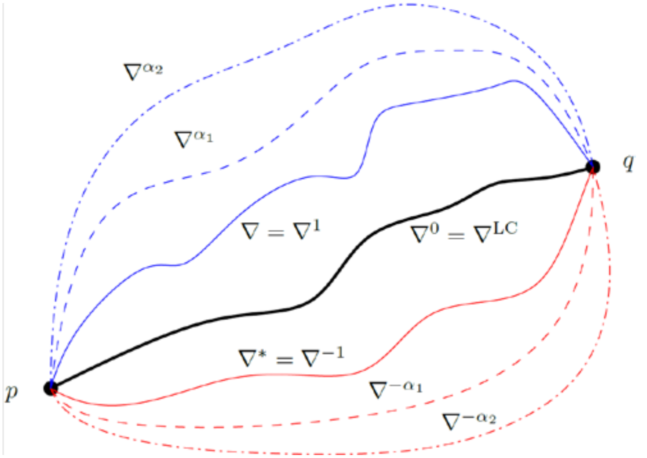
Information geometry for information sciences:
- A first intuitive overview -

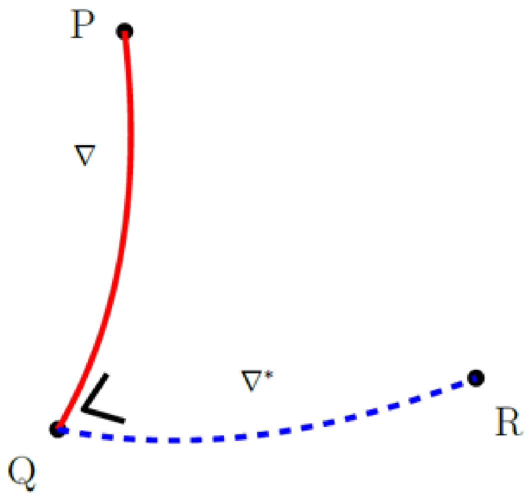


Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/



An elementary introduction to information geometry

https://arxiv.org/abs/1808.08271

The goal of this talk is to…

- Present the *main ideas* behind the dualistic structures of information geometry

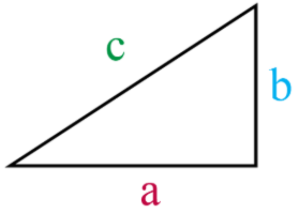
- Avoid common misconceptions and pitfalls

- 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



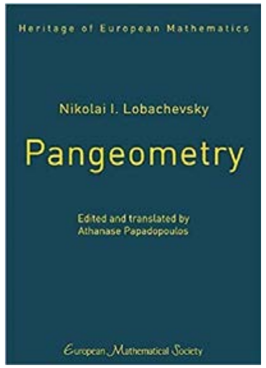
$a^2 + b^2 = c^2$

- Pythagoras's theorem (c570-495 BC)
- Euclid's axiomatization and deduction (c300 BC)

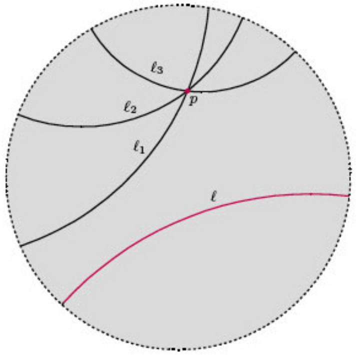
  Euclidean geometry

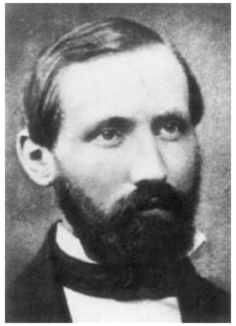
- Figures, congruences, construction with compass/rulers

Big bang!

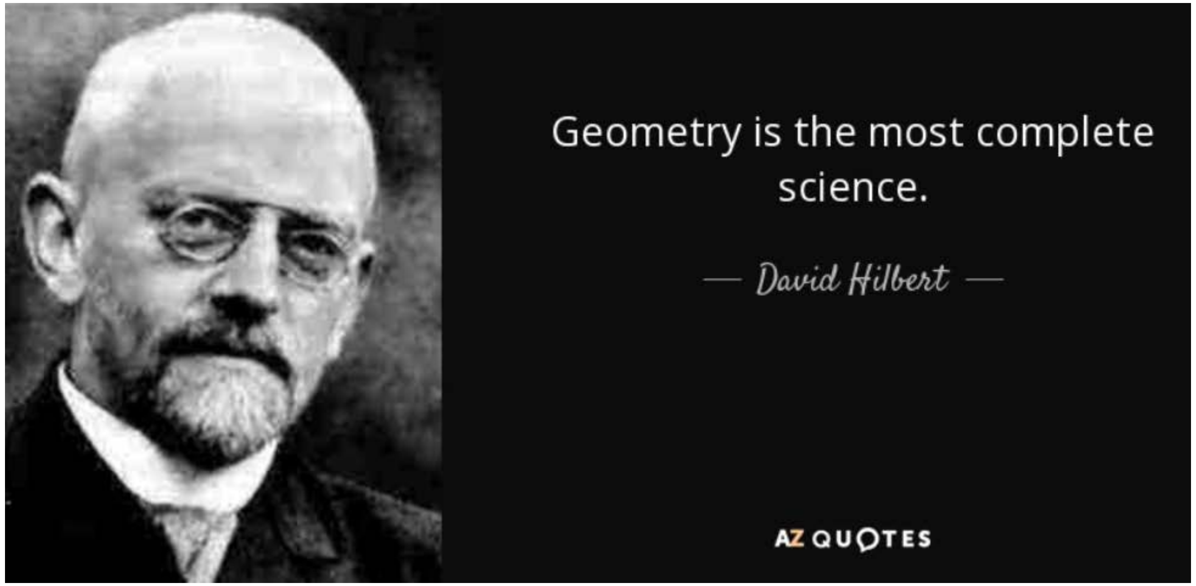


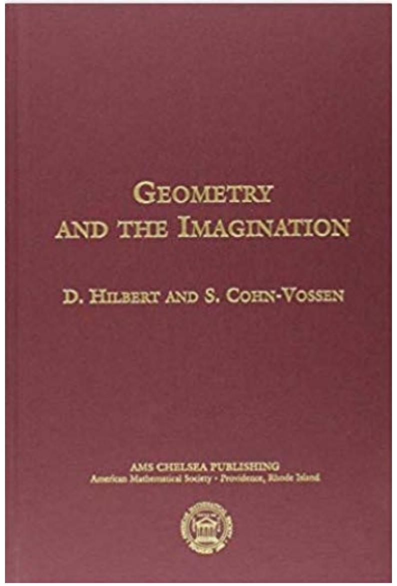
- Lobachevskian hyperbolic geometry is consistent (c1800)
- Riemannian geometry (c1850): <u>infinitely many</u> consistent differential geometries
- Klein's Erlangen program: classification (action of a group)
- Etc.

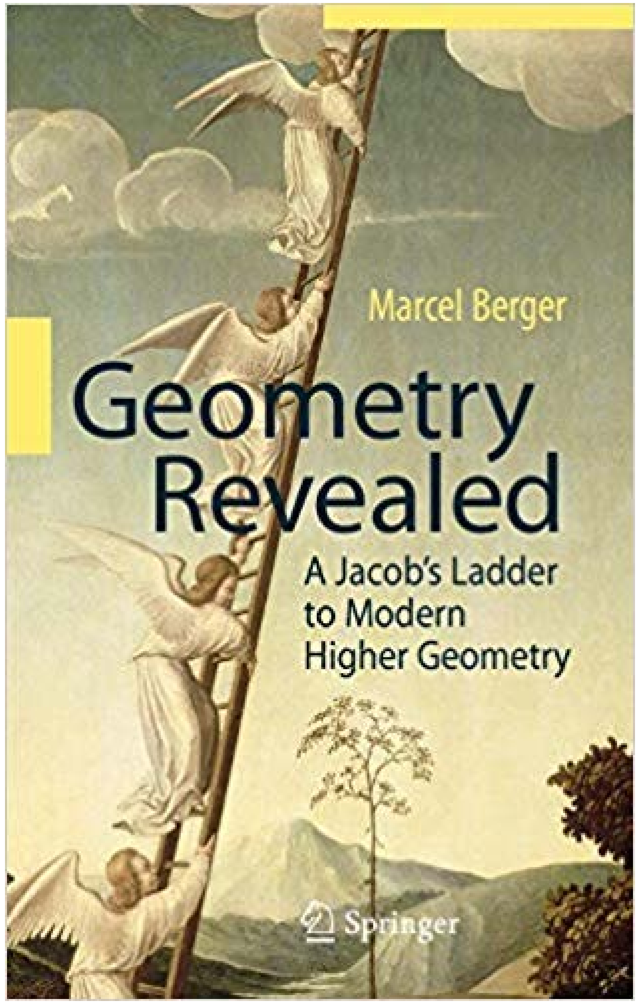


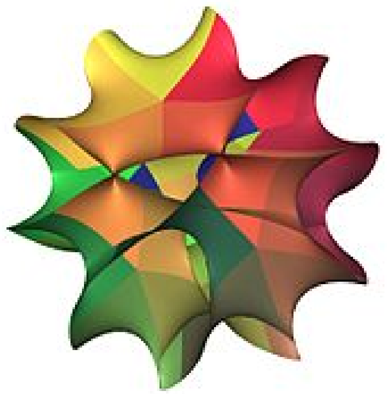


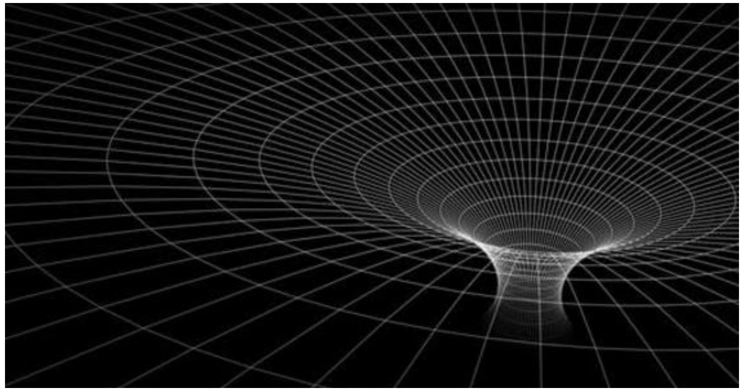
Geometry is an incredibly creative science!

