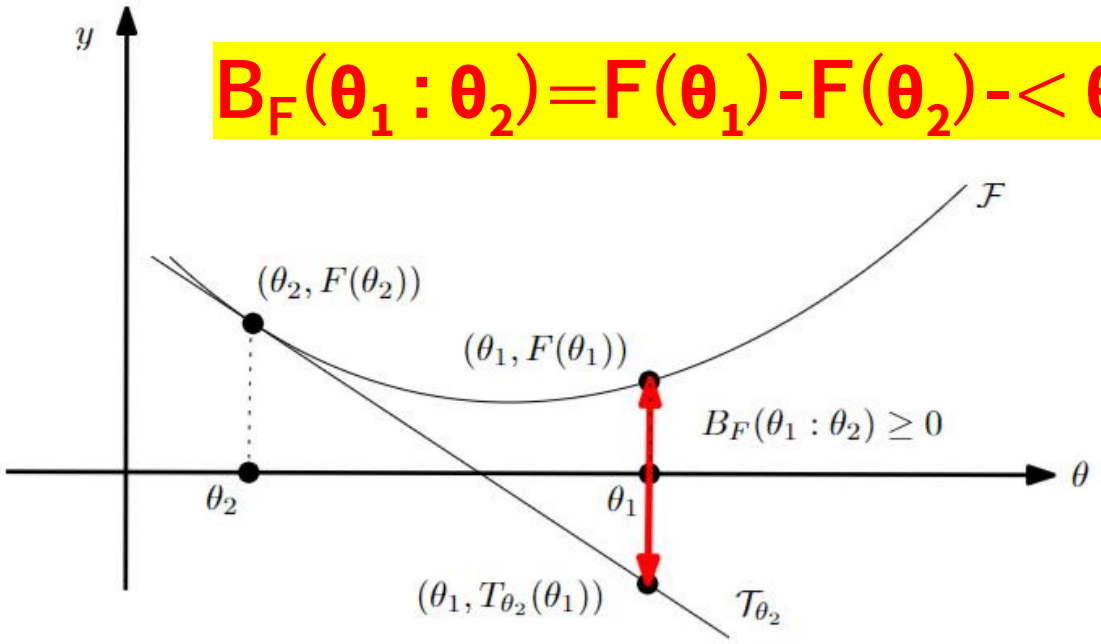
"The many faces of information geometry." *Not. Am. Math. Soc* 69.1 (2022): 36-45.

Dual Bregman/Fenchel-Young divergences

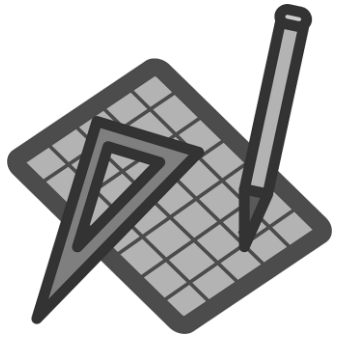
- Let $F: \Theta \subseteq \mathbb{R}^m \to \mathbb{R}$ be a strictly convex and smooth real-valued function on a Hilbert space $<.,.>$

Bregman divergence $B_F: \Theta \times \mathrm{Int}(\Theta) \to \mathbb{R}$

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - < \theta_1 - \theta_2, \nabla F(\theta_2) >$$



Popular in geometry, information theory, signal/sound processing!



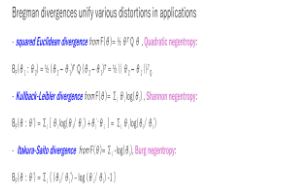
$$B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2) = Y_{F^*,F}(\eta_2, \theta_1) = B_{F^*}(\eta_2 : \eta_1)$$

BDs: A versatile class of dissimilarities

Bregman divergences unify various distortions in applications

- ***squared Euclidean divergence*** *from* $F(\theta) = \frac{1}{2}\ \theta^T Q\ \theta$, **Quadratic negentropy**:

$$B_F(\theta_1 : \theta_2) = \frac{1}{2}\ (\theta_2 - \theta_1)^T\ Q\ (\theta_2 - \theta_1)^T = \frac{1}{2}\ ||\ \theta_2 - \theta_1\ ||^2_Q$$



- ***Kullback-Leibler divergence*** *from* $F(\theta) = \sum_i \theta_i \log(\theta_i)$, **Shannon negentropy**:

$$B_F(\theta : \theta') = \sum_i \{\ \theta_i \log(\theta_i / \theta_i') + \theta_i' - \theta_i\ \} = \sum_i \theta_i \log(\theta_i / \theta_i')$$

- ***Itakura-Saito divergence*** *from* $F(\theta) = \sum_i -\log(\theta_i)$, **Burg negentropy**:

$$B_F(\theta : \theta') = \sum_i \{\ (\theta_i / \theta_i') - \log(\theta_i' / \theta_i) - 1\ \}$$

Statistical models with dual e/m flatness

- **Exponential families**: $\mathcal{E}_{t,\mu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta \rangle)\}_\theta,$

$$\begin{aligned} \mathrm{KL}(p(x;\theta_1) : p(x;\theta_2) &= B_F(\theta_2 : \theta_1), \\ &= B_{F^*}(\eta_1 : \eta_2), \end{aligned} \qquad F(\theta) := \log\left(\int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu(x)\right)$$

$$\eta = E_{p(x;\theta)}[t(x)] \qquad \text{FIM} \quad \begin{aligned} g_{ij}(\theta) &= \partial_i \partial_j F(\theta) \\ g &= \mathrm{Var}[t(X)] \end{aligned}$$

- **Mixture families**: $\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + (1 - \sum_{i=1}^{k-1} \eta_i) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}$

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta') \qquad G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta)\log m(x;\eta)\mathrm{d}\mu(x)$$

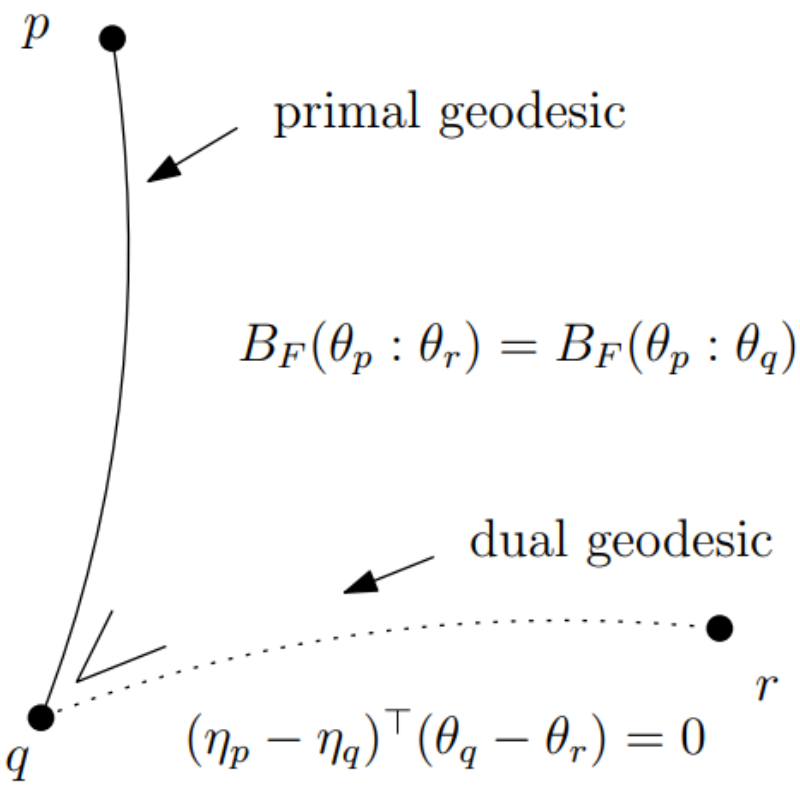
$$\text{FIM} \quad g_{ij}(\eta) = -\partial_i \partial_j h(\eta)$$

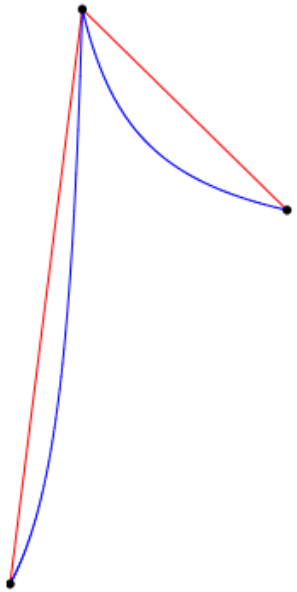
Remarks:

- Probability simplex is both a mixture and exponential family
- EFs and MFs are *not flat* with respect to the Levi-Civita g-connection
- Many statistical models are not e/m flat: Ex. Cauchy family

1803.07225

Bregman manifolds have dual Pythagorean theorems





Simultaneous dual Pythagoras theorems
(Burg/Itakura-Saito)

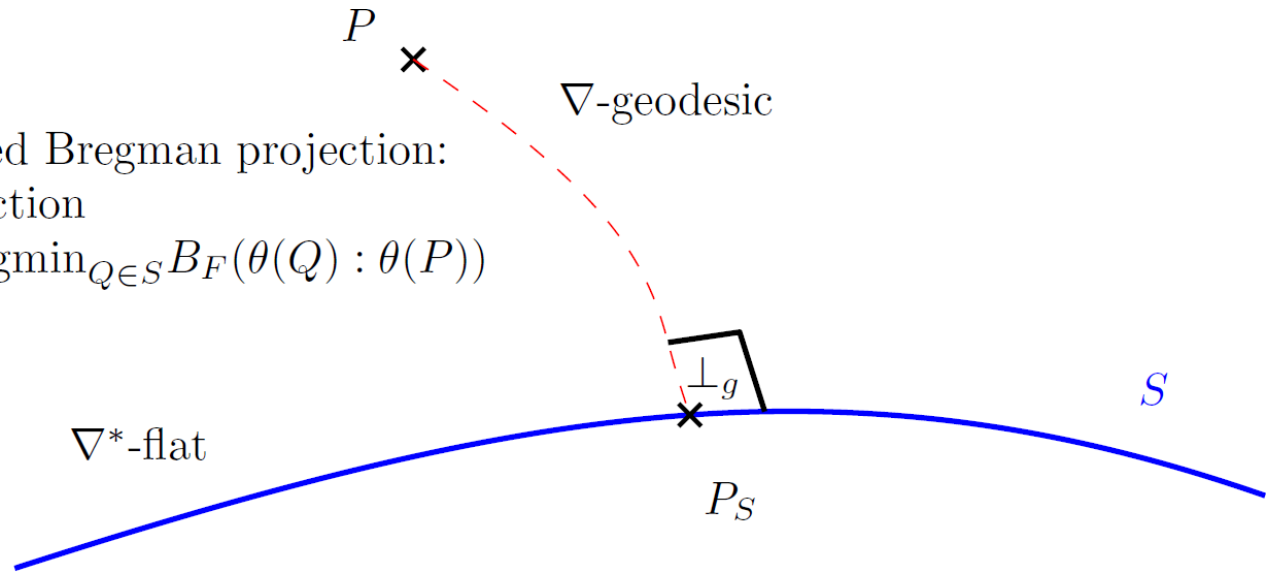
Bregman **3-parameter identity** (generalize law of cosines):

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3))$$

There is also a Bregman **4-parameter identity** (parallelogram law)

"On geodesic triangles with right angles in a dually flat space." Progress in information geometry, 2021.

Bregman projections



Remark: left/right sided divergence projections explains need of dual geodesics.

Theorem (Uniqueness of projections) *The* ∇-projection P_S *of* P *on* S *is* unique *if* S *is* ∇^*-flat *and minimizes the divergence* $D(\theta(P) : \theta(Q))$*:*

$$\nabla\text{-projection:} \quad P_S = \arg\min_{Q \in S} D(\theta(P) : \theta(Q)).$$

The dual ∇^**-projection* P_S^* *is* unique *if* $M \subseteq S$ *is* ∇-flat *and minimizes the divergence* $D(\theta(Q) : \theta(P))$*:*

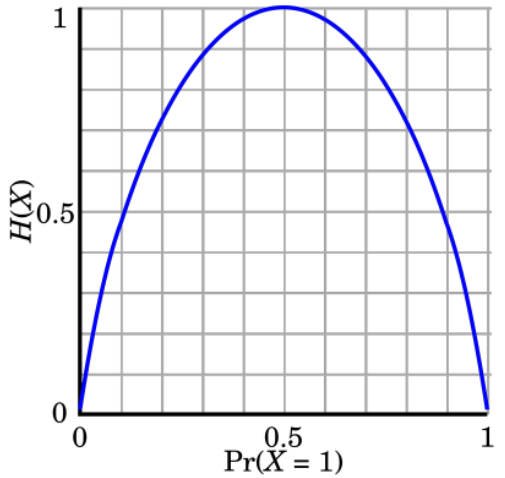
$$\nabla^*\text{-projection:} \quad P_S^* = \arg\min_{Q \in S} D(\theta(Q) : \theta(P)).$$

Shannon negentropy as Bregman potentials

$X \sim \text{Bernoulli}(p)$ with $p = \Pr(X = 1)$

$$H(X) = -(p \log p + (1-p) \log(1-p))$$

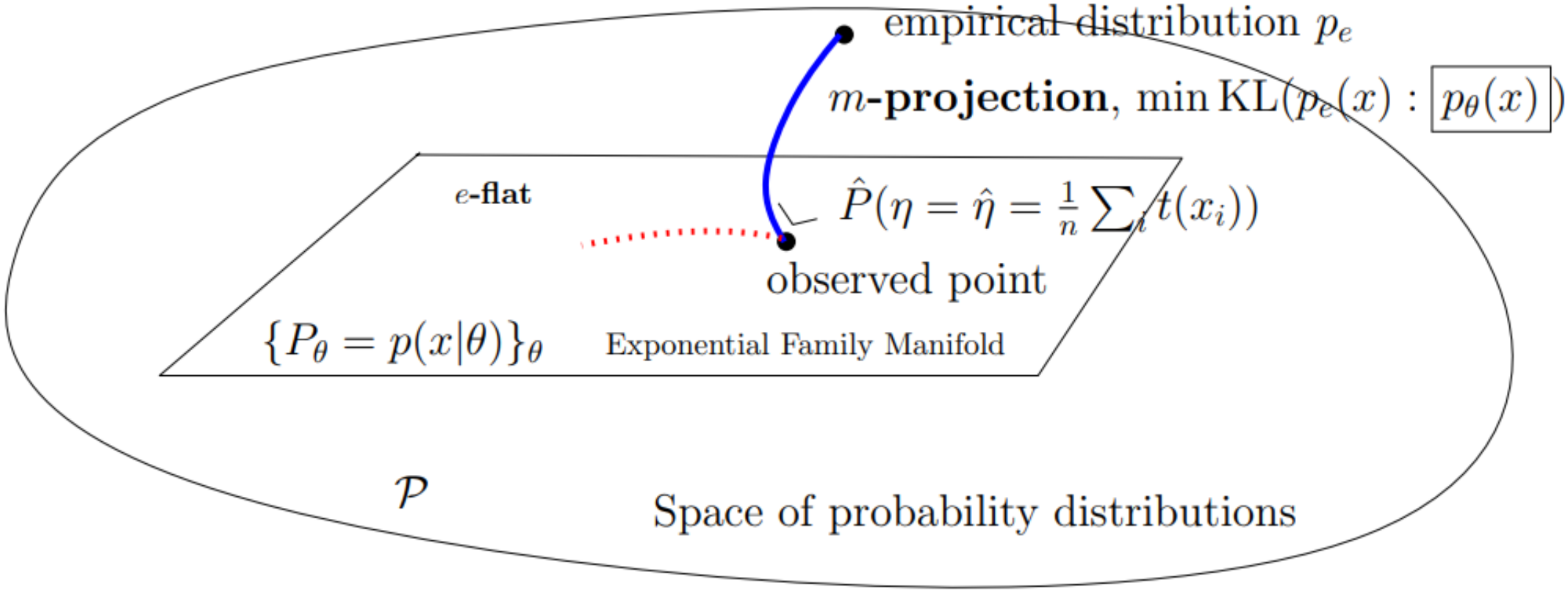
- Shannon entropy is **concave**



- **Discrete negentropy** of EFs -H(p) or **differential negentropy** of EFs are Legendre-type function = cumulant functions.

- Let us see now Bregman projections in action!

