

Information geometry for information sciences: - A first overview -

Frank Nielsen

Sony Computer Science Laboratories, Inc



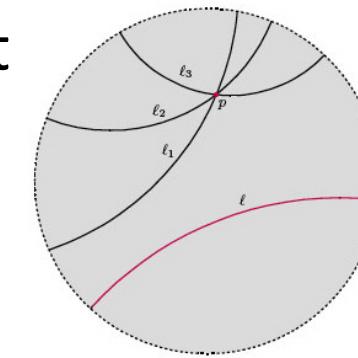
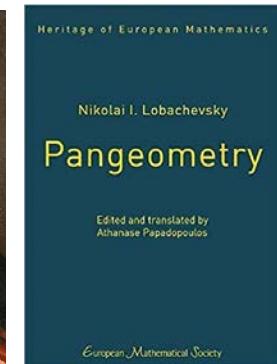
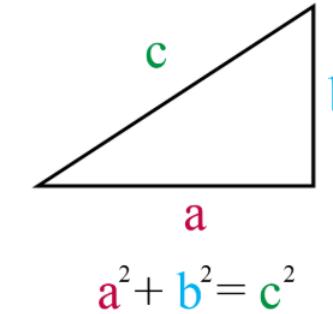
Sony CSL

The goal of this talk is to...

- Present the *main ideas* behind the **dualistic structures of information geometry**
- Avoid common misconceptions and pitfalls
- Decouple and explain the *interplay* of geometric structures with **distances/dissimilarities/divergences/diversities**
- Minimize the use of equations to introduce the **key concepts**

A (too) brief history of geometry

- Science for Earth measurements
- Pythagoras's theorem (c570-495 BC)
- Euclid's Axiomatization and deduction (c300 BC)
Euclidean geometry
- Figures, congruences, construction with compass/rulers
- Lobachevskian **hyperbolic geometry** is consistent (c1800)
- **Riemannian geometry** (c1850): infinitely many consistent differential geometries
- Klein's Erlangen program (action of a group)



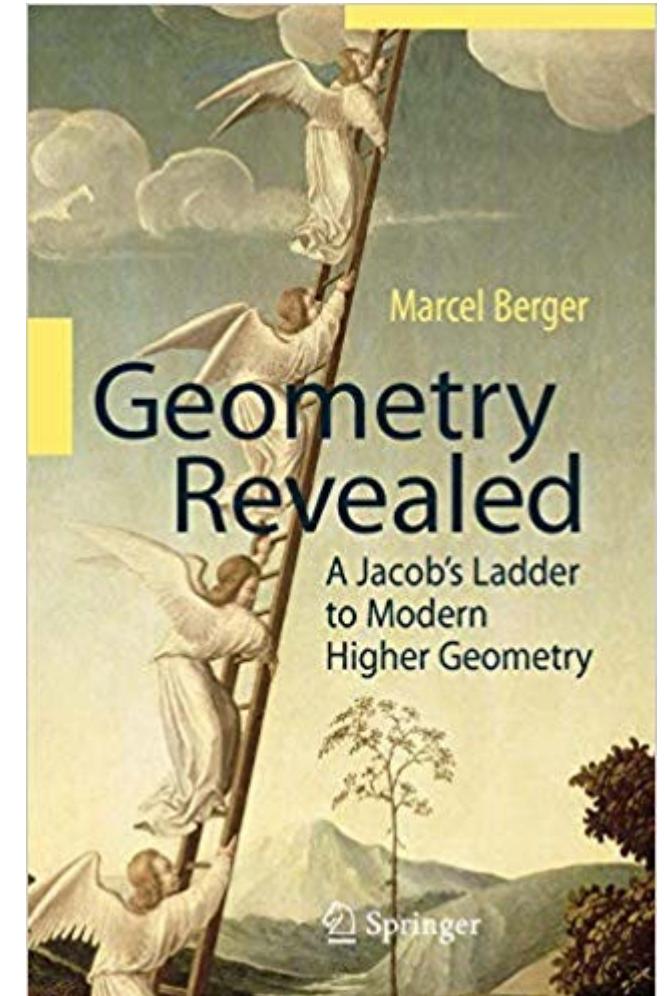
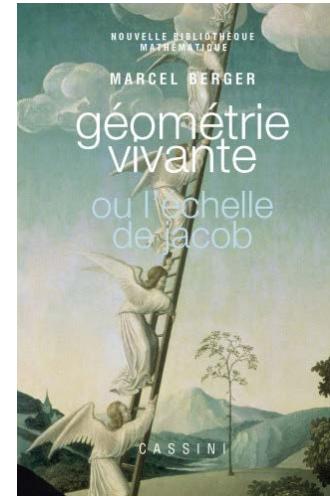
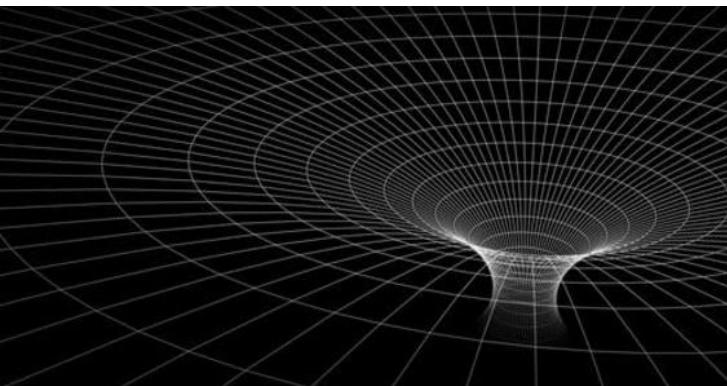
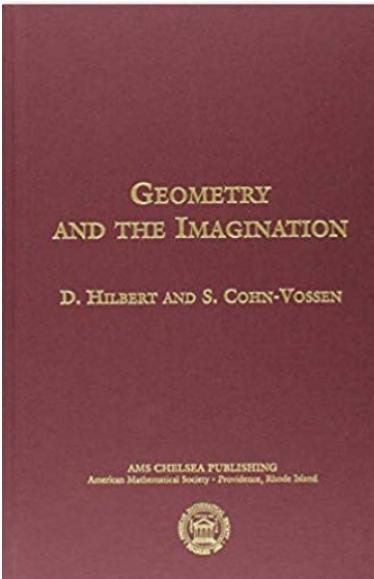
Geometry is an incredibly creative science!



Geometry is the most complete science.

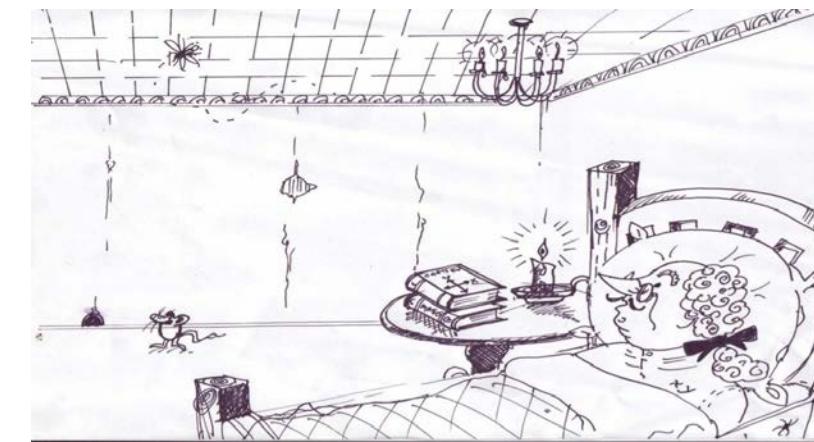
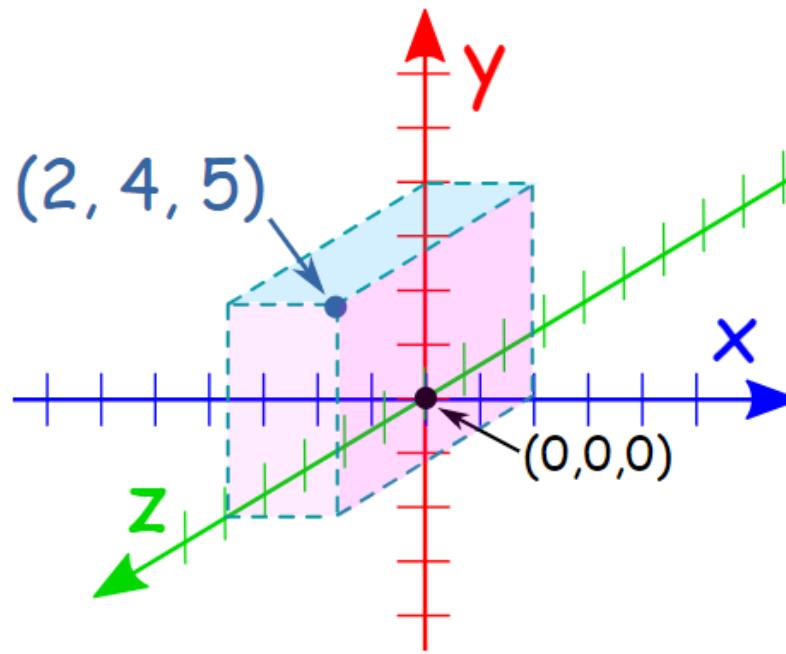
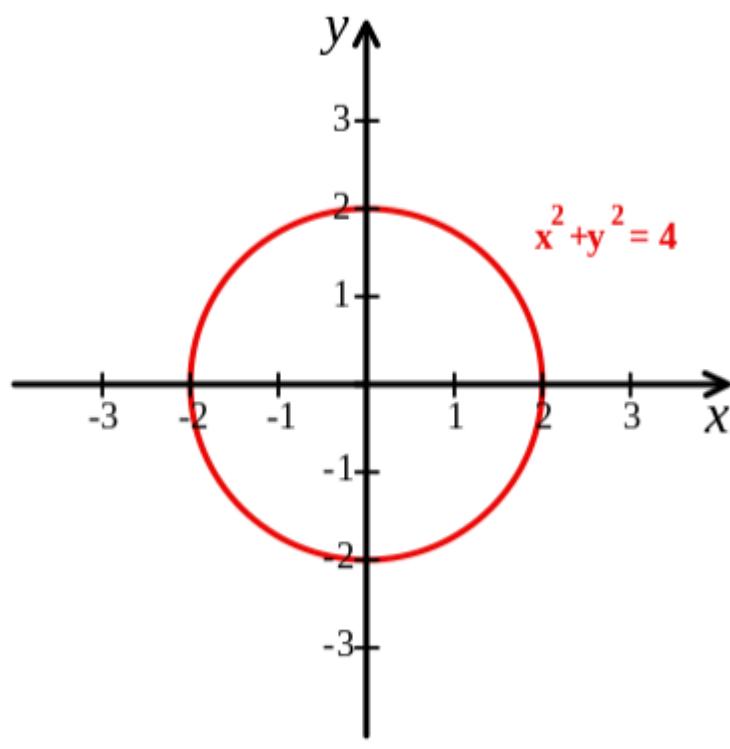
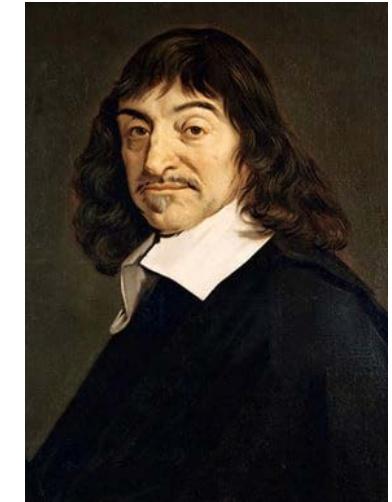
— David Hilbert —

AZ QUOTES



Analytic versus synthetic geometry

- Descartes (c1600) introduced the **Cartesian coordinates** and calculus in geometry

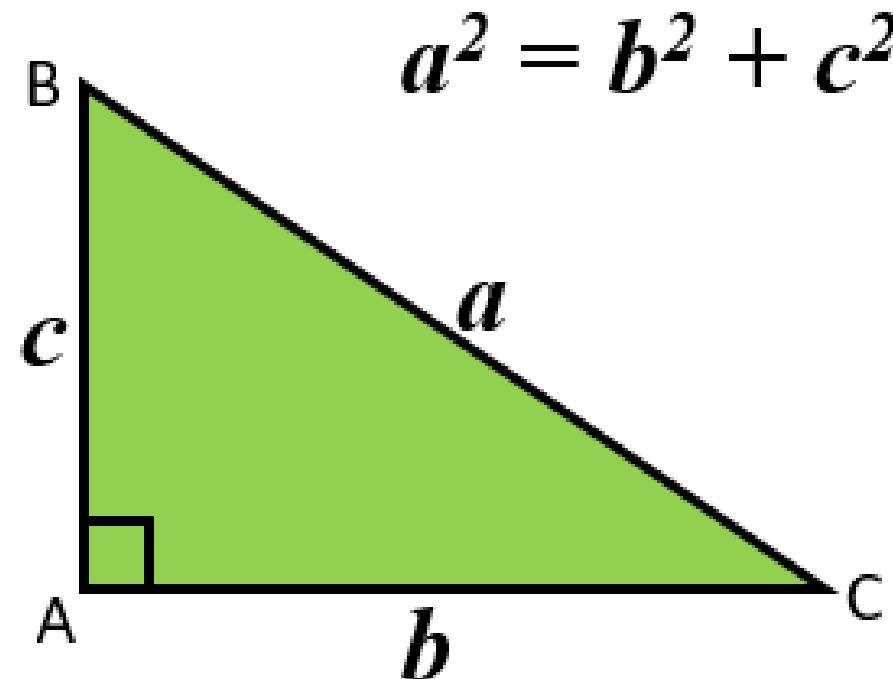




Pythagoras' theorem/Pythagorean thm

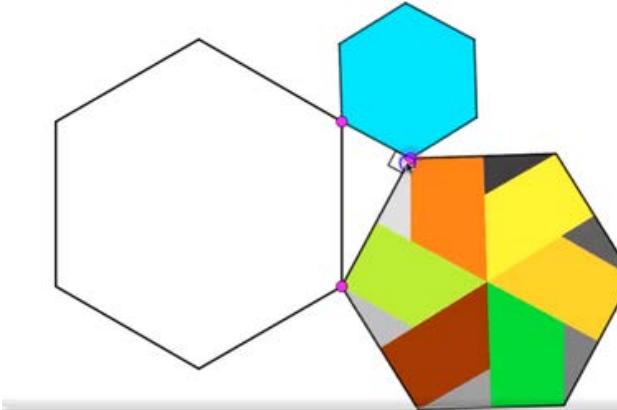
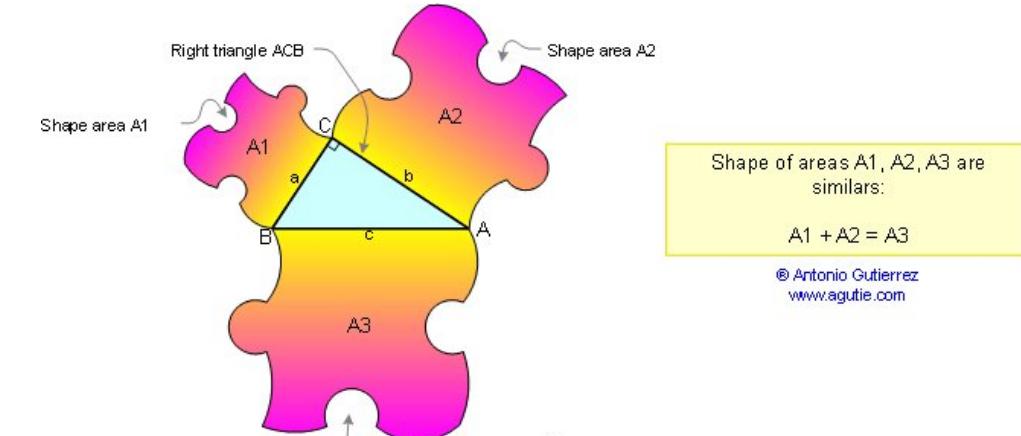
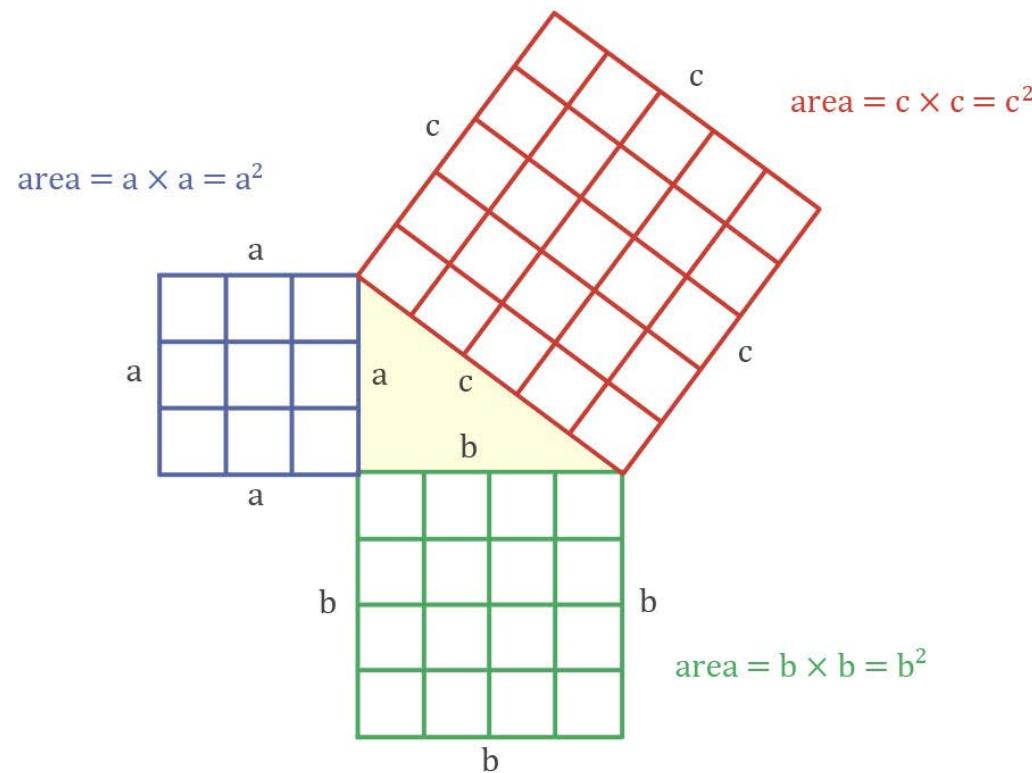
- Formula of Euclidean distance in Cartesian coordinate system

Circa 500 BC



Babylonian mathematics (2000-1600 BC)

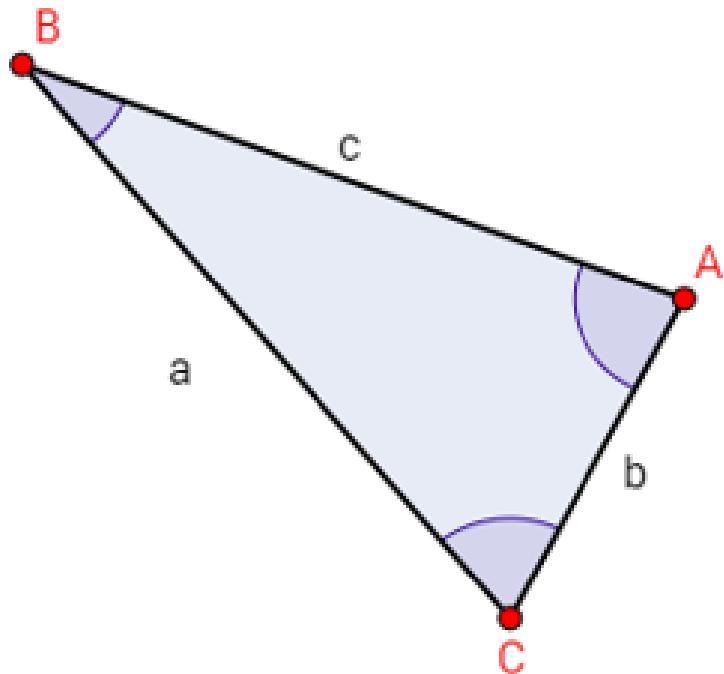
Sometimes presented with extra confusing information



Work for any homothets
(since area scales squarely)

Pythagoras' theorem generalized to the law of cosines for arbitrary triangles

Law of Cosines

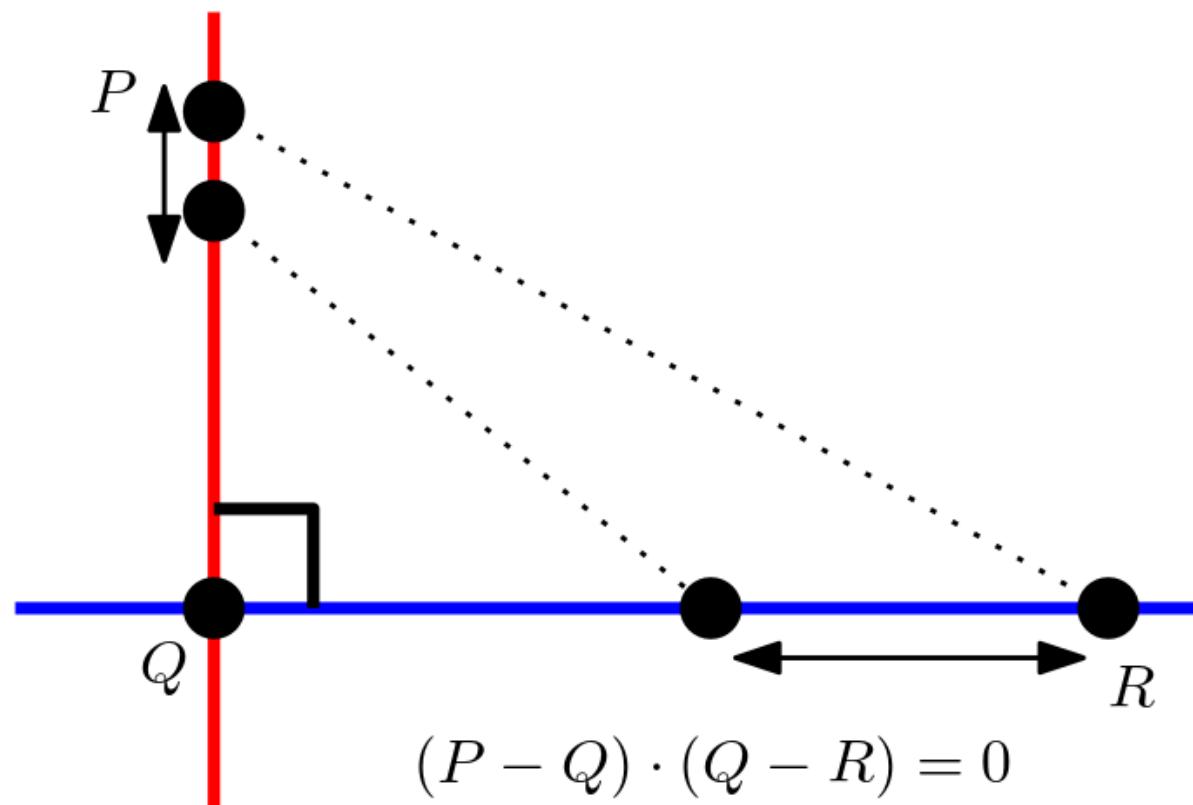


$$a^2 = b^2 + c^2 - 2bc \cos A$$

$$b^2 = a^2 + c^2 - 2ac \cos B$$

$$c^2 = a^2 + b^2 - 2ab \cos C$$

A modern view: A triangle PQR is rectangle iff straight lines perpendicular at Q induce distance identity

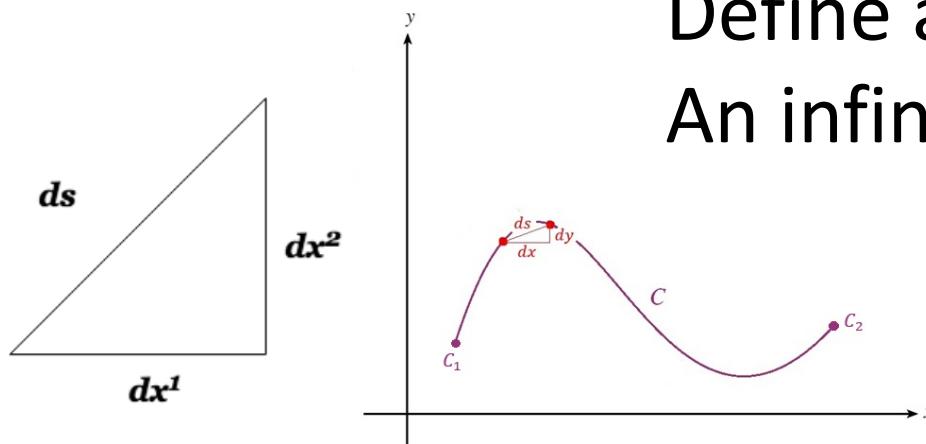


$$\|P - Q\|^2 + \|Q - R\|^2 = \|P - R\|^2$$

Riemannian differential geometry

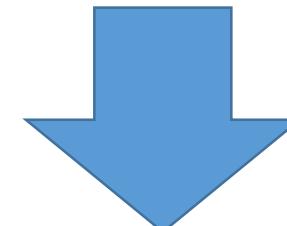


- Gauss pioneered the study of 3D surfaces and curvature



Length of a curve by integration

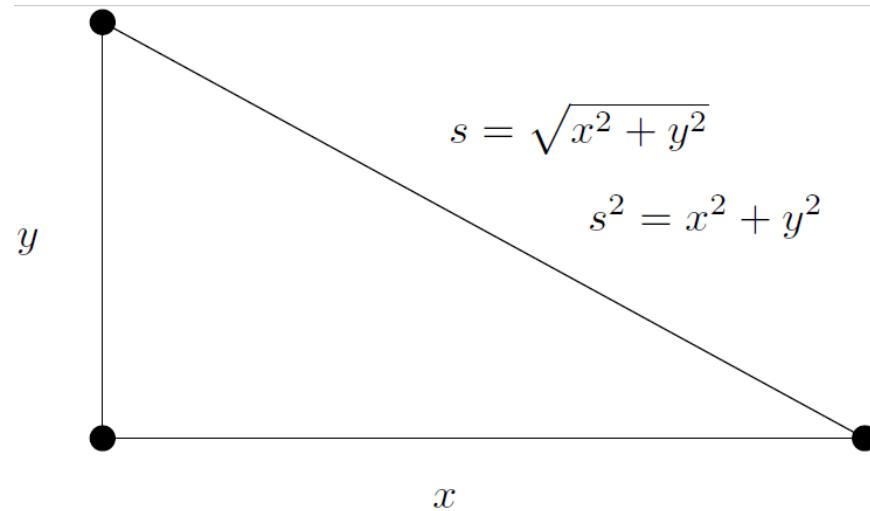
Introduce a **positive-definite matrix G**
Define a geometric object called **metric tensor**:
An infinitesimal Pythagoras theorem



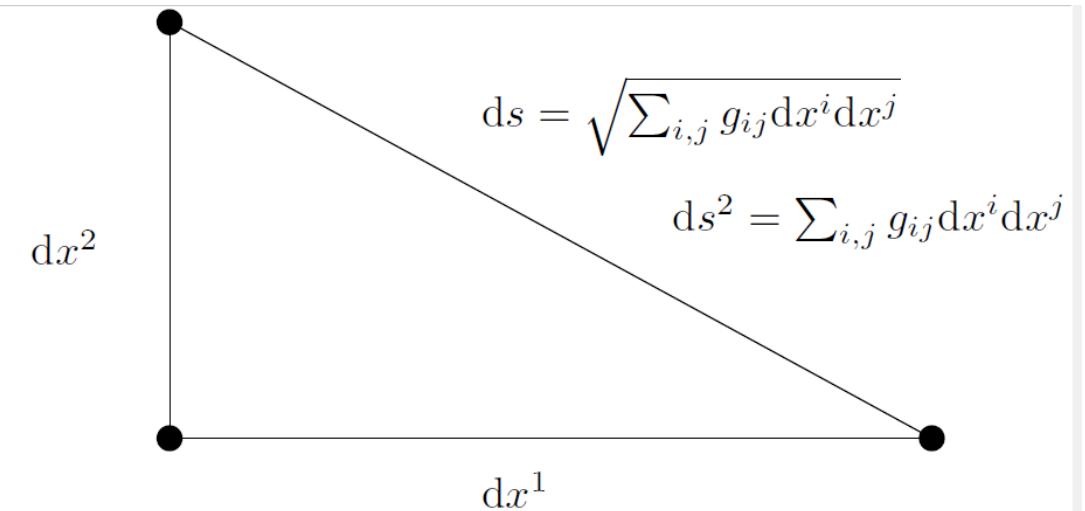
$$ds^2 = [dx \ dy] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix}$$

$$ds^2 = g_{11} du^2 + 2g_{12} du \ dv + g_{22} dv^2$$

Riemannian geometry: Infinitesimal Pythagorean theorem



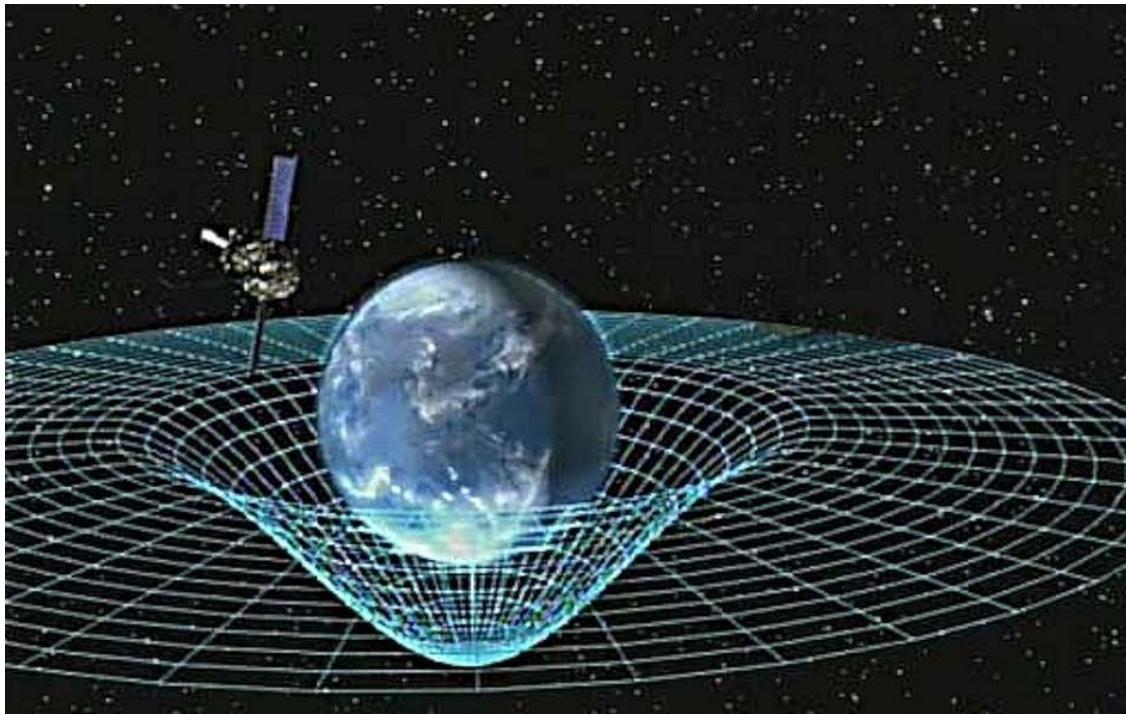
Pythagorean theorem



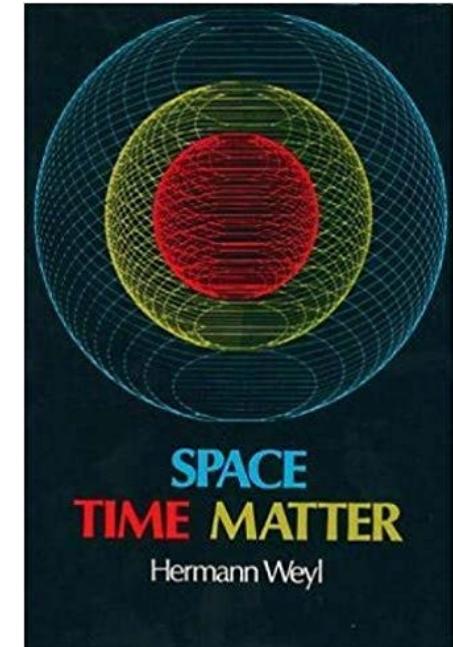
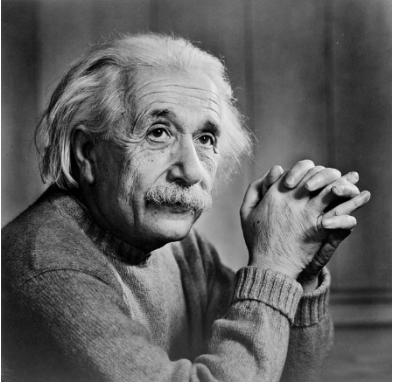
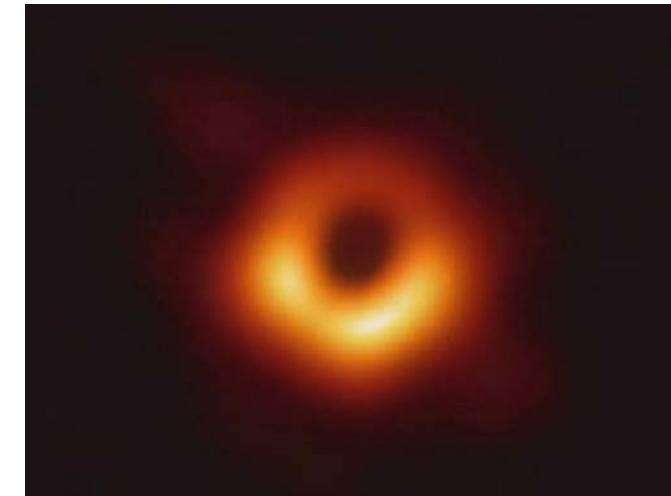
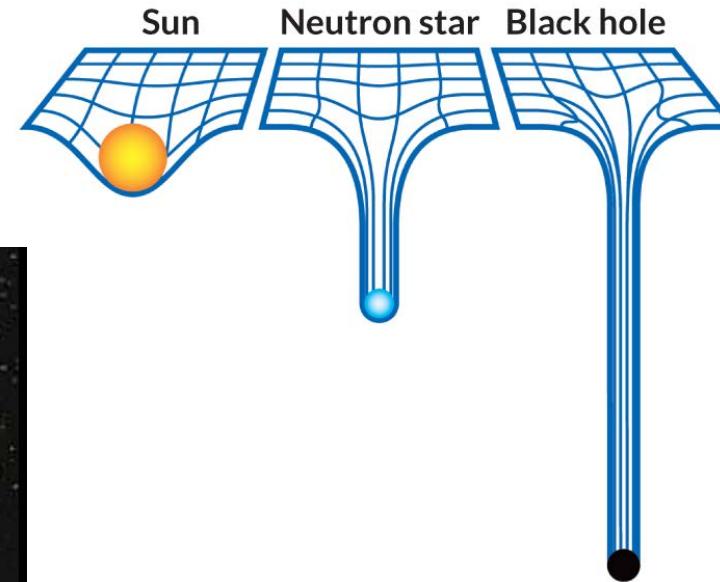
Infinitesimal Riemannian Pythagorean theorem

Riemannian geometry changed our perception of the universe and data analytics

- General relativity of spacetime



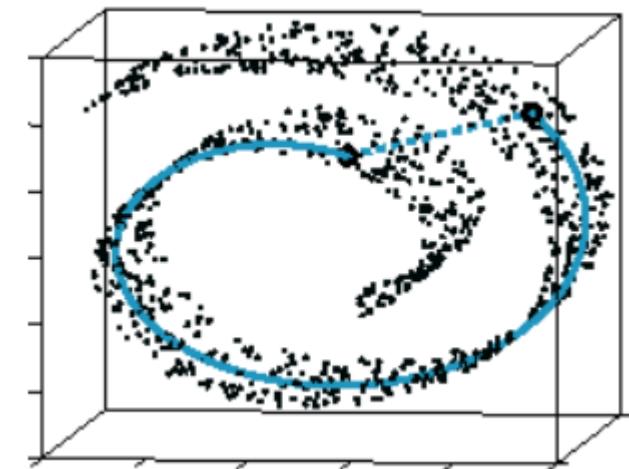
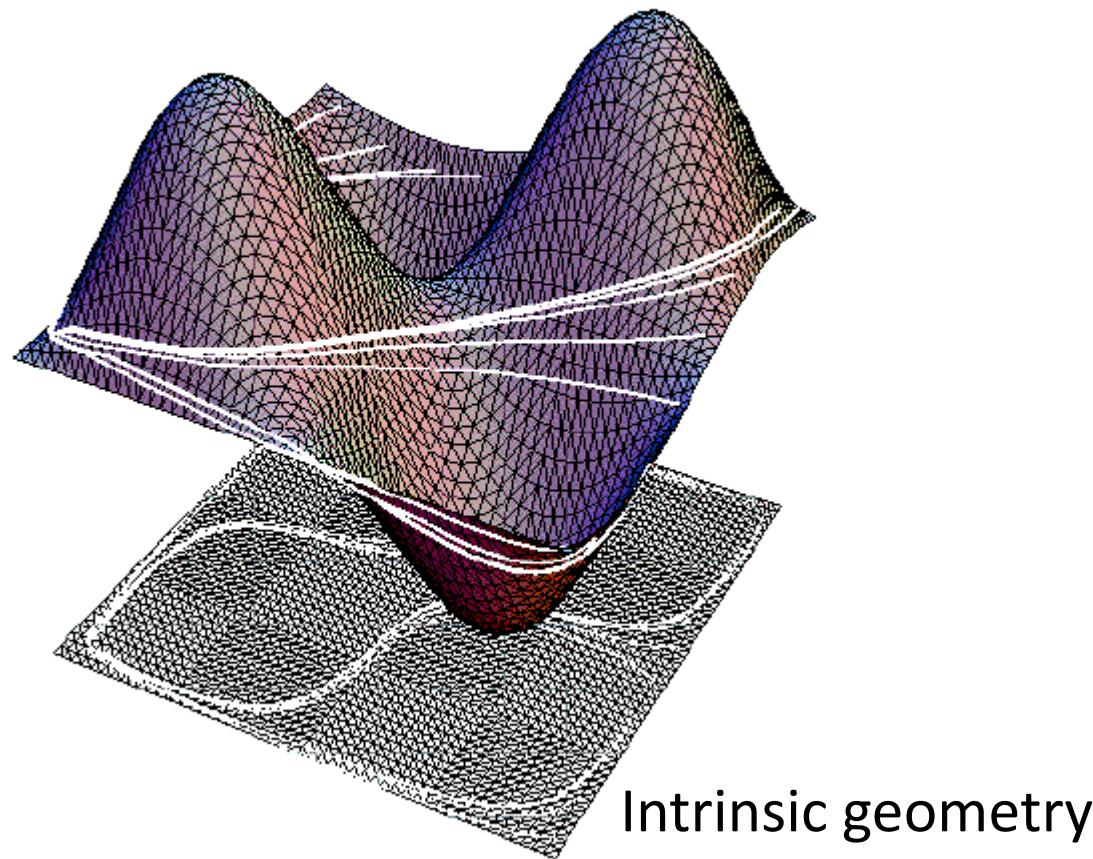
Spacetime+Matter



Riemannian manifolds

- Visualized extrinsically as smooth surfaces of the ambient Euclidean space

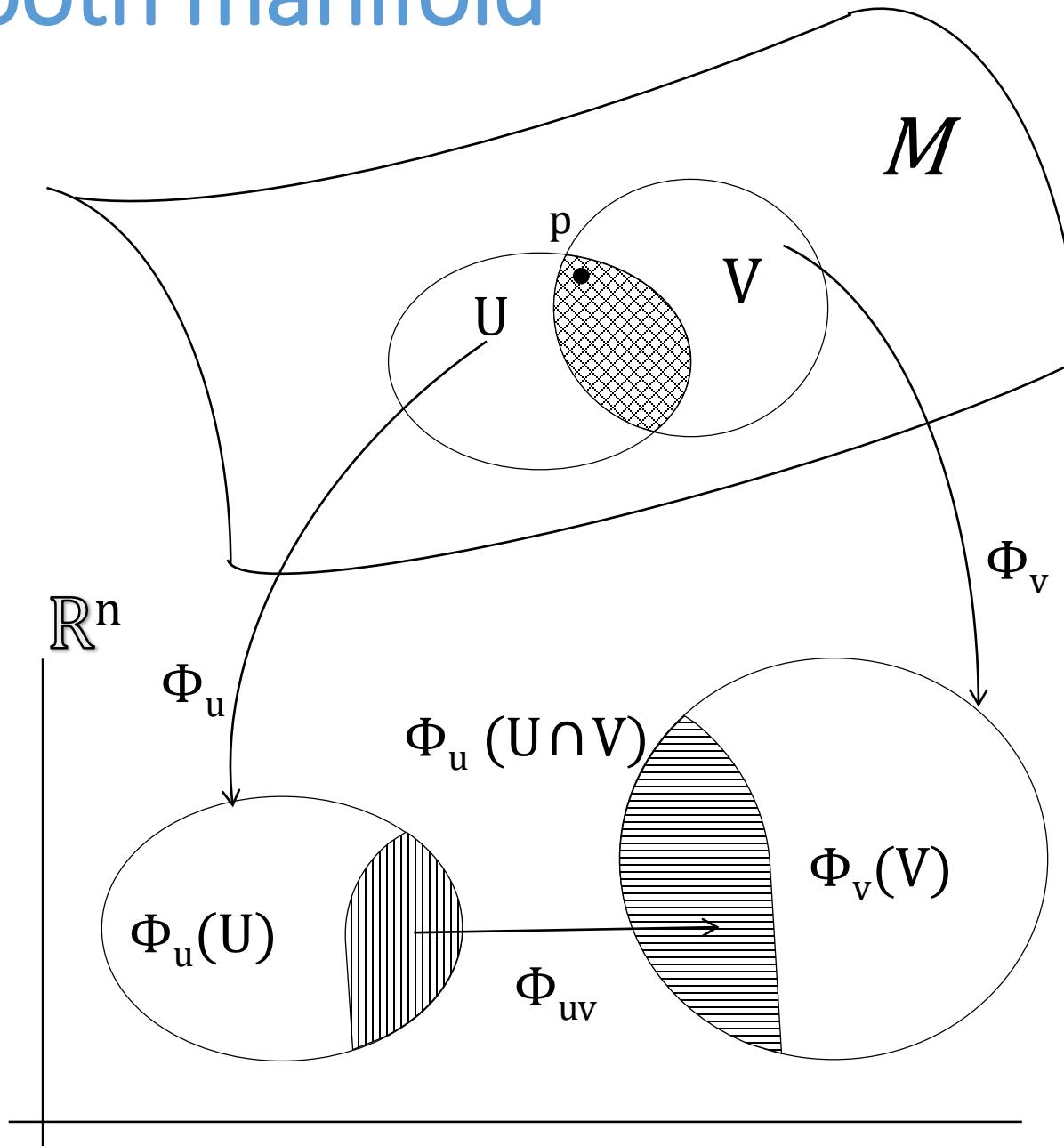
Isometric
embedding



Manifold learning
(Swiss roll)

Intrinsic versus isometric Whitney embeddings (in dim 2D)

Smooth manifold

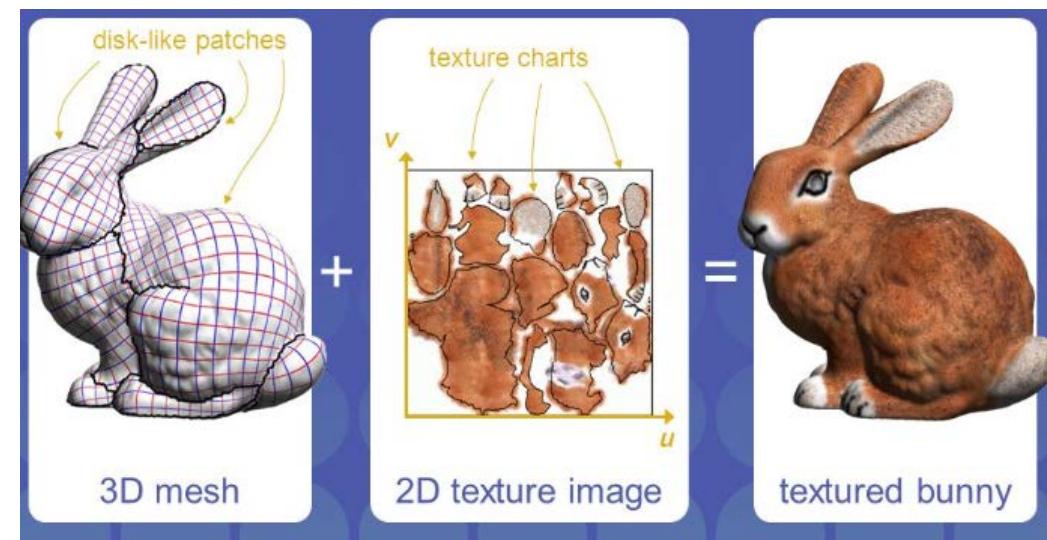


Global geometric objects

vs

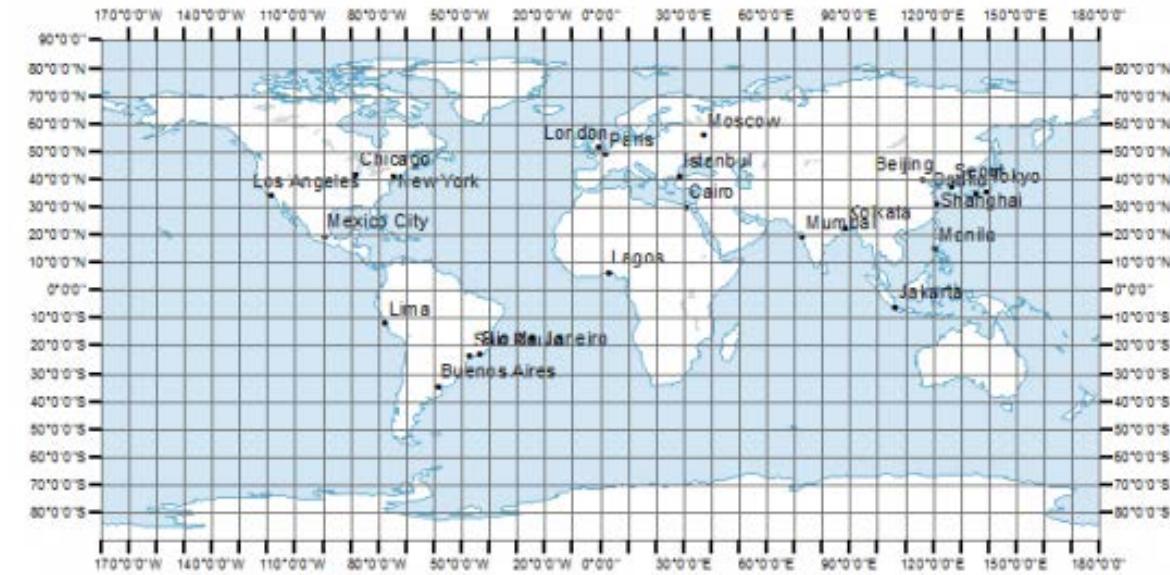
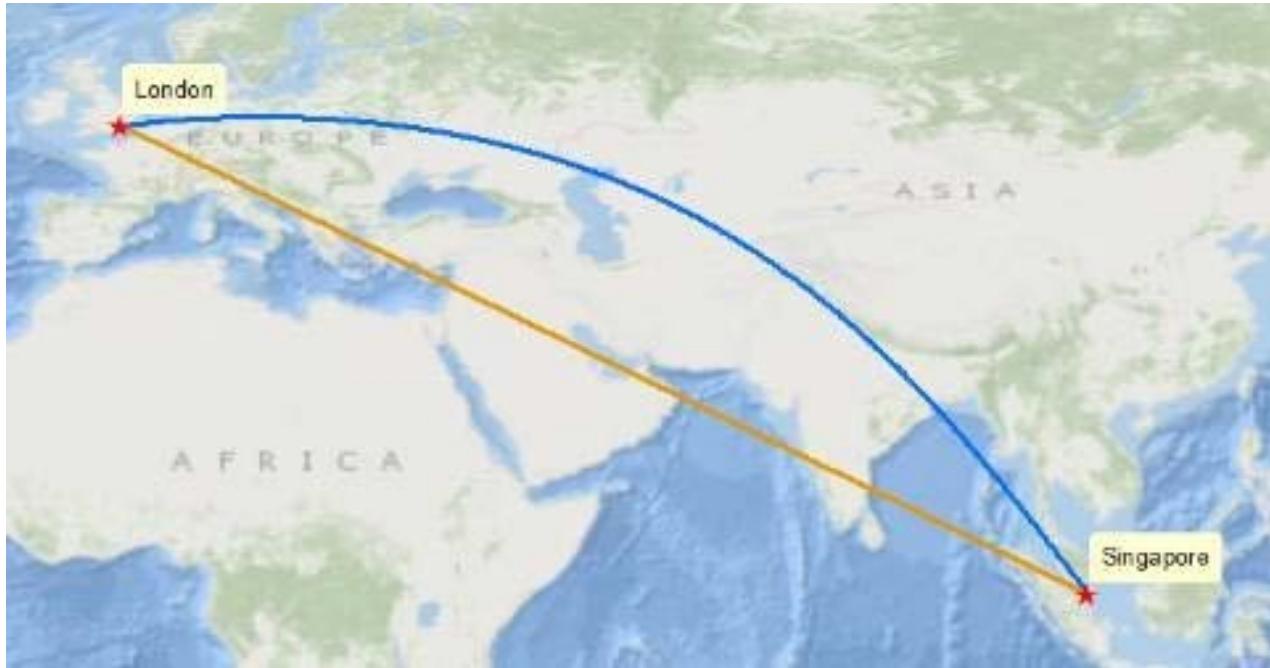
description in local coordinates

Atlas
Coordinate charts



UV mapping

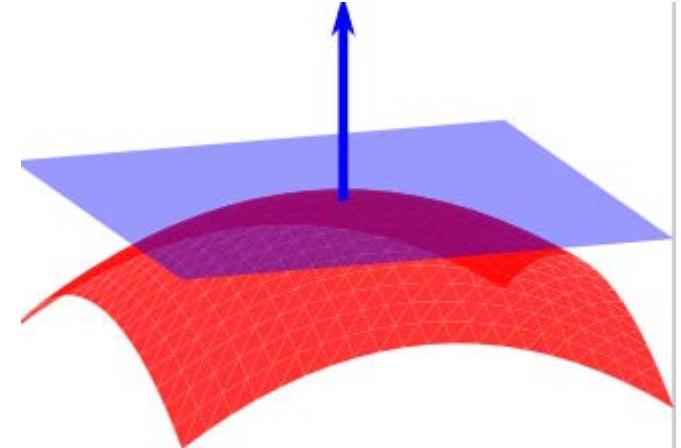
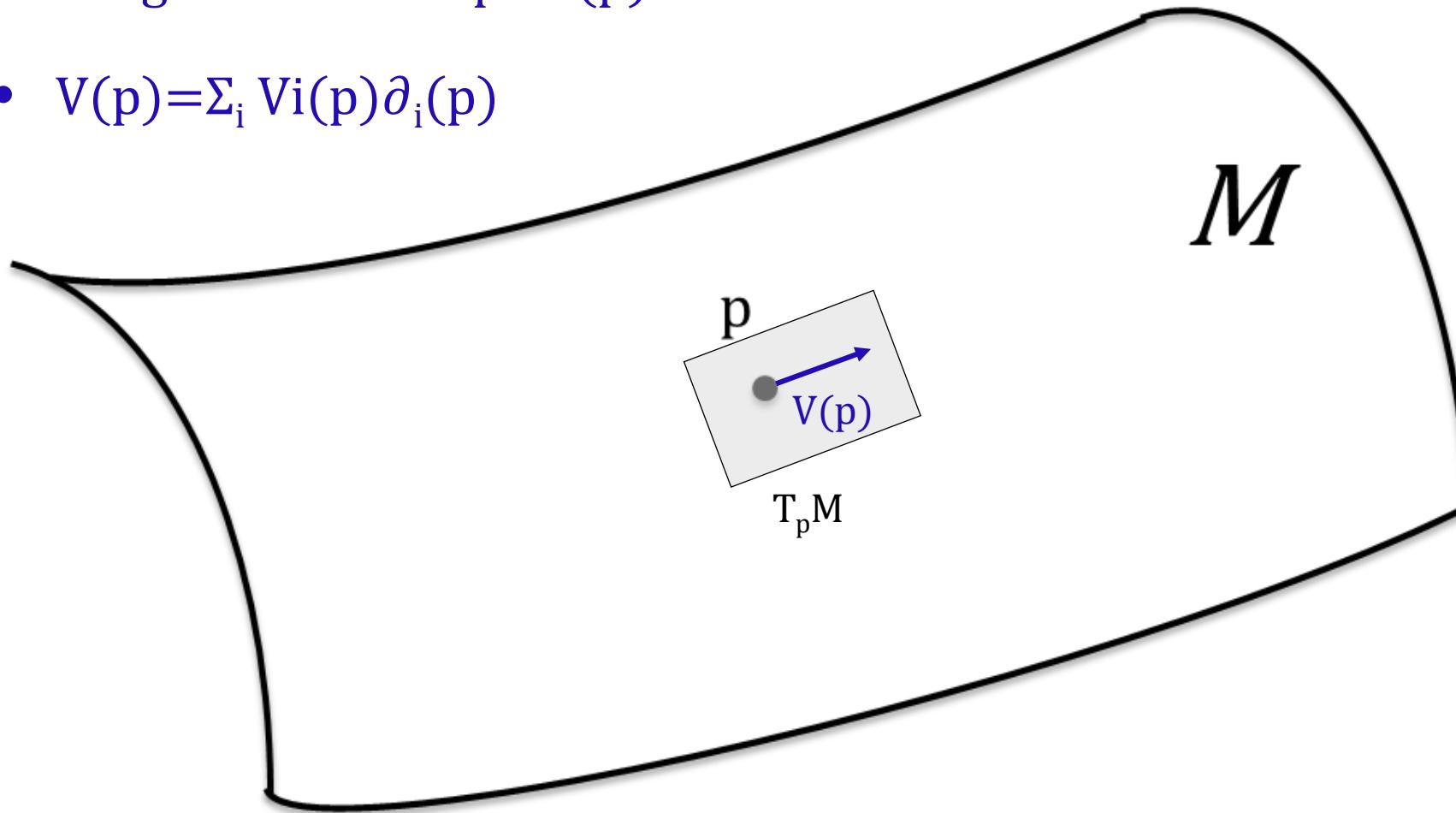
Visualizing paths in a chart (local coordinates)



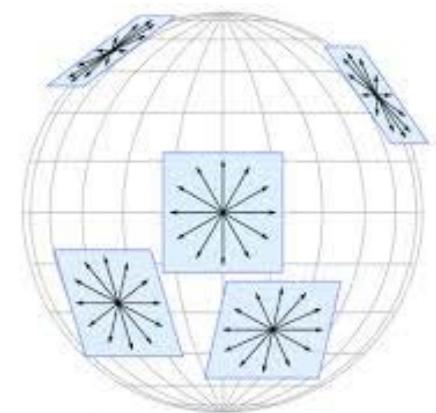
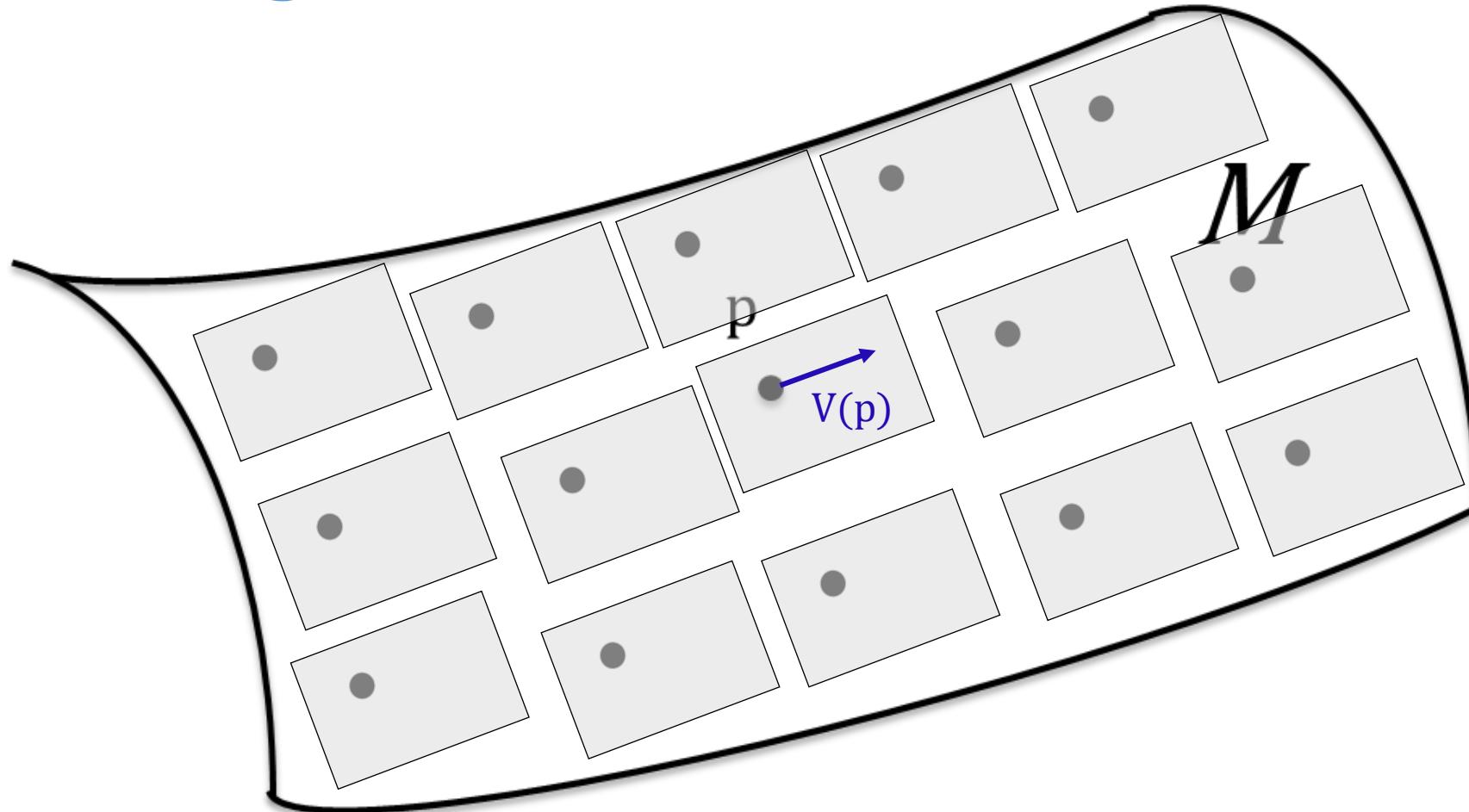
You can only visualize a geometry by rasterizing in a (local) coordinate chart or draw (conceptual) figures

Manifold: Tangent spaces

- Tangent space at p : $T_p M \cong \mathbb{R}^n$
- Tangent vector at p : $V(p)$
- $V(p) = \sum_i V_i(p) \partial_i(p)$



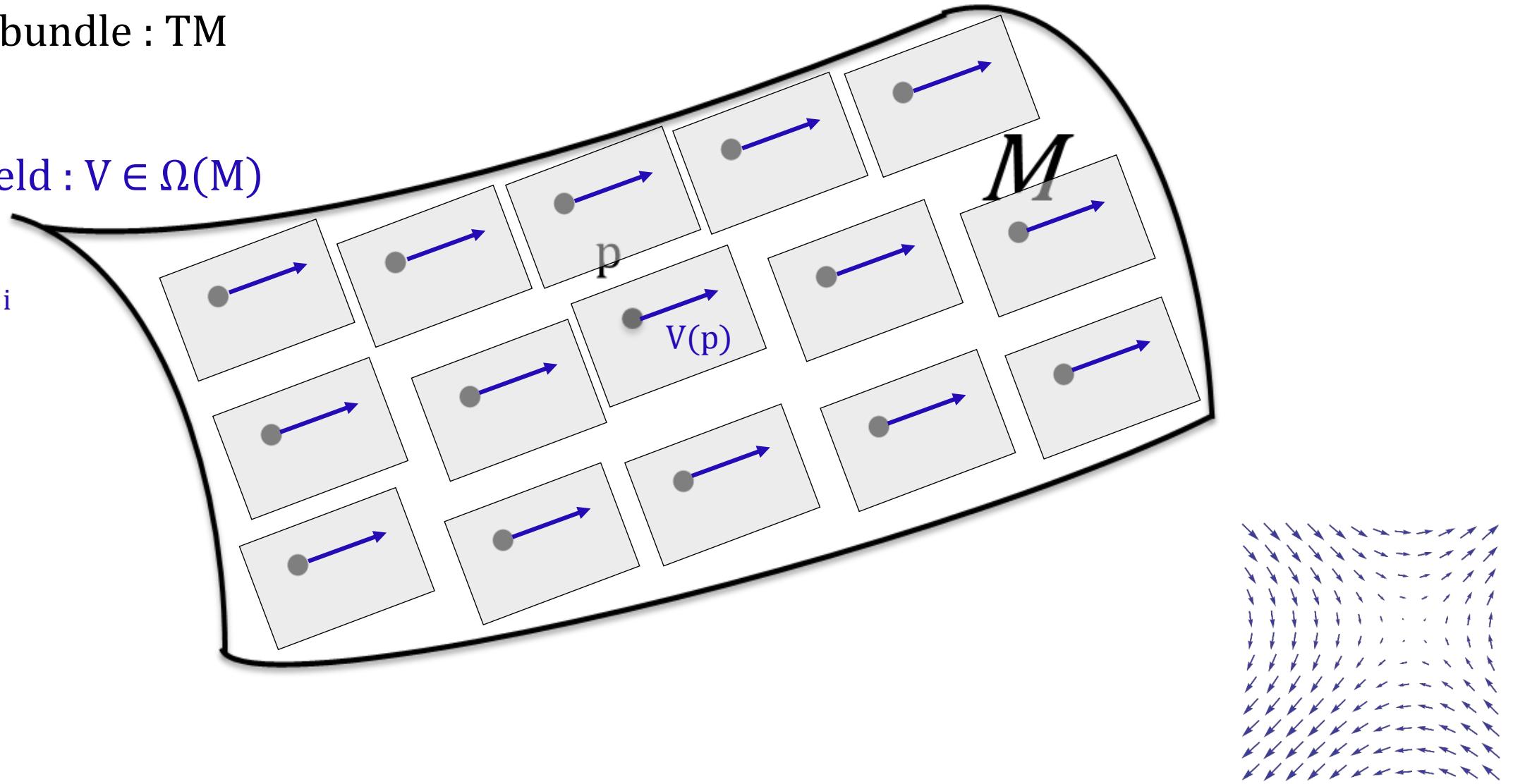
Tangent bundle : TM



Tangent bundle on a 2-sphere

Vector fields (cross-sections of tangent bundle)

- Tangent bundle : TM
- Vector field : $V \in \Omega(M)$
- $V = \sum_i V_i \partial_i$



Metric tensor field

(M,g) Riemannian manifold

$$g : \Omega(M) \times \Omega(M) \rightarrow F(M, \mathbb{R})$$

$$g : (V, W) \mapsto [g(V, W) : p \mapsto g_p(V, W)]$$

- Bilinear positive-definite

$$g(aU + V, W) = ag(U, W) + g(V, W)$$

- symmetric

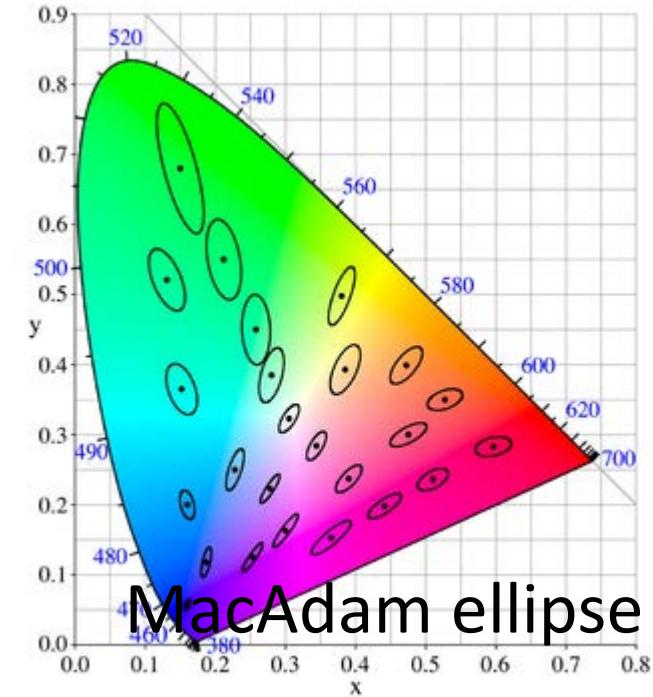
$$g(V, W) = g(W, V)$$

- nondegenerate

$$\forall p, \forall V \neq 0 \exists W, g_p(V, W) \neq 0$$

in coordinates:

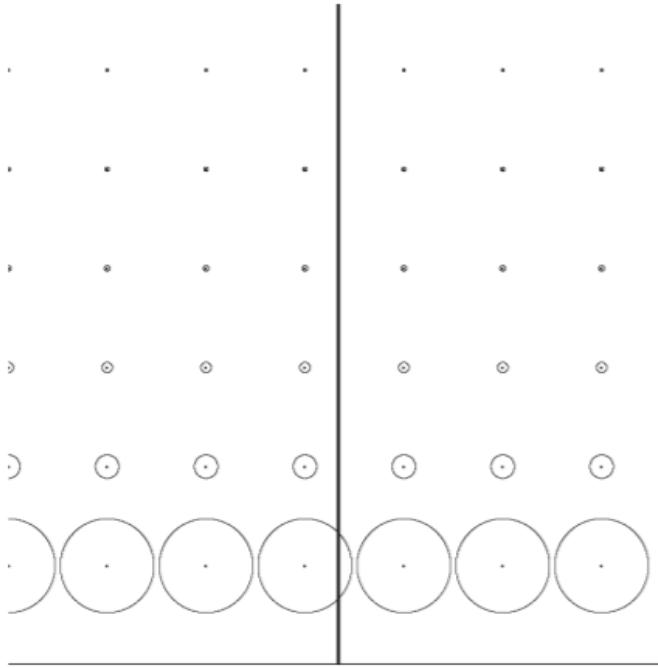
$$g_p = g_p(\partial_i(p), \partial_j(p)) = g_{ij}(p)$$



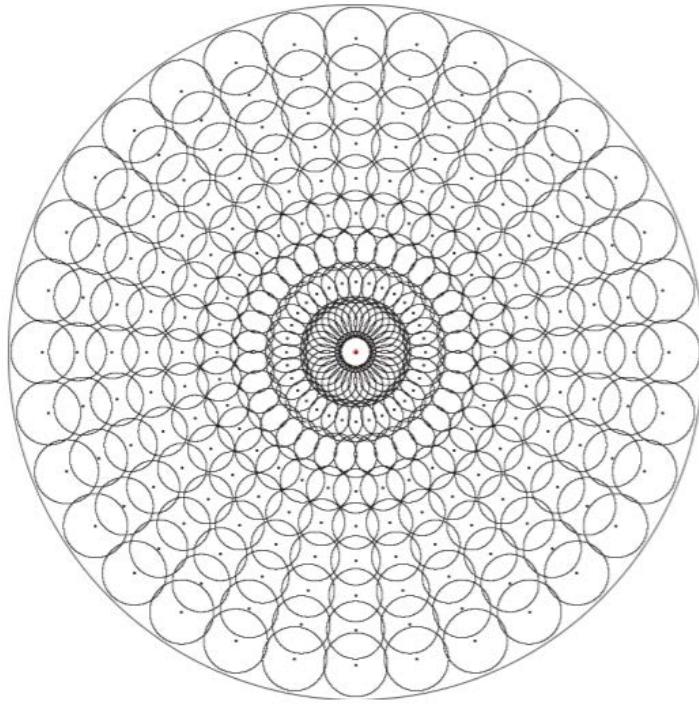
Metric tensor g allows you to:

- Measure angles between vectors
- Measure vector lengths

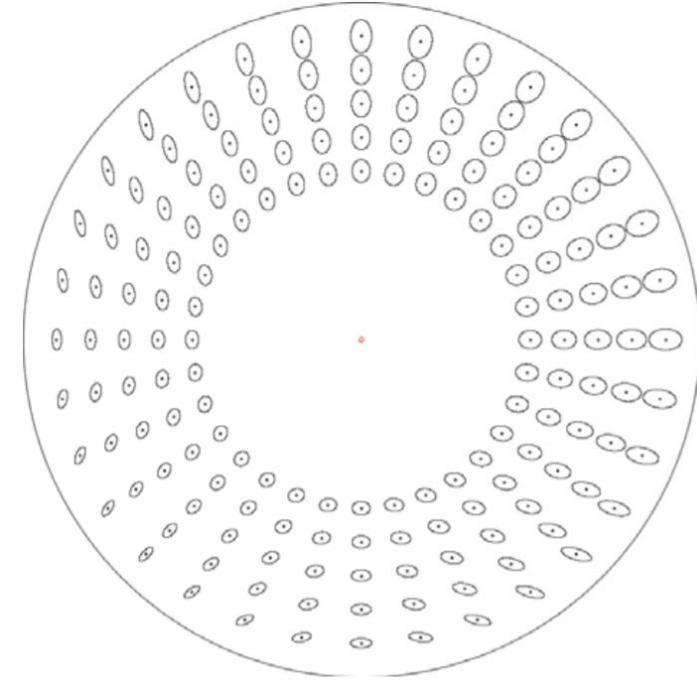
Conformal versus non-conformal metric tensor



Upper Poincare plane
(conformal)



Poincare disk
(conformal)



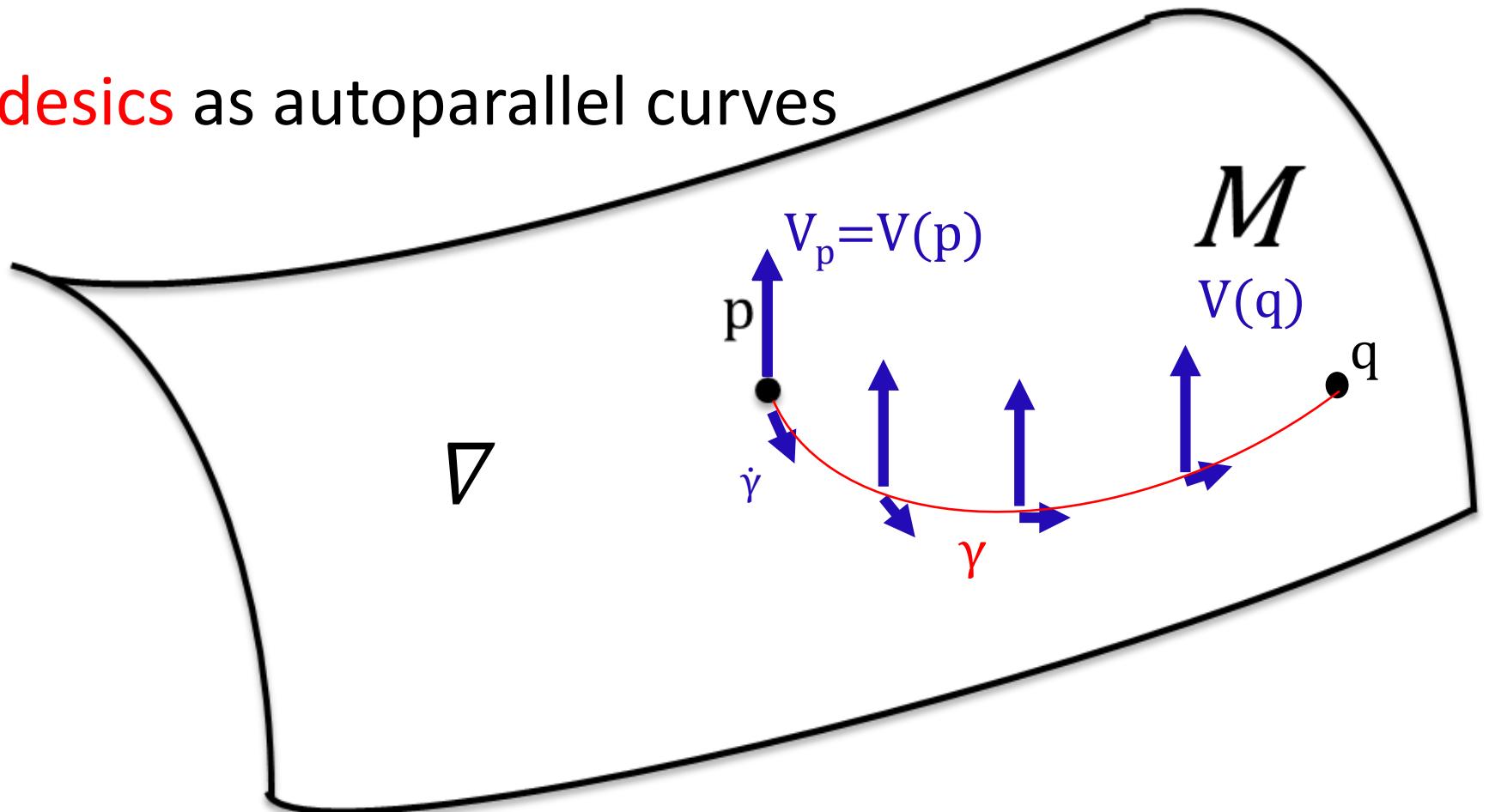
Klein disk
(non-conformal)

$$\hat{g}_p = e^{f(p)} g$$

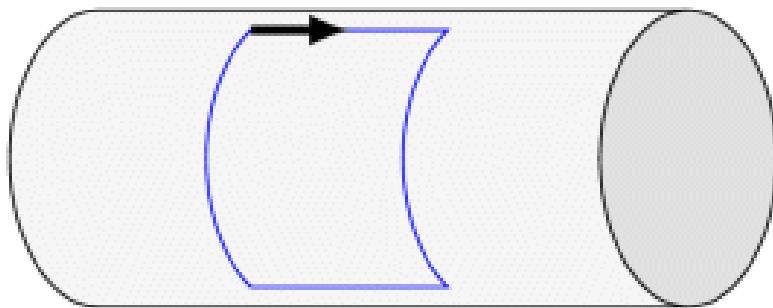
Conformal: metric tensor a scalar-value function of the Euclidean metric tensor
Can measure angles without distortions

Connection ∇

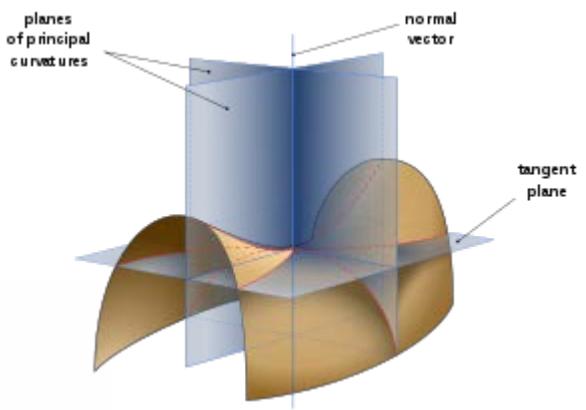
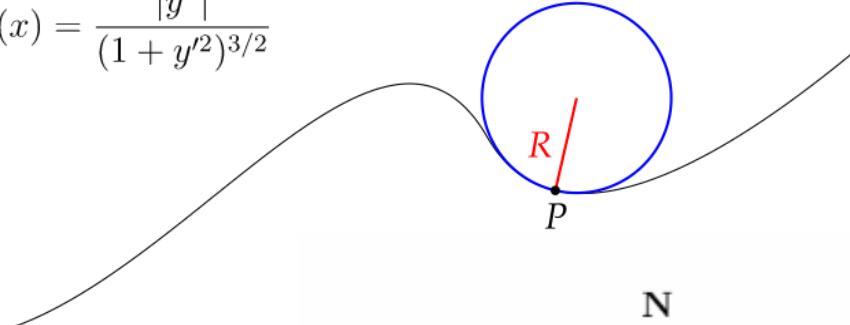
- Define how to **parallel transport** a vector from one tangent plane to another tangent plane by infinitesimally parallel shifting it along a curve
- Use to define **geodesics** as autoparallel curves



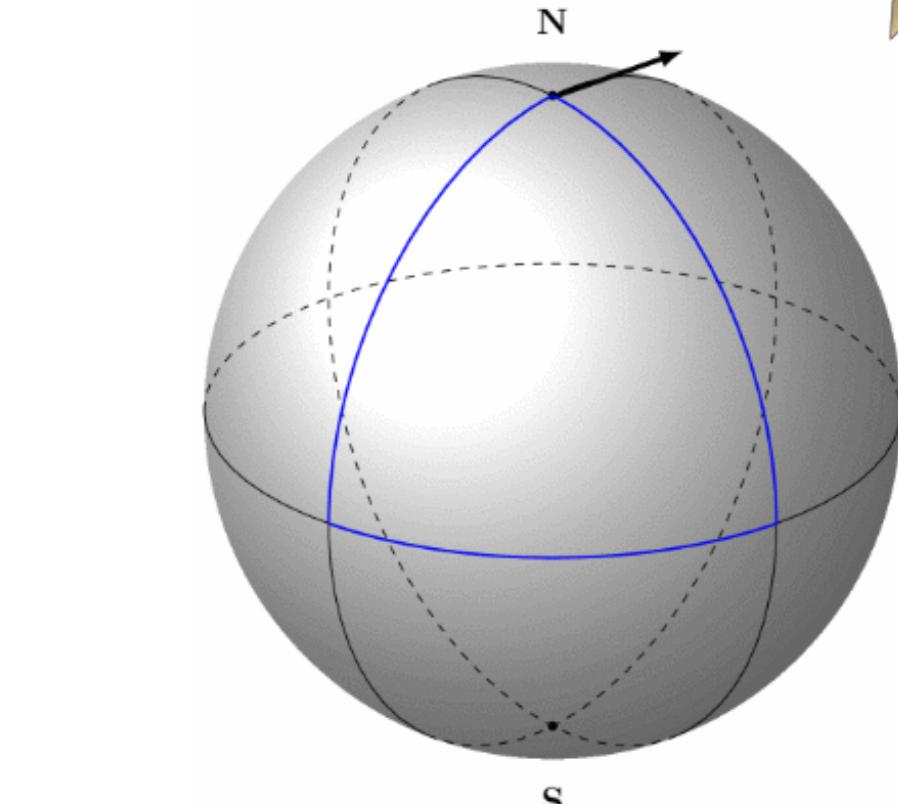
Curvature of ∇



$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$



Cylinder is flat
Parallel transport is
independent of path



Sphere has constant curvature
Parallel transport is path-dependent

A word about torsion of a connection ∇

Torsion measures the speed of rotation of the binormal vector

parallel transport “twists” vectors.

- For connections:

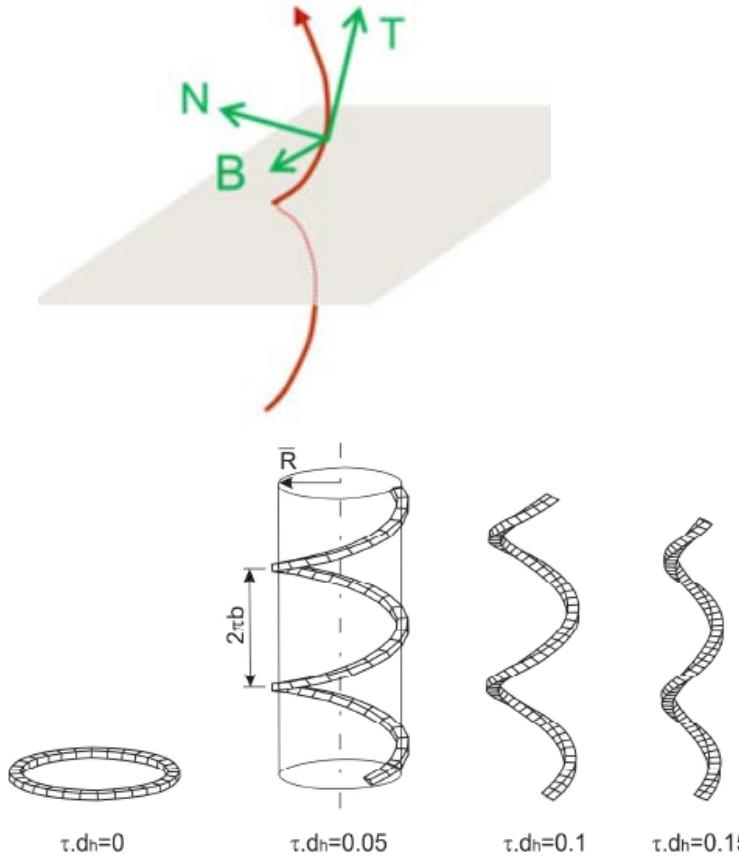


Figure 1. Helical channels with square cross section, constant curvature $\kappa.dh = 1$ and torsion $\tau.dh$ spanning from 0 to 0.15.