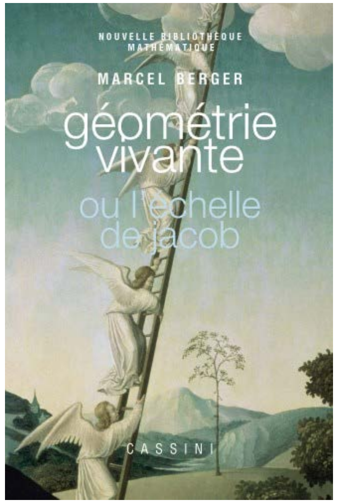






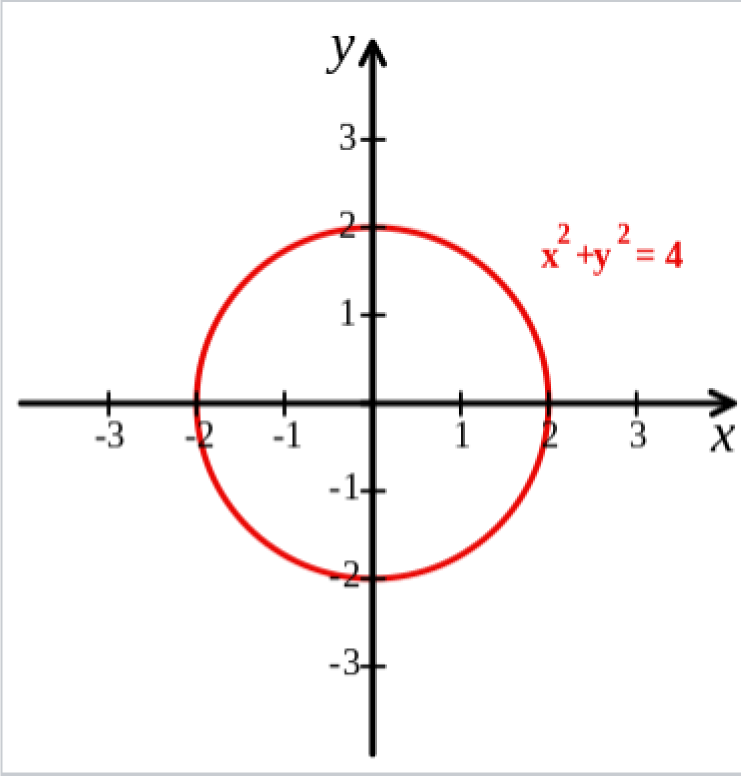


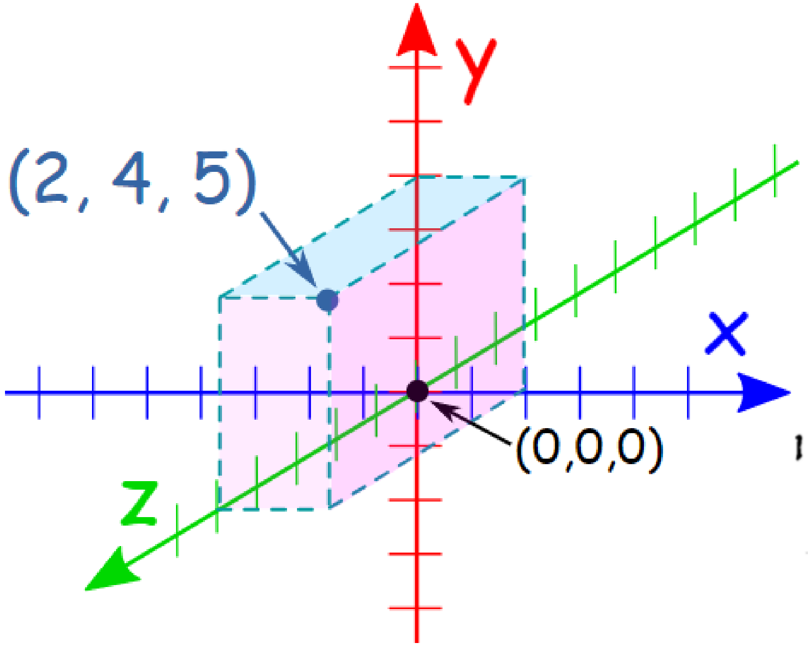


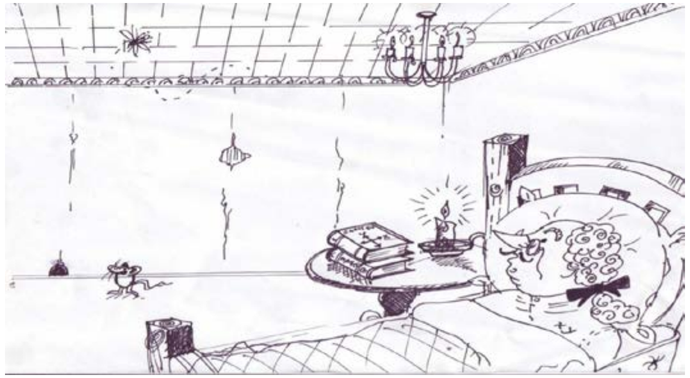


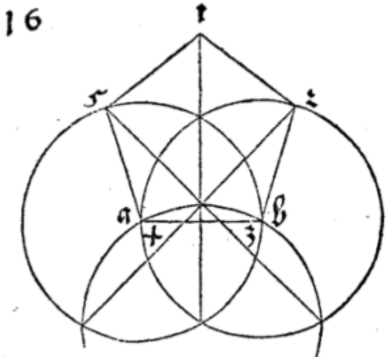
Analytic versus synthetic geometry

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







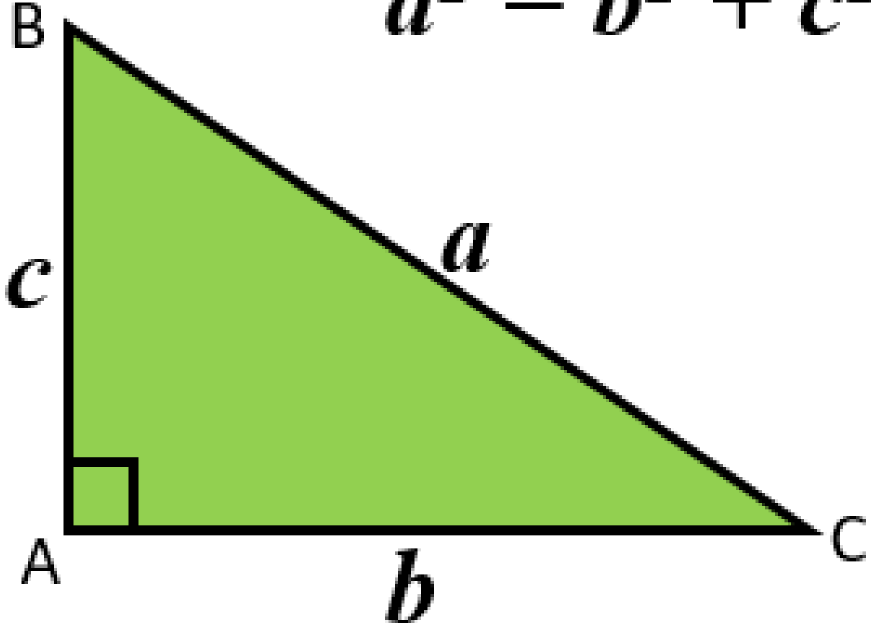


Pythagoras' / Pythagorean theorem

Circa 500 BC

- Yields formula of Euclidean distance in Cartesian coordinate system

$$a^2 = b^2 + c^2$$



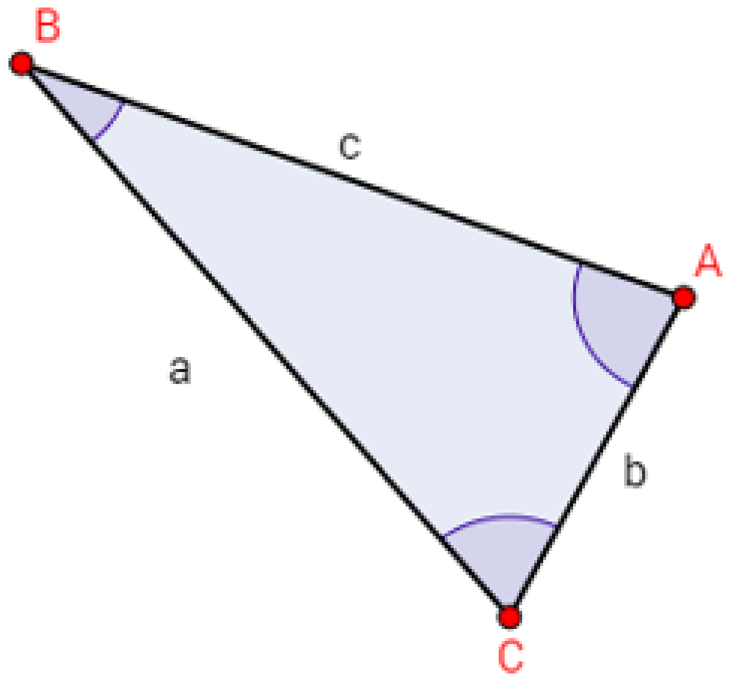
At the heart of the Euclidean distance

$$d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

Pythagoras' theorem allegedly know in Babylonian mathematics (2000-1600 BC)