Maximum likelihood estimator (MLE) on exponential families: Right-sided Bregman projections for Bregman = KLD



"What is an information projection?" Notices of the AMS 65.3 (2018)

Maximum entropy as right-sided Bregman projection for Bregman = reverse KLD (=left KL projection)

$$\min_{p \in \Delta^{D+1}} \sum_j p_j \log p_j$$

constraints:

$$\sum_j p_j t_i(x_j) = \eta_j, \quad \forall i \in [D]$$

$$p_j \geq 0, \quad \forall i \in [|\mathcal{X}|]$$

$$\sum_j p_j = 1$$

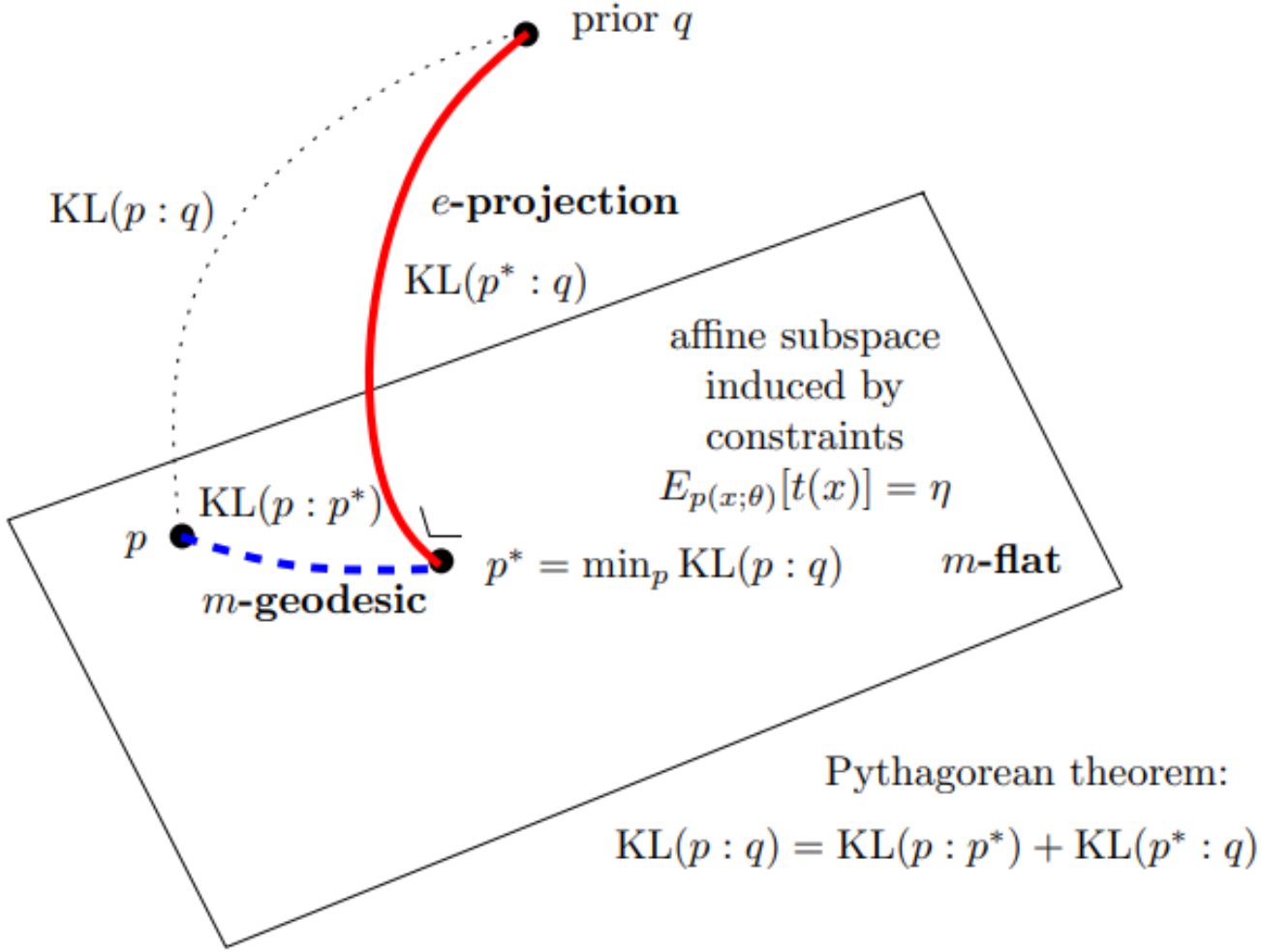
MaxEnt $H(P) \equiv$ **left-sided** $\min_P \mathrm{KL}(\boxed{\mathrm{P}} : U)$

$$\min_p \mathrm{KL}(p : h)$$

constraints:

$$\sum_j p_j t_i(x_j) = \eta_j, \quad \forall i \in [D]$$

$$p_j \geq 0, \quad \forall i \in [|\mathcal{X}|], \quad \sum_i p_j = 1$$



The moment constraints form a **m-flat submanifold**

The dual α-geometry of Amari and Nagaoka

Structure $(\mathcal{P}, g_F, \nabla^\alpha, \nabla^{-\alpha})$ Dual connections wrt Fisher metric

∇^α Defined by the Christoffel symbols $\Gamma^\alpha_{ij,k} = E_\theta\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right]$

Some α-connections:

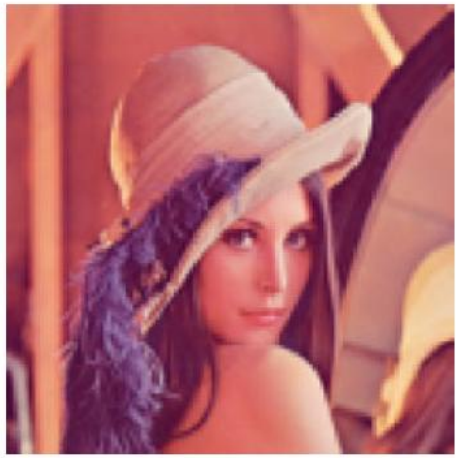
- **0-connection** = **Levi-Civita metric connection of Fisher metric : Fisher-Rao manifold**
- **1-connection** is called the **exponential connection** **[Efron 1975]**
- **-1 connection** is called the **mixture connection** **[Dawid 1975]**

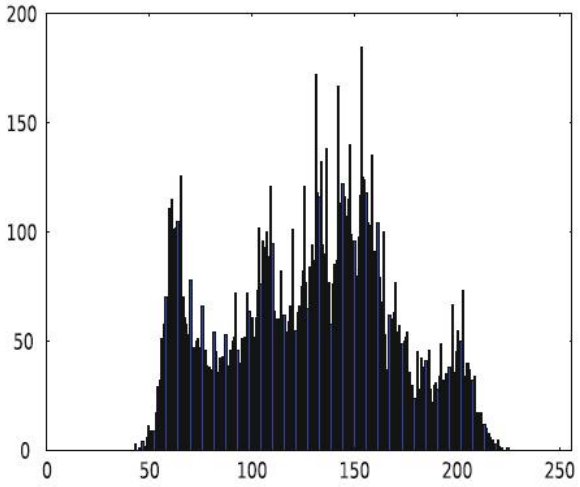
$$\nabla^\alpha = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m$$

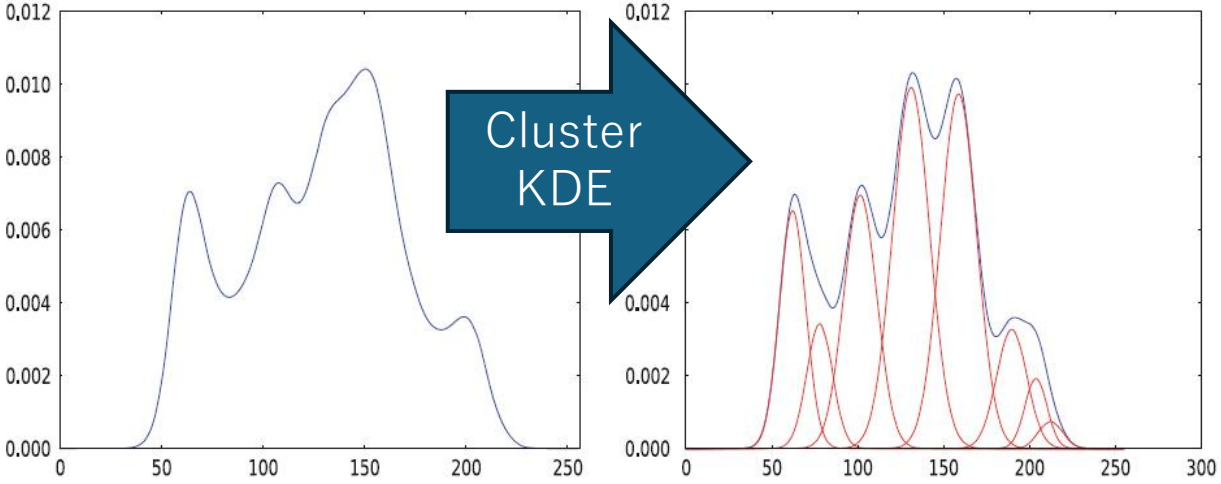
$$(\nabla^e)^* = \nabla^m$$
$$(\nabla^m)^* = \nabla^e$$

Dual geometry : study the duality between estimators/stat models

Application 1: Learning GMMs by clustering KDEs



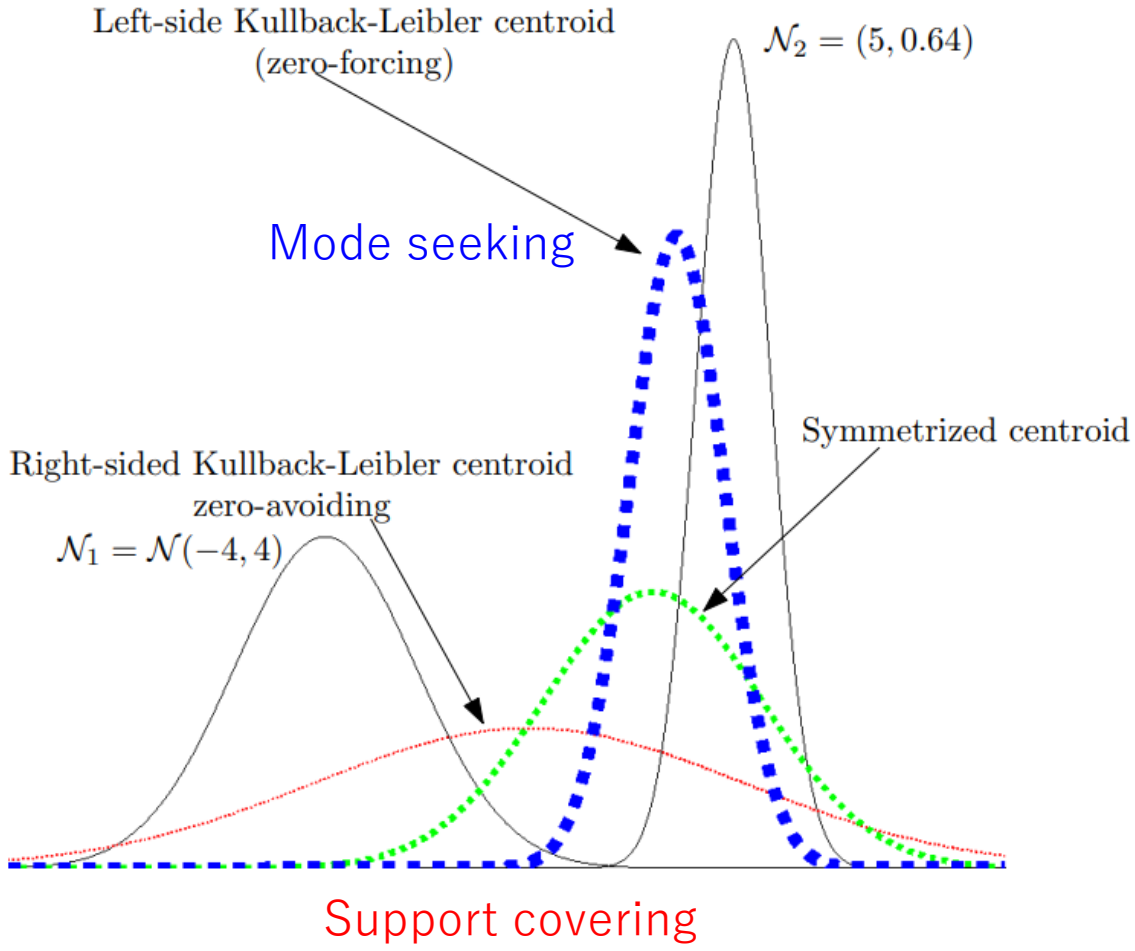




KDE

$$f(x) = \sum_{i=1}^{n} \omega_i \; g(x;\; \mu_i, \sigma_i^2)$$

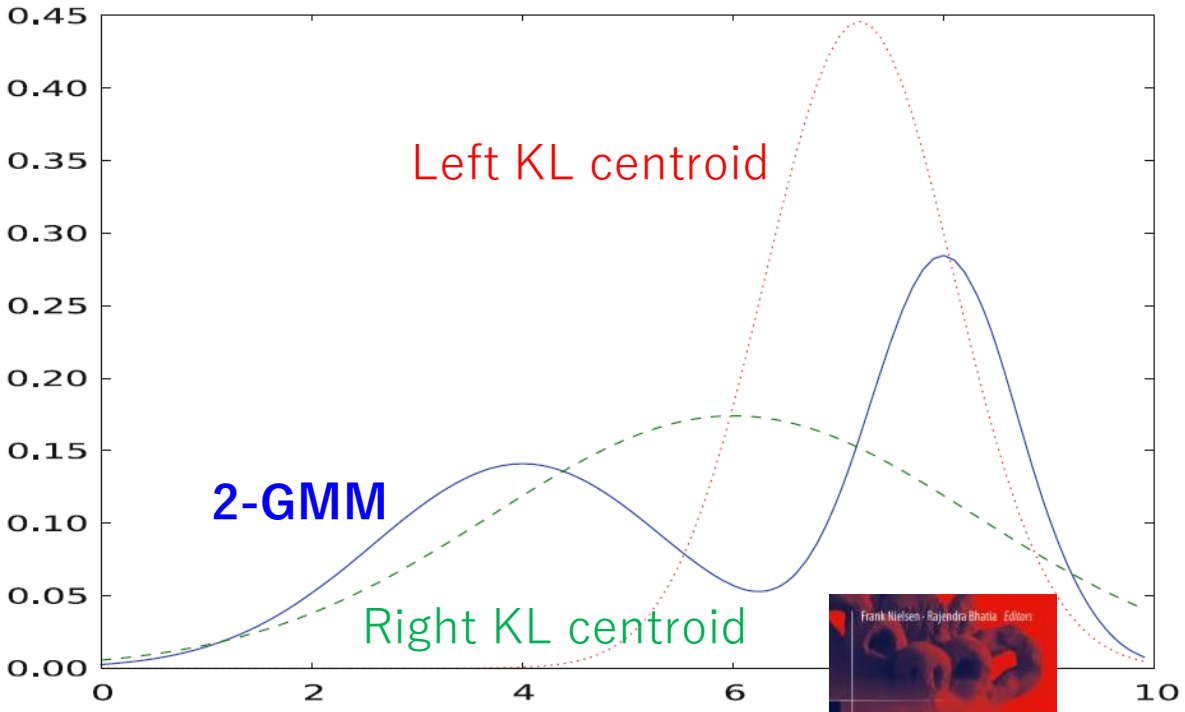
GMM

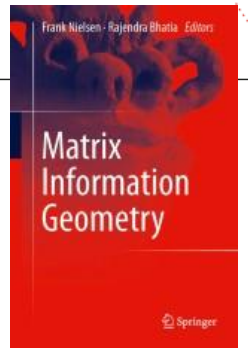


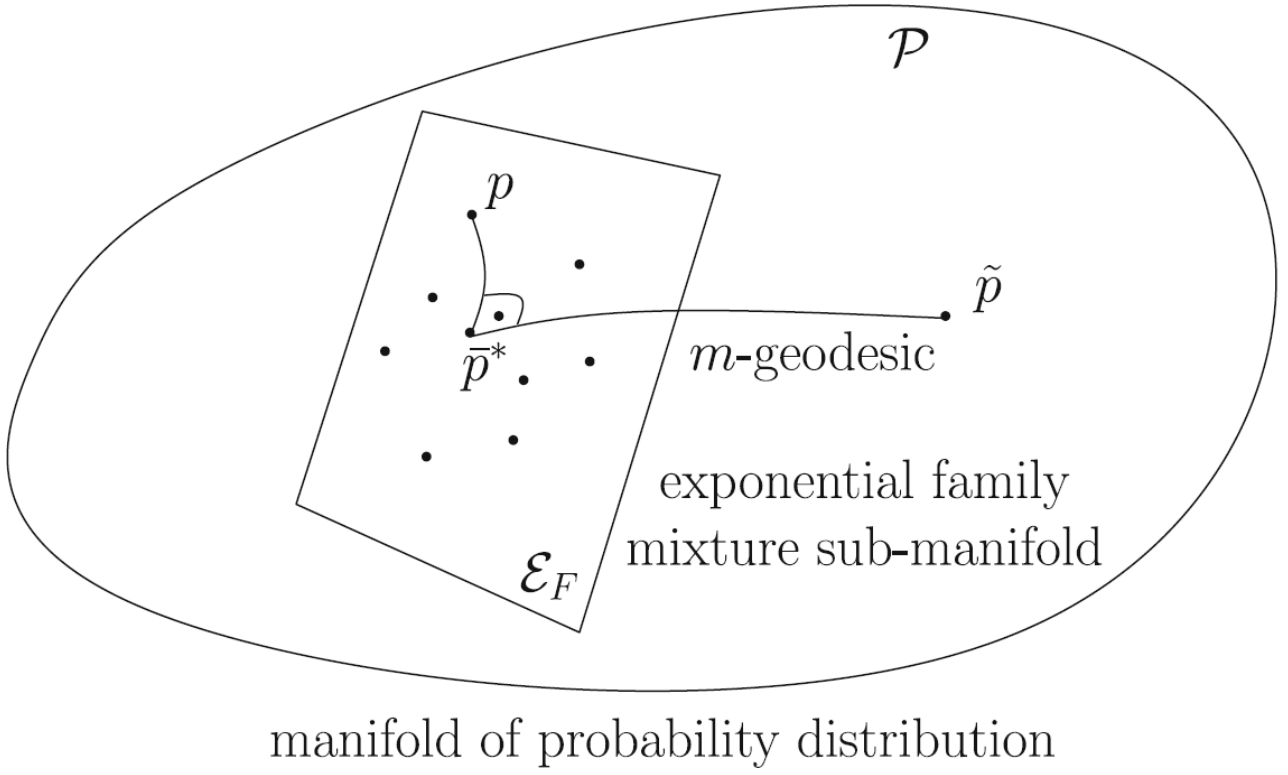
Sided and symmetrized Bregman centroids, IEEE Information Theory (2009)