Torsion in geometry and in field theory

3

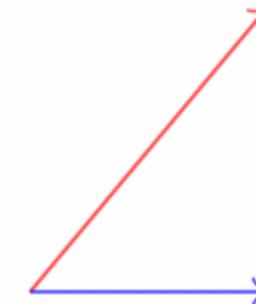
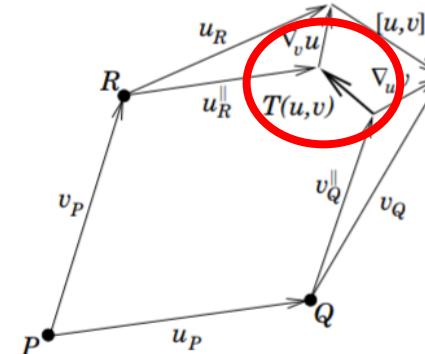
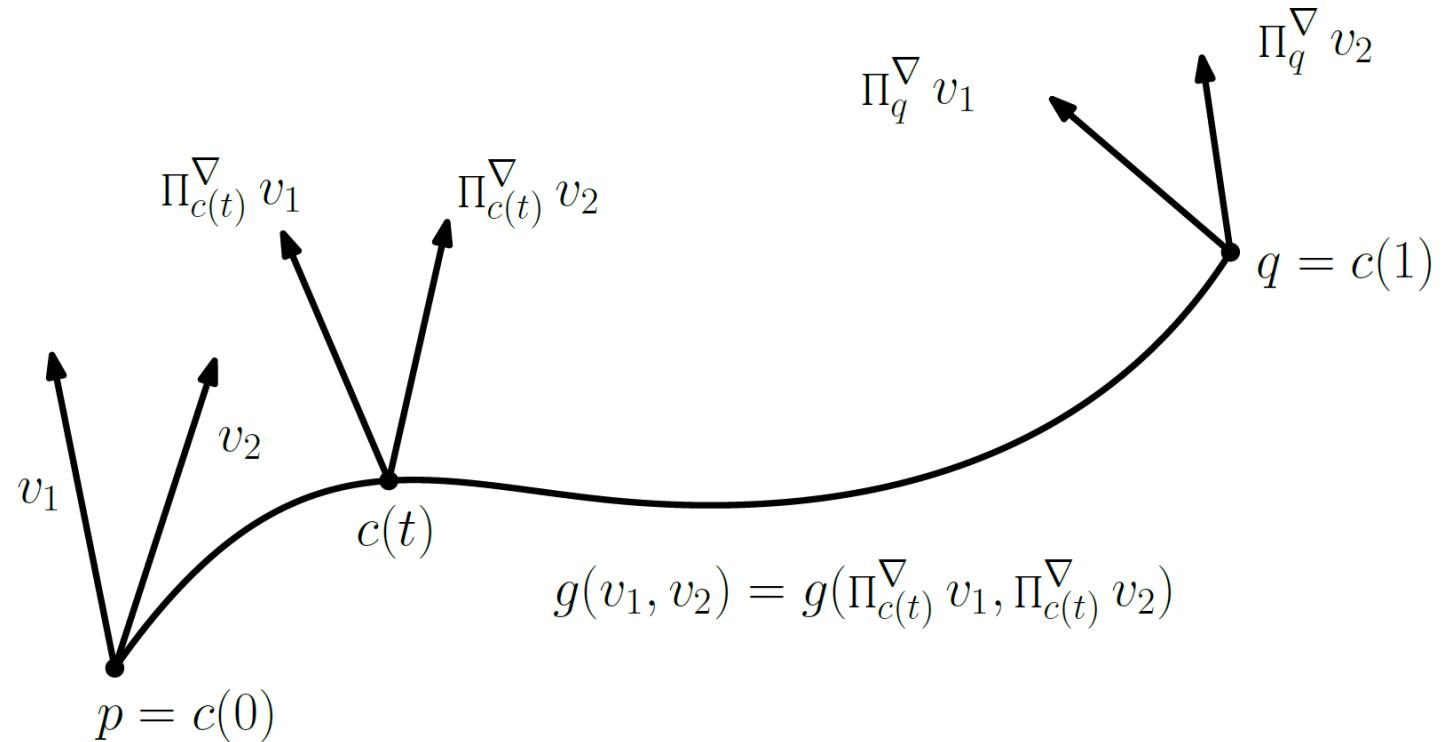


Figure 1: *On the geometrical interpretation of torsion*, see [39]: Two vector fields u and v are given. At a point P , we transport parallelly u and v along v or u , respectively. They become u_R^{\parallel} and v_Q^{\parallel} . If a torsion is present, they don't close, that is, a closure failure $T(u, v)$ emerges. This is a schematic view. Note that the points R and Q are infinitesimally near to P . A proof can be found in Schouten [88], p.127.

Connections differing by torsions have same geodesics
Pregeodesics

Metric-compatible connection ∇

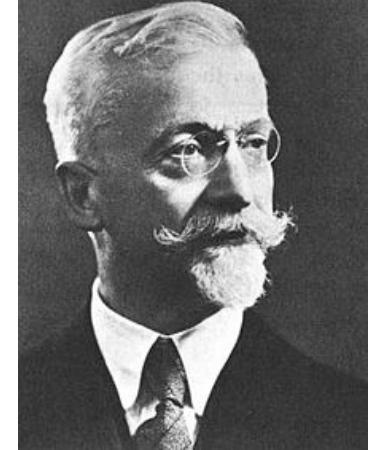
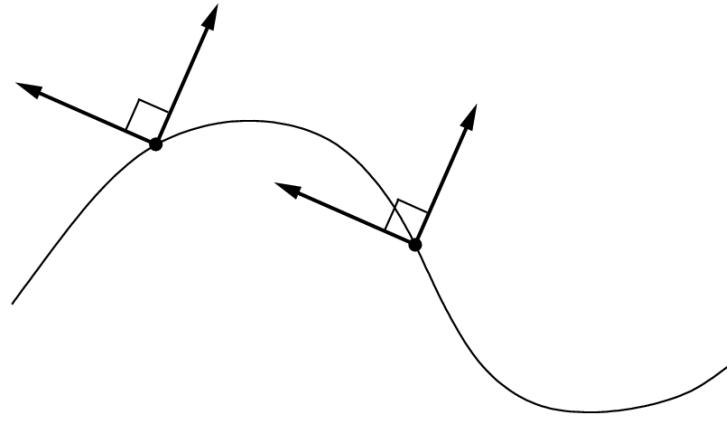
Preserves the inner product of vectors by parallel transport



$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \rightarrow c(t)}^{\nabla} u, \prod_{c(0) \rightarrow c(t)}^{\nabla} v \right\rangle_{c(t)} \quad \forall t$$

Fundamental theorem of Riemannian geometry

There exists a **unique** torsion-free connection that is **metric compatible** that is called the **Levi-Civita connection**. The connection is derived from g



Riemannian geometry: take the Levi-civita connection

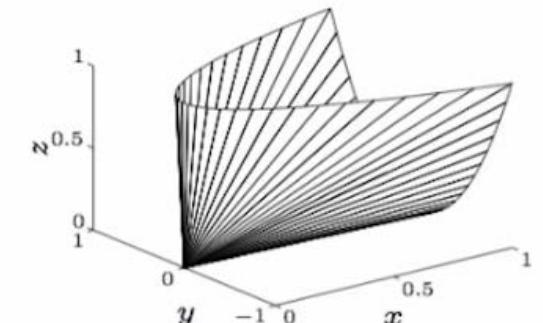
Affine differential geometry: take any affine connection (Elie Cartan)

Information geometry: take a pair of dual connections

Rationale for information spaces

- In traditional geometry, a space is a **vacuum**
- In physics, a spacetime contains **matter**
(torsion GR of Einstein-Cartan)
- An **information space** is a space packed with **entities**/models:
 - Space of matrices, symmetric matrices, positive-definite matrices
 - Space of parametric densities, non-parametric densities, positive densities
 - Etc.

example: $\begin{bmatrix} x & y \\ y & z \end{bmatrix} \in \mathbf{S}_+^2$

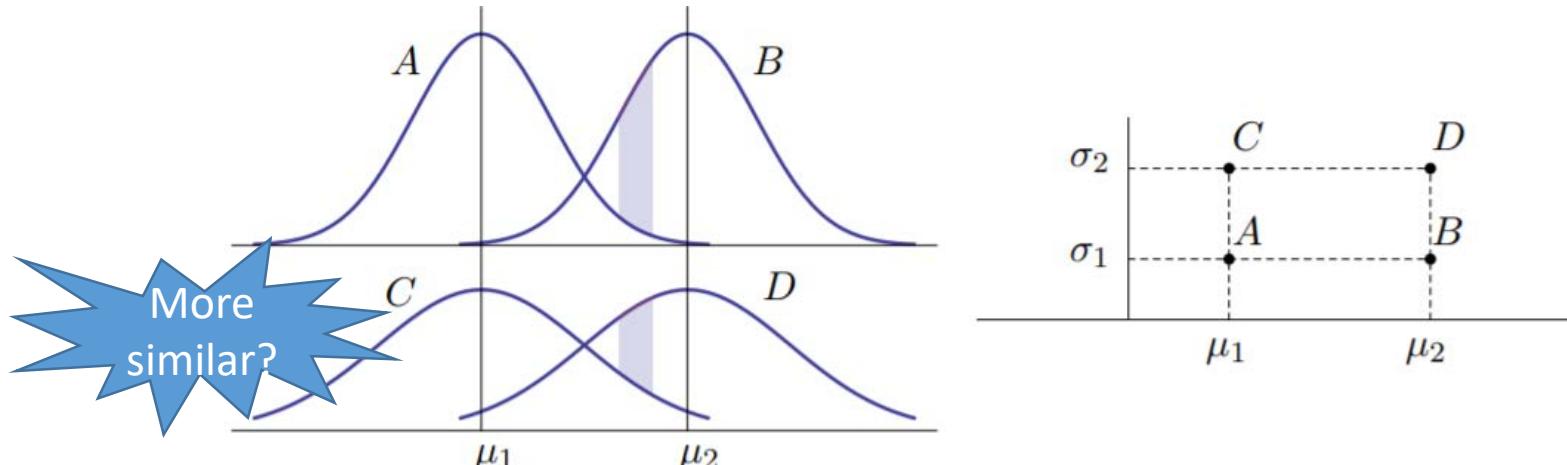


Rationale for Information Geometry (IG)

- What is **the/a** geometry of the space of Gaussians?
Distance, interpolation, closest Gaussian of a subfamily (projection)
Chosen geometry may depend on applications
- Discovered a **dualistic geometry** that can also be used in other **non-statistical contexts**
- Applications of the IG framework to information sciences (statistics, information theory, signal processing, machine learning, etc.)
- Wider scope of Geometric Science of Information (GSI)

What is the geometry of the Gaussian manifold?

- Euclidean geometry/distance yields this interpretation

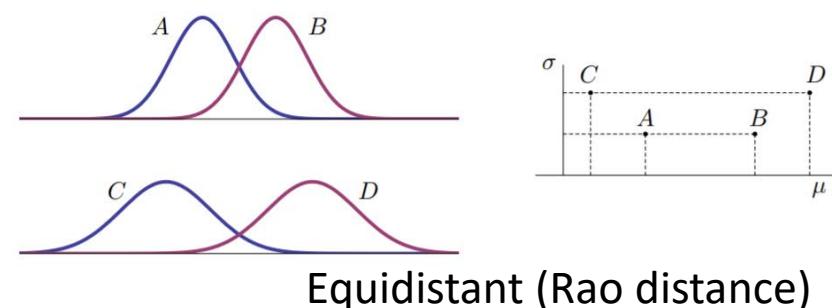


- Desiderata: Dissimilarity shall be **invariant of parameterization**:

Same distance for parameterizations $\{N(\mu, \sigma)\}$ or $\{N(\mu, \sigma^2)\}$

No geometry of the sample space

Invariant by sufficient statistics



Equidistant (Rao distance)

- Actually, Optimal Transport geometry of Gaussian manifold yields Euclidean geometry. But OT does not distinguish normal from any elliptical family

Fisher-Riemannian geometry (1930/1945)

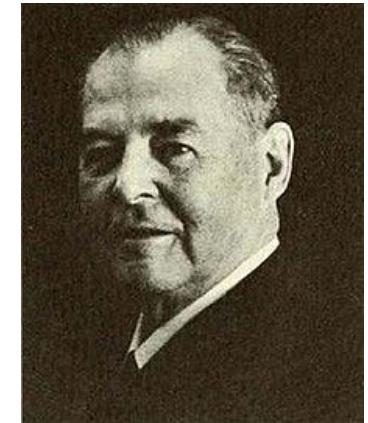
Spaces of Statistical Parameters.

By Harold Hotelling , Stanford University.

For a space of n dimensions representing the parameters p_1, \dots, p_n of a frequency distribution, a statistically significant metric is defined by means of the variances and



Use Fisher information
for the Riemannian
metric tensor

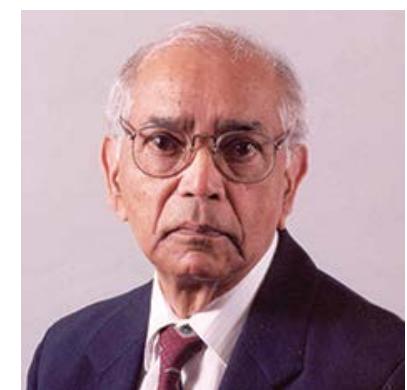
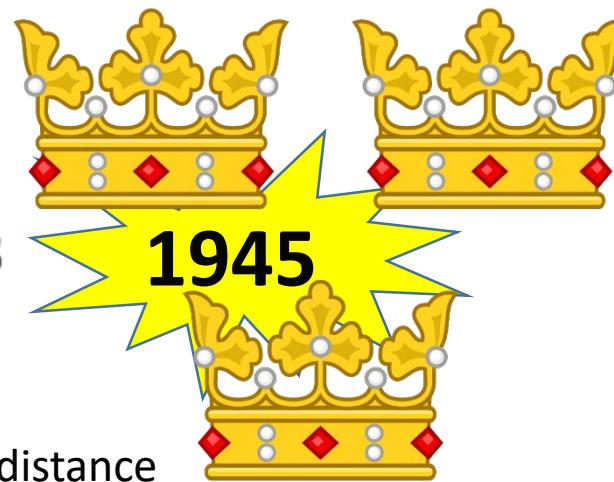


Harold Hotelling

Information and the Accuracy Attainable
in the Estimation of Statistical Parameters

C. Radhakrishna Rao

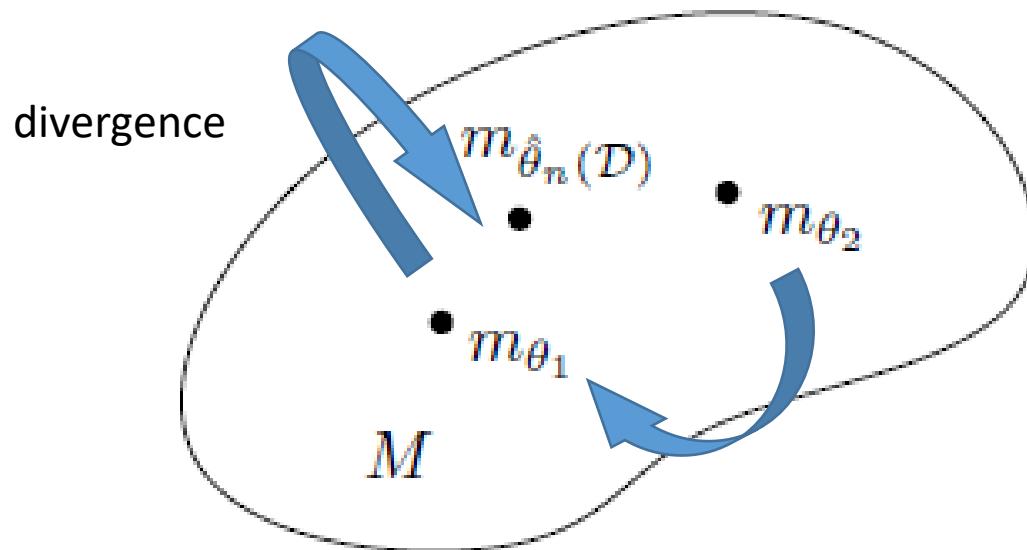
CRLB+Rao-Blackwellization+Fisher-Rao distance



C. R. Rao

Population space/parameter space

- Example in statistical hypothesis testing: estimate from observations and then classify wrt divergence to decide which hypothesis.



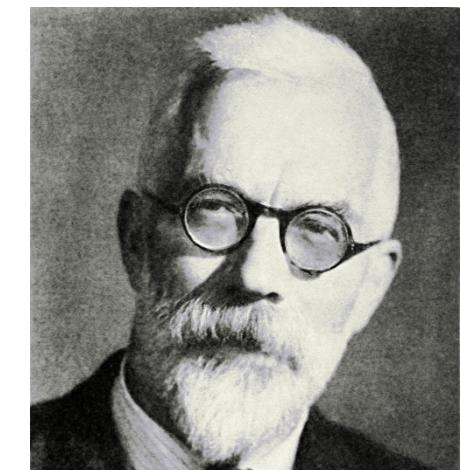
Needed to build better
Information sciences:

- Deal with model and data
- Deal with model and model

Fisher information metric/matrix (FIM)

$$g(\xi) = E_{\xi} \left[\frac{\partial}{\partial \xi} \log(p_{\xi}) \frac{\partial}{\partial \xi} \log(p_{\xi}) \right]$$

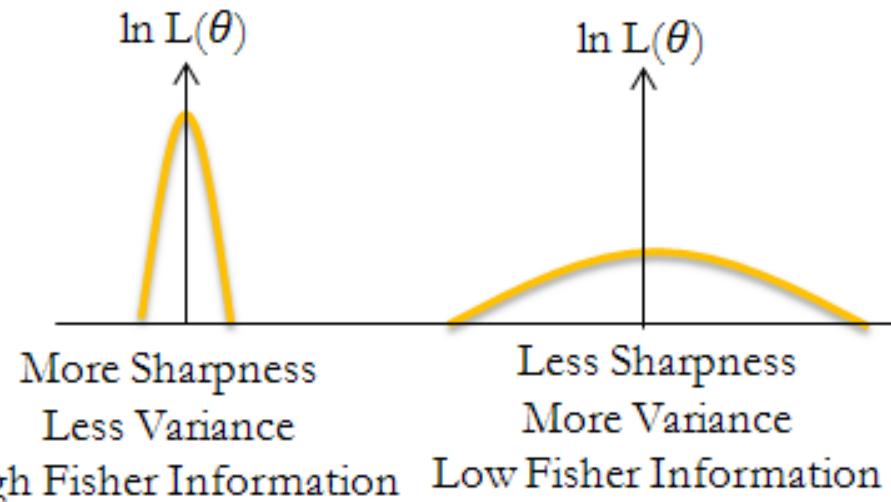
$$g_{ij}(\xi) = \int \frac{\partial}{\partial \xi} \log(p_{\xi}(x)) \frac{\partial}{\partial \xi} \log(p_{\xi}(x)) p_{\xi}(x) dx$$



Sir Ronald Fisher

FIM is **positive-semidefinite**, positive-definite for regular models

$$\text{Curvature} = - \frac{\partial^2}{\partial \theta^2} [\ln L(\theta)]$$

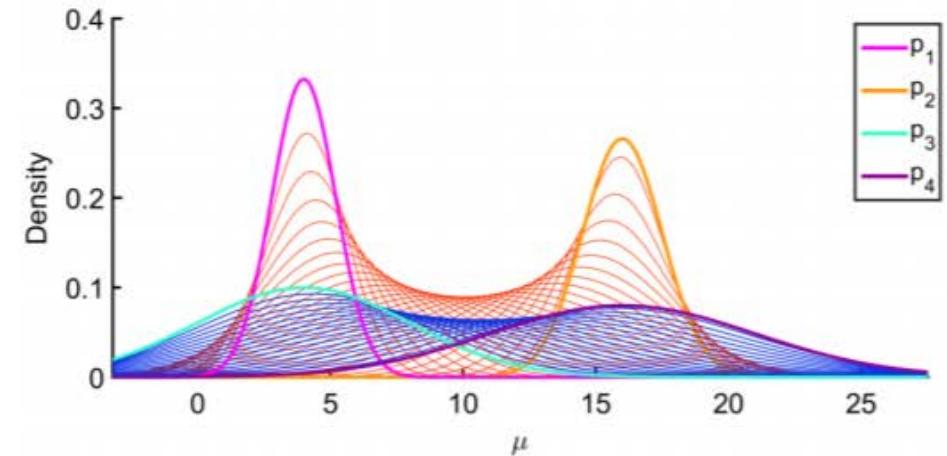
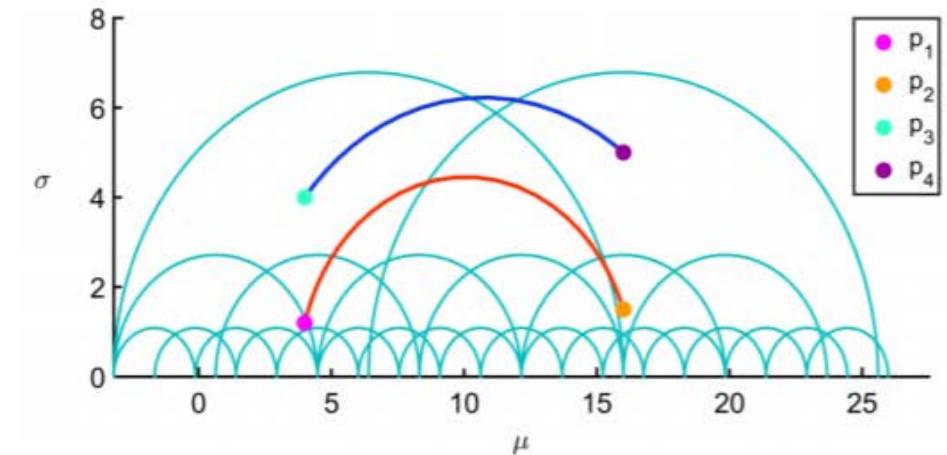
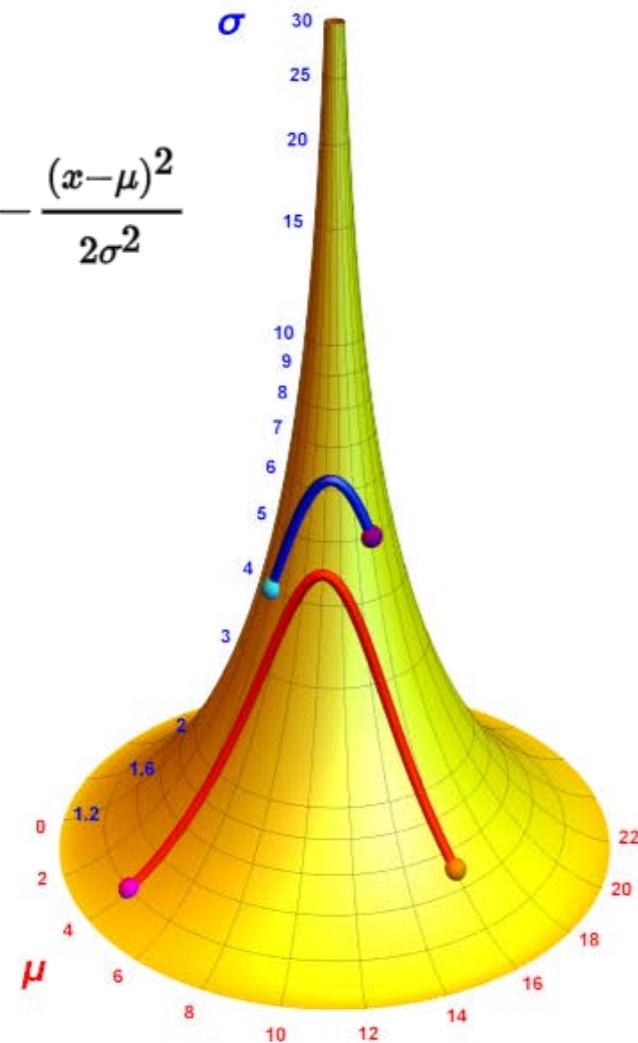


$$g_{ij}(\theta) = E \left\{ \frac{\partial}{\partial \theta_i} \log p(X | \theta) \frac{\partial}{\partial \theta_j} \log p(X | \theta) \right\}$$

Geometry of normal distributions (hyperbolic)

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Pseudo-sphere
(negative curvature)

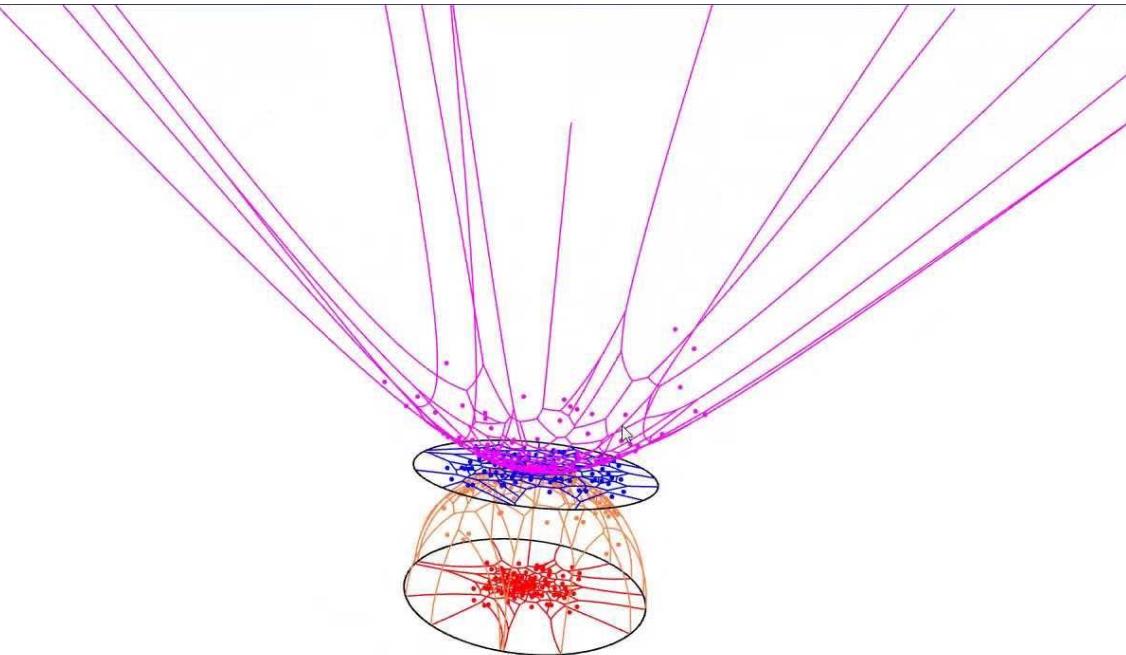


Hyperbolic geometry for location-scale families

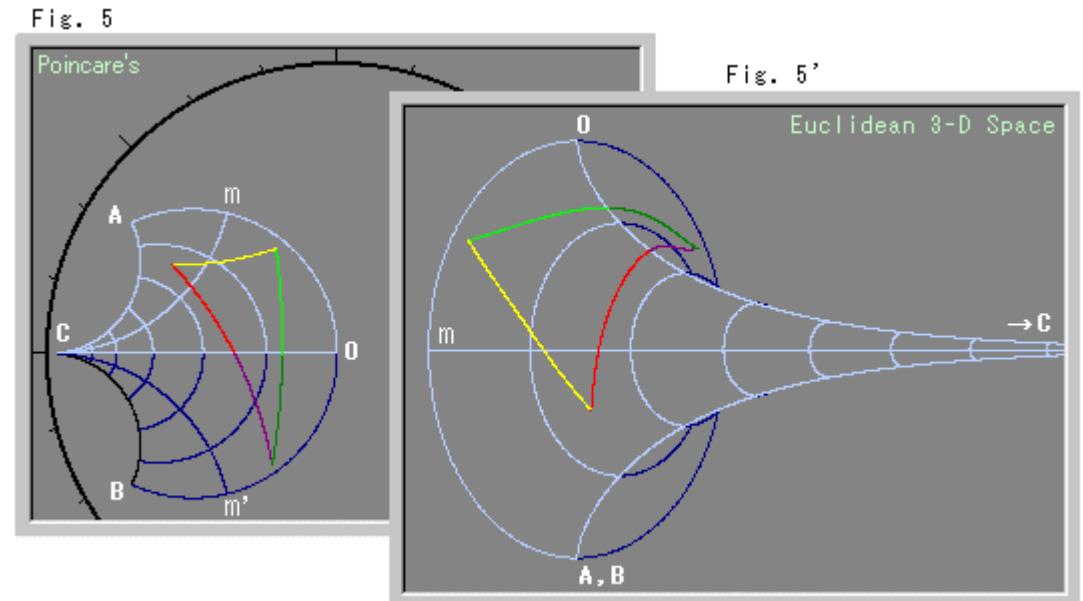
$$\mathcal{P} = \left\{ \frac{1}{s_1} p \left(\frac{x-l_1}{s_1} \right) : (l_1, s_1) \in \mathbb{H} \right\}$$

$\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$: open half-space of 2D (l, s) location-scale parameters

Several models of hyperbolic geometry (Klein, Poincare, Beltrami, pseudosphere)



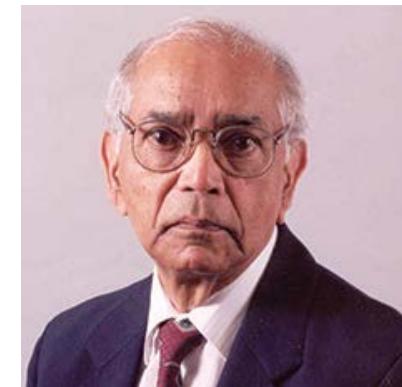
<https://www.youtube.com/watch?v=i9IUzNxeH4o>



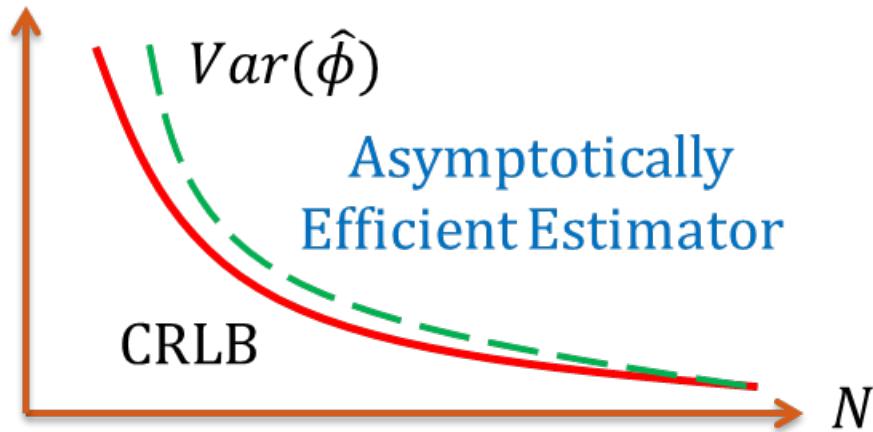
Cramer-Rao lower bound (CRLB)

The variance of any unbiased estimator is lower bounded by the inverse of the Fisher information

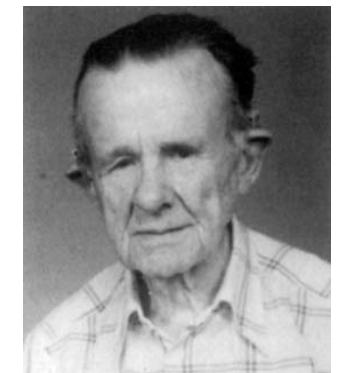
$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$



C. R. Rao



The covariance of any unbiased estimator is lower bounded by the inverse of the Fisher information matrix



Harald Cramer

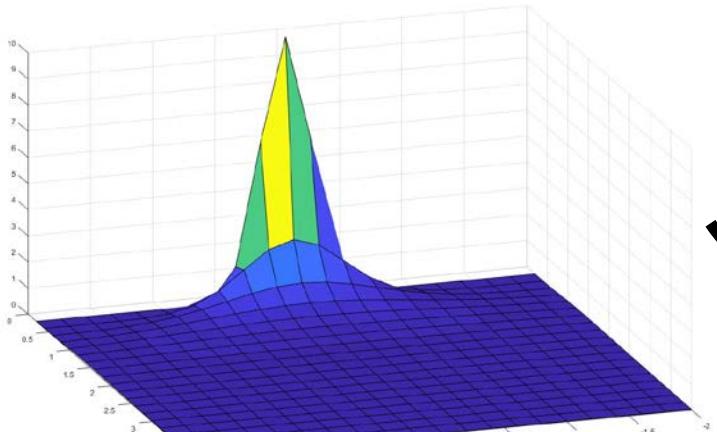
$$\mathbf{C}_{\hat{\theta}} - \mathbf{I}^{-1}(\theta) \geqslant 0$$

(here, positive-definite matrices, Lowner ordering)

Examples of statistical models (regular/identifiable)

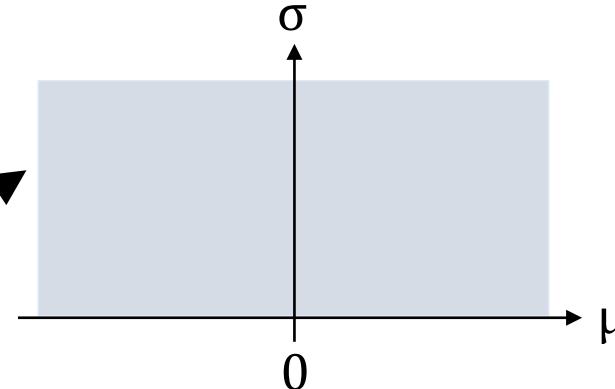
- $\mathcal{N}(\mu, \sigma)$

Negative curvature



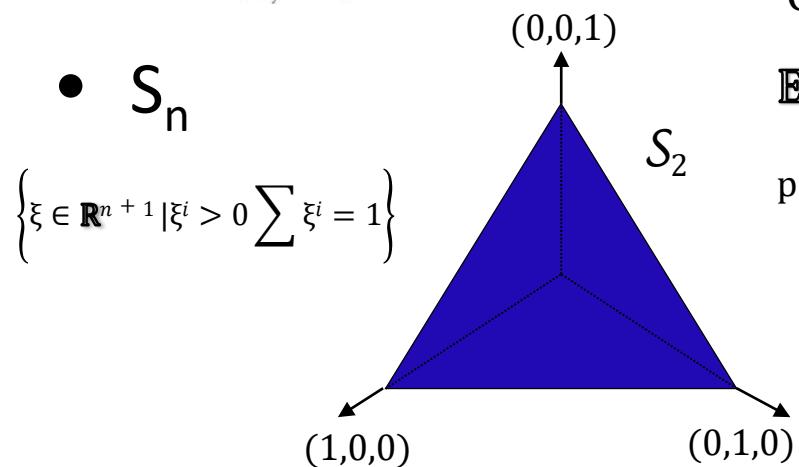
$$\begin{aligned} \Omega &= \mathbb{R} \\ \Xi &= \mathbb{R} \times \mathbb{R}^+ \\ p(x, \mu, \sigma) &\mapsto (\mu, \sigma) \end{aligned}$$

Φ



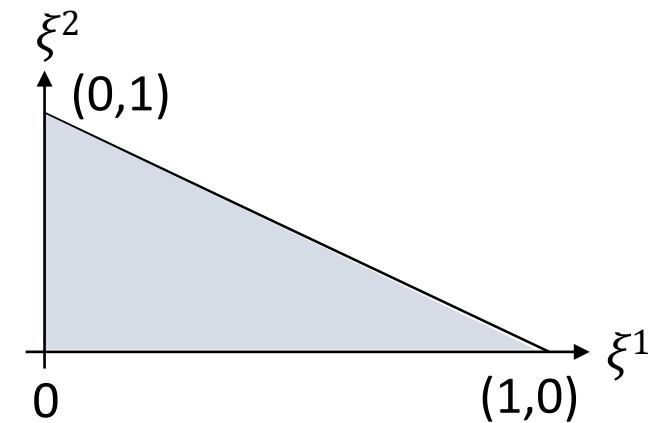
- S_n

Positive curvature



$$\begin{aligned} \Omega &= \{x_1, \dots, x_n\} \\ \Xi &= \left\{ \xi \in \mathbb{R}^n \mid \xi^i > 0 \sum \xi^i = 1 \right\} \\ p(x, \xi^1, \dots, \xi^{n+1}) &\mapsto (\xi^1, \dots, \xi^n) \end{aligned}$$

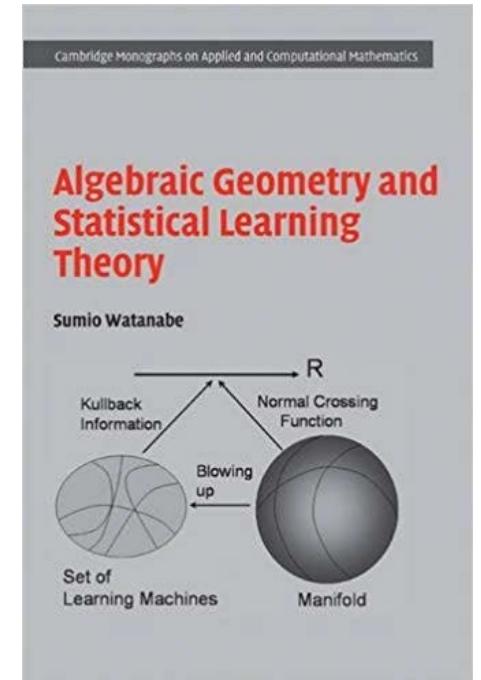
Φ



Exponential family : $p(x, \xi^1, \dots, \xi^n) = e^{C(x) + \xi^i F_i(x)} \cdot \psi(\xi) \mapsto (\xi^1, \dots, \xi^{n+1})$

Non-regular statistical models

- Not identifiable models
- Usually, **hierarchical models**:
 - Gaussian mixture models
 - Multi-layer perceptrons
- Singular Semi-Riemannian manifolds
- Cramer-Rao lower bounds does not hold, different theory for model selection (BIC, MDL)



Statistical curvature

Use of differential geometry
to study the information loss
in estimation

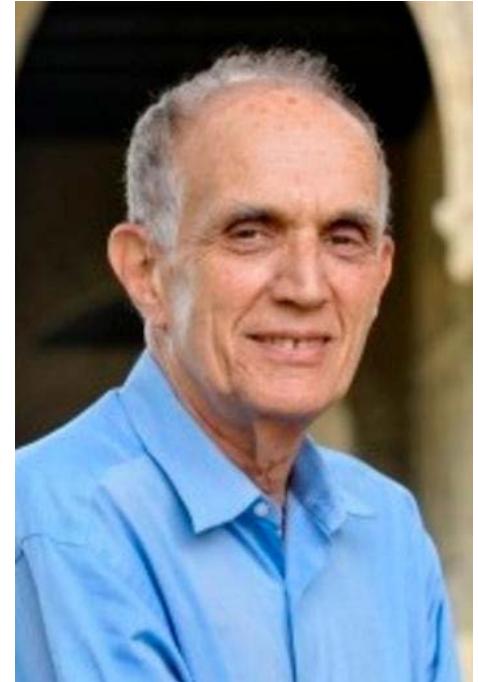
The Annals of Statistics
1975, Vol. 3, No. 6, 1189-1242

DEFINING THE CURVATURE OF A STATISTICAL PROBLEM (WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

BY BRADLEY EFRON

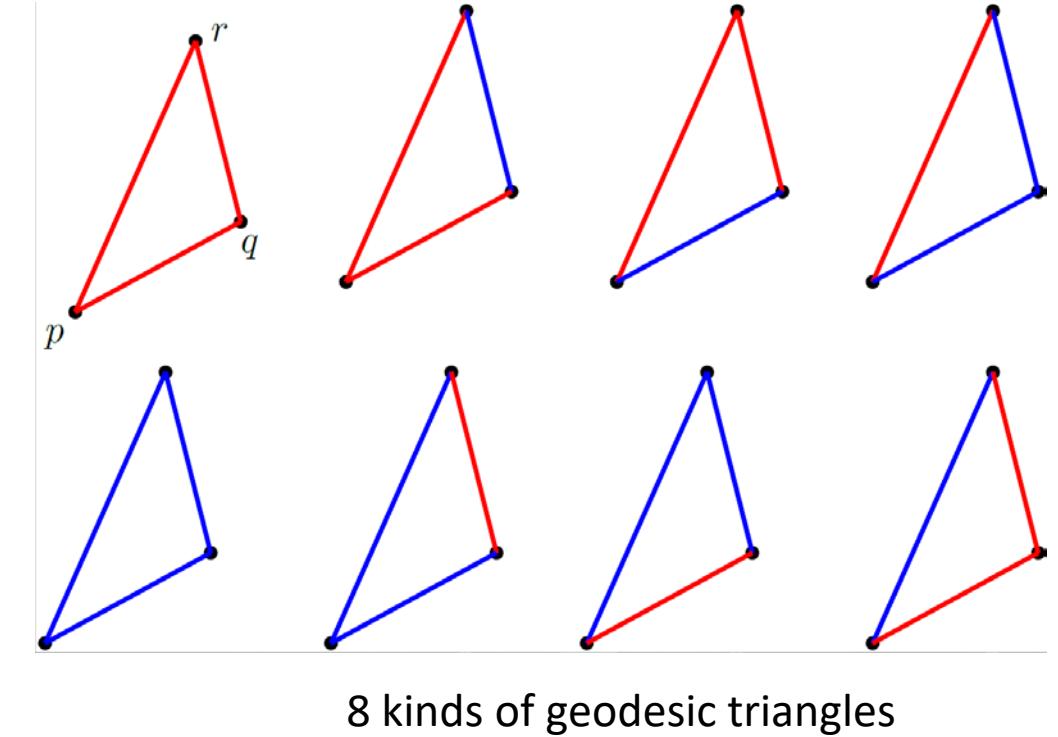
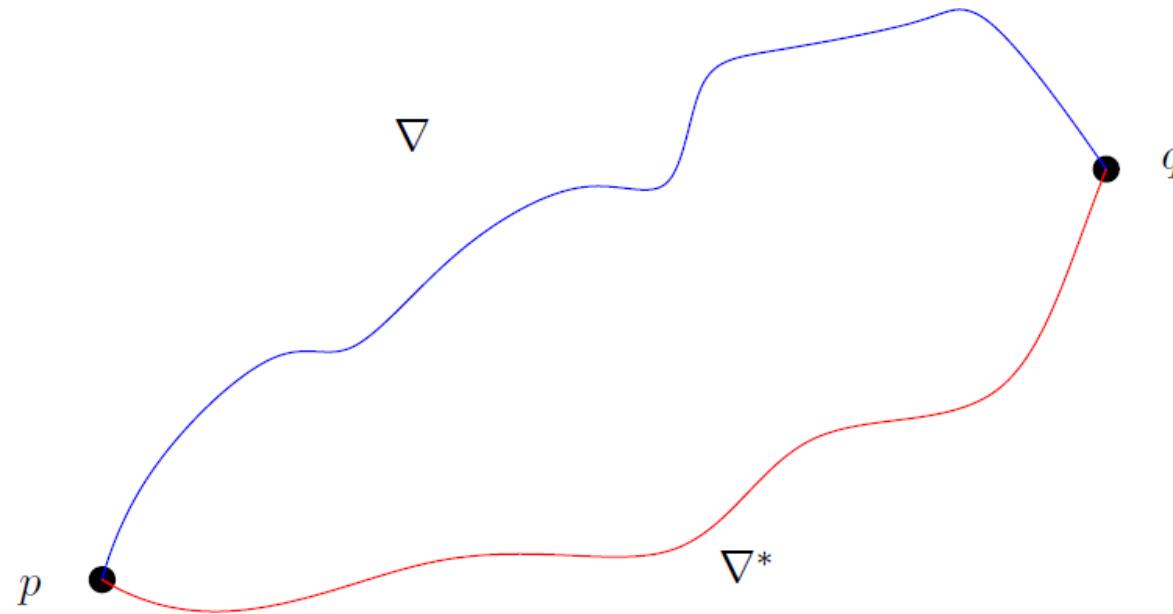
Stanford University

Statisticians know that one-parameter exponential families have very nice properties for estimation, testing, and other inference problems. Fundamentally this is because they can be considered to be “straight lines” through the space of all possible probability distributions on the sample space. We consider arbitrary one-parameter families \mathcal{F} and try to quantify how nearly “exponential” they are. A quantity called “the statistical curvature of \mathcal{F} ” is introduced. Statistical curvature is identically zero for exponential families, positive for nonexponential families. Our purpose is to show that families with small curvature enjoy the good properties of exponential families. Large curvature indicates a breakdown of these properties. Statistical curvature turns out to be closely related to Fisher and Rao’s theory of second order efficiency.



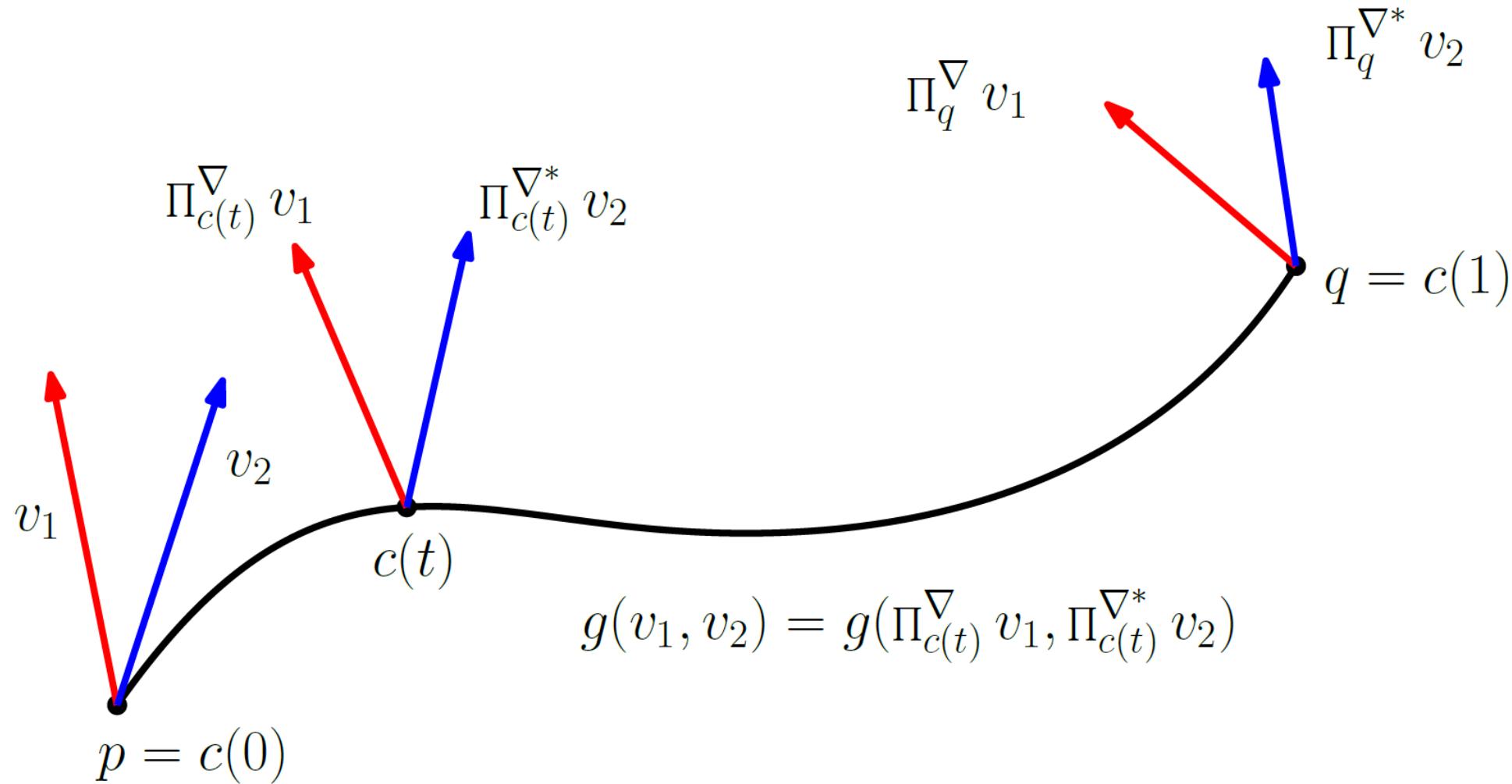
Bradley Efron

Dualistic structure of information geometry



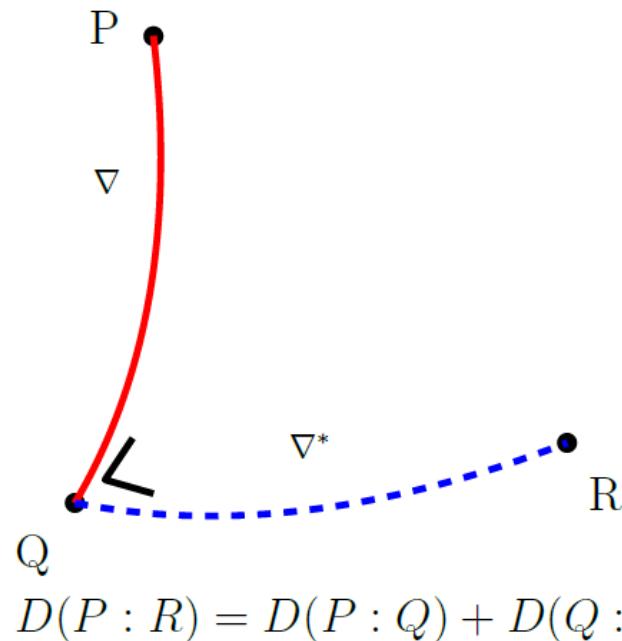
Two conjugate torsion-free affine connections coupled with the metric
Dual parallel transport is metric-compatible
There is not necessarily a distance, 2^k types of k -gons (eg, 8 triangles)

Dual parallel transport is metric-compatible

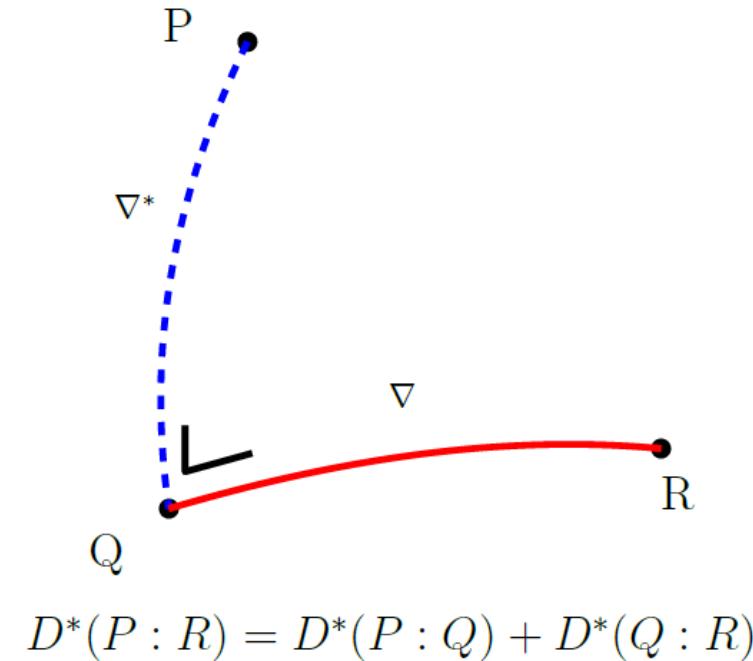


Dually flat space: Pythagoras' theorem

$$\gamma^*(P, Q) \perp_F \gamma(Q, R)$$



$$\gamma(P, Q) \perp_F \gamma^*(Q, R)$$



Two affine coordinate systems coupled by **Legendre-Fenchel transformation**

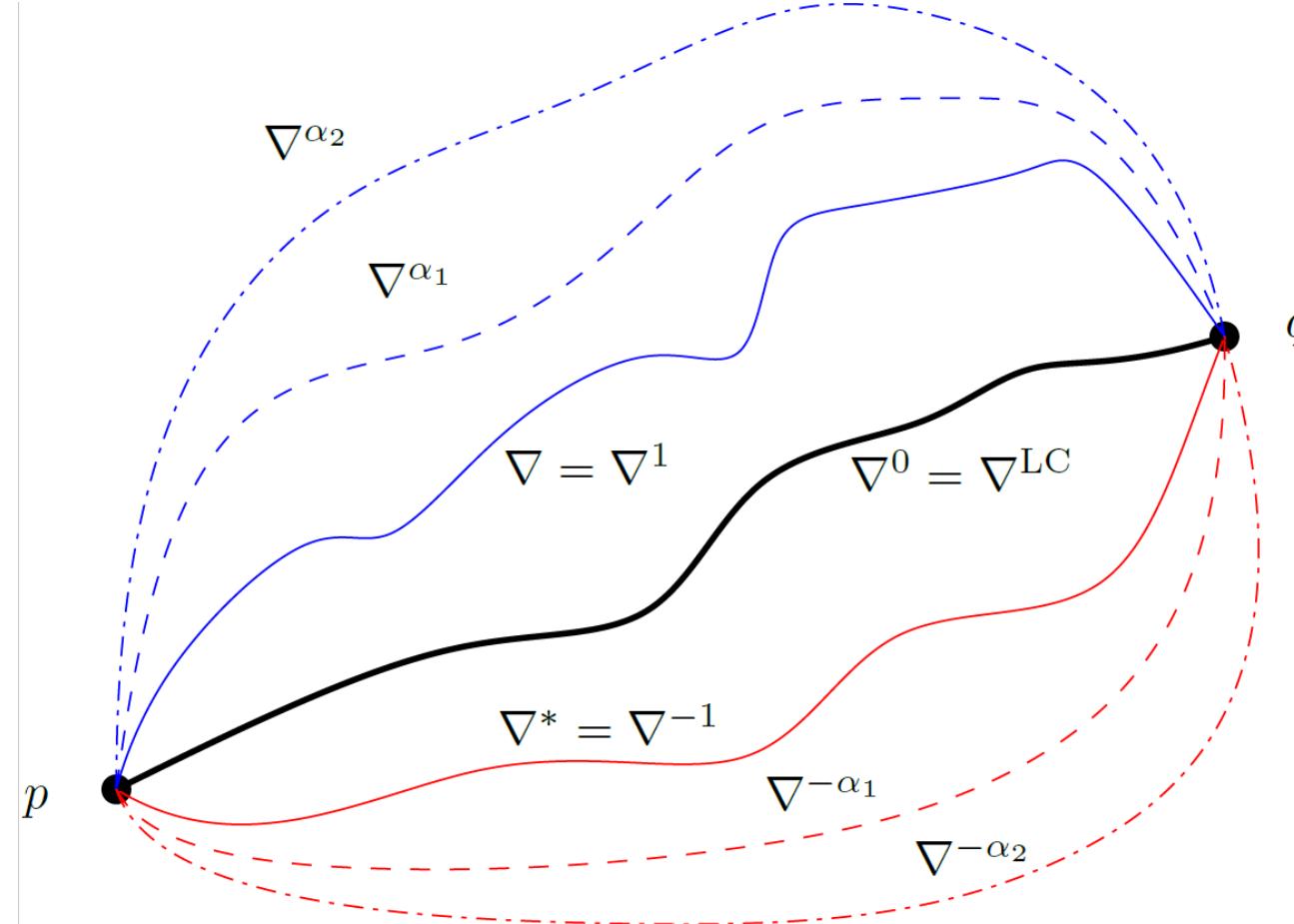
Two dual flat connections

Canonical distance = Bregman divergence induced by convex generator F

Bregman manifold/Hessian manifold

Generalize Euclidean space, very practical for computing!

From any dualistic structure to a 1-family of duality structures: α -geometries



How to choose α depending on applications?

Amari's expected α -geometry

- Given a parametric family of distributions, consider the Fisher information matrix and a family of connections: **α connections**

$$g_{p_\xi}(\nabla^\alpha \partial_i \partial_j(p_\xi), \partial_k(p_\xi)) = \Gamma^\alpha_{ijk}(p_\xi) = E_\xi \left[\left(\frac{\partial}{\partial \xi} \frac{\partial}{\partial \xi} \log(p_\xi) + \frac{1-\alpha}{2} \frac{\partial}{\partial \xi} \log(p_\xi) \frac{\partial}{\partial \xi} \log(p_\xi) \right) \frac{\partial}{\partial \xi^k} \log(p_\xi) \right]$$

- No associated distance in the alpha-expected geometry

Levi-Civita connection : $\nabla^0 = \nabla^{\text{LC}}$

From a dualistic structure to a 1-family of dually structures

- Let (M, g, ∇, ∇^*) be a dualistic structure: A dual pair of connections coupled to the metric (dual parallel transport is metric-compatible)
- We can build a **1-family of dualistic structures** $(M, g, \nabla^{-\alpha}, \nabla^\alpha)$
so that $\frac{\nabla^{-\alpha} + \nabla^\alpha}{2} = \nabla^0 = \nabla^{\text{LC}}$
- No distance associated here with the dualistic structure.

In particular, when $\alpha = 0$, $(M, g, \nabla^0, \nabla^0) = (M, g)$ the Riemannian geometry.
Thus information geometry generalizes Riemannian geometry

How to get dual connections?

- Historically, built the **e-connection** (exponential, $\alpha=1$) and **m-connection** (mixture, $\alpha=-1$) for statistical models

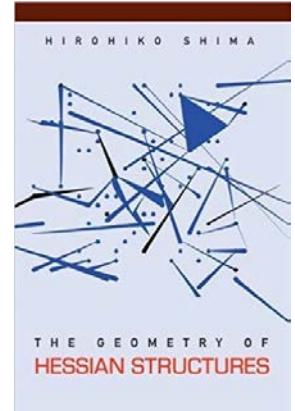
Log-likelihood $\ell(p_\xi)(x) = \ln p_\xi(x)$.

e-connection

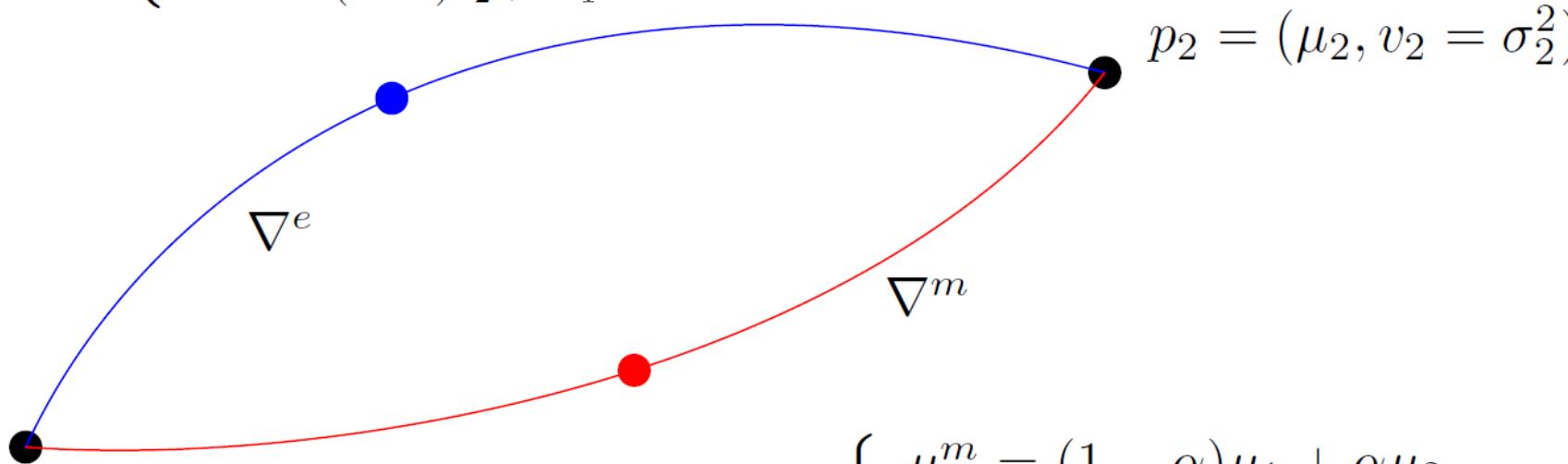
m-connection $g(\nabla_{\partial_i}^{(-1)} \partial_j, \partial_k) = \Gamma_{ij,k}^{(-1)} = E_\xi[(\partial_i \partial_j \ell + \partial_i \ell \partial_j \ell) (\partial_k \ell)]$

Dual with respect to Fisher information (Riemannian) metric

Example of dual e-/m-connections for the univariate Gaussian 2D manifold



$$(p_1 p_2)_{\alpha}^e = \begin{cases} \mu_{\alpha}^e = \frac{(1-\alpha)\mu_1 v_2 + \alpha \mu_2 v_1}{(1-\alpha)v_2 + \alpha v_1} \\ v_{\alpha}^e = \frac{v_1 v_2}{(1-\alpha)v_2 + \alpha v_1} \end{cases}$$



$$p_1 = (\mu_1, v_1 = \sigma_1^2)$$

$$(p_1 p_2)_{\alpha}^m = \begin{cases} \mu_{\alpha}^m = (1 - \alpha)\mu_1 + \alpha \mu_2 \\ v_{\alpha}^m = (1 - \alpha)v_1 + \alpha v_2 - \alpha(1 - \alpha)(\mu_1 - \mu_2)^2 \end{cases}$$

Misconception: The m-geodesic between two Gaussians of a Gaussian manifold is a Gaussian (and not a mixture of Gaussian!)
The Gaussian is obtained from linear interpolation on the moment parameters

Dual connections from a divergence $(M, {}_Dg, {}_D\nabla, {}_D\nabla^*)$

Dual connections from any smooth parametric distance, called a **divergence** D : D is not necessarily symmetric

- a tensor metric g : $g_{ij}(p_\xi) = \frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi^j} D(p_{\xi_1}, p_{\xi_2})|_{\xi_1=\xi_2=\xi}$

- a torsion-less connection ∇ :

$$\Gamma_{ijk}(p_\xi) = - \frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi^j} \frac{\partial}{\partial \xi^k} D(p_{\xi_1}, p_{\xi_2})|_{\xi_1=\xi_2=\xi}$$

Dual divergences $D^*(p_{\xi_1}, p_{\xi_2}) = D(p_{\xi_2}, p_{\xi_1})$

Symmetric divergences yields the same connection: The Levi-Civita connection

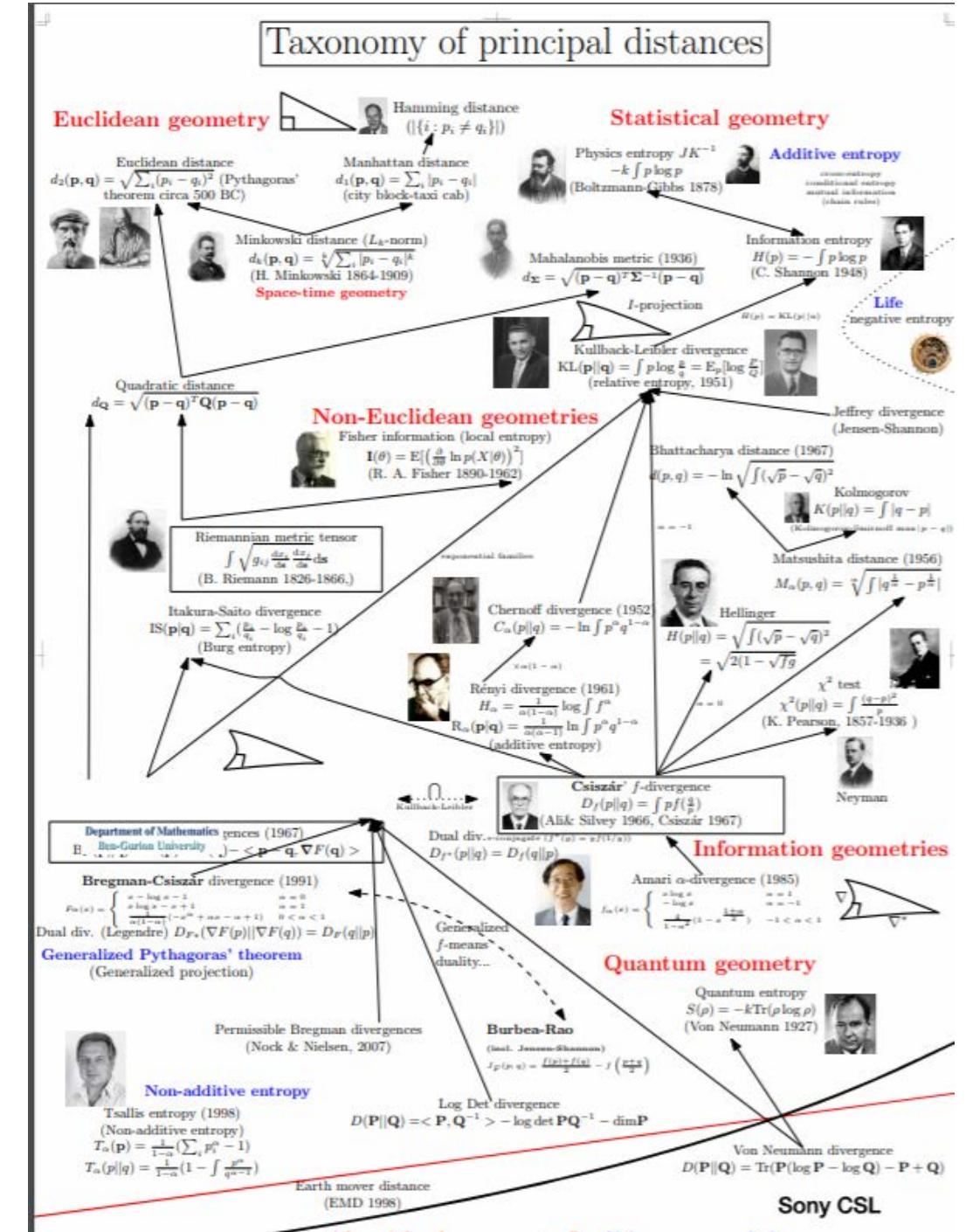
Many distances/divergences in information sciences

Divergence= discrepancy, dissimilarity, deviance between two probability distributions

Nowadays, smooth parametric dissimilarities (contrast function)

Distance is often thought as a **metric distance**:

- (a) $d(p, q) > 0$ if $p \neq q$; $d(p, p) = 0$;
- (b) $d(p, q) = d(q, p)$;
- (c) $d(p, q) \leq d(p, r) + d(r, q)$,



Divergences

- In information theory, relative entropy called **Kullback-Leibler divergence**

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

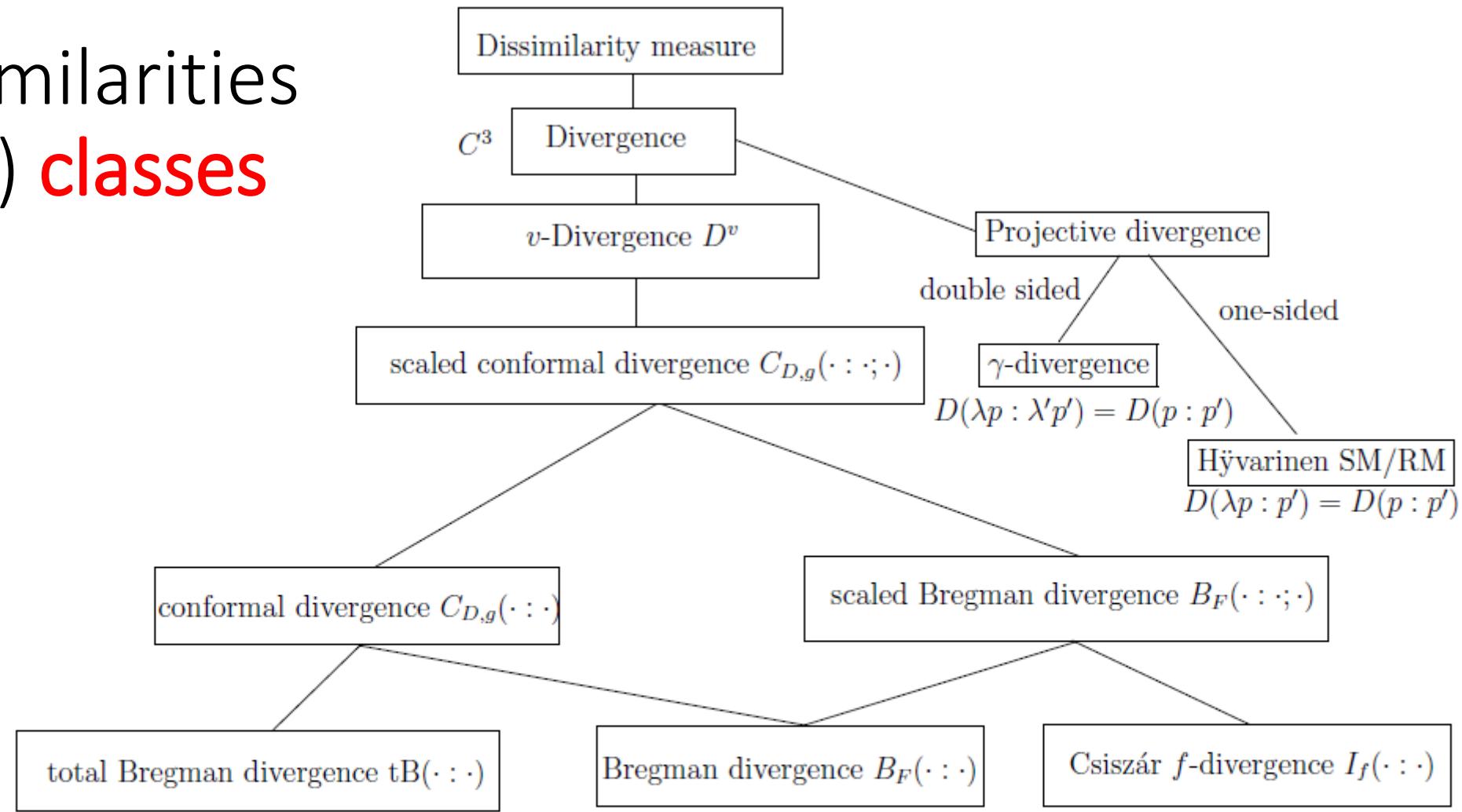
- Can be extended to **f-divergences**

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

$$D_f(P \parallel Q) = \int_{\Omega} f \left(\frac{p(x)}{q(x)} \right) q(x) d\mu(x).$$

- Distances can be scale-invariant (eg, Itakura-Saito), homogeneous, projective (work on unnormalized probability densities), etc.

Organize dissimilarities in (exhaustive) **classes**



$$D^v(P : Q) = D(v(P) : v(Q))$$

$$I_f(P : Q) = \int p(x) f\left(\left(\frac{q(x)}{p(x)}\right)\right) d\nu(x)$$

$$B_F(P : Q) = F(P) - F(Q) - \langle P - Q, \nabla F(Q) \rangle$$

$$tB_F(P : Q) = \frac{B_F(P : Q)}{\sqrt{1 + \|\nabla F(Q)\|^2}}$$

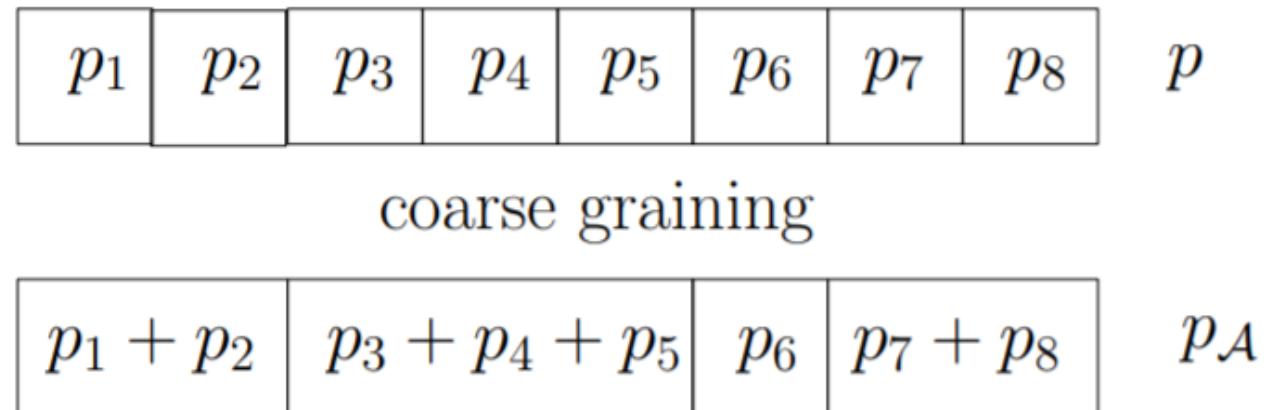
$$C_{D,g}(P : Q) = g(Q)D(P : Q)$$

$$B_{F,g}(P : Q; W) = WB_F\left(\frac{P}{Q} : \frac{Q}{W}\right)$$

Invariant divergence = f-divergences

- Lump or coarse-bin a separable distance, and ask for
information monotonicity

$$D(\theta_{\bar{\mathcal{A}}} : \theta'_{\bar{\mathcal{A}}}) \leq D(\theta : \theta')$$

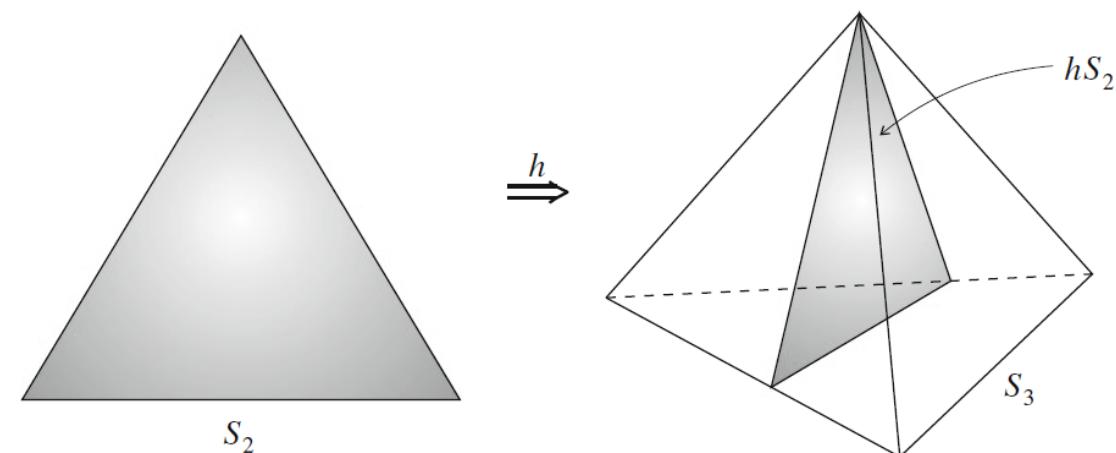


Theorem: The only monotone separable divergences are f-divergences (except the curious case of binary alphabets)

F-divergences are **invariant by diffeomorphisms** of the sample space

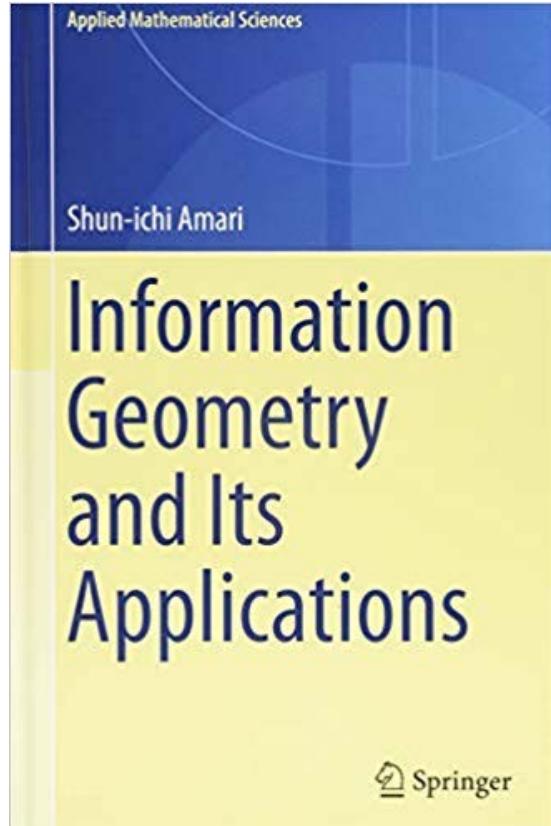
$$\begin{aligned} D_f(q_i, q_j) &= \int_y q_j(y) f\left(\frac{q_i(y)}{q_j(y)}\right) dy \\ &= \int_x p_j(x) |\mathcal{J}(x)|^{-1} f\left(\frac{p_i(x)|\mathcal{J}(x)|^{-1}}{p_j(x)|\mathcal{J}(x)|^{-1}}\right) |\mathcal{J}(x)| dx \\ &= \int_x p_j(x) f\left(\frac{p_i(x)}{p_j(x)}\right) dx = D_f(p_i, p_j). \end{aligned}$$

Statistical invariance



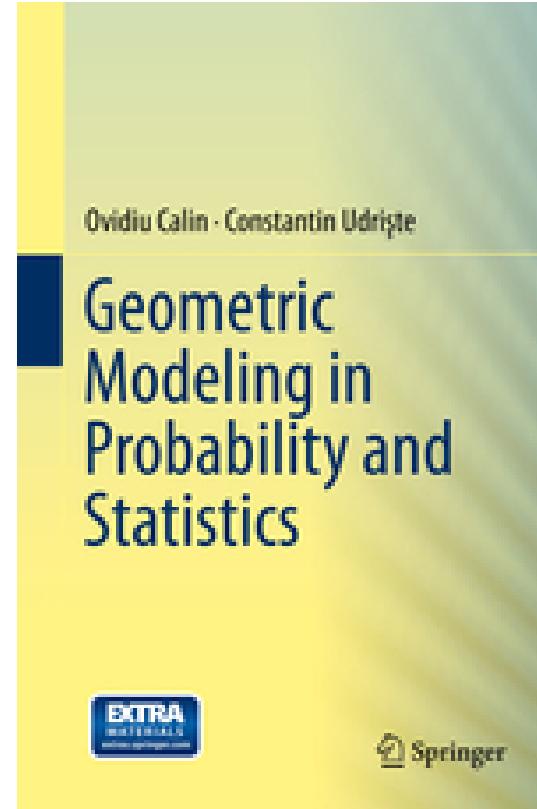
- Fisher-Rao distance is **independent of parameterization** (but FIM is covariant)
Same Fisher-Rao distance for parameterizations $\{N(\mu, \sigma)\}$ or $\{N(\mu, \sigma^2)\}$
- Fisher information metric is the only invariant metric tensor (up to a scale factor)
- Metric tensor induced by any **standard f-divergence** coincides with the Fisher information metric
- Dual connections induced by any f-divergence yield expected alpha-connections

Recommended textbooks+overview survey



2016

Very nice up-to-date survey including
Applications by the pioneer S.-i. Amari



2014

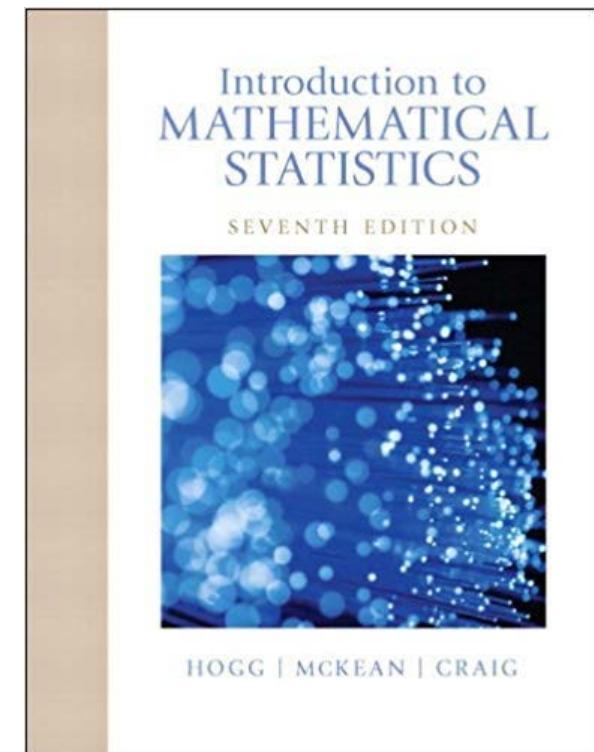
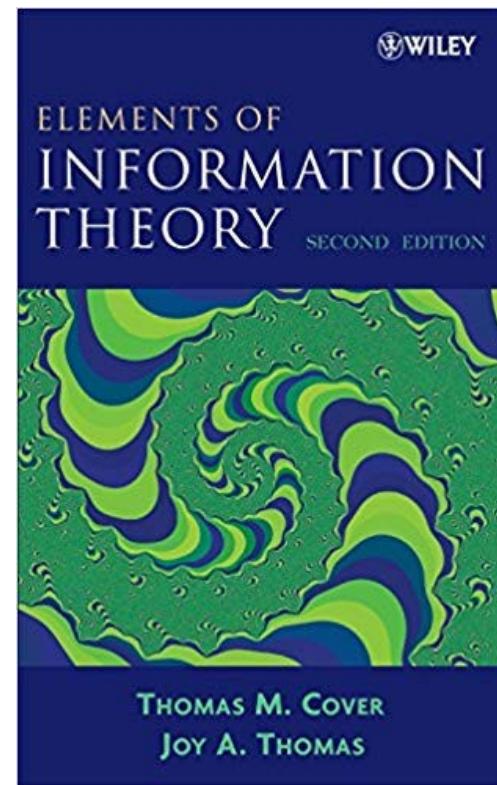
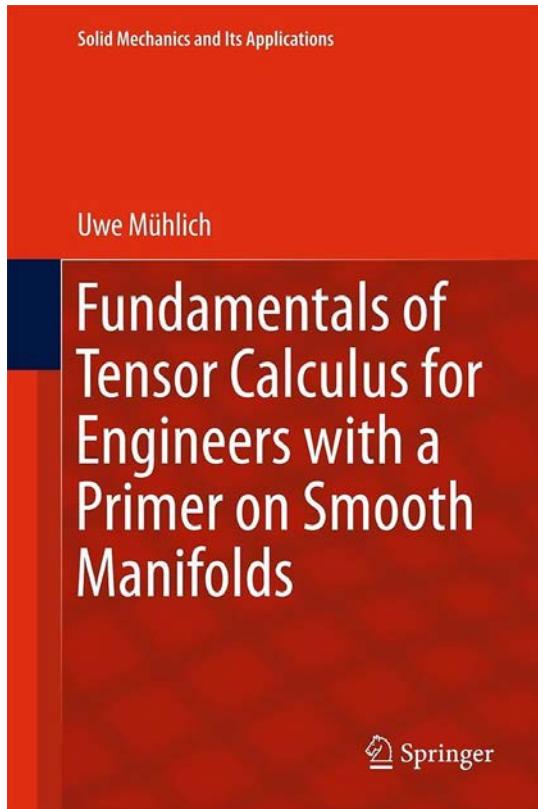
More details on differential geometry
with exercises

An elementary introduction
to information geometry

<https://arxiv.org/abs/1808.08271>

Prerequisite: Information sciences + Differential geometry

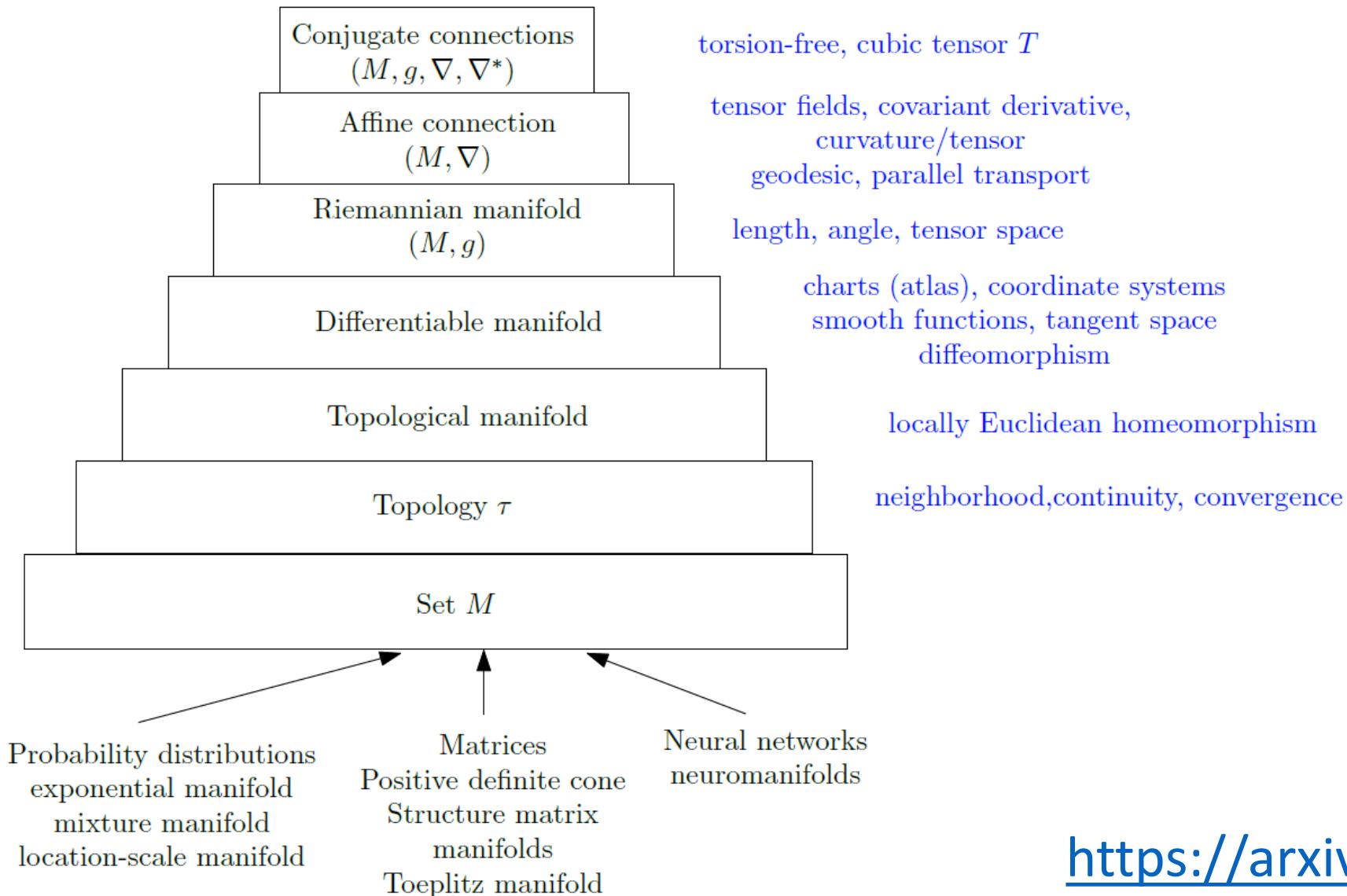
- Tensors + Manifolds
- Statistics + Information theory



Part I. Background

- Probability and statistical inference
 - Measures, random variables, Fisher information, exponential families
- Information theory and maximum entropy
 - Entropy, relative entropy (Kullback-Leibler divergence), maximum entropy principle
- Distances
 - Metrics, divergences, properties, information monotonicity, parametric families, f-divergences, Bregman divergences, Jensen divergences
- Geometry
 - Algebraic structures (dual vector/covector spaces, tensors), affine space, differential geometry (Riemannian, affine: uncoupling metric/connection)

Part II. Information-geometric structures



Information
geometry

Tensor analysis

Tensor algebra

Analysis

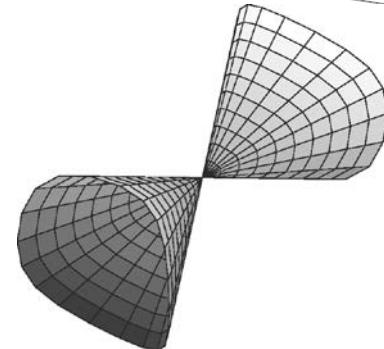
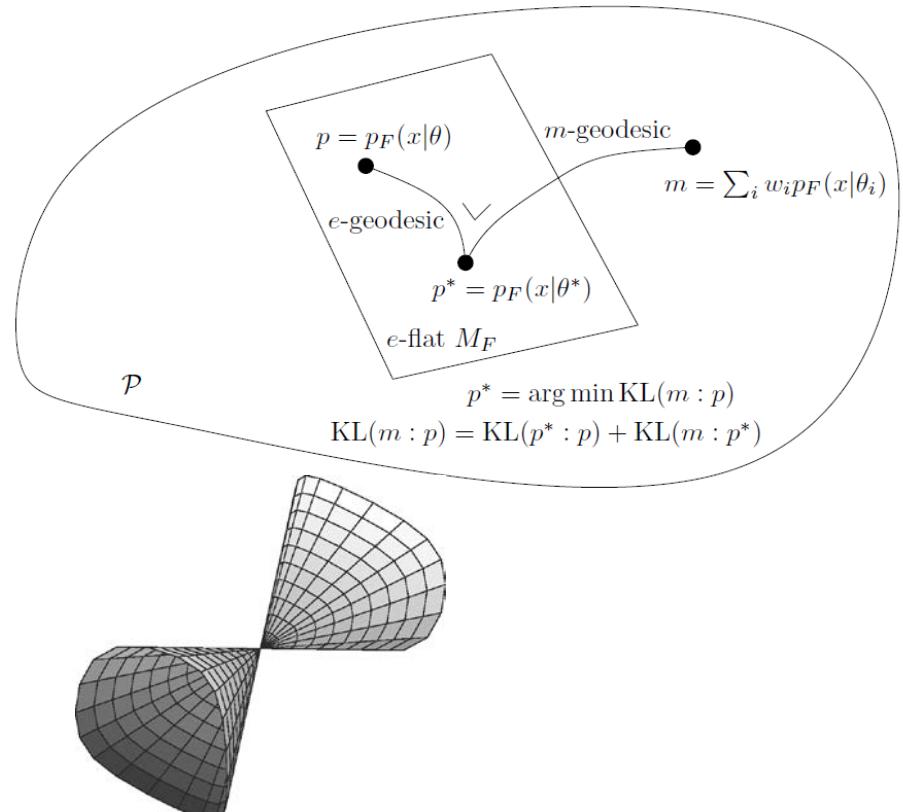
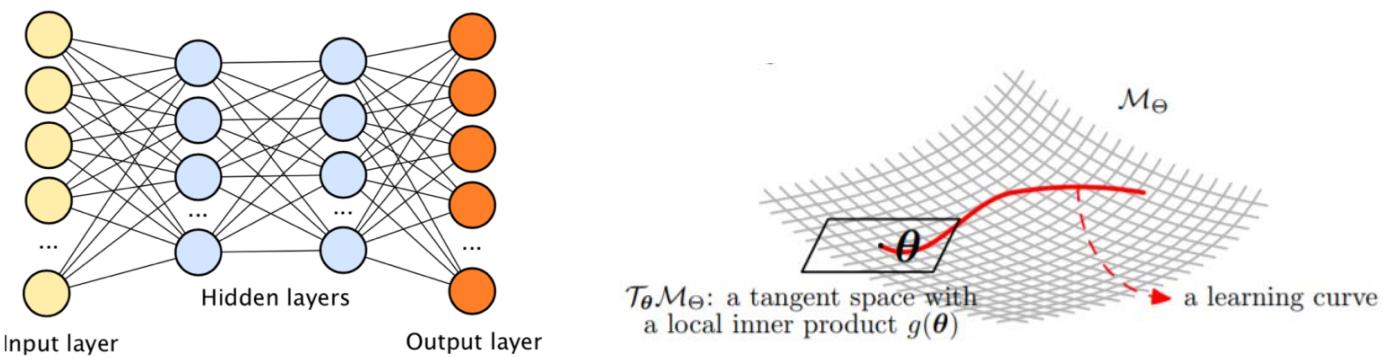
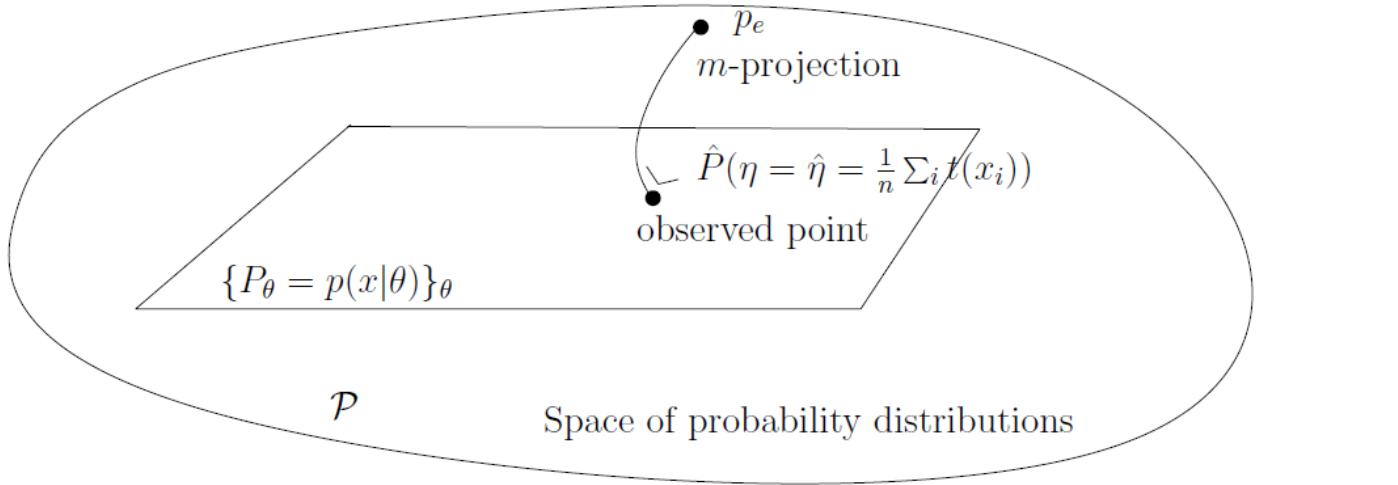
Topology

Part III. Applications

WHAT IS...

an Information Projection?