Pythagoras' theorem generalizes to the law of cosines for *arbitrary* triangles

Law of Cosines



$$a^2 = b^2 + c^2 - 2bc\,\mathrm{Cos}A$$

$$b^2 = a^2 + c^2 - 2ac\,\mathrm{Cos}B$$

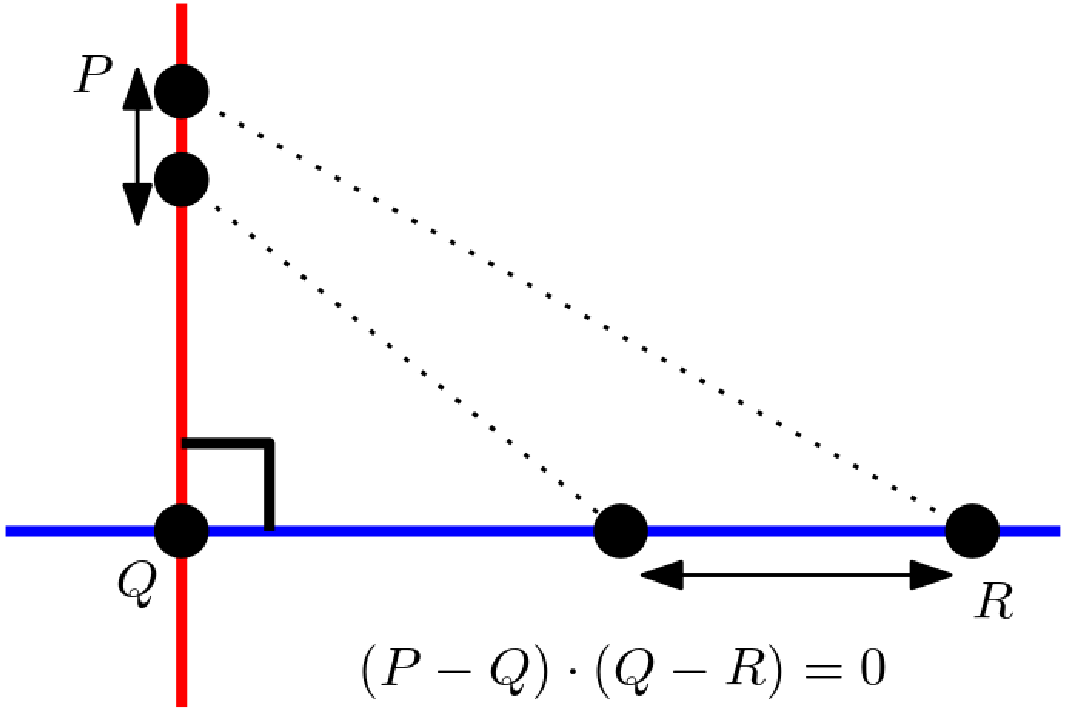
$$c^2 = a^2 + b^2 - 2ab\,\mathrm{Cos}C$$

We shall see that for Bregman manifolds in information geometry …

… we have dual Pythagorean theorems with generalized law of cosines

A modern view of Pythagoras' theorem: A triangle PQR is rectangle if and only if straight lines perpendicular at Q induce distance identity

Squared Euclidean distance $$D_E(X,Y) = d_E^2(X,Y) = \|X - Y\|^2$$



$$(P-Q)\cdot(Q-R)=0$$

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

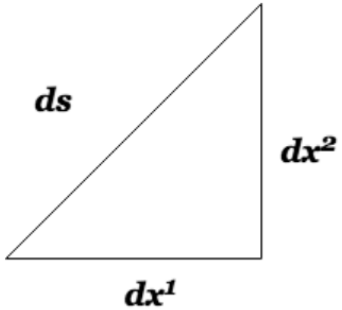
Riemannian differential geometry

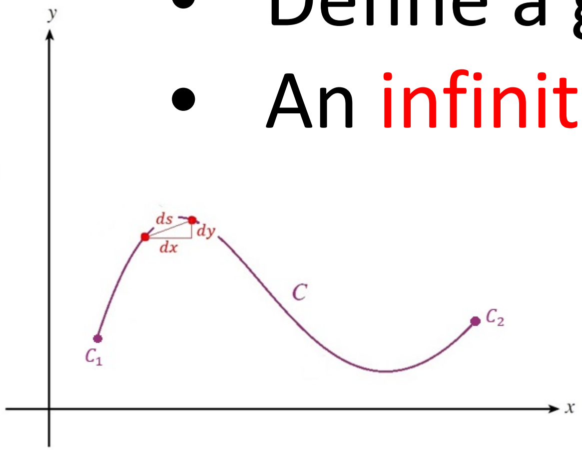
- Gauss pioneered the study of 3D surfaces and curvature

Gauss

Riemann

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



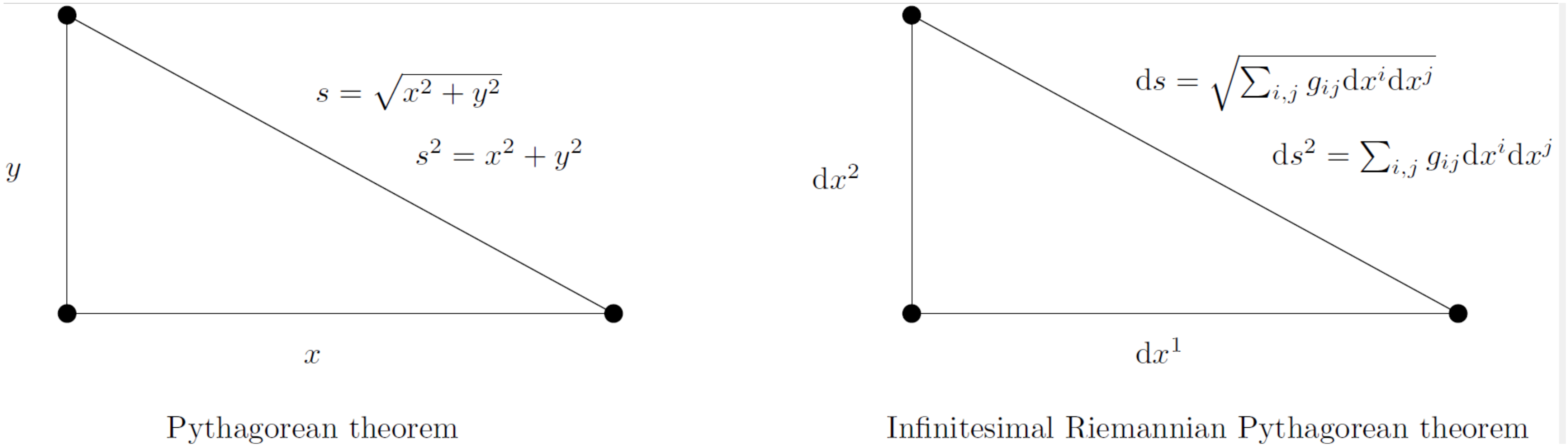


Length of a curve by integration

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

Infinitesimal length element: $ds^2 = g_{11} du^2 + 2 g_{12} du\, dv + g_{22} dv^2$

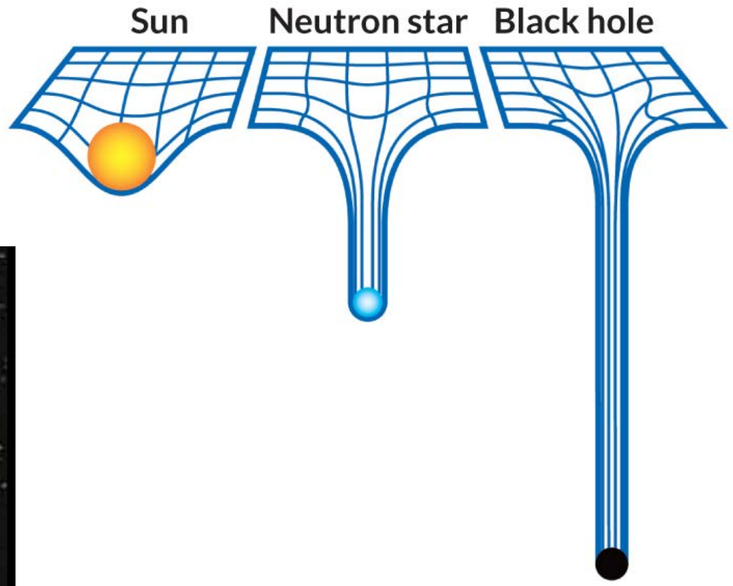
Riemannian geometry: Infinitesimal Pythagorean theorem

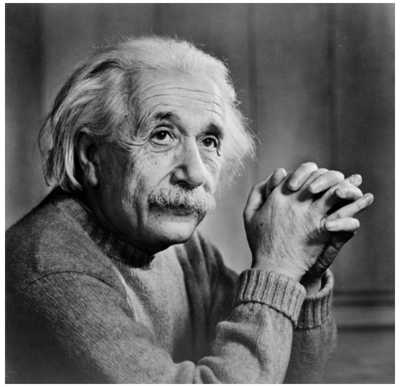


Infinitesimal length element ds

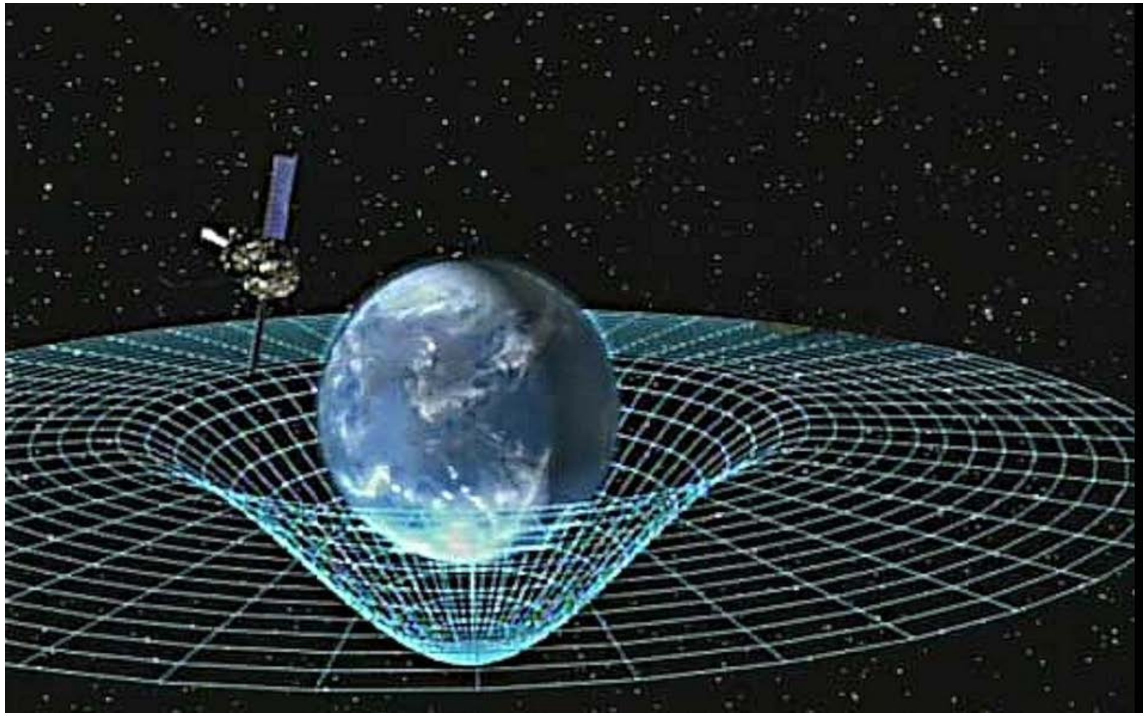
Riemannian distance is (locally) **length of shortest path**