Simplifying a Gaussian w-mixture density to a single Gaussian as a projection: $\bar{p}^* = \arg\min_{p \in \mathcal{E}_F} \mathrm{KL}(\tilde{p} : p)$

Right KLD minimization ≡ Left Bregman minimization
(=centroid of natural parameters)







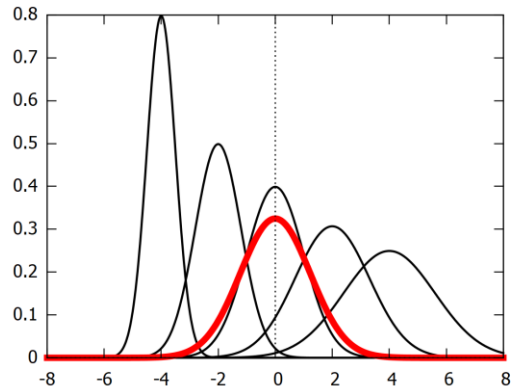
Application 1: Clustering normals/KDEs with respect to the Kullback-Leibler divergence

Since KLD between two w-mixtures amounts to a Bregman (KL) divergence,
right Bregman centroid is **center of mass** (= average of weights) ≡ left KLD centroid

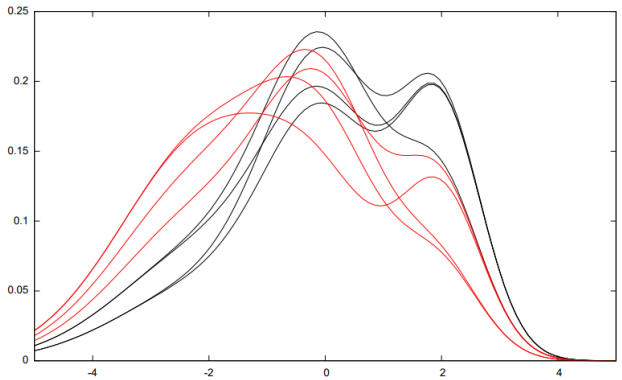
Maximize likelihood ≡ Minimize average dual Bregman divergence = Bregman centroid
In general, **Bregman k-means** can be interpreted as **k-MLE classification EM**

Duality between regular EFs and 'regular' BDs: $\log p_F(x;\theta) \propto -B_{F^*}(x:\eta)+F^*(x)$

Exponential Family $p_F(x\|\theta)$	⇔	Dual Bregman divergence B_{F^*}
Spherical Gaussian	⇔	Squared Euclidean divergence
Multinomial	⇔	Kullback-Leibler divergence
Poisson	⇔	I-divergence
Geometric	⇔	Itakura-Saito divergence
Wishart	⇔	log-det/Burg matrix divergence



Bregman right centroid
(left KLD centroid)



Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point in the natural parameter space

Two clusters: black and red

1203.5181

Application 2: Jensen-Shannon centroid

- **Jensen-Shannon divergence between categorical distributions (discrete mixtures of k-1 order because normalized to 1)**

$$\mathrm{JS}(p,q) := \frac{1}{2}\left(\mathrm{KL}\left(p:\frac{p+q}{2}\right)+\mathrm{KL}\left(q:\frac{p+q}{2}\right)\right) \qquad p_1 = m_{\theta_1} \text{ and } p_2 = m_{\theta_2}$$

$$\mathrm{JS}(p_1,p_2) = H\left(\frac{1}{2}p_1+\frac{1}{2}p_2\right) - \frac{H(p_1)+H(p_2)}{2}, \qquad H(p) = -\int p\log p\,\mathrm{d}\mu$$

- **amounts to a Jensen divergence for the Shannon negentropy generator**

$$\mathrm{JS}(p_1,p_2) = J_F(\theta_1,\theta_2)$$

lhs is divergence between densities
rhs is divergence between parameters

$$J_F(p,q) = \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right)$$

Task: Given a set of discrete distributions (categorical distributions, normalized histograms), calculate its **Jensen-Shannon centroid**

$$\min_p \sum_i \mathrm{JS}(p_i, p),$$

$$\min_\theta \sum_i J_F(\theta_i, \theta),$$

$$\min_\theta \sum_i \frac{F(\theta_i) + F(\theta)}{2} - F\left(\frac{\theta_i + \theta}{2}\right),$$

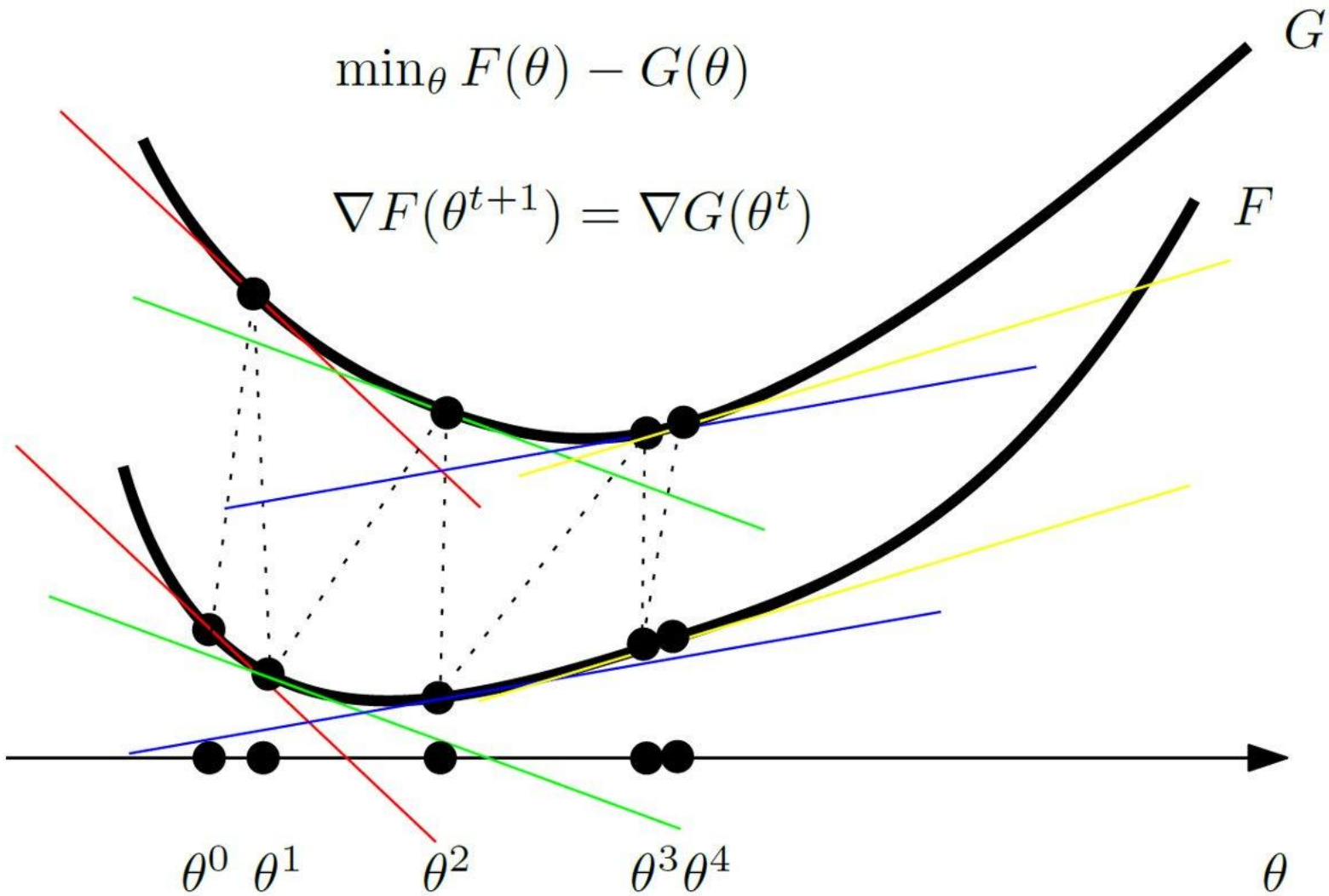
$$\equiv \min_\theta \frac{1}{2}F(\theta) - \frac{1}{n}\sum_i F\left(\frac{\theta_i + \theta}{2}\right) := E(\theta)$$

OPT is a sum of convex – convex function

Need to minimize a **difference of convex functions**
DCA or **ConCave Convex algorithm** or **DCA**

On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid, *Entropy* 22.2 (2020)

ConCave-Convex Procedure (CCCP)

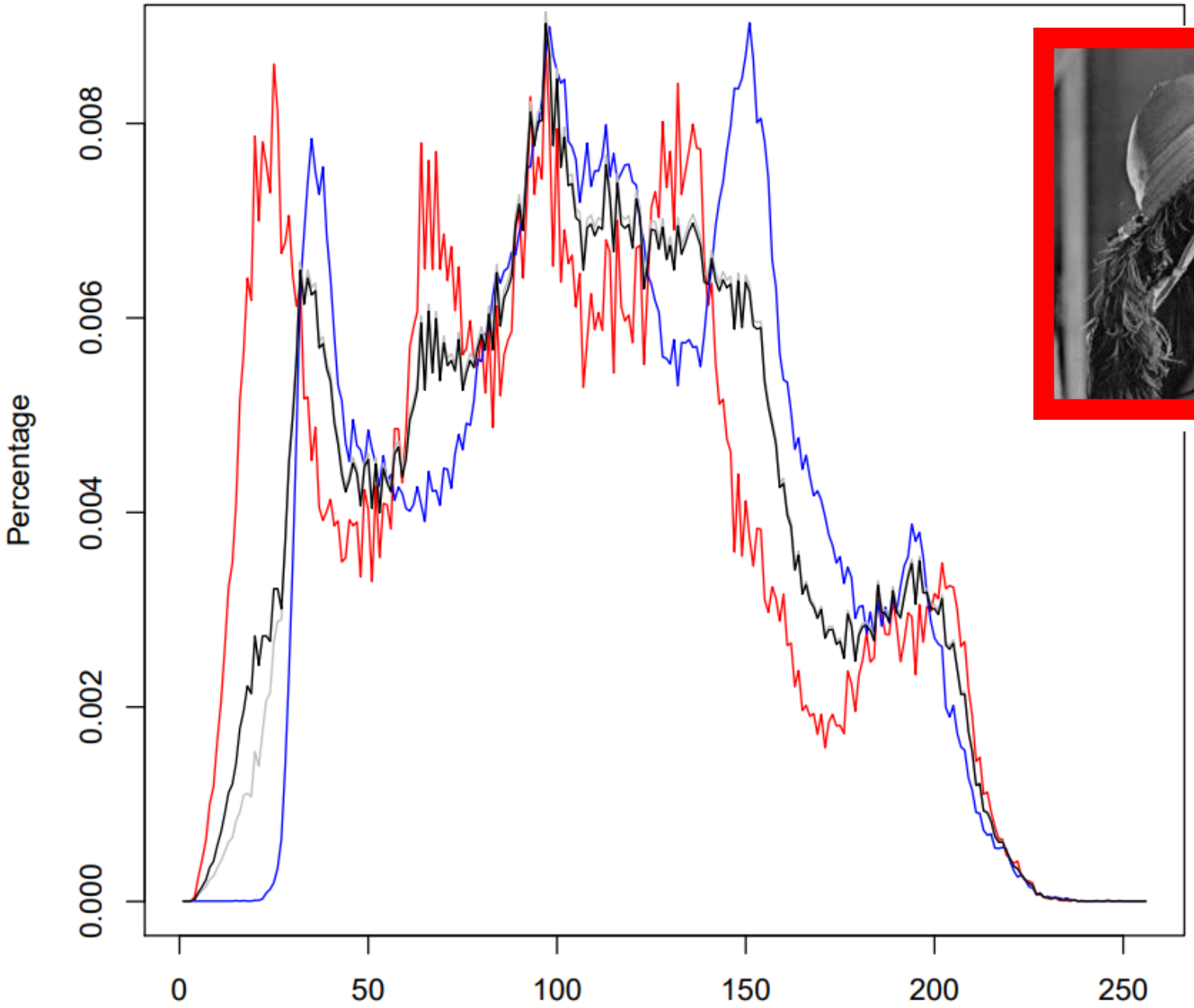


- Write any energy/loss with lower bounded Hessian as the sum of a **convex function F** plus a **concave function -G**
- Optimization to a local minimum by matching points of the graph plots which have the same tangent hyperplane

  (no learning rate!)

$$\nabla F(\theta^{t+1}) = \nabla G(\theta^t)$$

Yuille, Alan L., and Anand Rangarajan. "The concave-convex procedure (CCCP)." , NeurIPS 2002.
The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory* 57.8 (2011)



Jensen–Shannon centroid

Jeffreys/SKL centroid

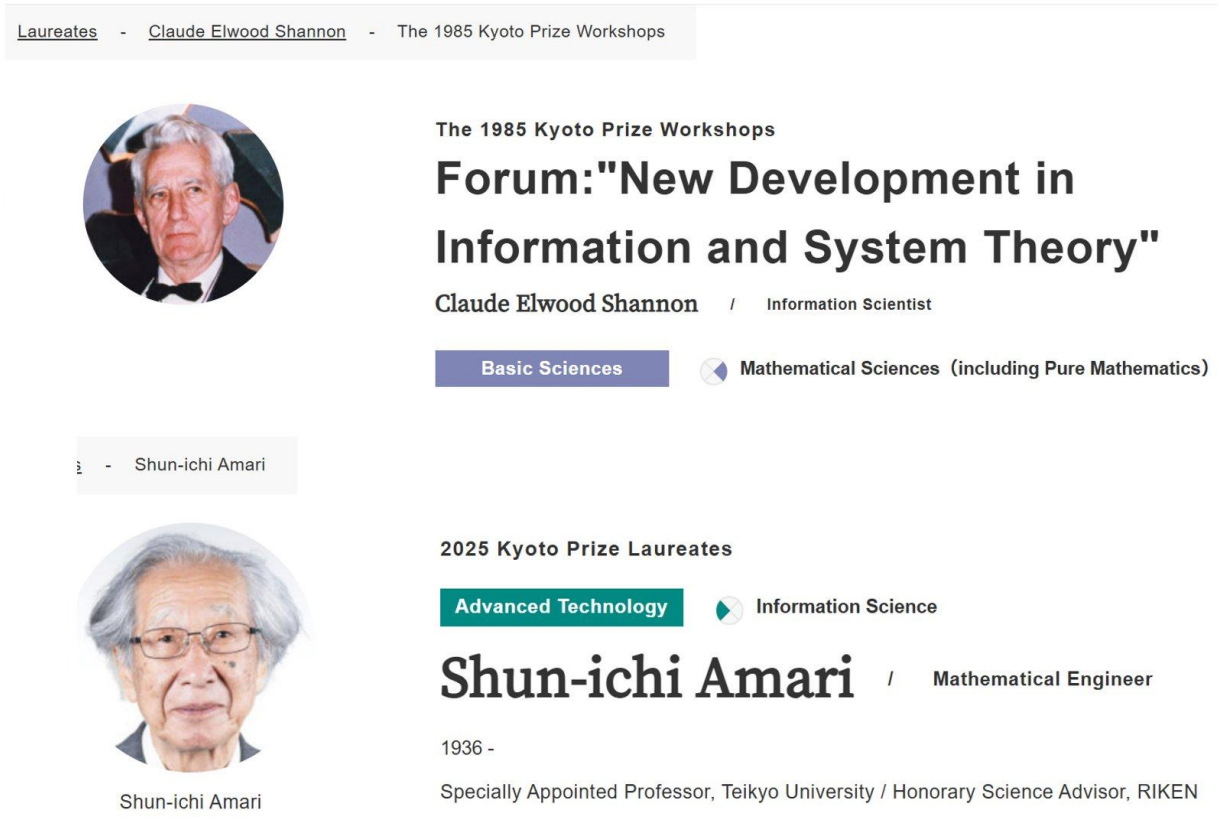
Jensen–Shannon centroid do not require same support