Empirical distribution : $p_e(X) = \frac{1}{n} \sum_{i=1}^n \delta(X - X(i))$
MLE = *m*-projection from p_e to the model submanifold



Singularities in neuromanifolds

Frank Nielsen

Communicated by Cesar E. Silva

Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

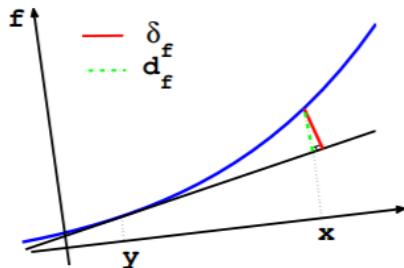
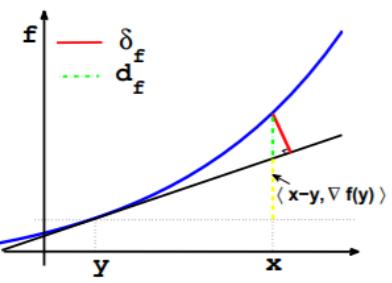
Definition The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,

$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}},$$

$\langle \cdot, \cdot \rangle$ is inner product
 $\langle \nabla f(y), \nabla f(y) \rangle$ generally.

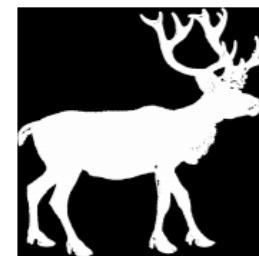
and $\|\nabla f(y)\|^2 =$

| X | $f(x)$ | $\delta_f(x, y)$ | t -center | ℓ_1 -norm BD center | Remark |
|-----------------------------|-----------------------------|--|---|--------------------------|------------------------------|
| \mathbb{R} | x^2 | $\frac{(x-y)^2}{\sqrt{1+4y^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total square loss (tSL) |
| $\mathbb{R} - \mathbb{R}_-$ | $x \log x$ | $\frac{x \log \frac{x}{y} + \bar{x} \log \frac{\bar{x}}{\bar{y}}}{\sqrt{1+y(1+\log y)^2 + \bar{y}(1+\log \bar{y})^2}}$ | $\prod_i (x_i)^{w_i}$ | $\sum_i x_i$ | total logistic loss |
| $[0, 1]$ | $-\log x$ | $\frac{\frac{x}{y} - \log \frac{x}{y} - 1}{\sqrt{1+y^{-2}}}$ | $\frac{\sum_i (x_i/(1-x_i))^{w_i}}{1 + \sum_i (x_i/(1-x_i))^{w_i}}$ | $\sum_i x_i$ | total Itakura-Saito distance |
| \mathbb{R}_+ | $-\log x$ | $\frac{\frac{x}{y} - \log \frac{x}{y} - 1}{\sqrt{1+y^{-2}}}$ | $\frac{1}{\sum_i w_i/x_i}$ | $\sum_i x_i$ | total squared Euclidean |
| \mathbb{R} | e^x | $\frac{e^x - e^y - (x-y)e^y}{\sqrt{1+e^{2y}}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total Mahalanobis distance |
| \mathbb{R}^d | $\ x\ ^2$ | $\frac{\ x-y\ ^2}{\sqrt{1+4\ y\ ^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total KL divergence (tKL) |
| \mathbb{R}^d | $x^t Ax$ | $\frac{(x-y)^t A(x-y)}{\sqrt{1+4\ Ay\ ^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total squared Frobenius |
| Δ^d | $\sum_{j=1}^d x_j \log x_j$ | $\frac{\sum_{j=1}^d x_j \log \frac{x_j}{y_j}}{\sqrt{1+\sum_{j=1}^d y_j(1+\log y_j)^2}}$ | $c \prod_i (x_i)^{w_i}$ | $\sum_i x_i$ | |
| $\mathbb{C}^{m \times n}$ | $\ x\ _F^2$ | $\frac{\ x-y\ _F^2}{\sqrt{1+4\ y\ _F^2}}$ | $\frac{\ x-y\ _F^2}{\sqrt{1+4\ y\ _F^2}}$ | $\sum_i x_i$ | |

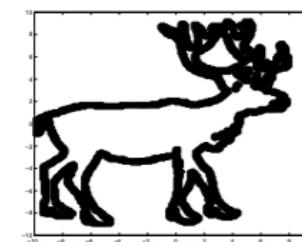


t-center: $\bar{x} = \arg \min_x \delta_f^1(x, E) = \arg \min_x \sum_{i=1}^n \delta_f(x, x_i)$

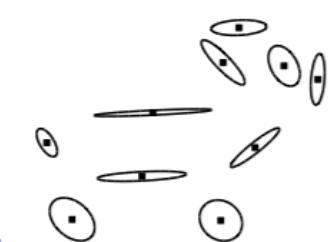
Robust to noise/outliers



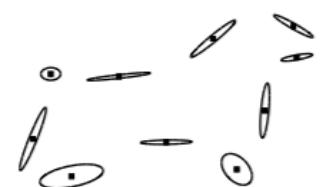
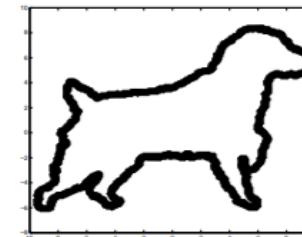
(m)



(n)



(o)



Total Bregman divergence and its applications to DTI analysis

IEEE Transactions on medical imaging, 30(2), 475-483, 2010.

Definition The total Bregman divergence (TBD) δ_f associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,

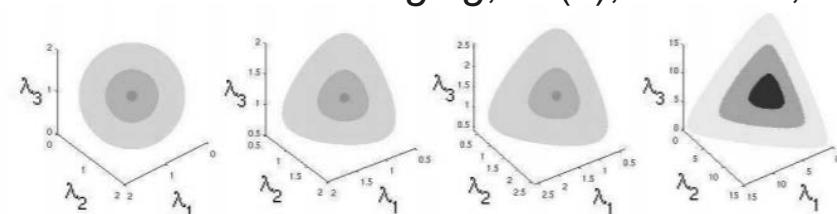
$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}}, \quad (2)$$

$\langle \cdot, \cdot \rangle$ is inner product as in definition II.1, and $\|\nabla f(y)\|^2 = \langle \nabla f(y), \nabla f(y) \rangle$ generally.

$$tKL(P, Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1 + \int (1 + \log q)^2 q dx}} \\ = \frac{\log(\det(P^{-1}Q)) + \text{tr}(Q^{-1}P) - n}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{n(1+\log 2\pi)}{2} \log(\det Q)}}$$

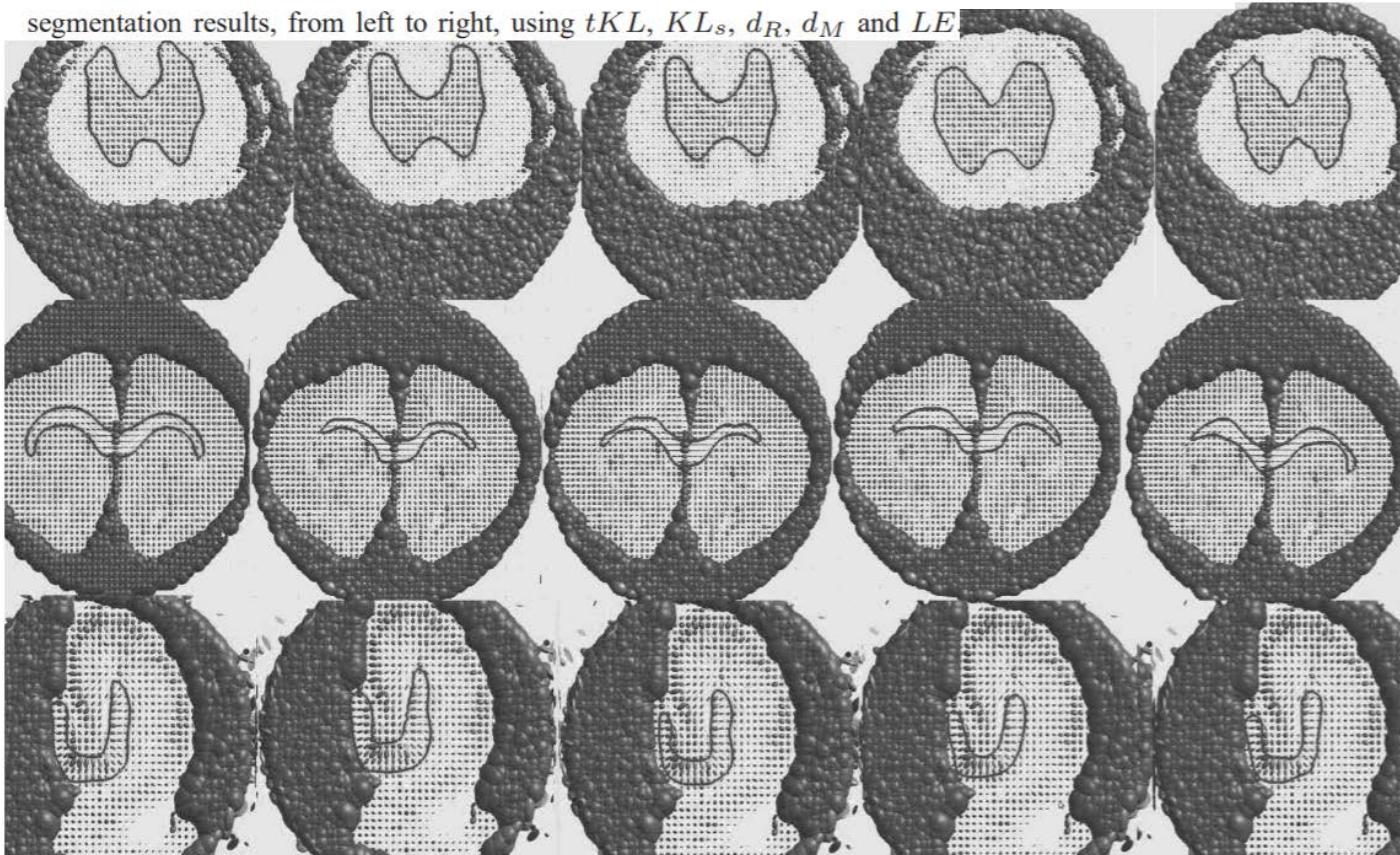
$$tKL(P, Q) = tKL(A'PA, A'QA), \quad \forall A \in SL(n),$$

$$tSL(P, Q) = \frac{\int (p - q)^2 dx}{\sqrt{1 + \int (2q)^2 q dx}} = \\ \frac{1/\sqrt{\det(2P)} + 1/\sqrt{\det(2Q)} - 2/\sqrt{\det(P+Q)}}{(2\pi)^n + 4\sqrt{(2\pi)^n}/\sqrt{\det(3Q)}}$$



The isosurfaces of $d_F(P, I) = r$, $d_R(P, I) = r$, $KL_s(P, I) = r$ and $tKL(P, I) = r$ shown from left to right. The three axes are eigenvalues of P .

segmentation results, from left to right, using tKL , KL_s , d_R , d_M and LE .



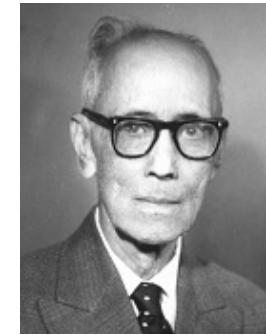
The origin of dual connections

- Aleksander P. Norden (1904-1993), **relative geometry**
(equiaffine torsion-free connection)

Russian book "Spaces with an affine connection" (1976)



Александр Петрович
НОРДЕН
Norden



Sen

- Rabindra Nath Sen (1896-1974), “**Senian geometry**”

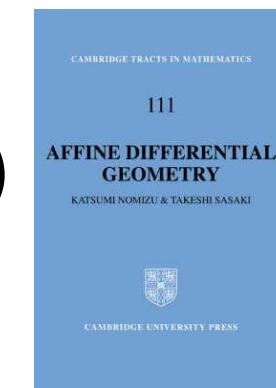


Nomizu

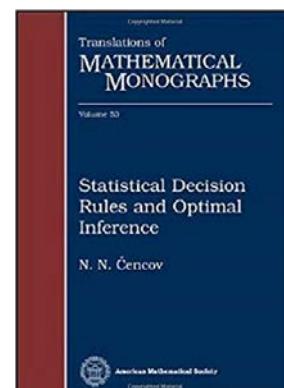


Amari

- Nomizu and Sasaki's **Affine differential geometry** (geometry of immersions)



- **Information geometry** (Chentsov's category approach and Amari)



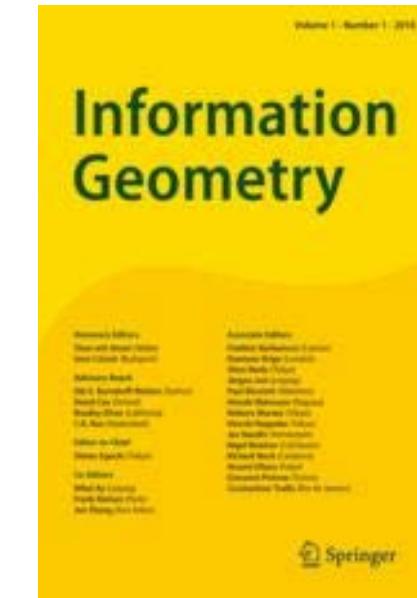
Thank you.



Frank Nielsen

- [What is new @FrnkNlsn](#)
- [Publications ResearchGate](#) [DBLP Slides \[video\]](#)
- [Blog](#)
- Textbooks:
 - [Introduction to HPC with MPI for Data Science](#), Springer 2018
 - [A Concise and Practical Introduction to Programming in Python](#), Springer 2019
 - [Visual Computing: Geometry, Graphics, and Vision](#), Springer 2019
- Edited books:
 - [Geometric Structures of Information](#), Springer 2019
 - [Computational Information Geometry For Image and Pattern Recognition](#), Springer 2019
 - [Differential Geometrical Theory of Statistics](#), MDPI 2019
 - [Geometric Theory of Information](#), Springer 2014

<https://franknielsen.github.io/>



<http://forum.cs-dc.org/category/72/geometric-science-of-information>

The screenshot shows the homepage of the CS-DC forum. At the top, there is a navigation bar with icons for user profile, search, and other site functions. To the right of the navigation bar are 'Register' and 'Login' buttons. Below the navigation bar, the page title 'Home / Geometric Science of Information' is displayed. A 'SUBCATEGORIES' section follows, listing several categories with their respective counts of topics and posts. To the right of the categories, there is a list of recent forum posts. The overall layout is clean and modern, with a white background and light blue accents for links.

| SUBCATEGORIES | TOPICS | POSTS | LAST POST | MESSAGE | |
|--|---|-------|-----------|-------------------------------|---|
| Register | Working group and Diffusion - submit events - new paper - projects - jobs - seminar (...) | 148 | 148 | Niltoida 4 months ago | AngularJS is a toolkit for creating frameworks, fully extensible and works well |
| Jobs offers - Call for projects | | 148 | 148 | Niltoida 4 months ago | AngularJS is a toolkit for creating frameworks, fully extensible and works well |
| GSI FORGE | Packages for data analysis and modelling | 14 | 15 | Geo-Sci-Info 5 months ago | UMAP - Leland McInnes, John Healy, James Melville |
| Partners - Friends - GdRs GeoSto - MIA - ISIS - Azimuth Project | Azimuth http://www.azimuthproject.org/ - ISIS http://gdr-isis.fr/ - MIA https://fadili.users.greyc.fr/mia | 6 | 6 | Geo-Sci-Info 3 years ago | |
| Preprints - Books - Archivs - Journal special edition (Entropy...) | Call for paper Entropy - new books - new papers - preprints | 12 | 12 | Geo-Sci-Info about a year ago | Special Issue MDPI "Joseph Fourier 250th Birthday: Modern Fourier Analysis and |
| e-room - visio-conference - seminar streaming - reservation | | 1 | 1 | Geo-Sci-Info 3 years ago | The CS-DC put at disposition of the |

Acknowledgements

- My collaborators (incl. Jean-Daniel Boissonnat, Gaetan Hadjeres, Richard Nock, Ke Sun, Olivier Schwander, ...)
- Images of these slides were mostly courtesy of the Internet.
Copyrights hold by owners
- Some figures were drawn in powerpoint by Joffrey Poitevin

Geometry and its language affordance

- What is geometry?
 - Science of measurements
 - Science of figures (ruler and compass construction)
 - Axioms, consistency and deductive theorems (Euclidean/hyperbolic)
 - Science of invariance (congruence of figures/Erlangen program)
 - Etc.
- Geometry has its own human language for reasoning
 - What is the distance between two points?
 - What is the midpoint between two points?
 - What is the closest point of a surface from a given point? (projection)
 - Balls and space of balls binary operations (CSG construction)

Probability and statistics

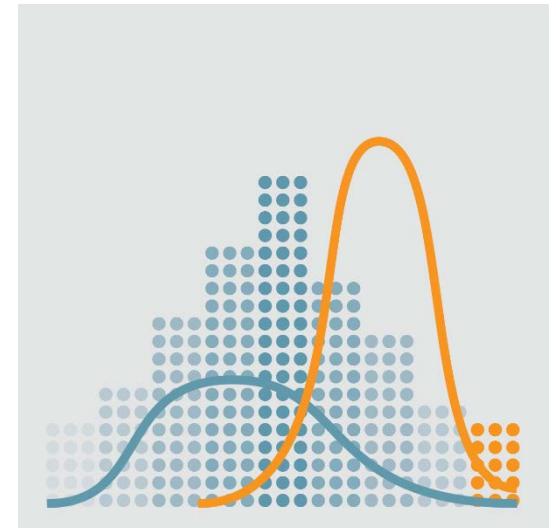


Frank Nielsen



Sony CSL

Background



Outline

- Classic probability theory
- Modern theory: Probability measures
- Statistical inference:
 - method of moments,
 - Maximum Likelihood Estimator (MLE),
 - sufficient statistics,
 - Fisher information (curvature interpretation)
- Exponential families

Discrete random variables $X \sim f(x)$

Jacob Bernoulli
(1654,1705)



- Bernoulli distribution (coin tossing), Binomial distribution (tossing a coin n times), multinomial distribution (throwing a dice n times), Poisson distributions, etc.

- Sample space, probability of events

$$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q$$

- Probability mass function

$$f(k; p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases} \quad f(k; p) = p^k (1 - p)^{1-k} \quad \text{for } k \in \{0, 1\}$$

- Cumulative distribution function (CDF)

$$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$$

- Expectation

$$\mathbb{E}[X] = \Pr(X = 1) \cdot 1 + \Pr(X = 0) \cdot 0 = p \cdot 1 + q \cdot 0 = p.$$

- Variance

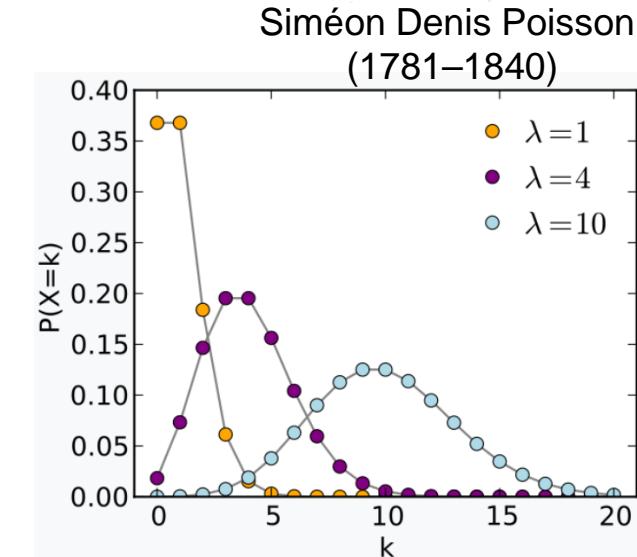
$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p) = pq$$

Discrete random variable $X \sim f(x)$



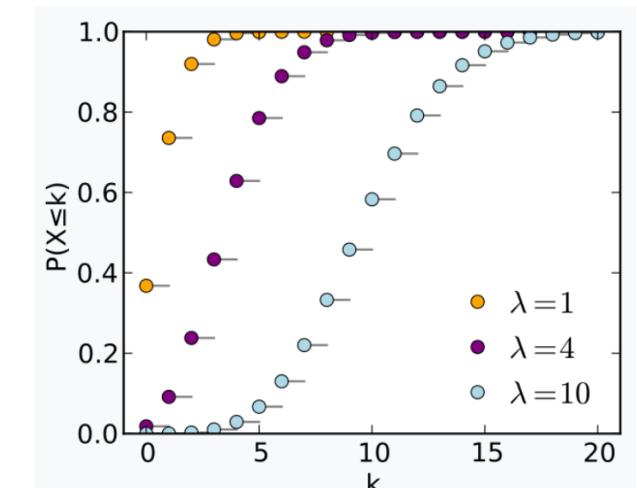
- Support $0, 1, \dots$
- Probability mass function:

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



- Cumulative distribution function

$$e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$$



- Mean and variance: $\lambda = \text{E}(X) = \text{Var}(X)$

Continuous random variable $X \sim f(x)$



1777-1855

- Probability density function (PDF)

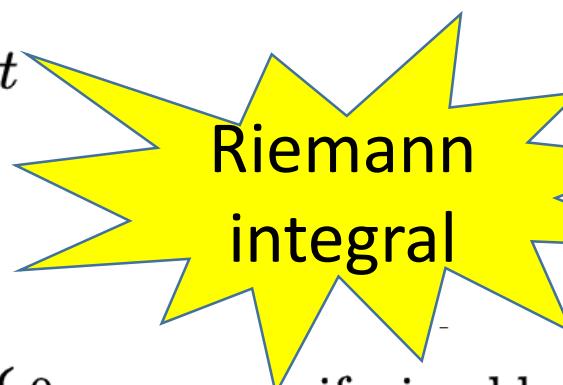
$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Normal or Gaussian distribution

$$f(x | \mu, \sigma^2) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right) \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- CDF of standard normal distribution $N(0,1)$

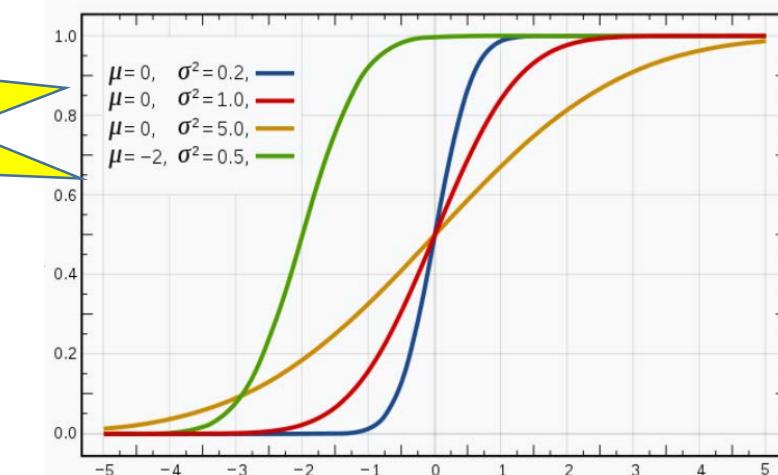
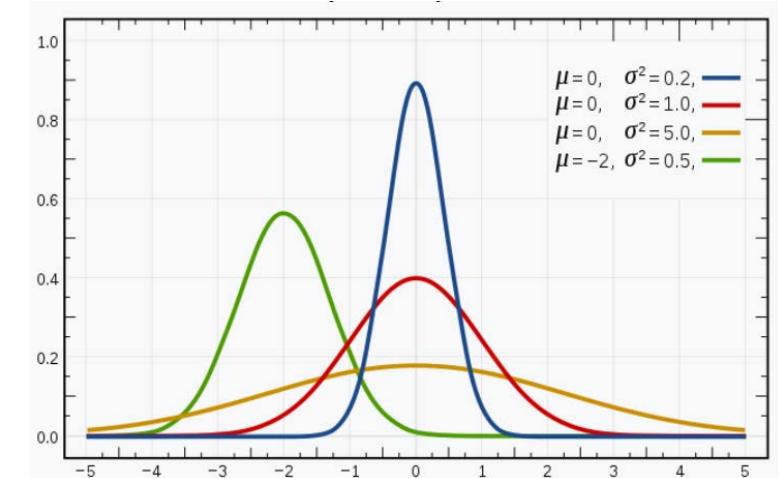
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$



- Expectation and moments

$$\mathbb{E}[X] = \int x f(x) dx.$$

$$\mathbb{E}[X^p] = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even.} \end{cases}$$



Continuous random variable $X \sim f(x)$



- PDF:

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma} \right)^2 \right]} = \frac{1}{\pi\gamma} \left[\frac{\gamma^2}{(x-x_0)^2 + \gamma^2} \right]$$

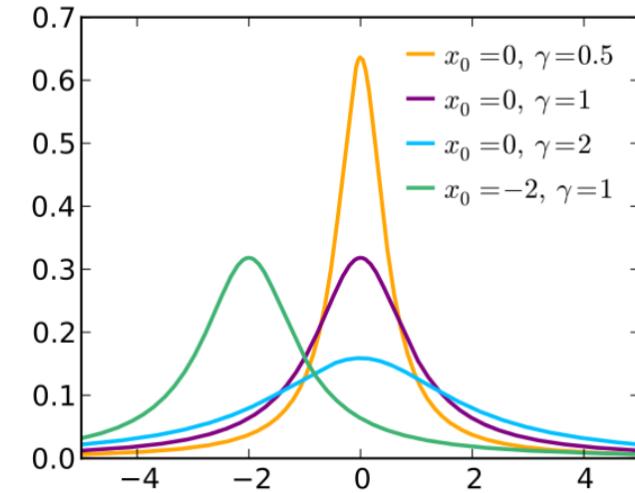
- CDF:

$$F(x; x_0, \gamma) = \frac{1}{\pi} \arctan \left(\frac{x - x_0}{\gamma} \right) + \frac{1}{2}$$

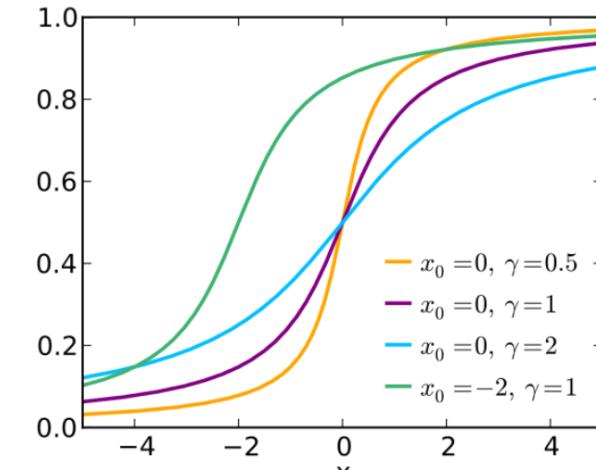
- Cauchy distributions **do not have finite moments** of any order! No expectation (improper integral)
- Location-scale family, standard Cauchy

$$g(x | \mu, \sigma) = \frac{1}{\sigma} \psi \left(\frac{x - \mu}{\sigma} \right) = \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-\mu}{\sigma} \right)^2}, \quad -\infty < x < \infty$$

Augustin-Louis Cauchy
(1789-1857)



$$\psi(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$



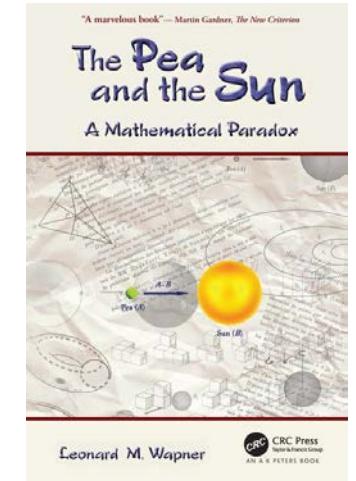
Probability measures

- additive law of probability for possibly countably infinite pairwise mutually exclusive events

$$\Pr(\cup_i E_i) = \sum_i \Pr(E_i)$$

- Interpreted as volumes of events for disjoint events

$$\mu(E) = \sum_i \mu(E_i)$$



- But Banach-Tarsky's paradox kicks in: for an uncountably sample space there exists a set S which can be partitioned into two disjoint congruent sets S₁ and such that

$$\mu(S) = \mu(S_1) + \mu(S_2) = 2\mu(S)$$



Measure theory: σ -algebra

- Cannot consider the power set for continuous sample spaces
- Let us define an algebra of measurable events: the **σ -algebra**
 1. $X \in \mathcal{A}$,
 2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$, and
 3. $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$.

A σ -algebra \mathcal{A} is an algebra that is closed under countably many unions:

 4. $\forall i \in \mathbb{N}, A_i \in \mathcal{A} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}$.
- σ -algebra generated by a set S : $\sigma(\mathcal{S})$

Smallest σ -algebra with respect to set inclusion

Measure space $(\mathbb{X}, \mathcal{A}, \mu)$

A measure μ is defined on a *measurable space* $(\mathbb{X}, \mathcal{A})$ as a map $\mu : \mathcal{A} \rightarrow [0, \infty]$ that is countably additive for pairwise disjoint subsets A_i 's:

$$\mu(\cup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mu(A_i).$$

- **Borel sets** $\mathcal{B}(\mathbb{R}^d)$ σ -algebra generated by all open intervals

$$\sigma(\mathcal{S}) \quad \mathcal{S} := \{(a, b) \in \mathbb{R} : a < b\}$$

- **Counting measure:** σ -algebra is the power set $2^{\mathbb{X}}$ and the measure is defined by cardinality $\mu_c(A) = |A|$

- **Lebesgue measure:** Volume for open boxes

$$\mu(A) = \prod_{i=1}^d (b_i - a_i)$$
$$A = \{x \in \mathbb{R}^d : \forall i \in [d], a_i < x_i < b_i\}$$

In the probability space $(\mathbb{X}, \mathcal{B}(\mathbb{R}^d), \mu_L)$ there are subsets that are not measurable

Measurable function

- Consider two measurable spaces: $(\mathbb{X}, \mathcal{A})$ $(\mathbb{Y}, \mathcal{B})$
- Preimage: $f^{-1}(B) := \{x \in \mathbb{X} : f(x) \in B\}$
- Measurable function: $f : (\mathbb{X}, \mathcal{A}) \rightarrow (\mathbb{Y}, \mathcal{B})$
If and only if the preimages $f^{-1}(B)$ of $B \in \mathcal{B}$ are in \mathcal{A} for all B
- Indicator function: $I_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$
- Simple function:
 $X(\omega) = \sum_{i=1}^k x_i I_{A_i}(\omega)$, where $x_i \in [0, \infty)$, $A_i \in \mathcal{A}$ with $A_i \cap A_j = \emptyset$

Lebesgue integration

- Riemann integral (signed area under the curve) not enough
- Integral of a simple function

$$\int X(\omega) \mu(d\omega) := \mu(X) = \sum_{i=1}^k x_i \mu(A_i).$$

- Other notations:

$$\int X d\mu(\omega) \quad \int X d\mu$$

- Integral of positive measurable functions:

$$\mu(X) = \int X(\omega) \mu(d\omega) = \sup\{\mu(X^*) : X^* \text{ is simple, } X^* \leq X\}.$$

- In general, for a measure, decompose into positive/negative measures:

$$\mu(X) = \int X(\omega) \mu(d\omega) = \int X^+(\omega) \mu(d\omega) - \int X^-(\omega) \mu(d\omega).$$

Random variables and expectations

- A **random variable** X is a real-valued measurable function:

$$X(\omega) : (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

- **Probability:**

$$\Pr(\omega \in A) = \mu(A)$$

- The **expectation** of a discrete or a continuous random variable writes similarly using probability measure theory:

$$E[X_1] = \int X_1(\omega) \mu_c(d\omega),$$

$$E[X_2] = \int X_2(\omega) \mu_L(d\omega).$$

Density and dominating measure

- For a measure space $(\mathbb{X}, \mathcal{A}, \mu)$ and a measurable function f , define the measure $\nu(A) := \int_A f d\mu = \int 1_A(x) f(x) \mu(dx).$

For example, the Gaussian density is formed from the Lebesgue density

$$(\mathbb{X}, \mathcal{B}(\mathbb{R}), \mu_L) \quad f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- **Absolute continuity:** $\nu \ll \mu \quad \forall A \in \mathcal{A}, \quad \mu(A) = 0 \Rightarrow \nu(A) = 0.$
 ν is dominated by μ

Let $\lambda = \frac{\mu+\nu}{2}$ then $\mu, \nu \ll \lambda$

Radon-Nikodym theorem, density

Theorem 1 (Radon-Nikodym) Let $(\mathbb{X}, \mathcal{A}, \mu)$ be a σ -finite measure space. Assume $\nu \ll \mu$. Then there exists f such that

$$\nu(A) = \int_A f d\mu,$$

Thus when $\nu \ll \mu$, ν has a density f wrt to μ denoted by $f = \frac{d\nu}{d\mu}$.

Many properties: If $\nu \ll \mu \ll \lambda$, then

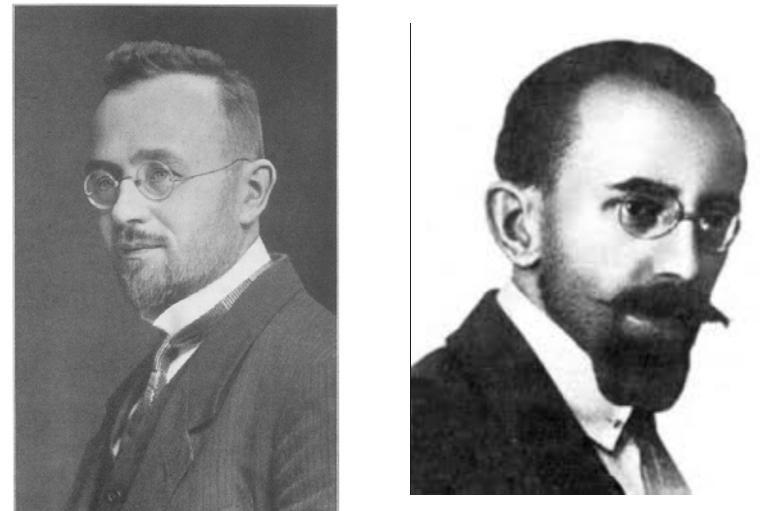
$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu} \frac{d\mu}{d\lambda} \quad \lambda\text{-almost everywhere.}$$

In particular, if $\mu \ll \nu$ and $\nu \ll \mu$, then

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu} \right)^{-1} \quad \nu\text{-almost everywhere.}$$

If $\mu \ll \lambda$ and g is a μ -integrable function, then

$$\int_X g d\mu = \int_X g \frac{d\mu}{d\lambda} d\lambda.$$



D. J. Radon

Summary

- Probability measure theory overcomes the *limitations* (analysis: convergence in limits) and *pitfalls* of classic probability theory (non-measurable sets due to Banach-Tarsky paradox, bypassed by the introduction of sigma-algebra)
- Estimators: method of moments, maximum likelihood
- Fisher's mathematical statistics define consistency and bias of estimators, efficiency by matching the Cramer-Rao lower bound. Asymptotic normality and equivariance of MLE
- Sufficient statistics contain all information about the parameter to infer: Statistical lossless compression. Exponential families have finite sufficient statistics and dual parameterization for MLE

Statistical inference: Estimators

- Given n independent and identically distributed observations, estimate the underlying distribution (probability density)
- Assume the density is **parametric**
- One of the oldest method is the **method of moments**:

Match the **distribution moments** with the **sample moments**

Consider the **uniform distribution** on the interval $[a, b]$, $U(a, b)$. If $W \sim U(a, b)$ then we have

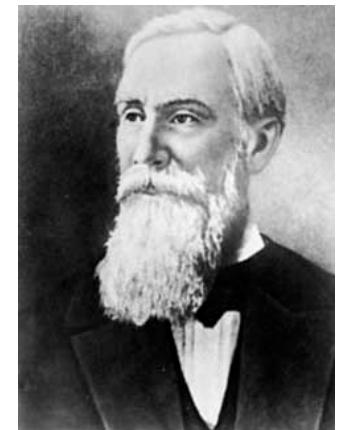
$$\mu_1 = E[W] = \frac{1}{2}(a + b)$$

$$\mu_2 = E[W^2] = \frac{1}{3}(a^2 + ab + b^2)$$

Solving these equations gives

$$\hat{a} = \mu_1 \pm \sqrt{3(\mu_2 - \mu_1^2)}$$

$$\hat{b} = 2\mu_1 - a$$



Pafnuty Chebyshev
(1821-1894)

- Infinitely many (point) estimators!

Maximum likelihood estimator (MLE)



Parametric family:

$$\mathcal{F} = \{p_\theta(x) \mid \theta \in \Theta\}$$

- **Likelihood function:** Function of the parameter $\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$

$$\mathcal{L}(\theta \mid x) = f_\theta(x),$$

- **Maximum likelihood estimate:**

$$\hat{\theta} \in \{\arg \max_{\theta \in \Theta} \mathcal{L}(\theta; x)\}$$

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n p(x_i; \theta)$$

$$\frac{1}{n} \sum_{i=1}^n \log p(x_i; \theta) \quad I(x; \theta) = \log p(x; \theta)$$

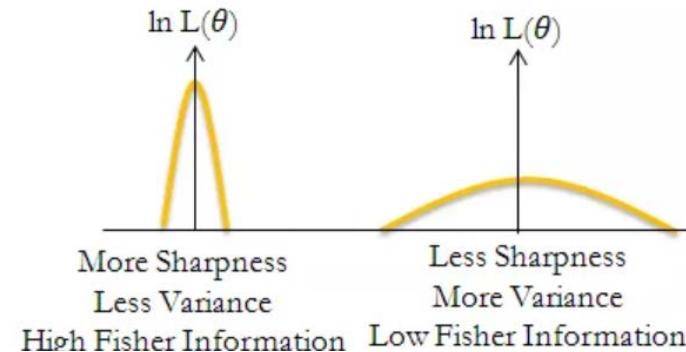
$$\begin{aligned} l(\mu, \sigma^2; x_1, \dots, x_n) &= \ln(L(\mu, \sigma^2; x_1, \dots, x_n)) \\ &= \ln\left((2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2\right)\right) \\ &= \ln\left((2\pi\sigma^2)^{-n/2}\right) + \ln\left(\exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2\right)\right) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 \end{aligned}$$

- **Consistent** method: converge in probability to the true value

$$\hat{\theta}_{\text{mle}} \xrightarrow{P} \theta_0$$

Fisher information

$$\text{Curvature} = -\frac{\partial^2}{\partial \theta^2} [\ln L(\theta)]$$



$$I(\theta) = E\left[\left(\frac{\partial \ell(x; \theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ell(x; \theta)}{\partial \theta^2}\right]$$

- measure the **amount of information** that an observable random variable X carries about an unknown parameter θ

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2 \middle| \theta\right] = \int \left(\frac{\partial}{\partial \theta} \log f(x; \theta)\right)^2 f(x; \theta) dx.$$

- Fisher information interpreted as the **curvature** of the graph of the log-likelihood: Near the MLE, high Fisher information indicates that the maximum is sharp, low Fisher information indicates that the maximum is shallow (many nearby values with a similar log-likelihood).

Cramer-Rao lower bound (CRLB)

- The variance of any unbiased estimator is lower bounded by the inverse of the Fisher information:

$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- Fisher information:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ell(x; \theta)}{\partial \theta}\right)^2\right]$$

Cramer-Rao lower bound: Multivariate case

Löwner partial ordering on positive-semi-definite matrices: $A \succeq B \Leftrightarrow A - B \succeq 0$

CRLB Theorem:

$$\text{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta_0)^{-1}$$

$$\begin{aligned}[I(\theta)]_{ij} &= E_\theta \left[\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right], \\ &= \int \left(\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right) p_\theta(x) dx.\end{aligned}$$

Under regularity conditions:

$$[I(\theta)]_{ij} = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x) \right]$$

Equivalent representation $[I(\theta)]_{ij} = 4 \int_{x \in \mathcal{X}} \frac{\partial}{\partial \theta_i} \sqrt{p_\theta(x)} \frac{\partial}{\partial \theta_j} \sqrt{p_\theta(x)} dx.$

Properties of the Maximum Likelihood Estimator (MLE)

- **Consistency:** $\hat{\theta}_n \rightarrow \theta_0$
- **Efficiency:** Variance of estimator matches the Cramer-Rao lower bound (CRLB)
- **Equivariance:** MLE estimator of Gaussian variance σ^2 is equivariant to MLE estimator of deviation σ
$$\widehat{f(\theta)} = f(\hat{\theta})$$
- **Asymptotic normality** (convergence in distribution):

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, I^{-1}(\theta))$$

Some properties of the Fisher Information Matrix

$$g_{ij}(\xi) = E_\xi[\partial_{\xi^i} \ln p_\xi \cdot \partial_{\xi^j} \ln p_\xi] = \int_{\mathcal{X}} \partial_{\xi^i} \ln p_\xi(x) \cdot \partial_{\xi^j} \ln p_\xi(x) \cdot p_\xi(x) dx.$$

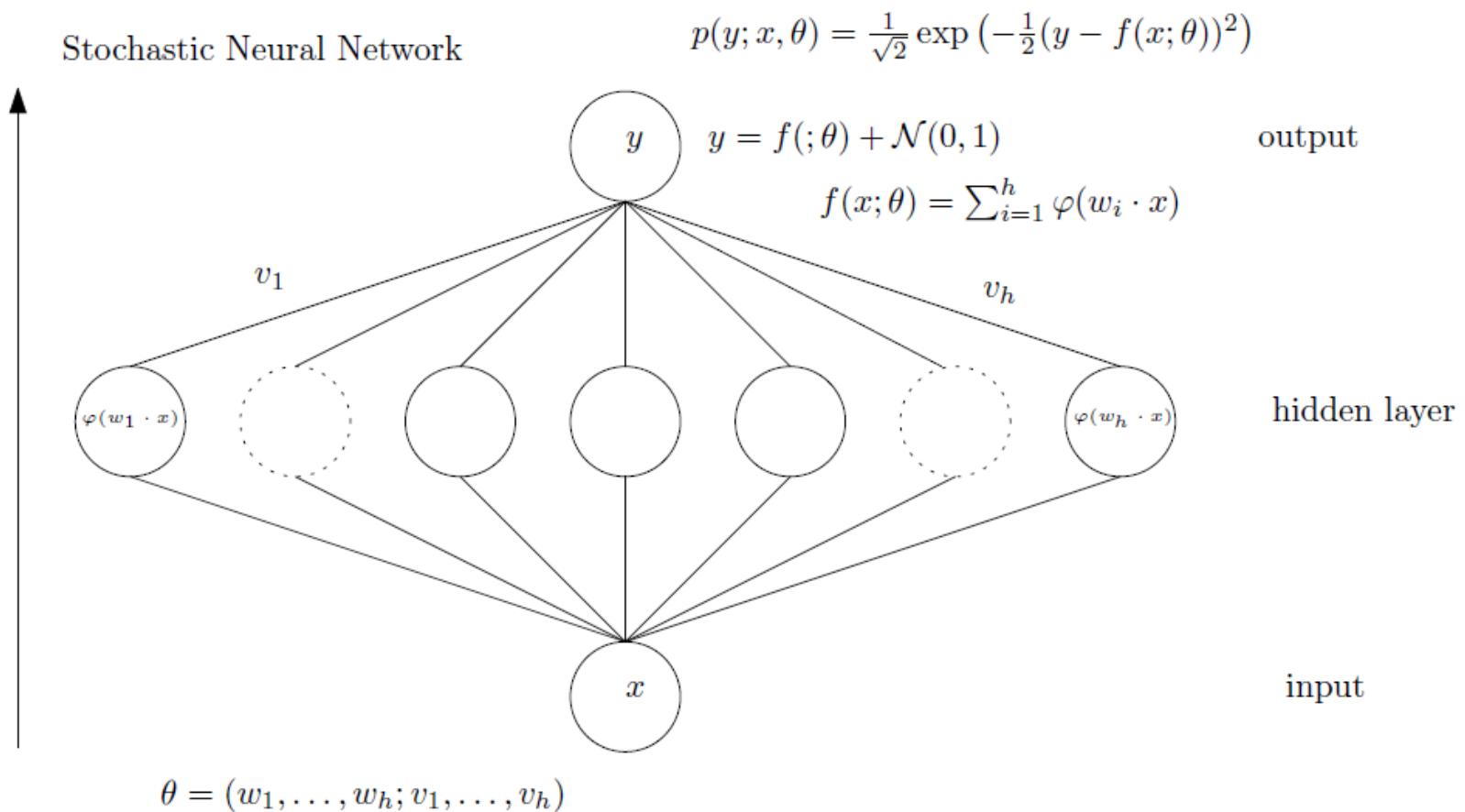
$$g_{ij}(\xi) = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ell(\xi)] = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ln p_\xi].$$

- Positive semi-definite FIM:
- Positive-definite FIM for **regular** models (identifiable)
- FIM **is invariant** under reparametrizations of the sample space.
- **Covariant** under reparameterization (later, a 2-covariant tensor metric)

Regular versus non-regular models

Regular models: 1-to-1 correspondence of parameters with distributions

Hierarchical models are usually non-regulars (eg., mixtures, multilayer perceptron)



Sufficient statistics

- A **statistic** is a function of a random vector (e.g., mean, variance)
- A **sufficient statistic** collect and concentrate from a random sample all necessary information for recovering/estimating the parameters. Statistical lossless compression
- Definition: conditional distribution of X given t does not depend on θ

$$\Pr(x|\theta) = \Pr(x|t)$$

- **Fisher-Neyman factorization theorem:** Statistic $t(x)$ sufficient iff the density can be decomposed as:

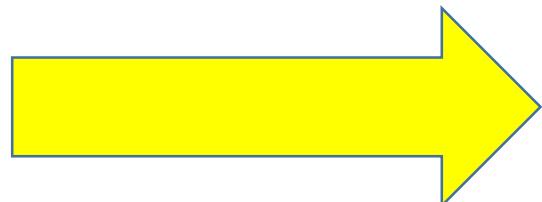
$$p(x; \lambda) = a(x)b_\lambda(t(x))$$

Example of sufficient statistics:

Fisher-Neyman factorization: $p(x; \lambda) = a(x)b_\lambda(t(x))$

For Poisson distributions:

$$p(x_1, \dots, x_n | \lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \underbrace{\prod_{i=1}^n \frac{1}{x_i}}_{a(x)} \underbrace{e^{-n\lambda} \lambda^{\sum x_i}}_{b(\sum x_i, \lambda)}$$



$\sum_{i=1}^n x_i$ is a sufficient statistic for λ .

Natural exponential families (NEF)

- Consider a **positive measure** μ
- An **exponential family** is a parametric family of densities that write as

$$p(x; \theta) = \exp(\theta x - F(\theta))$$

where F is real-analytic, strictly convex and differentiable:

$$F(\theta) = \log \int \exp(\theta x) d\mu(x)$$


**Log-Laplace
transform**

Natural parameter space $\Theta = \left\{ \theta : \int \exp(\theta x) d\mu(x) < \infty \right\}$

F : **Log-normalizer** (partition function, cumulant function, etc.)

Exponential families

- Consider a (sufficient) statistic $t(x)$
- Consider an additional carrier measure $k(x)$
- Consider an inner product between $t(x)$ and θ
(usual scalar/dot product)

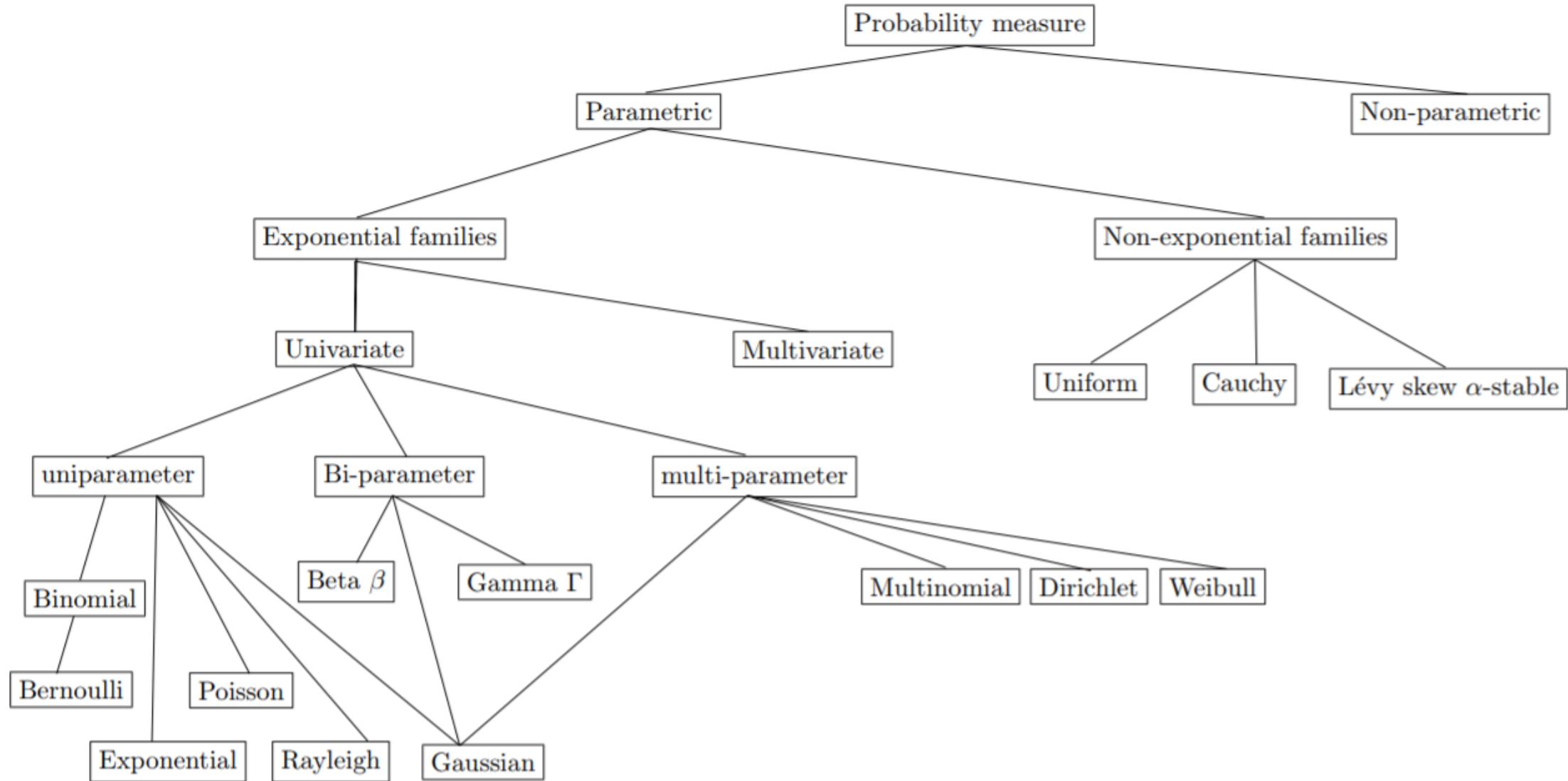
$$p_\theta(x) = \exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))$$

$$E[t(X)] = \nabla F(\theta)$$

Properties: $\text{Cov}[t(X)] = \nabla^2 F(\theta) = I(\theta)$

Exponential families have finite moments of any order

Many common distributions are exponential families



Maximum likelihood estimator for exponential families

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^n p_F(x_i; \theta).$$

Average log-likelihood: $\bar{l}(\theta; x_1, \dots, x_n) = \langle \theta, \sum_{i=1}^n t(x_i) \rangle - F(\theta) + \sum_{i=1}^n k(x_i)$

MLE equation

$$\nabla F(\theta) = \sum_{i=1}^n t(x_i)$$

$$\operatorname{var}(\hat{\theta}) \geq I^{-1}(\theta)$$

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

$$p_{\theta}(x) = \exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))$$

$$[I(\theta)]_{ij} = -E_{\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_{\theta}(x) \right]$$

Regular EFs and steepness of exponential families

- An exponential family is **regular** when the natural parameter space is open $\Theta = \text{int}(\Theta)$
- Closed convex hull of $\{t(x)\}$: $\mathcal{C} = \overline{\text{co}(\mathcal{S})}$
- Map $\eta(\theta) = E_\theta[t] = \nabla F(\theta)$ is **one-to-one**
- Consider the **expectation/moment parameter space**: $H : \{\eta(\theta) : \theta \in \Theta\}$
- Family is **steep** if $H = \text{int}(\mathcal{C})$
- MLE **exists** and is **unique** for regular and steep EFs when $\bar{t} = \sum_{i=1}^n t(x_i) \in \mathcal{C}$

$$\hat{\theta} = (\nabla F)^{-1} \left(\frac{1}{n} \sum_{i=1}^n t(x_i) \right)$$

Example of non-steep family: Singly-truncated Gaussian family

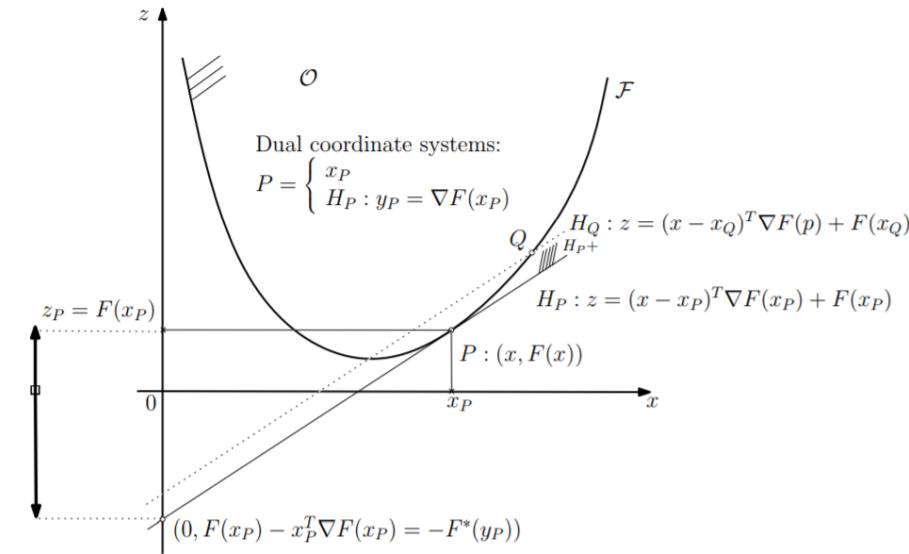
Moment/expectation parameterization

- For a regular EF density, let $\eta = \nabla F(\theta)$

- denote the **dual parameterization**

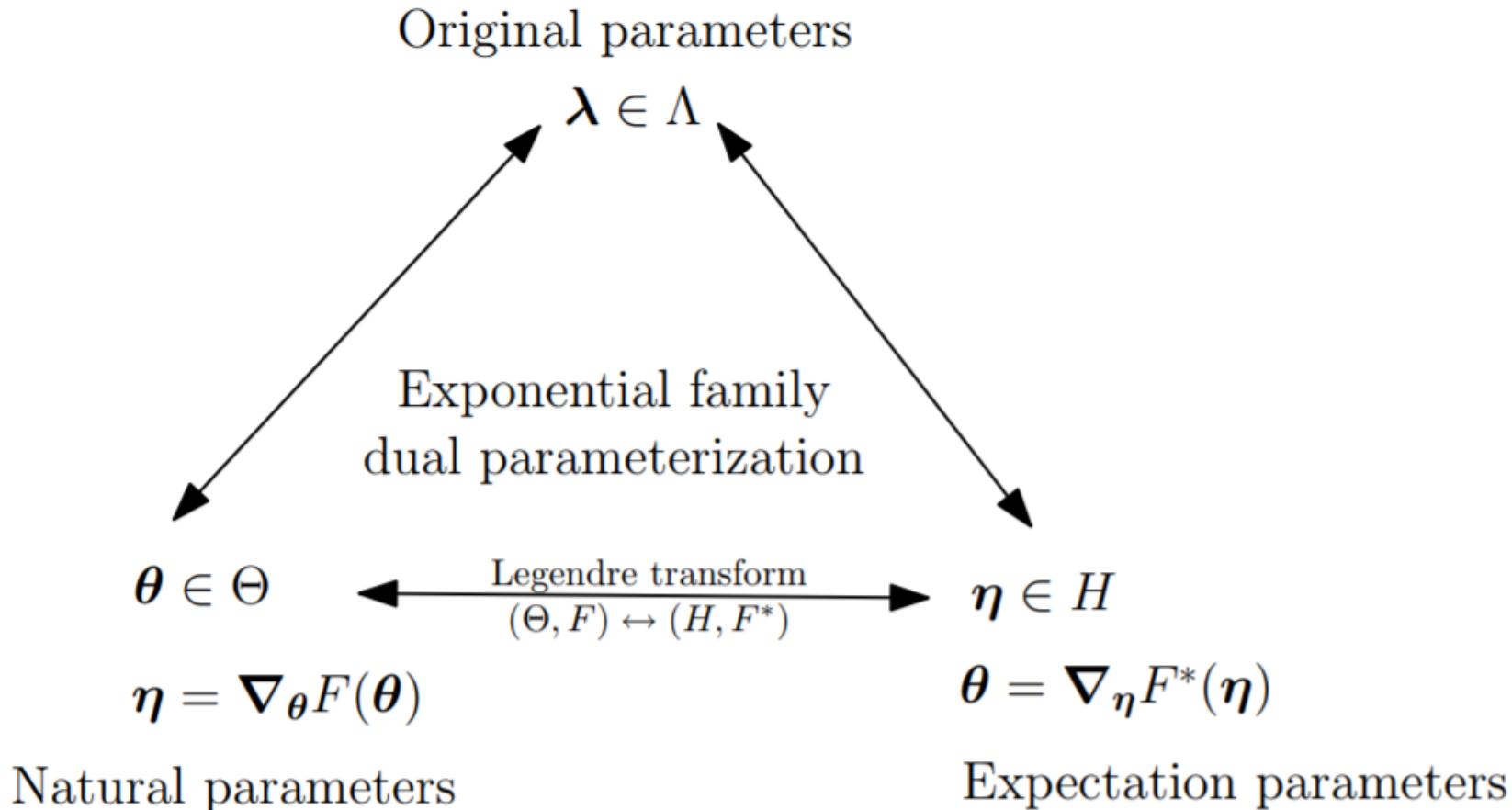
- Related to the **Legendre-Fenchel convex conjugate**:

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



- **Moreau biconjugate theorem:** when F is proper, lower semi-continuous, and convex function: $(F^*)^* = F$

Dual parameterization of exponential families



Legendre-Fenchel conjugate



- We have $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$
- The **convex conjugate** is defined by

$$F^*(\eta) = (\nabla F)^{-1}(\eta)^\top \eta - F((\nabla F)^{-1}(\eta))$$

- Crouzeix identity for convex conjugates

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

The identity matrix

Crouzeix, J.P. A Relationship Between The Second Derivatives of a Convex Function and of Its Conjugate. Math. Program. 1977, 3, 364–365.

Convex conjugates at the heart of Bregman manifolds

- Young's inequality states that

$$F(\theta) + F^*(\theta) \geq \theta^\top \eta$$

- It yields the Fenchel-Young divergence

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

.... that is equivalent to a Bregman divergence

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

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

| | |
|----------------------------------|---|
| PDF expression | $f(x; p) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$ |
| Kullback-Leibler divergence | $D_{\text{KL}}(f_1 \ f_2) = \log\left(\frac{1-p_1}{1-p_2}\right) - p_1 \log\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)$ |
| MLE | $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$ |
| Source parameters | $\Lambda = p \in [0, 1]$ |
| Natural parameters | $\Theta = \theta \in \mathbb{R}^+$ |
| Expectation parameters | $\mathbf{H} = \eta \in [0, 1]$ |
| $\Lambda \rightarrow \Theta$ | $\Theta = \log\left(\frac{p}{1-p}\right)$ |
| $\Theta \rightarrow \Lambda$ | $\Lambda = \frac{\exp \theta}{1+\exp \theta}$ |
| $\Lambda \rightarrow \mathbf{H}$ | $\mathbf{H} = p$ |
| $\mathbf{H} \rightarrow \Lambda$ | $\Lambda = \eta$ |
| $\Theta \rightarrow \mathbf{H}$ | $\mathbf{H} = \nabla F(\Theta)$ |
| $\mathbf{H} \rightarrow \Theta$ | $\Theta = \nabla G(\mathbf{H})$ |
| Log normalizer | $F(\Theta) = \log(1 + \exp \theta)$ |
| Gradient log normalizer | $\nabla F(\Theta) = \frac{\exp \theta}{1+\exp \theta}$ |
| G | $G(\mathbf{H}) = \log\left(\frac{\eta}{1-\eta}\right) \eta - \log\left(\frac{1}{1-\eta}\right) + C$ |
| Gradient G | $\nabla G(\mathbf{H}) = \log\left(\frac{\eta}{1-\eta}\right)$ |
| Sufficient statistics | $t(x) = x$ |
| Carrier measure | $k(x) = 0$ |

Bernoulli family Order 1

| | |
|----------------------------------|--|
| PDF expression | $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$ |
| Kullback-Leibler divergence | $D_{\text{KL}}(f_P \ f_Q) = \frac{1}{2} \left(2 \log \frac{\sigma_Q}{\sigma_P} + \frac{\sigma_P^2}{\sigma_Q^2} + \frac{(\mu_Q - \mu_P)^2}{\sigma_Q^2} - 1 \right)$ |
| MLE | $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}$ |
| Source parameters | $\Lambda = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ |
| Natural parameters | $\Theta = (\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}^-$ |
| Expectation parameters | $\mathbf{H} = (\eta_1, \eta_2) \in \mathbb{R} \times \mathbb{R}^+$ |
| $\Lambda \rightarrow \Theta$ | $\Theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$ |
| $\Theta \rightarrow \Lambda$ | $\Lambda = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} \right)$ |
| $\Lambda \rightarrow \mathbf{H}$ | $\mathbf{H} = (\mu, \sigma^2 + \mu^2)$ |
| $\mathbf{H} \rightarrow \Lambda$ | $\Lambda = (\eta_1, \eta_2 - \eta_1^2)$ |
| $\Theta \rightarrow \mathbf{H}$ | $\mathbf{H} = \nabla F(\Theta)$ |
| $\mathbf{H} \rightarrow \Theta$ | $\Theta = \nabla G(\mathbf{H})$ |
| Log normalizer | $F(\Theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\theta_2}\right)$ |
| Gradient log normalizer | $\nabla F(\Theta) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} \right)$ |
| G | $G(\mathbf{H}) = -\frac{1}{2} \log(\eta_1^2 - \eta_2) + C$ |
| Gradient G | $\nabla G(\mathbf{H}) = \left(-\frac{\eta_1}{\eta_1^2 - \eta_2}, \frac{1}{2(\eta_1^2 - \eta_2)} \right)$ |
| Sufficient statistics | $t(x) = (x, x^2)$ |
| Carrier measure | $k(x) = 0$ |

Univariate Gaussian family Order 2

| | |
|----------------------------------|--|
| PDF expression | $f(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ for $x \in \mathbb{N}^+$ |
| Kullback-Leibler divergence | $D_{\text{KL}}(f_P \ f_Q) = \lambda_Q - \lambda_P \left(1 + \log \left(\frac{\lambda_Q}{\lambda_P} \right) \right)$ |
| MLE | $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i$ |
| Source parameters | $\Lambda = \lambda \in \mathbb{R}^+$ |
| Natural parameters | $\Theta = \theta \in \mathbb{R}$ |
| Expectation parameters | $\mathbf{H} = \eta \in \mathbb{R}^+$ |
| $\Lambda \rightarrow \Theta$ | $\Theta = \log \lambda$ |
| $\Theta \rightarrow \Lambda$ | $\Lambda = \exp \theta$ |
| $\Lambda \rightarrow \mathbf{H}$ | $\mathbf{H} = \lambda$ |
| $\mathbf{H} \rightarrow \Lambda$ | $\Lambda = \eta$ |
| $\Theta \rightarrow \mathbf{H}$ | $\mathbf{H} = \nabla F(\Theta)$ |
| $\mathbf{H} \rightarrow \Theta$ | $\Theta = \nabla G(\mathbf{H})$ |
| Log normalizer | $F(\Theta) = \exp \theta$ |
| Gradient log normalizer | $\nabla F(\Theta) = \exp \theta$ |
| G | $G(\mathbf{H}) = \eta \log \eta - \eta + C$ |
| Gradient G | $\nabla G(\mathbf{H}) = \log \eta$ |
| Sufficient statistics | $t(x) = x$ |
| Carrier measure | $k(x) = -\log(x!)$ |

Poisson family Order 1

| | |
|----------------------------------|---|
| PDF expression | $f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \Sigma ^{1/2}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right)$ for $x \in \mathbb{R}^d$ |
| Kullback-Leibler divergence | $D_{\text{KL}}(f_P \ f_Q) = \frac{1}{2} \left(\log \left(\frac{\det \Sigma_Q}{\det \Sigma_P} \right) + \text{tr} \left(\Sigma_Q^{-1} \Sigma_P \right) \right) + \frac{1}{2} \left((\mu_Q - \mu_P)^T \Sigma_Q^{-1} (\mu_Q - \mu_P) - d \right)$ |
| MLE | $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$ |
| Source parameters | $\Lambda = (\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$ |
| Natural parameters | $\Theta = (\theta, \Theta)$ |
| Expectation parameters | $\mathbf{H} = (\eta, H)$ |
| $\Lambda \rightarrow \Theta$ | $\Theta = (\Sigma^{-1} \mu, \frac{1}{2} \Sigma^{-1})$ |
| $\Theta \rightarrow \Lambda$ | $\Lambda = (\frac{1}{2} \Theta^{-1} \theta, \frac{1}{2} \Theta^{-1})$ |
| $\Lambda \rightarrow \mathbf{H}$ | $\mathbf{H} = (\mu, -(\Sigma + \mu \mu^T))$ |
| $\mathbf{H} \rightarrow \Lambda$ | $\Lambda = (\eta, -(H + \eta \eta^T))$ |
| $\Theta \rightarrow \mathbf{H}$ | $\mathbf{H} = \nabla F(\Theta)$ |
| $\mathbf{H} \rightarrow \Theta$ | $\Theta = \nabla G(\mathbf{H})$ |
| Log normalizer | $F(\Theta) = \frac{1}{4} \text{tr}(\Theta^{-1} \theta \theta^T) - \frac{1}{2} \log \det \Theta + \frac{d}{2} \log \pi$ |
| Gradient log normalizer | $\nabla F(\Theta) = (\frac{1}{2} \Theta^{-1} \theta, -\frac{1}{2} \Theta^{-1} - \frac{1}{4} (\Theta^{-1} \theta)(\Theta^{-1} \theta)^T)$ |
| G | $G(\mathbf{H}) = -\frac{1}{2} \log (1 + \eta^T H^{-1} \eta) - \frac{1}{2} \log \det(-H) - \frac{d}{2} \log(2\pi e)$ |
| Gradient G | $\nabla G(\mathbf{H}) = (-(H + \eta \eta^T)^{-1} \eta, -\frac{1}{2} (H + \eta \eta^T)^{-1})$ |
| Sufficient statistics | $t(x) = (x, -xx^T)$ |
| Carrier measure | $k(x) = 0$ |

Multivariate Gaussian family Order

$$\frac{d(d+3)}{2}$$

Compound parameter:
 Vector part
 Matrix part

Inner product defined by:

$$\langle \theta, \theta' \rangle = \theta_v^\top \theta'_v + \text{tr} \left(\theta_M'{}^\top \theta_M \right)$$

Statistical exponential families: A digest with flash cards. arXiv:0911.4863 (2009)

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means, Entropy 2019

Summary

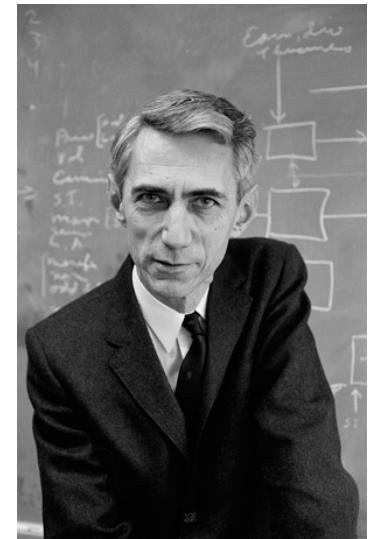
- **Probability measure** bypasses the Banach-Tarsky paradox by fixing a sigma-algebra of measurable events, and unifies discrete/continuous random variables as measurable functions
- **Fisher information** (FI) measures the sensitivity of the log-likelihood (curvature), invariant to reparametrization of sample space, covariant to reparameterization of parameter space
- **Cramer-Rao bound** provides a lower bound on the variance of unbiased estimator (non-asymptotic) based on the inverse of FI
- MLE has asymptotic normality for regular models
- Sufficient statistics is statistical lossless compression of random vectors
- Exponential families: Dual parameterizations via **Legendre-Fenchel conjugation**, MLE and FIM in closed-form

Information Theory

Frank Nielsen



Sony CSL



Background

Outline

- Entropy and differential entropy
- Relative entropy known as the Kullback-Leibler divergence
- Maximum entropy principle
 - (MaxEnt distributions = exponential families)
- Bounding the differential entropy of statistical mixtures
- KL of location-scale families

Entropy

- Quantifies the **uncertainty** of a discrete random variable X

$$H(X) = \sum_{i=1} p_i \log \frac{1}{p_i} = - \sum_{i=1} p_i \log p_i$$

$p_i = P(X = x_i)$

- Can be derived axiomatically from **Kinchin's axioms**

Theorem 2.1. Let the function $\mathcal{S}_n : \Delta_n \rightarrow \mathbb{R}^+$ satisfy the following Shannon-Khinchin axioms, for all $n \in \mathbb{N}$, $n > 1$:

[SA1] \mathcal{S}_n is continuous in Δ_n ;

[SA2] \mathcal{S}_n takes its largest value for the uniform distribution, $U_n = (1/n, \dots, 1/n) \in \Delta_n$, i.e. $\mathcal{S}_n(P) \leq \mathcal{S}_n(U_n)$, for any $P \in \Delta_n$;

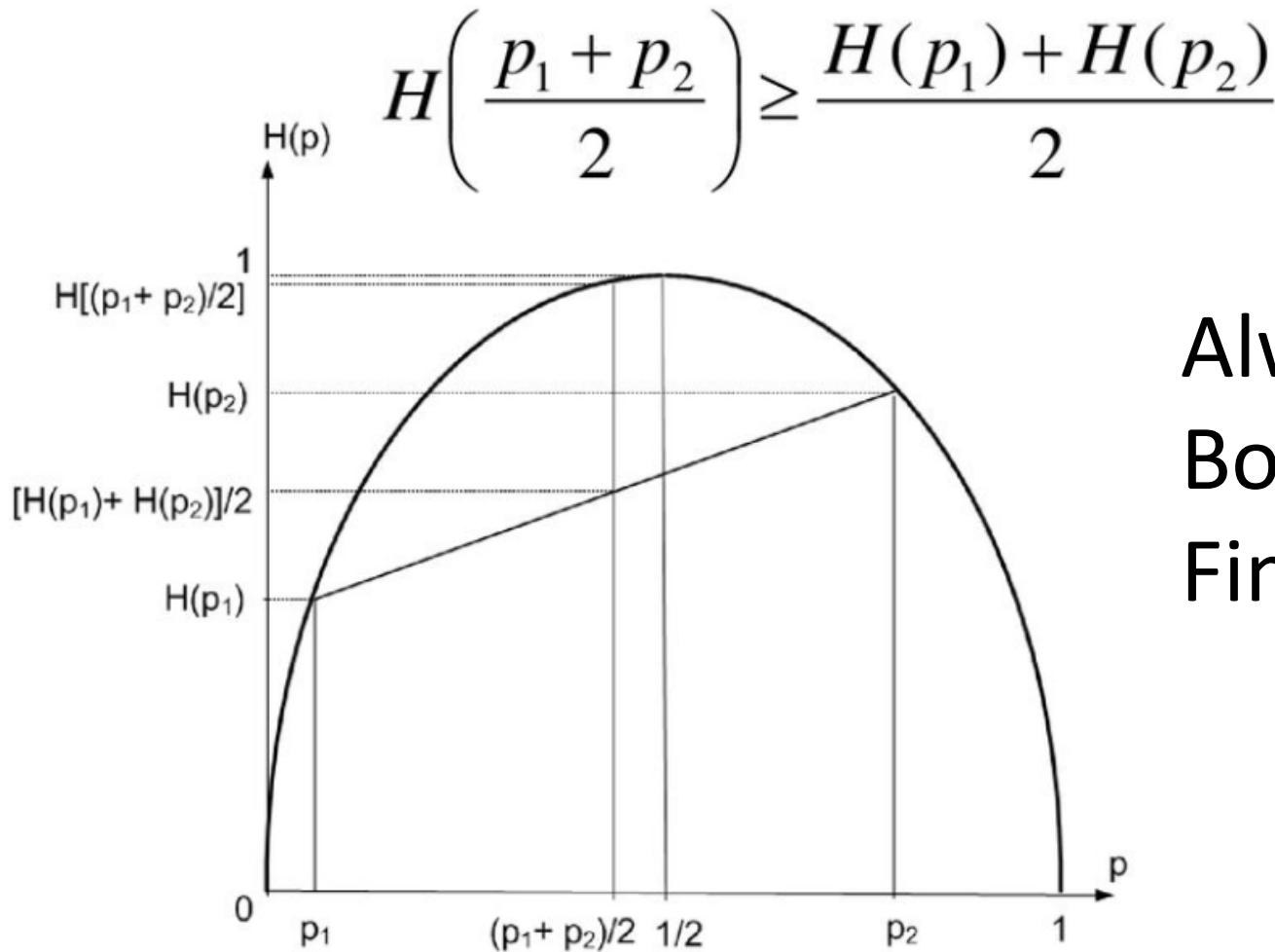
[SA3] \mathcal{S}_n is expandable: $\mathcal{S}_{n+1}(p_1, p_2, \dots, p_n, 0) = \mathcal{S}_n(p_1, p_2, \dots, p_n)$ for all $(p_1, \dots, p_n) \in \Delta_n$;

[SA4] Let $P = (p_1, \dots, p_n) \in \Delta_n$, $PQ = (r_{11}, r_{12}, \dots, r_{nm}) \in \Delta_{nm}$, $n, m \in \mathbb{N}$, $n, m > 1$ such that $p_i = \sum_{j=1}^m r_{ij}$, and $Q|k = (q_{1|k}, \dots, q_{m|k}) \in \Delta_m$, where $q_{i|k} = r_{ik}/p_k$. Then,

$$\mathcal{S}_{nm}(PQ) = \mathcal{S}_n(P) + \mathcal{S}_m(Q|P), \quad \text{where} \quad \mathcal{S}_m(Q|P) = \sum_k p_k \cdot \mathcal{S}_m(Q|k).$$

Then, the function \mathcal{S}_n is the Shannon entropy

Shannon's entropy is a concave function



Always positive
Bounded by $\log(n)$
Finite for fixed-size alphabets

The negentropy is called **Shannon information** (a convex function)

Differential entropy is different from discrete entropy

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx$$

- Can be **negative**
- Can be **infinite** when the integral diverges $h(X) = +\infty$

$X \sim p(x) = \frac{\log(2)}{x \log^2 x}$ for $x > 2$, with support $\mathcal{X} = (2, \infty)$

- For **Dirac distribution**, the entropy is: $X \sim p(x) = \delta(x), h(X) = -\infty$

For Gaussian distributions, the entropy is **independent of location**:

$$h(X) = \frac{1}{2} \log(2\pi e \sigma^2), \quad X \sim N(\mu, \sigma)$$

Entropy of a probability measure

- Random variable (measurable function)

$$X \sim P \ll \mu$$

$$H(X) = - \int_{\mathcal{X}} \log \frac{dP}{d\mu} dP$$

$$H(X) = - \int_{\mathcal{X}} p(x) \log p(x) d\mu(x), \quad p = \frac{dP}{d\mu}$$

Unifies discrete entropy (counting measure) and differential entropy (Lebesgue measure)

Relative entropy: Kullback-Leibler divergence

$$\text{KL}(P : Q) = \int p(x) \log \frac{p(x)}{q(x)} d\mu(x) \quad P, Q \ll \mu, \quad p = \frac{dP}{d\mu}, \quad \frac{dQ}{d\mu}$$

$$\text{KL}(P : Q) = H^\times(P : Q) - H(P)$$

Cross-entropy: $H^\times(P : Q) = - \int p \log q d\mu \quad H(P) = H^\times(P : P)$

KL = Relative entropy with respect to a reference distribution P

Not a metric distance because asymmetric and failing the triangle inequality

- $\text{KL}(P : Q) \geq 0$ (Gibb's inequality) and KL may be infinite:

$$p(x) = \frac{1}{\pi(1+x^2)} = \text{Cauchy distribution}$$

$$q(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) = \text{standard normal distribution}$$

$\text{KL}(p : q) = +\infty$ diverges while $\text{KL}(q : p) < \infty$ converges.