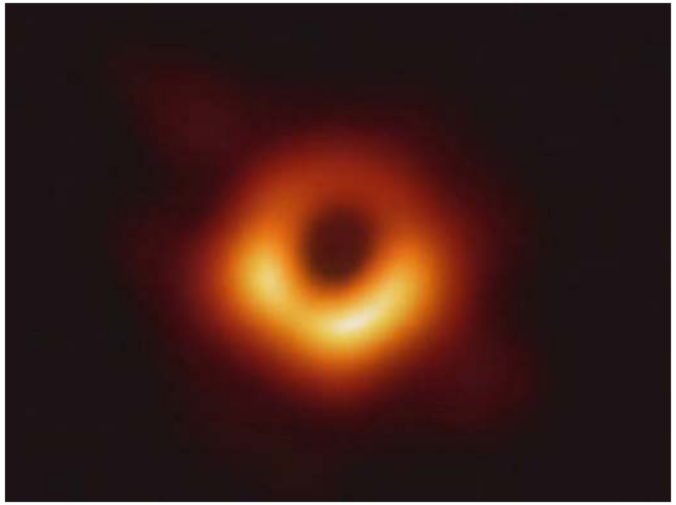
Riemannian geometry: A revolution that changed our perception of the universe and data science

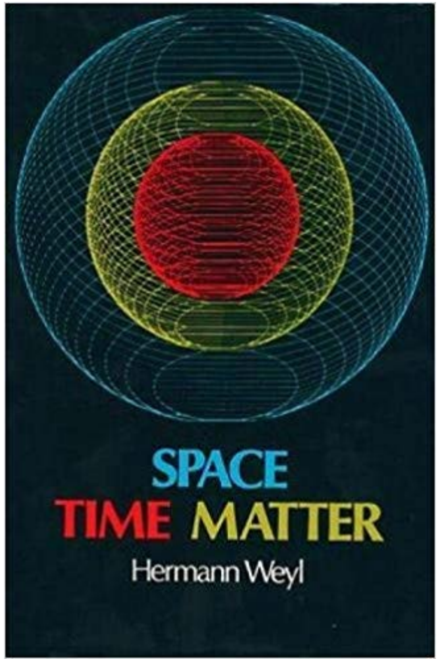




General relativity of spacetime



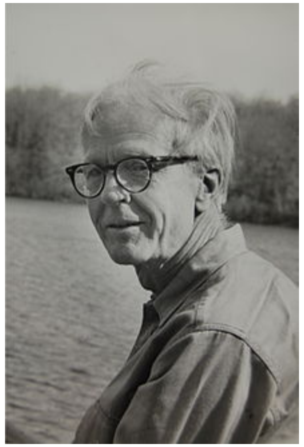




Spacetime+Matter

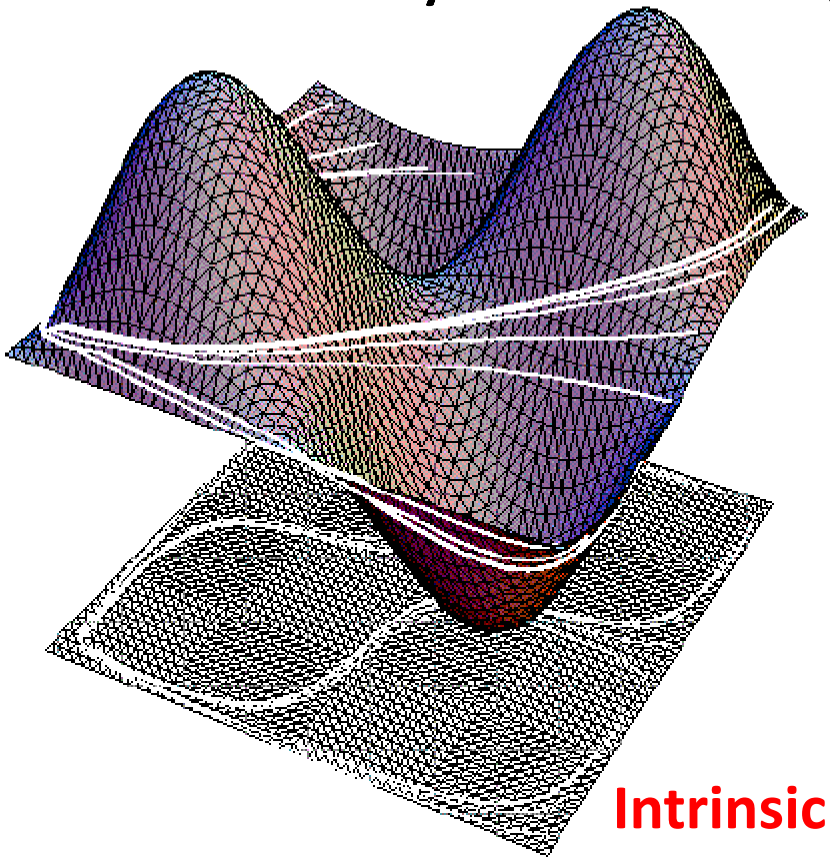
Riemannian manifolds: Extrinsic vs intrinsic views

Visualized **extrinsically** as smooth surfaces of the ambient Euclidean space: Whitney embedding theorem



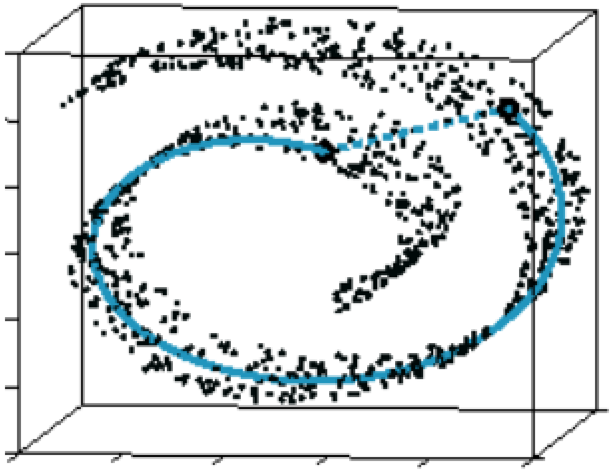
Hassler Whitney
(1907-1989)

Isometric embedding:



Extrinsic geometry

Intrinsic geometry

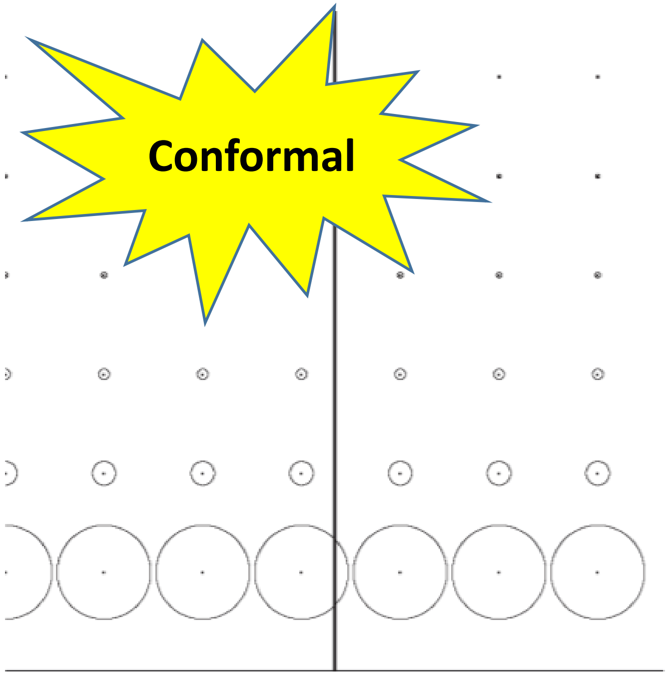


Manifold learning/reconstruction from data points (Swiss roll)

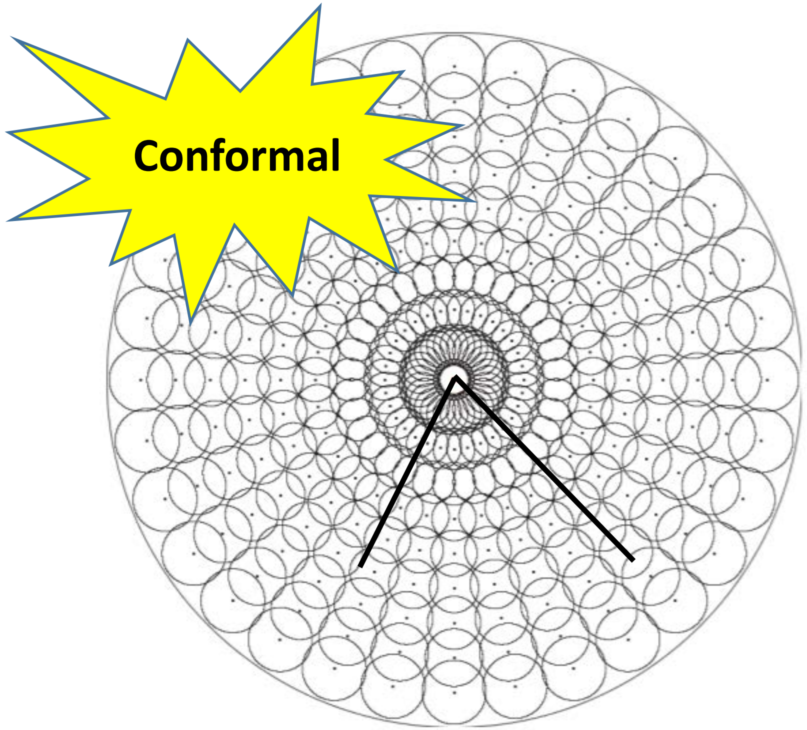
Intrinsic geometry versus isometric Whitney embedding (in dim 2D)