On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid, *Entropy* 22.2 (2020)

When Prof. Shun-ichi Amari met Claude Shannon…

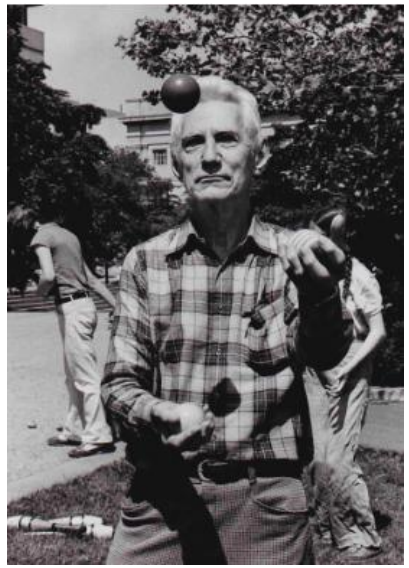
Shannon received the Kyoto prize in 1985 for
"New Development in Information and System Theory“

Prof. Amari attended the forum following the commemorative lecture of Shannon and gave a speech



End of part I: Framework of information geometry

- Amari, Shun-ichi. “Information geometry and its applications” Springer, 2016. *380 pages GOAT*
- "The many faces of information geometry" Not. Am. Math. Soc 69.1 (2022): 36-45. *9 pages OVERVIEW*
- "An elementary introduction to information geometry" Entropy 22.10 (2020): 1100. *61 pages TUTORIAL*
- "*k-MLE: A fast algorithm for learning statistical mixture models*," IEEE ICASSP, 2012.
- "*On geodesic triangles with right angles in a dually flat space*," Progress in information geometry: theory and applications. Springer 2021



8-Dan Honor for a Pioneer of Go AI

Geometric Information Theory

Hub to information sciences

Part II: Some recent results

Frank Nielsen

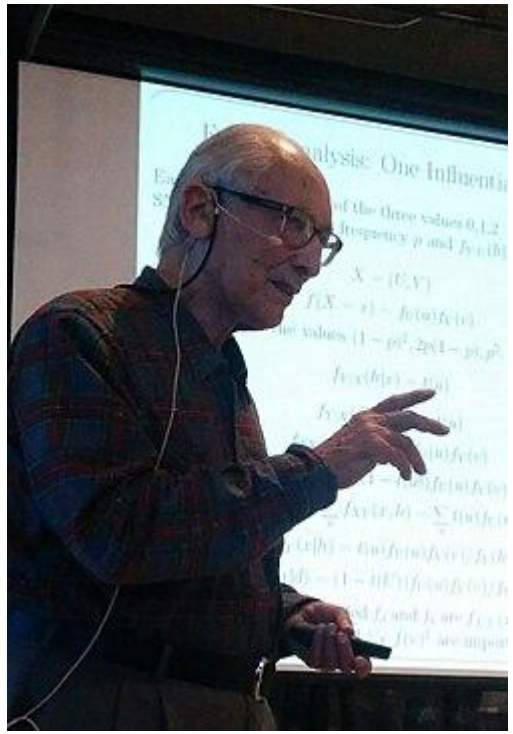
Sony Computer Science Laboratories, Inc

Sony CSL

July 6th, 2026

Information geometry of Chernoff information

Chernoff distribution, Chernoff point, LREF



July 1, 1923 (age 102)

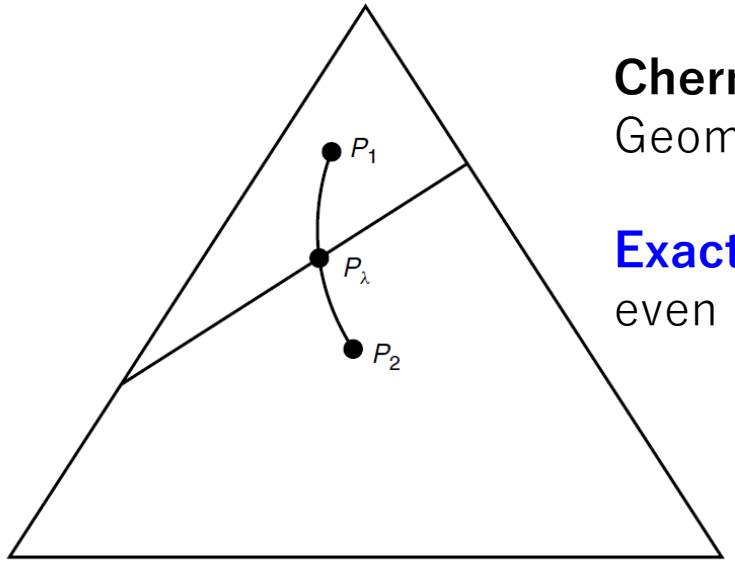
$$C(P,Q) = -\log \min_{\alpha\in(0,1)} \int p^{\alpha}(x) q^{1-\alpha}(x) \mathrm{d}\nu(x)$$

CI useful in statistics, information fusion, etc.

Chernoff information (Bayesian hypothesis testing)

$$C(P_1, P_2) \stackrel{\Delta}{=} -\min_{0\le\lambda\le 1} \log\left(\sum_x P_1^\lambda(x) P_2^{1-\lambda}(x)\right)$$
$$= D(P_{\lambda*}||P_1) = D(P_{\lambda*}||P_2)$$

$$C(P,Q) = -\log \min_{\alpha\in(0,1)} \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}\nu(x).$$



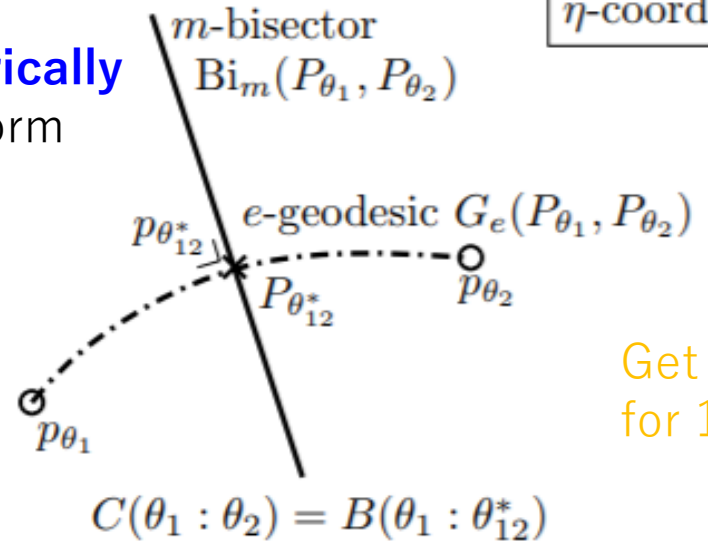
Chernoff distribution =
Geometric mixture for optimal exponent

Exactly characterized geometrically
even if not available in closed form

$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}$$

$$P^* = P_{\theta_{12}^*} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$$

η-coordinate system



$$C(\theta_1 : \theta_2) = B(\theta_1 : \theta_{12}^*)$$

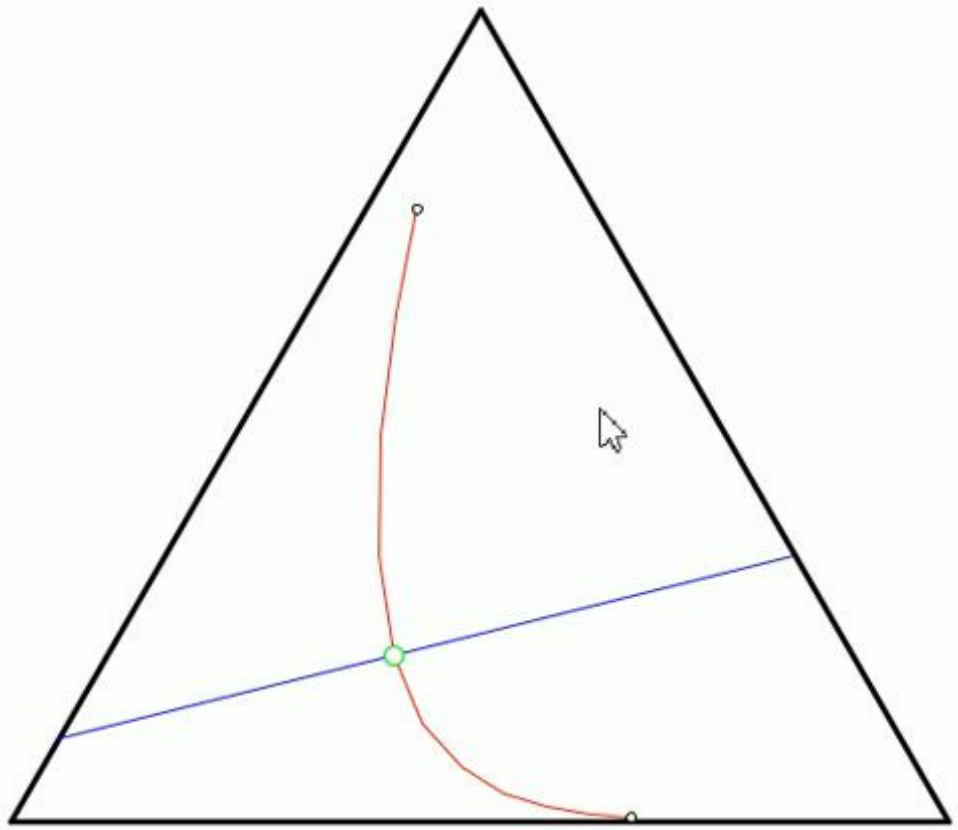
Get large formula
for 1D Gaussians!!!

Categorical/Shannon manifold

Exponential family manifold

In pure geometric term, Chernoff point intersection of primal geodesic with dual bisector

Unique intersection point of
the exponential geodesic
with
the dual mixture bisector

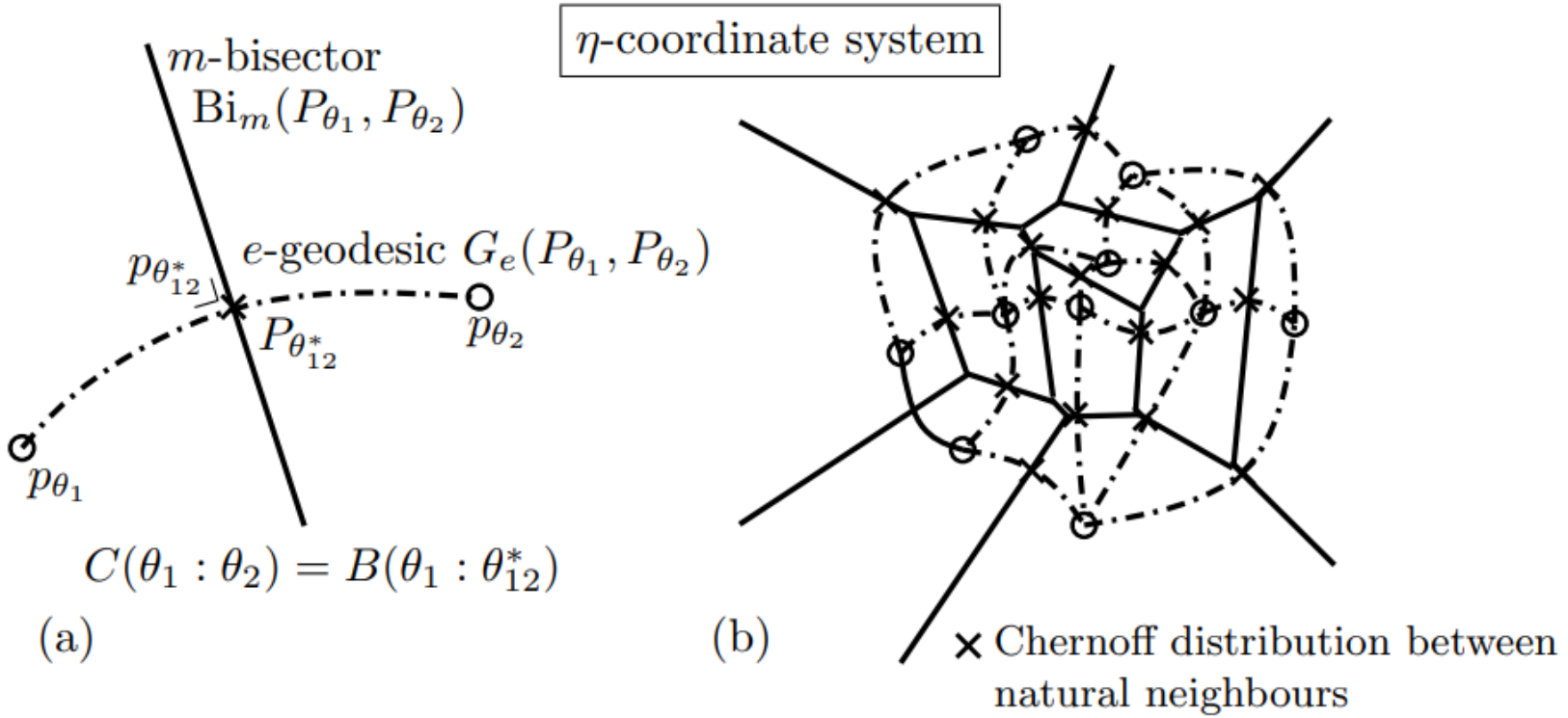


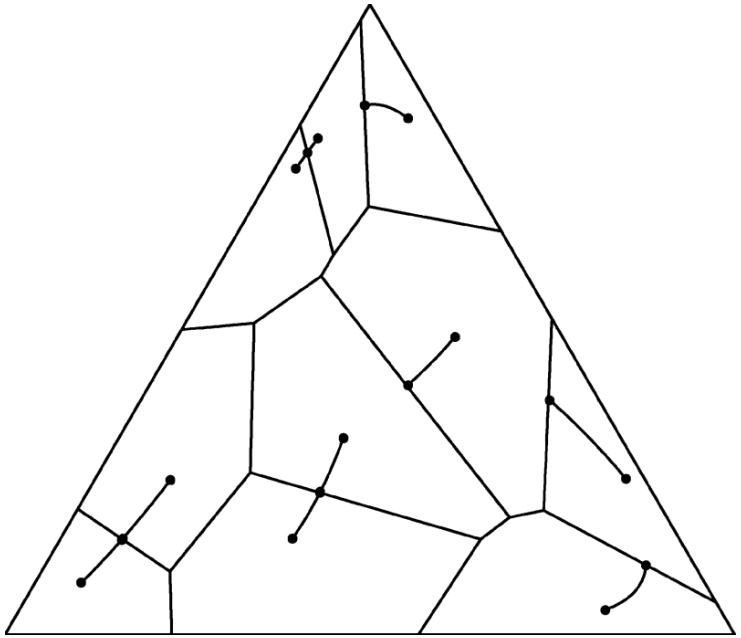
(Here 2D probability simplex of the family of categorical distributions with 3 choices)

Chernoff information of multiple distributions

Defined as the **minimum of pairwise Chernoff information**

Computed by inspecting **natural neighbors** in **Voronoi diagrams**





Westover, 2008

Bregman Voronoi diagrams

1102.2684

0709.2196

Likelihood ratio exponential families (LREFs)

General case for Chernoff information!

exponential open arc $\{g_\alpha(x) \ : \ \alpha \in (0,1)\}$

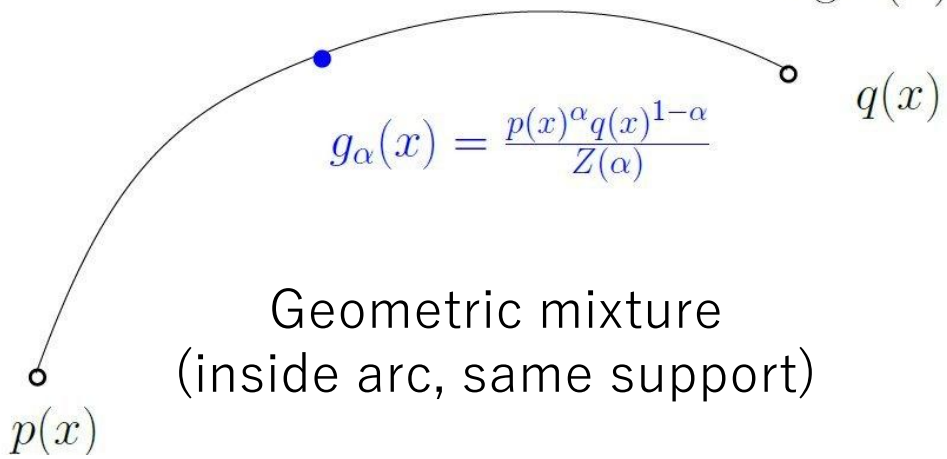
is an exponential family $g_\alpha(x) = \exp\left(\alpha \log \frac{p(x)}{q(x)} - \log Z(\alpha)\right) q(x)$

$$Z(\alpha) = \int p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}\mu(x)$$

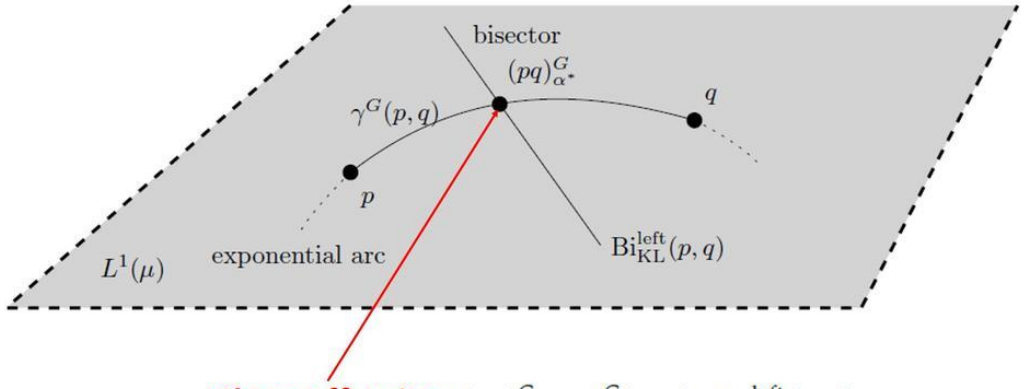
Convex

$$\log Z(\alpha) = -D_\alpha^{\text{Bhattacharyya}}[p:q]$$

Cumulant of LREF is negative Bhatt. dist.