Entropy for discrete/continuous exponential families

$$\exp\left(\sum_{i=1}^D t_i(x)\theta_i - F(\theta) + k(x)\right)$$

$$p(x; \theta) = \exp(\langle \theta, t(x) \rangle - F(\theta))$$

without carrier term $k(x)$

Using natural parameter θ :

$$H(P) = H_F(\theta_p) = F(\theta_p) - \langle \theta_p, \nabla F(\theta_p) \rangle - E_P[k(x)]$$

Using expectation parameter:

$$H(P) = -F^*(\eta) - E_P[k(x)]$$

Rayleigh distribution

$p(x; \sigma^2) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ that belongs to the exponential families for the log-normalizer $F(\theta) = -\log(-2\theta)$, natural parameter $\theta = -\frac{1}{2\sigma^2}$, sufficient statistic $t(x) = x^2$, gradient $F'(\theta) = -\frac{1}{\theta}$ and carrier measure $k(x) = \log x$. Let $X \sim \text{Rayleigh}(\sigma^2)$, we have: $H(X) = 1 + \ln \frac{\sigma}{\sqrt{2}} + \frac{\gamma}{2}$, where $\gamma = 0.57721566\dots$ stands for the Euler-Mascheroni constant. This is the term related to the carrier measure $\log x$ integrated over the distribution.

Consider yet another univariate exponential family: the Poisson distribution with probability mass function $p(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$. The entropy is $\lambda(1 - \log \lambda) - E[k(x)]$ Since $k(x) = -\log x!$ (see [4]), we have:

$$-E[k(x)] = \sum_{k=0}^{\infty} p_F(x; \lambda) \log k! = e^{-\lambda} \sum \frac{\lambda^k \log k!}{k!}.$$

Kullback-Leibler divergence for exponential families

Fenchel-Young divergence for exponential families

$$\text{KL}(p_{\theta_1} : p_{\theta_2}) = B(\theta_2 : \theta_1) = A(\theta_2 : \eta_1) = A^*(\eta_1 : \theta_2) = B^*(\eta_1 : \eta_2)$$

Fenchel-Young divergence (on mixed parameters):

$$A(\theta_2 : \eta_1) = F(\theta_2) + F^*(\eta_1) - \theta_2^\top \eta_1 \geq 0$$

Bregman divergence (on natural/expectation parameters):

$$B(\theta_2 : \theta_1) = F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F(\theta_1)$$

Jaynes' maximum entropy principle (MaxEnt)

- Jaynes's principle of **maximum ignorance**:

Underconstrained optimization problem

$$\max_p h(p) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$\sum_x p(x) t_i(x) = m_i, \quad \forall i \in \{1, \dots, D\}$$

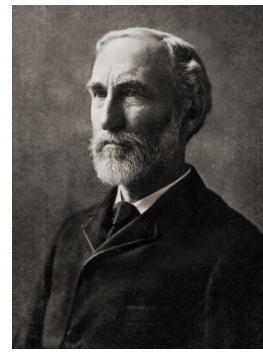
$$p(x) \geq 0, \quad \forall x \in \{1, \dots, n\}$$
$$\sum_x p(x) = 1$$



Edwin Thompson Jaynes
(1922–1998)

Maximizing a concave function subject to linear constraints (convex min optimization problem).

MaxEnt distributions (Boltzmann-Gibbs)



Constrained optimization problem:

Use **Lagrange multipliers** θ (but θ not in closed form)

Gibbs distribution, Maxwell-Boltzmann distribution in statistical mechanics:

$$p(x) = \frac{1}{Z(\theta)} \exp(\langle \theta, t(x) \rangle) q(x)$$

Gibbs distribution in statistical physics,
Titled distribution in probability

MaxEnt distributions are exponential families $\exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))$

Log-normalizer: $F(\theta) = \log Z(\theta)$ Free energy

Prior q gives the carrier measure: $q(x) = e^{k(x)}$

MaxEnt with Kullback-Leibler divergence

$$\min_p \text{KL}(p : q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\sum_x p(x)t_i(x) = m_i, \quad \forall i \in \{1, \dots, D\}$$

$$p(x) \geq 0, \quad \forall x \in \{1, \dots, n\}$$

$$\sum_x p(x) = 1$$

Maximum entropy distribution is the uniform prior: $q(x) = \frac{1}{n}$

Example: Fixed mean and fixed variance MaxEnt distribution

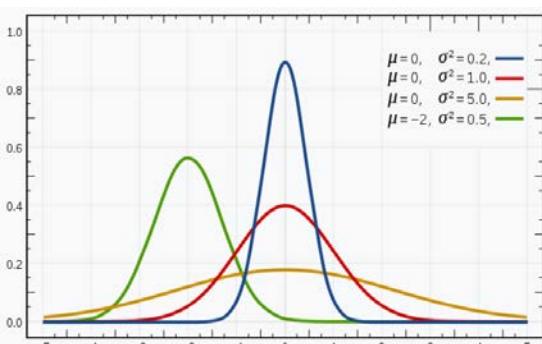
- Find the MaxEnt distributions with support the full real line and the first two moments prescribed

$$E[X] = m_1$$

$$E[X^2] = m_2$$

$$t(x) = (x, x^2)$$

$$p(x) \propto \exp(\theta_1 x + \theta_2 x^2)$$



$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



MLE has a right-sided KLD minimization

Recall that MaxEnt is KL left-sided minimization:

$$\min_p \text{KL}(p : q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- Empirical distribution

$$\begin{aligned} p_e(x) &= \frac{1}{m} \sum_{i=1}^m \delta_{s_i}(x) & \min \quad \text{KL}(p_e(x) : \boxed{p_\theta(x)}) \\ & &= \int p_e(x) \log p_e(x) dx - \int p_e(x) \log p_\theta(x) dx \\ & &= \min -H(p_e) - \underbrace{E_{p_e} [\log p_\theta(x)]}_{\text{}} \\ & &\equiv \max \frac{1}{n} \sum \delta(x - x_i) \log p_\theta(x) \\ & &= \max \frac{1}{n} \sum_i \log p_\theta(x_i) = \text{MLE} \end{aligned}$$

Upper bounding the differential entropy of mixtures

Key idea: compute the differential entropy of an exponential family with **given sufficient statistics** in **closed form**. Since it is a MaxEnt distribution, *any other* distribution has less entropy for the same moment expectations. In particular, this applies to statistical mixtures.

$$H(X) = \int_{\mathcal{X}} p(x) \log \frac{1}{p(x)} dx = - \int_{\mathcal{X}} p(x) \log p(x) dx$$

$$H(p(x; \theta)) = -F^*(\eta(\theta))$$

Absolute Monomial Exponential Family (AMEF): $p_I(x; \theta) = \exp \left(\theta |x|^I - F_I(\theta) \right)$

with log-normalizer

$$F_I(\theta) = \log 2 + \log \Gamma \left(\frac{1}{I} \right) - \log I - \frac{1}{I} \log(-\theta)$$

$$\Gamma(u) = \int_0^\infty x^{u-1} \exp(-x) dx$$

$$\Gamma(n) = (n-1)! \text{ for } n \in \mathbb{N}$$

Upper bounding the differential entropy of mixtures

$$p_I(x; \theta) = \exp\left(\theta|x|^I - F_I(\theta)\right)$$

$$\begin{aligned} H(p(x; \theta)) &= -F^*(\eta(\theta)) & H_I(\eta) &= \log 2 + \log \Gamma\left(\frac{1}{I}\right) - \log I + \frac{1}{I}(1 + \log I + \log \eta) \\ H_I(\theta) &= \log 2 + \log \Gamma\left(\frac{1}{I}\right) - \log I + \frac{1}{I}(1 - \log(-\theta)). \end{aligned}$$

Density of a Gaussian Mixture Model (GMM): $X \sim \sum_{c=1}^k w_c p(x; \mu_c, \sigma_c)$

$$H(X) \leq U_1(X) \quad U_1(X) = \log \left(2e \left(\sum_{c=1}^k w_c \left(\mu_c \left(1 - 2\Phi\left(-\frac{\mu_c}{\sigma_c}\right) \right) + \sigma_c \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2} \left(\frac{\mu_c}{\sigma_c}\right)^2\right) \right) \right) \right)$$

(MaxEnt distribution is Laplacian distribution)

$$H(X) \leq U_2(X) = \frac{1}{2} \log \left(2\pi e \sum_{c=1}^k w_c ((\mu_c - \bar{\mu})^2 + \sigma_c^2) \right) \quad \bar{\mu} = \sum_{c=1}^k w_c \mu_c$$

(MaxEnt distribution is Gaussian distribution)

A series of upper bounds for h (GMMs)

Zero-centered Gaussian Mixture Models:

$$H(X) \leq H_I^\eta(A_I(X)) = b_I + \frac{1}{I} \log z_I + \log \bar{\sigma}_I,$$

$$E_X[X^I] = \underbrace{2^{\frac{I}{2}} \frac{\Gamma(\frac{1+I}{2})}{\sqrt{\pi}}}_{z_I} \left(\sum_{i=1}^k w_i \sigma_i^I \right) = A_I(X).$$

$$\bar{\sigma}_I: I\text{-th power mean: } \bar{\sigma}_I = \left(\sum_{i=1}^k w_i \sigma_i^I \right)^{\frac{1}{I}}$$

Computing central/non-central geometric moments of Gaussians and GMMs

| Even I | $A_I = E[X ^I] = E[X^I] = \sum_{i=0}^{\lfloor \frac{I}{2} \rfloor} \binom{I}{2i} (2i-1)!! \mu^{I-2i} \sigma^{2i}$ |
|----------|--|
| 2 | $\mu^2 + \sigma^2$ |
| 4 | $\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$ |
| 6 | $\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6$ |
| 8 | $\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$ |
| 10 | $\mu^{10} + 45\mu^8\sigma^2 + 630\mu^6\sigma^4 + 3150\mu^4\sigma^6 + 4725\mu^2\sigma^8 + 945\sigma^{10}$ |

| Odd I | $A_I = E[X ^I] = C_I(\mu, \sigma) \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) + D_I(\mu, \sigma) \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |
|---------|--|
| 1 | $\sigma \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) + \mu \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |
| 3 | $(2\sigma^3 + \mu^2\sigma) \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) + (\mu^3 + 3\mu\sigma^2) \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |
| 5 | $(8\sigma^5 + 9\mu^2\sigma^3 + \mu^4\sigma) \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) + (\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4) \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |
| 7 | $(48\sigma^7 + 87\mu^2\sigma^5 + 20\mu^4\sigma^3 + \mu^6\sigma) \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) +$ $(\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6) \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |
| 9 | $(384\sigma^9 + 975\mu^2\sigma^7 + 345\mu^4\sigma^5 + 35\mu^6\sigma^3 + \mu^8\sigma) \sqrt{\frac{2}{\pi}} \exp(-\frac{\mu^2}{2\sigma^2}) +$ $(\mu^9 + 36\mu^7\sigma^2 + 378\mu^5\sigma^4 + 1260\mu^3\sigma^6 + 945\sigma^8) \operatorname{erf}(\frac{\mu}{\sqrt{2}\sigma})$ |

Computing the Kullback-Leibler divergence...

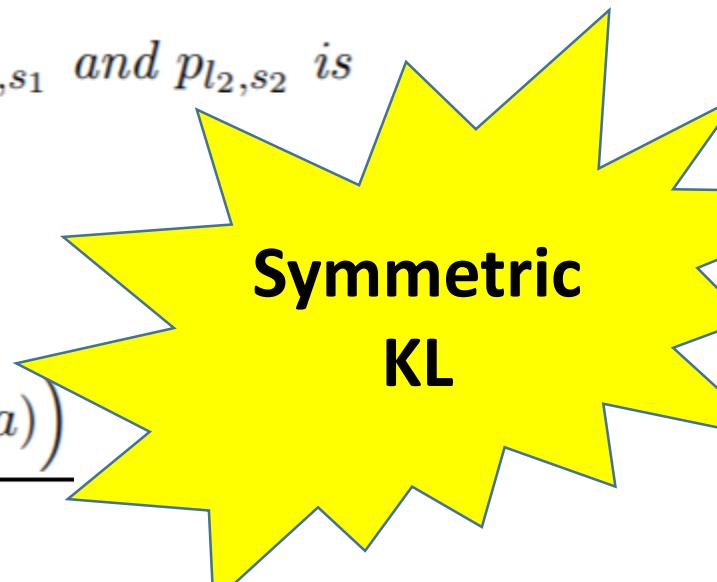
- In theory, Risch semi-algorithm reports whether a definite integral has a closed-form or not. Note that the KLD can also diverge.
- Symbolic calculations
- For example: Cauchy location-scale families . $p_{l,s}(x) = \frac{dP_{l,s}}{d\mu}(x) = \frac{s}{\pi(s^2 + (x - l)^2)}$

Theorem . *The Kullback-Leibler divergence between Cauchy density p_{l_1,s_1} and p_{l_2,s_2} is*

$$\text{KL}(p_{l_1,s_1} : p_{l_2,s_2}) = \log \frac{(s_1 + s_2)^2 + (l_1 - l_2)^2}{4s_1s_2}.$$

$$A(a, b, c; d, e, f) = \frac{2\pi \left(\log(2af - be + 2cd + \sqrt{4ac - b^2}\sqrt{4df - e^2}) - \log(2a) \right)}{\sqrt{4ac - b^2}}$$

A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions, arXiv:1905.10965



Kullback-Leibler divergence: Location-scale families

$$\mathcal{F}_1 = \left\{ p_{l_1, s_1}(x) = \frac{1}{s_1} p\left(\frac{x - l_1}{s_1}\right) : (l_1, s_1) \in \mathbb{H} \right\} \quad \mathcal{F}_2 = \left\{ q_{l_2, s_2}(x) = \frac{1}{s_2} q\left(\frac{x - l_2}{s_2}\right) : (l_2, s_2) \in \mathbb{H} \right\}$$

Location-scale group: $\mathbb{H} = \{(l, s) : l \in \mathbb{R} \times \mathbb{R}_{++}\}$

Property (Location-scale Kullback-Leibler divergence). *We have*

$$\begin{aligned} \text{KL}(p_{l_1, s_1} : q_{l_2, s_2}) &= h^\times \left(p : q_{\frac{l_2 - l_1}{s_1}, \frac{s_2}{s_1}} \right) - h(p) = \text{KL} \left(p : q_{\frac{l_2 - l_1}{s_1}, \frac{s_2}{s_1}} \right), \\ &= h^\times \left(p_{\frac{l_1 - l_2}{s_1}, \frac{s_1}{s_2}} : q \right) - h(p) + \log \frac{s_2}{s_1} = \text{KL} \left(p_{\frac{l_1 - l_2}{s_2}, \frac{s_1}{s_2}} : q \right). \end{aligned}$$

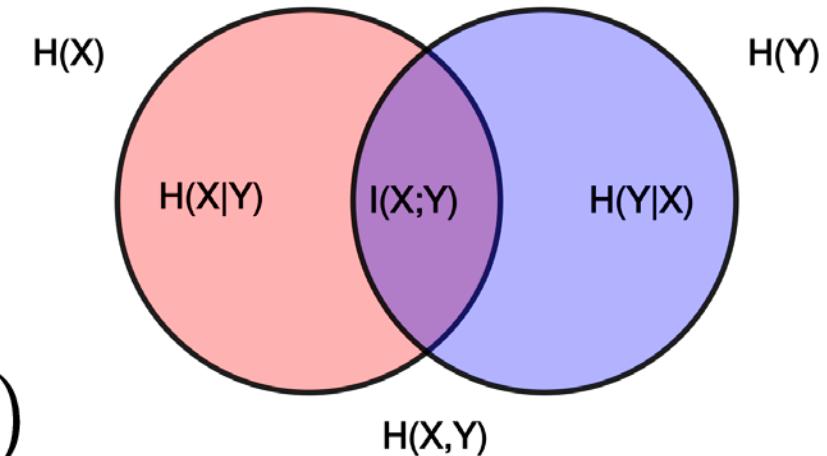
Interesting properties for the KL minimization:

$$\begin{aligned} \text{KL}(p_{l_1, s_1} : Q) &:= \min_{(l_2, s_2) \in \mathbb{H}} \text{KL}(p_{l_1, s_1} : q_{l_2, s_2}) \\ &\equiv \min_{(l_2, s_2) \in \mathbb{H}} \text{KL}(p : q_{\frac{l_2 - l_1}{s_1}, \frac{s_2}{s_1}}) \\ &\equiv \min_{(l, s) \in \mathbb{H}} \text{KL}(p : q_{l, s}) := \text{KL}(p : Q) \end{aligned}$$

Mutual information (MI)

- Consider two random variables X and Y.
- There are independent iff

$$p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$$



- Amount of mutual information quantified as the KL divergence between the joint distribution and the product of distributions

$$I(X; Y) = \text{KL} \left(P_{(X,Y)} \parallel P_X P_Y \right)$$

$$I(X; Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p_{(X,Y)}(x, y) \log \left(\frac{p_{(X,Y)}(x, y)}{p_X(x)p_Y(y)} \right) dx dy$$

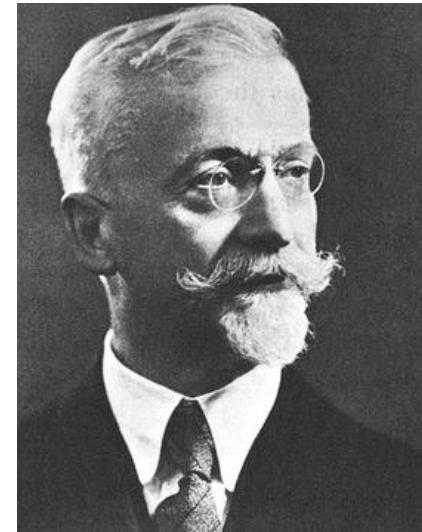
Not a metric distance but **symmetric distance between random variables**

Elements of differential geometry

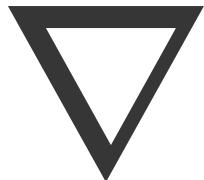
Frank Nielsen



Sony CSL

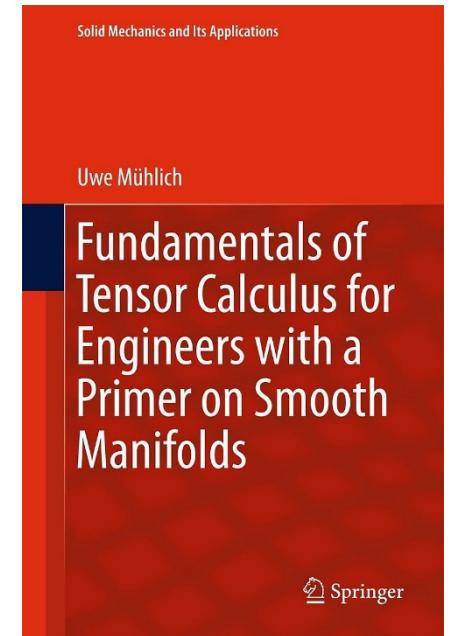


Elie Cartan
1869-1951



Outline

- Vector space and dual covector space
- Inner product space and metric tensor
(contravariant and covariant coordinates)
- Tensor fields
- Affine connection
- Riemannian metric connection



Finite dimensional real vector spaces

A *real vector space* is a set X with a **special element 0**, and **three operations** :

- **Addition:** Given two elements x, y in X , one can form the **sum** $x+y$, which is also an element of X .
- **Inverse:** Given an element x in X , one can form the inverse $-x$, which is also an element of X .
- **Scalar multiplication:** Given an element x in X and a real number c , one can form the product cx , which is also an element of X .

Operations must satisfy the following axioms:

- **Additive axioms.** For every x, y, z in X , we have
 - $x+y = y+x$.
 - $(x+y)+z = x+(y+z)$.
 - $0+x = x+0 = x$.
 - $(-x) + x = x + (-x) = 0$.
- **Multiplicative axioms.** For every x in X and real numbers c, d , we have
 - $0x = 0$
 - $1x = x$
 - $(cd)x = c(dx)$
- **Distributive axioms.** For every x, y in X and real numbers c, d , we have
 - $c(x+y) = cx + cy$.
 - $(c+d)x = cx + dx$.

Bases and dimension of a vector space V

- A set of D vectors $B = \{b_1, \dots, b_D\}$ is **linearly independent** iff

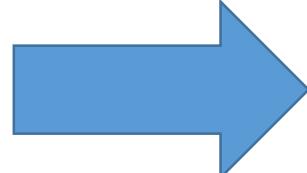
$$\sum_{i=1}^D \lambda_i b_i = 0 \quad \iff \quad \lambda_i = 0, \forall i \in [D]$$

- A **basis** is a set of maximal linearly independent vectors (wrt set inclusion)
- The **dimension** of the vector space is the cardinality of any basis (finite dimensional case)
- Vector v written in a basis B using coefficients $B = \{e_1, \dots, e_d\}$

$$v_{[B]} = (v^1, \dots, v^d) \quad v = \sum_{i=1}^d v^i e_i = v^i e_i$$

Einstein
summation
convention

Dual vector space V^* : Vector space of covectors

- **Linear form:** Linear mapping $\omega : V \rightarrow \mathbb{R}$ $\underline{\omega} : V \rightarrow \mathbb{R}$
 - **Dual vector space V^* :** = vector space of real-valued linear mappings
 - Same dimension: $\dim(V) = \dim(V^*)$
 - Isomorphism $V \simeq V^*$
-
- **Dual covector basis:** We have $\omega(v) = v^i \omega(e_i)$
 - Choose covector basis which reads vector components: $\underline{e}^i(v) = v^i$
- $\omega(v) = \omega_i \underline{e}^i(v), \omega_i = \underline{\omega}(e_i)$  $\underline{e}^i(e_j) = \delta_j^i$

Pairing product

Basis in vector space

$$B = \{e_1, \dots, e_d\}$$

Basis in covector space

$$B^* = \{e^1, \dots, e^d\}$$

- By notational definition:

$$\langle \omega, v \rangle := \omega(v)$$

- Vector components:

$$v^i = \langle e^i, v \rangle$$

- Covector components

$$\omega_i = \langle \omega, e_i \rangle$$

$$e^i(e_j) = \langle e^i, e_j \rangle = \delta_j^i$$

Inner product space (notion of lengths/angles/orthogonality)

Definition (*Inner product*) A mapping

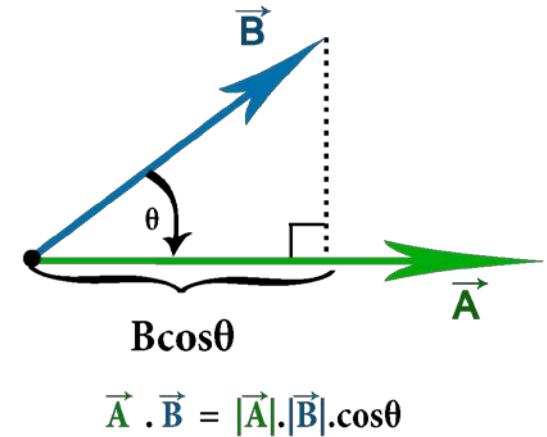
$$\cdot : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$$

$$(\mathbf{a}, \mathbf{b}) \mapsto \mathbf{a} \cdot \mathbf{b}$$

with the properties:

- (i) $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$
- (ii) $(\alpha\mathbf{a} + \beta\mathbf{b}) \cdot \mathbf{c} = \alpha\mathbf{a} \cdot \mathbf{c} + \beta\mathbf{b} \cdot \mathbf{c}$
- (iii) $\mathbf{a} \neq \mathbf{0} \Rightarrow \mathbf{a} \cdot \mathbf{a} > 0$

for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathcal{V}$ und $\alpha, \beta \in \mathbb{R}$ is called an inner product.



Orthogonality

$$v_1 \perp p_2 \Leftrightarrow \langle v_1, v_2 \rangle = 0$$

Norm and distance induced by an inner product

Definition

(*Norm*) A norm $\|\cdot\|$ on a vector space \mathcal{V} is a mapping with the properties:

- (i) $\|\alpha \mathbf{v}\| = |\alpha| \|\mathbf{v}\|$
- (ii) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
- (iii) $\|\mathbf{v}\| = 0$ implies $\mathbf{u} = \mathbf{0}$

for $\alpha \in \mathbb{R}$ and $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

Length of a vector \mathbf{v} is its norm

Distance (metric) induced by a norm:

$$D(v_1, v_2) = \|v_1 - v_2\|$$

Reciprocal basis

- Given an inner product $\langle \cdot, \cdot \rangle$, we can define a **reciprocal basis** of V

$e^j \in V$ such that $\langle e_i, e^j \rangle = \delta_i^j$

primal and reciprocal basis are **mutually orthogonal**

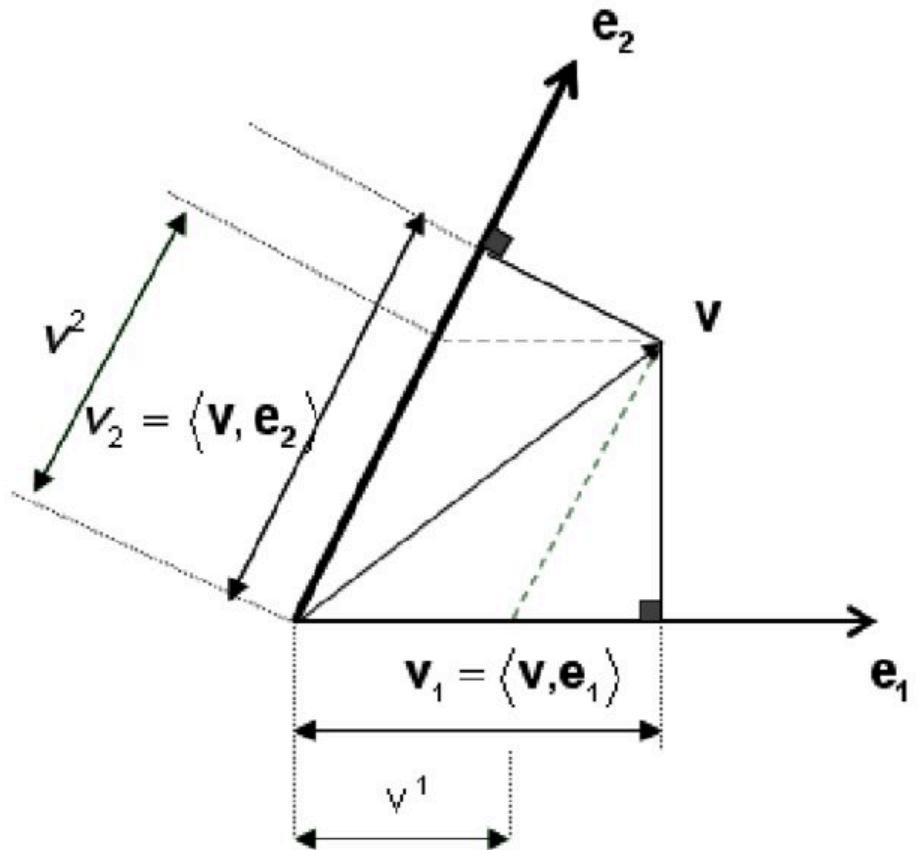
- The coefficients of a vector v in the **primal basis** are called the **contravariant coefficients**:

$$v = v^i e_i$$

- The coefficients of a vector v in the **reciprocal basis** are called the **covariant coefficients**:

$$v = v_i e^i$$

Geometric reading the covariant/contravariant coefficients of a vector



In a Cartesian coordinate system, the contravariant components match the covariant components

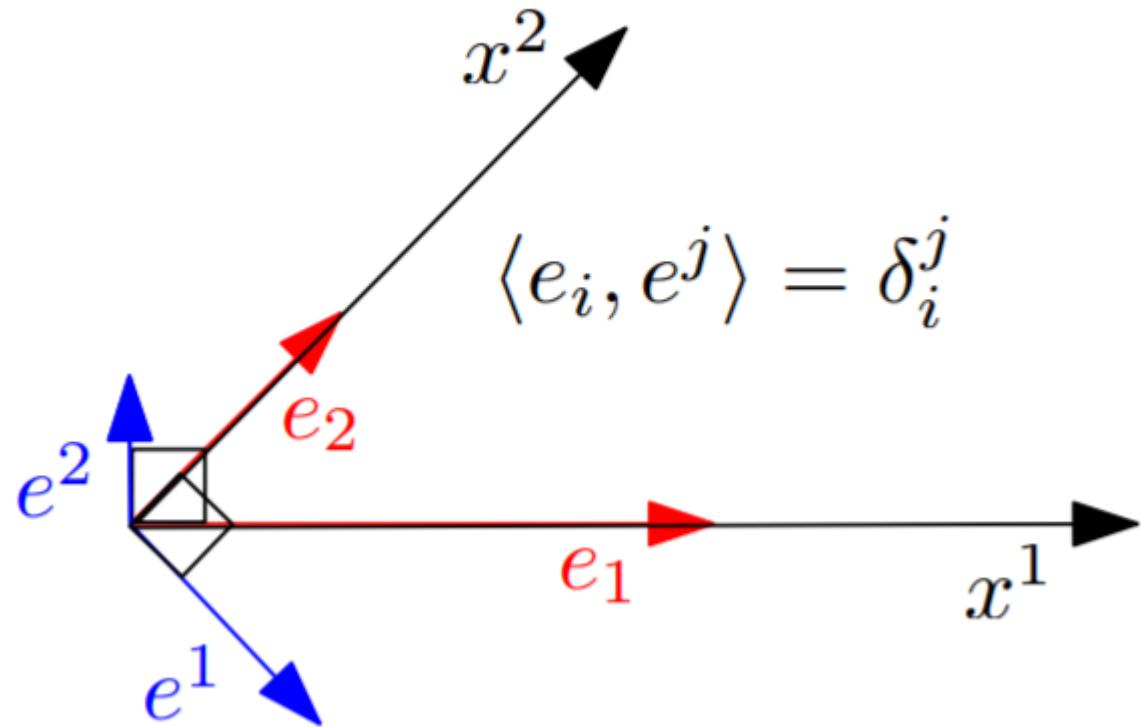
contravariant

$$v^i = \langle v, e_i \rangle$$

covariant

$$v_i = \langle v, e^i \rangle$$

Primal and reciprocal basis are mutually orthogonal



Scalar product and dual metric tensors

$$\langle u, v \rangle = u^i v_i = u_i v^i$$

$$g_{ij} = \langle e_i, e_j \rangle,$$

$$G = [g_{ij}]$$

$$g^{*ij} = g^{ij} = \langle e^i, e^j \rangle.$$

$$G^* = [g^{ij}]$$

- Scalars are tensors of order 0
- Vectors are contravariant tensors of order 1
- Covectors are covariant tensors of order 1

$$G \times G^* = I$$

Converting covariant \leftrightarrow contravariant components



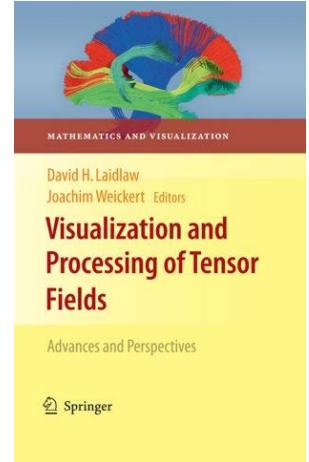
$$e^i = g^{*ij} e_j$$

$$e_i = g_{ij} e^j$$

$$v_i = g_{ij} v^i$$

$$v^i = g^{ij} v_i$$

Geometric tensors and tensor algebra



- Informally, tensor = multi-array of coefficients...
- Got attention in the media in deep learning with TensorFlow
- Tensors are **geometric objects** interpreted as **multilinear maps**

A tensor of type (r,s) $T : \underbrace{V^* \dots V^*}_r \times \underbrace{V \times \dots V}_s \rightarrow \mathbb{R}$

Components/coefficients

$$T_{i_1 \dots i_s}^{j_1 \dots j_r}$$

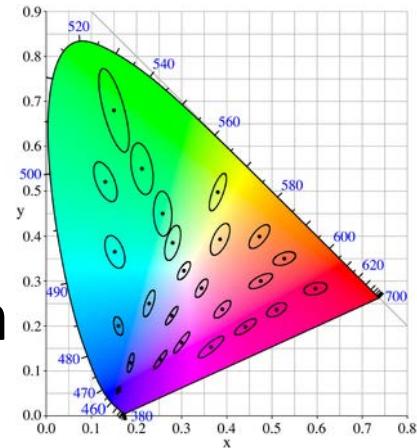
with respect to a basis

Later, we shall see that g is a 2-covariant tensor:

$$g = g_{ij} dx_i \otimes dx_j$$

Riemannian metric tensor g

- On a manifold, a smooth 2-covariant tensor field
- On each tangent space, define an inner product space
(extrinsic=embedded versus intrinsic visualization/interpretation)
- Union of all tangent spaces is called the **tangent bundle**



- Eat two vectors...

- **Bilinear positive-definite**
$$g(aU+V,W)=ag(U,W)+g(V,W)$$

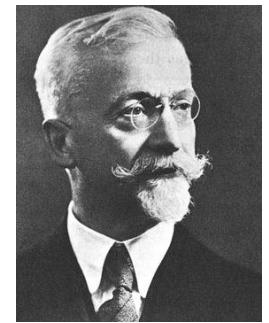


- **symmetric**
$$g(V,W) = g(W,V)$$

In (local) coordinates:
$$g_p = g_p(\partial_i(p), \partial_j(p)) = g_{ij}(p)$$

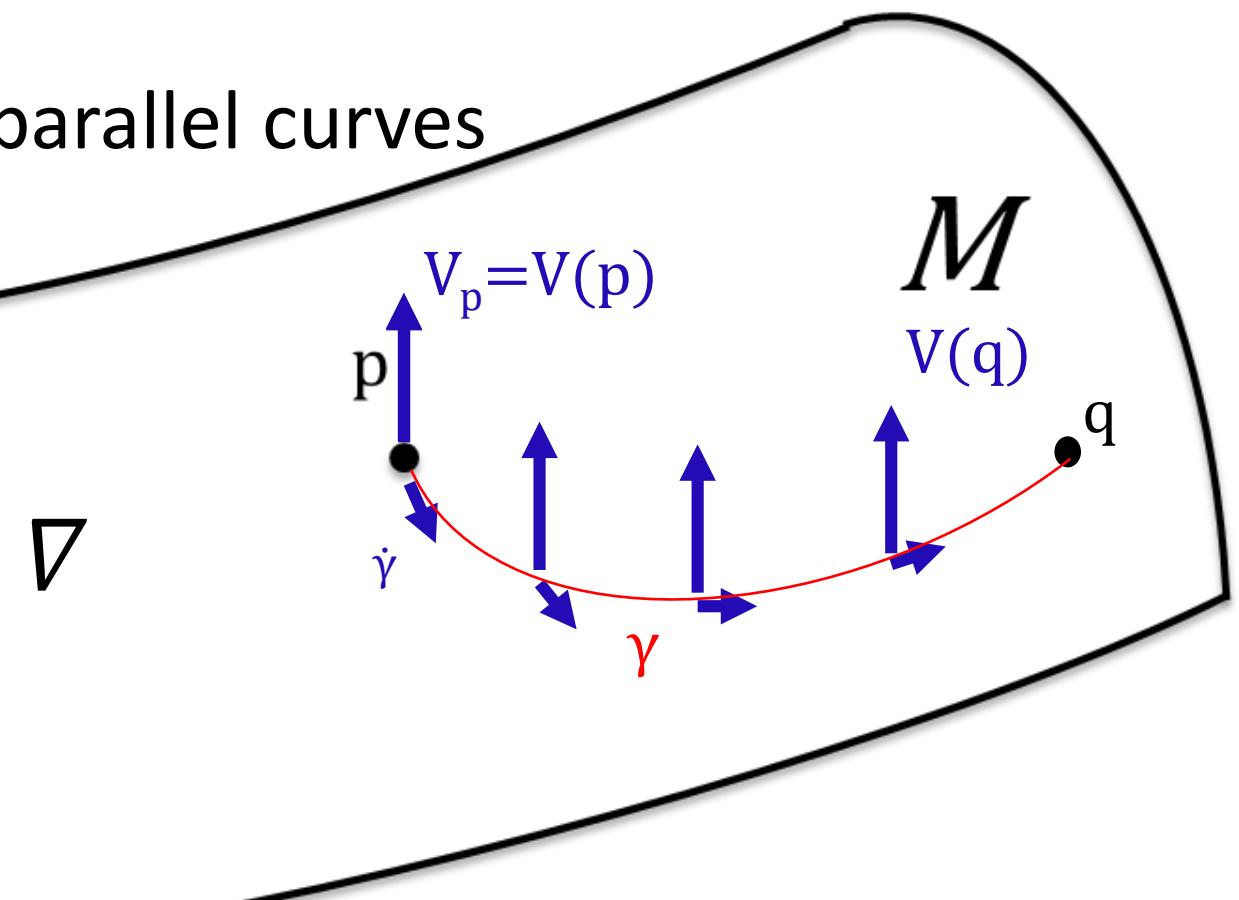
- **nondegenerate**
$$\forall p, \forall V \neq 0 \exists W, g_p(V,W) \neq 0$$

Affine connection



- Define how to **parallel transport** a vector from one tangent plane to another tangent plane by **infinitesimally parallel shifting** it along a curve
- Use to define **geodesics** as autoparallel curves

Also covariant derivative...



Defining an affine connection

- Report d^3 smooth functions, called **Christoffel symbols**
- In a local coordinate chart with natural basis, we have:

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k$$



Elwin Bruno Christoffel
(1829-1900)

- **Christoffel symbols** are not tensors, they do not obey the covariant/contravariant laws of change of basis

∇ -geodesics

- Geodesics are “straight lines”, auto-parallel lines

$$\nabla_{\dot{\gamma}} \dot{\gamma} = 0$$

- We find geodesics by solving a second-order Ordinary Differential Equations (ODE

$$\ddot{\gamma}(t) + \Gamma_{ij}^k \dot{\gamma}(t) \dot{\gamma}(t) = 0, \quad \gamma^l(t) = x^l \circ \gamma(t)$$

Riemannian metric-compatible connection

- A connection is metric compatible if for any smooth vectors fields X,Y,Z

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$$

- In local coordinates:

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

- Metric-compatible connection enjoys parallel transport with the property:

$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \rightarrow c(t)}^{\nabla} u, \prod_{c(0) \rightarrow c(t)}^{\nabla} v \right\rangle_{c(t)} \quad \forall t$$

Fundamental theorem of Riemannian geometry

- There exists a **unique torsion-free affine connection compatible with the metric** called the Levi-Civita connection:

$$\nabla^{\text{LC}}$$

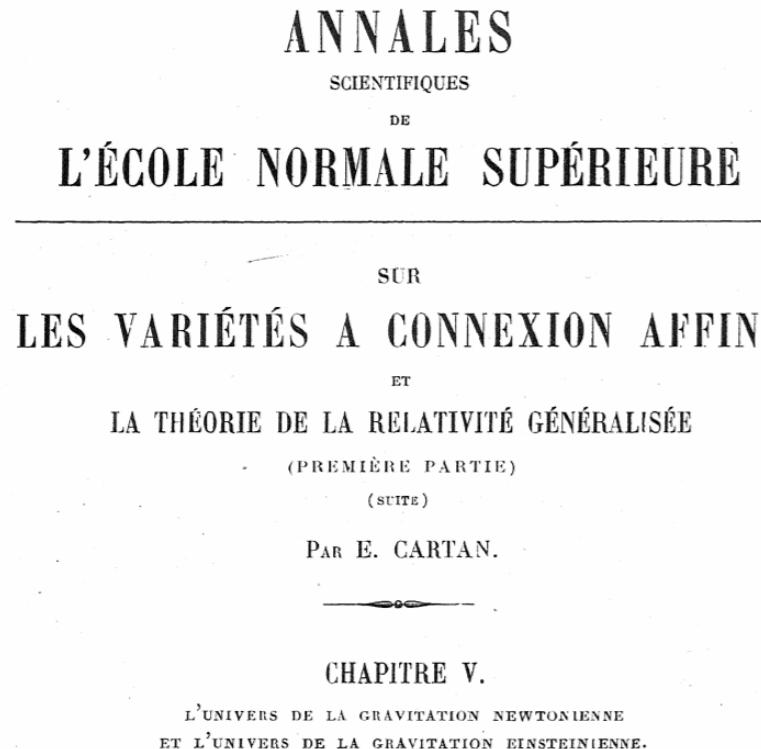
- The **Christoffel symbols** of the Levi-Civita connection are calculated from the metric tensor in local coordinates :

$${}^{\text{LC}}\Gamma_{ij}^k = \frac{1}{2}g^{kl} (\partial_i g_{il} + \partial_j g_{il} - \partial_l g_{ij})$$

- Or in coordinate-free equation by Koszul formula:

$$2g(\nabla_X Y, Z) = X(g(Y, Z)) + Y(g(X, Z)) - Z(g(X, Y)) + g([X, Y], Z) - g([X, Z], Y) - g([Y, Z], X)$$

Elie Cartan's study of affine connections



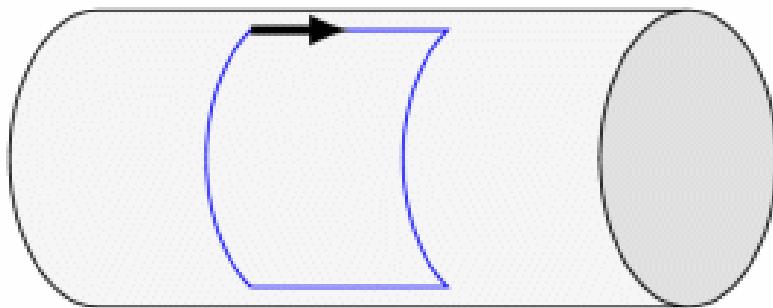
Cartan-Einstein manifold

La forme invariante des lois de la gravitation newtonienne.

70. Nous avons vu au Chapitre I qu'il était possible, et d'une infinité de manières, de ramener la gravitation newtonienne à la Géométrie en attribuant à l'Univers une connexion affine convenable. Dans cette

E. Cartan, Sur les variétés à connexion affine, et la théorie de la relativité généralisée , Ann. Ec. Norm. Sup. 40 (1923)

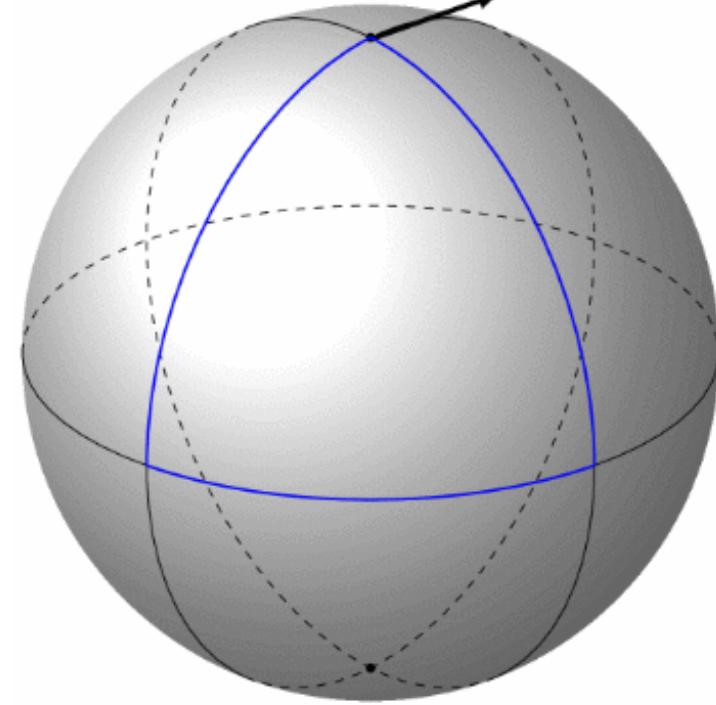
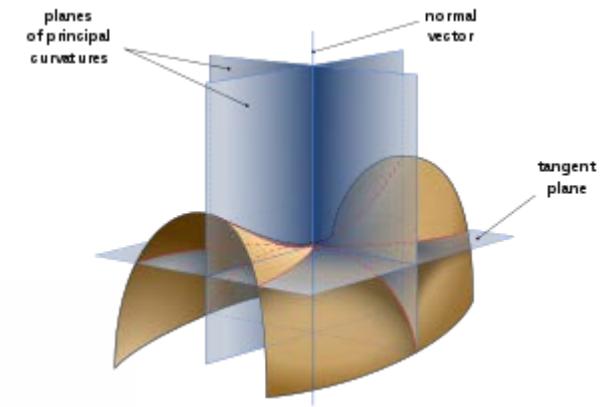
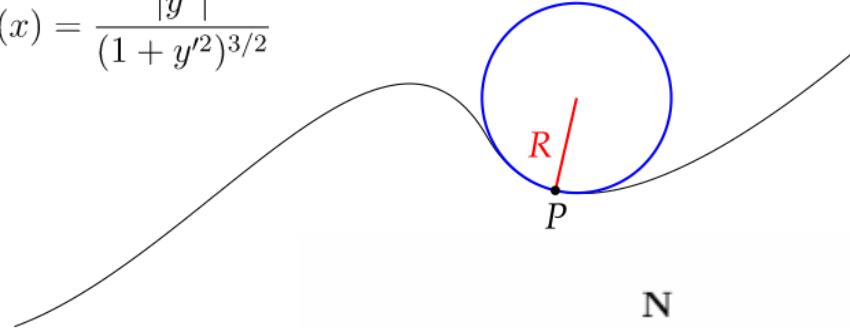
Curvature of ∇



Cylinder is flat

Parallel transport is
independent of path

$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$



S

Sphere has constant curvature
Parallel transport is path-dependent

A word about torsion of a connection ∇

Torsion measures the speed of rotation of the binormal vector

parallel transport “twists” vectors.

- For connections:

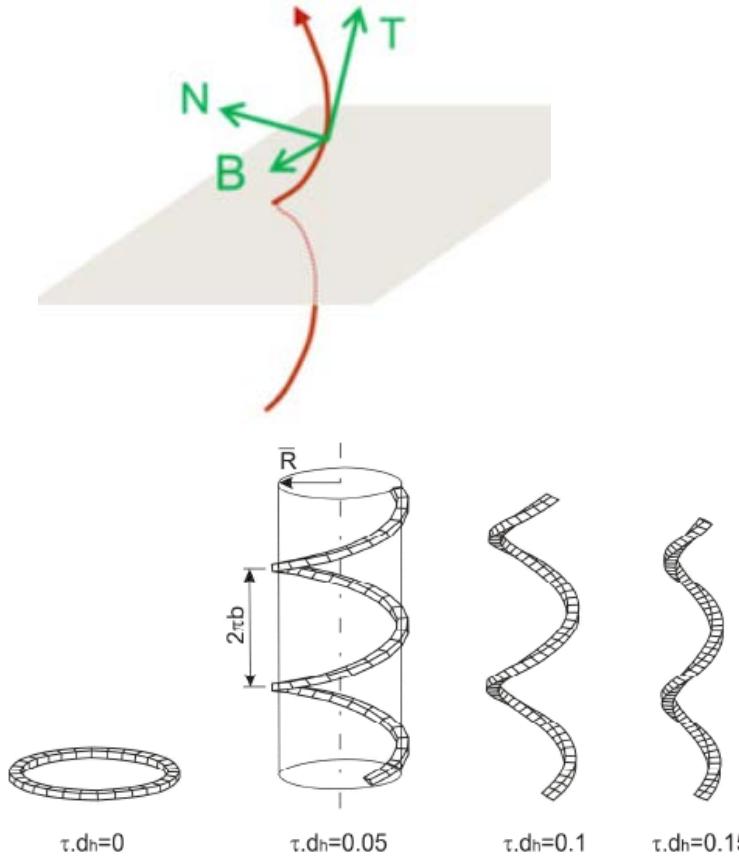


Figure 1. Helical channels with square cross section, constant curvature $\kappa.dh = 1$ and torsion $\tau.dh$ spanning from 0 to 0.15.

Torsion in geometry and in field theory

3

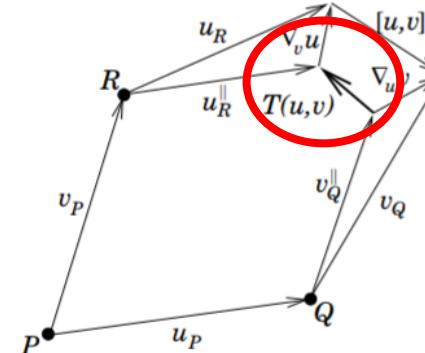
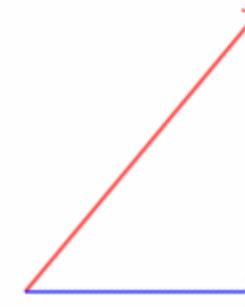


Figure 1: *On the geometrical interpretation of torsion*, see [39]: Two vector fields u and v are given. At a point P , we transport parallelly u and v along v or u , respectively. They become u_R^{\parallel} and v_Q^{\parallel} . If a torsion is present, they don't close, that is, a closure failure $T(u, v)$ emerges. This is a schematic view. Note that the points R and Q are infinitesimally near to P . A proof can be found in Schouten [88], p.127.



Connections differing by torsions have same geodesics
Pregeodesics

Summary

- Algebraic structures: Vector and dual covector spaces with natural pairing, inner product space and contravariant/covariant coordinates, tensor space and dyadic product
- Manifold with an affine connection: tensor fields, parallel transport, geodesics, curvature and torsion

Distances and entropies



Frank Nielsen



Sony CSL

Distances

- Too many **synonyms** and **ambiguities** in the literature
(two-point function, notion of distinguishability, discrepancy, divergence, metric, relative entropy, measure of discrimination, coefficient of divergence, etc.)
- Distance between points, densities, random variables, etc.
- **Statistical divergence** versus **parameter divergence**
- Principal distances and main **classes of distances**
- Generalized entropies and relative entropies

Metric distances and metric spaces (X,D)

A **metric** D is a (distance) function that satisfies the following axioms:

- M1. (Non-negativity) $D(p_1, p_2) \geq 0$
- M2. (Identity of the indiscernibles) $D(p_1, p_2) = 0 \Leftrightarrow p_1 = p_2$
- M3. (Symmetry) $D(p_1, p_2) = D(p_2, p_1)$
- M_4. (Triangle inequality/subadditivity)

$$D(p_1, p_2) + D(p_2, p_3) \geq D(p_1, p_3)$$

Examples of metric spaces

- Euclidean distance

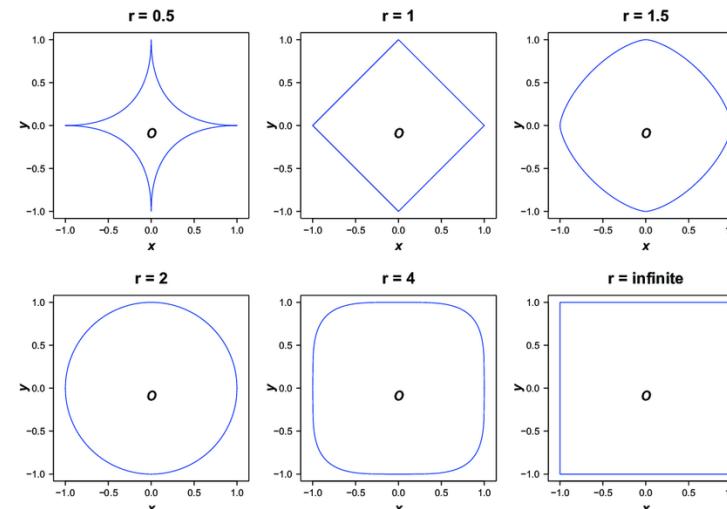
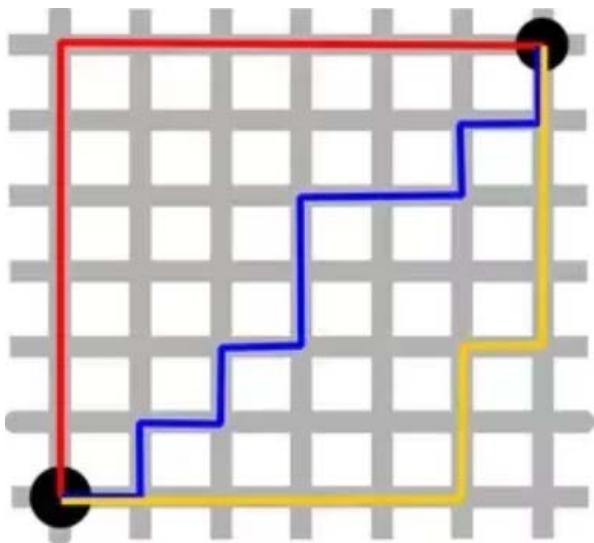
$$D_E(p, q) = \sqrt{\sum_{i=1}^d (p_i - q_i)^2}$$

- Manhattan/Taxi cab distance

$$M_1(p, q) = \sum_{i=1}^d |p_i - q_i|$$

- Minkowski metric distances

$$M_\alpha(p, q) = \left(\sum_{i=1}^d |p_i - q_i|^\alpha \right)^{\frac{1}{\alpha}}, \quad \alpha \geq 1$$



Inner product, induced norms and induced distance

- Inner product $\langle x, y \rangle_G$
- Induced norm $\|x\|_G = \sqrt{\langle x, x \rangle_G}$
- Induced metric distance $D_G(p, q) = \|p - q\|_G$
- Example with Euclidean distance and its dot/scalar product

$$\langle x, y \rangle_E = \sum_{i=1}^d x_i y_i \quad \longrightarrow \quad D_E(p, q) = \|p - q\|_E = \|p - q\|_2$$

- Example with Minkowski norms

$$\|x\|_\alpha = \left(\sum_i |x_i|^\alpha \right)^{\frac{1}{\alpha}} \quad \longrightarrow \quad M_\alpha(p, q) = \|p - q\|_\alpha$$

Distances and notational conventions

- Distances between strings, vectors, matrices (tensors), graphs, probability densities, cumulative distribution functions, random variables (mutual information), etc.
- : to indicate that the distance is oriented, asymmetric: $D(p : q)$
$$D(p : q) \neq D(q : p)$$

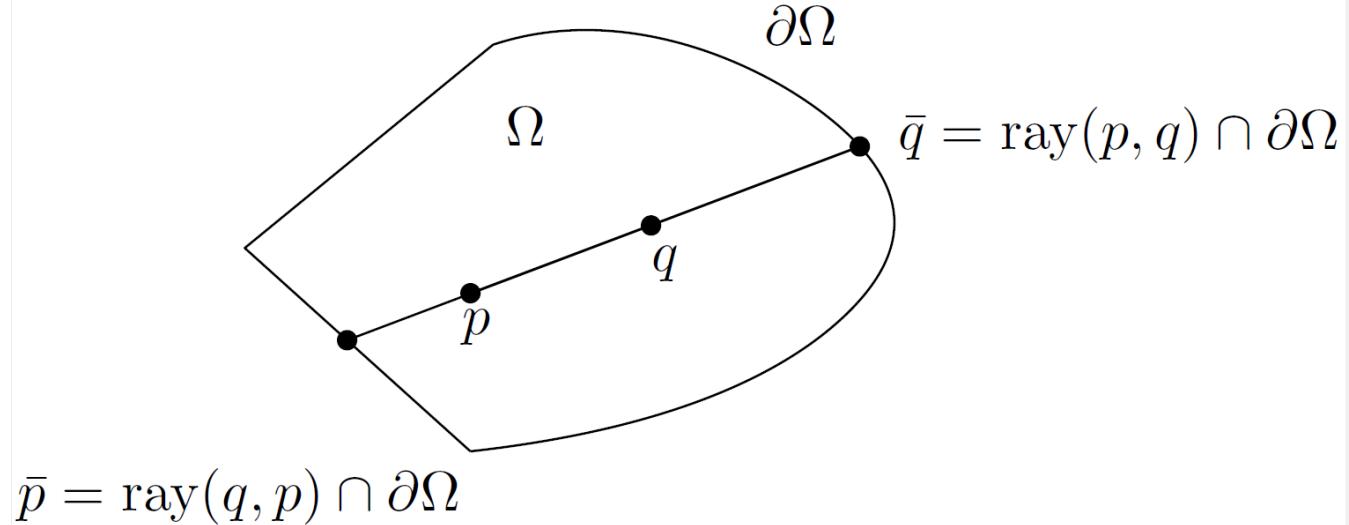
Stemmed from **information theory** $D(p \| q)$ to avoid confusion with joint variables $H(X, Y)$

- ; to indicate a symmetric but non-metric distance: $D(p; q)$

Example: Mutual information

- Bracket to indicate a statistical distance: $D[p : q]$
- For a parametric family, a statistical distance amount to a parameter distance:
$$D_{\mathcal{P}}(\theta_1 : \theta_2) = D[p_{\theta_1} : p_{\theta_2}]$$

Signed distances (failing non-negativity)



Hilbert-cross ratio metric
(signed)

$$H_\Omega(p, q) = \log \frac{\|\bar{q}-p\| \|\bar{p}-q\|}{\|\bar{q}-q\| \|\bar{p}-p\|}$$
$$H_\Omega(p, q) = \log \text{CR}(\bar{p}, p, q, \bar{q}) = \log \frac{\|\bar{q}-p\| \|\bar{p}-q\|}{\|\bar{q}-q\| \|\bar{p}-p\|}$$
$$H_\Omega(p, q) = \log |\text{CR}(\bar{p}, p, q, \bar{q})|$$

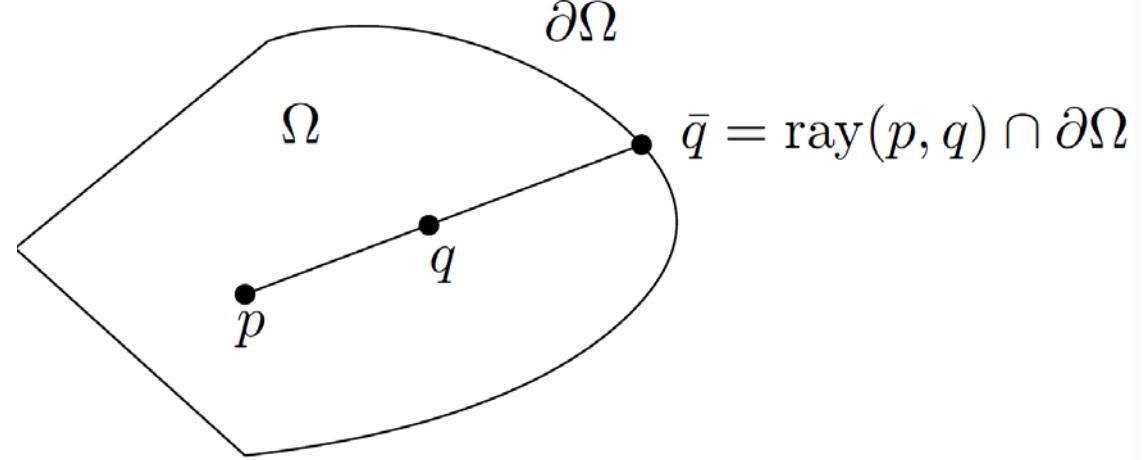
Pseudo-metrics: Failing the identity of the indiscernibles

- For example, we would like that the distance of a substring s' to a string s containing s' is zero but not the converse.
- Schubert distance: To give a geometric example, consider the distances between subspaces, where a k -dimensional subspace S of \mathbb{R}^d is represented by a (d, k) matrix S that consists of the k orthonormal base vectors arranged in column in S . The *Schubert distance* between k_1 -dimensional subspace S_1 and k_2 -dimensional subspace S_2 is defined by

$$\delta_S(S_1, S_2) = \sqrt{\sum_{i=1}^{\min\{k_1, k_2\}} \theta_i(S_1, S_2)^2},$$

where $\theta_i(S_1, S_2) = \arccos \lambda_i(S_1^\top S_2)$ is the i -th principal angle and $\lambda_i(X)$ denotes the i -th largest eigenvalue of matrix X . We have $\delta_S(S_1, S_2) = 0$ whenever S_1 is a subspace of S_2 (an asymmetric property).

Failing symmetry: E.g., Funk oriented distance



Hilbert cross-ratio metric is the arithmetic
Symmetrization of Funk distances

$$F_\Omega(p, q) = \log \frac{\|p - \bar{q}\|}{\|q - \bar{q}\|}$$

$$H_\Omega(p_1, p_2) = \frac{F_\Omega(p_1, p_2) + F_\Omega^r(p_1, p_2)}{2}$$

Reverse or dual distance (reference duality)

$$D^r(p : q) = D^*(p : q) = D(q : p)$$

Failing triangle inequality/subadditivity:

- Kullback-Leibler divergence between two pmfs:

$$\text{KL}(p : q) = \sum_i p_i \log \frac{p_i}{q_i}$$

- Notice that the squared Euclidean distance fails the triangle inequality

Scale-invariant distances



- Itakura-Saito divergence:

Fumitada Itakura

$$D_{\text{IS}}(p : q) = \sum_i \frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1$$

- Scale-invariant:

$$D_{\text{IS}}(\lambda p : \lambda q) = D_{\text{IS}}(p : q), \quad \lambda > 0$$

- Often used in music applications (spectrum)

Projective distances: E.g., Birkhoff's distance

- Distance independent of both argument scaling factors
- A cone that induces a partial order $p \preceq_C q \Leftrightarrow q - p \in C$

$$B_C(p, q) = \log \frac{M(p:q)}{m(p:q)} = \log M_C(p : q) M_C(q : p)$$

$$M_C(p : q) = \inf\{\beta \in \mathbb{R} : p \preceq_C \beta q\}$$

$$m_C(p : q) = \sup\{\alpha \in \mathbb{R} : \alpha q \preceq_C p\}$$

- For the positive orthant cone, we have Birkhoff's distance:

$$\tilde{\delta}(p, q) = \log \max_{i,j} \frac{p_i q_j}{p_j q_i} \quad \tilde{\delta}(\lambda_1 p, \lambda_2 q) = \tilde{\delta}(p, q), \quad \forall \lambda_1, \lambda_2 > 0$$

Statistical distance: Total Variation (TV) metric

$$\text{TV}(P, Q) = \sup_{E \in \mathcal{F}} |P(E) - Q(E)|$$

- The TV measures the **largest probability difference of an event E** of the σ -algebra of the sample space.
- When P and Q admit Radon-Nikodym densities p and q wrt μ , respectively, we have

$$\text{TV}(p, q) = \frac{1}{2} \|p(x) - q(x)\| d\mu(x)$$

$$\text{TV}(p, q) = \frac{1}{2} \|p - q\|_1$$

- Synonyms: city block distance, overlap distance

Kolmogorov metric distance



- A distance between distribution functions, less than TV:

$$K(F_X, F_Y) = \sup_{u \in \mathbb{R}} |F_X(u) - F_Y(u)|.$$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(X_i)$$

Kolmogorov–Smirnov test

$$D_n = \sup_x |F_n(x) - F(x)|$$

Classes of distances: Csiszar's f-divergence

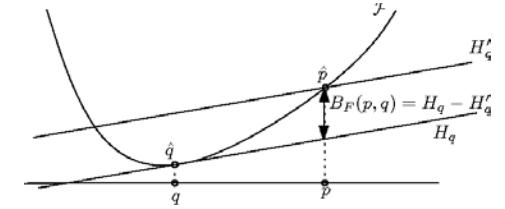
- Function f convex, strictly convex at 1, with $f(1)=0$

$$I_f(p : q) = \int p f\left(\frac{q}{p}\right) d\mu \geq f(1)$$

- Include the Kullback-Leibler divergence for $f(u)=-\log u$
- Invariant divergence** in information geometry (information monotone)

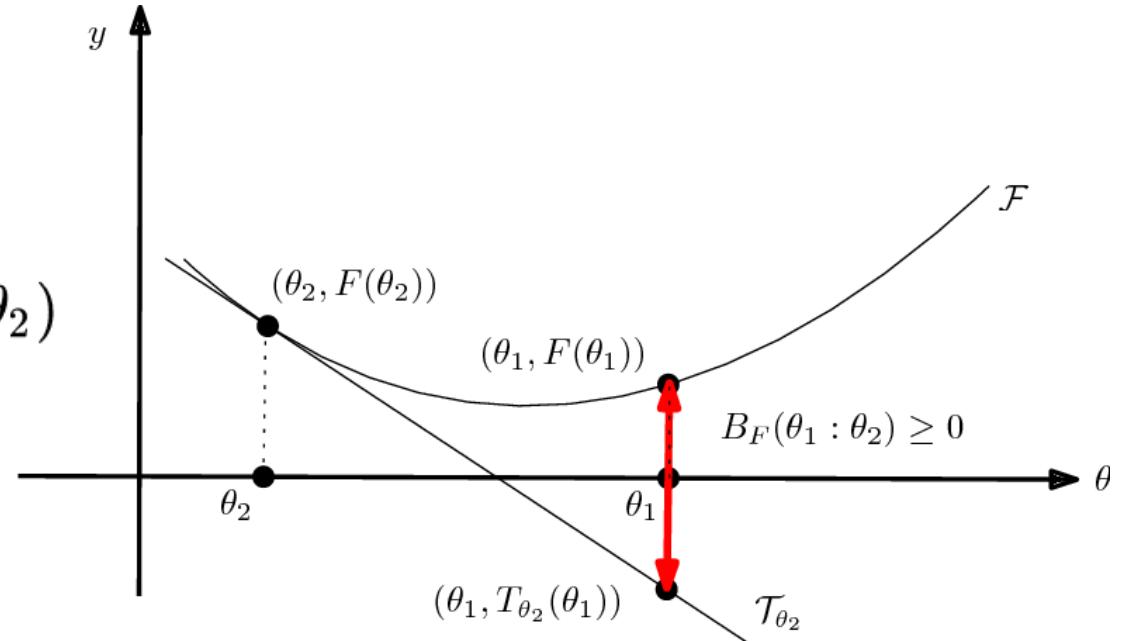
| Name of the f -divergence | Formula $I_f(P : Q)$ | Generator $f(u)$ with $f(1) = 0$ |
|-----------------------------|--|---|
| Total variation (metric) | $\frac{1}{2} \int p(x) - q(x) d\nu(x)$ | $\frac{1}{2} u - 1 $ |
| Squared Hellinger | $\int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\nu(x)$ | $(\sqrt{u} - 1)^2$ |
| Pearson χ_P^2 | $\int \frac{(q(x) - p(x))^2}{p(x)} d\nu(x)$ | $(u - 1)^2$ |
| Neyman χ_N^2 | $\int \frac{(p(x) - q(x))^2}{q(x)} d\nu(x)$ | $\frac{(1-u)^2}{u}$ |
| Pearson-Vajda χ_P^k | $\int \frac{(q(x) - \lambda p(x))^k}{p^{k-1}(x)} d\nu(x)$ | $(u - 1)^k$ |
| Pearson-Vajda $ \chi _P^k$ | $\int \frac{ q(x) - \lambda p(x) ^k}{p^{k-1}(x)} d\nu(x)$ | $ u - 1 ^k$ |
| Kullback-Leibler | $\int p(x) \log \frac{p(x)}{q(x)} d\nu(x)$ | $-\log u$ |
| reverse Kullback-Leibler | $\int q(x) \log \frac{q(x)}{p(x)} d\nu(x)$ | $u \log u$ |
| α -divergence | $\frac{4}{1-\alpha^2} (1 - \int p^{\frac{1-\alpha}{2}}(x) q^{1+\alpha}(x) d\nu(x))$ | $\frac{4}{1-\alpha^2} (1 - u^{\frac{1+\alpha}{2}})$ |
| Jensen-Shannon | $\frac{1}{2} \int (p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)}) d\nu(x)$ | $-(u+1) \log \frac{1+u}{2} + u \log u$ |

Classes of distances: Bregman divergence



- Bregman divergence between parameters for a strictly convex and differentiable convex function F

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



- The canonical divergence of dually flat spaces
- Extend to other types (matrices, functions, etc)

Mining matrix data with Bregman matrix divergences for portfolio selection." *Matrix Information Geometry*. Springer, Berlin, Heidelberg, 2013. 373-402.

Jensen difference/Jensen divergence (Burbea-Rao)

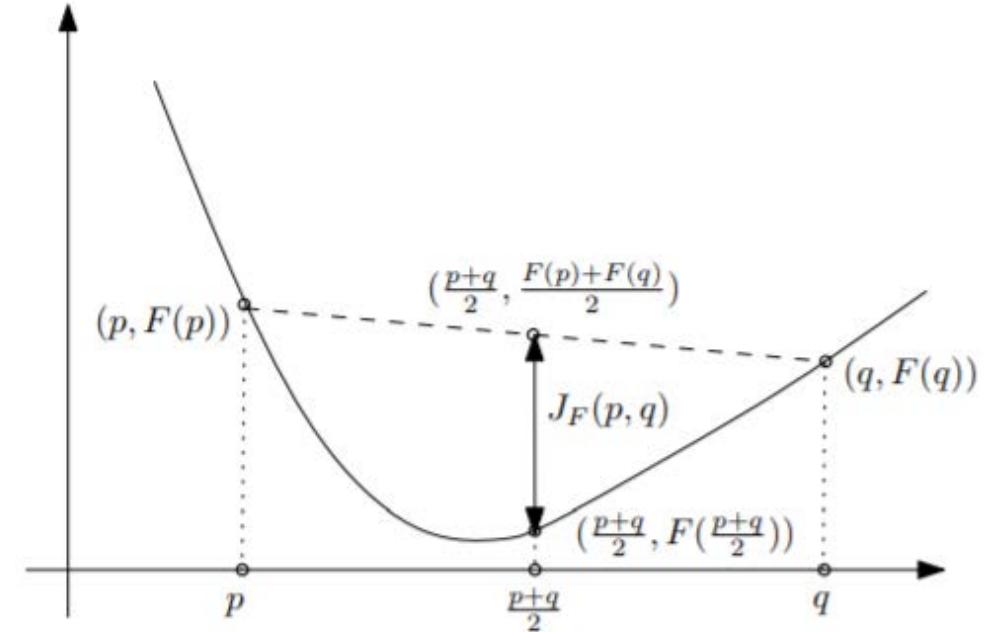
- Introduced by Burbea and Rao
- Vertical gap induced by Jensen inequality

$$J_F(\theta_1, \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right) \geq 0$$

Asymptotic scaled Jensen divergence amount to a Bregman or reverse Bregman divergence

$$J_\alpha^F(\theta_1 : \theta_2)$$

$$= \begin{cases} \frac{1}{\alpha(1-\alpha)} J'^F(\theta_1 : \theta_2) & \alpha \neq \{0, 1\} \\ B_F(\theta_1 : \theta_2) & \alpha = 1 \\ B_F(\theta_2 : \theta_1) & \alpha = 0 \end{cases}$$

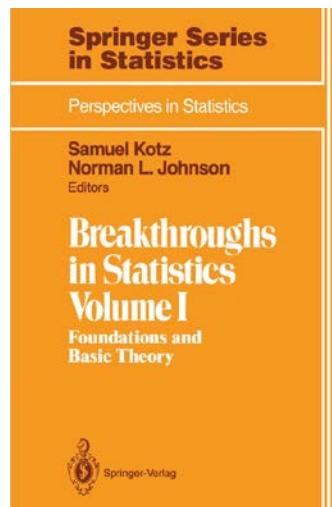


The Burbea-Rao and Bhattacharyya centroids." *IEEE Transactions on Information Theory* 57.8 (2011): 5455-5466.
Bregman chord divergence: <https://arxiv.org/abs/1810.09113>

Summary

- Distance measures the **separation of entities** (vectors, probability measures, probability densities, cumulative distribution functions, random variables, matrices, functions, etc.)
- A **metric** is a symmetric non-negative distance that satisfies the law of the indiscernibles and the triangle inequality
- A **divergence** originally meant a statistical distance (eg., probability metric), and means a smooth parametric distance in information geometry
- Statistical divergences on parametric densities amount to parameter divergences
- Three classes of **non-mutually exclusive parametric distances**: The f-divergences, Bregman divergences, and Jensen divergences, that are non-mutually exclusive
- But also Wasserstein distance in optimal transport, etc.

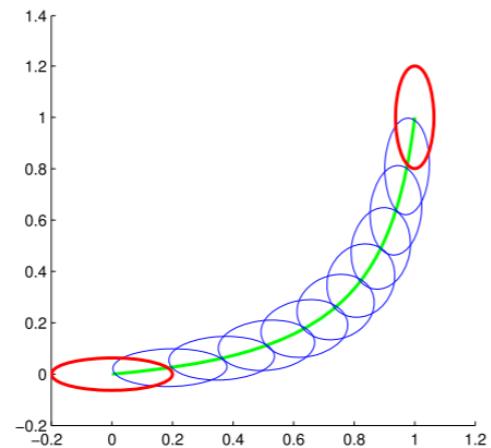
Fisher-Rao Riemannian geometry



Frank Nielsen



Sony CSL



Recalling the Fisher information metric

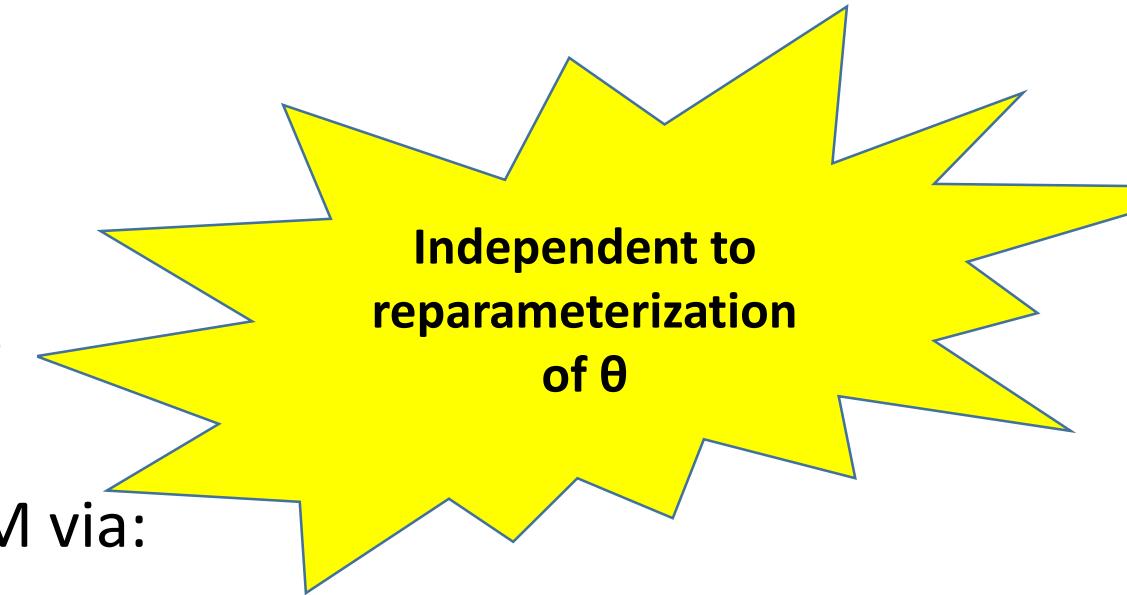
- Fisher Information Metric

$$g_{jk}(\theta) = \int_X \frac{\partial \log p(x, \theta)}{\partial \theta_j} \frac{\partial \log p(x, \theta)}{\partial \theta_k} p(x, \theta) dx.$$

- Infinitesimally, the KLD is related to the FIM via:

$$D_{\text{KL}}[P(\theta_0) \| P(\theta)] = \frac{1}{2} \sum_{jk} \Delta \theta^j \Delta \theta^k g_{jk}(\theta_0) + O(\Delta \theta^3).$$

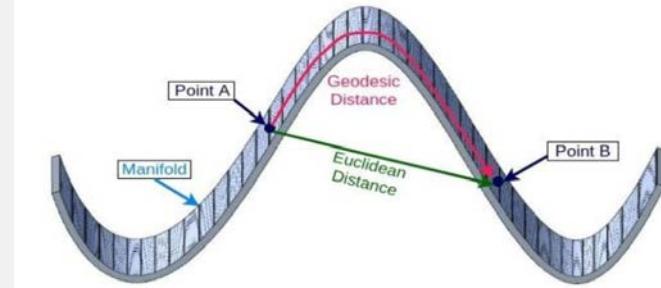
(Kind of squared Mahalanobis distance; holds for any standard f-divergence)



Rao distance is Riemannian geodesic distance

- ▶ Infinitesimal length element :

$$ds^2 = \sum_{ij} g_{ij}(\theta) d\theta_i d\theta_j = d\theta^T I(\theta) d\theta$$



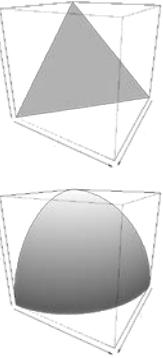
- ▶ Geodesic and distance are hard to explicitly calculate :

$$\rho(p(x; \theta_1), p(x; \theta_2)) = \min_{\substack{\theta(s) \\ \theta(0)=\theta_1 \\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{ds} \right)^T I(\theta) \frac{d\theta}{ds}} ds$$

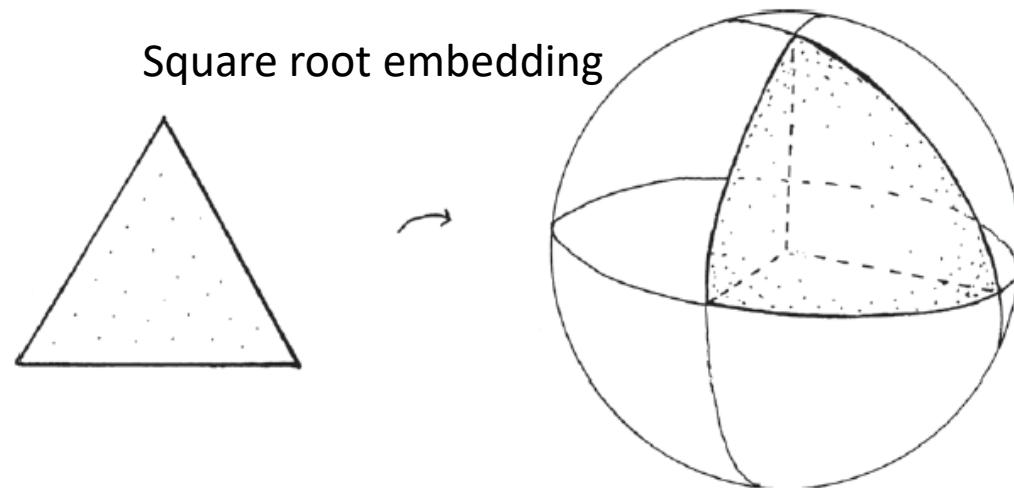
Riemannian
geodesics locally
minimize
lengths

- ▶ Metric property of ρ , many tools [1] : Riemannian Log/Exp tangent/manifold mapping

Simplex (categorical distribution)



- Trinomial (trinoulli)



Embedding to the sphere orthant

Information metric:

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

(Hotelling)-Fisher-Rao distance:

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

In practice, calculating Rao's distance is difficult

$$d(\theta^1, \theta^2) = \min_{\theta(t)} \int_{t_1}^{t_2} \sqrt{\sum_{i=1}^p \sum_{j=1}^p g_{ij}(\theta(t)) \frac{d\theta_i(t)}{dt} \frac{d\theta_j(t)}{dt}} dt.$$

- Need to solve the ODE for find the **geodesic**:

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

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{m=1}^p \left(\frac{\partial g_{im}(\theta)}{\partial \theta_j} + \frac{\partial g_{jm}(\theta)}{\partial \theta_i} - \frac{\partial g_{ij}(\theta)}{\partial \theta_m} \right) g^{mk}(\theta), \quad i, j, k = 1, \dots, p,$$

- Need to **integrate** the infinitesimal length elements along the geodesics...