Conformal versus non-conformal metric tensor field:

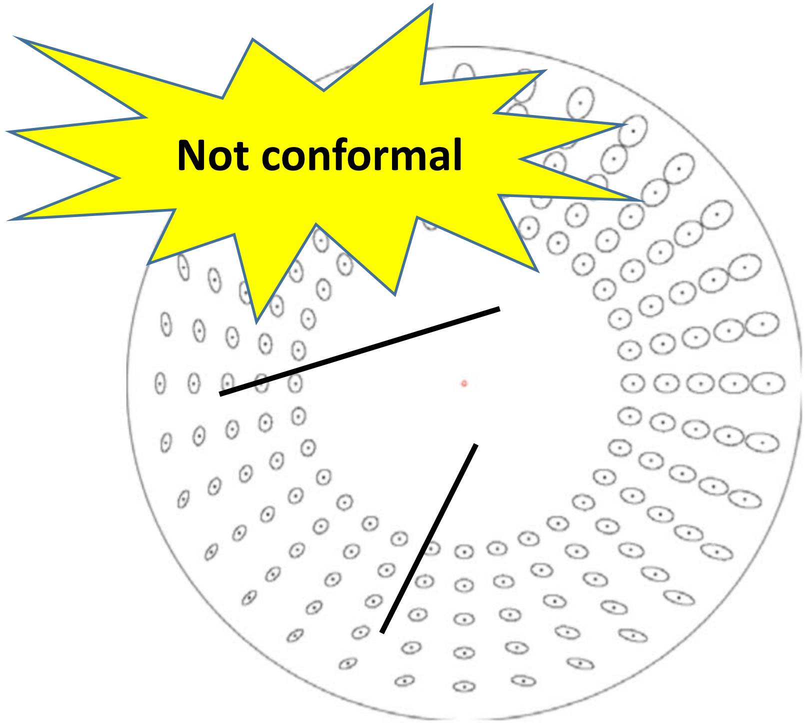
Hyperbolic geometry



Upper Poincare plane
(conformal)



Poincare disk
(conformal)



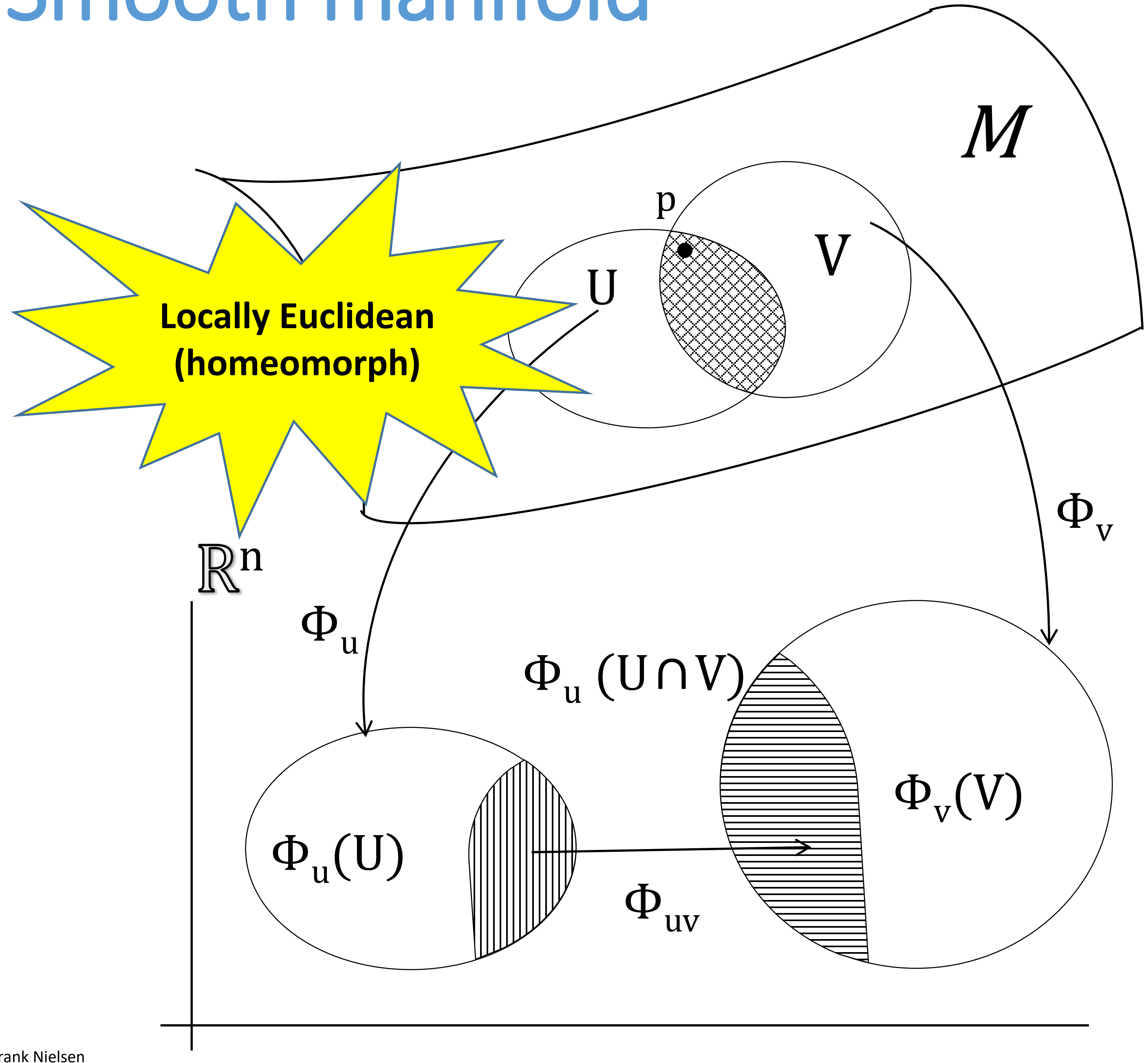
Klein disk
(non-conformal)

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

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

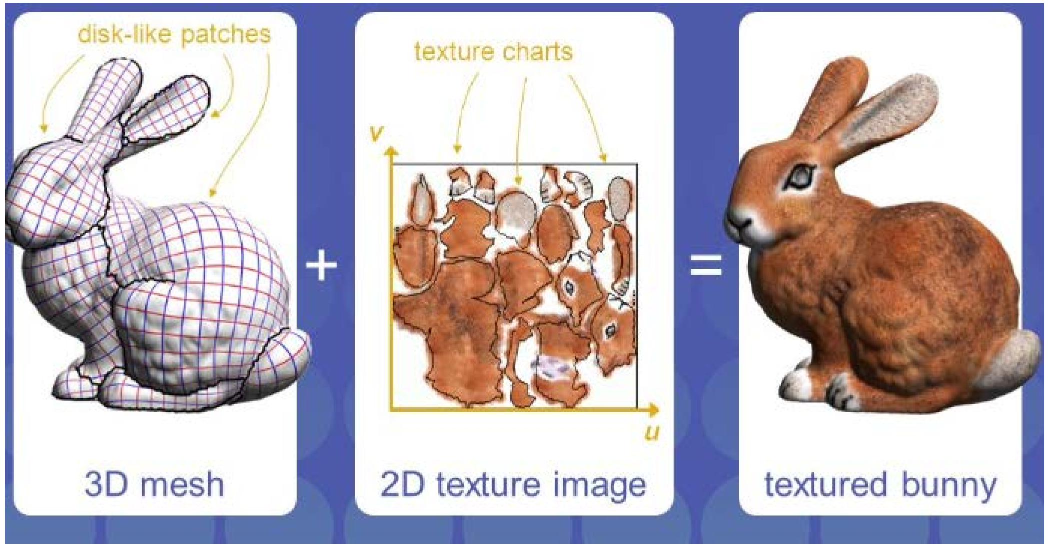
In conformal geometry, we can measure angles without distortions

Smooth manifold



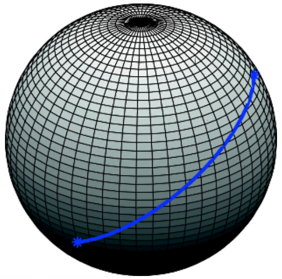
Global geometric objects
vs
Local descriptions
in local chart coordinates

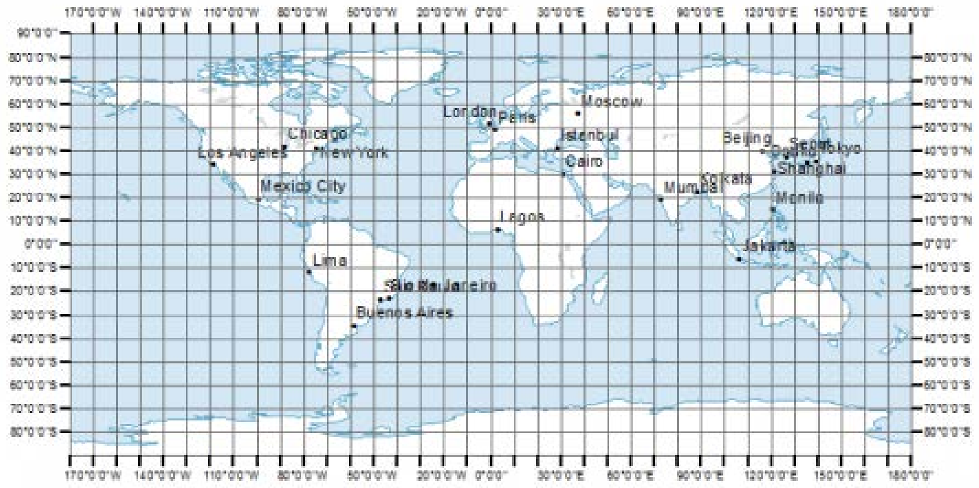
Atlas
Coordinate charts



UV mapping in computer graphics

Visualizing (shortest) paths in a chart: (i.e., in local coordinates)