Geometric mixture
(inside arc, same support)



Bhattacharya distance is not a metric distance

$$D_{\text{Bhat},\alpha}[p:q] := -\log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \, \mathrm{d}\mu(x)$$

Concave in α

Chernoff point $(pq)^G_{\alpha^*} = \gamma^G(p,q) \cap \text{Bi}^{\text{left}}_{\text{KL}}(p,q)$

$$\gamma^G(p,q) := \left\{(pq)^G_\alpha \ : \ \alpha \in [0,1]\right\}$$

$$\text{Bi}^{\text{left}}_{\text{KL}}(p,q) := \left\{r \in L^1(\mu) \ : \ D_{\text{KL}}[r:p] = D_{\text{KL}}[r:q]\right\}$$

"Revisiting Chernoff information with likelihood ratio exponential families." Entropy (2022)

Fisher Information Matrix (FIM)
vs
Fisher metric & Euclidean metric

How to recognize when Fisher metric is Hessian
(and get a dually flat geometry)

Reparameterizing the Fisher Information matrix

$$\eta(\theta) \Leftrightarrow \theta(\eta) \qquad I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial \theta_i}{\partial \eta_j}\right]$$

Q: Given FIM in some given parameterization, can we check that the Fisher metric yields Hessian geometry?
That is, is there a parameterization θ and potential function F so that $I(\theta)=\nabla^2 F(\theta)$

Many Fisher-Rao manifolds of **2D statistical models** are Hessians but the induced dual connections by F and F* may not be the e/m connections, nor the Hessian metric may correspond to the Fisher information matrix

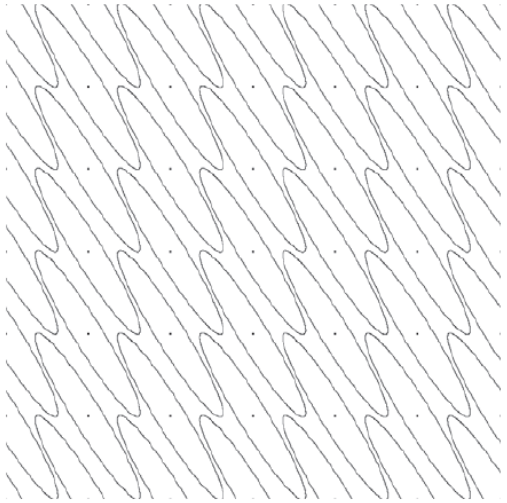
Example: t-Student yield Hessian geometry, location-scale family…

2006.07020

Mahalanobis geometry is Euclidean geometry

Rao distance between same-covariance normal distributions:

$$\rho(p_1, p_2) = \sqrt{(p_1 - p_2)^\top \Sigma^{-1} (p_1 - p_2)}, \quad g(p) = \Sigma^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$



non-conformal geometry : $g(p) \neq f(p)I$
(Visualization with Tissot indicatrix)

By using Cholesky decomposition, we get

$$\begin{aligned} D_\Sigma(p : q) &= (p - q)^\top \Sigma (p - q) \\ &= (p - q)^\top L L^\top (p - q) \end{aligned}$$

Euclidean distance on affinely transformed parameters:

$$\|L^\top p - L^\top q\| = D_E(L^\top p, L^\top q)$$

$$x' = L^\top x$$

Affine transformation is a gauge freedom
Represent the same Euclidean tangent plane geometry

Separable Bregman divergences (product of 1D EFs…)

… **always** yield **Euclidean geometry** (in non-Cartesian coords)

$$\delta_\Phi(\boldsymbol{x},\boldsymbol{x}') = \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}') - (\boldsymbol{x}-\boldsymbol{x}')^\top \nabla\Phi(\boldsymbol{x}')$$

$$\Phi(\boldsymbol{x}) = \sum_{j=1}^K \phi(x_j) \text{ with } \phi : \mathcal{J} \to \mathbb{R}$$

$$\delta_\Phi(x,x') = \sum_{j=1}^K \delta_\phi(x_j,x'_j)$$

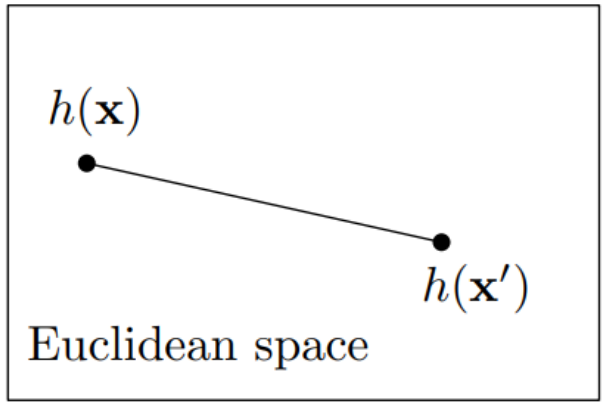
Geodesic distance (Rao distance): $d_\phi(\boldsymbol{x},\boldsymbol{x}')^2 = \sum_{j=1}^K \left(h(x_j)-h(x'_j)\right)^2$

where $h(x) = \int \sqrt{\phi''(x)}$ $\quad g_{i,j}(\boldsymbol{x}) = \phi''(x_i)\delta_{i,j}$ Diagonal matrix

geodesics $\gamma_i(t) = h^{-1}\Big((1-t)h(x_i) + t h(x'_i)\Big)$

$$I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial\theta_i}{\partial\eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial\theta_i}{\partial\eta_j}\right]$$

Jacobian matrix y=h(x)



Euclidean dist.
on h-representation

1810.10770

Metric tensor g using covariant/contravariant notations

metric tensor in local primal coordinates:

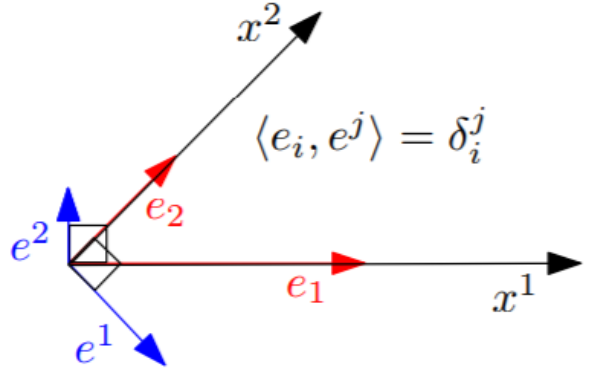
$$g_{ij}(\theta) = \nabla^2 F(\theta)$$

Contravariant g

Dual metric tensor in dual local coordinates:

$$g^{ij}(\eta) = g^{*ij}(\eta) = \nabla^2 F^*(\eta)$$

covariant g



Reciprocal basis:
Mutually orthogonal basis
e_i perpendicular to e^i

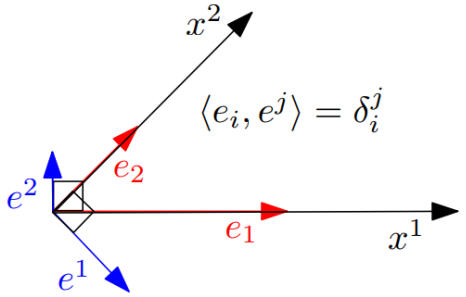
Crouzeix's identity:

$$\nabla^2 F(\theta) \nabla^2 F^*(\eta) = I$$

An elementary introduction to information geometry." Entropy 22.10 (2020)

Squaring Hessian potential matrices yield Euclidean geometry…

Crouzeix identity: $\nabla_\eta^2 F^*(\eta)^\top \nabla^2 F(\theta(\eta)) = I_{m,m}$



$$G_\xi(\xi) = J_\xi(\theta)^\top G(\theta(\xi))\, J_\xi(\theta) \qquad J_\xi(\theta) = \left[\frac{\partial \theta_i}{\partial \xi_j}\right]$$

In particular, we have $J_\eta(\theta) = \nabla_\eta \theta(\eta) = \nabla_\eta \nabla_\eta F^*(\theta) = \nabla_\eta^2 F^*(\eta)$

Squared any Hessian matrix to get **another SPD matrix** representing a metric:

$$G_{\mathrm{sqr}}(\eta) = \nabla_\eta^2 F^*(\eta)^\top (\nabla^2 F(\theta(\eta)))^2 \nabla_\eta^2 F^*(\eta) = I_{m,m}.$$

Geodesic distance is the **Euclidean distance on the dual parameters**:

$$\rho_{g_{\mathrm{sqr}}}(P_1, P_2) = \|\eta_1 - \eta_2\|_2 = \|\nabla F(\theta_1) - \nabla F(\theta_2)\|_2$$

$$\eta_i = \nabla F(\theta(P_i))$$

Instead of Cholesky decomp of SPD, we can take symmetric square root:
It is SPD matrix and square it. Mahalanobis is thus Euclidean geometry 2511.21173

Reconstructing statistical divergences from Bregman divergences with integral generators

$$F(\theta) = F_{\mathcal{E}}(p_\theta) = \log\left(\int \exp(\theta^\top t(x))\mathrm{d}\mu(x)\right)$$

$$F_{\mathcal{M}}(m_\theta) = -h(m_\theta) = \int m_\theta(x)\log m_\theta(x)\mathrm{d}\mu(x)$$

$$F(\theta) = F_{\mathcal{E}}(p_\theta) = \log\left(\int \exp(\theta^\top t(x))\mathrm{d}\mu(x)\right) \qquad F^*(\eta) = -h(p_\theta) = \int p(x)\log p(x)\mathrm{d}\mu(x)$$

$$\eta_2(i) = E_{p_{\theta_2}}[t_i(x)]$$

$$\textstyle\sum_i \theta_1(i)\eta_2(i) = E_{p_{\theta_2}}\left[\sum_i \theta_1(i)t_i(x)\right]$$

$$\textstyle\sum_i \theta_1(i)t_i(x) = (\log p_{\theta_1}(x)) + F(\theta_1)$$

hence $\theta_1^\top \eta_2 = E_{p_{\theta_2}}[\log p_{\theta_1} + F(\theta_1)] = F(\theta_1) + E_{p_{\theta_2}}[\log p_{\theta_1}]$

$$\begin{aligned} B_{F,\mathcal{E}}[p_{\theta_1} : p_{\theta_2}] &= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2, \\ &= F(\theta_1) - h(p_{\theta_2}) - E_{p_{\theta_2}}[\log p_{\theta_1}] - F(\theta_1) \\ &= E_{p_{\theta_2}}\left[\log \frac{p_{\theta_2}}{p_{\theta_1}}\right] =: D_{\mathrm{KL}^*}[p_{\theta_1} : p_{\theta_2}], \end{aligned}$$

BD induced by cumulant function = reverse KLD on corresponding densities

Compare with: In ML, KLD between EF densities = BD on swapped parameter order

Reconstructing statistical divergences from Bregman divergences

- The **partition function** Z=exp(F) is **log-convex** and hence also convex. We get **two BDs** B_Z and B_F from exp. families.

$$
\begin{aligned}
D_{\mathrm{KL}}(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) &= \int \left(\tilde{p}_{\theta_1}(x) \log \frac{\tilde{p}_{\theta_1}(x)}{\tilde{p}_{\theta_2}(x)} + \tilde{p}_{\theta_2}(x) - \tilde{p}_{\theta_1}(x) \right) \mathrm{d}\mu(x), \\
&= \int \left(e^{\langle t(x), \theta_1 \rangle} \langle \theta_1 - \theta_2, t(x) \rangle + e^{\langle t(x), \theta_2 \rangle} - e^{\langle t(x), \theta_1 \rangle} \right) \mathrm{d}\mu(x), \\
&= \left\langle \int t(x) e^{\langle t(x), \theta_1 \rangle} \mathrm{d}\mu(x), \theta_1 - \theta_2 \right\rangle + Z(\theta_2) - Z(\theta_1), \\
&= \langle \theta_1 - \theta_2, \nabla Z(\theta_1) \rangle + Z(\theta_2) - Z(\theta_1) = B_Z(\theta_2 : \theta_1),
\end{aligned}
$$

$$p_\lambda(x) \propto \tilde{p}_\lambda(x) = \exp\left(\langle \theta(\lambda), t(x) \rangle\right) h(x)$$

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \tilde{p}_\lambda(x)$$

$$Z(\lambda) = \int \tilde{p}_\lambda(x) \mathrm{d}\mu(x)$$

Extended Kullback-Leibler divergence

2312.12849

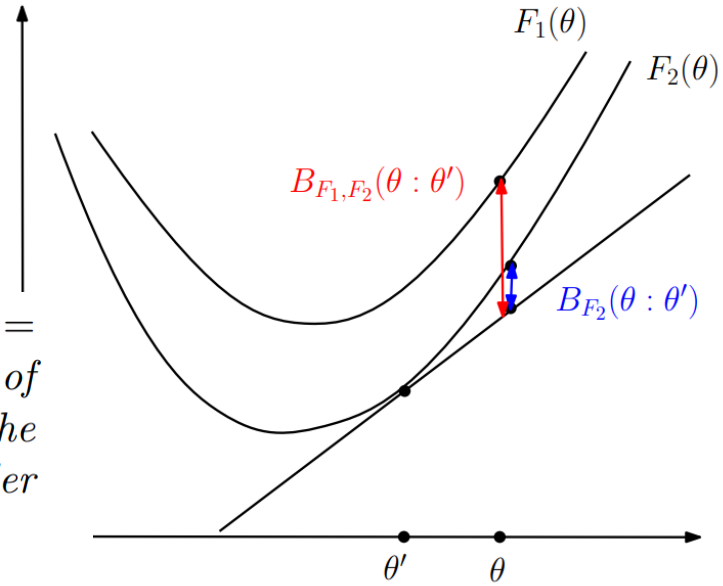
Truncated exponential families and KLDs

- **Truncated EFs are EFs** but not necessarily regular nor steep
- Ex.: Singly truncated normals $[0, \infty)$ are not regular and not steep
- Consider two generators F_1 F_2. Legendre transform reverse order

$$B_{F_1,F_2}(\theta : \theta') = F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta') \geq 0$$

$$Y_{F_1,F_2^*}(\theta, \eta') := F_1(\theta) + F_2^*(\eta') - \theta^\top \eta'$$

Duo pseudo-Bregman divergence
Duo pseudo Fenchel-Young divergence



Theorem (Kullback-Leibler divergence between nested exponential family densities). *Let $\mathcal{E}_2 = \{q_{\theta_2}\}$ be an exponential family with support $\mathcal{X}_2$, and $\mathcal{E}_1 = \{p_{\theta_1}\}$ a truncated exponential family of $\mathcal{E}_2$ with support $\mathcal{X}_1 \subset \mathcal{X}_2$. Let F_1 and F_2 denote the log-normalizers of $\mathcal{E}_1$ and $\mathcal{E}_2$ and η_1 and η_2 the moment parameters corresponding to the natural parameters θ_1 and θ_2. Then the Kullback-Leibler divergence between a truncated density of $\mathcal{E}_1$ and a density of $\mathcal{E}_2$ is*

$$D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = Y_{F_2,F_1^*}(\theta_2 : \eta_1) = B_{F_2,F_1}(\theta_2 : \theta_1) = B_{F_1^*,F_2^*}(\eta_1 : \eta_2) = Y_{F_1^*,F_2}(\eta_1 : \theta_2).$$

2202.10726

KL diverge between truncated normal densities

For illustration purpose only!!!