Hotelling's 1930 paper considered location-scale families

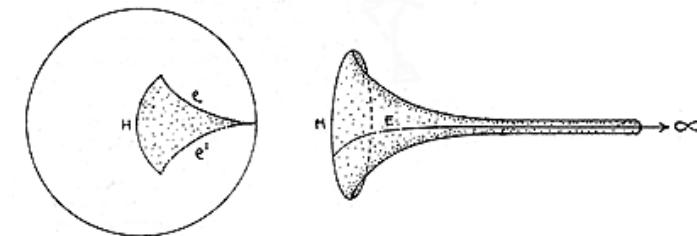
Spaces of Statistical Parameters.

By Harold Hotelling, Stanford University.

$$f(x|\mu, \sigma) = \frac{1}{\sigma} f((x - \mu)/\sigma)$$

For a space of n dimensions representing the parameters p_1, \dots, p_n of a frequency distribution, a statistically significant metric is defined by means of the variances and

- 2D FIM



- Constant curvature, isometric to hyperbolic geometry of curvatures

$$-\frac{1}{\beta^2}$$

$$\beta^2 := \int \left(x \frac{p'(x)}{p(x)} + 1 \right)^2 p(x) dx$$



Harold
Hotelling

Some common Fisher-Rao geodesic distances

| Distribution | Density | Geodesic Distance |
|---------------------------------------|--|---|
| Binomial | $\binom{n}{x} p^x (1-p)^{n-x}$ | $2\sqrt{n} \arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2}) $ |
| Poisson | $\frac{e^{-\lambda} \lambda^x}{x!}$ | $2 \sqrt{\lambda_1} - \sqrt{\lambda_2} $ |
| Geometric | $(1-p)p^x$ | $2 \log \frac{1-\sqrt{p_1 p_2} + \sqrt{p_1} - \sqrt{p_2} }{\sqrt{(1-p_1)(1-p_2)}}$ |
| Gamma | $\frac{e^{-\theta x} \theta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}$ | $\sqrt{\alpha} \log \theta_1 - \log \theta_2 $ |
| Normal (fixed variance) | $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\frac{ \mu_1 - \mu_2 }{\sigma}$ |
| Normal (fixed mean) | $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\sqrt{2} \log \sigma_1 - \log \sigma_2 $ |
| General Normal | $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $2\sqrt{2} \tanh^{-1} \sqrt{\frac{(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2}{(\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2)^2}}$ |
| p -Variate Normal (Σ fixed) | $\frac{e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)}}{(2\pi)^{p/2} \Sigma ^{1/2}}$ | $(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$ |
| p -Variate Normal (μ fixed) | $\frac{e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)}}{(2\pi)^{p/2} \Sigma ^{1/2}}$ | $\frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^p \log \lambda_i^2}$ |
| Multinomial | (here, $\{\lambda_i\}$ are the $\frac{n!}{\prod_{i=1}^k n_i!} p_i^{n_i}$ | roots of $ \Sigma_2 - \lambda \Sigma_1 = 0$ $2\sqrt{\pi} \arccos(\sum_{i=1}^k \sqrt{p_i \theta_i})$ |

Approximating geodesics for multivariate normal via geodesic shooting

Algorithm 1 Shooting method for minimal geodesics on $\mathcal{N}(n)$

Given: Initial point $P_0 = (\mu_0, \Sigma_0)$, final point $P_1 = (\mu_1, \Sigma_1)$.

Output: Minimal geodesic $P(t) = (\mu(t), \Sigma(t))$, $t \in [0, 1]$, such that $P(1) = (\mu_1, \Sigma_1)$.

Initialization: Choose initial velocities $V(0) = (\dot{\mu}(0), \dot{\Sigma}(0))$ (e.g., zeroes), initial values for ϵ (10^{-5}), error = 10^6 .

while $\text{error} \geq \epsilon$ **do**

Numerically integrate the geodesic equations (13), (14) for given initial conditions $(\mu_0, \Sigma_0, \dot{\mu}_0, \dot{\Sigma}_0)$ from $t = 0$ to $t = 1$.

Denote the solution by $(\mu(t), \Sigma(t))$;

Set $W(1) = (W_\mu(1), W_\Sigma(1)) = (\mu_1 - \mu(1), \Sigma_1 - \Sigma(1))$;

Calculate error = $\|W(1)\|_{P_1} = \sqrt{W_\mu(1)^T \Sigma_1^{-1} W_\mu(1) + \frac{1}{2} \text{tr}((\Sigma_1^{-1} W_\Sigma(1))^2)}$;

Numerically integrate the parallel transport equations (18) and (19) for given trajectory $(\mu(t), \Sigma(t))$ and final velocities $W(1)$, backward in time from $t = 1$ to $t = 0$;

Numerically calculate Jacobi field $J(1)$ from (22),

$$J(1) = \frac{\exp_{P_0}(V(0) + \alpha W(0)) - \exp_{P_0}(V(0))}{\alpha}, \text{ where } \alpha \text{ is sufficiently small value and we use } \frac{\epsilon}{\|W(0)\|_{P_0}}$$

Determine proper update size s :

$$s_1 = \frac{(W(1), J(1))_{P(1)}}{\|J(1)\|_{P(1)}^2}$$

if $\|W(1)\|_{P(1)} > 0.05$ **then**

$$s = 0.05 / \|W(1)\|_{P(1)} s_1;$$

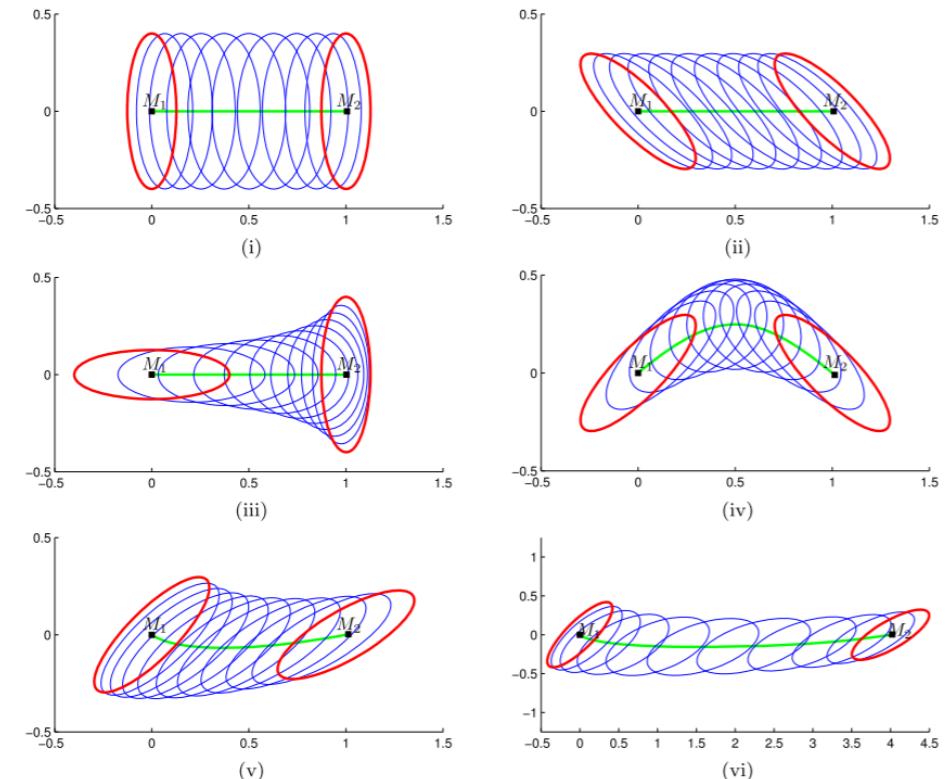
else

$$s = s_1;$$

end if

$$V(0) \leftarrow V(0) + s W(0);$$

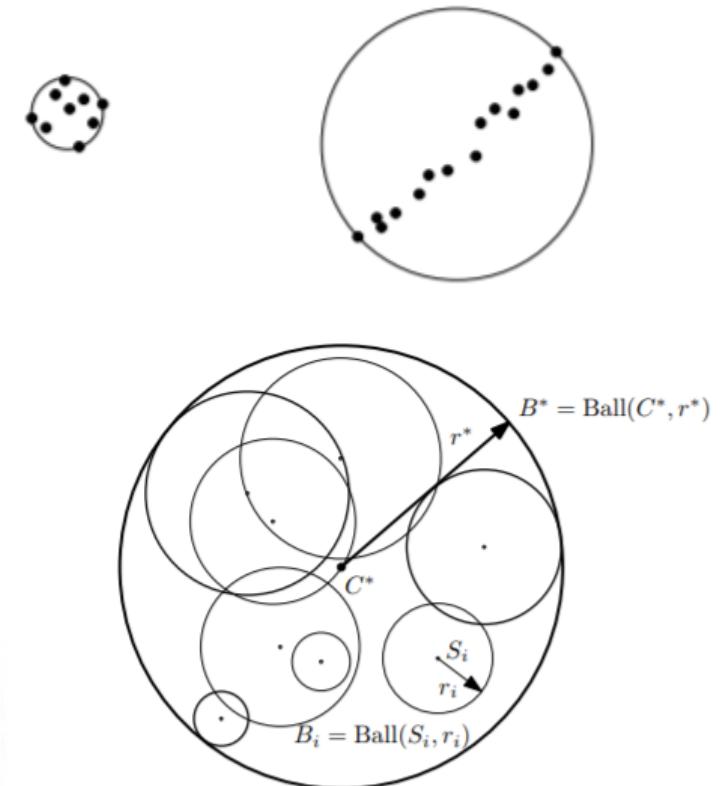
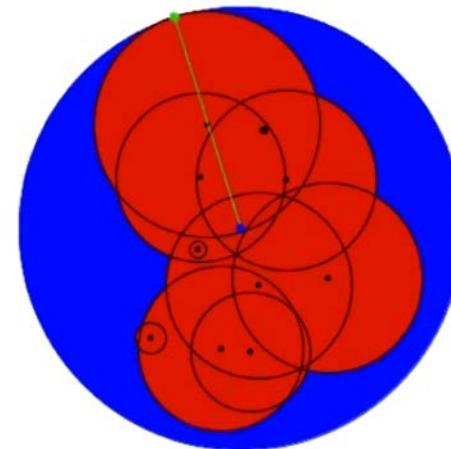
end while



Approximating the smallest enclosing ball

- Iterative algorithm that yields a **core-set**
- Extends to balls, etc.

```
1 Bădoiu -Clarkson( $\mathcal{S}, \epsilon$ );  
2 ▷ Compute a  $(1 + \epsilon)$ -approximation of the smallest enclosing ball ▷  
3 ▷ Return the circumcenter of a small enclosing ball in  $O(\frac{dn}{\epsilon^2})$  time ▷  
4  $C = S_1$ ;  
5   for  $i = 1$  to  $\lceil \frac{1}{\epsilon^2} \rceil$  do  
6     ▷ The core-set is the collection of furthest points ▷  
7     ▷ Furthest point is  $F_i = S_j$  ▷  
8      $j = \operatorname{argmax}_{i=1}^n \|CS_i\|$ ;  
9      $C = C + \frac{1}{i+1}CS_j$ ;  
9 return  $C$ ;
```

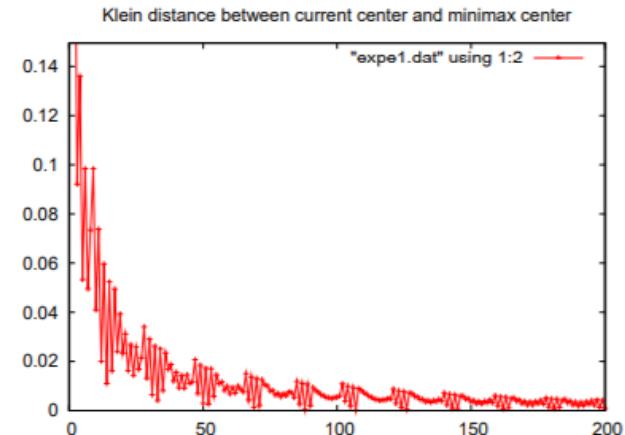


Riemannian minimum enclosing ball

$a \#_t^M b$: point $\gamma(t)$ on the geodesic line segment $[ab]$ wrt M .

Algorithm GeoA

```
c1 ← choose randomly a point in  $\mathcal{P}$ ;  
for  $i = 2$  to  $l$  do  
    // farthest point from  $c_i$   
     $s_i \leftarrow \arg \max_{j=1}^n \rho(c_i, p_j);$   
    // update the center: walk on the geodesic line  
    // segment  $[c_i, p_{s_i}]$   
     $c_{i+1} \leftarrow c_i \#_{\frac{1}{i+1}}^M p_{s_i};$   
end  
// Return the SEB approximation  
return Ball( $c_l, r_l = \rho(c_l, \mathcal{P})$ );
```



Hyperbolic geometry:

$$\rho(p, q) = \operatorname{arccosh} \frac{1 - p^\top q}{\sqrt{(1 - p^\top p)(1 - q^\top q)}}$$

$$T_p(T_{-p}(p) \#_\alpha T_{-p}(q)) = p \#_\alpha q$$

$$T_p(x) = \frac{(1 - \|p\|^2)x + (\|x\|^2 + 2\langle x, p \rangle + 1)p}{\|p\|^2\|x\|^2 + 2\langle x, p \rangle + 1}$$

Positive-definite matrices:

$$\rho(P, Q) = \|\log(P^{-1}Q)\|_F = \sqrt{\sum_i \log^2 \lambda_i}$$

$$\gamma_t(P, Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \right)^t P^{\frac{1}{2}}$$

f-divergence between isotropic Gaussians: monotic increasing function of Mahalanobis Smallest enclosing ball same for all f-divergences...

First, we consider the problem of divergence between two n -dimensional normal distributions with different mean vectors but the same variance matrix. Let these be $N(\mu_i, \Sigma)$, $i = 1, 2$. Mahalanobis's generalized distance is α^2 , where

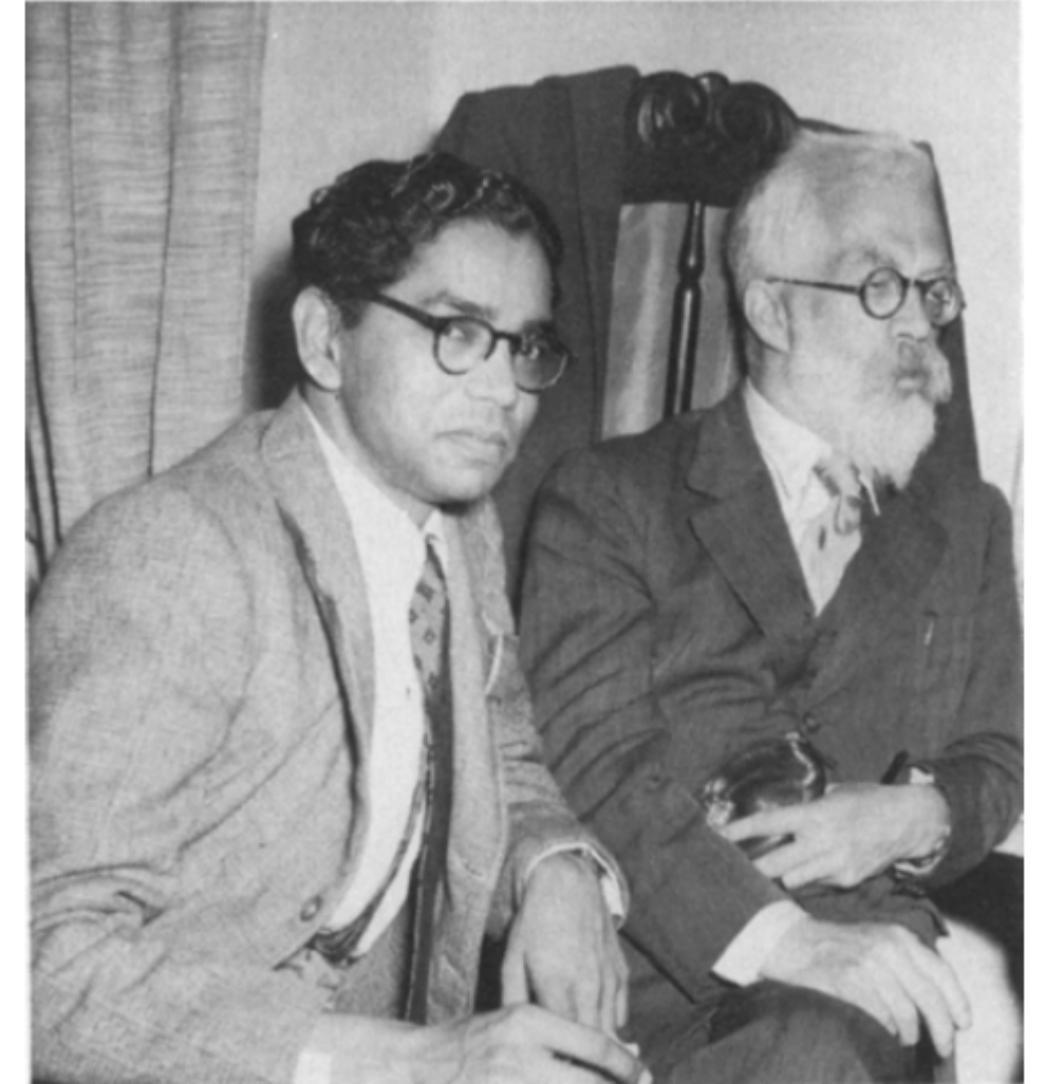
$$\alpha^2 = (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1).$$

α is a metric and a generally accepted measure of distance between the two distributions.

Now every coefficient in the class we are considering is an increasing function of α . This is easily demonstrated by considering the transformation

$$y = (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1) / \alpha$$

and so reducing the problem to that of the divergence of a $N(\alpha, 1)$ distribution from a $N(0, 1)$. The family $\{N(\alpha, 1) : \alpha \geq 0\}$ of distributions of y has monotonic increasing likelihood-ratio in y and it follows from Theorem 2 that if f is increasing and C convex then $f[E^*\{C(\phi)\}]$ is an increasing function of α .



C. R. Rao with Sir R. Fisher in 1956

STATISTICAL DATA ANALYSIS AND INFERENCE edited by Yadolah DODGE, 1989

Other differential metrics for parametric probability families

- Rao's quadratic entropy

$$Q(P) = \int K(x, y)dP(x)dP(y)$$

- Conditionally negative definite kernel

$$\begin{aligned} \sum_1^n \sum_1^n K(x_i, x_j) a_i a_j &\leq 0, & \text{for all } x_1, \dots, x_n \in \mathcal{X} \\ a_1 + \dots + a_n &= 0 \end{aligned}$$

Jensen-Shannon divergence: $D_Q(P_1 : P_2) = Q\left(\frac{P_1+P_2}{2}\right) - \frac{1}{2}Q(P_1) - \frac{1}{2}Q(P_2)$

Metric distance property: $\sqrt{D_Q(P_1 : P_2)}$

Rao, C.R. (1987). Differential metrics in probability spaces, in Differential Geometry in Statistical Inference, S.-I. Amari et al. Eds., IMS Lecture Notes and Monographs Series
Rao, C. R. "Quadratic entropy and analysis of diversity." *Sankhya A* 72.1 (2010): 70-80.

Summary

- By using the Fisher information matrix of a regular parametric model as the Riemannian metric tensor (= **information metric**), we get a Riemannian manifold for the probability model
- Statistical invariance by a 1-to-1 transformation of the sample space x , and by reparameterization of the parameter space
- The Fisher-Rao distance is the Riemannian metric distance: geodesic distance
- Difficult to calculate/approximate, even for the multivariate normal family

Berkane, Maia, Kevin Oden, and Peter M. Bentler. "Geodesic estimation in elliptical distributions." *Journal of Multivariate Analysis* 63.1 (1997): 35-46

Interview with Professor Calyampudi Radhakrishna Rao

1 DECEMBER 2016

4,635 VIEWS

NO COMMENT

Frank Nielsen

C. R. Rao has contributed to facets of modern statistics such as differential-geometric methods in statistics, score test, quadratic entropy, orthogonal arrays, multivariate analysis, and generalized inverse of a matrix (singular or not) and its applications. Frank Nielson—a professor of computer science at Ecole Polytechnique, Palaiseau, France, and a senior researcher at Sony Computer Science Laboratories, Inc.—interviewed Rao this past year to learn more about his life and work. What follows is what he discovered.



Can you briefly tell us about your family and education in India?

I was born on September 10, 1920, in a small town in Madras Presidency (under British rule known as Hadagali). I am the eighth child out of 10 (four girls and six boys) to my parents.

One of my sisters was a Telugu (my mother tongue) poet. Another sister was a business woman selling cars imported from Britain. The seventh child was a boy who had phenomenal memory. He received a gold medal on his anatomy exam for remembering the names of all the bones and other organs of the

C.R. Rao

Dualistic structures of information geometry

Frank Nielsen

Sony Computer Science Laboratories, Inc



An elementary introduction to information geometry

<https://arxiv.org/abs/1808.08271>

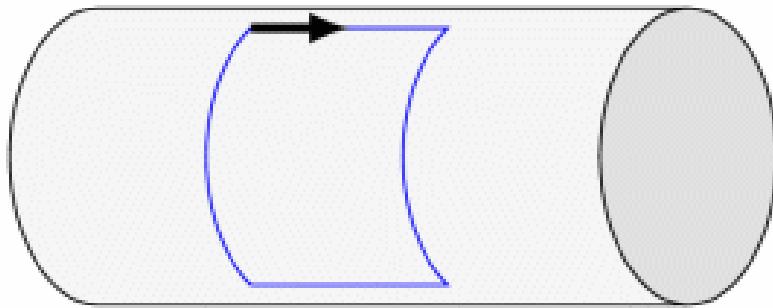
Covariant derivative ∇

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M)$$

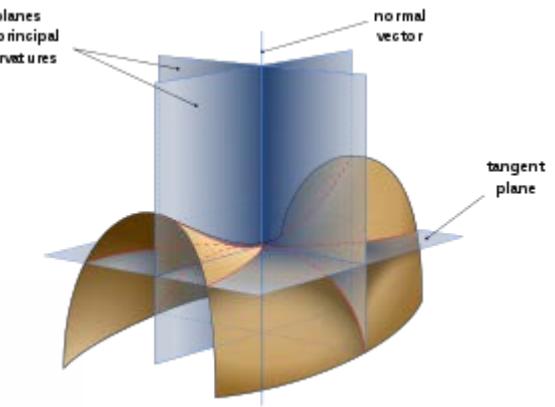
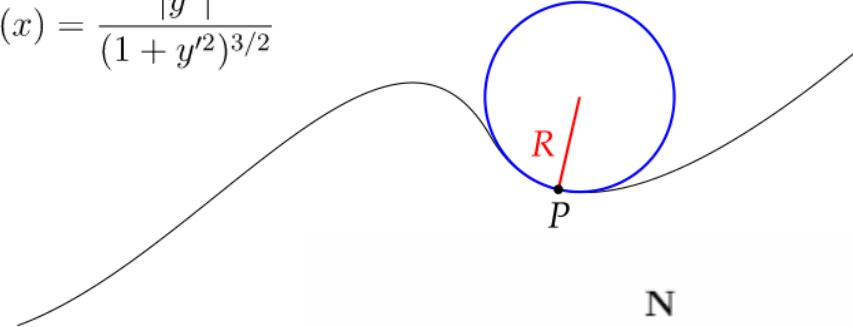
- calculate differentials of a vector field Y with respect to another vector field X : Namely, the covariant derivative $\nabla_X Y := \nabla(X, Y)$
- Defined by prescribing a dimension cubic number of smooth functions: The Christoffel symbols $\Gamma_{ij}^k = \Gamma_{ij}^k(p)$
- In local coordinates of a chart, we have $\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k$
- The k -th component $(\nabla_X Y)^k$ of the covariant derivative of vector field Y with respect to vector field X is given by

$$(\nabla_X Y)^k \stackrel{\Sigma}{=} X^i (\nabla_i Y)^k \stackrel{\Sigma}{=} X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma_{ij}^k Y^j \right)$$

Curvature of ∇



$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$



Cylinder is flat
Parallel transport is
independent of path

Sphere has constant curvature
Parallel transport is path-dependent

Curvature/torsion of an affine connection ∇

parallel transport “twists” vectors.



- Curvature of a tensor

$$R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X, Y]} Z$$

$$R(\partial_j, \partial_k) \partial_i \stackrel{\Sigma}{=} R^l_{jki} \partial_l \quad (\text{in local coordinates})$$

- Connection is said **flat** when $R=0$
- Symmetric connection: $\nabla_X Y - \nabla_Y X = [X, Y]$
In local coordinates: $\Gamma_{ij}^k = \Gamma_{ji}^k$
- (1,2)-torsion tensor: $T(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y]$

Conjugate connections or dual connections (∇, ∇^*)

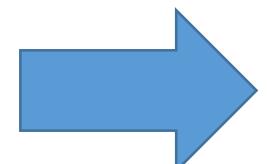
- For any three smooth vectors fields X, Y, Z of manifold M , conjugate affine torsion-free connection ∇^* of ∇ with respect to the metric tensor g

$$X\langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X^* Z \rangle, \quad \forall X, Y, Z \in \mathcal{X}(M)$$

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

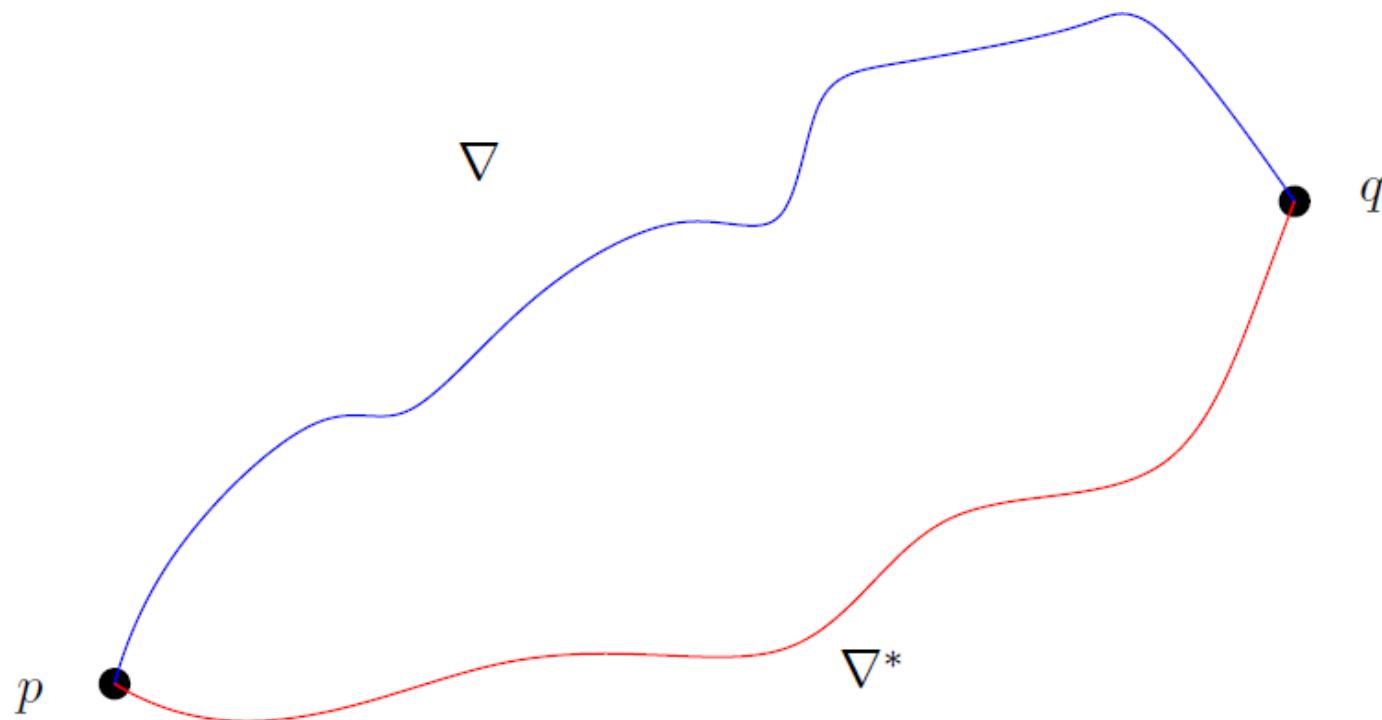
- NB: check that the right-hand-side is a scalar and that the left-hand-side is a directional derivative of a real-valued function, that is also a scalar. **Unique dual torsion-free affine connection ∇^***

- Involution: $(\nabla^*)^* = \nabla$



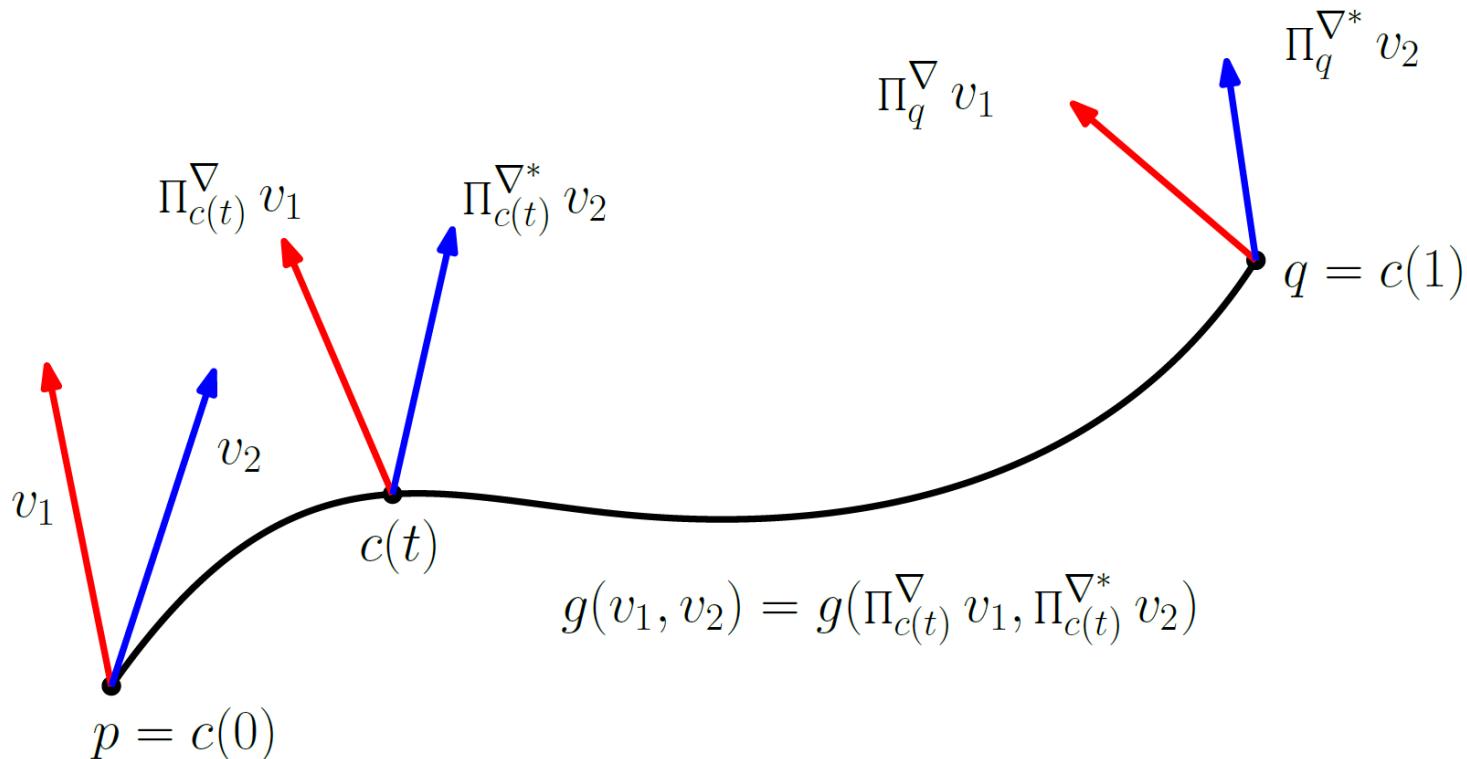
$$(M, g, \nabla, \nabla^*)$$

Dual ∇ -geodesic and ∇^* -geodesic



Property: Dual parallel transport of vectors preserves the metric

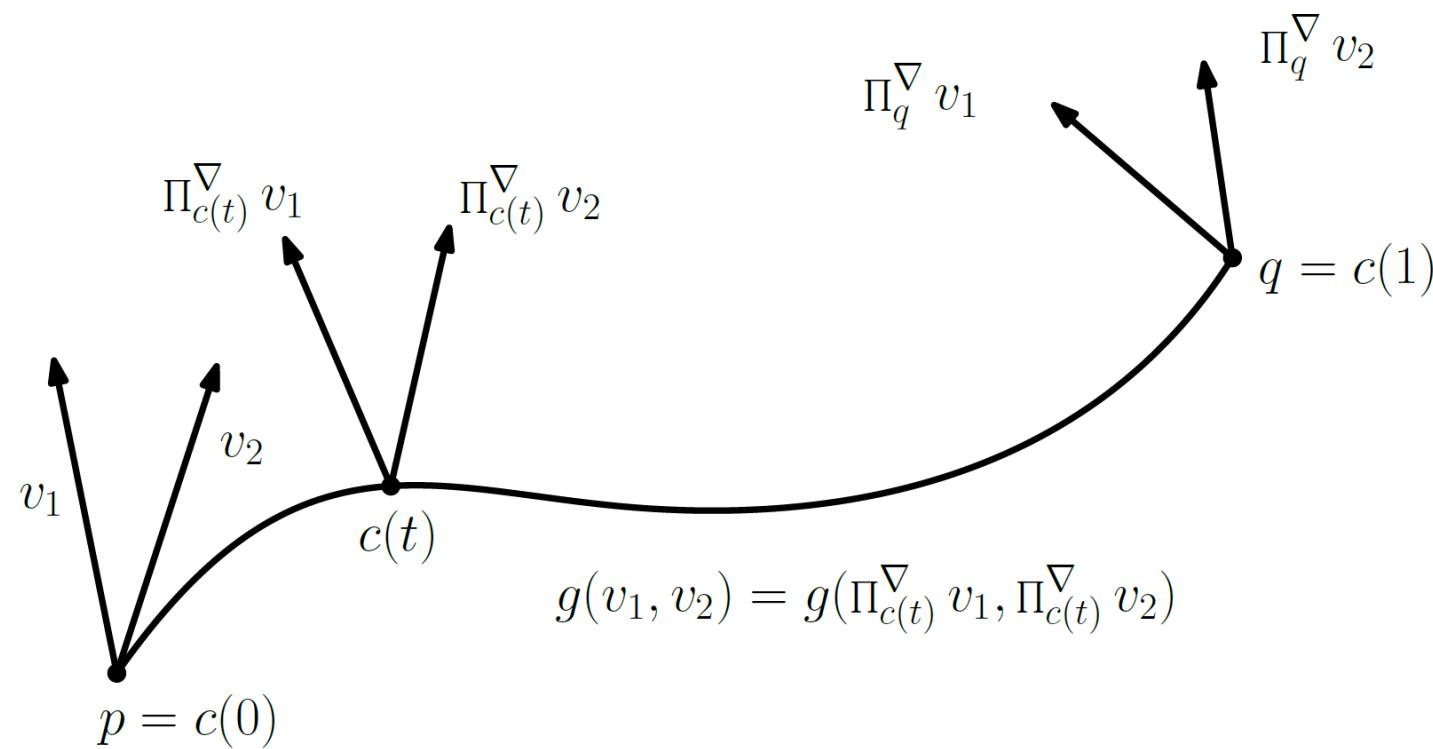
$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \rightarrow c(t)}^{\nabla} u, \prod_{c(0) \rightarrow c(t)}^{\nabla^*} v \right\rangle_{c(t)}$$



Metric Levi-Civita connection from averaging dual connections

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

$$\bar{\nabla} = \text{LC } \nabla \quad \bar{\Gamma}$$



Statistical manifolds: Structure

(M, g, C)

- Apply also to non-statistical contexts! Dualistic structure with metric tensor g and cubic tensor C



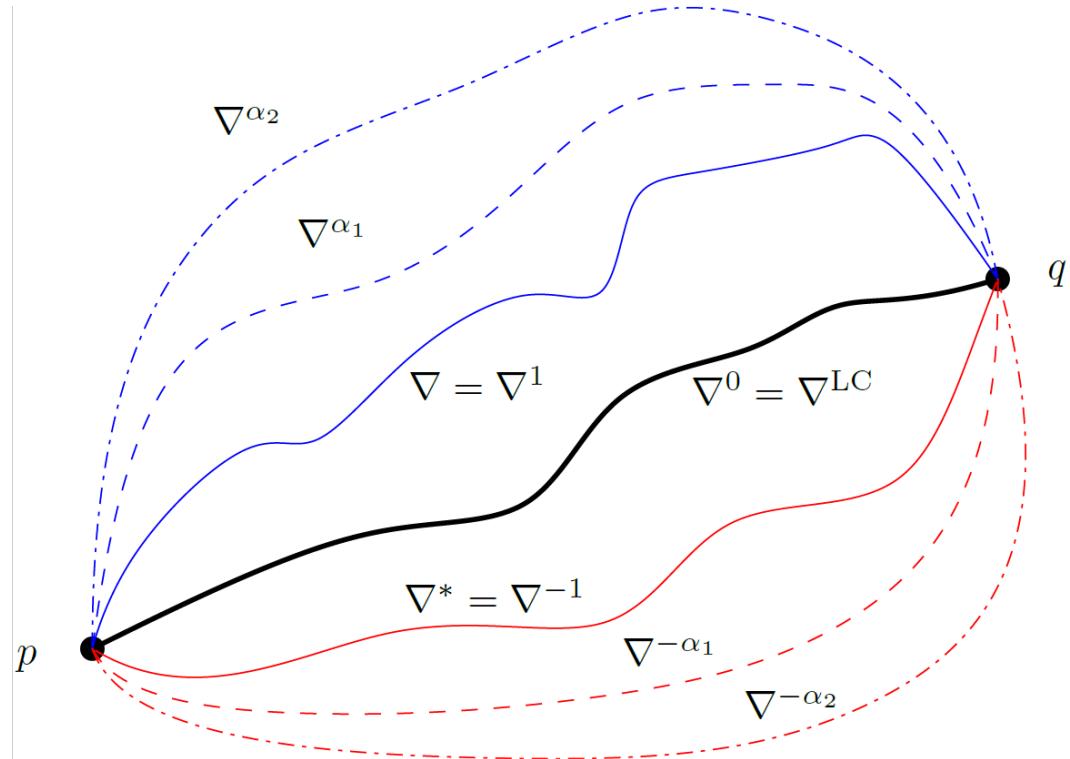
Steffen Lauritzen
(1987)

$$C(X, Y, Z) := \langle \nabla_X Y - \nabla_X^* Y, Z \rangle$$

$$C_{ijk} := \Gamma_{ij}^k - \Gamma_{ij}^{*k} \quad (\text{local coordinates})$$

In a local basis: $C_{ijk} = C(\partial_i, \partial_j, \partial_k) = \langle \nabla_{\partial_i} \partial_j - \nabla_{\partial_i}^* \partial_j, \partial_k \rangle$

From a statistical manifold to a 1-family of structures



$$\Gamma_{ij,k}^\alpha = \Gamma_{ij,k}^0 - \frac{\alpha}{2} C_{ij,k},$$
$$\Gamma_{ij,k}^{-\alpha} = \Gamma_{ij,k}^0 + \frac{\alpha}{2} C_{ij,k},$$
$$\Gamma_{ij,k}^\alpha = \frac{1+\alpha}{2} \Gamma_{ij,k} + \frac{1-\alpha}{2} \Gamma_{ij,k}^*$$

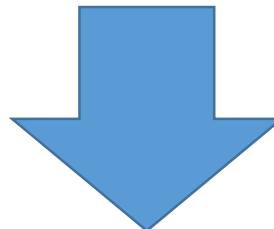
The α -connections $g(\nabla_X^\alpha Y, Z) = g(\nabla_X^{\text{LC}} Y, Z) + \frac{\alpha}{2} C(X, Y, Z), \forall X, Y, Z \in \mathfrak{X}(M)$



$(M, g, \nabla^{-\alpha}, \nabla^\alpha = (\nabla^{-\alpha})^*)$

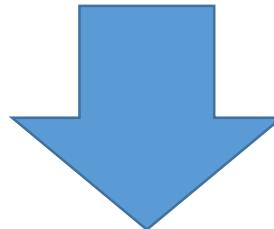
The fundamental theorem of information geometry

Theorem: If ∇ has constant curvature κ then its conjugate connection ∇^* has necessarily the same constant curvature κ



$\kappa=0$

A manifold $(M, g, \nabla^{-\alpha}, \nabla^\alpha)$ is ∇^α -flat if and only if it is $\nabla^{-\alpha}$ -flat.



$\kappa=0$

A manifold (M, g, ∇, ∇^*) is ∇ -flat if and only if it is ∇^* -flat

How to get initial dual connections?

- Historically, Amari's defined the statistical **expected exponential and mixture connections**, and then the **expected α -connections**
- Then Eguchi showed how to define dual connections from smooth parameter distances called divergences (originally, **contrast functions**). From that, we get a 1-family of **α -connections**

Definition of a parameter divergence

Definition (Divergence) A divergence $D : M \times M \rightarrow [0, \infty)$ on a manifold M with respect to a local chart $\Theta \subset \mathbb{R}^D$ is a C^3 -function satisfying the following properties:

1. $D(\theta : \theta') \geq 0$ for all $\theta, \theta' \in \Theta$ with equality holding iff $\theta = \theta'$ (law of the indiscernibles),
2. $\partial_{i,\cdot} D(\theta : \theta')|_{\theta=\theta'} = \partial_{\cdot,j} D(\theta : \theta')|_{\theta=\theta'} = 0$ for all $i, j \in [D]$,
3. $-\partial_{\cdot,i} \partial_{\cdot,j} D(\theta : \theta')|_{\theta=\theta'}$ is positive-definite.

$$\partial_{i,\cdot} f(x, y) = \frac{\partial}{\partial x^i} f(x, y), \partial_{\cdot,j} f(x, y) = \frac{\partial}{\partial y^j} f(x, y), \partial_{ij,k} f(x, y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial}{\partial y^k} f(x, y), \text{etc.}$$



Statistical divergence (deviance) like the Kullback-Leibler divergence
versus
Parameter divergence as a synonym of a contrast function

Statistical manifolds from divergences

- Reverse/dual parameter divergence

$$D^*(\theta : \theta') := D(\theta' : \theta) \quad (D^*)^* = D$$

- Statistical manifold structures:

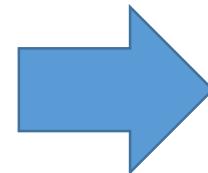
$$(M, {}^D g, {}^D \nabla, {}^{D^*} \nabla) \quad (M, {}^D g, {}^D C)$$

$${}^D g := -\partial_{i,j} D(\theta : \theta')|_{\theta=\theta'} = {}^{D^*} g,$$

$${}^D C_{ijk} = {}^{D^*} \Gamma_{ijk} - {}^D \Gamma_{ijk}$$

$${}^D \Gamma_{ijk} := -\partial_{ij,k} D(\theta : \theta')|_{\theta=\theta'},$$

$${}^{D^*} \Gamma_{ijk} := -\partial_{k,ij} D(\theta : \theta')|_{\theta=\theta'}.$$



$${}^D \nabla^* = {}^{D^*} \nabla$$

$$\rightarrow \left\{ (M, {}^D g, {}^D C^\alpha) \equiv (M, {}^D g, {}^D \nabla^{-\alpha}, ({}^D \nabla^{-\alpha})^* = {}^D \nabla^\alpha) \right\}_{\alpha \in \mathbb{R}}$$

Statistical manifolds from Bregman divergences

Bregman divergence (1967):

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$$

$$(M, F) \equiv (M, {}^{B_F} g, {}^{B_F} \nabla, {}^{B_F} \nabla^* = {}^{B_{F^*}} \nabla)$$

Dual Bregman divergence and Legendre-Fenchel transformation

$$B_F^*(\theta : \theta') = B_F(\theta' : \theta) = B_{F^*}(\eta' : \eta)$$

$$\eta = \nabla F(\theta), \theta = \nabla F(\theta)$$

Expected α -geometry for a parametric model

$$\mathcal{P} := \{p_\theta(x)\}_{\theta \in \Theta} \quad \xrightarrow{\hspace{1cm}} \quad \{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\}_{\alpha \in \mathbb{R}}$$

- Use Fisher information metric
- Define the expected α -connections:
- Amari-Chentsov cubic tensor

$$C_{ijk} := E_\theta [\partial_i l \partial_j l \partial_k l]$$
$$l(\theta; x) := \log L(\theta; x) = \log p_\theta(x)$$

$$\begin{aligned} {}_{\mathcal{P}}\Gamma^\alpha{}_{ij,k}(\theta) &:= E_\theta [\partial_i \partial_j l \partial_k l] + \frac{1-\alpha}{2} C_{ijk}(\theta), \\ &= E_\theta \left[\left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right]. \end{aligned}$$

Exponential family and mixture family

Example 1 (FIM of an exponential family \mathcal{E}) An exponential family [41] \mathcal{E} is defined for a sufficient statistic vector $t(x) = (t_1(x), \dots, t_D(x))$, and an auxiliary carrier measure $k(x)$ by the following canonical density:

$$\mathcal{E} = \left\{ p_\theta(x) = \exp \left(\sum_{i=1}^D t_i(x) \theta_i - F(\theta) + k(x) \right) \text{ such that } \theta \in \Theta \right\},$$

where F is the strictly convex cumulant function. Exponential families include the Gaussian family, the Gamma and Beta families, the probability simplex Δ , etc. The FIM of an exponential family is given by:

$$\varepsilon I(\theta) = \text{Cov}_{X \sim p_\theta(x)}[t(x)] = \nabla^2 F(\theta) = (\nabla^2 F^*(\eta))^{-1} \succ 0.$$

Example 2 (FIM of a mixture family \mathcal{M}) A mixture family is defined for $D + 1$ functions F_1, \dots, F_D and C as:

$$\mathcal{M} = \left\{ p_\theta(x) = \sum_{i=1}^D \theta_i F_i(x) + C(x) \text{ such that } \theta \in \Theta \right\},$$

where the functions $\{F_i(x)\}_i$ are linearly independent on the common support \mathcal{X} and satisfying $\int F_i(x) d\mu(x) = 0$. Function C is such that $\int C(x) d\mu(x) = 1$. Mixture families include statistical mixtures with prescribed component distributions and the probability simplex Δ . The FIM of a mixture family is given by:

$$\mathcal{M}I(\theta) = E_{X \sim p_\theta(x)} \left[\frac{F_i(x) F_j(x)}{(p_\theta(x))^2} \right] = \int_{\mathcal{X}} \frac{F_i(x) F_j(x)}{p_\theta(x)} d\mu(x) \succ 0.$$

Exponential e-connection and mixture m-connection: Examples of dually flat connections

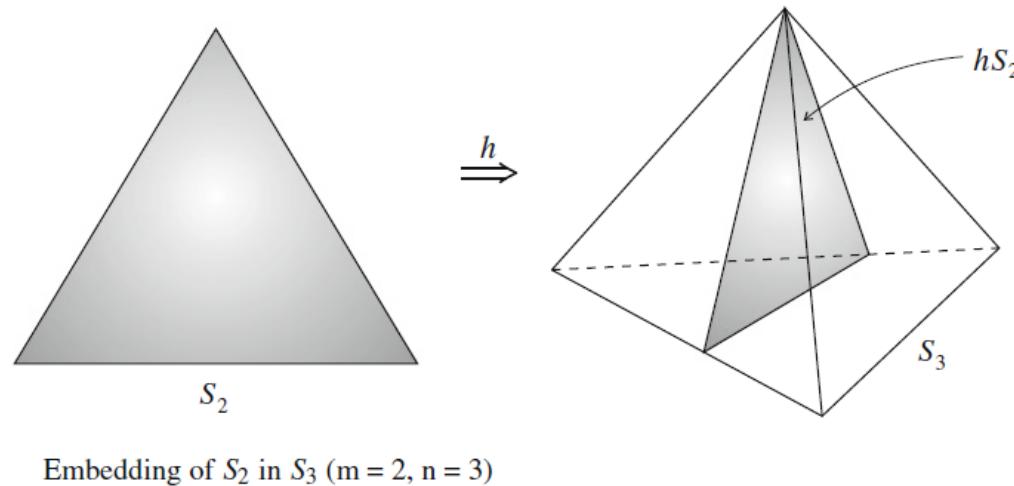
- For an exponential family, the e-connection is flat. Then by using the fundamental theorem of information geometry, we have the dual m-connection flat.
- For a mixture family, the m-connection is flat. Then by using the fundamental theorem of information geometry, we have the dual e-connection flat

Statistical invariance

- Which metric tensor to choose?
- Which dual connections to choose?
- How are statistical divergences related to geometric structures?

Statistical invariance: metric tensor

- The Fisher information metric is the unique invariant metric tensor under **Markov embeddings** up to a scaling constant.

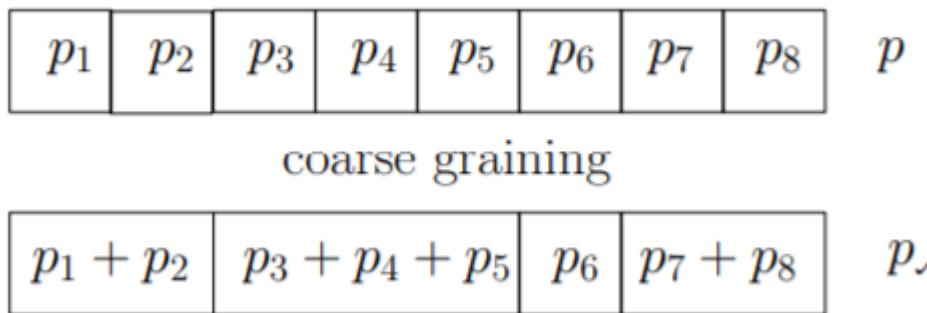


L. Lorne Campbell. An extended Cencov characterization of the information metric. ^{^\wedge}
Proceedings of the American Mathematical Society, 98(1):135–141, 1986.
Hong Van Le. The uniqueness of the Fisher metric as information metric. Annals of the
Institute of Statistical Mathematics, 69(4):879–896, 2017.

Statistical invariance: statistical divergence

- **Information monotonicity** of parameter divergences:

$$D(\theta_{\bar{\mathcal{A}}} : \theta'_{\bar{\mathcal{A}}}) \leq D(\theta : \theta')$$



Markov embeddings, Markov kernels, etc.



Statistical invariance: f-divergences

- **Separable** divergence: A separable divergence is a divergence that can be expressed as the sum of elementary scalar divergences

$$D(\theta_1 : \theta_2) = \sum_i d(\theta_1^i : \theta_2^j)$$

- Squared Euclidean distance is separable but not the Euclidean distance (because of the square root)
- The only invariant and decomposable divergences when $D > 2$ are f-divergences defined for a convex functional generator f :

$$I_f(\theta : \theta') = \sum_{i=1}^D \theta_i f\left(\frac{\theta'_i}{\theta_i}\right) \geq f(1), \quad f(1) = 0$$

Standard invariant f-divergences

- F strictly convex at 1 (for law of the indiscernibles)
- Choose $f(1)=0$ (for lower bound being 0)
- Choose $f'(1)=0$ to fix lambda in equivalent class of generators:

$$f_\lambda(u) = f(u) + \lambda(u - 1)$$

- Expansion of

$$I_f(p : p + dp) = f''(1) \frac{1}{2} dp^\top g(p) dp$$

- Choose $f''(1)=1$ to get **standard f-divergence** with infinitesimal distance expressed using the Fisher information matrix tensor
- **The α -connection for the standard f-divergence corresponds to the expected α -connection for**

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

Summary

- Geometry of parametric families of distributions:
 - Fisher Riemannian geometry
 - α -expected geometry
 - Statistical invariance
- α -geometry from any parameter divergence
- Dually flat geometry for +1/-1-geometry of exponential families or mixture families

Bregman dually flat manifolds and information projections

Frank Nielsen

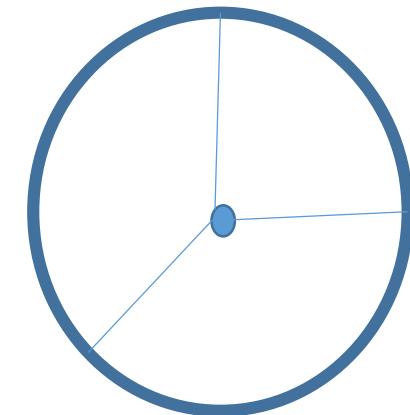
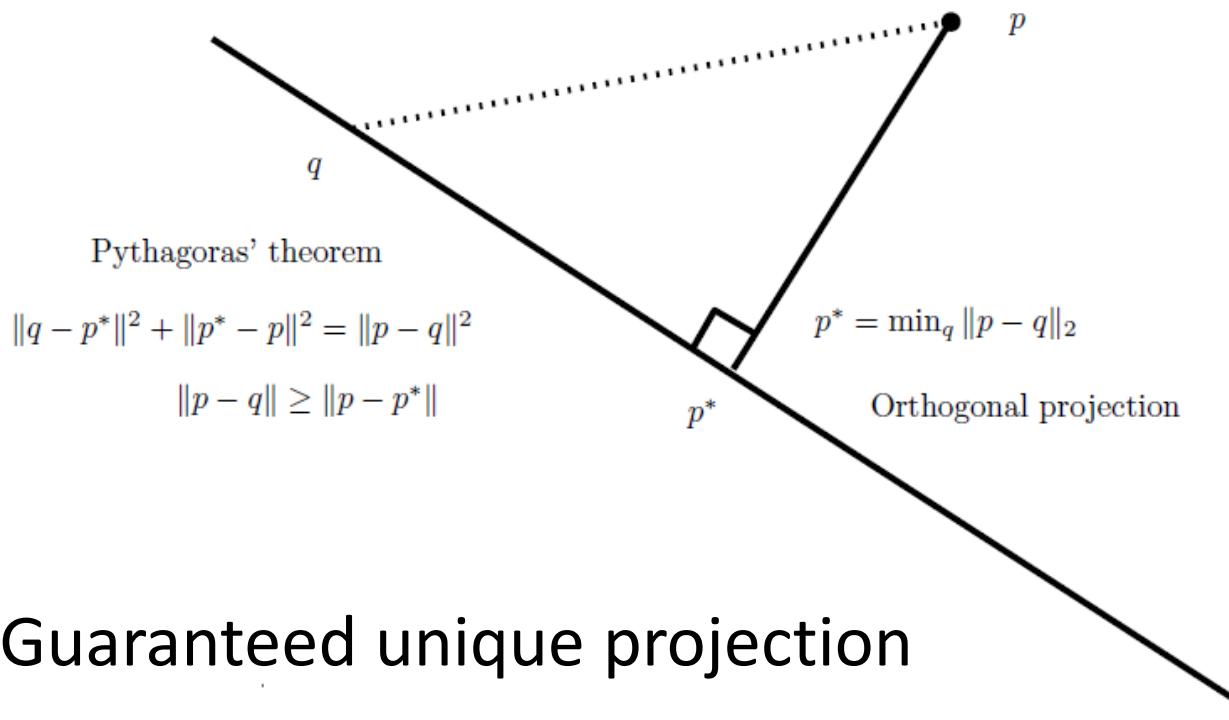


Sony CSL

Projection, orthogonality and Pythagoras' theorem

Recalling Euclidean geometry....

Distance, geodesic, orthogonality, uniqueness of projection



Non-unique projection

Goal: give geometric interpretations of MLE/MaxEnt of KL divergence minimizations

MaxEnt (with prior q)

$$\min_p \text{KL}(p : q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\sum_x p(x)t_i(x) = m_i, \quad \forall i \in \{1, \dots, D\}$$

$$p(x) \geq 0, \quad \forall x \in \{1, \dots, n\}$$

$$\sum_x p(x) = 1$$

Maximum Likelihood Estimate

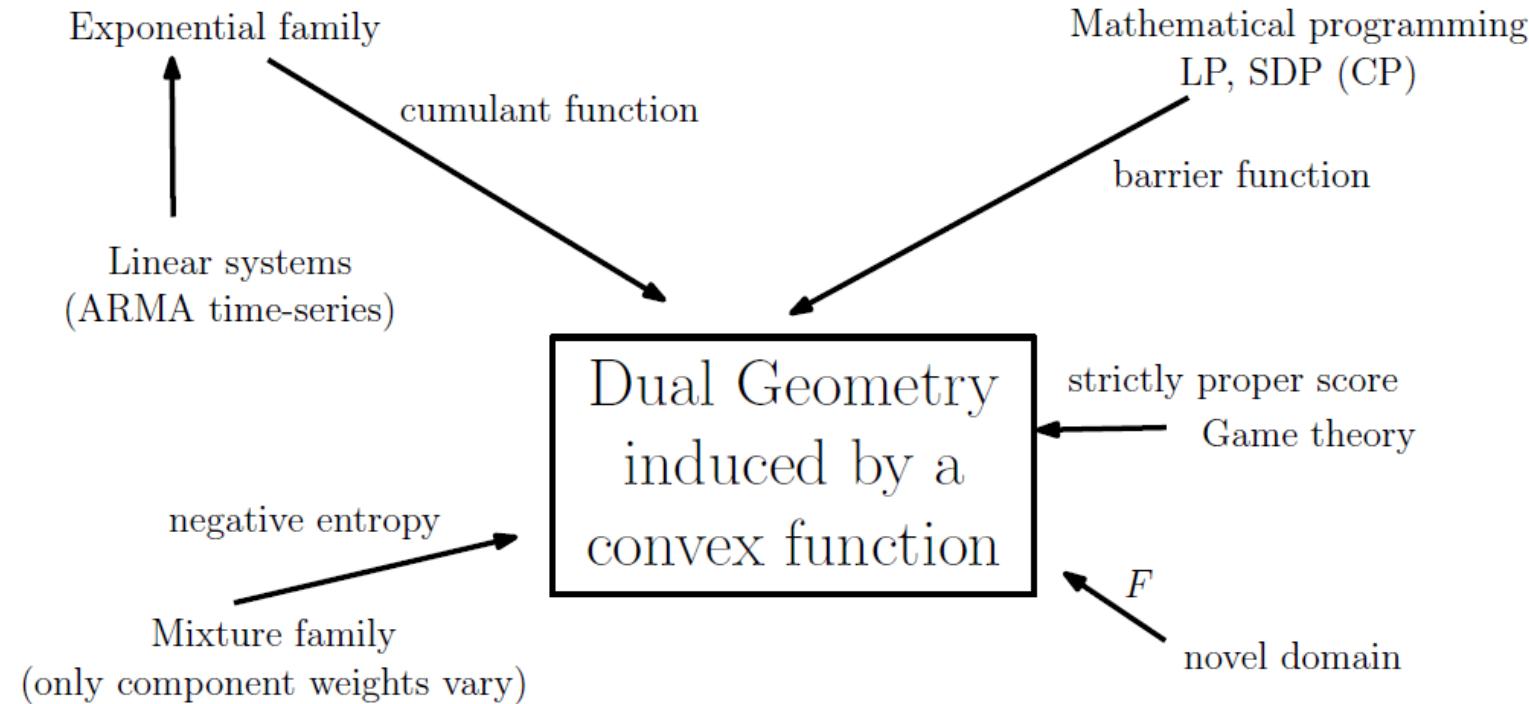
$$\begin{aligned} \min & \quad \text{KL}(p_e(x) : p_\theta(x)) \\ &= \int p_e(x) \log p_e(x) dx - \int p_e(x) \log p_\theta(x) dx \\ &= \min -H(p_e) - \underbrace{E_{p_e} [\log p_\theta(x)]}_{\equiv \max} \end{aligned}$$

$$\begin{aligned} &\equiv \max \frac{1}{n} \sum \delta(x - x_i) \log p_\theta(x) \\ &= \max \frac{1}{n} \sum_i \log p_\theta(x_i) = \text{MLE} \end{aligned}$$

Bregman manifolds in a nutshell

- From any smooth (C3) convex function F , we can build a dualistic information-geometric structure called a **dually flat manifold**.
- Duality emanates from **Legendre-Fenchel conjugation**
- There are **two global (affine) coordinate systems**: primal and dual
- We can associate a canonical divergence to dually flat manifolds: **Bregman divergences** or **Fenchel-Young divergences**
- There are two dual **Pythagoras theorems** (and law of cosines)
(Give a sufficient case where dual **information projections** are unique)
- Very well-suited to **computational geometry** (**Voronoi** and **proximity queries**)

Dually flat geometry from a convex function



Historically, the dualistic structure of information geometry was called By Lauritzen (1987) a **statistical manifold**.

But the structure can be used in non-statistical contexts.

Not necessarily related to statistical models, but can find a regular statistical models

Vân Lê, Hồng. "Statistical manifolds are statistical models." *Journal of Geometry* 84.1-2 (2006)

Dually flat manifold construction

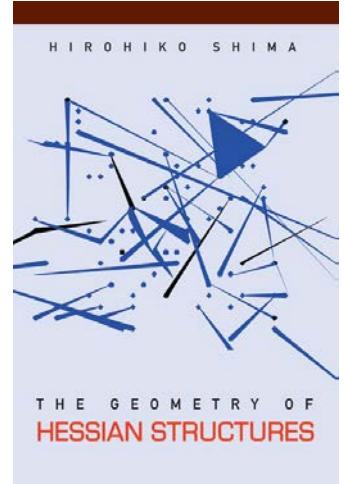
- A global coordinate system (single chart) θ
- Metric tensor g is the **Hessian of the potential function**:

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

- Geodesic of the connection ∇ are straight lines in the θ -coordinate system since

$${}^F \Gamma_{ijk}(\theta) = 0$$

- Bregman manifold is a special case of Hessian manifolds where the Hessian is the Hessian of a global function



Dually flat manifold construction

Duality emanates from the Legendre-Fenchel convex duality:

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

- Dual Riemannian metric tensor ${}^F g^*$
- Expressed in the dual coordinate system η : ${}^F g^* = \nabla^2 F^*(\eta)$
- Coordinate-free notation: ${}^F g^* = {}^{F^*} g$
- Geodesic of the connection ∇^* are straight lines since

$${}^F \Gamma^{*ijk}(\eta) = 0$$

Metric tensor using covariant/contravariant notations

2-covariant metric tensor in local coordinates:

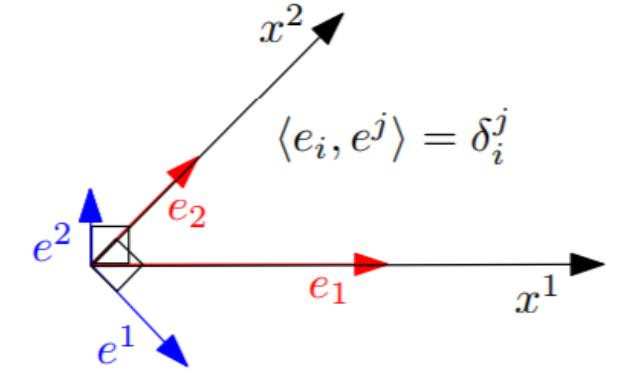
$$g_{ij}(\theta) = \nabla^2 F(\theta)$$

Dual metric tensor in local coordinates:

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

Crouzeix's identity of Hessians of convex conjugates:

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



α -geometry

$$(M, g, \nabla^{-\alpha}, \nabla^\alpha)$$

Cubic tensor:

$${}^F C_{ijk} = {}^F \Gamma_{ijk} - {}^F \Gamma_{ijk}^*$$

$${}^F C_{ijk} = \partial_i \partial_j \partial_k F(\theta)$$

$$\nabla^1 = \nabla \quad \nabla^{-1} = \nabla^* \quad \Gamma^0 = \Gamma^{\text{LC}}$$

Get the alpha-connections:

$$\Gamma_{ij,k}^\alpha = \Gamma_{ij,k}^0 - \frac{\alpha}{2} C_{ij,k}$$

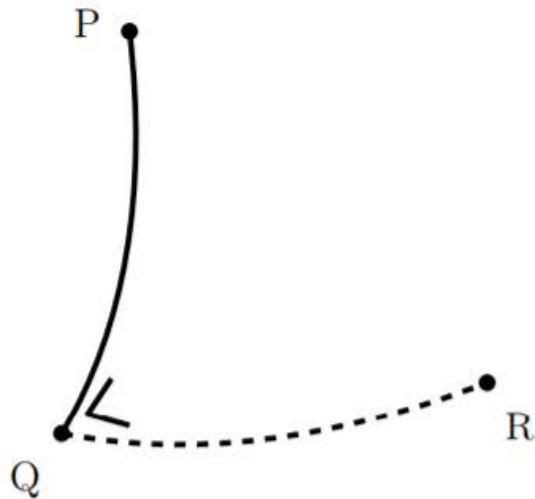
$$\Gamma_{ij,k}^{-\alpha} = \Gamma_{ij,k}^0 + \frac{\alpha}{2} C_{ij,k}$$

Dually flat manifolds

- F C3 smooth and convex function

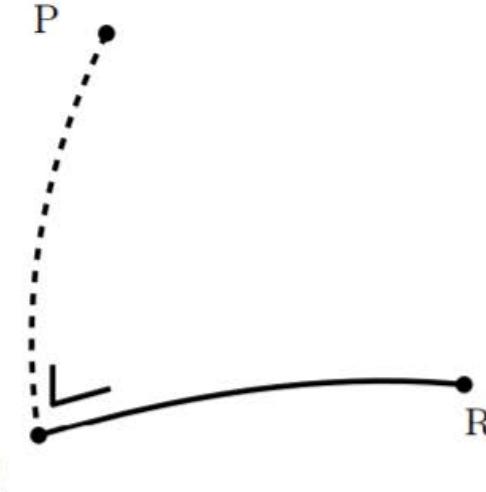
Dual Pythagoras' theorem

$$\gamma^*(P, Q) \perp_F \gamma(Q, R)$$



$$D(P : R) = D(P : Q) + D(Q : R)$$

$$\gamma(P, Q) \perp_F \gamma^*(Q, R)$$



$$D^*(P : R) = D^*(P : Q) + D^*(Q : R)$$

$$B_F(\theta(P) : \theta(R)) = B_F(\theta(P) : \theta(Q)) + B_F(\theta(Q) : \theta(R))$$

$$B_{F^*}(\eta(P) : \eta(R)) = B_{F^*}(\eta(P) : \eta(Q)) + B_{F^*}(\eta(Q) : \eta(R))$$

$$\gamma^*(P, Q) \perp \gamma(Q, R) \Leftrightarrow (\eta(P) - \eta(Q))^\top (\theta(Q) - \theta(R)) = (\eta_i(P) - \eta_i(Q))(\theta_i(Q) - \theta_i(R)) = 0$$

$$\gamma(P, Q) \perp \gamma^*(Q, R) \Leftrightarrow (\theta(P) - \theta(Q))^\top (\eta(Q) - \eta(R)) = (\theta_i(P) - \theta_i(Q))^\top (\eta_i(Q) - \eta_i(R)) = 0$$

Uniqueness of projections in dually flat spaces

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

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

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

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

Geometry of KL for exponential families or mixture families is dually flat

e-projection q_e^* is **unique** if $M \subseteq S$ is *m-flat* and minimizes the *m*-divergence $\text{KL}(\boxed{q} : p)$ (left-sided argument):

$$\text{e-projection: } q_e^* = \arg \min_q \text{KL}(\boxed{q} : p)$$

m-projection q_m^* is **unique** if $M \subseteq S$ is *e-flat* and minimizes the *e*-divergence $\text{KL}(p : \boxed{q})$ (right-sided argument):

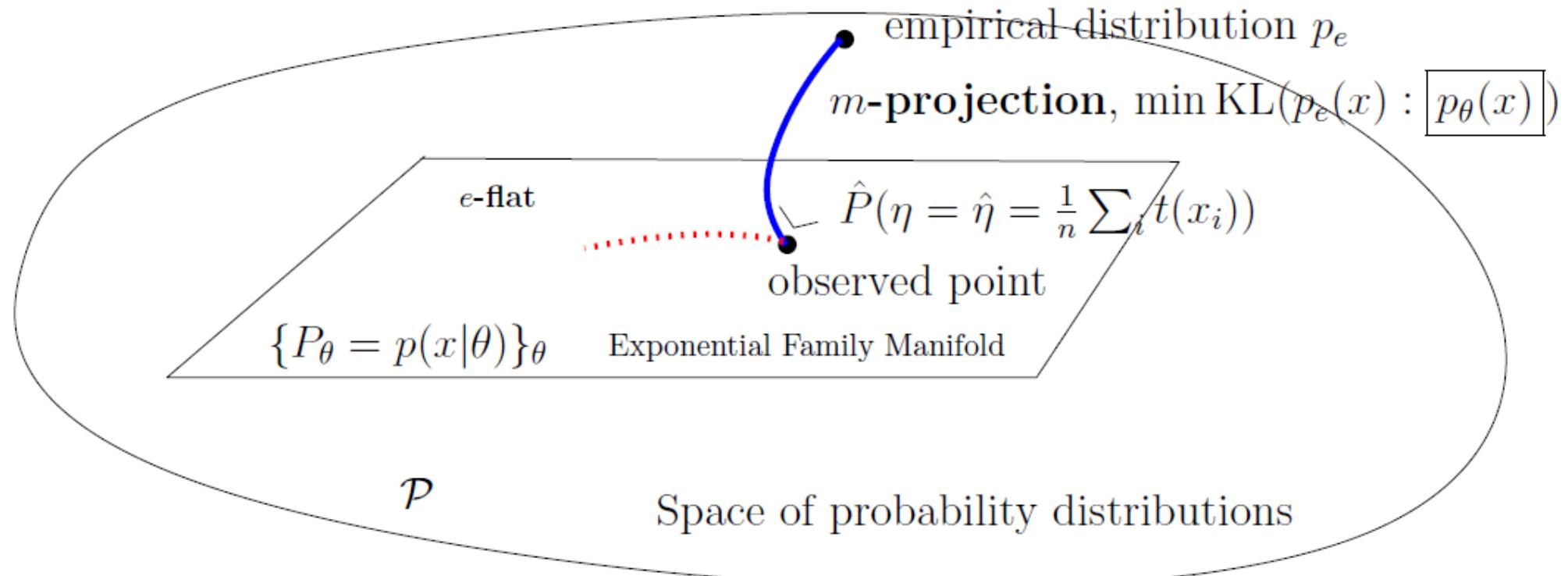
$$\text{m-projection: } q_m^* = \arg \min_q \text{KL}(p : \boxed{q})$$

I–projection, rI–projection, KL–projection

MLE for an exponential family as an information projection

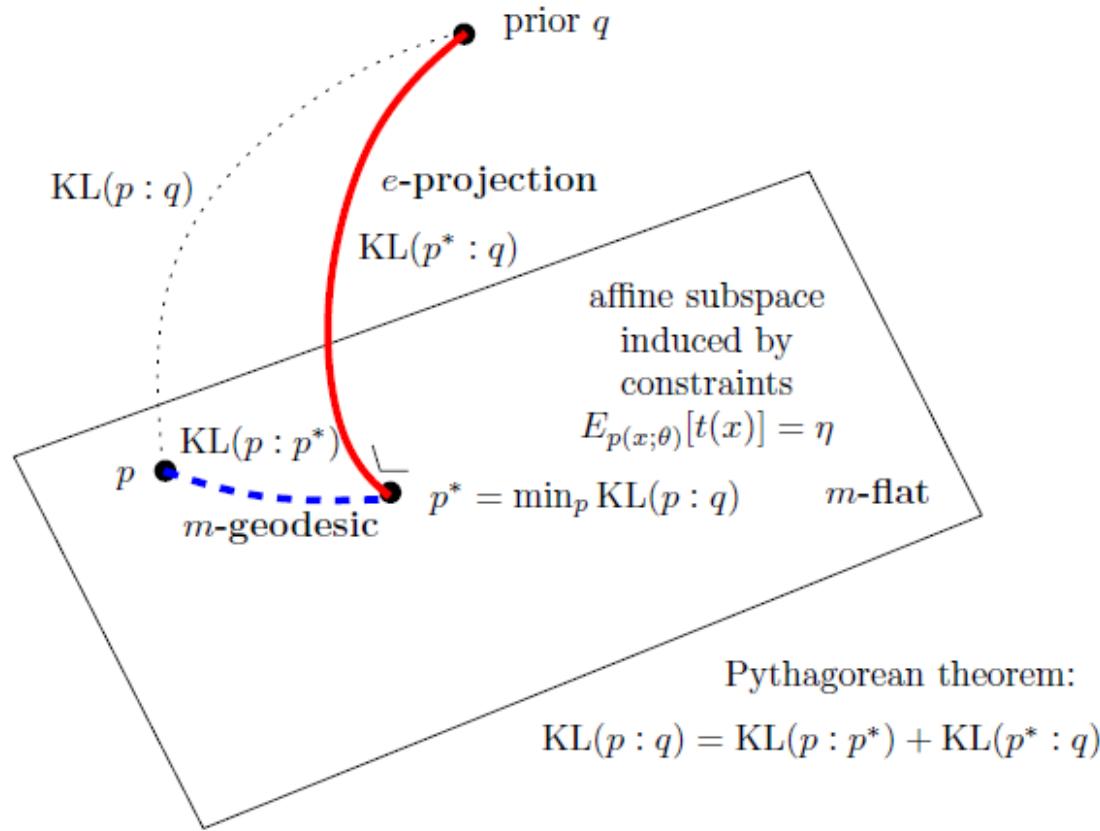
Exponential Family Manifold (EFM) is **e-flat**

Observed point



MaxEnt as an information projection

- MaxEnt linear constraints define a **m-flat**



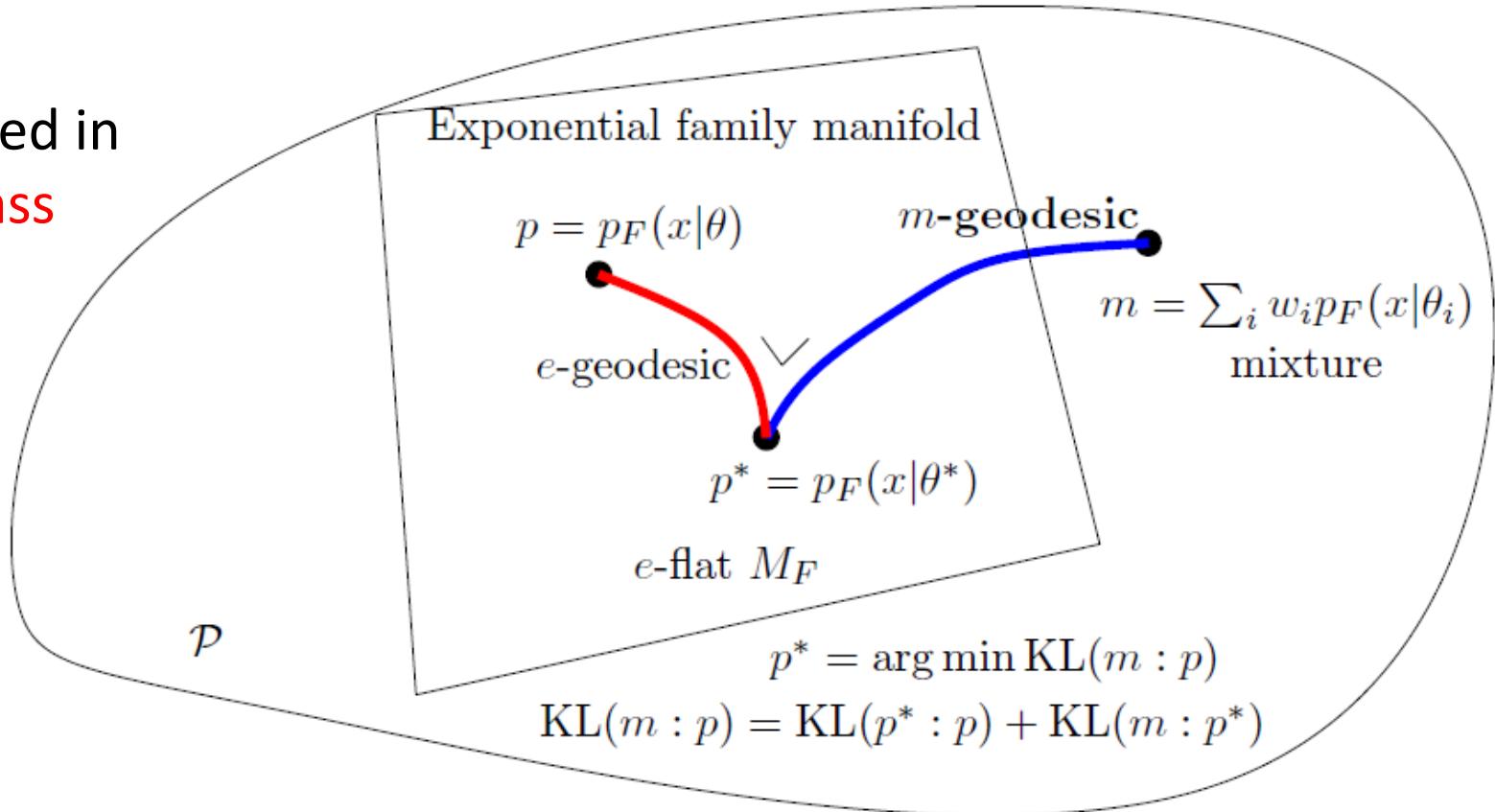
Pythagoras' theorem (Fisher orthogonality) $\gamma_m(p, p^*) \perp_{\text{FIM}} \gamma_e(p^*, q)$

Simplifying a mixture model to a single component

KL right-sided minimization problem

Best single distribution is expressed in
 η -coordinates as the **center of mass**

$$\bar{\eta} = \sum_i w_i \eta_i$$



Learning mixtures by simplifying kernel density estimators, 2012

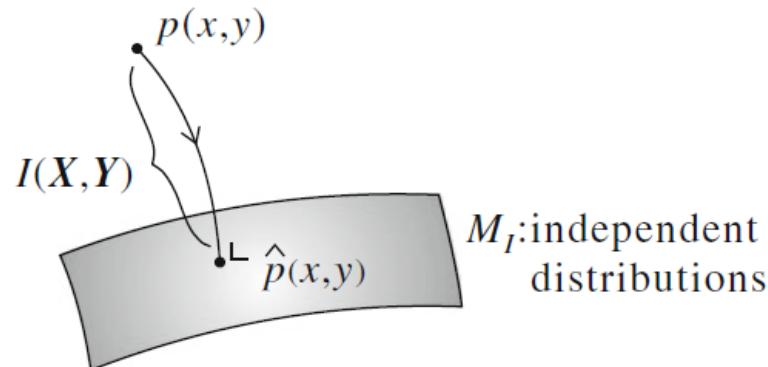
Model centroids for the simplification of kernel density estimators, ICASSP 2012

Information projection: Closest independent distribution

$$p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$$

- Independence of random variables X and Y: KL between joint (X,Y) and product of marginals

$$KL[p(x, y) : \hat{p}(x, y)] = \int p(x, y) \log \frac{p(x, y)}{\hat{p}(x, y)} dx dy$$

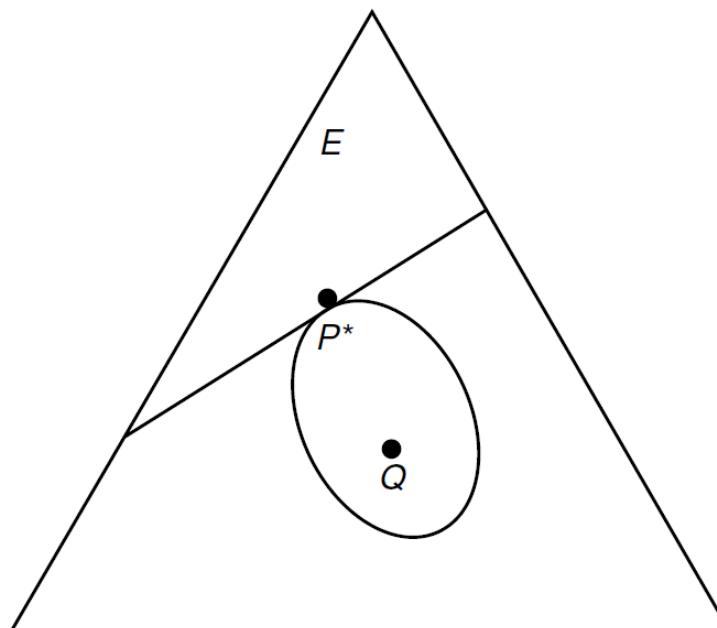


e-geodesic of two independent distributions is family of independent distributions

m-projection of $p(x, y)$ to Manifold of independent distributions

Sanov's theorem (large deviation theory)

- Probability simplex is both an exponential family and a mixture family



Theorem (Sanov's theorem) Let X_1, X_2, \dots, X_n be i.i.d. $\sim Q(x)$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)},$$

where

$$P^* = \arg \min_{P \in E} D(P||Q)$$

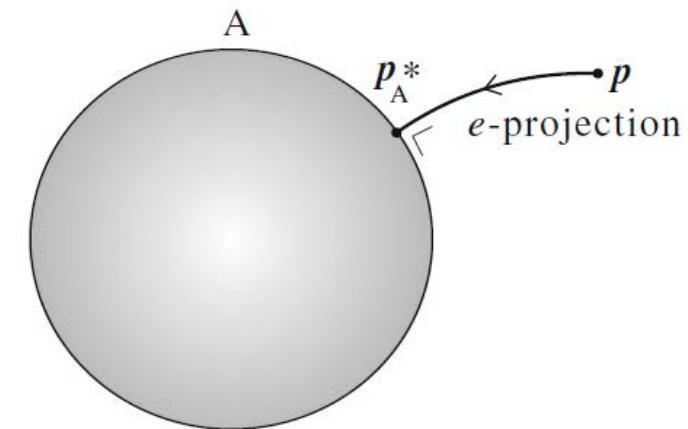
is the distribution in E that is closest to Q in relative entropy.
If, in addition, the set E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \rightarrow -D(P^*||Q).$$

Sanov's theorem (large deviation theory)

Empirical distribution from iid observations is MLE of categorical distributions

$$\hat{p}_i = \frac{1}{N} \sum_{t=1}^N \delta_i \{x(t)\} = \frac{N_i}{N}$$



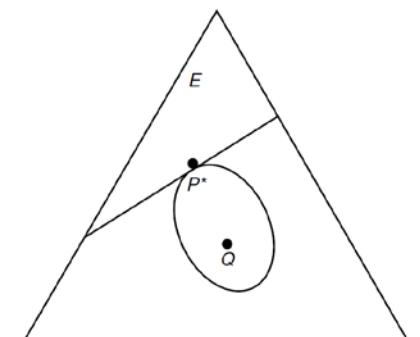
Large Deviation Theorem The probability that \hat{p} is included in A is given asymptotically by

$$\text{Prob} \{\hat{p} \in A\} = \exp \{-N D_{KL} [p_A^* : p]\},$$

where

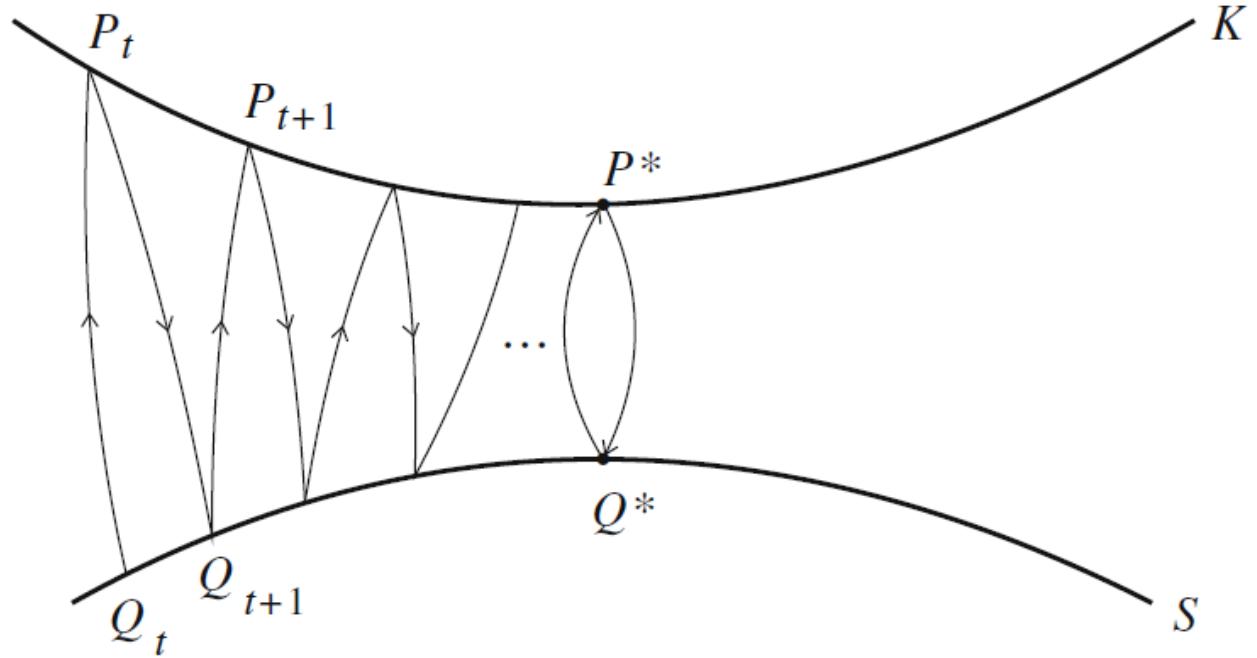
$$p_A^* = \arg \min_{q \in A} D_{KL} [q : p].$$

When A is a closed set having a boundary, p_A^* is given by e -projecting p to the boundary of A .