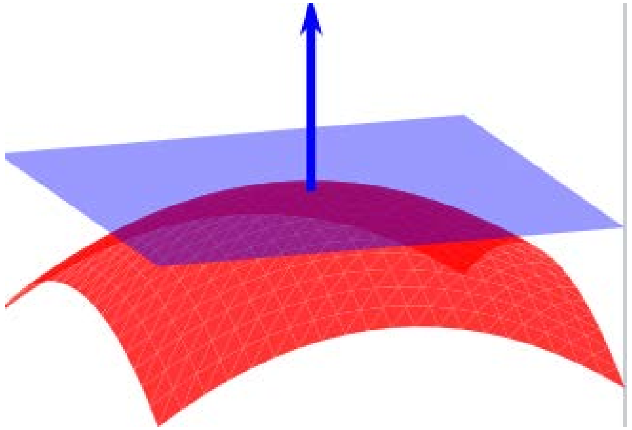


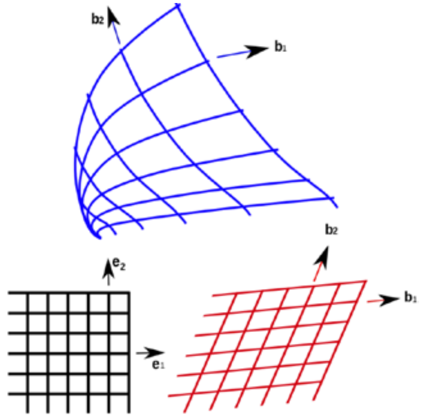


You can only visualize a geometry by rasterizing in a (local) coordinate chart or drawing (conceptual) figures, or much better imagining it in your head!

Lev Semenovich Pontryagin (1908–1988)
blind by accident at 14 yo

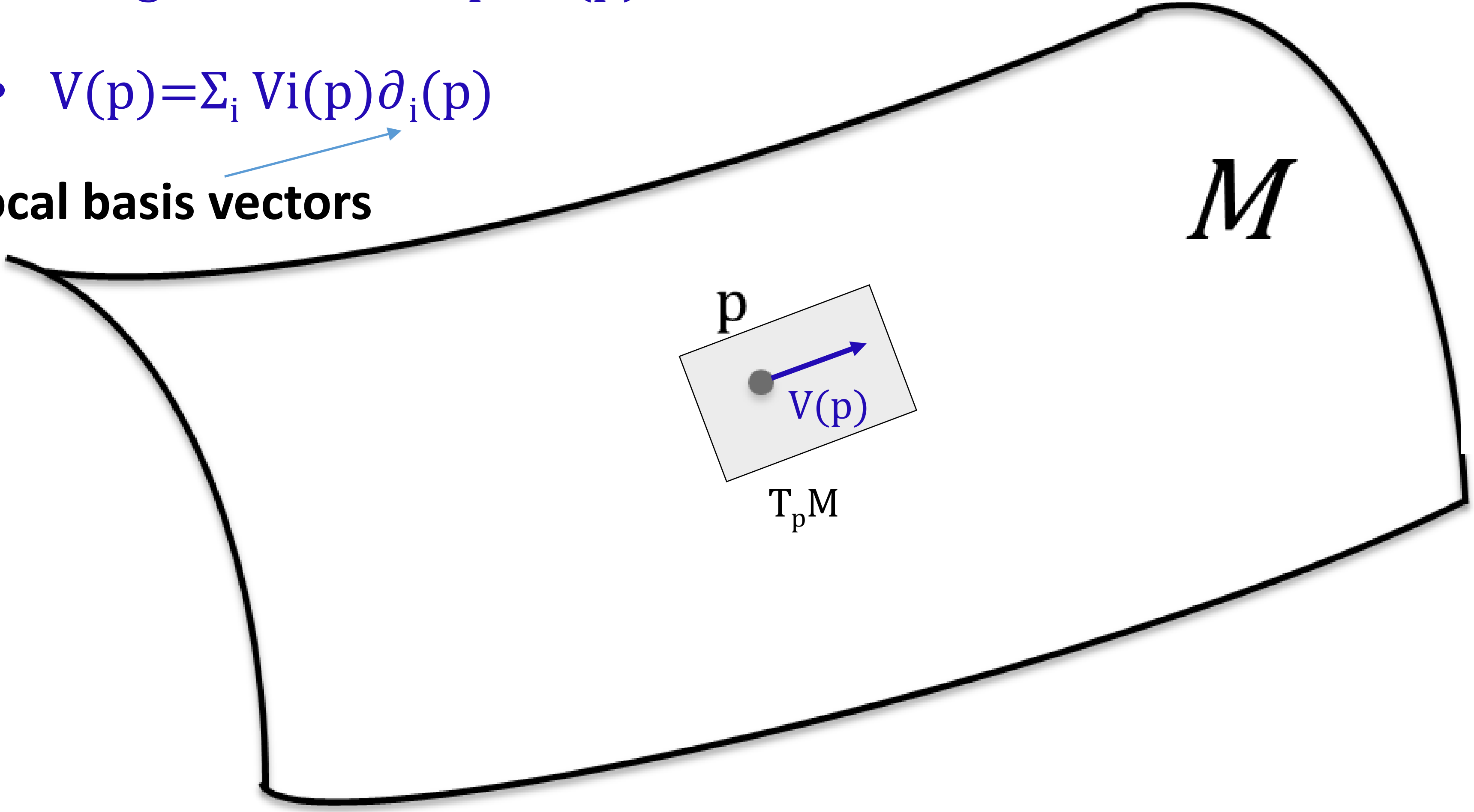
Manifold: Tangent spaces

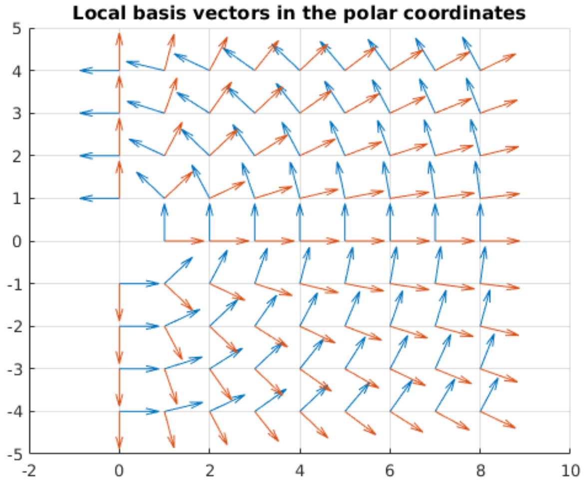




- Tangent space at p : $T_pM \cong \mathbb{R}^n$
- Tangent vector at p : V(p)
- $V(p)=\Sigma_i Vi(p)\partial_i(p)$

Local basis vectors

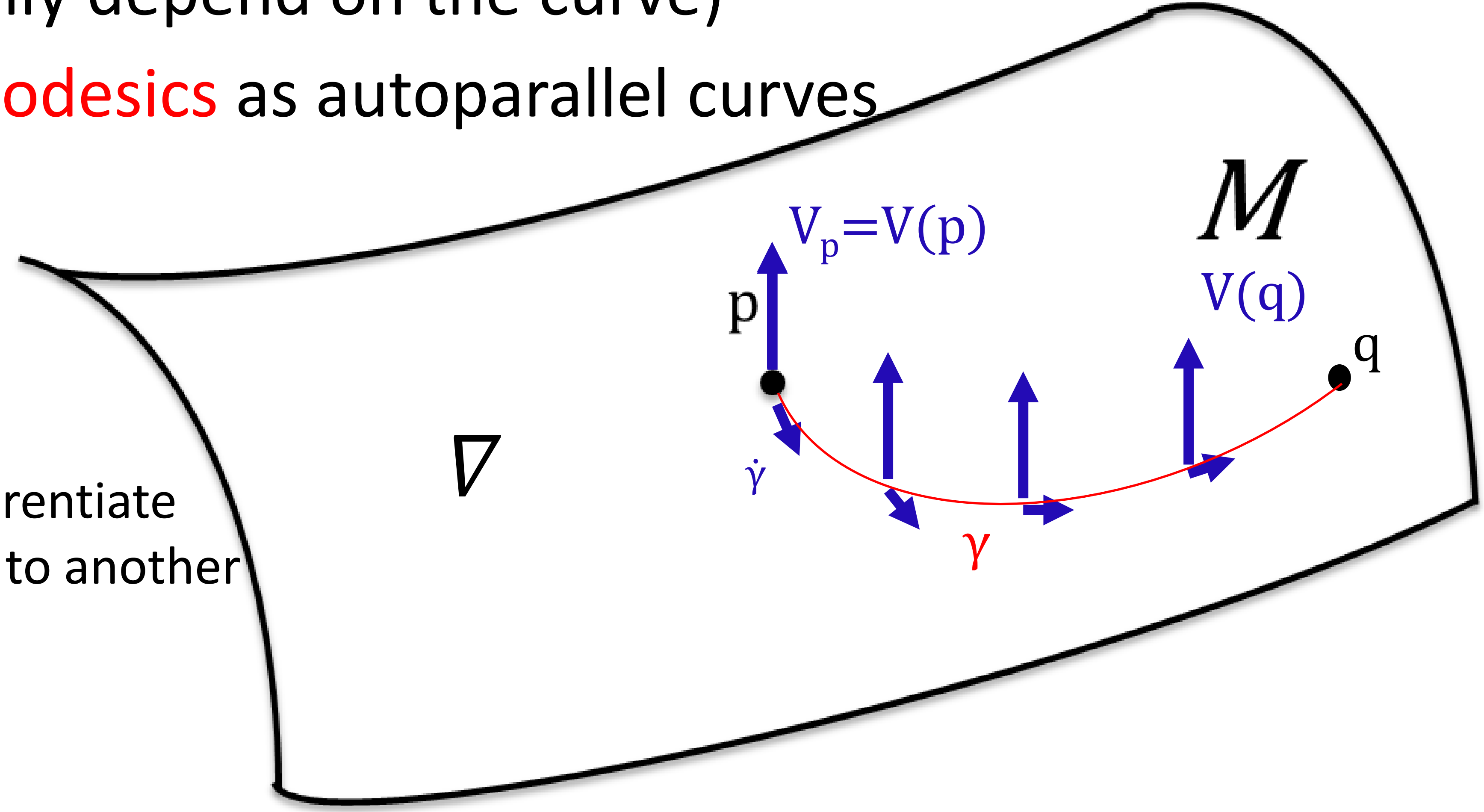




Intrinsic geometry view: interpret a vector as a **directional derivative** and not as an arrow