PDF of truncated nor (a,b):

$$p_{m,s}^{a,b}(x) = \frac{1}{\sqrt{2\pi}s\ (\Phi_{m,s}(b) - \Phi_{m,s}(a))} \exp\left(-\frac{(x-m)^2}{2s^2}\right)$$

$$\Phi_{m,s}(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-m}{\sqrt{2}s}\right)\right), \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t.$$

Truncated normal PDFs form an exp l family with log-normalizer :

$$F_{a,b}(m, \ldots) = \ldots^2 + \frac{1}{2}\log 2\pi s^2 + \log\left(\Phi_{m,s}(b) - \Phi_{m,s}(a)\right)$$

Moment parameters and mean & variance:

$$\mu(m, s; a, \ldots) = \ldots s\frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}, \qquad \phi(x) := \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

$$\eta_1(m, s; a, b) = E_{p_{m,s}^{a,b}}[x] = \mu(m, s; a, b),$$

$$\eta_2(m, s; a, b) = E_{p_{m,s}^{a,b}}[x^2] = \sigma^2(m, s; a, b) + \mu^2(m, s; a, b). \qquad \sigma^2(m, s; a, b) = s^2\left(1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right),$$

Kullback-Leibler divergence between nested truncated normal distributions:

$$D_{\mathrm{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = \frac{m_2}{2s_2^2} - \frac{m_1}{2s_1^2} + \log\frac{Z_{a_2,b_2}(m_2, s_2)}{Z_{a_1,b_1}(m_1, s_1)} - \left(\frac{m_2}{s_2^2} - \frac{m_1}{s_1^2}\right)\eta_1(m_1, s_1; a_1, b_1)$$

$$- \left(\frac{1}{2s_1^2} - \frac{1}{2s_2^2}\right)\eta_2(m_1, s_1; a_1, b_1) \quad \text{if nested distributions } (a_1, b_1) \subseteq (a_2, b_2)$$

$$D_{\mathrm{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = +\infty, (a_1, b_1) \not\subseteq (a_2, b_2) \quad \text{otherwise}$$

Curved Bregman divergences

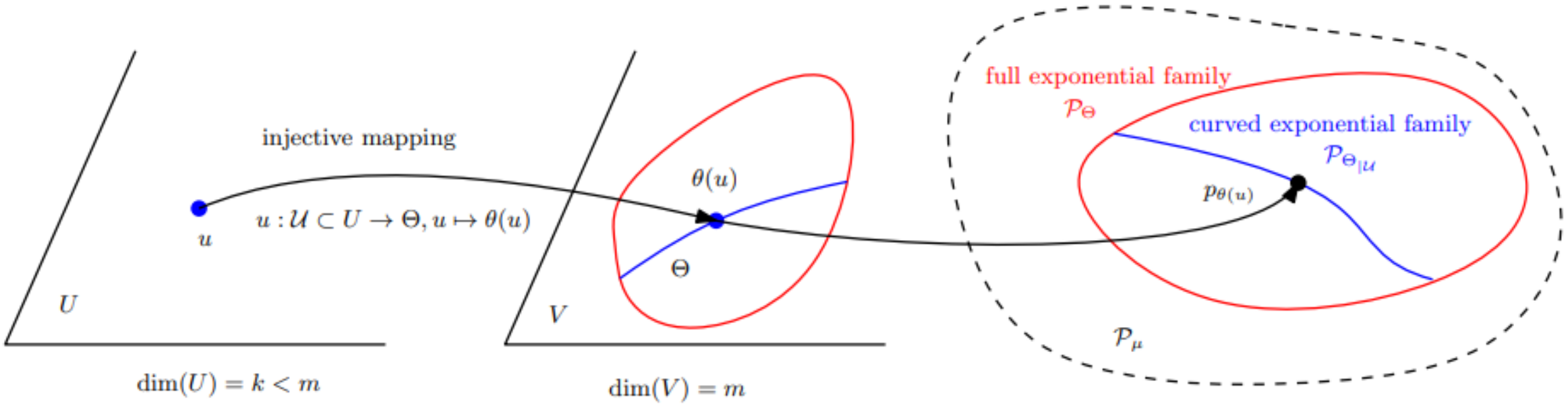
Symmetrized Bregman divergences

…

Curved Bregman divergences (cBDs)

Consider a domain U which maps to a subset of Θ by $\theta = c(u)$ with $\dim(U) < \dim(\Theta)$: $B_{F,c}(u_1 : u_2) := B_F(c(u_1) : c(u_2))$ is not Bregman when $\{c(u) \mid u \in U\}$ not convex. cBDs usually not BDs unless constraints c(.) are affine

By analogy to **curved exponential families**



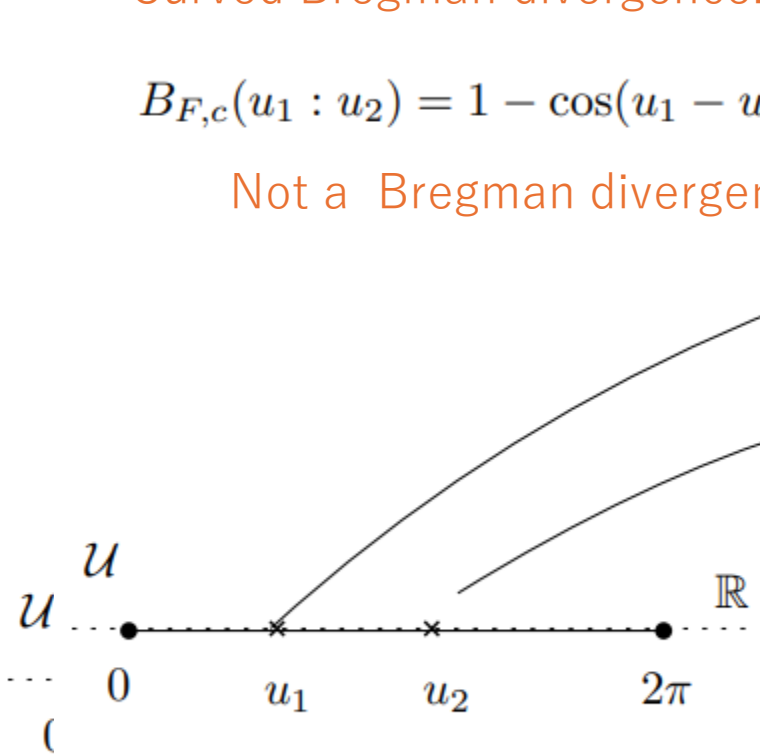
For example, U is real vector space, V is symmetric matrix vector space, CEF is sub-family of Gaussians

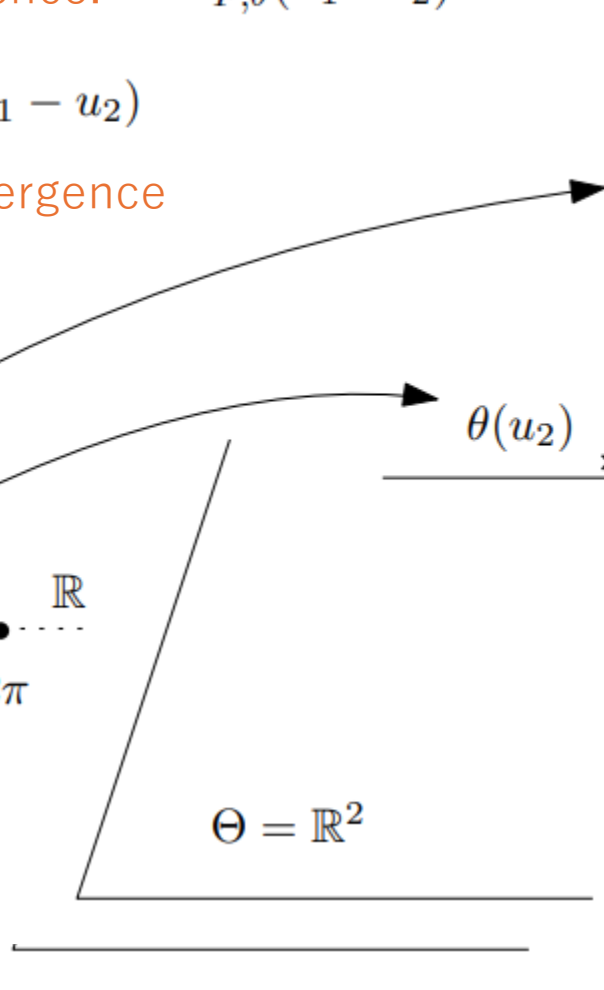
Curved Bregman divergence example: The cosine dissimilarity between unit vectors

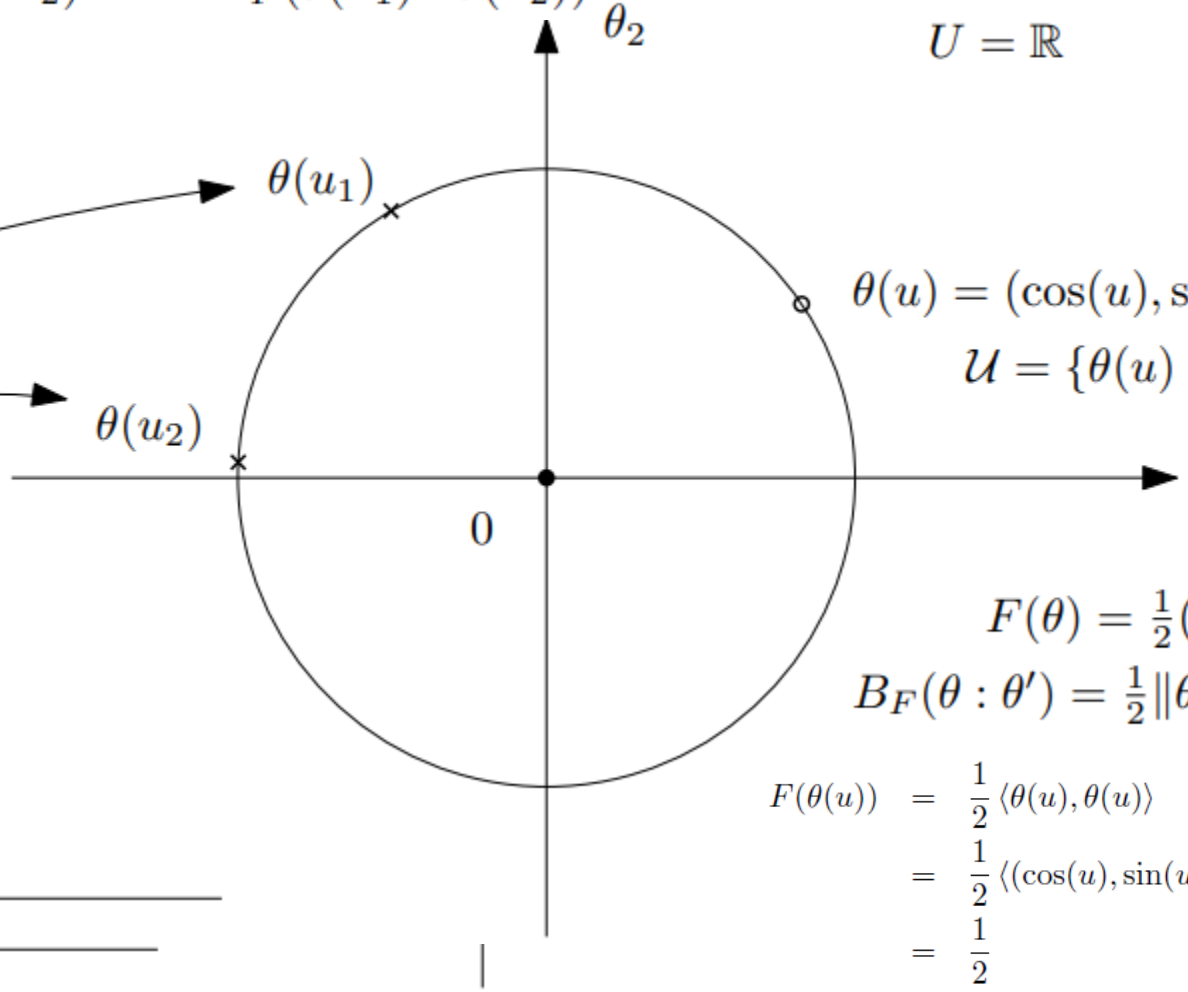
Curved Bregman divergence: $B_{F,\theta}(u_1 : u_2) = B_F(\theta(u_1) : \theta(u_2))$

$B_{F,c}(u_1 : u_2) = 1 - \cos(u_1 - u_2)$

Not a Bregman divergence







$U = \mathbb{R}$

$\theta(u) = (\cos(u), \sin(u))$

$\mathcal{U} = \{\theta(u) \ : u \in [0, 2\pi)\}$

$F(\theta) = \frac{1}{2}(\theta_1^2 + \theta_2^2)$

$B_F(\theta : \theta') = \frac{1}{2}\|\theta - \theta'\|^2$

$$\begin{aligned} F(\theta(u)) &= \frac{1}{2}\langle \theta(u), \theta(u) \rangle \\ &= \frac{1}{2}\langle (\cos(u), \sin(u)), (\cos(u), \sin(u)) \rangle \\ &= \frac{1}{2} \end{aligned}$$

arXiv:2504.05654

A constant on the circle submanifold, thus not a proper Bregman generator!

Example of curved BDs: Symmetrized BDs

= **Jeffreys-Bregman divergences** are curved Bregman divergences:

$S_F(\theta_1, \theta_2) = \langle \theta_1 - \theta_2, \eta_1 - \eta_2 \rangle$

$$\begin{aligned} S_F(\theta_1 : \theta_2) &= B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1), \\ &= B_F(\theta_1 : \theta_2) + B_{F^*}(\nabla F(\theta_1) : \nabla F(\theta_2)) \\ &= \check{B}_{F_\xi}(\xi(\theta_1) : \xi(\theta_2)), \end{aligned}$$

Curved domain: $\mathcal{U} = \{(\theta, \nabla F(\theta)) : \theta \in \Theta\}$

$\xi(\theta) = (\theta, \nabla F(\theta))$ $F_\xi(\theta, \eta) := F(\theta) + F^*(\eta)$ $F^*(\eta) = \langle \theta, \eta \rangle - F(\theta)$

m-dimensional submanifold in 2m-dimensional space
(usually not cvx affine space, hence not a Bregman divergence)
Only affine for quadratic generators (i.e., squared Mahalanobis divergences)

"Sided and symmetrized Bregman centroids." IEEE transactions on Information Theory 55.6 (2009)

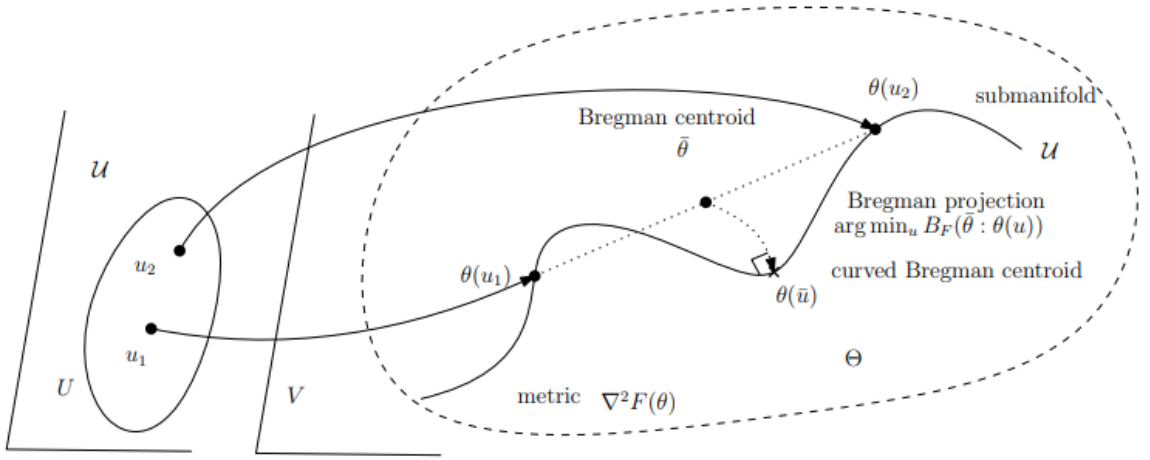
Theorem: Curved Bregman centroid is the Bregman projection of the full Bregman centroid

Right Bregman projection of Bregman centroid $\bar{\theta} = \sum_i w_i \theta_i$

$$\arg\min_{u \in \mathcal{U}} \sum_{i=1}^{n} w_i \, B_F(\theta_i : \theta(u)) = \arg\min_{u \in \mathcal{U}} B_F(\bar{\theta} : \theta(u))$$

N-point opt. $\theta_i = \theta(u_i)$

From n-point opt to One-point opt!