Divergence between two submanifolds

- Alternating minimization algorithm

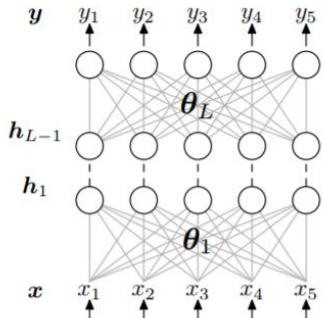


$$D(K : S) = \min_{P \in K, Q \in S} D(P : Q) = D(P^* : Q^*)$$

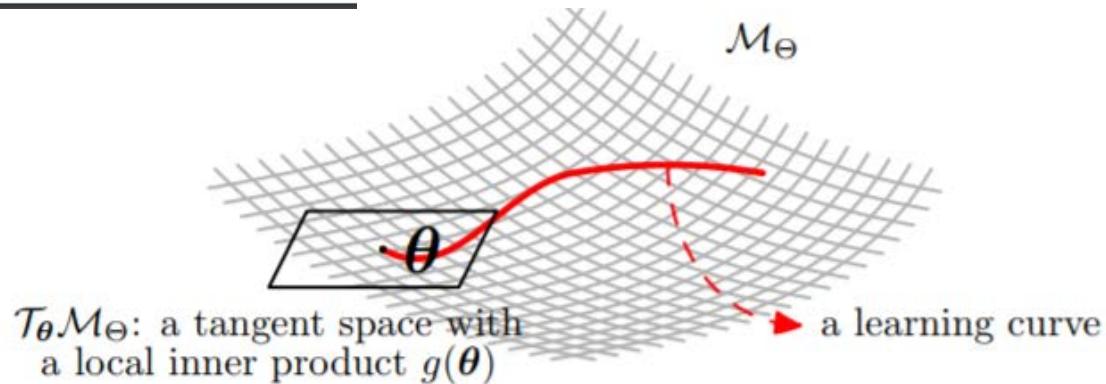
$$D(P_{t-1} : Q_t) \geq D(P_t : Q_t) \geq D(P_t : Q_{t+1})$$

Unique when S is flat and K is dual flat. Otherwise, converging point not necessarily unique.

$$p(\mathbf{y} | \mathbf{x}, \Theta) = \sum_{\mathbf{h}_1, \dots, \mathbf{h}_{L-1}} p(\mathbf{y} | \mathbf{h}_{L-1}, \theta_L) \cdots p(\mathbf{h}_2 | \mathbf{h}_1, \theta_2) p(\mathbf{h}_1 | \mathbf{x}, \theta_1),$$



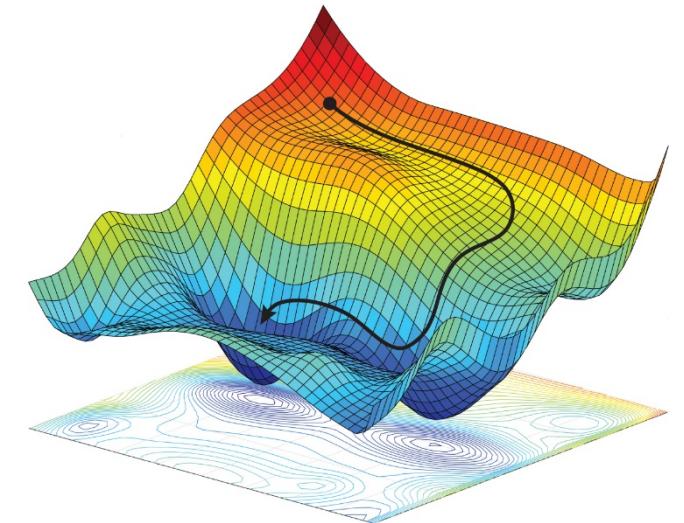
Natural gradient and mirror descent



Frank Nielsen



Sony CSL

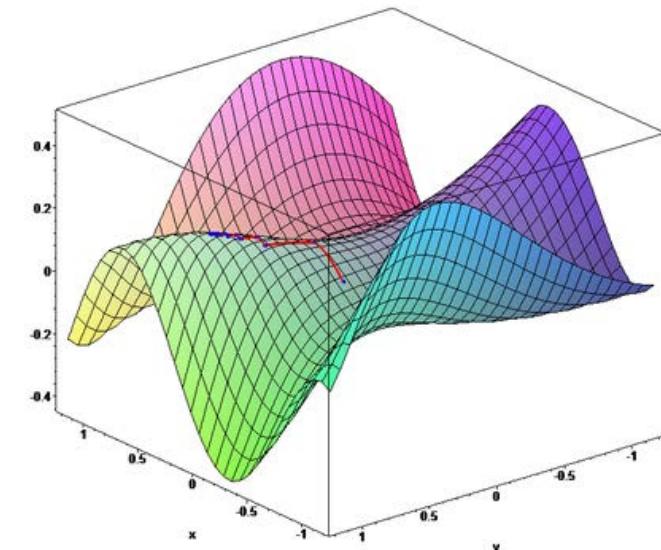
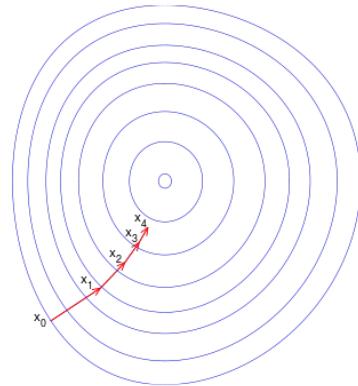
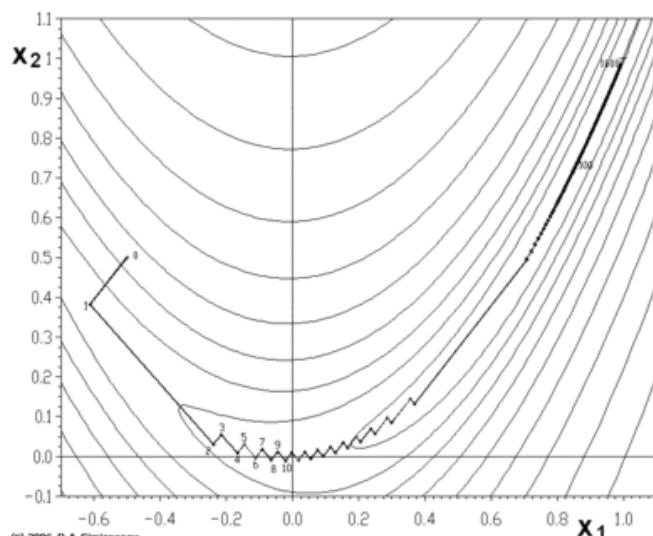


Steepest gradient descent method

- Iterative optimization algorithm
- Start from an initial parameter value θ_0
- Update the current parameter using a **learning rate α** (step size) and the **gradient of the energy function**:

$$\theta_{t+1} = \theta_t - \alpha \nabla E(\theta_t)$$

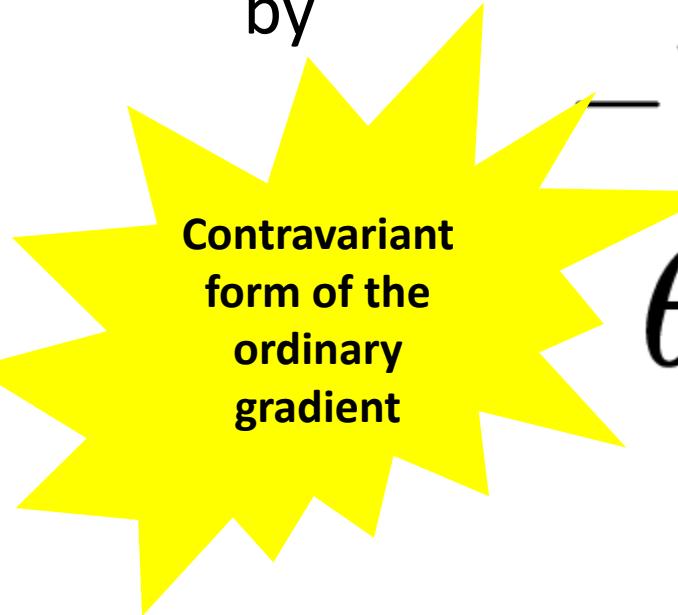
First order optimization method



Steepest descent in a Riemannian space

- The steepest descent direction of $E(\theta)$ in a **Riemannian space** is given by

$$-\tilde{\nabla} E(\theta) = -G^{-1}(\theta) \nabla E(\theta)$$
$$\theta_{t+1} = \theta_t - l_t \tilde{\nabla} E(\theta_t)$$



Contravariant
form of the
ordinary
gradient

Computing the inverse of the Fisher information matrix is tricky

Pros and cons of natural gradient

- Pros
 - Invariant (intrinsic) gradient
 - Not trapped in plateaus
 - Achieve Fisher efficiency in online learning
- Cons
 - Too expensive to compute (no closed-form FIM; need matrix inversion)
 - May be degenerate for irregular models (e.g., hierarchical models)

In a dually flat space, natural gradient is ordinary gradient for the dual coordinates

- In a dually flat space, we have

$$I_\theta(\theta) = \nabla_\theta^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$$

Natural gradient

$$\begin{aligned}\tilde{\nabla}_\theta L_\theta(\theta) &:= I_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta) \\ &= (\nabla_\theta \eta)^{-1} \nabla_\theta \eta \nabla_\eta L_\eta(\eta) \\ &= \nabla_\eta L_\eta(\eta)\end{aligned}$$

Ordinary gradient



Used in variational information

Mirror descent in non-Euclidean space

$$\theta_{t+1} = \theta_t - \alpha \nabla E(\theta_t)$$

Can be rewritten as

$$x_{k+1} = \operatorname{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

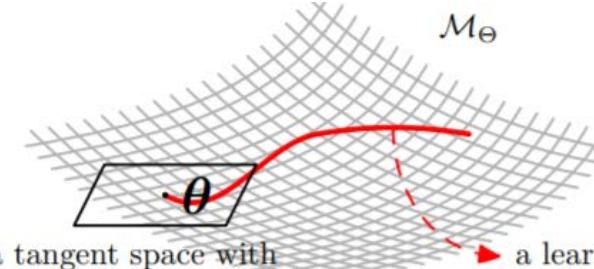
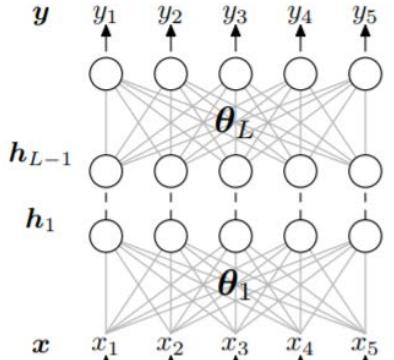
Replace squared loss with any Bregman divergence

$$x_{k+1} = \operatorname{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k} B_F(x : x_k) \right\}$$

Mirror descent for the Bregman divergence on the primal parameter amounts to natural gradient for the dual parameter

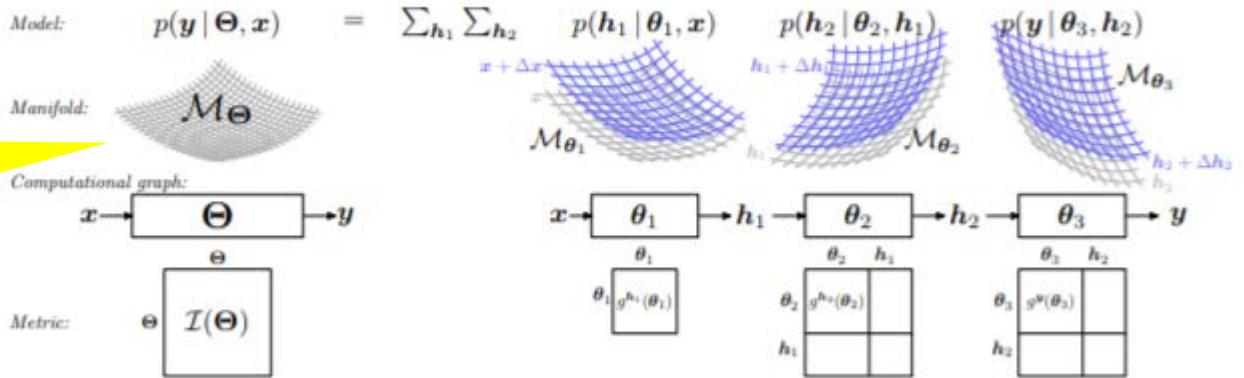
Relative Fisher Information Matrix (RFIM) and Relative Natural Gradient (RNG) for deep learning

$$p(\mathbf{y} | \mathbf{x}, \Theta) = \sum_{\mathbf{h}_1, \dots, \mathbf{h}_{L-1}} p(\mathbf{y} | \mathbf{h}_{L-1}, \theta_L) \cdots p(\mathbf{h}_2 | \mathbf{h}_1, \theta_2) p(\mathbf{h}_1 | \mathbf{x}, \theta_1),$$



$$\begin{aligned} g(\Theta) &= E_{\mathbf{x} \sim \hat{p}(X_n), \mathbf{y} \sim p(\mathbf{y} | \mathbf{x}, \Theta)} \left[\frac{\partial I}{\partial \Theta} \frac{\partial I}{\partial \Theta^\top} \right] \\ &= \frac{1}{n} \sum_{i=1}^n E_{p(\mathbf{y} | \mathbf{x}_i, \Theta)} \left[\frac{\partial I_i}{\partial \Theta} \frac{\partial I_i}{\partial \Theta^\top} \right] \end{aligned}$$

Relative Fisher IM: $g^h(\theta | \theta_f) = E_{p(h | \theta, \theta_f)} \left[\frac{\partial}{\partial \theta} \ln p(h | \theta, \theta_f) \frac{\partial}{\partial \theta^\top} \ln p(h | \theta, \theta_f) \right]$



The RFIMs of single neuron models, a linear layer, a non-linear layer, a soft-max layer, two consecutive layers all have simple closed form solutions

Neuromanifolds, Occam's Razor and Deep Learning

Question: Why do DNNs generalize well with huge number of free parameters?

Problem: Generalization error of DNNs is experimentally not U-shaped but a **double descent risk curve** (arxiv 1812.11118)

Occam's razor for Deep Neural Networks (DNNs):

(uniform width M, L layers, N #observations, d: dimension of screen distributions in lightlike neuromanifold)

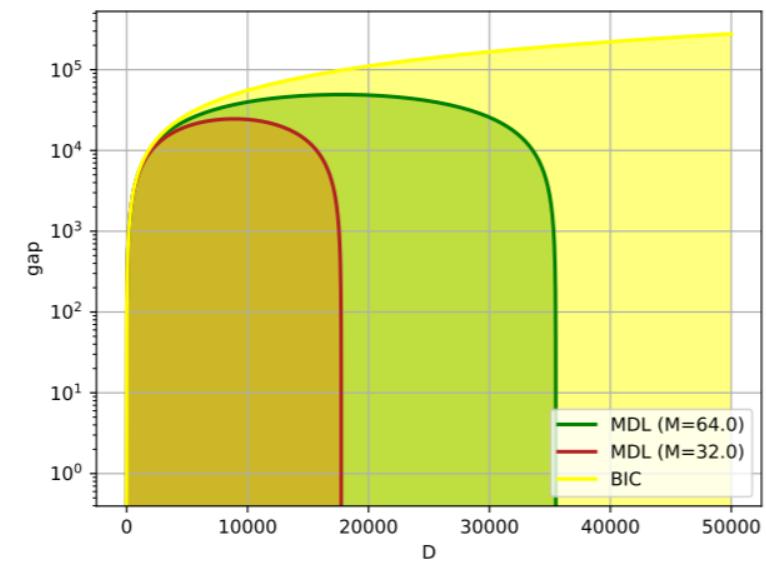
Θ : parameters of the DNN, $\hat{\Theta}$: estimated parameters

$$\mathcal{O} = -\log P(X | \hat{\Theta}) + \frac{d}{2} \log N + \frac{d}{2} \int_0^\infty \rho_{\mathcal{I}}(\lambda) \log \lambda d\lambda$$

$$\mathcal{O} \approx -\log P(X | \hat{\Theta}) + \frac{d}{2} \log N - \frac{d}{2} \gamma LM$$

$\rho_{\mathcal{I}}$ Spectrum density of the Fisher Information Matrix (FIM)

$$\mathcal{I}(\Theta) = E_p \left(\frac{\partial \log p(X | \Theta)}{\partial \Theta} \frac{\partial \log p(X | \Theta)}{\partial \Theta^T} \right)$$



Estimated generalisation gap (in log scale) against the number of free parameters.

<https://arxiv.org/abs/1905.11027>

Summary

- Natural gradient in a dually flat manifold is equivalent to ordinary gradient with respect to the dual parameter
- Mirror descent
- Random Matrix Theory (RMT) for the FIM
- Other alternatives

Clustering: Hard, Soft and Hierarchical

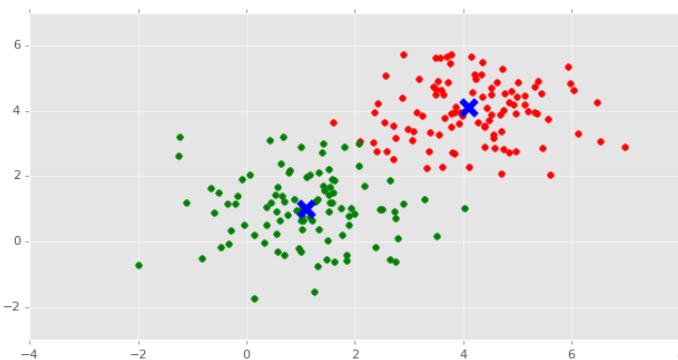
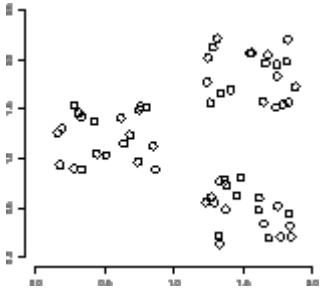
Frank Nielsen

Sony Computer Science Laboratories, Inc

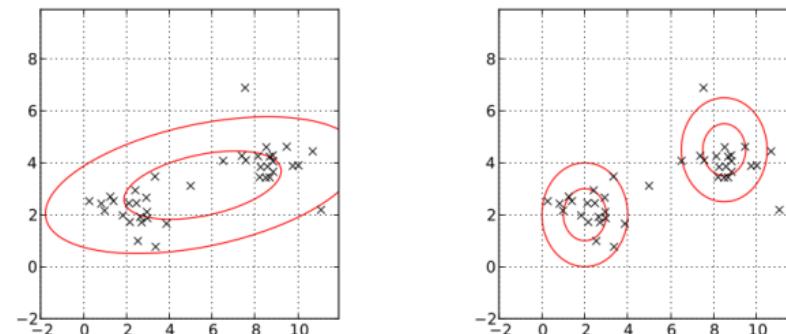
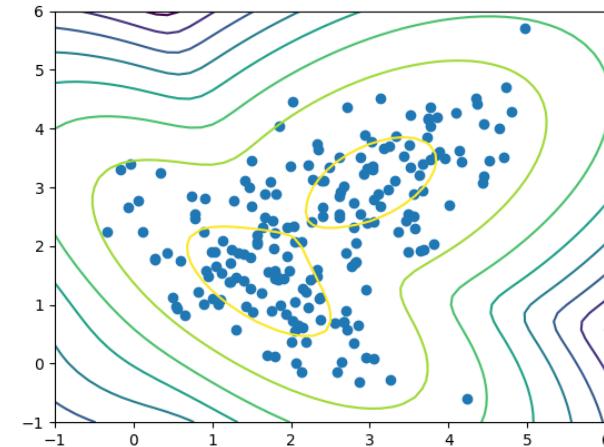


Sony CSL

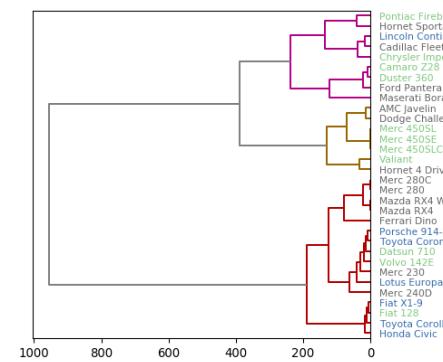
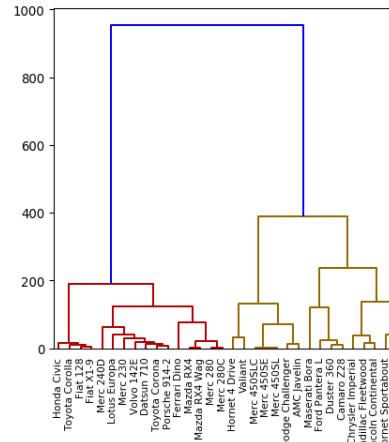
Finding structures (clusters) in datasets



Hard membership
Flat clustering (partitions)



Soft membership
Mixture models
Gaussian mixture models



Hierarchical clustering
Dendrograms
Agglomerative/divisive

Exploratory data science

Rationale

- Extend squared Euclidean distance-based clustering to **arbitrary Bregman divergence**: k-means, expectation-maximization (isotropic GMMs), hierarchical clustering

$$D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle$$

- Use **duality** of “regular” Bregman divergences with regular exponential families to learn mixtures of exponential families

$$\log p_F(x; \theta) = -B_{F^*}(t(x) : \eta) + F^*(t(x)) + k(x)$$

- Use **conformal Bregman divergences** (total Bregman divergences) to get robust clustering

Bregman k-mean clustering

- NP-complete when $k>1$ and $d>1$
- Local, global and probabilistic heuristics to find good k-means clustering
- Easy dynamic programming (DP) when $d=1$: Interval clustering

$$\underbrace{[x_1 \dots x_{l_2-1}]}_{c_1} \underbrace{[x_{l_2} \dots x_{l_3-1}]}_{c_2} \dots \underbrace{[x_{l_k} \dots x_n]}_{c_k}$$

- Speed calculation of mean/variance of clusters using Look-Up-Tables (summed area tables)
- Can perform model selection and also give constraints on cluster sizes

Bregman clustering ($d>1$)

Algorithm 1 Bregman Hard Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$, probability measure ν over \mathcal{X} , Bregman divergence $d_\phi : \mathcal{S} \times \text{ri}(\mathcal{S}) \mapsto \mathbb{R}$, number of clusters k .

Output: \mathcal{M}^\dagger , local minimizer of $L_\phi(\mathcal{M}) = \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)$ where $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$, hard partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of \mathcal{X} .

Method:

Initialize $\{\boldsymbol{\mu}_h\}_{h=1}^k$ with $\boldsymbol{\mu}_h \in \text{ri}(\mathcal{S})$ (one possible initialization is to choose $\boldsymbol{\mu}_h \in \text{ri}(\mathcal{S})$ at random)

repeat

{The Assignment Step}

Set $\mathcal{X}_h \leftarrow \emptyset$, $1 \leq h \leq k$

for $i = 1$ to n **do**

$\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{\mathbf{x}_i\}$

where $h = h^\dagger(\mathbf{x}_i) = \underset{h'}{\operatorname{argmin}} d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'})$

end for

{The Re-estimation Step}

for $h = 1$ to k **do**

$\pi_h \leftarrow \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$

$\boldsymbol{\mu}_h \leftarrow \frac{1}{\pi_h} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i \mathbf{x}_i$

end for

until convergence

return $M^\dagger \leftarrow \{\boldsymbol{\mu}_h\}_{h=1}^k$

Bregman centroids are centers of mass, independent of the generator

Compared to squared Euclidean k-means, only the assignment step changes

k-MLE: Inferring statistical mixtures a la k-Means

arxiv:1203.5181

Bijection between regular Bregman divergences
and regular (dual) exponential families

$$\log p_F(x; \theta) = -B_{F^*}(t(x) : \eta) + F^*(t(x)) + k(x)$$

Maximum log-likelihood estimate (exp. Family)
= dual Bregman centroid

$$\begin{aligned} \max_{\theta \in \mathbb{N}} \quad & \bar{l}(\theta; x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (\langle t(x_i), \theta \rangle - F(\theta) + k(x_i)) \\ \equiv \min_{\eta \in \mathbb{M}} \quad & \frac{1}{n} \sum_{i=1}^n B_{F^*}(t(x_i) : \eta) \end{aligned}$$

Classification Expectation-Maximization (CEM) yields a **dual Bregman k-means** for mixtures of exponential families (however, k-MLE is not consistent)

Online k-MLE for Mixture Modeling with Exponential Families, GSI 2015

On learning statistical mixtures maximizing the complete likelihood, AIP 2014

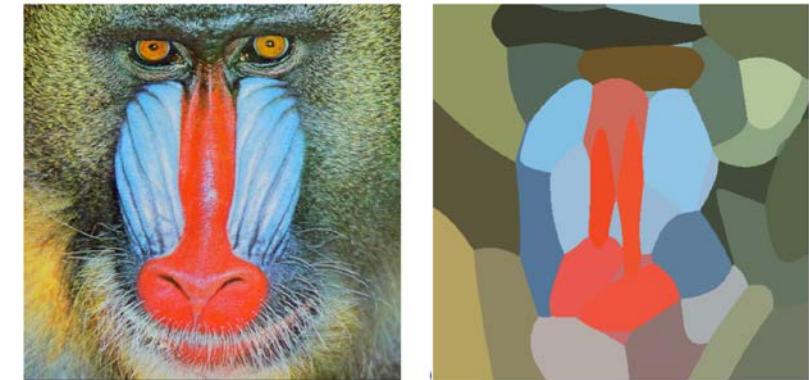
Hartigan's Method for k-MLE: Mixture Modeling with Wishart Distributions and Its Application to Motion Retrieval, GTI 2014

A New Implementation of k-MLE for Mixture Modeling of Wishart Distributions, GSI 2013

Fast Learning of Gamma Mixture Models with k-MLE, SIMBAD 2013

k-MLE: A fast algorithm for learning statistical mixture models, ICASSP 2012

k-MLE for mixtures of generalized Gaussians, ICPR 2012



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

MLE as a Bregman centroid for exponential families

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_{i=1}^n p_F(x_i; \theta) = \nabla F^{-1} \left(\sum_{i=1}^n t(x_i) \right).$$

Maximizing the average log-likelihood $\bar{l} = \frac{1}{n} \log L$, we have:

$$\max_{\theta \in \mathbb{N}} \quad \bar{l}(\theta; x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (\langle t(x_i), \theta \rangle - F(\theta) + k(x_i))$$

$$\max_{\theta \in \mathbb{N}} \quad \frac{1}{n} \sum_{i=1}^n -B_{F^*}(t(x_i) : \eta) + F^*(t(x_i)) + k(x_i)$$

$$\equiv \min_{\eta \in \mathbb{M}} \quad \frac{1}{n} \sum_{i=1}^n B_{F^*}(t(x_i) : \eta)$$

K-MLE: Classification Expectation-Maximization (CEM)

- 0. **Initialization:** $\forall i \in \{1, \dots, k\}$, let $w_i = \frac{1}{k}$ and $\eta_i = t(x_i)$
(Proper initialization is further discussed later on).
- 1. **Assignment:** $\forall i \in \{1, \dots, n\}, z_i = \operatorname{argmin}_{j=1}^k B_{F^*}(t(x_i) : \eta_j) - \log w_j$.
Let $\forall i \in \{1, \dots, k\} \mathcal{C}_i = \{x_j | z_j = i\}$ be the cluster partition: $\mathcal{X} = \bigcup_{i=1}^k \mathcal{C}_i$.
(some clusters may become empty depending on the weight distribution)
- 2. **Update the η -parameters:** $\forall i \in \{1, \dots, k\}, \eta_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} t(x)$.
(By convention, $\eta_i = \emptyset$ if $|\mathcal{C}_i| = 0$) **Goto step 1** unless local convergence of the complete likelihood is reached.
- 3. **Update the mixture weights:** $\forall i \in \{1, \dots, k\}, w_i = \frac{1}{n} |\mathcal{C}_i|$.
Goto step 1 unless local convergence of the complete likelihood is reached.

Additive Bregman Voronoi diagrams
Biased, not consistent

Bregman soft-clustering: Generalize expectation-maximization (EM) algorithm

Algorithm 2 Standard EM for Mixture Density Estimation

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, number of clusters k .

Output: Γ^\dagger : local maximizer of $L_{\mathcal{X}}(\Gamma) = \prod_{i=1}^n (\sum_{h=1}^k \pi_h p_{\psi, \theta_h}(\mathbf{x}_i))$ where $\Gamma = \{\theta_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$.

Method:

Initialize $\{\theta_h, \pi_h\}_{h=1}^k$ with some $\theta_h \in \Theta$, and $\pi_h \geq 0$, $\sum_{h=1}^k \pi_h = 1$

repeat

{The Expectation Step (E-step)}

for $i = 1$ to n **do**

for $h = 1$ to k **do**

$$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p_{(\psi, \theta_h)}(\mathbf{x}_i)}{\sum_{h'=1}^k \pi_{h'} p_{(\psi, \theta_{h'})}(\mathbf{x}_i)}$$

end for

end for

{The Maximization Step (M-step)}

for $h = 1$ to k **do**

$$\pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|\mathbf{x}_i)$$

$$\theta_h \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \log(p_{(\psi, \theta)}(\mathbf{x}_i)) p(h|\mathbf{x}_i)$$

end for

until convergence

return $\Gamma^\dagger = \{\theta_h, \pi_h\}_{h=1}^k$

Algorithm 3 Bregman Soft Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{S} \subseteq \mathbb{R}^d$, Bregman divergence $d_\phi : \mathcal{S} \times \text{ri}(\mathcal{S}) \mapsto \mathbb{R}$, number of clusters k .

Output: Γ^\dagger , local maximizer of $\prod_{i=1}^n (\sum_{h=1}^k \pi_h b_\phi(\mathbf{x}_i) \exp(-d_\phi(\mathbf{x}_i, \mu_h)))$ where $\Gamma = \{\mu_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$

Method:

Initialize $\{\mu_h, \pi_h\}_{h=1}^k$ with some $\mu_h \in \text{ri}(\mathcal{S})$, $\pi_h \geq 0$, and $\sum_{h=1}^k \pi_h = 1$

repeat

{The Expectation Step (E-step)}

for $i = 1$ to n **do**

for $h = 1$ to k **do**

$$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h \exp(-d_\phi(\mathbf{x}_i, \mu_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-d_\phi(\mathbf{x}_i, \mu_{h'}))}$$

end for

end for

{The Maximization Step (M-step)}

for $h = 1$ to k **do**

$$\pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|\mathbf{x}_i)$$

$$\mu_h \leftarrow \frac{\sum_{i=1}^n p(h|\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}$$

end for

until convergence

return $\Gamma^\dagger = \{\mu_h, \pi_h\}_{h=1}^k$

K-means++ probabilistic seeding

k-means++: Pick uniformly at random at first seed c_1 , and then iteratively choose the $(k - 1)$ remaining seeds according to the following probability distribution:

$$\Pr(c_j = p_i) = \frac{D(p_i, \{c_1, \dots, c_{j-1}\})}{\sum_{i=1}^n D(p_i, \{c_1, \dots, c_{j-1}\})} \quad (2 \leq j \leq k).$$

K-means++ probabilistic seeding

$$E_D(\Lambda, C) = \frac{1}{n} \sum_{i=1}^n \min_{j \in \{1, \dots, k\}} D(p_i : c_j)$$

Theorem (Generalized k -means++ performance). Let κ_1 and κ_2 be two constants such that κ_1 defines the quasi-triangular inequality property:

$$D(x : z) \leq \kappa_1 (D(x : y) + D(y : z)), \quad \forall x, y, z$$

and κ_2 handles the symmetry inequality:

$$D(x : y) \leq \kappa_2 D(y : x), \quad \forall x, y$$

Then the generalized k -means++ seeding guarantees with high probability a configuration C of cluster centers such that:

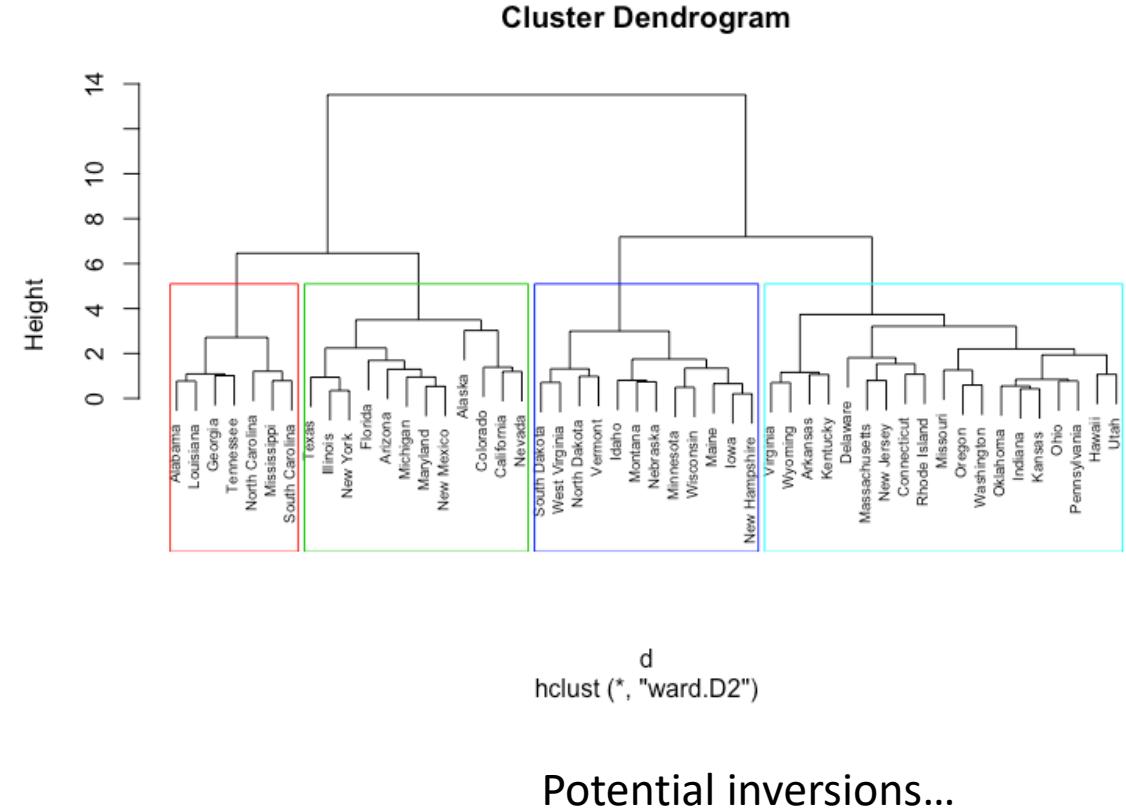
$$E_D(\Lambda, C) \leq 2\kappa_1^2(1 + \kappa_2)(2 + \log k)E_D^*(\Lambda, k).$$

Hierarchical clustering (Ward criterion)

1. Start with m clusters: $C_i := \{x_i\}$ for each i .
2. While at least two clusters remain:
 - (a) Choose $\{C_i, C_j\}$ with minimal $\Delta(C_i, C_j)$.
 - (b) Remove $\{C_i, C_j\}$, add in $C_i \cup C_j$.

$$\Delta_w(C_i, C_j) := \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\tau(C_i) - \tau(C_j)\|_2^2,$$

where $\tau(C)$ denotes the mean of cluster C .



Telgarsky, Matus, and Sanjoy Dasgupta. "Agglomerative Bregman Clustering." (2012).

Extending to Bregman divergences

$$B_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Consider more general directional derivatives

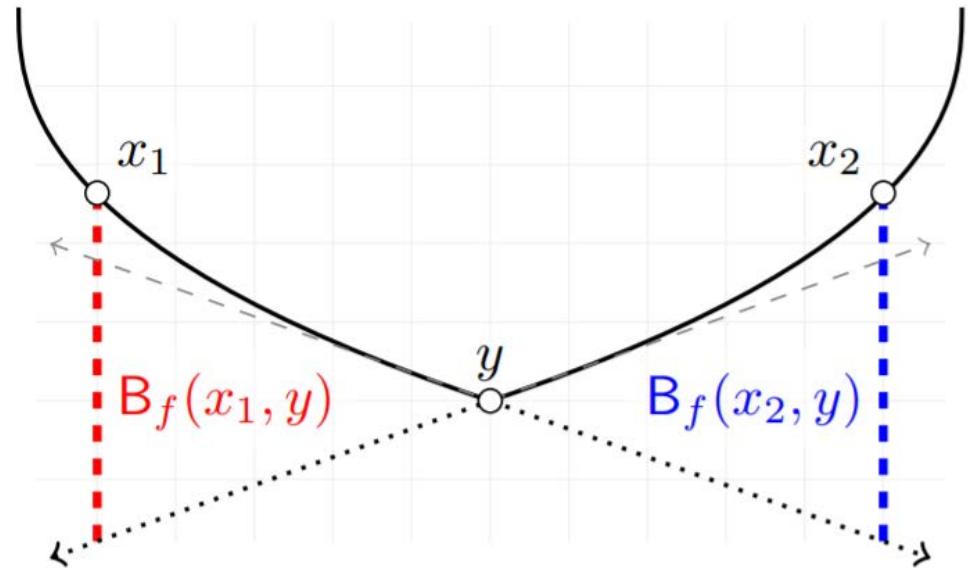
$$B_f(x, y) := f(x) - f(y) + f'(y; y - x).$$

Subgradient derivatives

Bregman
Ward
Criterion

Proposition 3.8. *Let a proper convex relatively differentiable f and two finite subsets C_1, C_2 of \mathcal{X} with $\tau(C_i) \in \text{ri}(\text{dom}(f))$ be given. Then*

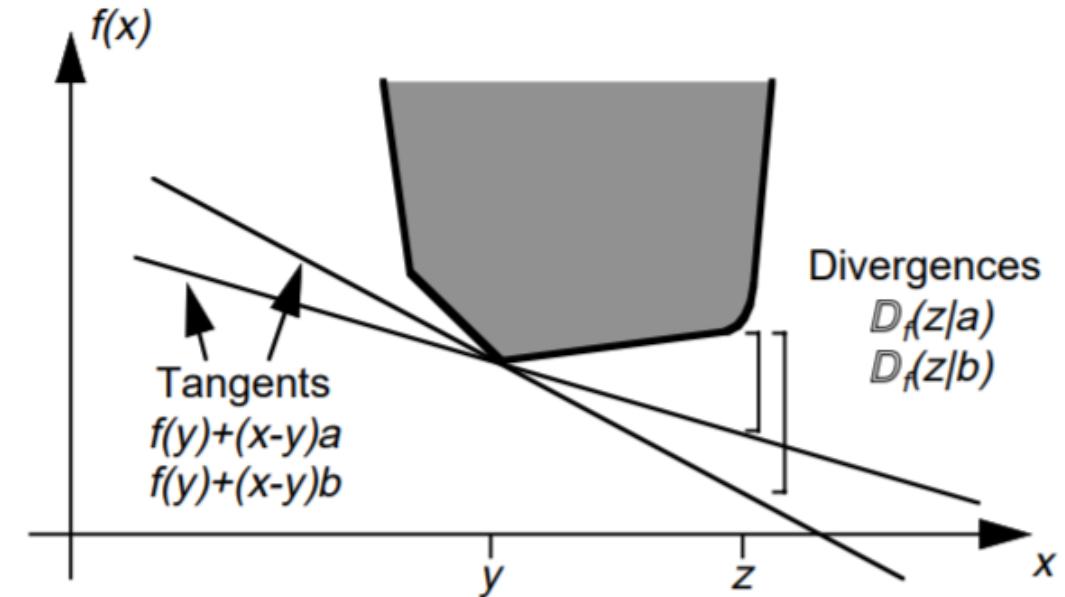
$$\Delta_{f,\tau}(C_1, C_2) = \sum_{j \in \{1,2\}} |C_j| B_f(\tau(C_j), \tau(C_1 \cup C_2)).$$



Another generalization of Bregman divergences

$$D_f(x, g) := f(x) + f^*(g) - \langle g, x \rangle. \quad g \in \partial f(y)$$

$$B_f(x, y) := \max\{D_f(x, g) : g \in \partial f(y)\}.$$





Clustering with mixed α -Divergences

$$M_{\lambda,\alpha}(p : x : q) = \lambda D_\alpha(p : x) + (1 - \lambda) D_\alpha(x : q) \quad \text{with} \quad D_\alpha(p : q) \doteq \sum_{i=1}^d \frac{4}{1 - \alpha^2} \left(\frac{1 - \alpha}{2} p^i + \frac{1 + \alpha}{2} q^i - (p^i)^{\frac{1-\alpha}{2}} (q^i)^{\frac{1+\alpha}{2}} \right)$$

K-means (hard/flat clustering)

Algorithm 1: Mixed α -seeding; MAS($\mathcal{H}, k, \lambda, \alpha$)

Input: Weighted histogram set \mathcal{H} , integer $k \geq 1$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} \leftarrow h_j$ with uniform probability ;

for $i = 2, 3, \dots, k$ **do**

Pick at random histogram $h \in \mathcal{H}$ with probability:

$$\pi_{\mathcal{H}}(h) \doteq \frac{w_h M_{\lambda,\alpha}(c_h : h : c_h)}{\sum_{y \in \mathcal{H}} w_y M_{\lambda,\alpha}(c_y : y : c_y)},$$

//where $(c_h, c_h) \doteq \arg \min_{(z,z) \in \mathcal{C}} M_{\lambda,\alpha}(z : h : z)$;

$\mathcal{C} \leftarrow \mathcal{C} \cup \{(h, h)\}$;

Output: Set of initial cluster centers \mathcal{C} ;

Input: Weighted histogram set \mathcal{H} , integer $k > 0$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \text{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

//Assignment

for $i = 1, 2, \dots, k$ **do**

$A_i \leftarrow \{h \in \mathcal{H} : i = \arg \min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

// Centroid relocation

for $i = 1, 2, \dots, k$ **do**

$r_i \leftarrow \left(\sum_{h \in A_i} w_i h^{\frac{1-\alpha}{2}} \right)^{\frac{2}{1-\alpha}}$;

$l_i \leftarrow \left(\sum_{h \in A_i} w_i h^{\frac{1+\alpha}{2}} \right)^{\frac{2}{1+\alpha}}$;

until convergence;

Output: Partition of \mathcal{H} in k clusters following \mathcal{C} ;

$$J_\alpha(\tilde{p} : \tilde{q}) = \frac{8}{1 - \alpha^2} \left(1 + \sum_{i=1}^d H_{\frac{1-\alpha}{2}}(\tilde{p}^i, \tilde{q}^i) \right)$$

$$H_\beta(a, b) = \frac{a^\beta b^{1-\beta} + a^{1-\beta} b^\beta}{2}$$

Heinz means interpolate the arithmetic and the geometric means

$$\sqrt{ab} = H_{\frac{1}{2}}(a, b) \leq H_\alpha(a, b) \leq H_0(a, b) = \frac{a+b}{2}$$

EM (soft/generative clustering)

Input: Histogram set \mathcal{H} with $|\mathcal{H}| = m$, integer $k > 0$, real $\lambda \leftarrow \lambda_{\text{init}} \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \text{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

//Expectation

for $i = 1, 2, \dots, m$ **do**

for $j = 1, 2, \dots, k$ **do**

$p(j|h_i) = \frac{\pi_j \exp(-M_{\lambda,\alpha}(l_j : h_i : r_j))}{\sum_{j'} \pi_{j'} \exp(-M_{\lambda,\alpha}(l_{j'} : h_i : r_{j'}))}$;

//Maximization

for $j = 1, 2, \dots, k$ **do**

$\pi_j \leftarrow \frac{1}{m} \sum_i p(j|h_i);$
 $l_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1+\alpha}{2}} \right)^{\frac{2}{1+\alpha}}$;
 $r_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1-\alpha}{2}} \right)^{\frac{2}{1-\alpha}}$;

//Alpha - Lambda

$\alpha \leftarrow \alpha - \eta_1 \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) \frac{\partial}{\partial \alpha} M_{\lambda,\alpha}(l_j : h_i : r_j)$;

if $\lambda_{\text{init}} \neq 0, 1$ **then**

$\lambda \leftarrow \lambda - \eta_2 \left(\sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_\alpha(l_j : h_i : r_j) - \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_\alpha(h_i : r_j) \right)$;

//for some small η_1, η_2 ; ensure that $\lambda \in [0, 1]$.

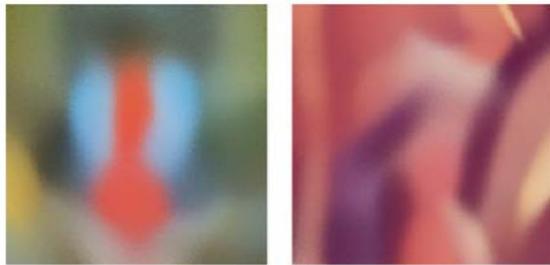
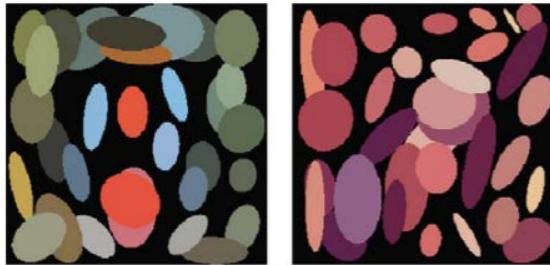
until convergence;

Output: Soft clustering of \mathcal{H} according to k densities $p(j|.)$ following \mathcal{C} ;

Hierarchical mixtures of exponential families

Hierarchical clustering with Bregman sided and symmetrized divergences

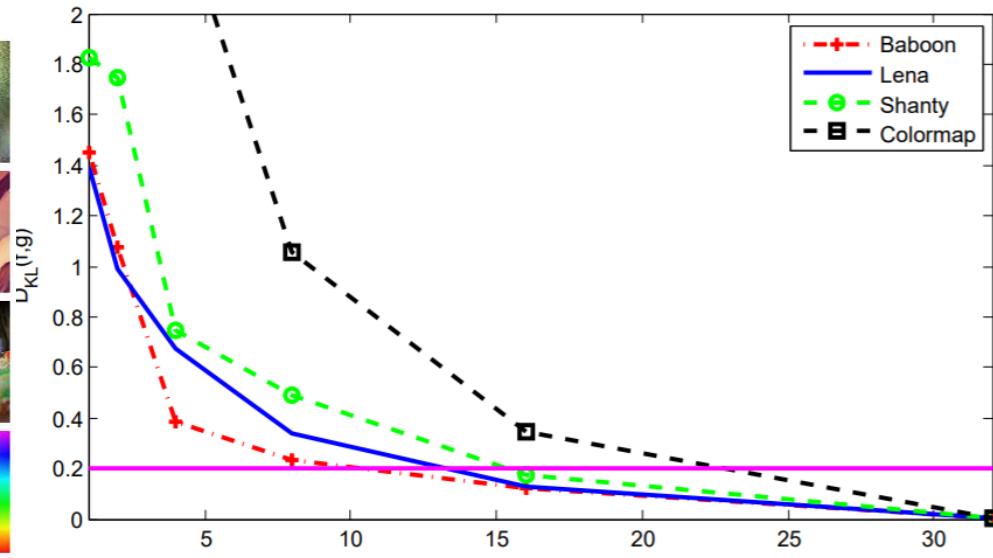
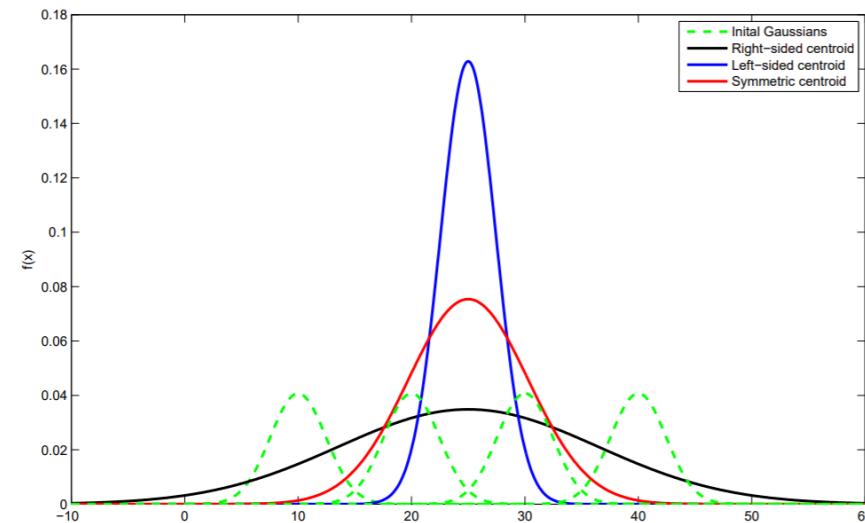
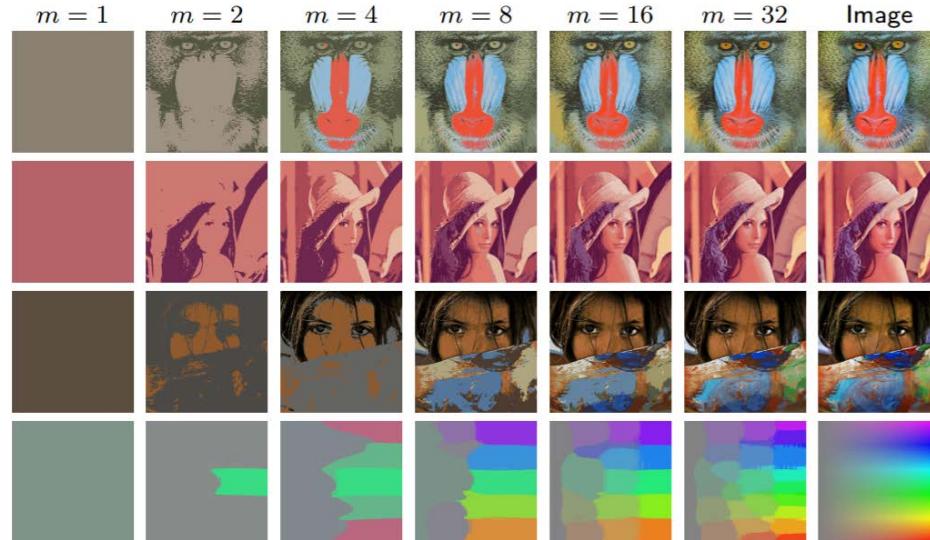
Learning & simplifying
Gaussian mixture models (GMMs)



- Agglomerative method:

- Find the two closest subsets \mathcal{S}_i and \mathcal{S}_j
- Merge the subsets \mathcal{S}_i and \mathcal{S}_j
- Go back to 1. until one single set remains

| Criterion | Formula |
|------------------|--|
| Minimum distance | $D_{\min}(A, B) = \min\{d(a, b) \mid a \in A, b \in B\}$ |
| Maximum distance | $D_{\max}(A, B) = \max\{d(a, b) \mid a \in A, b \in B\}$ |
| Average distance | $D_{\text{av}}(A, B) = \frac{1}{ A B } \sum_{a \in A} \sum_{b \in B} d(a, b)$ |



Conformal divergences

$$D'(p : q) = \rho(p, q) D(p : q)$$

$$\mathbf{D}_{F,\kappa} [\xi : \xi'] := \kappa(\xi) \mathbf{B}_F [\xi : \xi']$$

Consider the right-sided centroid: Amount to reweight the points according to a positive conformal factor.
Related to conformal geometry

- Total Bregman divergences, total Jensen divergences, etc.

On Conformal Divergences and Their Population Minimizers. *IEEE Trans. Information Theory* 62(1) (2016)

Total Jensen divergences: Definition, properties and clustering. [ICASSP 2015](#): 2016-2020

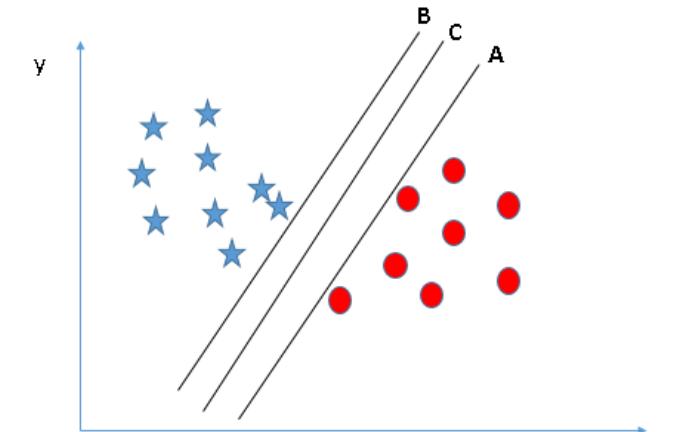
Shape Retrieval Using Hierarchical Total Bregman Soft Clustering. [IEEE Trans. Pattern Anal. Mach. Intell.](#) 34(12): 2407-2419 (2012)

Total Bregman Divergence and Its Applications to DTI Analysis. [IEEE Trans. Med. Imaging](#) 30(2): 475-483 (2011)

Conformal distances in machine learning: SVM

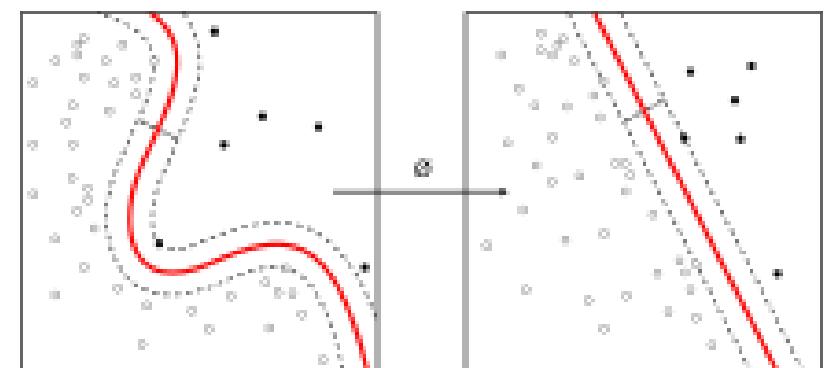
- Conformal kernel

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = D(\mathbf{x})D(\mathbf{x}')K(\mathbf{x}, \mathbf{x}'),$$



- Conformal Riemannian metric

$$\tilde{g}_{ij}(\mathbf{x}) = D(\mathbf{x})^2 g_{ij}(\mathbf{x}) + D_i(\mathbf{x})D_j(\mathbf{x}) + 2D_i(\mathbf{x})D(\mathbf{x})K_i(\mathbf{x}, \mathbf{x}),$$



Wu, Si, and Shun-ichi Amari. "Conformal Transformation of Kernel Functions: A Data-dependent Way to Improve Support Vector Machine Classifiers." *Neural Processing Letters* 15.1 (2002): 59-67.

Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

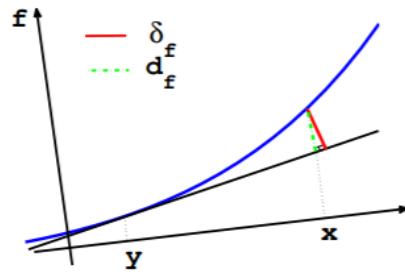
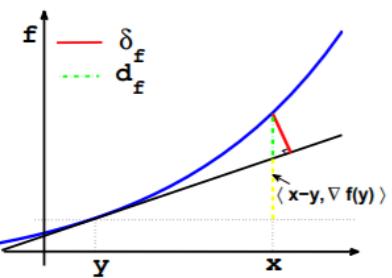
Definition The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,

$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}},$$

$\langle \cdot, \cdot \rangle$ is inner product
 $\langle \nabla f(y), \nabla f(y) \rangle$ generally.

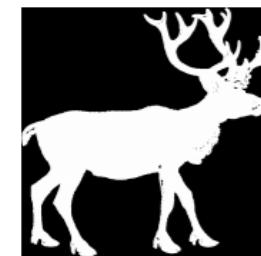
and $\|\nabla f(y)\|^2 =$

| X | $f(x)$ | $\delta_f(x, y)$ | t -center | ℓ_1 -norm BD center | Remark |
|-----------------------------|-----------------------------|--|---|--------------------------|------------------------------|
| \mathbb{R} | x^2 | $\frac{(x-y)^2}{\sqrt{1+4y^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total square loss (tSL) |
| $\mathbb{R} - \mathbb{R}_-$ | $x \log x$ | $\frac{x \log \frac{x}{y} + \bar{x} \log \frac{\bar{x}}{\bar{y}}}{\sqrt{1+y(1+\log y)^2 + \bar{y}(1+\log \bar{y})^2}}$ | $\prod_i (x_i)^{w_i}$ | $\sum_i x_i$ | total logistic loss |
| $[0, 1]$ | $-\log x$ | $\frac{\frac{x}{y} - \log \frac{x}{y} - 1}{\sqrt{1+y^{-2}}}$ | $\frac{\sum_i (x_i/(1-x_i))^{w_i}}{1 + \sum_i (x_i/(1-x_i))^{w_i}}$ | $\sum_i x_i$ | total Itakura-Saito distance |
| \mathbb{R}_+ | $-\log x$ | $\frac{\frac{x}{y} - \log \frac{x}{y} - 1}{\sqrt{1+y^{-2}}}$ | $\frac{1}{\sum_i w_i/x_i}$ | $\sum_i x_i$ | total squared Euclidean |
| \mathbb{R} | e^x | $\frac{e^x - e^y - (x-y)e^y}{\sqrt{1+e^{2y}}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total Mahalanobis distance |
| \mathbb{R}^d | $\ x\ ^2$ | $\frac{\ x-y\ ^2}{\sqrt{1+4\ y\ ^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total KL divergence (tKL) |
| \mathbb{R}^d | $x^t Ax$ | $\frac{(x-y)^t A(x-y)}{\sqrt{1+4\ Ay\ ^2}}$ | $\sum_i w_i x_i$ | $\sum_i x_i$ | total squared Frobenius |
| Δ^d | $\sum_{j=1}^d x_j \log x_j$ | $\frac{\sum_{j=1}^d x_j \log \frac{x_j}{y_j}}{\sqrt{1+\sum_{j=1}^d y_j(1+\log y_j)^2}}$ | $c \prod_i (x_i)^{w_i}$ | $\sum_i x_i$ | |
| $\mathbb{C}^{m \times n}$ | $\ x\ _F^2$ | $\frac{\ x-y\ _F^2}{\sqrt{1+4\ y\ _F^2}}$ | | $\sum_i x_i$ | |

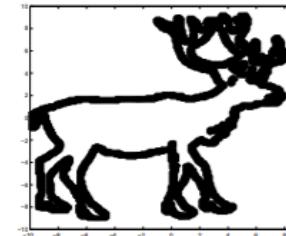


t-center: $\bar{x} = \arg \min_x \delta_f^1(x, E) = \arg \min_x \sum_{i=1}^n \delta_f(x, x_i)$

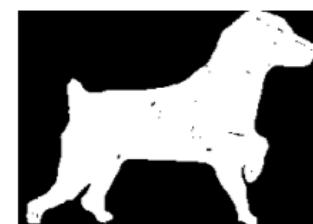
Robust to noise/outliers



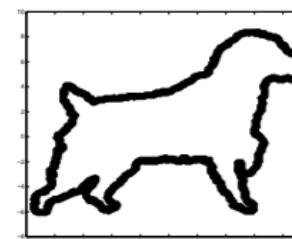
(m)



(n)



(o)



Total Bregman divergence and its applications to DTI analysis

IEEE Transactions on medical imaging, 30(2), 475-483, 2010.

Definition The total Bregman divergence (TBD) δ_f associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,

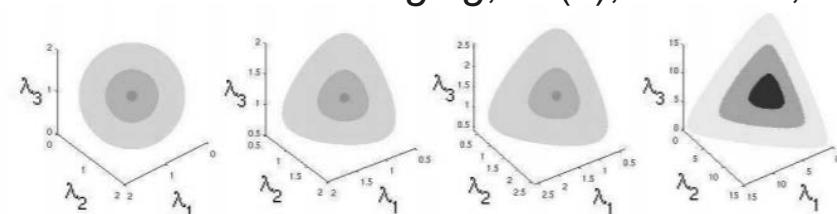
$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}}, \quad (2)$$

$\langle \cdot, \cdot \rangle$ is inner product as in definition II.1, and $\|\nabla f(y)\|^2 = \langle \nabla f(y), \nabla f(y) \rangle$ generally.

$$tKL(P, Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1 + \int (1 + \log q)^2 q dx}} \\ = \frac{\log(\det(P^{-1}Q)) + \text{tr}(Q^{-1}P) - n}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{n(1+\log 2\pi)}{2} \log(\det Q)}}$$

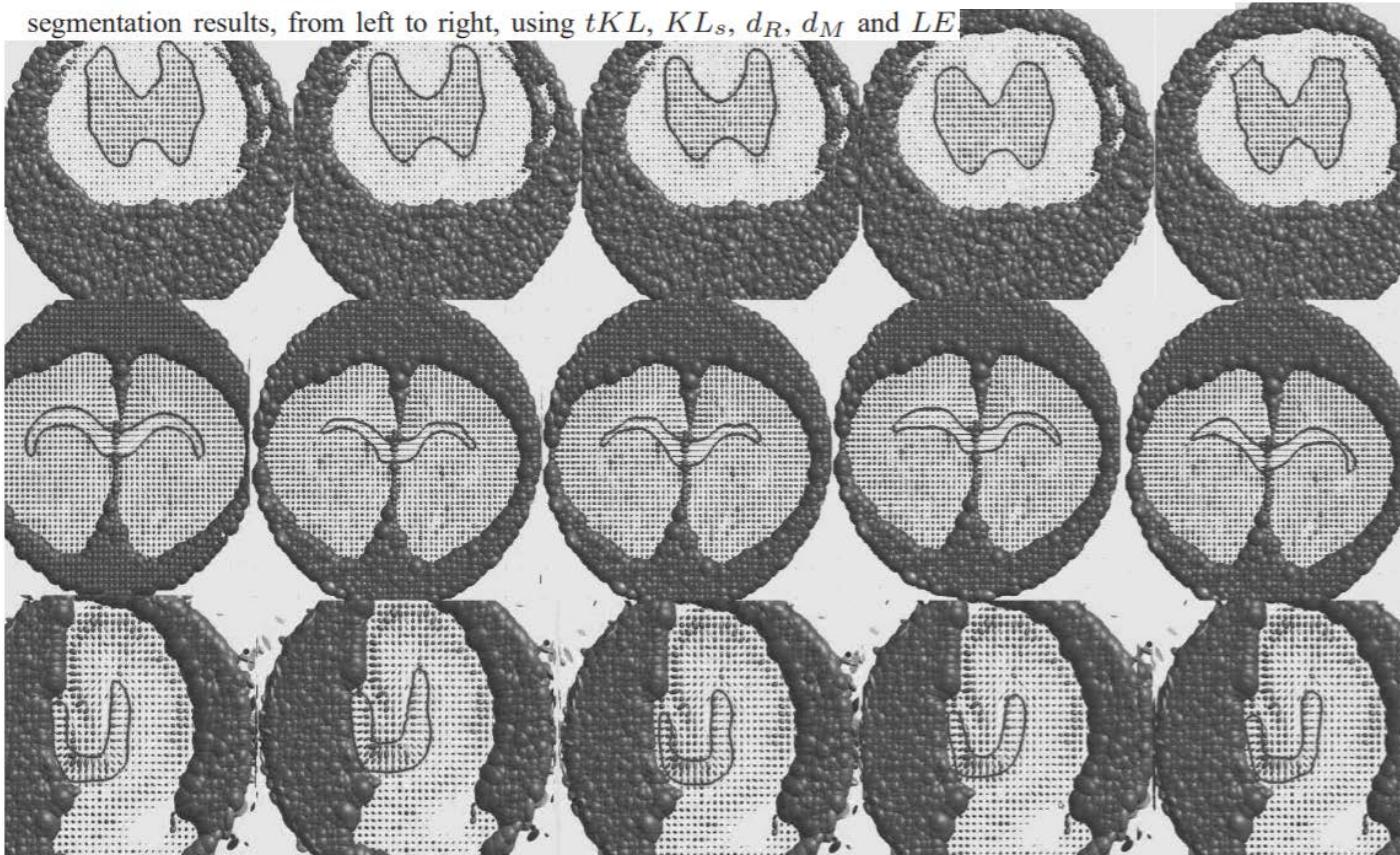
$$tKL(P, Q) = tKL(A'PA, A'QA), \quad \forall A \in SL(n),$$

$$tSL(P, Q) = \frac{\int (p - q)^2 dx}{\sqrt{1 + \int (2q)^2 q dx}} = \\ \frac{1/\sqrt{\det(2P)} + 1/\sqrt{\det(2Q)} - 2/\sqrt{\det(P+Q)}}{(2\pi)^n + 4\sqrt{(2\pi)^n}/\sqrt{\det(3Q)}}$$



The isosurfaces of $d_F(P, I) = r$, $d_R(P, I) = r$, $KL_s(P, I) = r$ and $tKL(P, I) = r$ shown from left to right. The three axes are eigenvalues of P .

segmentation results, from left to right, using tKL , KL_s , d_R , d_M and LE .



Axioms for a statistical distance (Ali & Silvey, 1966)

First property. The coefficient $d(P_1, P_2)$ should be defined for all pairs of measures P_1 and P_2 on the same sample space.

Second property. Suppose that $y = t(x)$ is a measurable transformation from $(\mathcal{X}, \mathcal{F})$ onto a measure space $(\mathcal{Y}, \mathcal{G})$. Then we should have

$$d(P_1, P_2) \geq d(P_1 t^{-1}, P_2 t^{-1}). \quad \begin{array}{l} \text{Coarser sigma-algebra} \\ \text{More distinguishability of stochastic processes} \end{array}$$

Here $P_i t^{-1}$ denotes the induced measure on \mathcal{Y} corresponding to P_i .

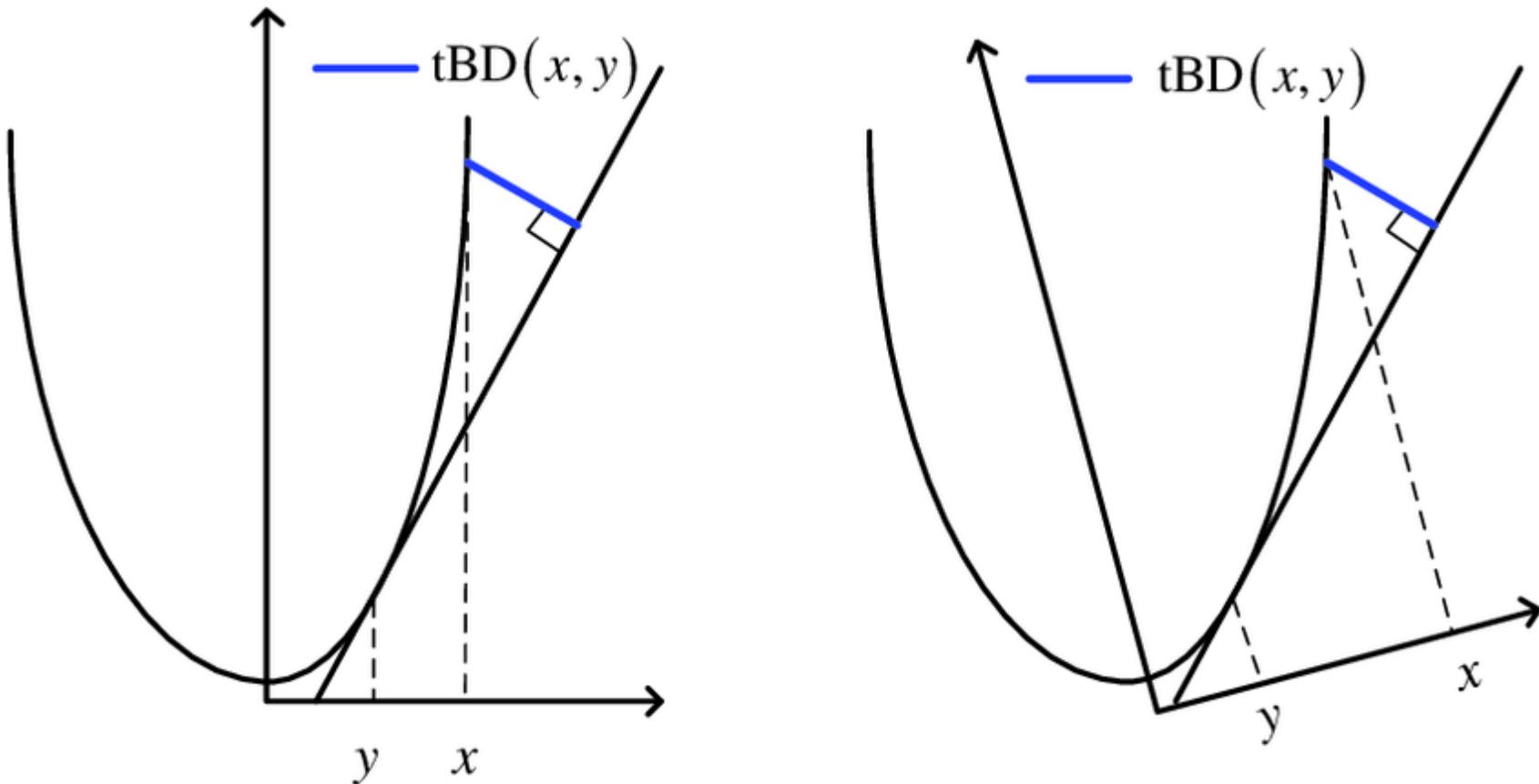
$$d(P_1^{(m)}, P_2^{(m)}) \leq d(P_1^{(n)}, P_2^{(n)}) \quad \text{for } m < n. \quad t(x_1, x_2, \dots, x_n) = (x_1, x_2, \dots, x_m).$$

Third property. $d(P_1, P_2)$ should take its minimum value when $P_1 = P_2$ and its maximum value when $P_1 \perp P_2$.