An essential concept: 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 (thus generally depend on the curve)
- Use to define ∇-geodesics as autoparallel curves

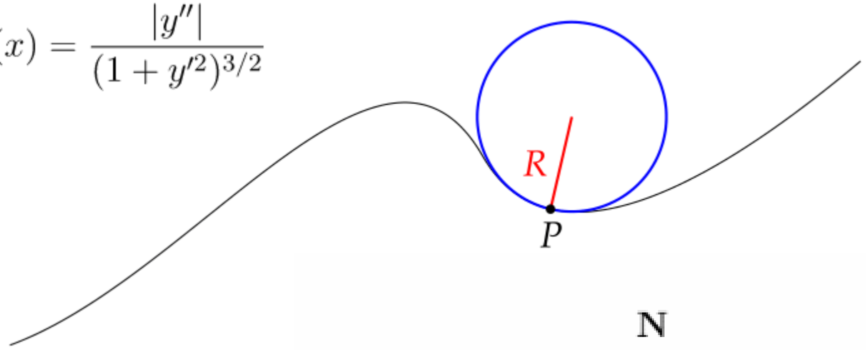


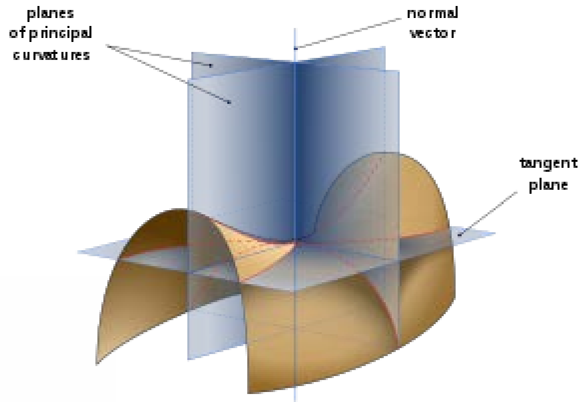
Also provide a way to differentiate a vector field with respect to another vector field called the **covariant derivative**

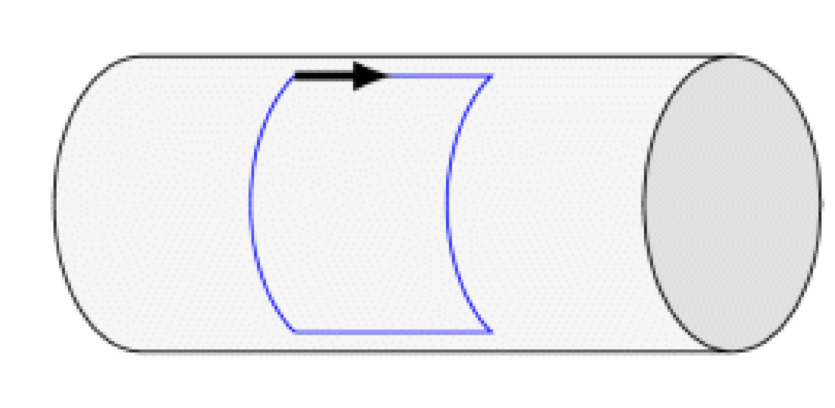
https://arxiv.org/abs/1808.08271

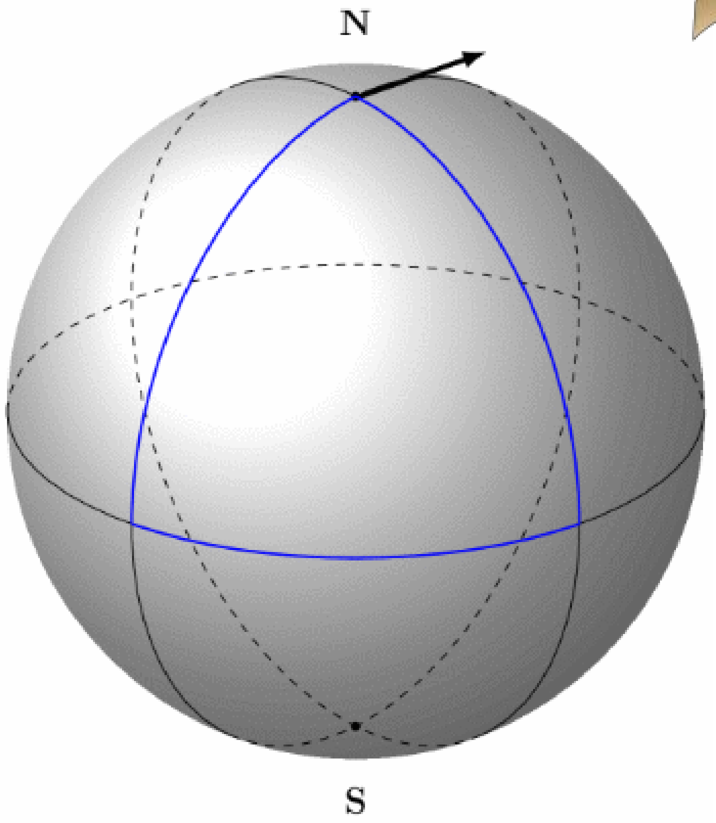
Curvature of a connection ∇

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









Cylinder is flat:
Parallel transport is
path-independent

Sphere has constant curved curvature:
Parallel transport is path-dependent

A word about the torsion of a connection ∇

Torsion measures the speed of rotation of the binormal vector

parallel transport “twists” vectors.

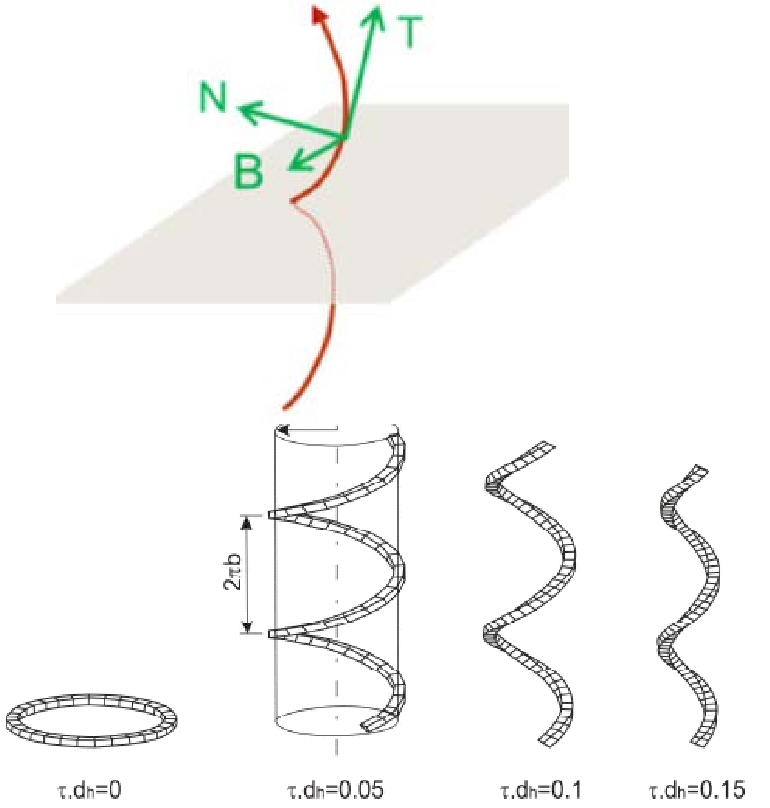


Figure 1. Helical channels with square cross section, constant curvature κ.dh = 1 and torsion τ.dh spanning from 0 to 0.15.

Torsion in geometry and in field theory 3

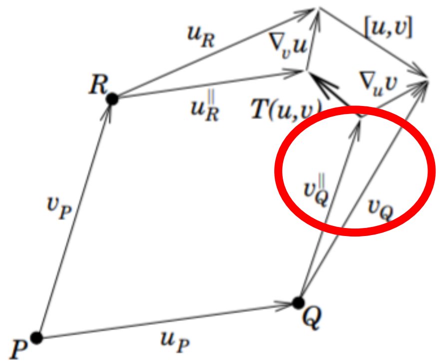
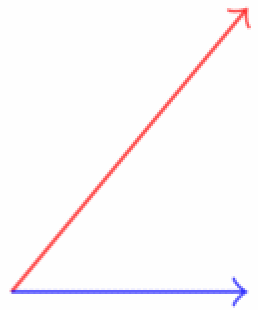


Figure 1: *On the geometrical interpretation of torsion*, see [39]: Two vector fields u and v are given. At a point P, we transport parallelly u and v along v or u, respectively. They become $u_R^{\|}$ and $v_Q^{\|}$. If a torsion is present, they don't close, that is, a *closure failure* $T(u,v)$ emerges. 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.