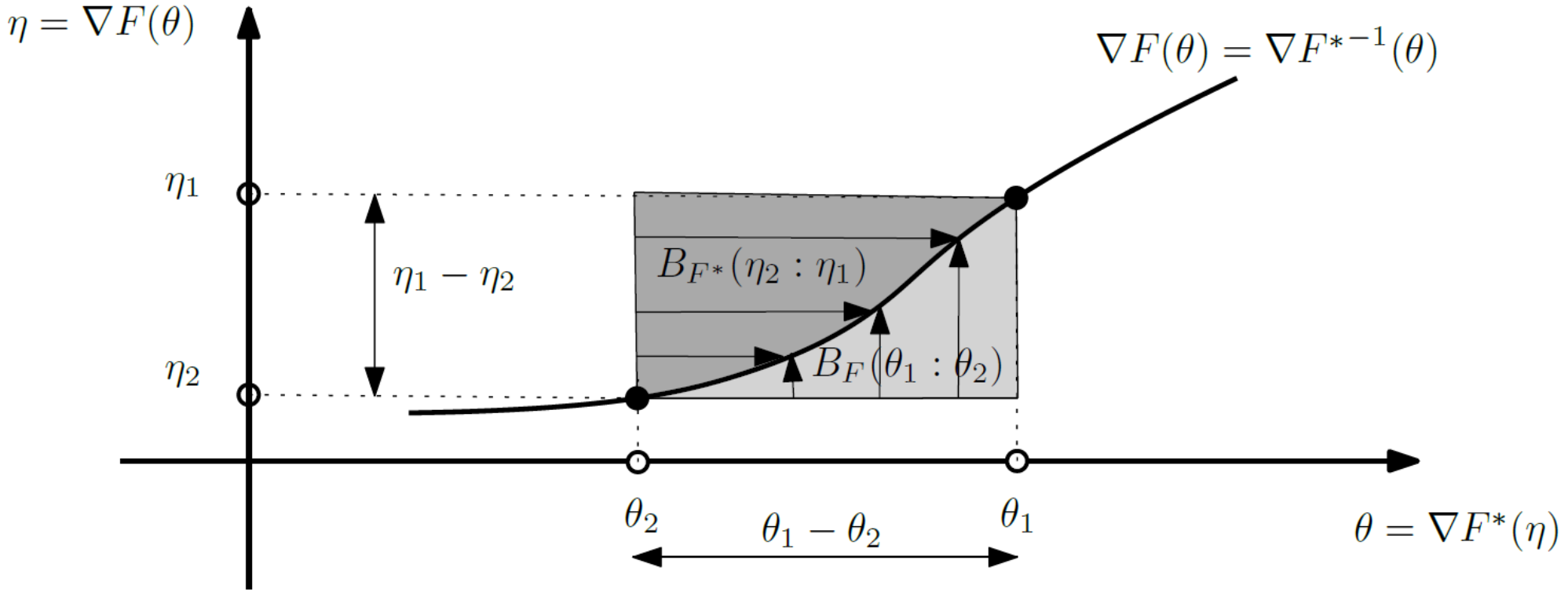


Simple Proof.

$$\begin{aligned}
\min_{u \in \mathcal{U}} \sum_{i=1}^{n} w_i \, B_F(\theta_i : \theta(u)) &= \sum_{i=1}^{n} w_i (F(\theta_i) - F(\theta(u)) - \langle \theta_i - \theta(u), \nabla F(\theta(u)) \rangle), \\
&\equiv -F(\theta(u)) - \langle \bar{\theta} - \theta(u), \nabla F(\theta(u)) \rangle, \\
&\equiv F(\bar{\theta}) - F(\theta(u)) - \langle \bar{\theta} - \theta(u), \nabla F(\theta(u)) \rangle \\
&= B_F(\bar{\theta} : \theta(u)).
\end{aligned}$$

"What is... an information projection?" Notices of the AMS 65.3 (2018): 321-324.

Symmetrized Bregman divergence: Geometric reading



$$
\begin{aligned}
B_F(\theta_1 : \theta_2) &= \int_{\theta_2}^{\theta_1} (F'(\theta) - F'(\theta_2))\mathrm{d}\theta \\
B_{F^*}(\eta_2 : \theta_1) &= \int_{\eta_1}^{\eta_2} (F^{*\prime}(\eta) - F^{*\prime}(\eta_1))\mathrm{d}\eta
\end{aligned}
$$

$$
\begin{aligned}
S_F(\theta_1, \theta_2) &= B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) \\
&= B_F(\theta_1 : \theta_2) + B_{F^*}(\eta_1 : \eta_2) \\
&= (\theta_1 - \theta_2)^\top (\eta_1 - \eta_2)
\end{aligned}
$$

2107.05901

Estimating KLDs/f-divergences
extended to positive densities

Monte Carlo estimation of ext. KL/f-divergences

$$D_{\mathrm{KL}}(p:q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)$$

$x_1, \cdots, x_n$ sampled from p

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i)}{q(x_i)}$$

Problem: Estimated KLD may be negative. Bug/error in practice!

Definition 1 (Extended f-divergence) *The extended f-divergence for a convex generator f, strictly convex at 1 and satisfying $f(1) = 0$ is defined by*

$$I_f^e(p:q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right) \mathrm{d}\mu(x).$$

$$= \int p(x) \underbrace{B_f\left(\frac{q(x)}{p(x)} : 1\right)}_{\geq 0} \mathrm{d}\mu(x) \geq 0.$$

f-divergences are equivalent modulo affine terms

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1 \right) \geq 0$$

Non-negative Monte Carlo estimation of f-divergences, 2020

Sigma points for the KLD between EF densities

- Bregman divergences $B_F = B_G$ iff $F \equiv G$ modulo an **affine term**
- Thus consider the equivalent generator $\boxed{F_\omega(\theta) := -\log(p_\theta(\omega))}$ for any $\omega \in \mathcal{X}$:

$$F_\omega(\theta) := -\log(p_\theta(\omega)) = F(\theta) - \underbrace{(\theta^\top t(\omega) + k(\omega))}_{\text{affine term in } \theta}$$

- Since we have $\nabla F_\omega(\theta(\lambda_1)) = -(t(\omega) - \nabla F(\theta(\lambda_1)))$:

$$\begin{aligned} D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] &= B_{F_\omega}(\theta(\lambda_2) : \theta(\lambda_1)), \\ &= \log\left(\frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)}\right) - (\theta(\lambda_2) - \theta(\lambda_1))^\top \nabla_\theta F_\omega(\theta(\lambda_1)) \end{aligned}$$

$$\boxed{D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \log\left(\frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)}\right) + (\theta(\lambda_2) - \theta(\lambda_1))^\top (t(\omega) - \nabla F(\theta(\lambda_1))), \quad \forall \omega \in \mathcal{X}}$$

KLD for multi-order exponential families

- Choose $s \leqslant D+1$ values for ω (i.e., $\omega_1, \ldots, \omega_s$), and **average**:

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s} \sum_{i=1}^{s} \log \left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)} \right) + (\theta(\lambda_2) - \theta(\lambda_1))^\top \left(\frac{1}{s} \sum_{i=1}^{s} t(\omega_i) - \nabla F(\theta(\lambda_1)) \right)$$

Theorem (KLD as sum of log density-ratios)

The Kullback-Leibler divergence between two densities p_{λ_1} and p_{λ_2} belonging to a full regular exponential family $\mathcal{E}$ of order D can be expressed as the average sum of logarithms of density ratios:

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s} \sum_{i=1}^{s} \log \left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)} \right) \text{ such that } \frac{1}{s} \sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]$$

where $\omega_1, \ldots, \omega_s$ are $s \leqslant D+1$ distinct points of $\mathcal{X}$ chosen such that $\frac{1}{s} \sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]$.

Example: KLD between two multivariate normals (MVNs)

For illustration purpose only!!!

- Exponential family of d-dimensional no[illegible]sities: **order** $D = \frac{d(d+3)}{2}$.
- Use $s = 2d \leqslant D + 1$ **vectors** ω_i to express [illegible] $\Sigma_1 : p_{\mu_2,\Sigma_2}]$:

$$\omega_i = \mu_1 - \sqrt{d\lambda_i} e_i, \omega_{i+d} = \mu_1 + \sqrt{d\lambda_i} e_i, [illegible]\{1, \ldots, d\}$$

where the λ_i's are the **eigenvalues** of Σ_1 with the e_i's [illegible]esponding **eigenvectors**.

- We have $\frac{1}{2d} \sum_{i=1}^{2d} \omega_i = E_{p_{\lambda_1}}[x] = \mu_1$ and $\frac{1}{2d} \sum_{i=1}^{2d} \omega_i \omega_i^\top = \mu_1 \mu_1^\top$
- Let $\sqrt{\lambda_i} e_i = [\sqrt{\Sigma_1}]_{.,i}$, the i-th column of the **square root of the covariance matrix** of Σ_1. We have:

$$D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] = \frac{1}{2d} \sum_{i=1}^{d} \left(\log \left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1 - [\sqrt{d\Sigma_1}]_{.,i}\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1 - [\sqrt{d\Sigma_1}]_{.,i}\right)} \right) + \log \left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1 + [\sqrt{d\Sigma_1}]_{.,i}\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1 + [\sqrt{d\Sigma_1}]_{.,i}\right)} \right) \right)$$

where $[\sqrt{d\Sigma_1}]_{.,i} = \sqrt{\lambda_i} e_i$ denotes the vector extracted from the i-th column of the square root matrix of $d\Sigma_1$.

Characterizing Monte Carlo estimation error

- Monte Carlo KLD estimation: $\tilde{D}_{\mathrm{KL}}^{\mathcal{S}_m}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{m}\sum_{i=1}^{m} \log \frac{p_{\lambda_1}(x_i)}{p_{\lambda_2}(x_i)}$

- The error can be **exactly quantified** for EF densities as

$$\left|\tilde{D}_{\mathrm{KL}}^{\mathcal{S}_m}[p_{\lambda_1} : p_{\lambda_2}] - D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}]\right| = \left|(\theta(\lambda_2) - \theta(\lambda_1))^{\top}\left(\frac{1}{m}\sum_{i=1}^{m} t(x_i) - E_{p_{\lambda_1}}[t(x)]\right)\right|$$

$$\text{When } m \to \infty, \text{ we have } \frac{1}{m}\sum_{i=1}^{m} t(x_i) \to E_{p_{\lambda_1}}[t(x)]$$

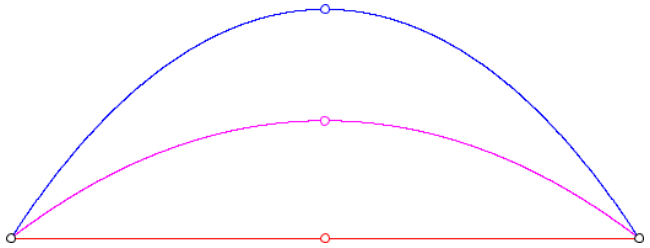
Consistent estimator

"Computing statistical divergences with sigma points." GSI 2021, Springer LNCS.

Implementing information geometry

Stochastic Bregman generator / Random information geometry
Numerical information geometry / Neural information geometry

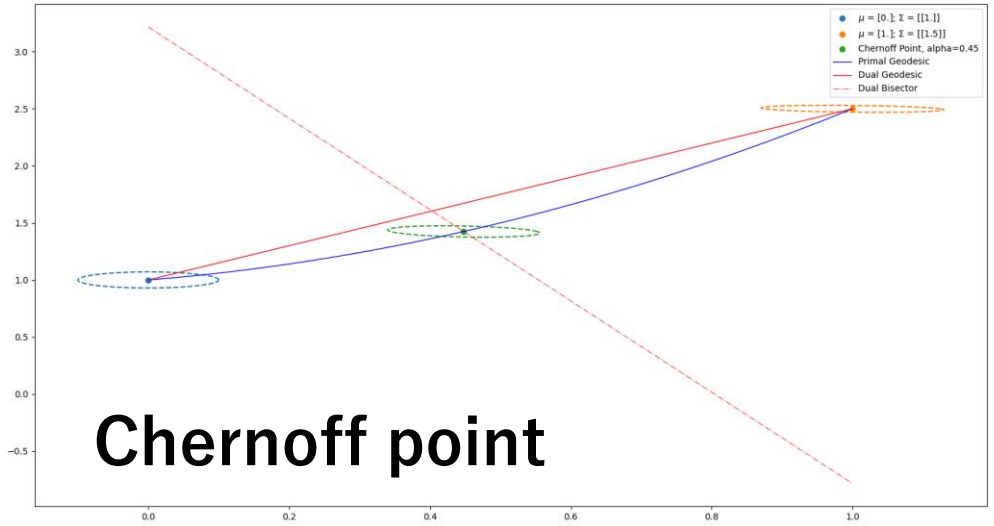
2606.09077



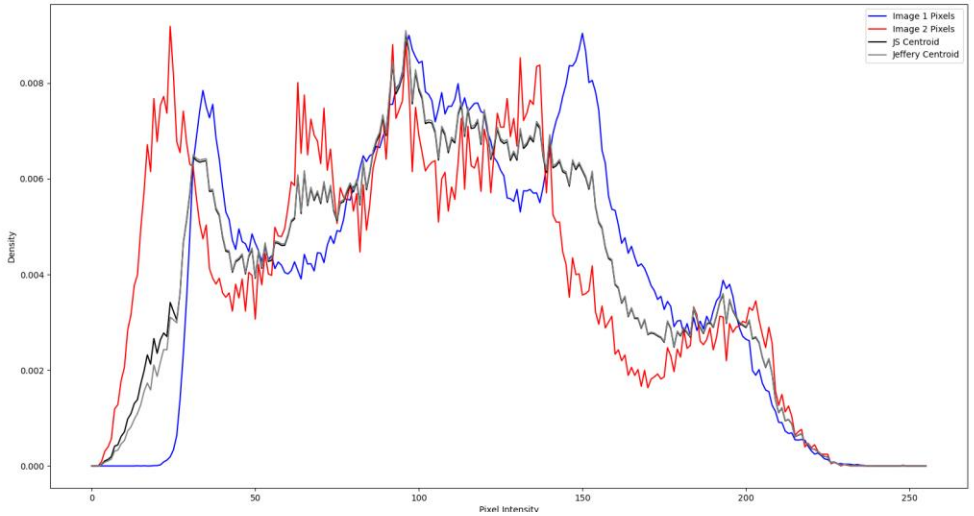
A Python library for geometric computing on Bregman Manifolds

pyBregMan

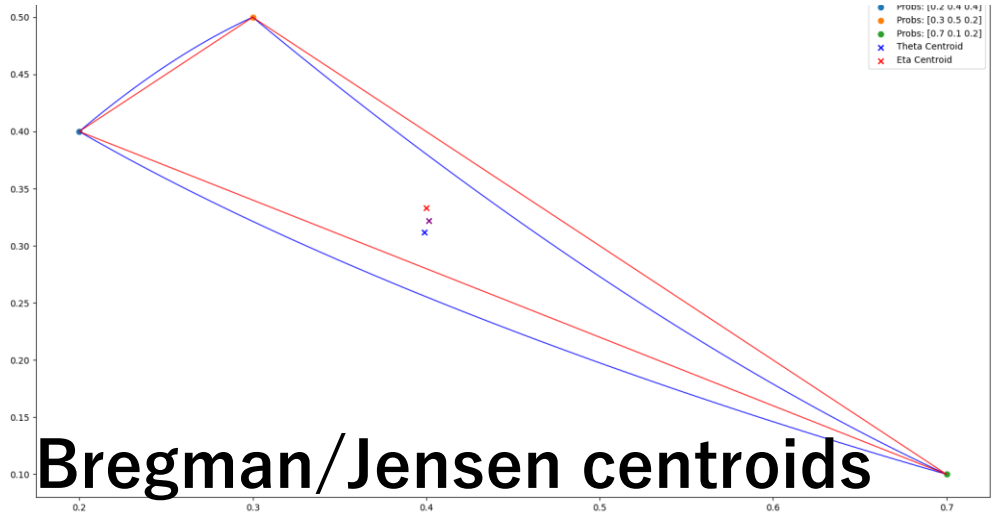
https://franknielsen.github.io/pyBregMan/