Fourth property. Let θ be a real parameter and let $\{P_\theta; \theta \in (a, b)\}$ be a family of equivalent (mutually absolutely continuous) distributions on the real line such that the family of densities $p_\theta(x)$ with respect to a fixed measure μ has monotone likelihood ratio in x (see Lehmann, 1959, p. 68). Then if $a < \theta_1 < \theta_2 < \theta_3 < b$, we should have

$$d(P_{\theta_1}, P_{\theta_2}) \leq d(P_{\theta_1}, P_{\theta_3}).$$

Total Bregman divergence



$$TBD(p : q) = \frac{\varphi(p) - \varphi(q) - \nabla \varphi(q) \cdot (p - q)}{\sqrt{1 + |\nabla \varphi(q)|^2}}$$

Aitchison distance in the simplex

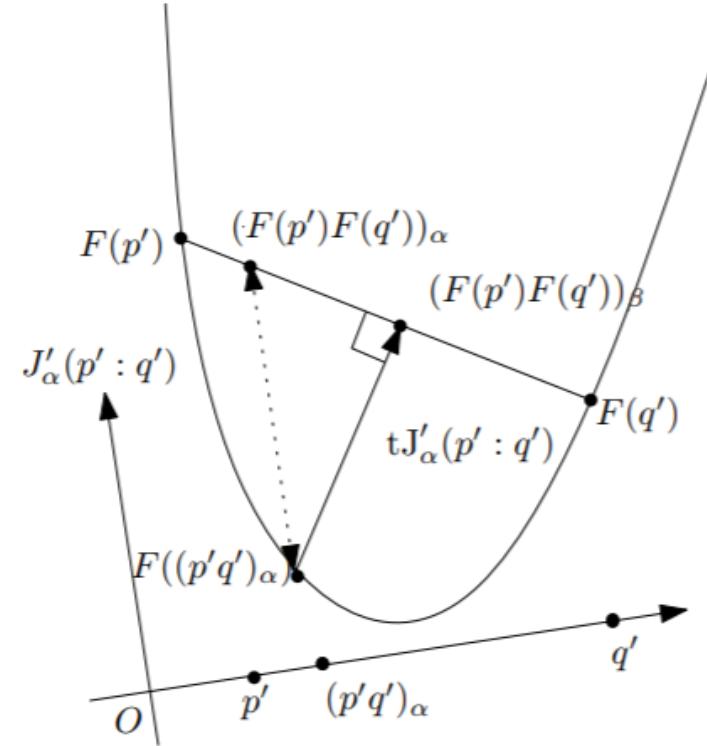
- Non-separable (experimentally monotone distance)

$$D_{\Delta}(x_i, x_j) = \left[\sum_{k=1}^D \left(\log\left(\frac{x_{ik}}{g(\mathbf{x}_i)}\right) - \log\left(\frac{x_{jk}}{g(\mathbf{x}_j)}\right) \right)^2 \right]^{\frac{1}{2}}$$

- Invariant by permutation, by scaling, by subcompositional dominance

Total Jensen divergence

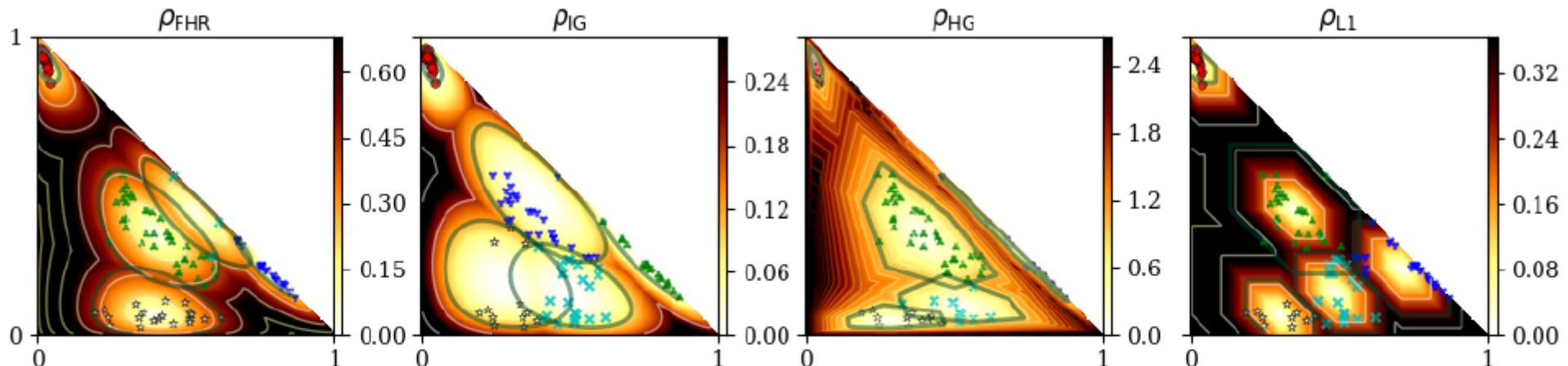
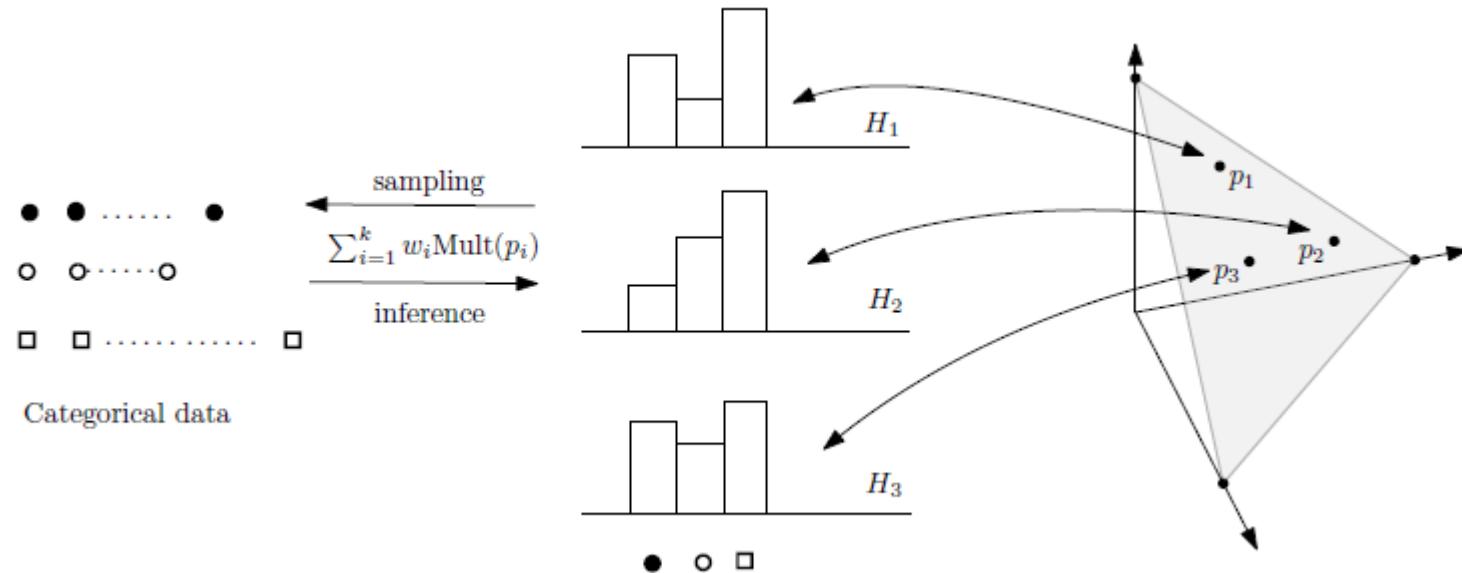
Invariant to axis rotation

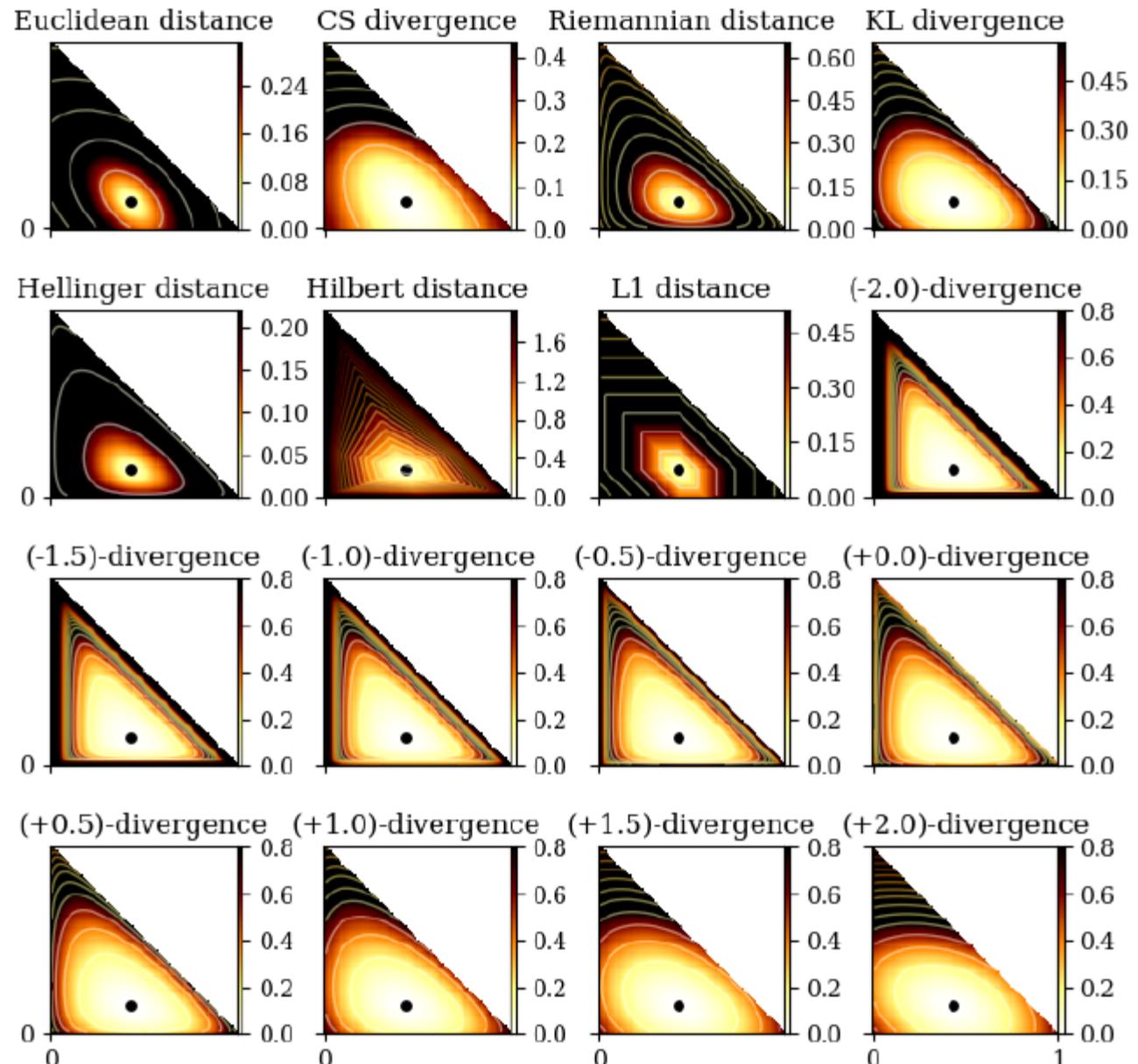


$$tB(p : q) = \rho_B(q)B(p : q), \quad \rho_B(q) = \sqrt{\frac{1}{1 + \langle \nabla F(q), \nabla F(q) \rangle}}$$

$$tJ_\alpha(p : q) = \rho_J(p, q)J_\alpha(p : q), \quad \rho_J(p, q) = \sqrt{\frac{1}{1 + \frac{(F(p) - F(q))^2}{\langle p - q, p - q \rangle}}}$$

Clustering categorical distributions

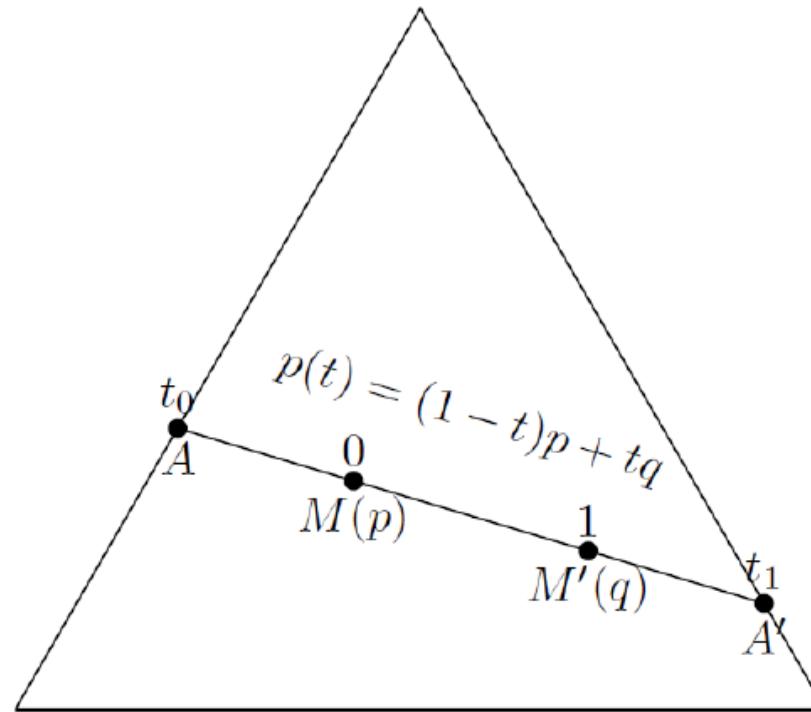




Reference point $(\frac{3}{7}, \frac{3}{7}, \frac{1}{7})$

Hilbert log cross-ratio metric

$$\rho_{\text{HG}}(M, M') = \begin{cases} \left| \log \frac{|A'M||AM'|}{|A'M'||AM|} \right|, & M \neq M', \\ 0 & M = M'. \end{cases}$$



Geodesics are straight lines but not unique

Isometry of Hilbert simplex geometry with a normed vector space $(\Delta^d, \rho_{\text{HG}}) \cong (V^d, \|\cdot\|_{\text{NH}})$

- ▶ $V^d = \{v \in \mathbb{R}^{d+1} : \sum_i v^i = 0\} \subset \mathbb{R}^{d+1}$
- ▶ Map $p = (\lambda^0, \dots, \lambda^d) \in \Delta^d$ to $v(x) = (v^0, \dots, v^d) \in V^d$:

$$v^i = \frac{1}{d+1} \left(d \log \lambda^i - \sum_{j \neq i} \log \lambda^j \right) = \log \lambda^i - \frac{1}{d+1} \sum_j \log \lambda^j.$$

$$\lambda^i = \frac{\exp(v^i)}{\sum_j \exp(v^j)}.$$

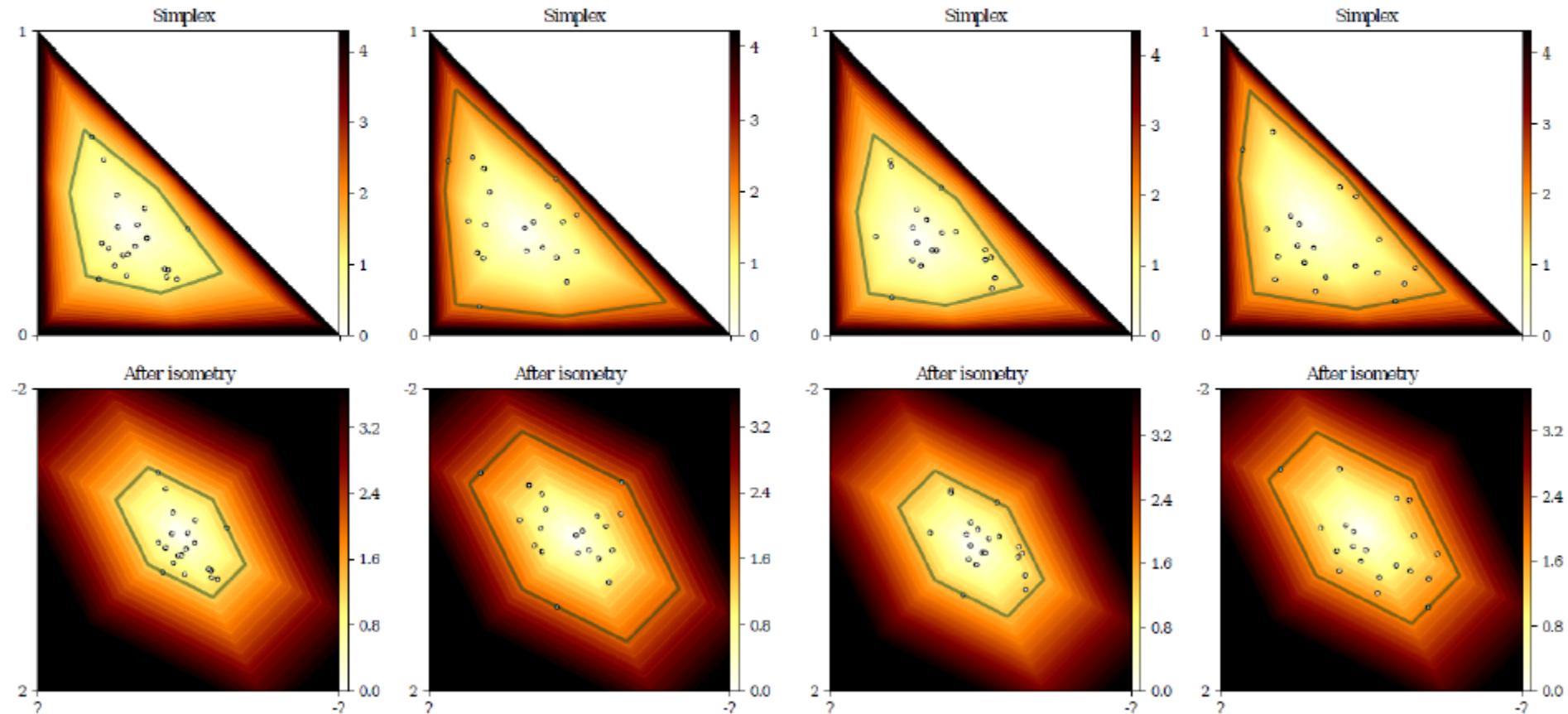
- ▶ Norm $\|\cdot\|_{\text{NH}}$ in V^d defined by the shape of its unit ball $B_V = \{v \in V^d : |v^i - v^j| \leq 1, \forall i \neq j\}$.

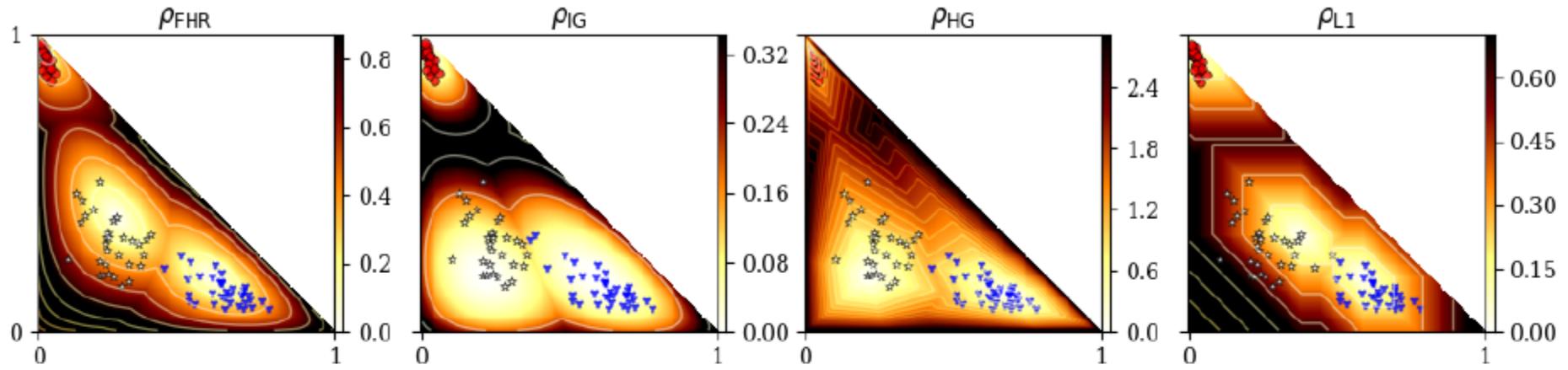
- ▶ Polytopal norm-induced distance:

$$\rho_V(v, v') = \|v - v'\|_{\text{NH}} = \inf \{\tau : v' \in \tau(B_V \oplus \{v\})\},$$

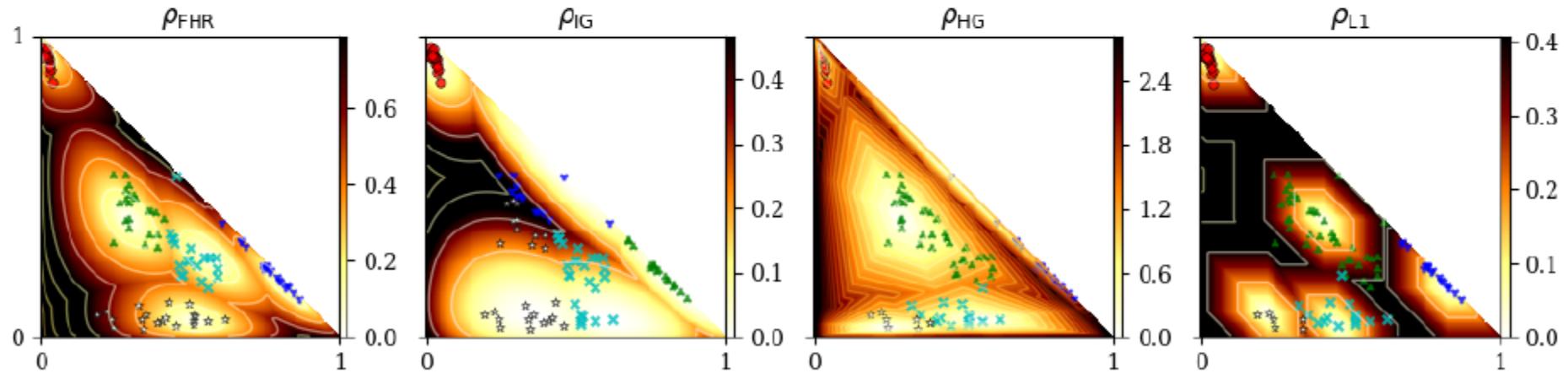
- ▶ Norm does not satisfy parallelogram law (no inner product)

Visualizing the isometry: $(\Delta^d, \rho_{\text{HG}}) \cong (V^d, \|\cdot\|_{\text{NH}})$





$k = 3$ clusters



$k = 5$ clusters

K-center clustering

Algorithm : A 2-approximation of the k -center clustering for any metric distance ρ .

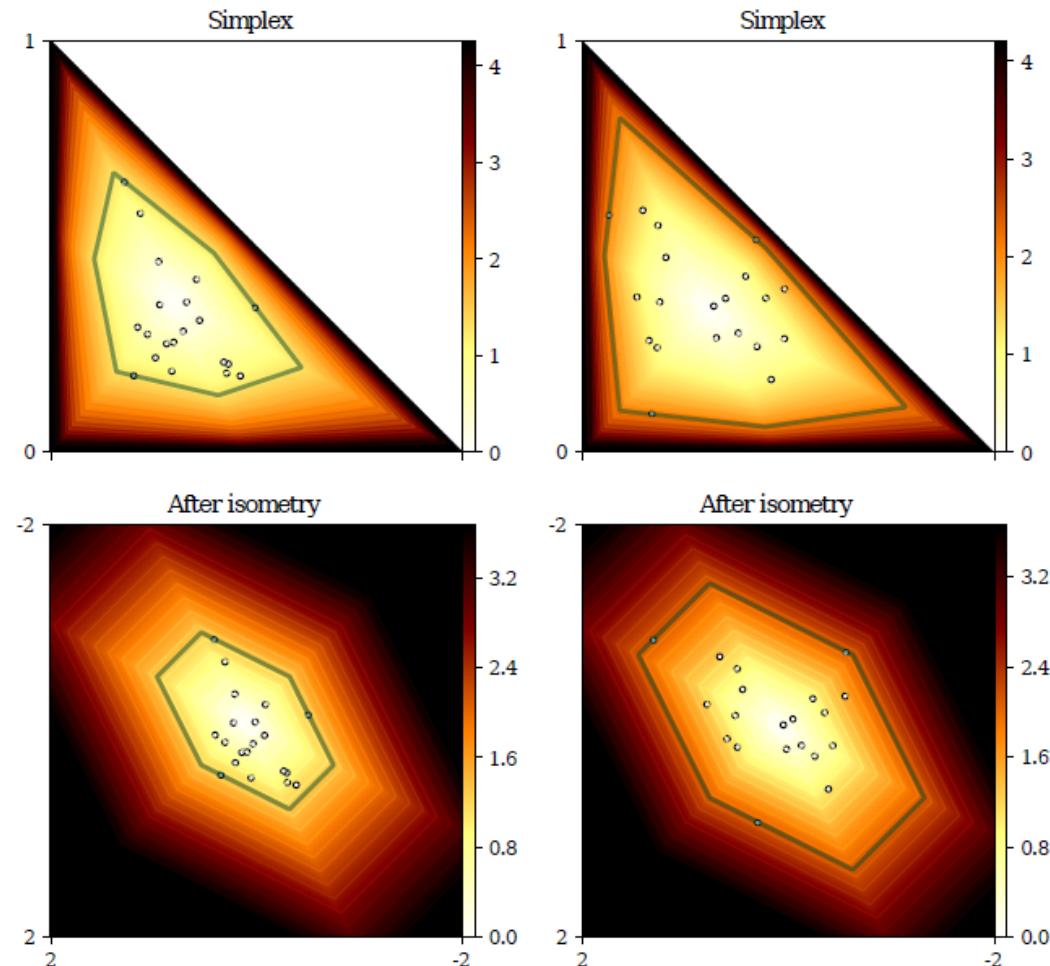
Data: A set Λ ; a number k of clusters; a metric distance ρ .

Result: A 2-approximation of the k -center clustering

```
1 begin
2    $c_1 \leftarrow \text{ARandomPointOf}(\Lambda);$ 
3    $C \leftarrow \{c_1\};$ 
4   for  $i = 2, \dots, k$  do
5      $c_i \leftarrow \arg \max_{p \in \Lambda} \rho(p, C);$ 
6      $C \leftarrow C \cup \{c_i\};$ 
7 Output  $C;$ 
```

- Guaranteed performance: 2-factor for any metric

Smallest enclosing ball



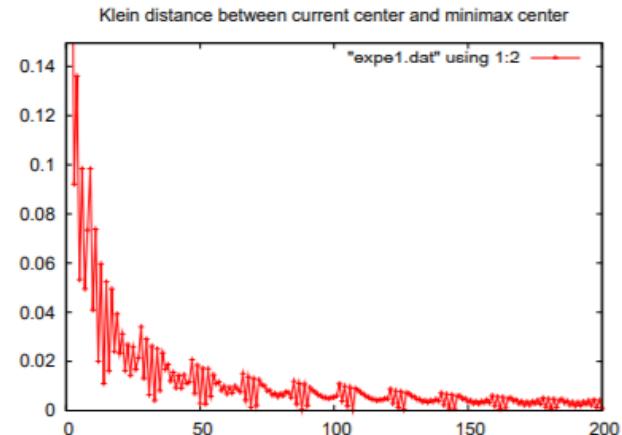
3 points on the border

Riemannian minimum enclosing ball

$a \#_t^M b$: point $\gamma(t)$ on the geodesic line segment $[ab]$ wrt M .

Algorithm GeoA

```
c1 ← choose randomly a point in  $\mathcal{P}$ ;  
for  $i = 2$  to  $l$  do  
    // farthest point from  $c_i$   
     $s_i \leftarrow \arg \max_{j=1}^n \rho(c_i, p_j);$   
    // update the center: walk on the geodesic line  
    // segment  $[c_i, p_{s_i}]$   
     $c_{i+1} \leftarrow c_i \#_{\frac{1}{i+1}}^M p_{s_i};$   
end  
// Return the SEB approximation  
return Ball( $c_l, r_l = \rho(c_l, \mathcal{P})$ );
```



Hyperbolic geometry:

$$\rho(p, q) = \operatorname{arccosh} \frac{1 - p^\top q}{\sqrt{(1 - p^\top p)(1 - q^\top q)}}$$

$$T_p(T_{-p}(p) \#_\alpha T_{-p}(q)) = p \#_\alpha q$$

$$T_p(x) = \frac{(1 - \|p\|^2)x + (\|x\|^2 + 2\langle x, p \rangle + 1)p}{\|p\|^2\|x\|^2 + 2\langle x, p \rangle + 1}$$

Positive-definite matrices:

$$\rho(P, Q) = \|\log(P^{-1}Q)\|_F = \sqrt{\sum_i \log^2 \lambda_i}$$

$$\gamma_t(P, Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \right)^t P^{\frac{1}{2}}$$

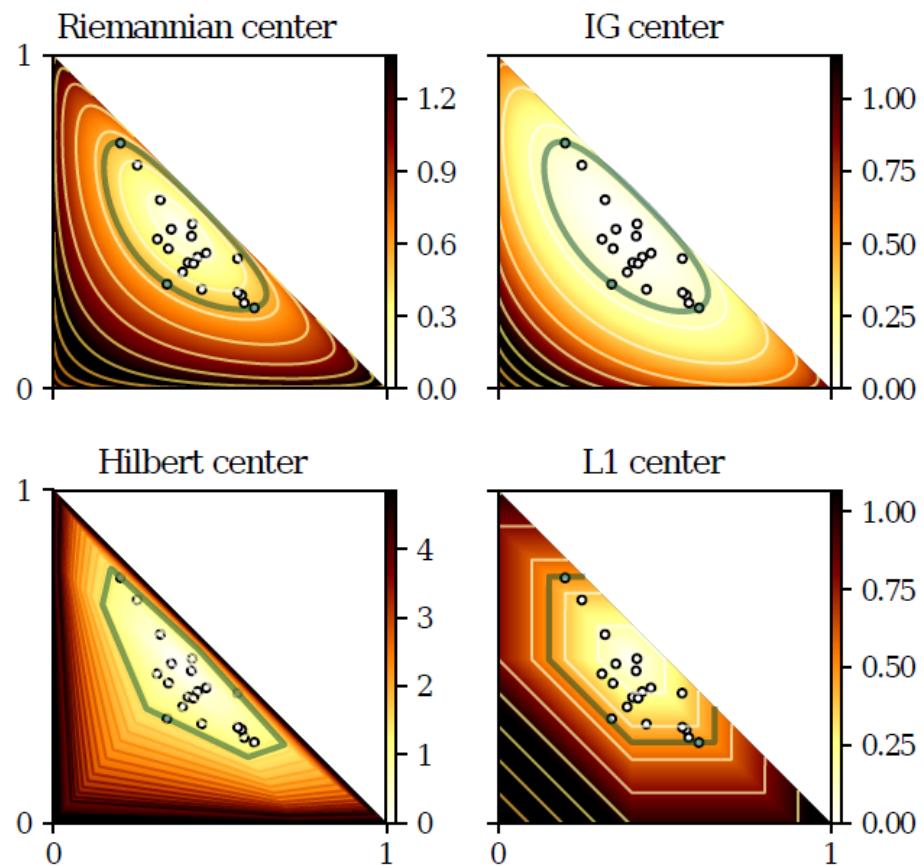
Approximating the smallest enclosing ball

Algorithm 4: Geodesic walk for approximating the Hilbert minimax center, generalizing [11]

Data: A set of points $p_1, \dots, p_n \in \Delta^d$. The maximum number T of iterations.

Result: $c \approx \arg \min_c \max_i \rho_{\text{HG}}(p_i, c)$

```
1 begin
2    $c_0 \leftarrow \text{ARandomPointOf}(\{p_1, \dots, p_n\});$ 
3   for  $t = 1, \dots, T$  do
4      $p \leftarrow \arg \max_{p_i} \rho_{\text{HG}}(p_i, c_{t-1});$ 
5      $c_t \leftarrow c_{t-1} \#_{1/(t+1)}^\rho p;$ 
6   Output  $c_T;$ 
```



K-means

| k | n | d | σ | ρ_{FHR} | ρ_{IG} | ρ_{HG} | ρ_{EUC} | ρ_{L1} |
|-----|-----|-----|----------|-----------------------------------|--------------------|-----------------------------------|---------------------|-----------------------------------|
| 3 | 50 | 9 | 0.5 | 0.62 ± 0.22 | 0.60 ± 0.22 | 0.71 ± 0.23 | 0.45 ± 0.20 | 0.54 ± 0.22 |
| | | | 0.9 | 0.29 ± 0.17 | 0.27 ± 0.16 | 0.39 ± 0.19 | 0.17 ± 0.13 | 0.25 ± 0.15 |
| | | 255 | 0.5 | 0.70 ± 0.25 | 0.69 ± 0.26 | 0.74 ± 0.25 | 0.37 ± 0.29 | 0.70 ± 0.26 |
| | 100 | 9 | 0.5 | 0.42 ± 0.25 | 0.35 ± 0.20 | 0.40 ± 0.19 | 0.03 ± 0.08 | 0.44 ± 0.26 |
| | | 9 | 0.5 | 0.63 ± 0.22 | 0.61 ± 0.22 | 0.71 ± 0.22 | 0.46 ± 0.19 | 0.56 ± 0.20 |
| | | | 0.9 | 0.29 ± 0.15 | 0.26 ± 0.14 | 0.38 ± 0.20 | 0.18 ± 0.12 | 0.24 ± 0.14 |
| 5 | 50 | 9 | 0.5 | 0.71 ± 0.26 | 0.69 ± 0.27 | 0.75 ± 0.25 | 0.31 ± 0.28 | 0.70 ± 0.27 |
| | | | 0.9 | 0.41 ± 0.26 | 0.33 ± 0.20 | 0.38 ± 0.18 | 0.02 ± 0.06 | 0.43 ± 0.26 |
| | | 255 | 0.5 | 0.64 ± 0.15 | 0.61 ± 0.14 | 0.70 ± 0.14 | 0.48 ± 0.14 | 0.57 ± 0.15 |
| | 100 | 9 | 0.5 | 0.31 ± 0.12 | 0.29 ± 0.12 | 0.41 ± 0.15 | 0.20 ± 0.09 | 0.26 ± 0.10 |
| | | 9 | 0.5 | 0.74 ± 0.17 | 0.72 ± 0.17 | 0.77 ± 0.16 | 0.41 ± 0.20 | 0.74 ± 0.17 |
| | | | 0.9 | 0.44 ± 0.17 | 0.37 ± 0.16 | 0.44 ± 0.15 | 0.04 ± 0.06 | 0.47 ± 0.17 |
| | 100 | 9 | 0.5 | 0.62 ± 0.14 | 0.61 ± 0.14 | 0.71 ± 0.14 | 0.46 ± 0.13 | 0.54 ± 0.14 |
| | | | 0.9 | 0.30 ± 0.10 | 0.27 ± 0.11 | 0.40 ± 0.13 | 0.19 ± 0.08 | 0.25 ± 0.09 |
| | | 255 | 0.5 | 0.73 ± 0.18 | 0.70 ± 0.18 | 0.75 ± 0.16 | 0.37 ± 0.20 | 0.73 ± 0.17 |
| | | 255 | 0.9 | 0.43 ± 0.16 | 0.35 ± 0.14 | 0.41 ± 0.12 | 0.03 ± 0.06 | 0.46 ± 0.18 |

K-center

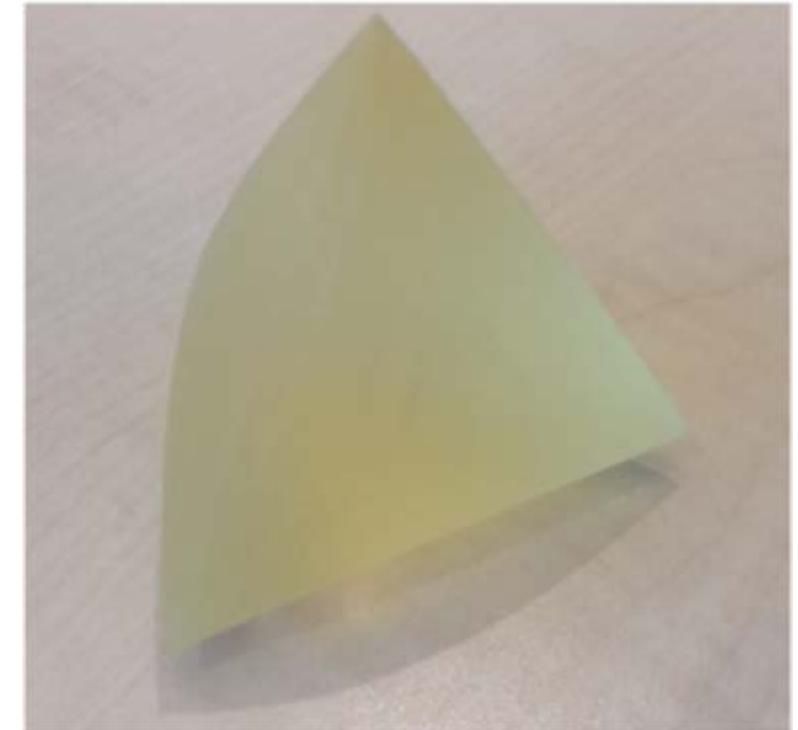
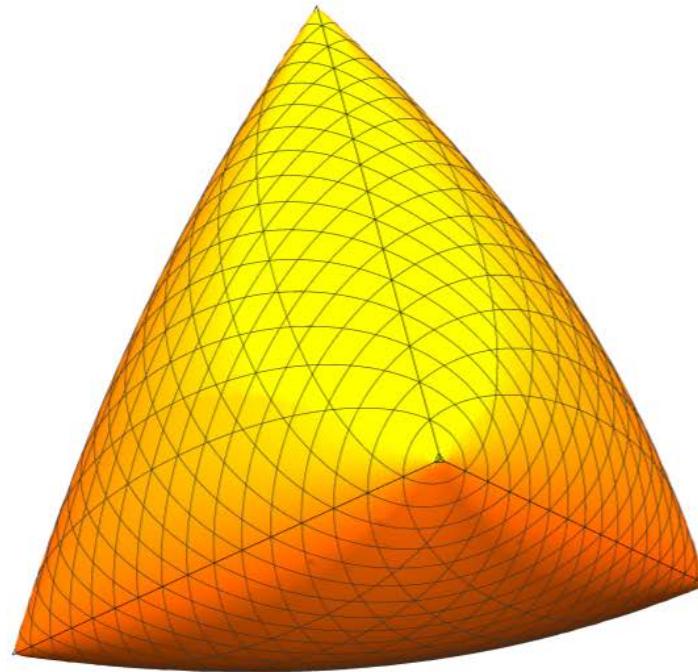
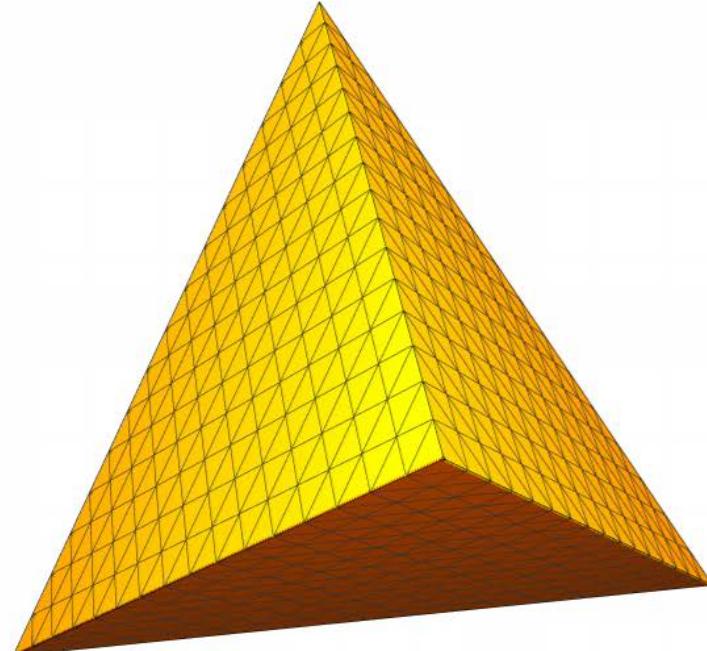
| k | n | d | σ | ρ_{FHR} | ρ_{IG} | ρ_{HG} | ρ_{EUC} | ρ_{L1} |
|-----|-----|-----|----------|-----------------------------------|-----------------------------------|-----------------------------------|---------------------|-----------------------------------|
| 50 | 9 | 0.5 | 0.5 | 0.87 ± 0.19 | 0.85 ± 0.19 | 0.92 ± 0.16 | 0.72 ± 0.22 | 0.80 ± 0.20 |
| | | 0.9 | 0.9 | 0.54 ± 0.21 | 0.51 ± 0.21 | 0.70 ± 0.23 | 0.36 ± 0.17 | 0.44 ± 0.19 |
| | 100 | 0.5 | 0.5 | 0.93 ± 0.16 | 0.92 ± 0.18 | 0.95 ± 0.14 | 0.89 ± 0.18 | 0.90 ± 0.19 |
| | | 0.9 | 0.9 | 0.76 ± 0.24 | 0.72 ± 0.26 | 0.82 ± 0.24 | 0.50 ± 0.28 | 0.76 ± 0.25 |
| 3 | 9 | 0.5 | 0.5 | 0.88 ± 0.17 | 0.86 ± 0.18 | 0.93 ± 0.14 | 0.70 ± 0.20 | 0.80 ± 0.20 |
| | | 0.9 | 0.9 | 0.53 ± 0.20 | 0.49 ± 0.19 | 0.70 ± 0.22 | 0.33 ± 0.14 | 0.41 ± 0.18 |
| | 100 | 0.5 | 0.5 | 0.93 ± 0.16 | 0.92 ± 0.17 | 0.95 ± 0.13 | 0.88 ± 0.19 | 0.93 ± 0.16 |
| | | 0.9 | 0.9 | 0.81 ± 0.22 | 0.75 ± 0.24 | 0.83 ± 0.22 | 0.47 ± 0.28 | 0.79 ± 0.22 |
| 50 | 9 | 0.5 | 0.5 | 0.82 ± 0.13 | 0.81 ± 0.13 | 0.89 ± 0.12 | 0.67 ± 0.13 | 0.75 ± 0.13 |
| | | 0.9 | 0.9 | 0.50 ± 0.13 | 0.47 ± 0.13 | 0.66 ± 0.15 | 0.34 ± 0.11 | 0.40 ± 0.12 |
| | 100 | 0.5 | 0.5 | 0.92 ± 0.11 | 0.91 ± 0.12 | 0.93 ± 0.11 | 0.87 ± 0.13 | 0.92 ± 0.12 |
| | | 0.9 | 0.9 | 0.77 ± 0.15 | 0.71 ± 0.17 | 0.85 ± 0.17 | 0.54 ± 0.19 | 0.74 ± 0.16 |
| 5 | 9 | 0.5 | 0.5 | 0.83 ± 0.12 | 0.81 ± 0.13 | 0.89 ± 0.11 | 0.67 ± 0.11 | 0.76 ± 0.13 |
| | | 0.9 | 0.9 | 0.48 ± 0.12 | 0.46 ± 0.12 | 0.66 ± 0.15 | 0.33 ± 0.09 | 0.39 ± 0.10 |
| | 100 | 0.5 | 0.5 | 0.93 ± 0.10 | 0.92 ± 0.11 | 0.94 ± 0.09 | 0.89 ± 0.11 | 0.92 ± 0.11 |
| | | 0.9 | 0.9 | 0.81 ± 0.14 | 0.74 ± 0.15 | 0.84 ± 0.16 | 0.52 ± 0.19 | 0.79 ± 0.14 |

generator 1

Clustering correlation matrices (elliptope)

- Covariance matrices with unit diagonal, correlation coefficients

$$\mathcal{C}^d = \{ C_{d \times d} : C \succ 0; C_{ii} = 1, \forall i \}$$



Some distances between correlation matrices

- Hilbert log cross-ratio distance

$$\rho_{\text{HG}}(C_1, C_2) = \left| \log \frac{\|C_1 - C'_2\| \|C'_1 - C_2\|}{\|C_1 - C'_1\| \|C_2 - C'_2\|} \right|.$$

- L1-norm
- L2-norm
- Log-det divergence

$$\rho_{\text{LD}}(C_1, C_2) = \text{tr}(C_1 C_2^{-1}) - \log |C_1 C_2^{-1}| - d.$$

| ν_1 | ν_2 | ρ_{HG} | ρ_{EUC} | ρ_{L1} | ρ_{LD} |
|---------|---------|-----------------------------------|---------------------|--------------------|-----------------------------------|
| 4 | 10 | 0.62 ± 0.22 | 0.57 ± 0.21 | 0.56 ± 0.22 | 0.58 ± 0.22 |
| 4 | 30 | 0.85 ± 0.18 | 0.80 ± 0.20 | 0.81 ± 0.19 | 0.82 ± 0.20 |
| 4 | 50 | 0.89 ± 0.17 | 0.87 ± 0.17 | 0.86 ± 0.18 | 0.88 ± 0.18 |
| 5 | 10 | 0.50 ± 0.21 | 0.49 ± 0.21 | 0.48 ± 0.20 | 0.47 ± 0.21 |
| 5 | 30 | 0.77 ± 0.20 | 0.75 ± 0.21 | 0.75 ± 0.21 | 0.75 ± 0.21 |
| 5 | 50 | 0.84 ± 0.19 | 0.82 ± 0.19 | 0.82 ± 0.20 | 0.84 ± 0.18 |

$$(p_1 p_2)_\alpha^m = \begin{cases} \mu_\alpha^m = (1 - \alpha)\mu_1 + \alpha\mu_2 \\ v_\alpha^m = (1 - \alpha)v_1 + \alpha v_2 + \alpha(1 - \alpha)(\mu_1 - \mu_2)^2 \end{cases}$$

$$(p_1 p_2)_\alpha^e = \begin{cases} \mu_\alpha^e = \frac{(1-\alpha)\mu_1 v_2 + \alpha\mu_2 v_1}{(1-\alpha)v_2 + \alpha v_1} \\ v_\alpha^e = \frac{v_1 v_2}{(1-\alpha)v_2 + \alpha v_1} \end{cases}$$

$$(p_1 p_2)_\alpha^m = \begin{cases} \mu_\alpha^m = (1 - \alpha)\mu_1 + \alpha\mu_2 \\ \Sigma_\alpha^m = \bar{\Sigma}_\alpha + (1 - \alpha)\mu_1 \mu_1^\top - \alpha\mu_2 \mu_2^\top - \bar{\mu}_\alpha \bar{\mu}_\alpha^\top \end{cases}$$

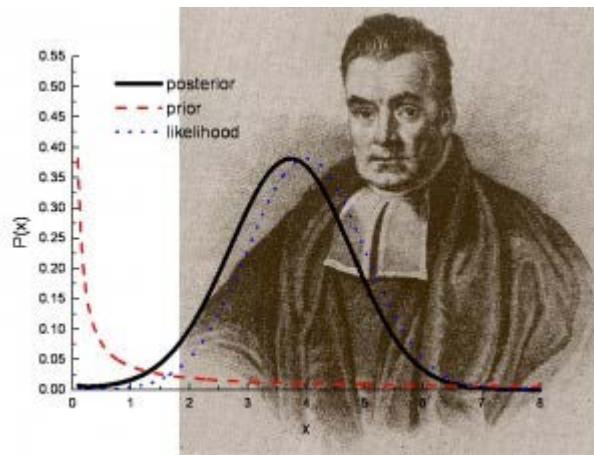
$$(p_1 p_2)_\alpha^e = \begin{cases} \mu_\alpha^e = \Sigma_\alpha^e ((1 - \alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2) \\ \Sigma_\alpha^e = ((1 - \alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1} \end{cases}$$

Bayesian Binary/Multiple Hypothesis testing



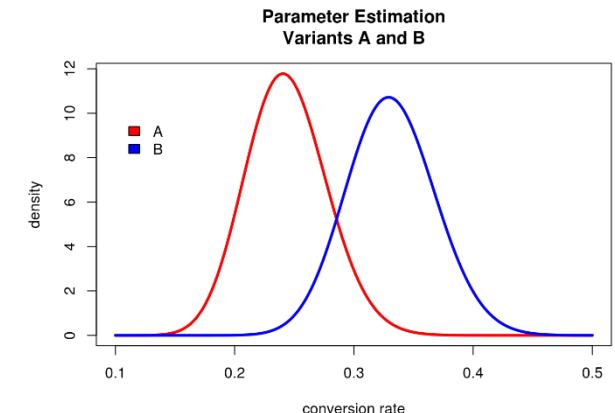
Bayes' error (probability of error), Total Variation,
Chernoff Information and its generalizations

Detecting signal from noise



Exponential Family Manifold

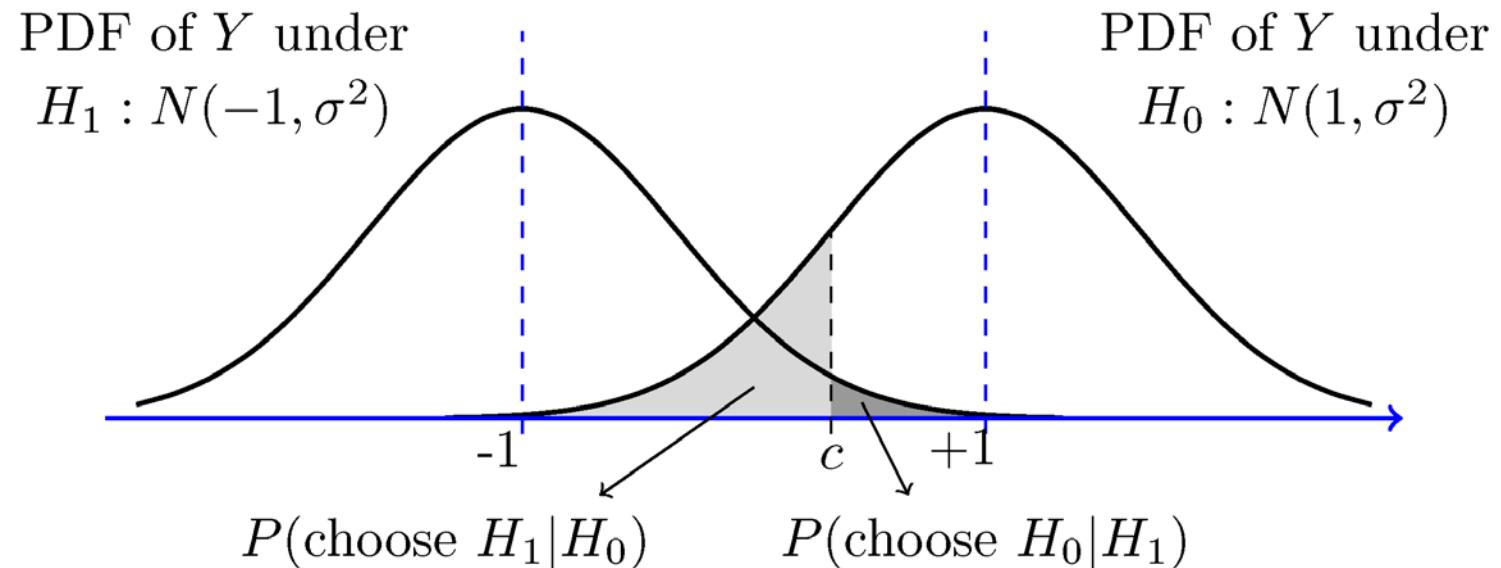
Frank Nielsen



An information-geometric characterization of Chernoff information, IEEE Signal Processing Letters (2013)
Hypothesis Testing, Information Divergence and Computational Geometry. GSI 2013
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)
Computational Information Geometry for Binary Classification of High-Dimensional Random Tensors, Entropy (2018)

Bayesian binary hypothesis testing

- Decide given an iid sample set whether it emanates from the distribution of the null hypothesis H_0 or the alternative hypothesis H_1
-> unavoidable **probability of error**

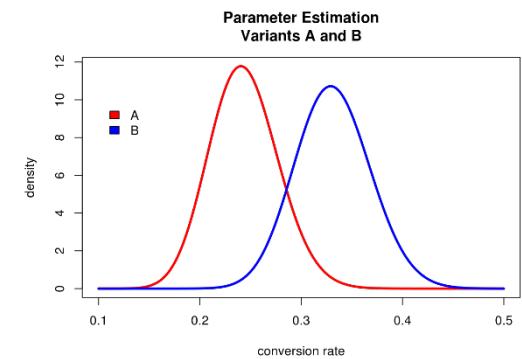


Among the many decision rules, the best rule is the **Maximum A Posteriori (MAP)** rule

$$P(H_0 | X = x) \geq P(H_1 | X = x)$$

Probability of error

(Bayes' error for diagonal cost matrix)



- Confusion matrix
- Cost design matrix, where errors uniformly account (diagonal matrix)
- Probability of error

$$P_{\text{error}} = P(\text{choose } H_1 | H_0) P(H_0) + P(\text{choose } H_0 | H_1) P(H_1)$$

- A priori probabilities of classes: $w_0 = P(H_0)$ and $w_1 = P(H_1)$
- Theorem: MAP rule minimizes the probability of error among all decision rules:

$$\text{MAP}(x) = \operatorname{argmax}_{i \in \{1, \dots, n\}} w_i p_i(x)$$

Class conditional probabilities

Probability of error with equal priors ($w_1=w_2=1/2$)

$$P_{error} = \int_{x \in \mathcal{X}} p(x) \min(\Pr(H_1|x), \Pr(H_2|x)) d\nu(x)$$

From Bayes' rule: $\Pr(H_i|X=x) = \frac{\Pr(H_i)\Pr(X=x|H_i)}{\Pr(X=x)} = \frac{w_i p_i(x)}{p(x)}$

It follows that we have:

$$P_{error} = \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) d\nu(x)$$

This is also called **histogram intersection similarity** in computer vision

Bounding the probability of error



Trick:

$$\boxed{\min(a, b) \leq \min_{\alpha \in (0,1)} a^\alpha b^{1-\alpha}} \text{ for } a, b > 0, \text{ upper bound } P_e:$$

$$\begin{aligned} P_e &= \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) d\nu(x) \\ &\leq \frac{1}{2} \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) d\nu(x). \end{aligned}$$

Define **Chernoff information** :

$$C(P_1, P_2) = -\log \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) d\nu(x) \geq 0,$$

For alpha=1/2, we get the Bhattacharyya distance, skewed Bhattacharyya distance: $B_\alpha(p, q) = -\ln \int_x p^\alpha(x) q^{1-\alpha}(x) dx$

Then it comes that

$$P_e \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C(P_1, P_2)} \leq e^{-C(P_1, P_2)}$$

Chernoff information: A statistical distance

- For m iid samples

$$P_{\text{correct}}^m = 1 - P_{\text{error}}^m = 1 - P_e^m$$

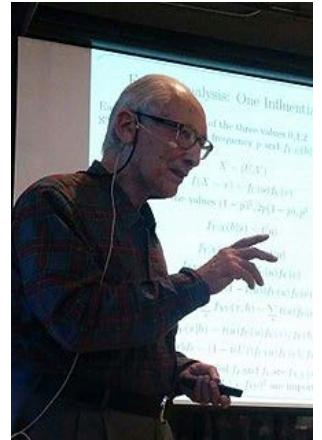
- **Asymptotic regime** when $m \rightarrow \infty$

$$\alpha = -\frac{1}{m} \log P_e^m$$

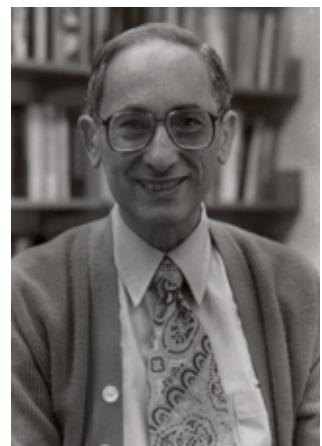
- **Best error exponent:**

$$P_e \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C(P_1, P_2)} \leq e^{-C(P_1, P_2)}$$

$$C(P_1, P_2) = -\log \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) d\nu(x) \geq 0,$$



Herman Chernoff
(1923, 95 yo)
pic 2015



Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations,"
Ann. Math. Statist., vol. 23, pp. 493–507, 1952

$$D(Y) = -\log \left[\inf_{0 < t < 1} \int [f_1(x)]^t [f_0(x)]^{1-t} d\nu(x) \right]$$

Hypothesis testing: exponential family manifold

The manifold of an exponential family is dually flat

By using the bijection between log-likelihood and Bregman divergence:

$$\log p_{\theta_i}(x) = -B^*(t(x) : \eta_i) + F^*(t(x)) + k(x), \quad \eta_i = \nabla F(\theta_i)$$

The map rule induces an **additive Bregman Voronoi diagram**

$$\begin{aligned} \text{MAP}(x) &= \operatorname{argmax}_{i \in \{1, \dots, n\}} w_i p_i(x) \\ &= \arg \min_{i \in \{1, \dots, n\}} B^*(t(x) : \eta_i) - \log w_i \end{aligned}$$

Geometry of the best error exponent for exponential families

$$c_\alpha(P_{\theta_1} : P_{\theta_2}) = \int p_{\theta_1}^\alpha(x) p_{\theta_2}^{1-\alpha}(x) d\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2)),$$

Jensen divergence:

$$J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1 - \alpha) F(\theta_2) - F(\theta_{12}^{(\alpha)}),$$

Theorem: At best exponent, the Chernoff information amounts to an equivalent Bregman divergence:

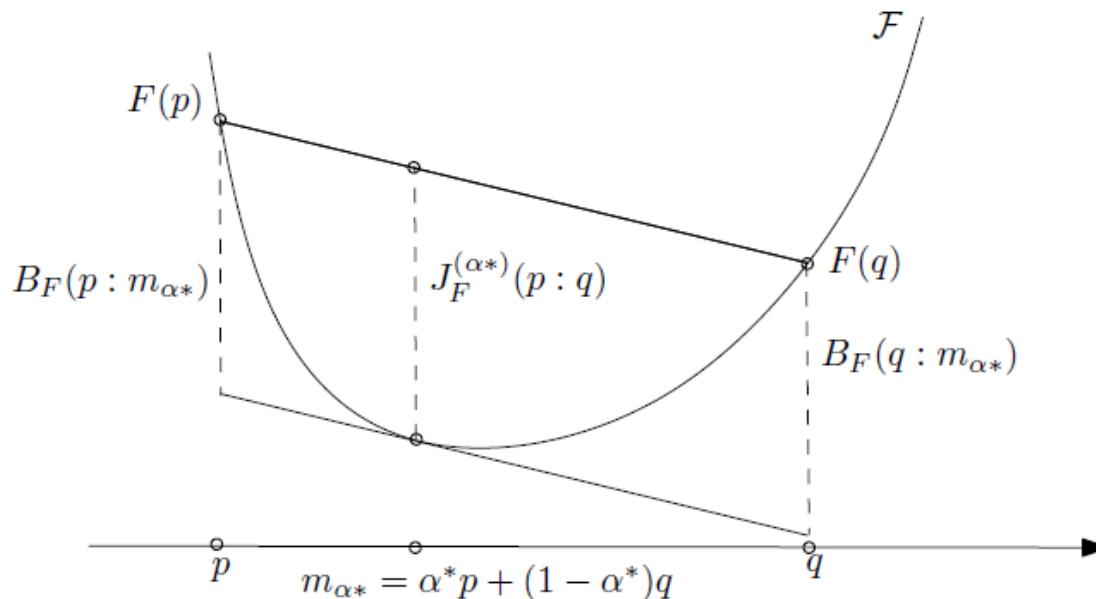
$$C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})$$

Maximizing skew Jensen divergence yields a Bregman divergence

$$\alpha^* = \arg \max_{0 < \alpha < 1} J_F^{(\alpha)}(p : q)$$

$$J_F^{(\alpha^*)}(p : q) = B_F(p : m_{\alpha^*}) = B_F(q : m_{\alpha^*})$$

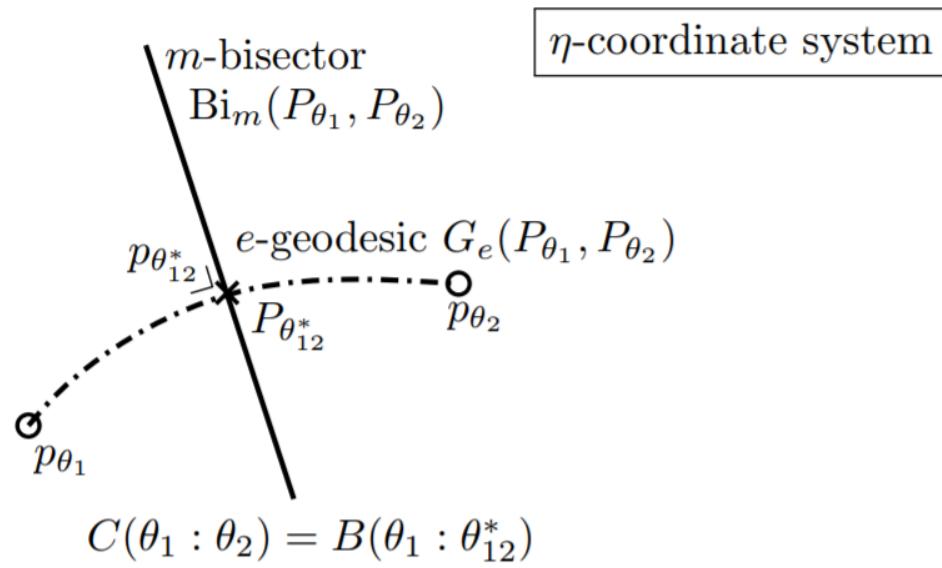
$m_\alpha = \alpha p + (1 - \alpha)q$: α -mixing of p and q .



Bayesian hypothesis testing: A geometric characterization of the best error exponent

Dually flat Exponential Family Manifold (EFM)

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

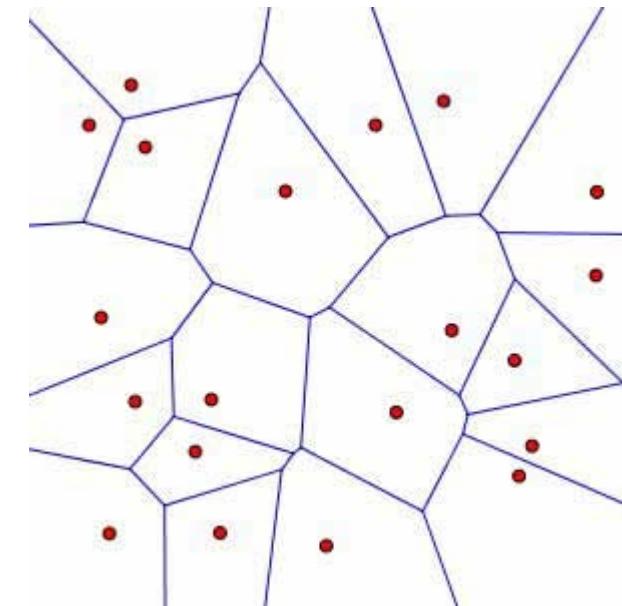


This characterization yields to an **exact closed-form solution in 1D EFs**,
and a **simple geodesic bisection** search for arbitrary dimension

Multiple hypothesis testing

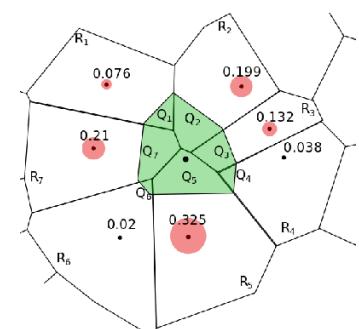
- Minimum pairwise Chernoff information distance

$$C(P_1, \dots, P_n) = \min_{i,j \neq i} C(P_i, P_j)$$



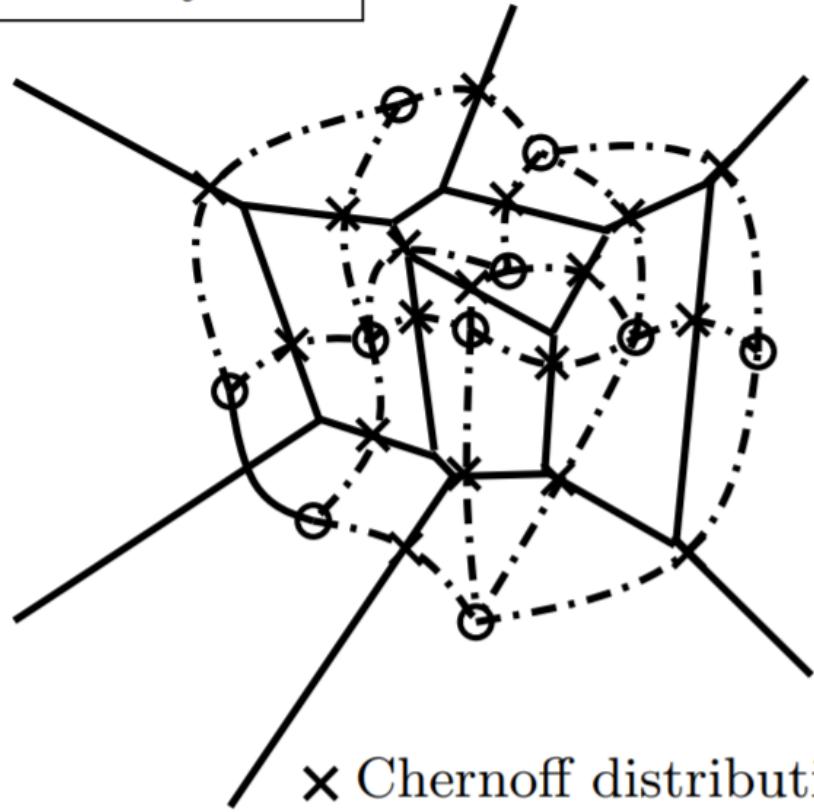
$$P_e^m \leq e^{-mC(P_{i^*}, P_{j^*})}, \quad (i^*, j^*) = \arg \min_{i,j \neq i} C(P_i, P_j)$$

- In the (additive) **Bregman Voronoi diagram**, check only the **natural neighbors** (with Voronoi cells sharing a common facet)



Multiple hypothesis testing on EFM

η -coordinate system



Bregman Voronoi diagram is affine in the eta (moment/expectation) coordinate system

Natural neighbors

Link between the Probability of error and the Total Variation (TV)



Use the trick

$$\min(a, b) = \frac{a+b}{2} - \frac{1}{2}|b-a|,$$

$$P_{\text{error}} = \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) d\nu(x)$$

$$P_e = \frac{1}{2} - \text{TV}(w_1 p_1, w_2 p_2).$$

$$P_e = \frac{1}{2} (1 - \text{TV}(p_1, p_2)). \quad (\text{same weights here})$$

Computing Total Variation can be difficult...

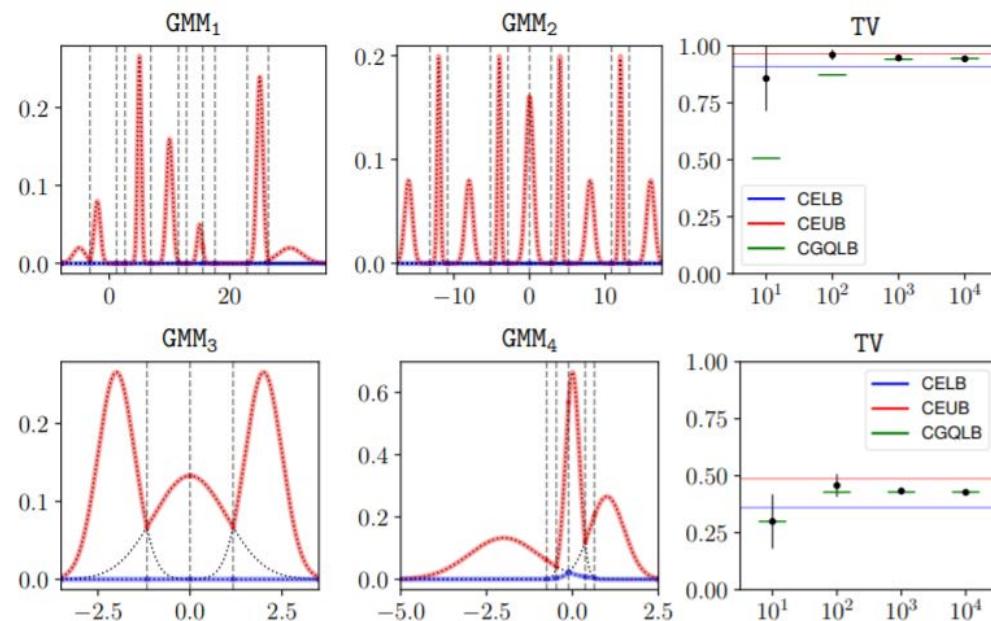
Pe between two multivariate Gaussians with same positive semi-definite covariance matrix

$$\text{TV}(p_1, p_2) = \frac{1}{2} \left| \operatorname{erf} \left(\frac{x_1 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \operatorname{erf} \left(\frac{x_1 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right| + \frac{1}{2} \left| \operatorname{erf} \left(\frac{x_2 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \operatorname{erf} \left(\frac{x_2 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right|.$$

$$P_e = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{1}{2\sqrt{2}} \|(\Sigma^+)^{\frac{1}{2}}(\mu_2 - \mu_1)\| \right)$$

$$\frac{1}{2}|a - b| = \frac{a+b}{2} - \min(a, b) = \max(a, b) - \frac{a+b}{2},$$

$$\begin{aligned} \text{TV}(p, q) &= \int_{\mathcal{X}} \left(\frac{p(x) + q(x)}{2} - \min(p(x), q(x)) \right) d\mu(x), \\ &:= 1 - \int_{\mathcal{X}} \min(p(x), q(x)) d\mu(x) = \int_{\mathcal{X}} \max(p(x), q(x)) d\mu(x) - 1. \end{aligned}$$



From geometric mean to other means

Remember the trick:

Geometric weighted mean
is greater than the minimum

$$\begin{aligned} P_e &= \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) d\nu(x) \\ &\leq \frac{1}{2} \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) d\nu(x). \end{aligned}$$

Interness property of any mean M $\min(a, b) \leq M(a, b) \leq \max(a, b)$

Consider quasi arithmetic means for a strictly monotone function f
(with well-defined inverse function)

$$M_f(a, b; \alpha) = f^{-1}(\alpha f(a) + (1 - \alpha) f(b))$$

Chernoff information with quasi-arithmetic means

$$M_f(a, b; \alpha) = f^{-1}(\alpha f(a) + (1 - \alpha)f(b))$$

Definition 2. The Chernoff-type information for a strictly monotonous function f is defined by:

$$\begin{aligned} C_f(p_1, p_2) &= -\log \rho_*^f(p_1, p_2) \\ &= \max_{\alpha \in [0,1]} -\log \int M_f(p_1(x), p_2(x); \alpha) dx \geq 0. \end{aligned}$$

Geometric means and exponential families

we consider the *geometric mean* obtained for $f(x) = \log x$. Since $p_1(x) = \exp(x^\top \theta_1 - F(\theta_1))$ and $p_2(x) = \exp(x^\top \theta_2 - F(\theta_2))$ belong to the exponential families, we get:

$$M_f(w_1 p_1(x), w_2 p_2(x); \alpha) = e^{\alpha \log w_1 p_1(x) + (1-\alpha) \log w_2 p_2(x)}, \quad (60)$$

$$= w_1^\alpha w_2^{1-\alpha} p_1^\alpha(x) p_2^{1-\alpha}(x). \quad (61)$$

$$f^{-1}(m_\alpha(x; \theta_1, \theta_2)) = e^{F(\alpha\theta_1 + (1-\alpha)\theta_2) - \alpha F(\theta_1) - (1-\alpha) F(\theta_2)} \\ \times p(x; \alpha\theta_1 + (1-\alpha)\theta_2),$$

$$= e^{-J_F^{(\alpha)}(\theta_1, \theta_2)} \underbrace{p(x; \alpha\theta_1 + (1-\alpha)\theta_2)}_{\theta_{12}^{(\alpha)}}$$

1 since natural parameter space is convex

Thus

$$P_e \leq w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1, \theta_2)} \int p(x; \alpha\theta_1 + (1-\alpha)\theta_2) dx.$$

$$P_e \leq \min_{\alpha \in [0,1]} w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1, \theta_2)}.$$

Harmonic mean for Cauchy distributions

- Cauchy family is a location-scale family

$$p(x; s) = \frac{1}{\pi} \frac{s}{x^2 + s^2}$$

- Choose harmonic mean with generator

$$f(x) = f^{-1}(x) = \frac{1}{x}$$

$$\begin{aligned} P_e &\leq \int M_H\left(\frac{1}{2}p_1(x), \frac{1}{2}p_2(x); \alpha\right) dx, \\ &\leq \frac{1}{2} \int \frac{p_1(x)p_2(x)}{(1-\alpha)p_1(x) + \alpha p_2(x)} dx, \\ &\leq \frac{1}{2} \int \frac{\frac{s_1}{\pi(x^2+s_1^2)} \frac{s_2}{\pi(x^2+s_2^2)}}{(1-\alpha)\frac{s_1}{\pi(x^2+s_1^2)} + \alpha\frac{s_2}{\pi(x^2+s_2^2)}} dx, \\ &\leq \frac{1}{2} \int \frac{s_1 s_2}{\pi((1-\alpha)s_1(x^2+s_2^2) + \alpha s_2(x^2+s_1^2))} dx, \\ &\leq \frac{1}{2} \int \frac{s_1 s_2}{\pi(((1-\alpha)s_1 + \alpha s_2)x^2 + (1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2)} dx, \\ &\leq \frac{1}{2} \frac{s_1 s_2}{((1-\alpha)s_1 + \alpha s_2)s_\alpha} \underbrace{\int \frac{1}{\pi} \frac{s_\alpha}{x^2 + s_\alpha^2} dx}_{=1}, \end{aligned}$$

Probability of error for Cauchy hypothesis

$$\text{TV}(p_1, p_2) = \frac{2}{\pi} \left(\arctan \left(\sqrt{\frac{s_2}{s_1}} \right) - \arctan \left(\sqrt{\frac{s_1}{s_2}} \right) \right).$$

$$\begin{aligned} P_e &= \frac{1}{2} - \frac{1}{\pi} \left(\arctan \left(\sqrt{\lambda} \right) - \arctan \left(\sqrt{1/\lambda} \right) \right), \\ &= 1 - \frac{2}{\pi} \arctan \left(\sqrt{\lambda} \right), \quad \lambda = \frac{s_2}{s_1}. \end{aligned}$$

$$P_e \leq \frac{1}{2} \frac{s_1 s_2}{((1-\alpha)s_1 + \alpha s_2) \sqrt{\frac{(1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2}{(1-\alpha)s_1 + \alpha s_2}}}.$$

HT with Pearson type VII distributions

$$p(x; \mu, \Sigma, \lambda) = \pi^{-\frac{d}{2}} \frac{\Gamma(\lambda)}{\Gamma(\lambda - \frac{d}{2})} |\Sigma|^{-\frac{1}{2}} \left(1 + (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)^{-\lambda},$$

Consider the α -weighted f -mean with $f(x) = x^{-\frac{1}{\lambda}}$, for prescribed $\lambda > \frac{d}{2}$ (and $f^{-1}(x) = x^{-\lambda}$).

$$\begin{aligned} P_e &\leq \frac{1}{2} (\alpha |\Sigma_1|^{\frac{1}{2\lambda}} + (1 - \alpha) |\Sigma_2|^{\frac{1}{2\lambda}})^{-\lambda} |\Sigma_\alpha|^{\frac{1}{2}} \underbrace{\int p(x; \Sigma_\alpha) dx}_{=1}, \\ &= \frac{1}{2} (\alpha |\Sigma_1|^{\frac{1}{2\lambda}} + (1 - \alpha) |\Sigma_2|^{\frac{1}{2\lambda}})^{-\lambda} |\Sigma_\alpha|^{\frac{1}{2}}, \end{aligned}$$

since $\Sigma_\alpha \in \Theta$.

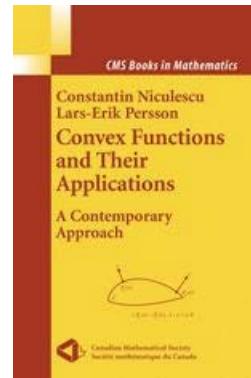
New Bregman divergences from abstract means

A function is **(M,N)-convex (comparative convexity)** iff

$$F(M(p, q)) \leq N(F(p), F(q)), \quad \forall p, q \in \mathcal{X}$$

A mean is **regular** if it is:

1. homogeneous
2. symmetric,
3. continuous
4. increasing in each variable.



Skewed (M,N)-Jensen-divergence for regular means:

$$J_F^{M,N}(p, q) = N(F(p), F(q))) - F(M(p, q)) \quad J_{F,\alpha}^{M,N}(p : q) \geq 0$$

Example of non-regular means: Lehmer mean (also Bajraktarevic mean) $L_\delta(x_1, \dots, x_n; w_1, \dots, w_n) = \frac{\sum_{i=1}^n w_i x_i^{\delta+1}}{\sum_{i=1}^n w_i x_i^\delta}$

(M,N)-Bregman divergences from comparative convexity

(M,N) Bregman divergences obtained in the scaled limit case of Jensen divergence:

$$B_F^{M,N}(p : q) = \lim_{\alpha \rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p : q) = \lim_{\alpha \rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} (N_\alpha(F(p), F(q))) - F(M_\alpha(p, q)))$$

Quasi-arithmetic Bregman divergences obtained

$$B_F^{\rho,\tau}(p : q) = \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p) - \rho(q)}{\rho'(q)} F'(q).$$

$$B_F^{\rho,\tau}(p : q) = \kappa_\tau(F(q) : F(p)) - \kappa_\rho(q : p) F'(q)$$

For example, the power mean Bregman divergences:

$$B_F^{\delta_1, \delta_2}(p : q) = \frac{F^{\delta_2}(p) - F^{\delta_2}(q)}{\delta_2 F^{\delta_2-1}(q)} - \frac{p^{\delta_1} - q^{\delta_1}}{\delta_1 q^{\delta_1-1}} F'(q)$$

$$M_f(p, q) = f^{-1} \left(\frac{f(p) + f(q)}{2} \right)$$

| Type | γ | $\kappa_\gamma(x : y) = \frac{\gamma(y) - \gamma(x)}{\gamma'(x)}$ |
|---------------------------|-------------------------------|---|
| A | $\gamma(x) = x$ | $y - x$ |
| G | $\gamma(x) = \log x$ | $x \log \frac{y}{x}$ |
| H | $\gamma(x) = \frac{1}{x}$ | $x^2 \left(\frac{1}{y} - \frac{1}{x} \right)$ |
| $P_\delta, \delta \neq 0$ | $\gamma_\delta(x) = x^\delta$ | $\frac{y^\delta - x^\delta}{\delta x^{\delta-1}}$ |

Jensen-Shannon divergences

$$\begin{aligned}\text{JS}(p; q) &:= \frac{1}{2} \left(\text{KL} \left(p : \frac{p+q}{2} \right) + \text{KL} \left(q : \frac{p+q}{2} \right) \right), \\ &= \frac{1}{2} \int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q} \right) d\mu.\end{aligned}$$

$$\text{JS}(p; q) = h \left(\frac{p+q}{2} \right) - \frac{h(p) + h(q)}{2}.$$

Jensen-Shannon divergence is the total divergence to the average divergence
Always bounded by $\log 2$, and the square root of JS is a metric

(M,N) Jensen-Shannon divergences

Symmetrizing the KL divergence

Resistor average divergence

$$\begin{aligned}\frac{1}{R(p;q)} &= \frac{1}{2} \left(\frac{1}{\text{KL}(p : q)} + \frac{1}{\text{KL}(q : p)} \right), \\ R(p;q) &= \frac{2(\text{KL}(p : q) + \text{KL}(q : p))}{\text{KL}(p : q)\text{KL}(q : p)} = \frac{2J(p;q)}{\text{KL}(p : q)\text{KL}(q : p)}.\end{aligned}$$

Jeffreys divergence

$$J(p;q) := \text{KL}(p : q) + \text{KL}(q : p) = \int (p - q) \log \frac{p}{q} d\mu = J(q;p).$$

Jensen-Bregman divergence as a Jensen divergence

$$\begin{aligned}\text{JB}_F(\theta : \theta') &:= \frac{1}{2} \left(B_F \left(\theta : \frac{\theta + \theta'}{2} \right) + B_F \left(\theta' : \frac{\theta + \theta'}{2} \right) \right), \\ &= \frac{F(\theta) + F(\theta')}{2} - F \left(\frac{\theta + \theta'}{2} \right) =: J_F(\theta : \theta'),\end{aligned}$$

$$\begin{aligned}\text{JB}_F^\alpha(\theta : \theta') &:= (1 - \alpha)B_F \left(\theta : (\theta\theta')_\alpha \right) + \alpha B_F \left(\theta' : (\theta\theta')_\alpha \right), \\ &= (F(\theta)F(\theta'))_\alpha - F((\theta\theta')_\alpha) =: J_F^\alpha(\theta : \theta'),\end{aligned}$$

$$\text{in} \; M^h_\alpha(x,y) \!:=\! h^{-1}\left((1-\alpha)h(x)+\alpha h(y)\right) \in I.$$

$$A_{\alpha}(x,y)=(1-\alpha)x+\alpha y$$

$$G_{\alpha}(x,y)=x^{1-\alpha}y^{\alpha}$$

$$H_{\alpha}(x,y)=\tfrac{xy}{(1-\alpha)y+\alpha x}$$

M-statistical mixture

$$(pq)_\alpha^M(x) := \frac{M_\alpha(p(x), q(x))}{Z_\alpha^M(p : q)}$$

$$Z_\alpha^M(p : q) = \int_{t \in \mathcal{X}} M_\alpha(p(t), q(t)) d\mu(t)$$

$$(p_1 \dots p_k)_\alpha^M := \frac{p_1(x)^{\alpha_1} \times \dots \times p_k(x)^{\alpha_k}}{Z_\alpha(p_1, \dots, p_k)}$$

$$\text{JS}_D^{M_\alpha}(p : q) := (1 - \alpha)D\left(p : (pq)_\alpha^M\right) + \alpha D\left(q : (pq)_\alpha^M\right)$$

$$\text{JS}^{M_\alpha}(p : q) := (1 - \alpha)\text{KL}\left(p : (pq)_\alpha^M\right) + \alpha\text{KL}\left(q : (pq)_\alpha^M\right)$$

The M-JSD is upper bounded by $\log \frac{Z_\alpha^M(p,q)}{1-\alpha}$ when $M \geq A$.

$$\text{JS}_D^{M_\alpha, N_\beta}(p : q) := N_\beta \left(D \left(p : (pq)_\alpha^M \right), D \left(q : (pq)_\alpha^M \right) \right)$$

Closed-form formula for exponential families

$$\begin{aligned}\text{KL} \left(p_\theta : (p_{\theta_1} p_{\theta_2})_\alpha^G \right) &= \text{KL} \left(p_\theta : p_{(\theta_1 \theta_2)_\alpha} \right) \\ &= B_F((\theta_1 \theta_2)_\alpha : \theta).\end{aligned}$$

$$\begin{aligned}\text{JS}_\alpha^G(p_{\theta_1} : p_{\theta_2}) &:= (1 - \alpha)\text{KL}(p_{\theta_1} : (p_{\theta_1} p_{\theta_2})_\alpha^G) + \alpha\text{KL}(p_{\theta_2} : (p_{\theta_1} p_{\theta_2})_\alpha^G), \\ &= (1 - \alpha)B_F((\theta_1 \theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1 \theta_2)_\alpha : \theta_2).\end{aligned}$$

$$\begin{aligned}\text{JS}_{\text{KL}*}^{G_\alpha}(p_{\theta_1} : p_{\theta_2}) &:= (1 - \alpha)\text{KL}((p_{\theta_1} p_{\theta_2})_\alpha^G : p_{\theta_1}) + \alpha\text{KL}((p_{\theta_1} p_{\theta_2})_\alpha^G : p_{\theta_2}), \\ &= (1 - \alpha)B_F(\theta_1 : (\theta_1 \theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1 \theta_2)_\alpha) = \text{JB}_F^\alpha(\theta_1 : \theta_2), \\ &= (1 - \alpha)F(\theta_1) + \alpha F(\theta_2) - F((\theta_1 \theta_2)_\alpha), \\ &= J_F^\alpha(\theta_1 : \theta_2).\end{aligned}$$

Case study of multivariate Gaussians

$$\text{KL}(p_{(\mu_1, \Sigma_1)} : p_{(\mu_2, \Sigma_2)}) = \frac{1}{2} \left\{ \text{tr}(\Sigma_2^{-1} \Sigma_1) + \Delta_\mu^\top \Sigma_2^{-1} \Delta_\mu + \log \frac{|\Sigma_2|}{|\Sigma_1|} - d \right\}$$

$$\begin{aligned} \text{JS}^{G_\alpha}(p_{(\mu_1, \Sigma_1)} : p_{(\mu_2, \Sigma_2)}) &= (1 - \alpha)\text{KL}(p_{(\mu_1, \Sigma_1)} : p_{(\mu_\alpha, \Sigma_\alpha)}) + \alpha\text{KL}(p_{(\mu_2, \Sigma_2)} : p_{(\mu_\alpha, \Sigma_\alpha)}), \\ &= (1 - \alpha)B_F((\theta_1 \theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1 \theta_2)_\alpha : \theta_2), \\ &= \frac{1}{2} \left(\text{tr} \left(\Sigma_\alpha^{-1} ((1 - \alpha)\Sigma_1 + \alpha\Sigma_2) \right) + \log \frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha} |\Sigma_2|^\alpha} + \right. \\ &\quad \left. (1 - \alpha)(\mu_\alpha - \mu_1)^\top \Sigma_\alpha^{-1} (\mu_\alpha - \mu_1) + \alpha(\mu_\alpha - \mu_2)^\top \Sigma_\alpha^{-1} (\mu_\alpha - \mu_2) - d \right) \end{aligned}$$

$$\begin{aligned} \text{JS}_*^{G_\alpha}(p_{(\mu_1, \Sigma_1)} : p_{(\mu_2, \Sigma_2)}) &= (1 - \alpha)\text{KL}(p_{(\mu_\alpha, \Sigma_\alpha)} : p_{(\mu_1, \Sigma_1)}) + \alpha\text{KL}(p_{(\mu_\alpha, \Sigma_\alpha)} : p_{(\mu_2, \Sigma_2)}), \\ &= (1 - \alpha)B_F(\theta_1 : (\theta_1 \theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1 \theta_2)_\alpha), \\ &= J_F(\theta_1 : \theta_2), \\ &= \frac{1}{2} \left(\text{tr} \left(\Sigma_\alpha^{-1} ((1 - \alpha)\Sigma_1 + \alpha\Sigma_2) \right) + \alpha u_2^\top \Sigma_\alpha^{-1} u_2 - u_\alpha^\top \Sigma_\alpha^{-1} u_\alpha + \log \frac{|\Sigma_1|^{1-\alpha} |\Sigma_2|^\alpha}{|\Sigma_\alpha|} \right) \end{aligned}$$

$$\Sigma_\alpha = (\Sigma_1 \Sigma_2)_\alpha^\Sigma = \left((1 - \alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1} \right) \quad \mu_\alpha = (\mu_1 \mu_2)_\alpha^\mu = \Sigma_\alpha \left((1 - \alpha)\Sigma_1^{-1} \mu_1 + \alpha\Sigma_2^{-1} \mu_2 \right)$$

Case study of Cauchy family: Harmonic mean

$$\mathcal{C}_\Gamma := \left\{ p_\gamma(x) = \frac{1}{\gamma} p_{\text{std}} \left(\frac{x}{\gamma} \right) = \frac{\gamma}{\pi(\gamma^2 + x^2)} : \gamma \in \Gamma = (0, \infty) \right\}$$

$$(p_{\gamma_1} p_{\gamma_2})_{\frac{1}{2}}^H(x) = \frac{H_\alpha(p_{\gamma_1}(x) : p_{\gamma_2}(x))}{Z_\alpha^H(\gamma_1, \gamma_2)} = p_{(\gamma_1 \gamma_2)_\alpha}$$

$$Z_\alpha^H(\gamma_1, \gamma_2) := \sqrt{\frac{\gamma_1 \gamma_2}{(\gamma_1 \gamma_2)_\alpha (\gamma_1 \gamma_2)_{1-\alpha}}} = \sqrt{\frac{\gamma_1 \gamma_2}{(\gamma_1 \gamma_2)_\alpha (\gamma_2 \gamma_1)_\alpha}}$$

$$\begin{aligned} \text{JS}^H(p : q) &= \frac{1}{2} \left(\text{KL} \left(p : (pq)_{\frac{1}{2}}^H \right) + \text{KL} \left(q : (pq)_{\frac{1}{2}}^H \right) \right), \\ \text{JS}^H(p_{\gamma_1} : p_{\gamma_2}) &= \frac{1}{2} \left(\text{KL} \left(p_{\gamma_1} : p_{\frac{\gamma_1 + \gamma_2}{2}} \right) + \text{KL} \left(p_{\gamma_2} : p_{\frac{\gamma_1 + \gamma_2}{2}} \right) \right) \\ &= \log \frac{(3\gamma_1 + \gamma_2)(3\gamma_2 + \gamma_1)}{8\sqrt{\gamma_1 \gamma_2}(\gamma_1 + \gamma_2)}. \end{aligned}$$

Kullback-Leibler divergence between Cauchy densities

Cauchy density $p_{l,s}(x) = \frac{dP_{l,s}}{d\mu}(x) = \frac{s}{\pi(s^2 + (x - l)^2)}$

$$\text{KL}(p_{l_1,s_1} : p_{l_2,s_2}) = \log \frac{(s_1 + s_2)^2 + (l_1 - l_2)^2}{4s_1s_2}$$

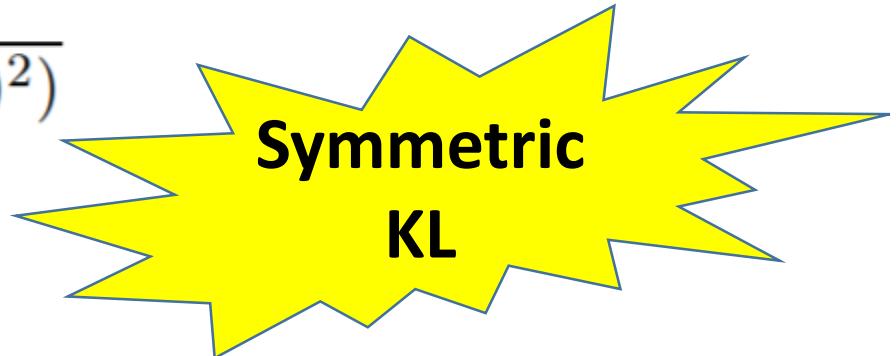
Cross-entropy $h^\times(p_{l_1,s_1} : p_{l_2,s_2}) = \log \frac{\pi((s_1 + s_2)^2 + (l_1 - l_2)^2)}{s_2}$

Differential entropy $h(p_{l,s}) = h^\times(p_{l,s} : p_{l,s}) = \log 4\pi s,$

$$A(a, b, c; d, e, f) = \int_{-\infty}^{\infty} \frac{\log(dx^2 + ex + f)}{ax^2 + bx + c} dx,$$

Relies on this definite integral
with

$$A(a, b, c; d, e, f) = \frac{2\pi \left(\log(2af - be + 2cd + \sqrt{4ac - b^2}\sqrt{4df - e^2}) - \log(2a) \right)}{\sqrt{4ac - b^2}}$$



Kullback-Leibler divergence between location-scale densities

Property: The f-divergence between location-scale densities reduces to the f-divergence between a standard density and another location-scale density

$$I_f(p_{l_1,s_1} : q_{l_2,s_2}) = I_f \left(p : q_{\frac{l_2-l_1}{s_1}, \frac{s_2}{s_1}} \right) = I_f \left(p_{\frac{l_1-l_2}{s_2}, \frac{s_1}{s_2}} : q \right)$$

Proof by change of variable

$$y = \frac{x-l_1}{s_1}$$

$$dx = s_1 dy$$

$$x = s_1 y + l_1$$

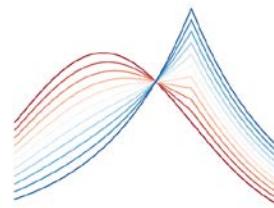
$$\frac{x-l_2}{s_2} = \frac{s_1 y + l_1 - l_2}{s_2} = \frac{y - \frac{l_2-l_1}{s_1}}{\frac{s_2}{s_1}}$$

$$\begin{aligned} I_f(p_{l_1,s_1} : q_{l_2,s_2}) &:= \int_{\mathcal{X}} p_{l_1,s_1}(x) f\left(\frac{q_{l_2,s_2}(x)}{p_{l_1,s_1}(x)}\right) dx, && \text{Location-scale group} \\ &= \int_{\mathbb{Y}} \frac{1}{s_1} p(y) f\left(\frac{\frac{1}{s_2} q\left(\frac{y - \frac{l_2-l_1}{s_1}}{\frac{s_2}{s_1}}\right)}{\frac{1}{s_1} p(y)}\right) s_1 dy, && \mathbb{H} = \{(l, s) : l \in \mathbb{R} \times \mathbb{R}_{++}\} \\ &= \int p(y) f\left(\frac{q_{l_2-l_1, \frac{s_2}{s_1}}(y)}{p(y)}\right) dy, \\ &= I_f \left(p : q_{\frac{l_2-l_1}{s_1}, \frac{s_2}{s_1}} \right). \end{aligned}$$

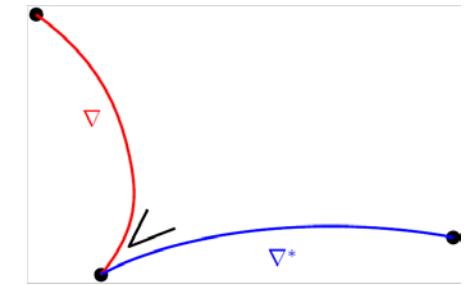
Structures of location-scale families

$$\min_{(l_1, s_1) \in \mathbb{H}} I_f(p_{l_1, s_1} : q_{l_2, s_2}) = \min_{(l, s) \in \mathbb{H}} I_f(p : q_{l, s}) := I_f(p : Q).$$

Distances and information geometry of finite statistical mixtures



of

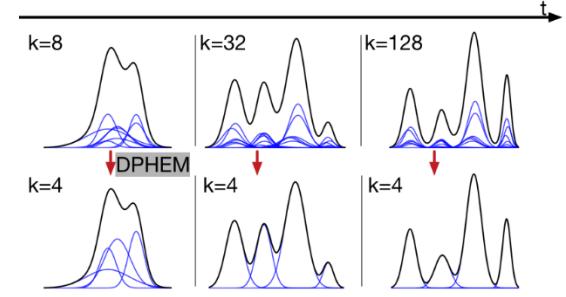
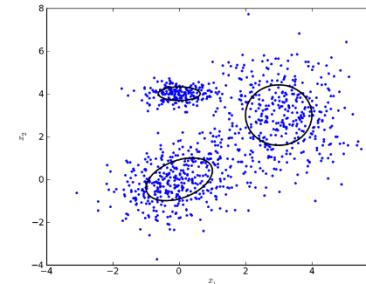


Frank Nielsen



Sony CSL

Finite statistical mixtures

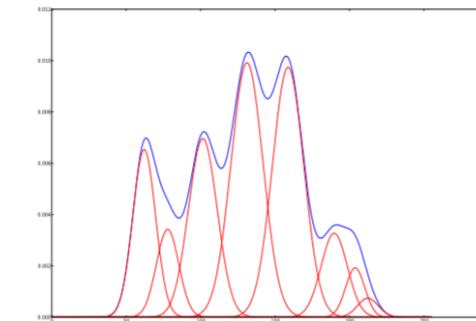
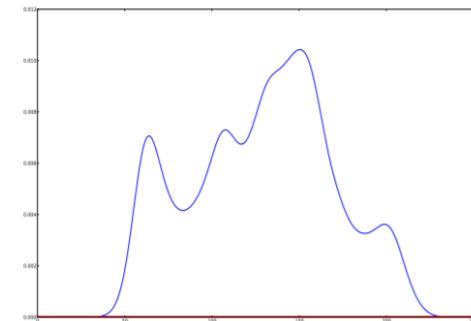
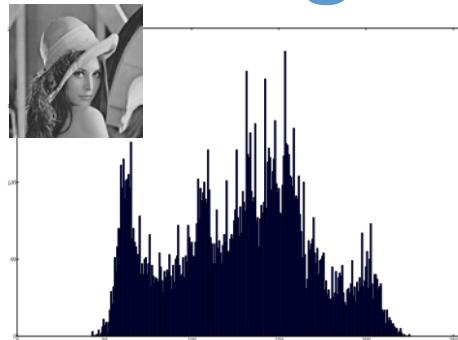


- **Semi-parametric** models, universal estimators of smooth densities
- Gaussian mixture models (GMMs), Exponential family mixture models (EFMMs), etc.

$$f(x) = \sum_{i=1}^n \omega_i g(x; \mu_i, \sigma_i^2) \quad \text{with} \quad g(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- But **Non-identifiable**/non-regular !!! (not 1-to-1 parameter/density)
- Usually learn GMMs by Expectation-Maximization (EM, local optimum)
- But also can learn mixtures by **simplifying** a Kernel Density estimator

Learning a mixture by simplifying a kernel density estimator



Original histogram
raw KDE (14400 components)
simplified mixture (8 components)

$$f(x) = \sum_{i=1}^n \omega_i g(x; \mu_i, \sigma_i^2) \quad g(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Usual **centroids** based on **Kullback-Leibler** sided/symmetrized

$$\arg \min_c \sum_i \omega_i KLD(c, x_i)$$

$$\begin{aligned} KLD(f_p, f_q) &= \frac{1}{2} \log \left(\frac{\det \Sigma_p}{\det \Sigma_q} \right) \\ &\quad + \frac{1}{2} \text{tr}(\Sigma_q^{-1} \Sigma_p) \\ &\quad + \frac{1}{2} (\mu_q - \mu_p)^T \Sigma_q^{-1} (\mu_q - \mu_p) - \frac{d}{2} \end{aligned}$$

$$\arg \min_c \sum_i \omega_i SKL(x_i, c)$$

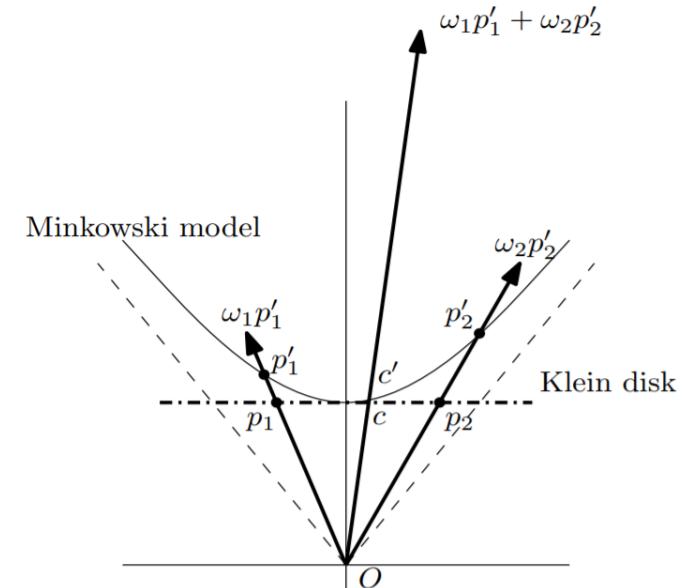
$$\arg \min_c \sum_i \omega_i FRD(f_p, f_q) =$$

$$\sqrt{2} \ln \frac{|(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)| + |(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)|}{|(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)| - |(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)|}$$

Problem:

No closed-form FR/SKL centroids!!!