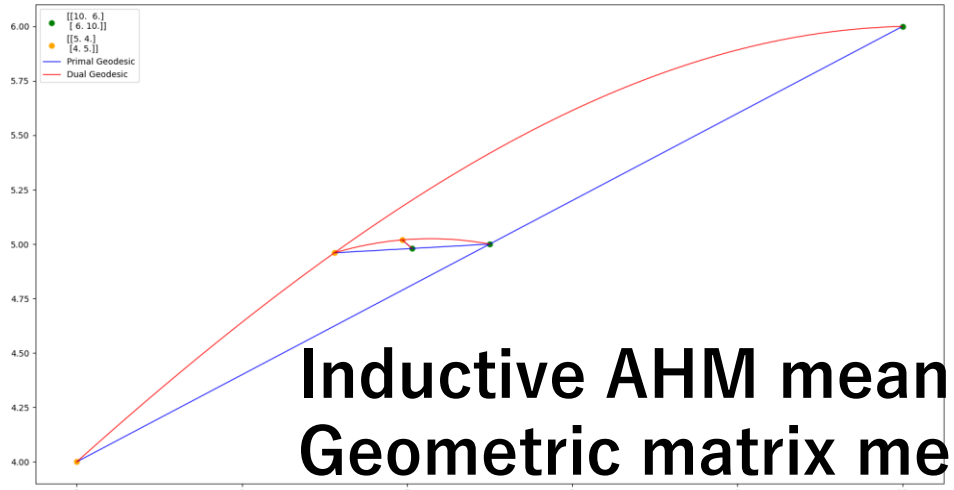
Chernoff point



Jensen-Shannon centroid



Bregman/Jensen centroids



Inductive AHM mean
Geometric matrix mean

Joint work of Frank Nielsen and Alexander Soen

Stochastic/random information geometry

- In many cases, dual structures of information geometry are not available in closed-forms or comp. tractable (polynomial EF, continuous mixture family). **With high probability**, we can sample the generator integrals and get proper Bregman generators

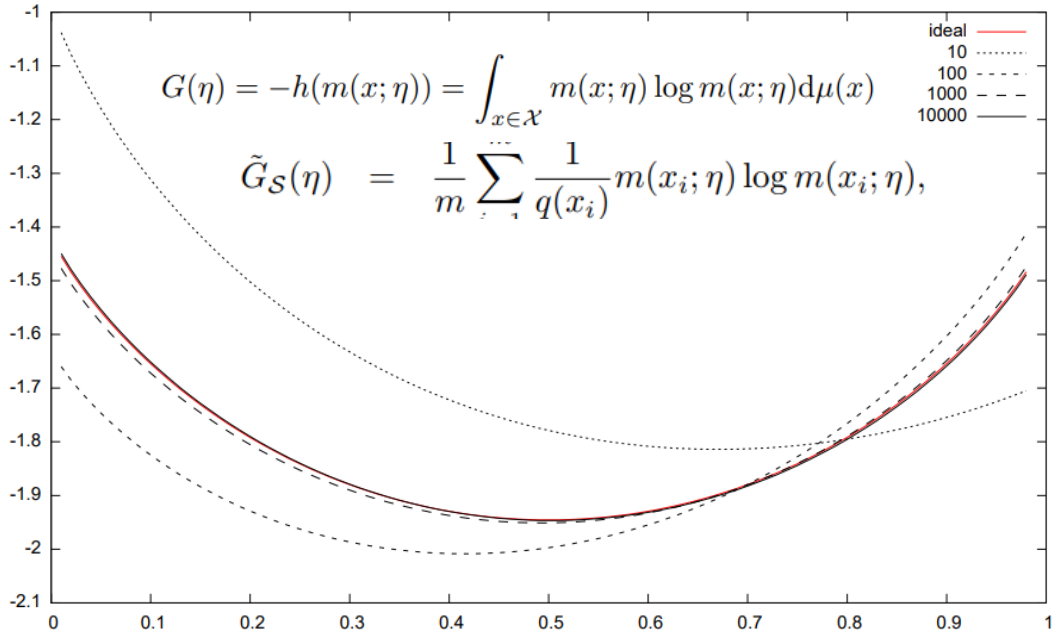
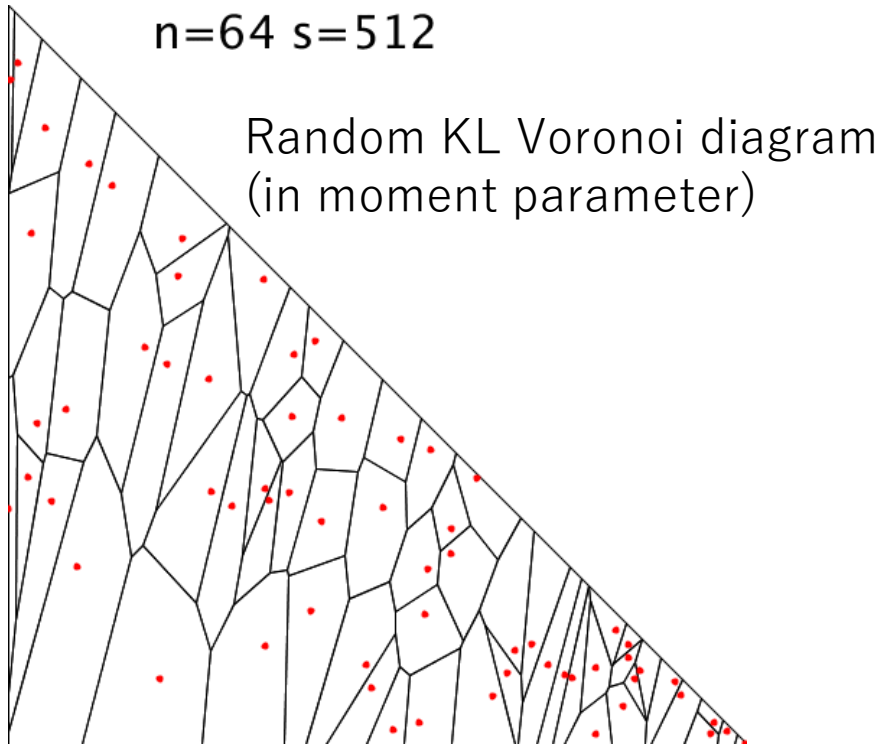
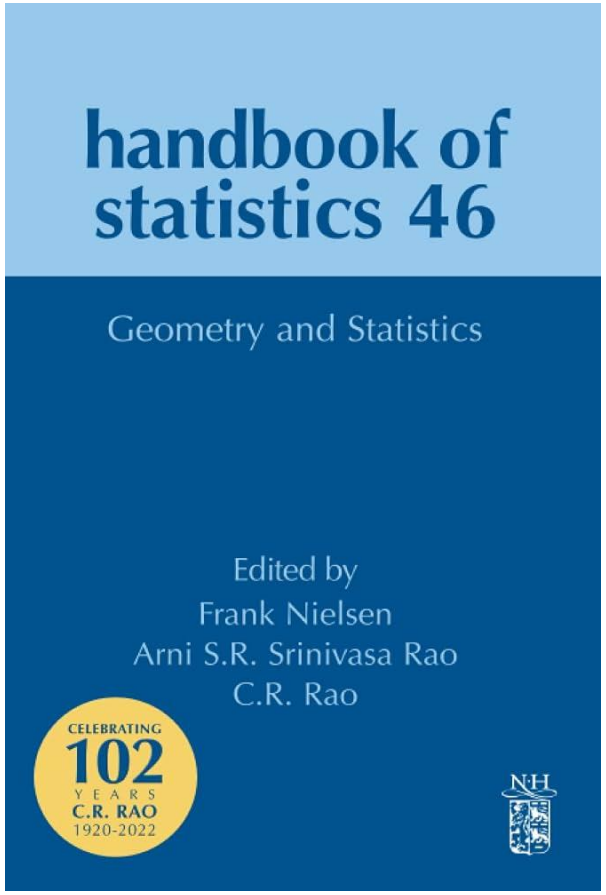


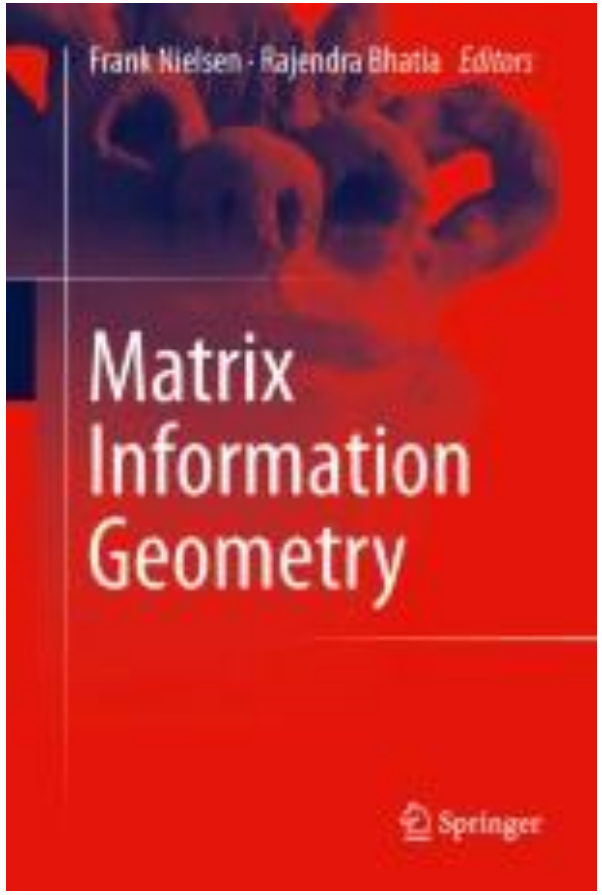
Figure 2: A series $G_{\mathcal{S}}(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x;\eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x;\eta) = (1-\eta)p(x;0,3) + \eta p(x;2,1)$ for the proposal distribution $q(x) = m(x;\frac{1}{2})$.

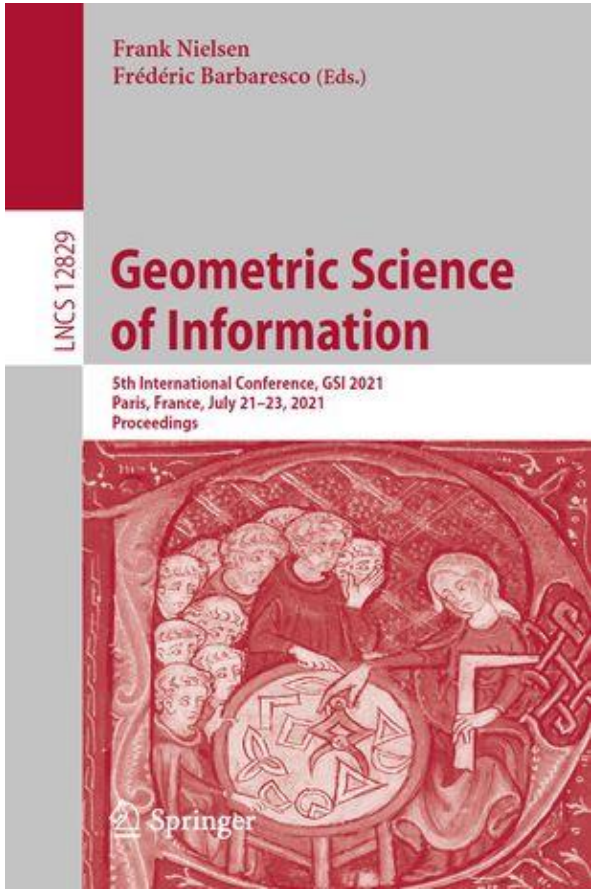
1803.07225

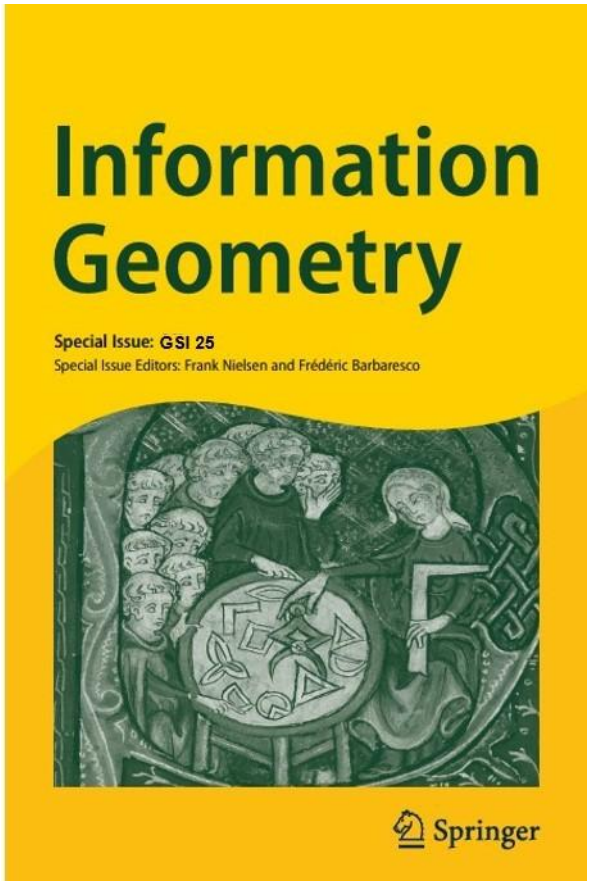


Geometric Science of Information









https://franknielsen.github.io/GSI/

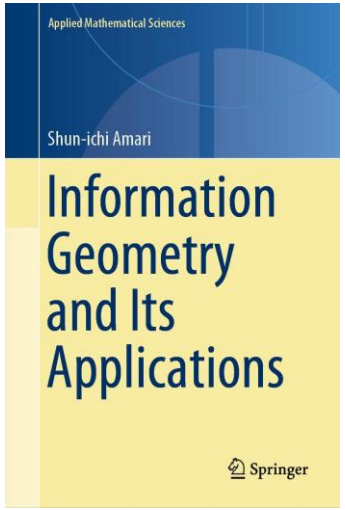
Geometric information theory: Hub to information sciences

divergence

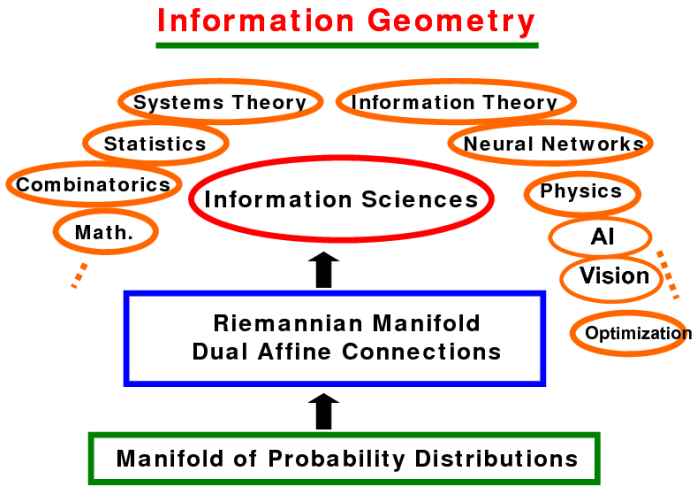
geometry

The fabric of information geometry

and the **untangling** of its geometry, divergence, statistical models



Professor Amari pioneered IG for information sciences



Some references

- "**An information-geometric characterization of Chernoff information**", IEEE Signal Processing Letters 20.3 (2013)
- "**Revisiting Chernoff information with likelihood ratio exponential families**," Entropy 24.10 (2022)
- "**Variational representations of annealing paths: Bregman information under monotonic embedding**", Information Geometry 7.1 (2024)
- "**Sided and symmetrized Bregman centroids**," IEEE transactions on Information Theory 55.6 (2009)
- "**On conformal divergences and their population minimizers**," IEEE Transactions on Information Theory 62.1 (2015)
- "**Relative Fisher information and natural gradient for learning large modular models**," International Conference on Machine Learning, 2017.
- "**Tractable structured natural-gradient descent using local parameterizations**." International Conference on Machine Learning, 2021.
- "**Bregman Voronoi diagrams**", Discrete & Computational Geometry 44.2 (2010)
- "**Divergences induced by the cumulant and partition functions of exponential families and their deformations induced by comparative convexity**," Entropy 26.3 (2024)
- "**Curved representational Bregman divergences and their applications**," International Conference on Geometric Science of Information, 2025
- "**Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences**," Entropy 24.3 (2022)
- "**A geometric modeling of Occam's razor in deep learning**", Information Geometry (2025)
- "**Computing statistical divergences with sigma points**," GSI 2021.
- "**Generalized Legendre Transforms Have Roots in Information Geometry**," Entropy 28.1 (2025)
- "**Monte Carlo information-geometric structures**," Geometric Structures of Information, 2018