Galperin's model centroid (HG)



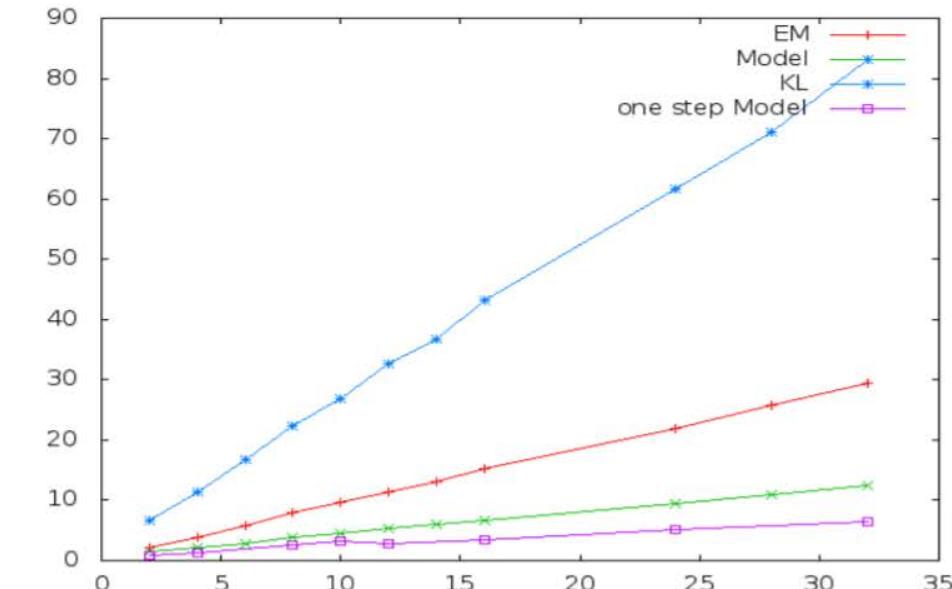
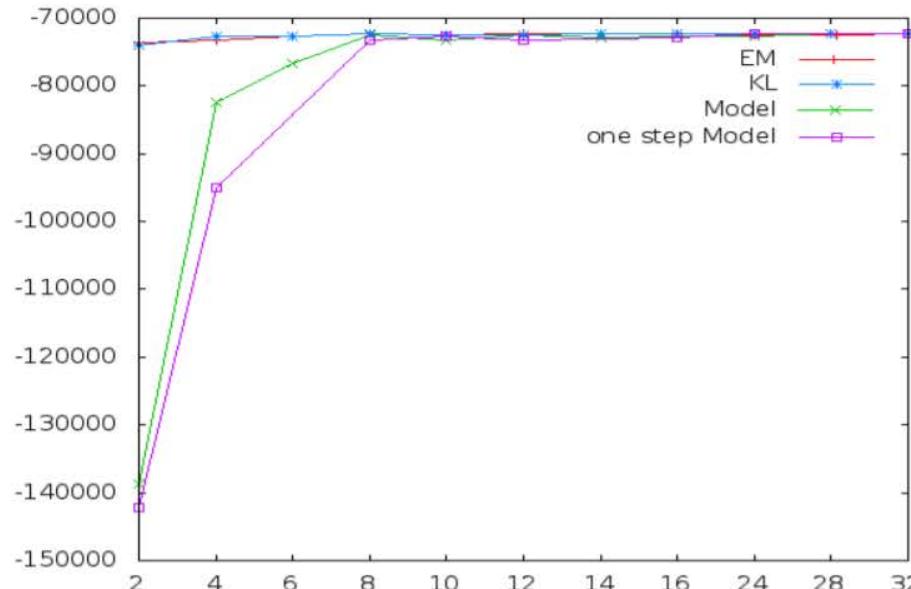
Simple model centroid algorithm:

Embed Klein points to points of the Minkowski hyperboloid

Centroid = center of mass c , scaled back to c' of the hyperboloid

Map back c' to Klein disk

Experiments



Log-likelihood of the simplified models and computation time

Dataset: intensity histogram of Lena image

KL with right-sided centroids

Full k-means or only one iteration

While achieving same log-likelihood, model centroid is the fastest method, significantly faster than EM.

Distances and geometry of statistical mixtures

- Many common statistical distances are **not in closed-form** when dealing with statistical mixtures (eg., KLD between GMMs not even analytic!).
- Need **approximation algorithms** to calculate mixture distances
- Or design **novel principled statistical distances** that admit closed forms or approximate probabilistically/deterministically statistical distances (e.g, Cauchy-Schwarz divergence, Jensen-Renyi divergence , etc.)
- Geometry of **mixtures family** in information geometry is **dually flat**: intractable Bregman manifold and **tractable Monte Carlo Bregman manifold**

Batch learning of mixtures and lightspeed distance calculations

$$m(x) = \sum_{i=1}^{k_1} \omega_i p_F(x; \eta_i) \quad m'(x) = \sum_{i=1}^{k_2} \omega'_i p_F(x; \eta'_i)$$

$$\text{KL}_{\text{MC}}(m \| m') = \frac{1}{n} \sum_{i=1}^n \log \frac{m(x_i)}{m'(x_i)}$$

Kullback-Leibler divergence

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

$$= H(p, q) - H(p)$$

Monte-Carlo stochastic estimation (iid sampling from m)

- Hungarian best bipartite matching of components (Goldberger)

$$\text{KL}_{\text{Gold}}(m \| m') = \arg \min_{\sigma} \text{KL}(\omega \| \sigma(\omega'))$$

$$+ \sum \omega_i \text{KL}(p_F(\cdot \| \eta_i) \| p_F(\cdot \| \eta'_{\sigma(i)}))$$

$$\text{KL}_{\text{var}}(m \| m') = \sum_i \omega_i \log \frac{\sum_j \omega_j e^{-\text{KL}(p_F(\cdot \| \eta_i) \| p_F(\cdot \| \eta_j))}}{\sum_j \omega'_j e^{-\text{KL}(p_F(\cdot \| \eta_i) \| p_F(\cdot \| \eta'_j))}}$$

Definition A co-mixture of exponential families (a *comix*) with K components is a set of S statistical mixture models of the form:

$$\left\{ \begin{array}{l} m_1(x; \omega_i^{(1)} \dots \omega_K^{(1)}) = \sum_{i=1}^K \omega_i^{(1)} p_F(x; \eta_i) \\ m_2(x; \omega_i^{(2)} \dots \omega_K^{(2)}) = \sum_{i=1}^K \omega_i^{(2)} p_F(x; \eta_i) \\ \dots \\ m_S(x; \omega_i^{(S)} \dots \omega_K^{(S)}) = \sum_{i=1}^K \omega_i^{(S)} p_F(x; \eta_i) \end{array} \right.$$

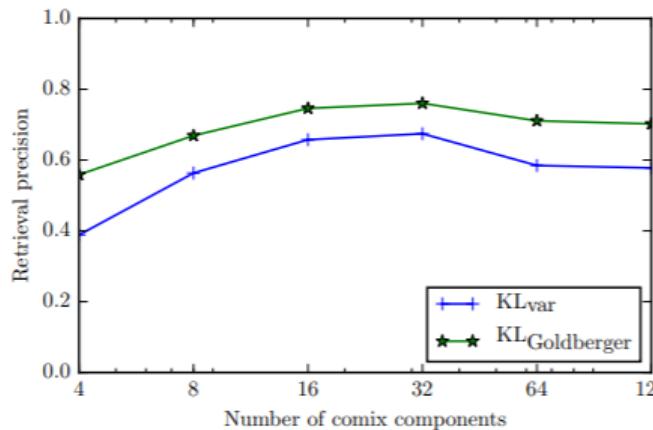
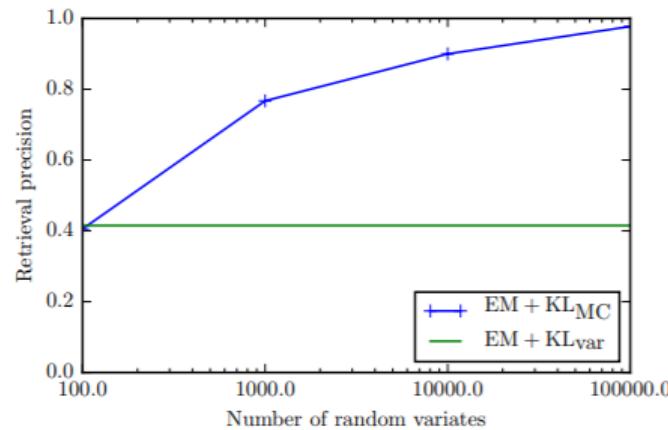
Extend Expectation-Maximization algorithms for **batch learning of co-mixtures**
(co-EM, adapt Bregman soft clustering)

Precompute the matrix $D_{ij} = \text{KL}(p_F(\cdot \| \eta_i) \| p_F(\cdot \| \eta_j))$.

Comix: Joint estimation and lightspeed comparison of mixture models. ICASSP 2016

Bag-of-components: an online algorithm for batch learning of mixture models, GSI 2015

Experiments on comixs



mean average precision (mAP) over all the possible queries (by successively taking each mixture as the query and looking at the retrieved mixtures in a short list of size 10)

Fig. *Left:* mAP of KLMC between EM mixtures wrt the sample size and result from variational KL. *Right:* mAP wrt the number of components of variational Kullback-Leibler and Goldberger between co-EM mixtures.

| k | co-EM | Speed-up between co-EM and EM8 | KL _{var} on comix | Speed-up between KL _{var} on comix and KL _{var} on EM8 | Speed-up between KL _{var} on comix and KL _{MC100} on EM8 | Goldberger on comix |
|-----|-------|--------------------------------|----------------------------|--|--|---------------------|
| 4 | 51s | ×1.5 | 0.00020s | ×180 | × 20 | 0.00015s |
| 8 | 99s | ×0.77 | 0.00044s | ×84 | × 5.8 | 0.00030s |
| 16 | 48s | ×1.6 | 0.0012s | ×28 | × 1.6 | 0.00059s |
| 32 | 150s | ×0.49 | 0.0040s | ×9.1 | × 0.41 | 0.0012s |
| 64 | 450s | ×0.17 | 0.014s | ×2.5 | × 0.10 | 0.0024s |
| 128 | 600s | ×0.12 | 0.046s | ×0.80 | ×0.026 | 0.0049s |

Table Absolute times for computation on comix and speed-up when compared to the times of the equivalent computation on individual mixtures. Times for co-EM are compared with the total time for all the individual EM.

Chain Rule Optimal Transport (CROT) distance

$$m_1(x) = \sum_{i=1}^{k_1} \alpha_i p_i(x)$$

$$m_1 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j}$$

$$m_2(x) = \sum_{i=1}^{k_2} \beta_i q_i(x)$$

$$m_2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} q_{i,j}$$

$$p_{i,j} = p_i$$

$$q_{i,j} = q_j$$

Solve the **Linear Program**: $H_\delta(p, q) = \sum \sum w_{ij} \delta(p_i, q_j)$
 with the following constraints:

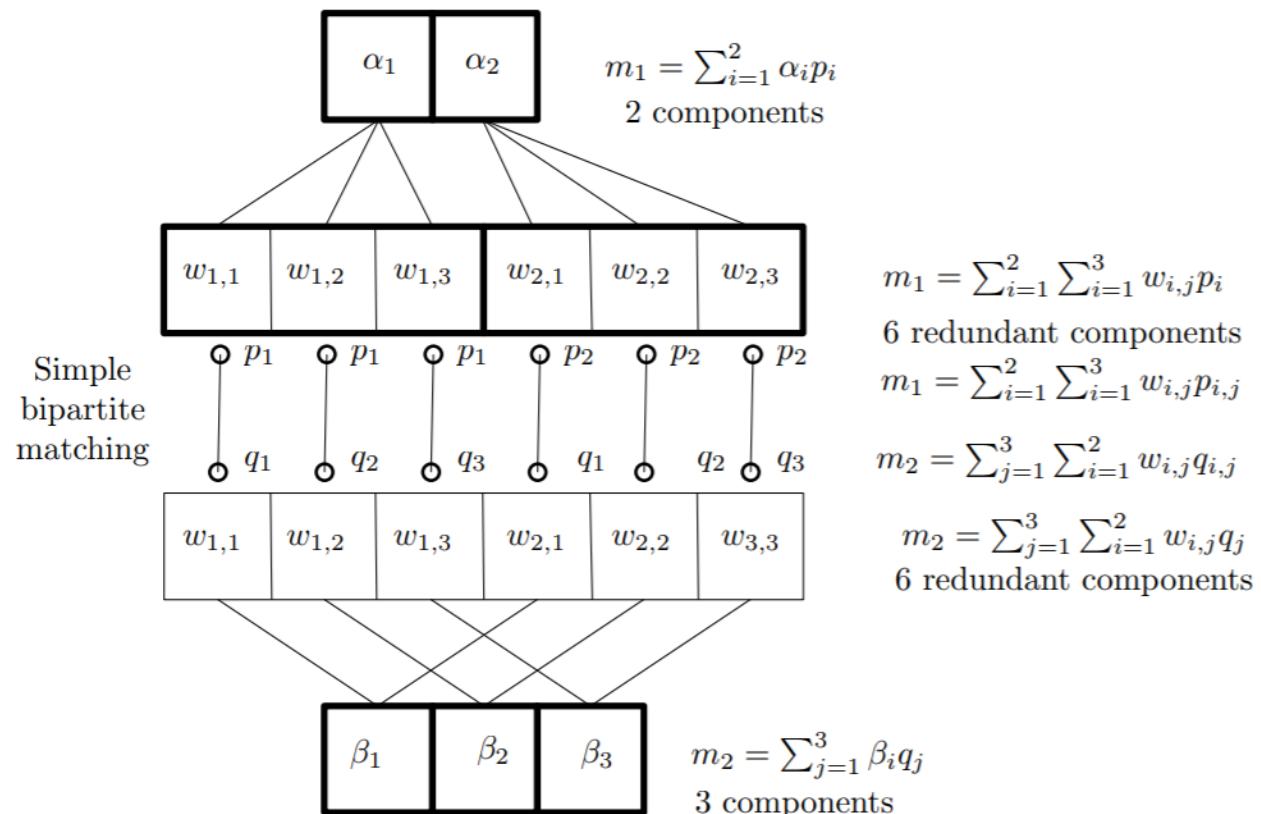
$$w_{ij} \geq 0, \quad \forall i \in [k_1], j \in [k_2]$$

$$\sum_{l=1}^{k_2} w_{il} = \alpha_i, \quad \forall i \in [k_1]$$

$$\sum_{l=1}^{k_1} w_{lj} = \beta_j, \quad \forall j \in [k_2].$$

Equivalent **optimal transport** problem:

$$H_\delta(m_1 : m_2) = \min_{W \in U(\alpha, \beta)} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{ij} \delta(p_i, q_j).$$

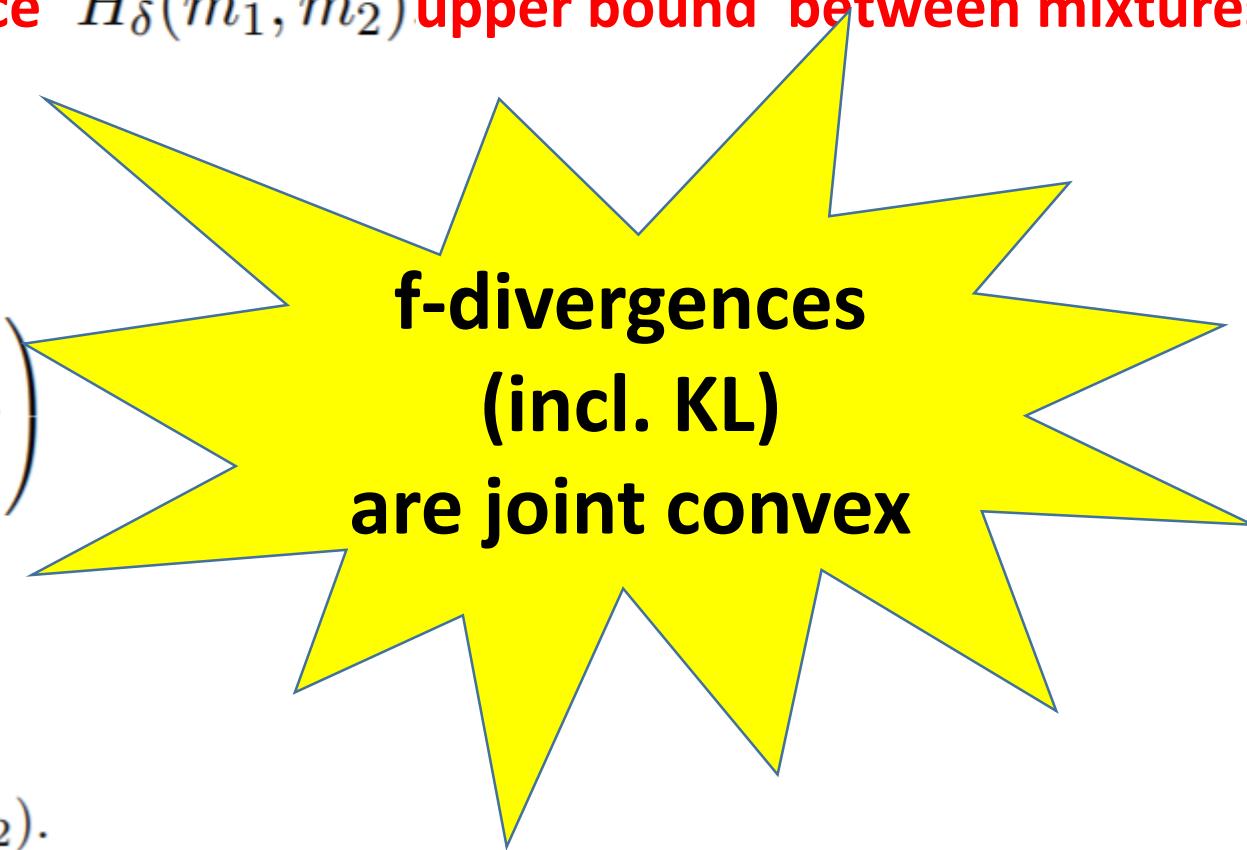


Chain Rule Optimal Transport (CROT) distance

For any joint convex distance $\delta(m_1 : m_2)$,

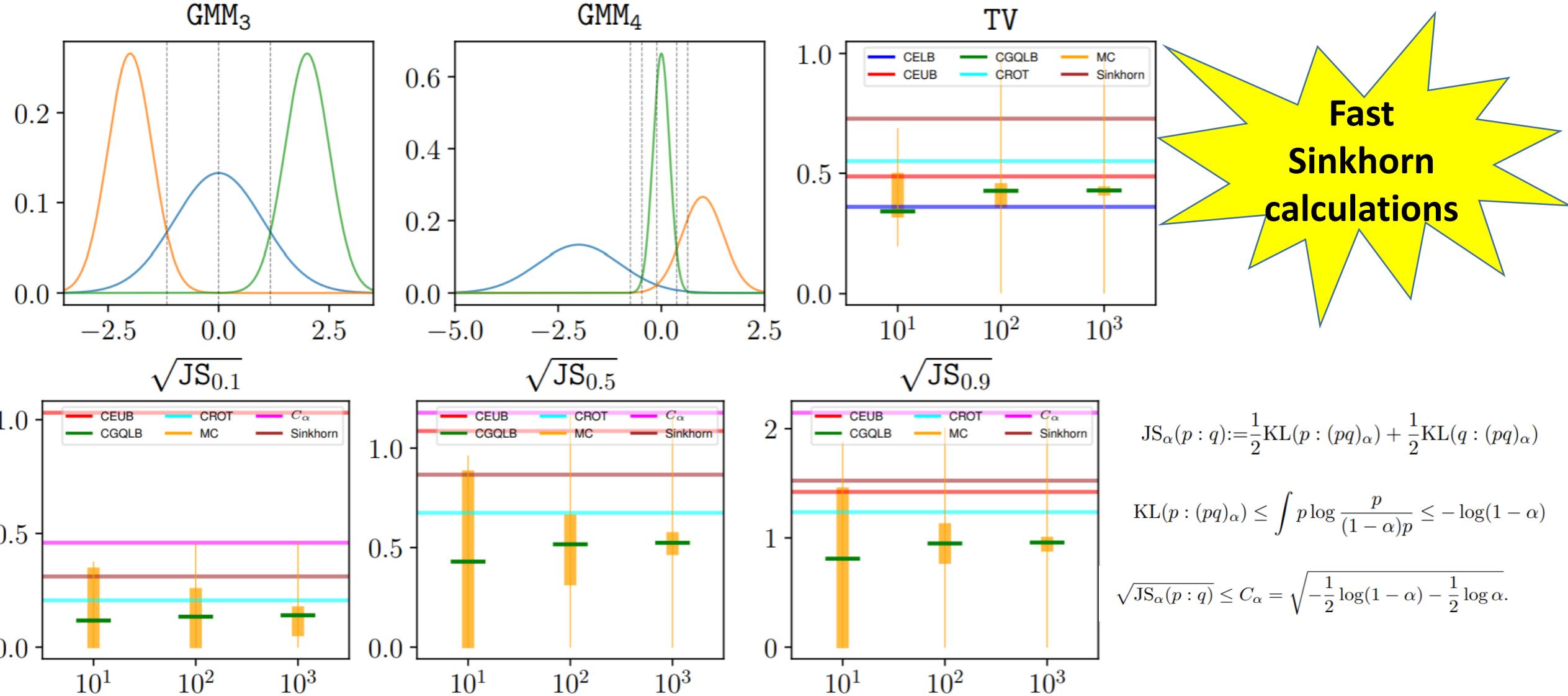
the CROT distance $H_\delta(m_1, m_2)$ upper bound between mixtures

$$\begin{aligned}\delta(m_1 : m_2) &= \delta\left(\sum_{i=1}^{k_1} \alpha_i p_i, \sum_{j=1}^{k_2} \beta_j q_j\right) \\ &= \delta\left(\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j} : \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} q_{i,j}\right) \\ &\leq \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \delta(p_{i,j} : q_{i,j}), \\ &\leq \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \delta(p_i : q_j) =: H_\delta(m_1, m_2).\end{aligned}$$



But also the p-powered Wasserstein distances,
Etc.

Chain Rule Optimal Transport (CROT) distance



$$\text{JS}_\alpha(p : q) := \frac{1}{2} \text{KL}(p : (pq)_\alpha) + \frac{1}{2} \text{KL}(q : (pq)_\alpha)$$

$$\text{KL}(p : (pq)_\alpha) \leq \int p \log \frac{p}{(1-\alpha)p} \leq -\log(1-\alpha)$$

$$\sqrt{\text{JS}_\alpha(p : q)} \leq C_\alpha = \sqrt{-\frac{1}{2} \log(1-\alpha) - \frac{1}{2} \log \alpha}$$

Statistical mixtures versus mixture families

- In statistics, finite statistical mixtures are **irregular models**
(non-identifiable)

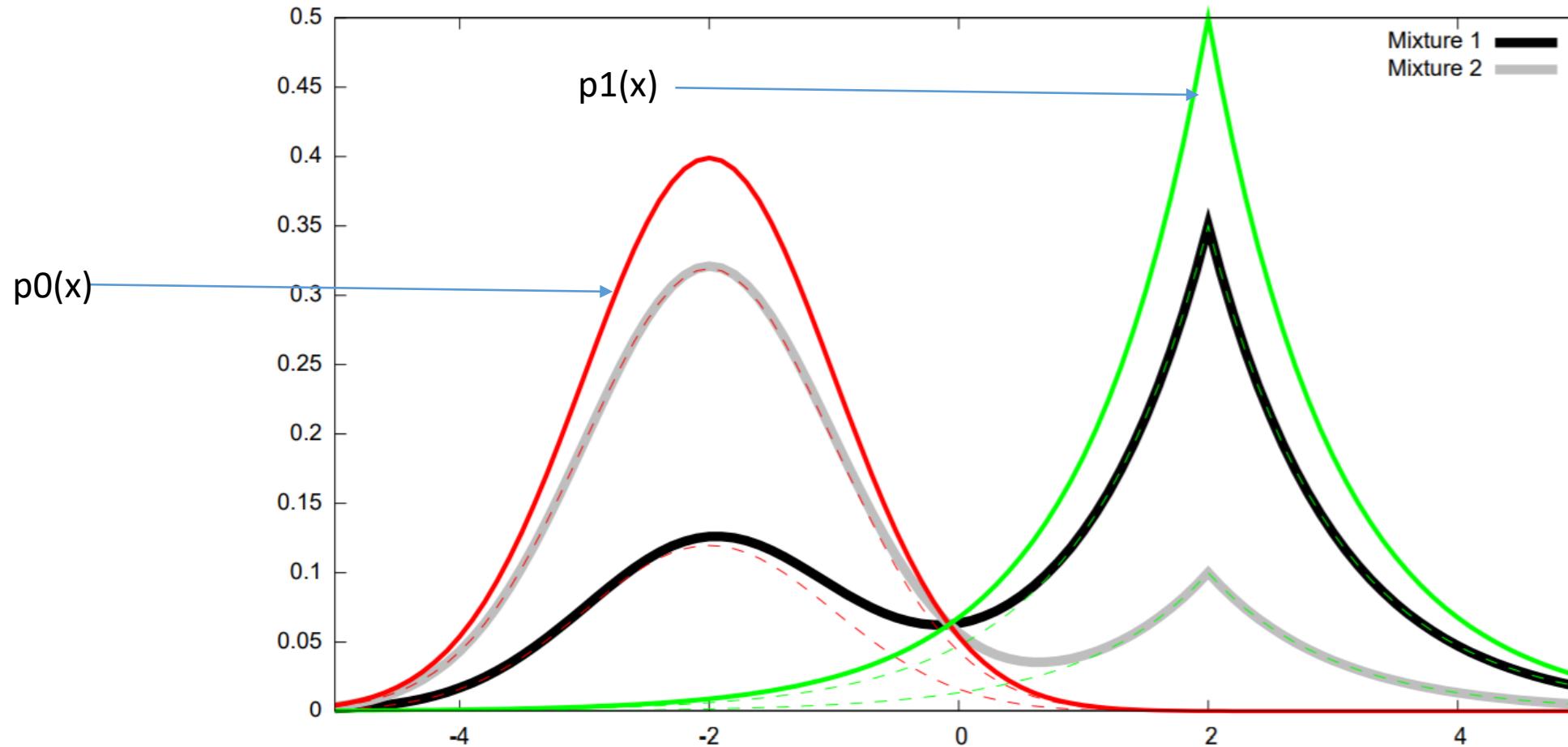
$$m(x; w) := \sum_{i=0}^{k-1} w_i p_i(x),$$

- Information geometry primarily considers **regular models**
- In information geometry, **mixture families are regular parametric models**

$$\mathcal{M} := \{m(x; w) , w \in \Delta_{k-1}^\circ\} \quad f_i(x) = p_i(x) - p_0(x) \quad c(x) = p_0(x)$$
$$\mathcal{M} = \left\{ m(x; \eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x), \eta \in \mathbb{R}_{++}^{k-1} \right\} \quad \mathcal{M} = \left\{ m(x; \eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x), \eta \in H^\circ \right\}$$

- Statistical mixtures with prescribed distinct component distributions form mixture families

A mixture family of order 1 (2 components)



$$\mathcal{M} = \left\{ m(x; \eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x), \eta \in \mathbb{R}_{++}^{k-1} \right\}$$

A mixture family of order 2 (3 components)

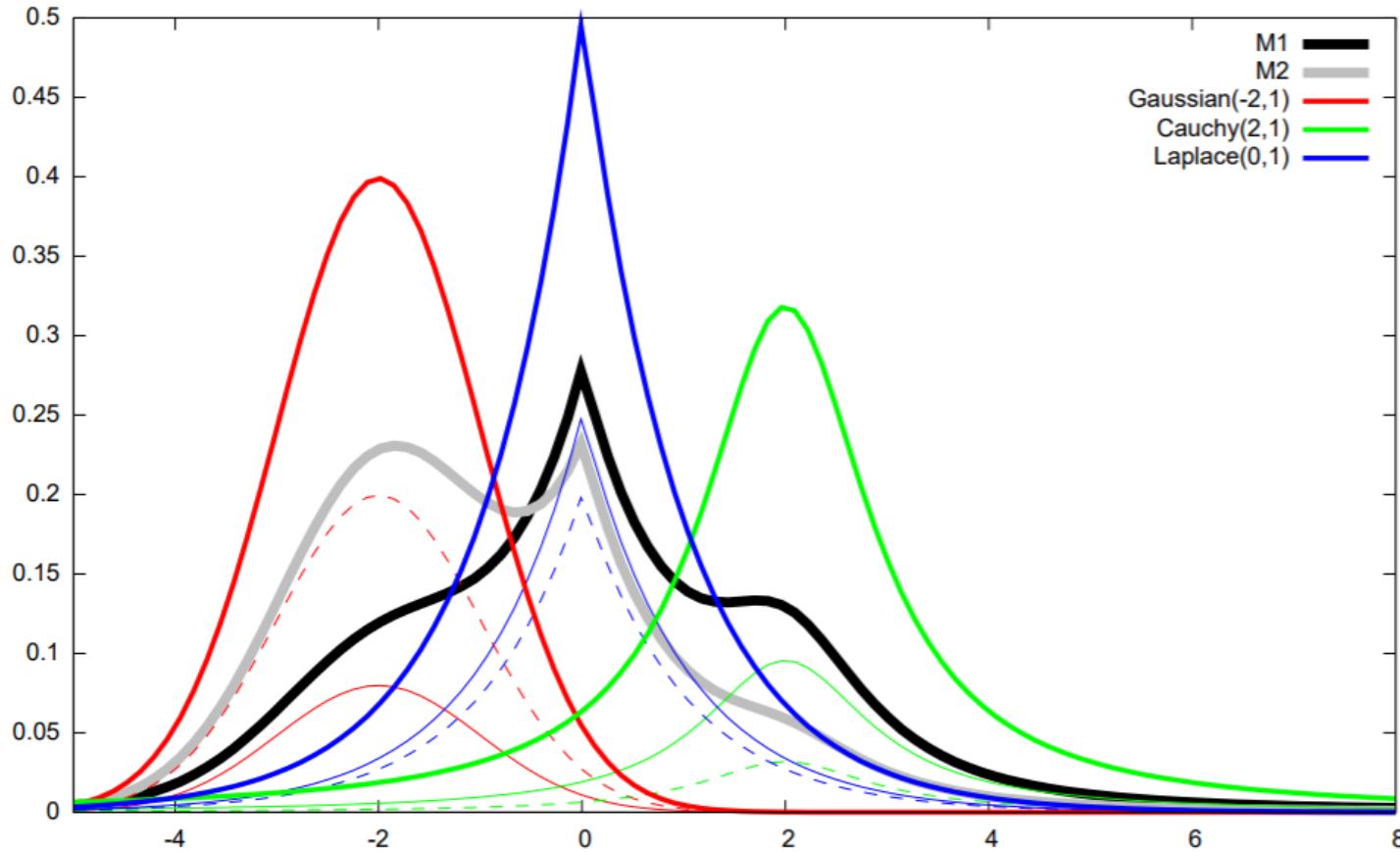


Figure : Example of a mixture family of order $D = 2$ ($k = 3$): $p_0(x) \sim \text{Gaussian}(-2, 1)$ (red), $p_1(x) \sim \text{Laplace}(0, 1)$ (blue) and $p_2(x) \sim \text{Cauchy}(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = (0.3, 0.5)$ and $m_2(x) = m(x; \eta)$ (gray) with $\eta = (0.1, 0.4)$.

A mixture family is a Bregman (Hessian) manifold

- Two global coordinate systems related by Legendre-Fenchel transformation
- Two flat connections that are coupled to the metric tensor (Hessian of a potential function)
- Primal/dual geodesics are straight lines in the primal/dual coordinate system

| Manifold (\mathcal{M}, F) | Primal structure | Dual structure |
|--|---|---|
| Affine coordinate system Conversion $\theta \leftrightarrow \eta$ | $\theta(\cdot)$ $\theta(\eta) = \nabla F^*(\eta)$ | $\eta(\cdot)$ $\eta(\theta) = \nabla F(\theta)$ |
| Potential function | $F(\theta) = \langle \theta, \nabla F(\theta) \rangle - F^*(\nabla F(\theta))$ | $F^*(\eta) = \langle \eta, \nabla F^*(\eta) \rangle - F(\nabla F^*(\eta))$ |
| Metric tensor g | $G(\theta) = \nabla^2 F(\theta)$ | $G^*(\eta) = \nabla^2 F^*(\eta)$ |
| Geodesic ($\lambda \in [0, 1]$) | $g_{ij} = \partial_i \partial_j F(\theta)$ $\gamma(P, Q) = \{(PQ)_\lambda = (1 - \lambda)\theta(P) + \lambda\theta(Q)\}_\lambda$ | $g^{ij} = \partial^i \partial^j F^*(\eta)$ $\gamma^*(P, Q) = \{(PQ)^*_\lambda = (1 - \lambda)\eta(P) + \lambda\eta(Q)\}_\lambda$ |

Two prominent examples of Bregman manifolds

| | Exponential Family | Mixture Family |
|--|--|--|
| Density | $p(x; \theta) = \exp(\langle \theta, x \rangle - F(\theta))$ | $m(x; \eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x)$ $f_i(x) = p_i(x) - p_0(x)$ |
| Family/Manifold Convex function ($\equiv ax + b$) | $\mathcal{M} = \{p(x; \theta) : \theta \in \Theta^\circ\}$ F : cumulant | $\mathcal{M} = \{m(x; \eta) : \eta \in H^\circ\}$ F^* : negative entropy |
| Dual coordinates | moment $\eta = E[t(x)]$ | $\theta^i = h^\times(p_0 : m) - h^\times(p_i : m)$ |
| Fisher Information $g = (g_{ij})_{ij}$ | $g_{ij}(\theta) = \partial_i \partial_j F(\theta)$ $g = \text{Var}[t(X)]$ | $g_{ij}(\eta) = \int_{\mathcal{X}} \frac{f_i(x) f_j(x)}{m(x; \eta)} d\mu(x)$ |
| Christoffel symbol | $\Gamma_{ij,k} = \frac{1}{2} \partial_i \partial_j \partial_k F(\theta)$ | $g_{ij}(\eta) = -\partial_i \partial_j h(\eta)$ $\Gamma_{ij,k} = -\frac{1}{2} \int_{\mathcal{X}} \frac{f_i(x) f_j(x) f_k(x)}{m^2(x; \eta)} d\mu(x)$ |
| Entropy | $-F^*(\eta)$ | $-F^*(\eta)$ |
| Kullback-Leibler divergence | $B_F(\theta_2 : \theta_1)$ $= B_{F^*}(\eta_1 : \eta_2)$ | $B_{F^*}(\eta_1 : \eta_2)$ $= B_F(\theta_2 : \theta_1)$ |

A mixture family is a dually flat manifold

- The canonical divergence of any dually flat manifold is a **Bregman divergence**

$$\text{KL}(m(x; \eta) : m(x; \eta')) = B_G(\eta : \eta')$$

The KL between two mixtures with prescribed components amounts to a Bregman divergence

- Strictly convex and differential convex generator:

$$G(\eta) = -h(m(x; \eta)) = \int_{x \in \mathcal{X}} m(x; \eta) \log m(x; \eta) d\mu(x)$$

- However, G not in closed-form, event **not analytic!**
- A Bregman divergence is always finite, and so is the KL between two members of the same mixture family.

Tractability of Bregman manifolds

| Algorithm | $F(\theta)$ | $\eta(\theta) = \nabla F(\theta)$ | $\theta(\eta) = \nabla F^*(\eta)$ | $F^*(\eta)$ |
|--|-------------|-----------------------------------|-----------------------------------|-------------|
| Right-sided Bregman clustering | ✓ | ✓ | ✗ | ✗ |
| Left-sided Bregman clustering | ✗ | ✗ | ✓ | ✓ |
| Symmetrized Bregman centroid | ✓ | ✓ | ✓ | ✓ |
| Mixed Bregman clustering | ✓ | ✓ | ✓ | ✓ |
| Maximum Likelihood Estimator for EFs | ✗ | ✗ | ✓ | ✗ |
| Bregman soft clustering (\equiv EM) | ✗ | ✓ | ✓ | ✓ |

| Type | F | ∇F^* | Example |
|--------|-------------------|-----------------|------------------------------------|
| Type 1 | closed-form | closed-form | Gaussian (exponential) family |
| Type 2 | closed-form | not closed-form | Beta (exponential) family |
| Type 3 | comp. intractable | not closed-form | Ising family [49] |
| Type 4 | not closed-form | not closed-form | Polynomial exponential family [39] |
| Type 5 | not analytic | not analytic | mixture family |

Random Bregman manifolds

- If any time we want to compute integral-based generators or Bregman divergences, we used stochastic Monte-Carlo estimators, we get **inconsistencies** and **faulty algorithms**

$$\widehat{\text{KL}}_m(p : q) := \frac{1}{m} \sum_{i=1}^m \log \frac{p(x_i)}{q(x_i)}$$

- Solution: use the **same variates** for all integral-based evaluations
- It turns out that this scheme is similar to defining a random Bregman generator that is with high probability a **proper Bregman generator**. Geometric algorithms run inside that randomized manifold are **consistent** by construction

Random 1D mixture manifolds

Monte Carlo Mixture Family Generator 1D:

$$\begin{aligned}\tilde{G}_{\mathcal{S}}(\eta) &= \frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} m(x_i; \eta) \log m(x_i; \eta), \\ \tilde{G}'_{\mathcal{S}}(\eta) &= \theta = \frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} (p_1(x_i) - p_0(x_i))(1 + \log m(x_i; \eta)), \\ \tilde{G}''_{\mathcal{S}}(\eta) &= \frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} \frac{(p_1(x_i) - p_0(x_i))^2}{m(x_i; \eta)}.\end{aligned}$$

Theorem: With high-probability, $\tilde{G}_{\mathcal{S}}(\eta)$ a Bregman generator

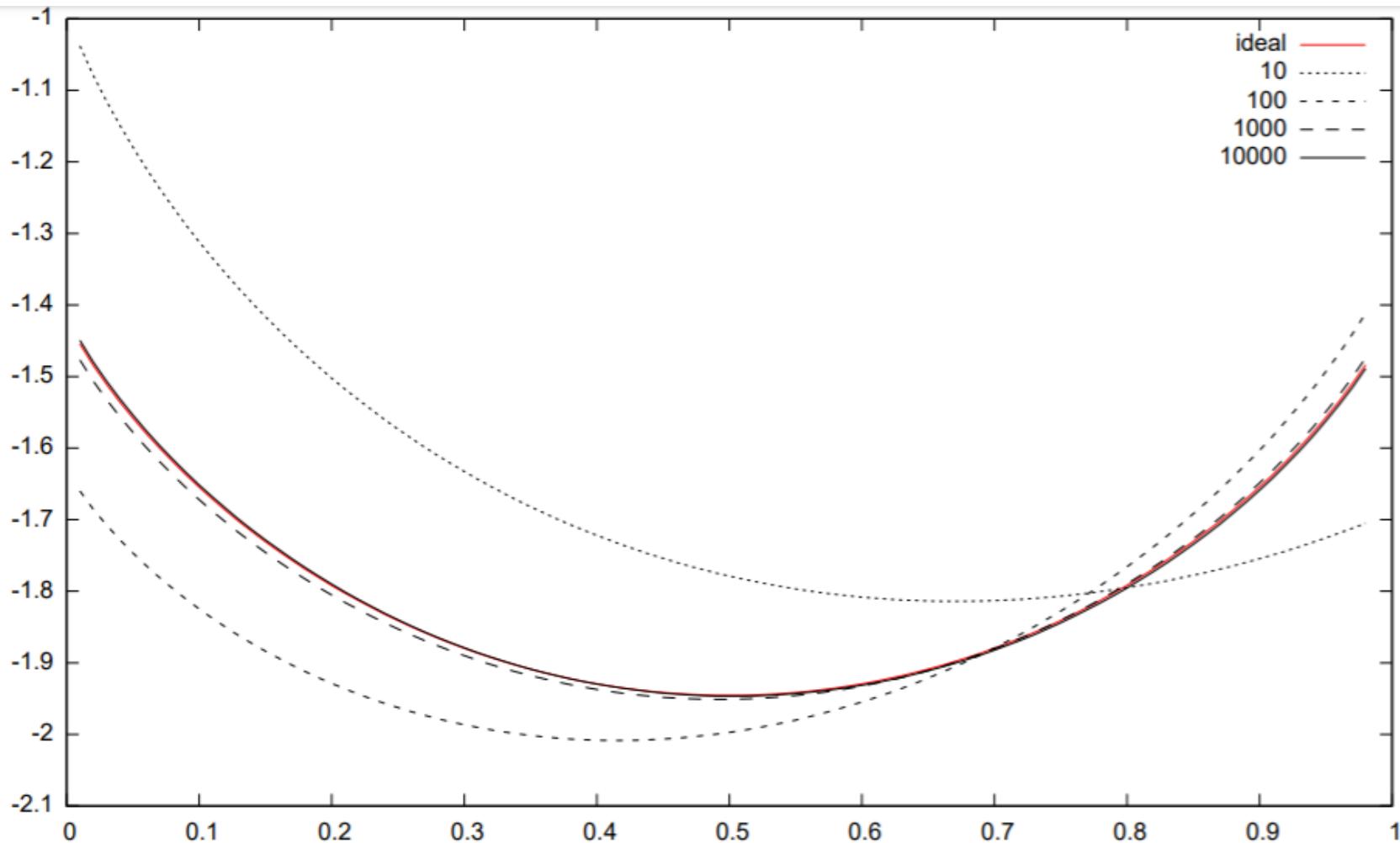


Figure 2: A series $G_S(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x; \eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x; \eta) = (1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

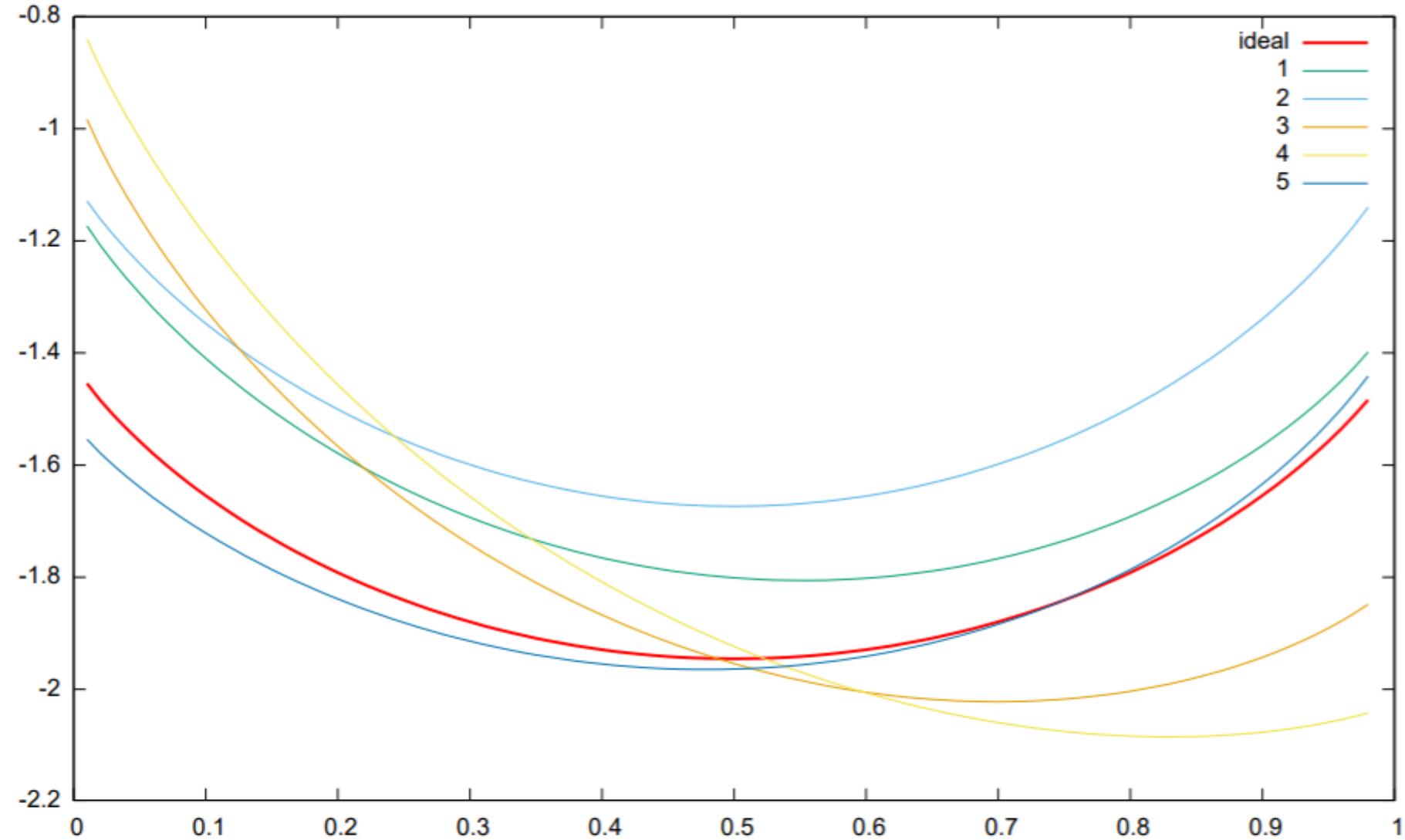


Figure : The Monte Carlo Mixture Family Generator \hat{G}_{10} (MCMFG) considered as a random variable: Here, we show five realizations (i.e., $\mathcal{S}_1, \dots, \mathcal{S}_5$) of the randomized generator for $m = 5$. The ideal generator is plotted in thick red.

Application to clustering Gaussian mixtures (with prescribed Gaussian components)

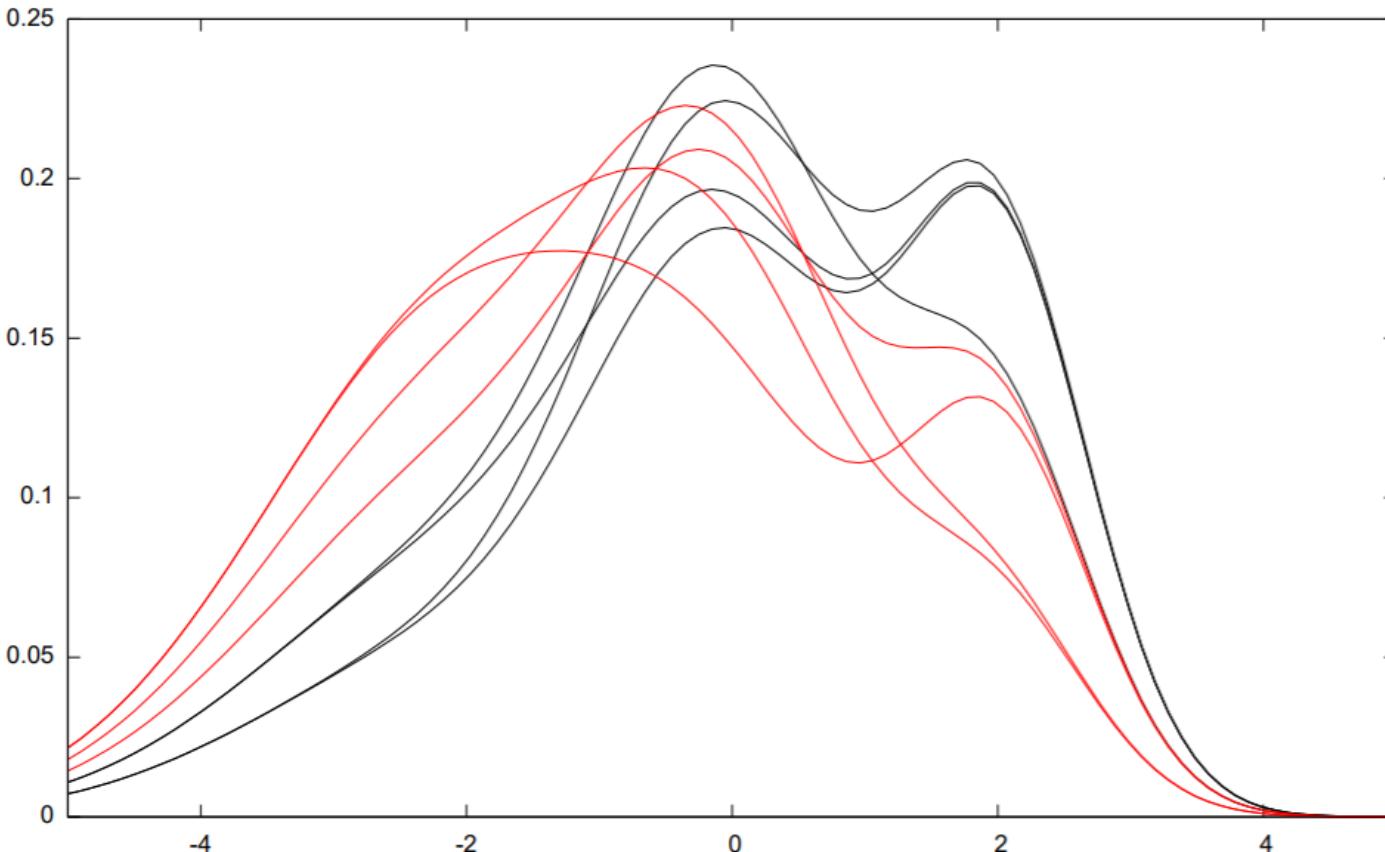


Figure 6: Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point on the mixture family manifold. The Kullback-Leibler divergence is equivalent to an integral-based Bregman divergence that is computationally untractable: The Bregman generator is stochastically approximated by Monte Carlo sampling.

Random dD mixture manifolds

$$\tilde{G}_S(\eta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} m(x_i; \eta) \log m(x_i; \eta).$$

$$\partial^i \partial^j \tilde{G}_S(\eta) = \frac{1}{m} \sum_{l=1}^m \frac{1}{q(x_l)} \frac{(p_i(x_l) - p_0(x_l))(p_j(x_l) - p_0(x_l))}{m(x_l; \eta)}.$$

Theorem (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo multivariate function $\tilde{G}_S(\eta)$ is always convex and twice continuously differentiable, and strictly convex almost surely.*

Random Exponential Family Manifolds

$$\mathcal{E} := \{p(x; \theta) = \exp(t(x)\theta - F(\theta) + k(x)) : \theta \in \Theta\}$$

$$F(\theta) = \log \left(\int \exp(t(x)\theta + k(x)) d\mu(x) \right)$$

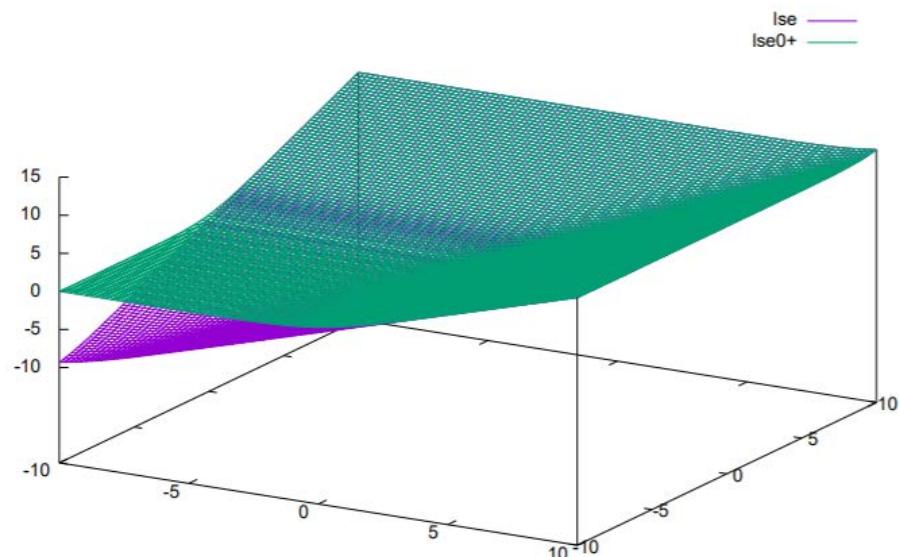
$$F(\theta) \simeq \tilde{F}_{\mathcal{S}}^\dagger(\theta) := \log \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} \exp(t(x_i)\theta + k(x_i)) \right)$$

$$\tilde{F}_{\mathcal{S}}^\dagger(\theta) \equiv \tilde{F}_{\mathcal{S}}(\theta),$$

$$\tilde{F}_{\mathcal{S}}(\theta) = \log \left(1 + \sum_{i=2}^m \exp((t(x_i) - t(x_1))\theta + k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1)) \right)$$

$$= \log \left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i) \right),$$

$$:= \text{lse}_0^+(a_2\theta + b_2, \dots, a_m\theta + b_m),$$



Polynomial Exponential Families

$$p(x; \theta) = \exp(\langle \theta, t(x) \rangle - F(\theta))$$

- Estimate a PEF with **score matching/summed area table**
- Use **projective gamma-divergence** (Monte-Carlo)

$$D_\gamma(p, q) = \frac{1}{\gamma(1 + \gamma)} \log I_\gamma(p, p) - \frac{1}{\gamma} \log I_\gamma(p, q) + \frac{1}{1 + \gamma} \log I_\gamma(q, q),$$

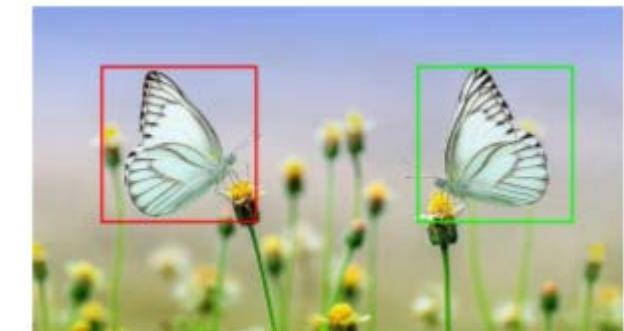
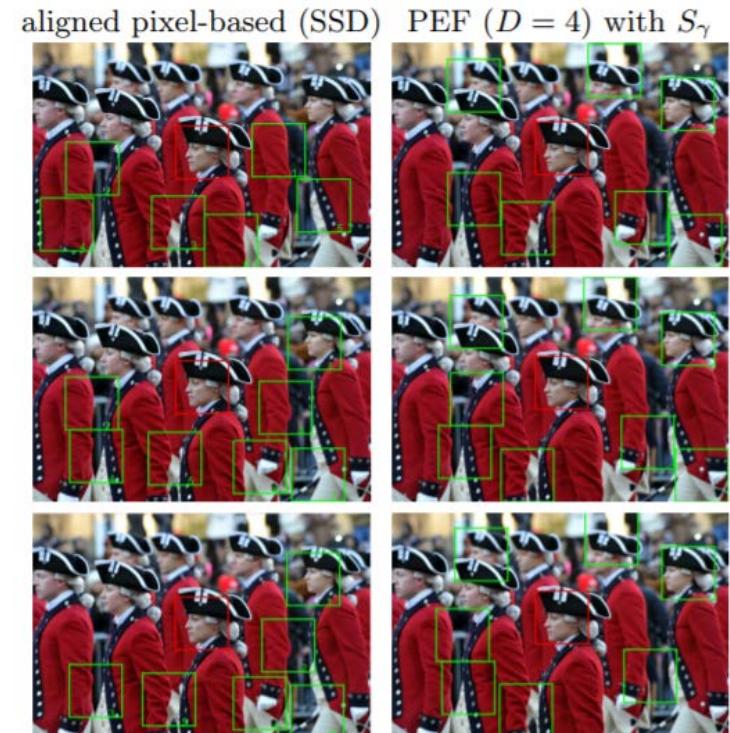
where

$$I_\gamma(p, q) = \int_{x \in \mathcal{X}} p(x)q(x)^\gamma dx.$$

When $\gamma \rightarrow 0$, $D_\gamma(p, q) \rightarrow \text{KL}(p, q)$.

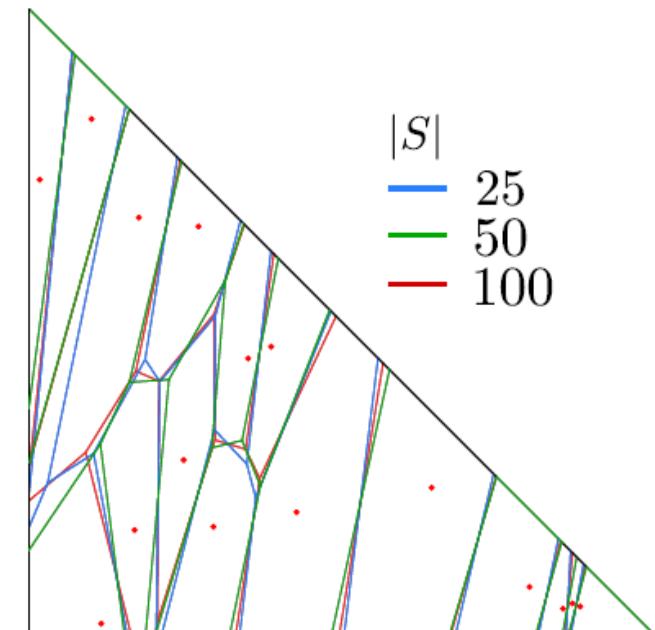
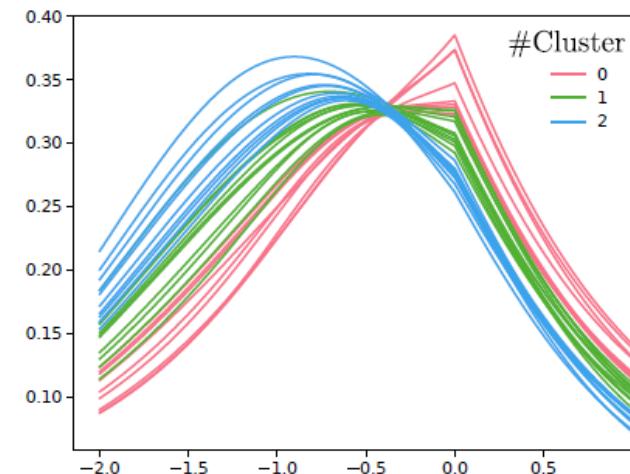
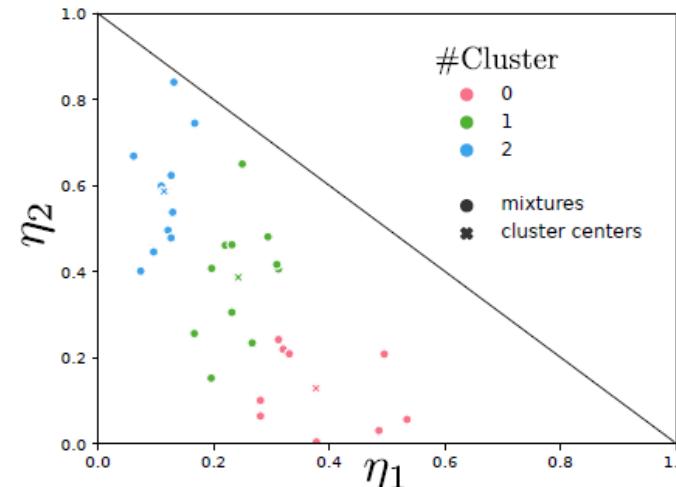
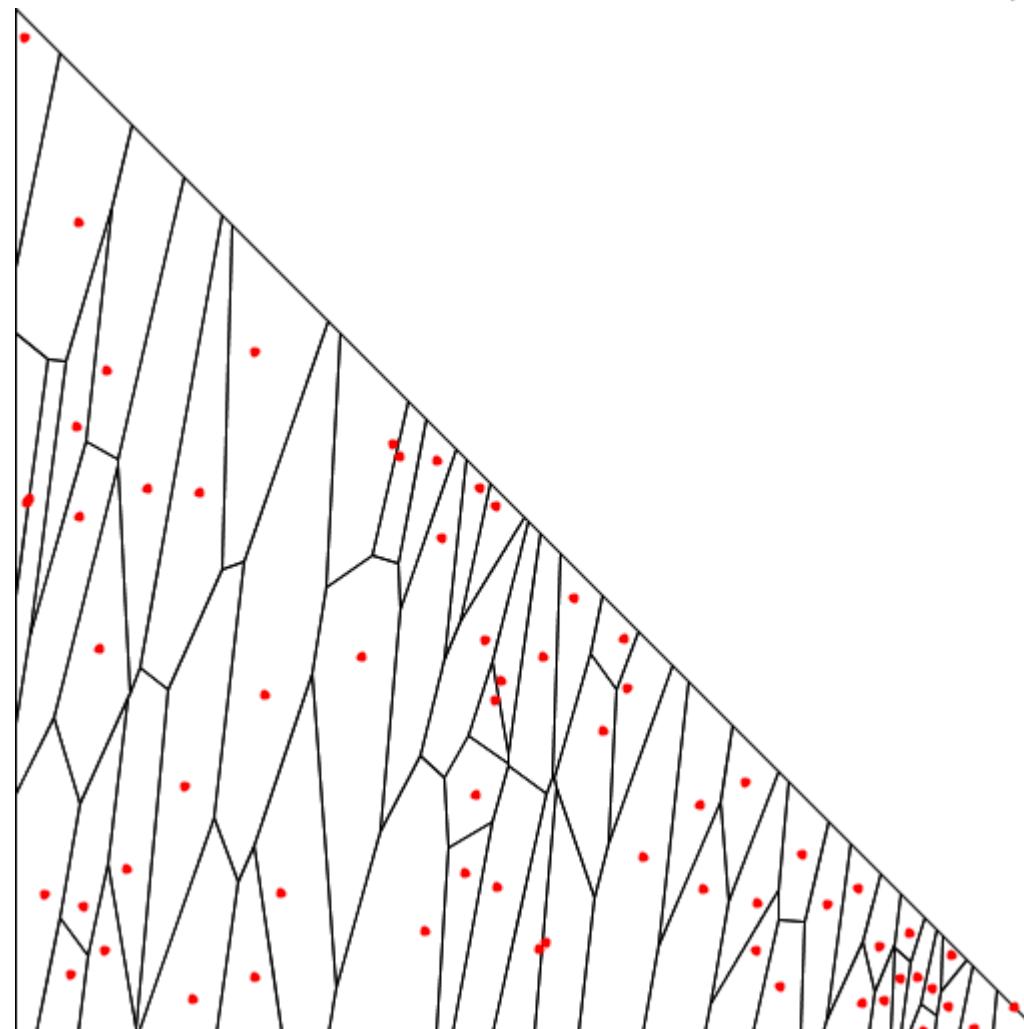
$$I_\gamma(\theta_p, \theta_q) = \exp(F(\theta_p + \gamma\theta_q) - F(\theta_p) - \gamma F(\theta_q)).$$

$$I_\gamma(p, q) = \int_{x \in \mathcal{X}} p(x)q(x)^\gamma dx \simeq \frac{1}{m} \sum_{i=1}^m q(x_i)^\gamma$$



Random/Monte Carlo Bregman Voronoi diagrams

$p_1 = \text{Laplace}(0, 1), p_2 = \mathcal{N}(-1, 1), p_0 = \text{Cauchy}(-0.5, 1)$.



Some statistical distances with closed-form expressions for statistical mixtures

- **Cauchy-Schwarz divergence:** $\text{CS}(P : Q) = -\log \frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2dx \int q(x)^2dx}},$

- For mixtures of exponential families with conic natural parameter space:

$$\begin{aligned}\int m(x)m'(x)dx &= \sum_{i=1}^k \sum_{j=1}^{k'} w_i w'_j \int p_F(x; \theta_i) p_F(x; \theta'_j) dx \\ \int p_F(x; \theta_i) p_F(x; \theta'_j) dx &= e^{F(\theta_i + \theta'_j) - (F(\theta_i) + F(\theta'_j))} \underbrace{\int e^{\langle t(x), \theta_i + \theta'_j \rangle - F(\theta_i + \theta'_j)} dx}_{=1},\end{aligned}$$

$$\int m(x)m'(x)dx = \sum_{i=1}^k \sum_{j=1}^{k'} w_i w'_j e^{F(\theta_i + \theta'_j) - (F(\theta_i) + F(\theta'_j))}$$

When natural parameter space
Is a cone

Examples of conic exponential families (CEFs)

$$\int m(x)m'(x)dx = \sum_{i=1}^k \sum_{j=1}^{k'} w_i w'_j e^{\Delta_F(\theta_i, \theta'_j)}, \quad \Delta_F(\theta_i, \theta'_j) = F(\theta_i + \theta'_j) - (F(\theta_i) + F(\theta'_j)).$$

Bernoulli. $p(x; \lambda) = \lambda^x(1 - \lambda)^{1-x}$ (with $\lambda \in (0, 1)$), $\theta = \log \frac{\lambda}{1-\lambda}$, $\Theta = \mathbb{R}$, $F(\theta) = \log(1 + e^\theta)$.

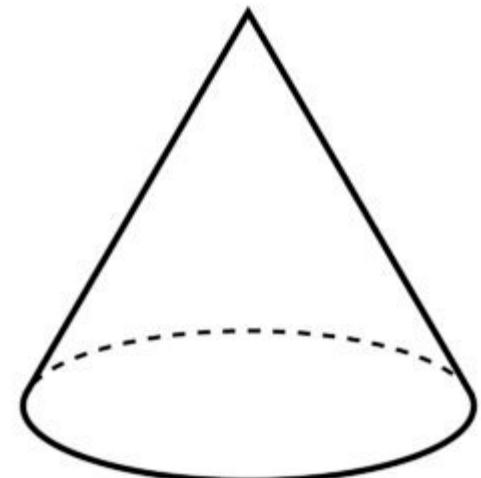
$$\Delta_{\text{Bernoulli}}(\lambda_i, \lambda_j) = \log \frac{1 + \frac{\lambda_i + \lambda_j}{1 - \lambda_i - \lambda_j}}{(1 + \frac{\lambda_i}{1 - \lambda_i})(1 + \frac{\lambda_j}{1 - \lambda_j})}$$

Wishart $p(x; n, S) = \frac{|X|^{\frac{n-d-1}{2}} e^{-\frac{1}{2} \text{tr}(S^{-1} X)}}{2^{\frac{nd}{2}} |S|^{\frac{n}{2}} \Gamma_d(\frac{n}{2})}$, with $S \succ 0$ the scale matrix

and $n > d - 1$ the number of degrees of freedom, where Γ_d is the multivariate Gamma function $\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma(x + (1-j)/2)$. $\theta = (\theta_s, \theta_M) = (\frac{n-d-1}{2}, S^{-1})$ with $\Theta = \mathbb{R}_+ \times S_{++}^d$ the cone of positive definite matrices. $F(\theta) = \frac{(2\theta_s + d + 1)d}{2} \log 2 + (\theta_s + \frac{d+1}{2}) \log |\theta_M| + \log \Gamma_d(\theta_s + \frac{d+1}{2})$.

Zero-centered Laplacian. $p(x; \sigma) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}$, $\theta = -\frac{1}{\sigma}$, $\Theta = (-\infty, 0)$, $F(\theta) = \log(\frac{2}{-\theta})$.

$$\Delta_{\text{Laplacian}}(\sigma_i, \sigma_j) = \log \frac{1}{2(\sigma_i + \sigma_j)}$$



Gaussian. $p(x; \mu, \Sigma) =$

$$\frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left(-\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2} \right),$$

$\theta = (\theta_v, \theta_M) = (\Sigma^{-1}\mu, \Sigma^{-1})$, $\Theta = \mathbb{R}^d \times S_{++}^d$ where S_{++}^d denotes the cone of positive definite matrices of dimension $d \times d$,

$$F(\theta) = \frac{1}{2} \theta_v^T \theta_M^{-1} \theta_v - \frac{1}{2} \log |\theta_M| + \frac{d}{2} \log 2\pi.$$

$$\begin{aligned} \Delta_{\text{Gaussian}}((\mu_i, \Sigma_i), (\mu_j, \Sigma_j)) &= \frac{1}{2} (\\ &\mu_{ij}^T \Sigma_{ij}^{-1} \mu_{ij} - (\mu_i^T \Sigma_i^{-1} \mu_i + \mu_j^T \Sigma_j^{-1} \mu_j) \\ &- \log \frac{|\Sigma_i^{-1} + \Sigma_j^{-1}|}{|\Sigma_i^{-1}| |\Sigma_j^{-1}|} - d \log 2\pi) \end{aligned}$$

References

- Monte Carlo Information-Geometric Structures, Geometric Structures of Information, 2019 (arXiv:1803.07225)
- On the Geometry of Mixtures of Prescribed Distributions. ICASSP 2018 (arxiv:1708.00568)
- On The Chain Rule Optimal Transport Distance. arXiv:1812.08113, 2018
- Patch matching with polynomial exponential families and projective divergences, SISAP, 2016
- Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities, Entropy, 2016
- Bag-of-components: an online algorithm for batch learning of mixture models, GSI 2015
- Model centroids for the simplification of Kernel Density estimators, ICASSP 2012
- Closed-form information-theoretic divergences for statistical mixtures, ICPR 2012
- Comix: Joint estimation and lightspeed comparison of mixture models. ICASSP 2016

Advanced topics and perspectives

Frank Nielsen



Sony CSL

Outline

- Representations and basis of the Fisher information matrix
- Affine differential geometry and immersions
- Wong's logarithmic-divergence
- Kernel exponential families
- Some limitations

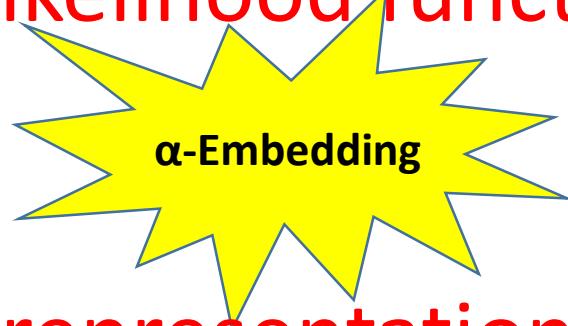
α -representations of the FIM

We introduced the
FIM in two ways
formerly

$$I(\theta) := (I_{ij}(\theta)), \quad I_{ij}(\theta) := E_{p(x;\theta)}[\partial_i l(x; \theta) \partial_j l(x; \theta)].$$

$$I'_{ij}(\theta) := 4 \int \partial_i \sqrt{p(x; \theta)} \partial_j \sqrt{p(x; \theta)} d\nu(x)$$

α -likelihood function $l^{(\alpha)}(x; \theta) := k_\alpha(p(x; \theta))$



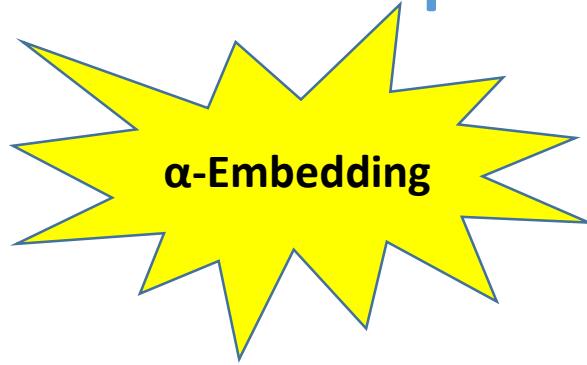
α -representation of the FIM

$$k_\alpha(u) = \begin{cases} \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1 \\ \log u, & \text{if } \alpha = 1. \end{cases}$$

$$I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)}(x; \theta) \partial_j l^{(-\alpha)}(x; \theta) d\nu(x)$$

Corresponds to a basis choice in the tangent space (α -base)

α -representations of the FIM



$$I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)}(x; \theta) \partial_j l^{(-\alpha)}(x; \theta) d\nu(x)$$

- 0-representation (square root) : $I'_{ij}(\theta) := 4 \int \partial_i \sqrt{p(x; \theta)} \partial_j \sqrt{p(x; \theta)} d\nu(x)$
- 1-representation (log): $I_{ij}(\theta) := E_{p(x; \theta)}[\partial_i l(x; \theta) \partial_j l(x; \theta)]$
- Under mild regularity conditions:
$$I_{ij}^{(\alpha)}(\theta) = -\frac{2}{1+\alpha} \int p(x; \theta)^{\frac{1+\alpha}{2}} \partial_i \partial_j l^{(\alpha)}(x; \theta) d\nu(x)$$
- Coefficients of the connection: $\Gamma_{ij,k}^{(\alpha)} = \int \partial_i \partial_j l^{(\alpha)} \partial_k l^{(-\alpha)} d\nu(x)$

The α -representations of the Fisher Information Matrix, 2017

(ρ, τ) -representations of the FIM

Smooth convex function and convex conjugates: $f^*(t) = t(f')^{-1}(t) - f((f')^{-1}(t))$

τ -representation

$$\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p))$$

ρ -representation

$$\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p))$$

(ρ, τ) -FIM

$$g_{ij}(\theta) = E_\mu \left\{ f'' \left(\rho(p(\zeta|\theta)) \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \right) \right\}$$

(ρ, τ) - α -connections

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} f'''(\rho(p(\zeta|\theta))) A_{ijk} + f''(\rho(p(\zeta|\theta))) B_{ijk} \right\}$$

$$A_{ijk}(\zeta, \theta) = \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k}, \quad B_{ijk}(\zeta, \theta) = \frac{\partial^2 \rho(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k}$$

Affine differential geometry

Copyrighted Material

CAMBRIDGE TRACTS IN MATHEMATICS

111

**AFFINE DIFFERENTIAL
GEOMETRY**

KATSUMI NOMIZU & TAKESHI SASAKI



CAMBRIDGE UNIVERSITY PRESS

Copyrighted Material

Applications in statistical physics

- Scalar curvature of Fisher-Rao information metric explains phase transitions: Information geometry, one, two, three (and four), 2003

Wong's logarithmic divergence

Wong, Ting-Kam Leonard. "Logarithmic divergences from optimal transport and Rényi geometry." *Information Geometry* 1.1 (2018): 39-78.

Limitations of parametric frameworks

- The f-divergence between 1-to-1 smooth transformations of variables yields the same parametric divergence, and the same information geometry
- Eg., KL and f-divergences between normal or log-normal have same formula (via $y=\log x$)
- Fisher-Rao between elliptical distributions with fixed dispersion matrix is proportional to Mahalanobis distance
- Optimal transport formula is the same for elliptical distributions and coincide with the formula for Gaussian measures

Perspectives: Not covered topics

- Deformed exponential families
- Kernel exponential families, deep exponential families
- Non-parametric information geometry
- Quantum information geometry