Failing to close a “parallelogram”

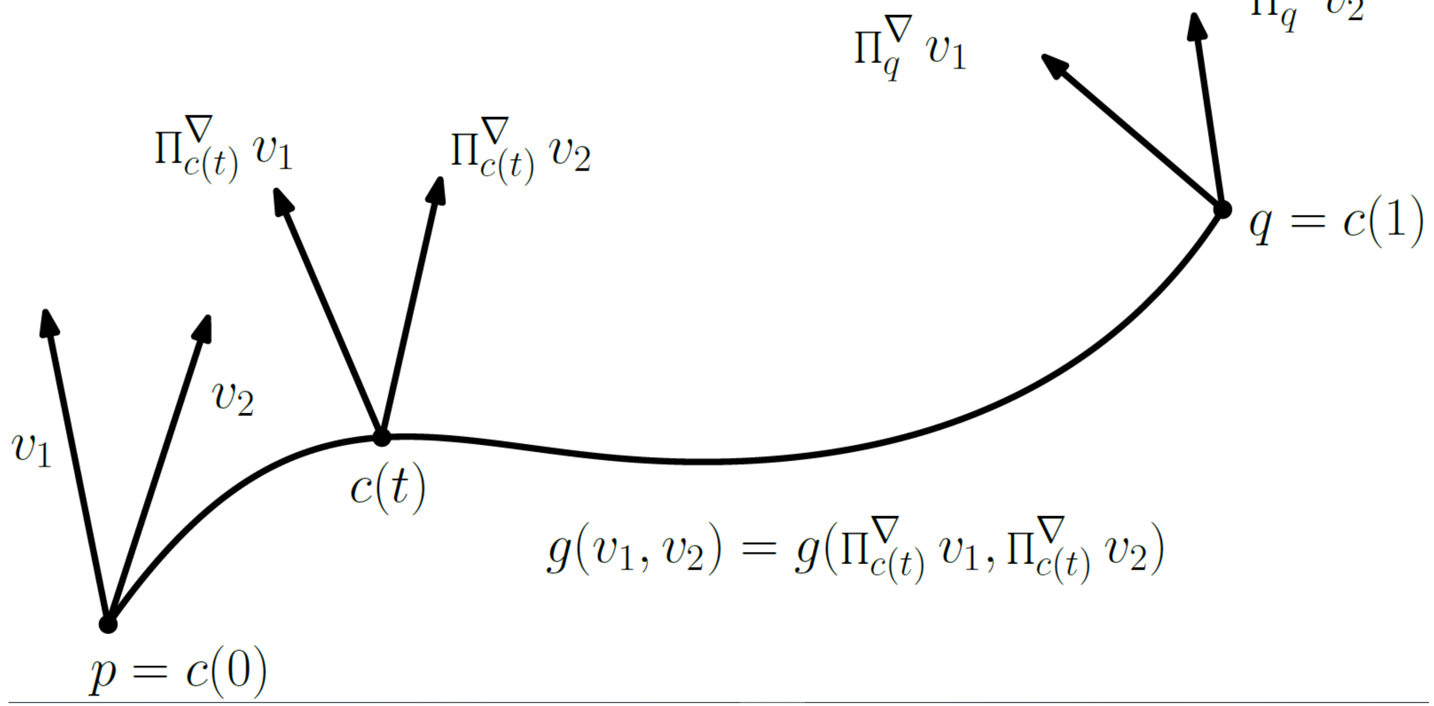
Connections differing by torsions have same geodesics

Pregeodesics= geodesic shapes without parameterization

Metric-compatible connection ∇

Preserves the “inner product” of vectors by parallel transport

Preserves the metric



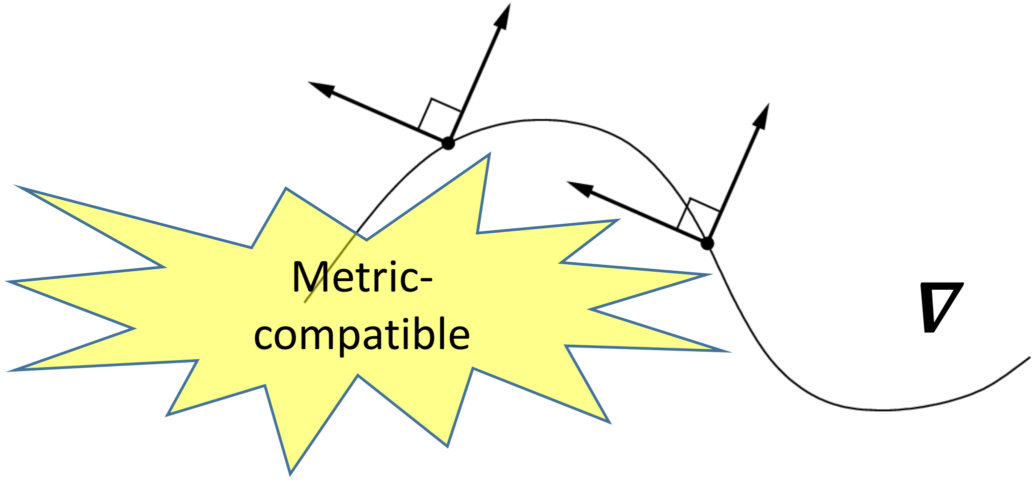
You can measure lengths or angles consistently at any tangent plane

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

The fundamental theorem of Riemannian geometry

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

Tullio Levi-Civita (1873-1941)



Elie Joseph Cartan (1869-1951)

Riemannian geometry: take the Levi-civita metric connection

Differential geometry: take any affine connection (Elie Cartan)

Information geometry: take a pair of "dual" connections

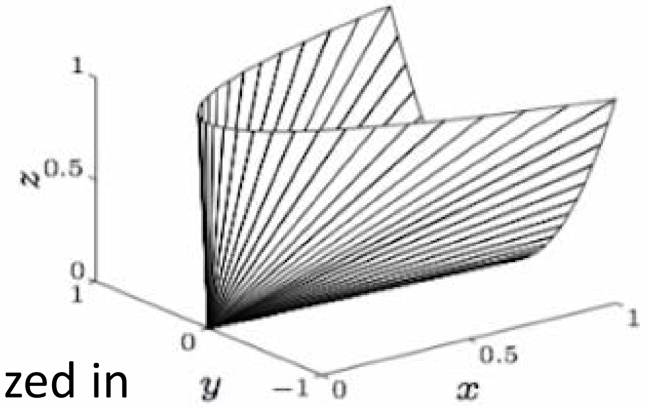
Rationale for information spaces

- In traditional geometry, a space is an empty vacuum
- In physics, a spacetime contains matter

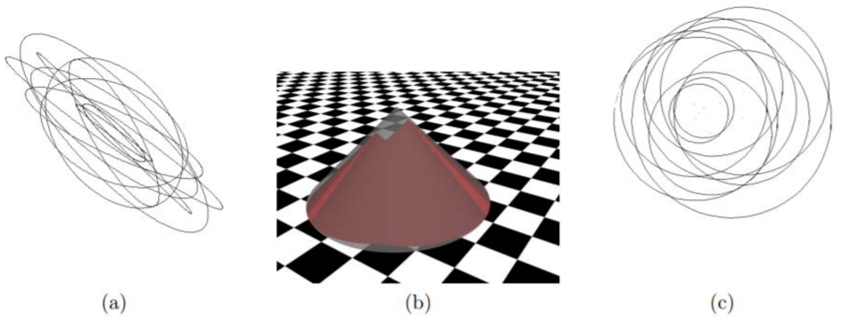
(torsion in General Relativity of Einstein-Cartan)

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

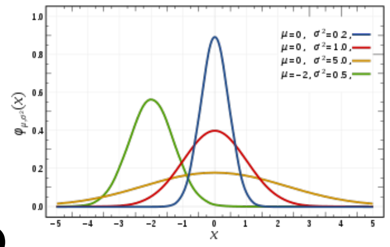


Cone of positive-definite 2x2 matrices visualized in 3D



https://arxiv.org/abs/1604.01592

Rationale for Information Geometry (IG)

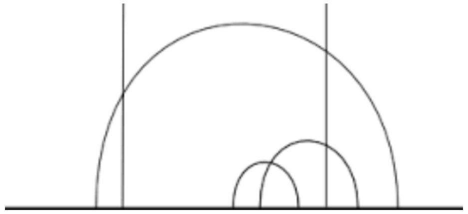


- What is the/a geometry of the space of Gaussian densities?

  Distance, interpolation, closest Gaussian of a subfamily (projection)?

  Note that appropriate geometry may depend on applications

- IG discovered a **dualistic geometry** that can also be used in other non-statistical contexts too!

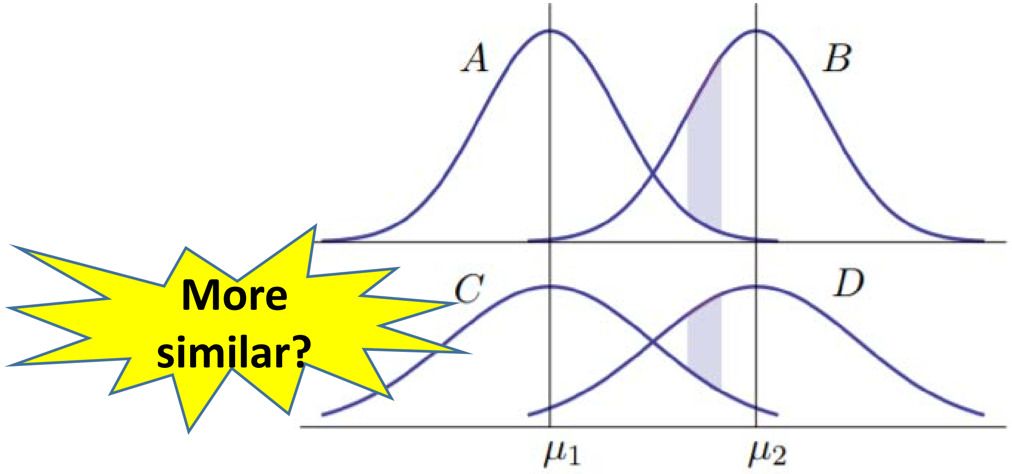


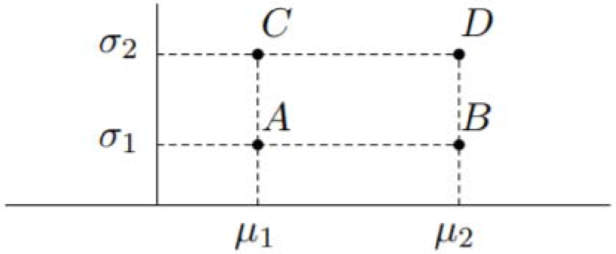
- Applications of the IG framework to information sciences (statistics, information theory, signal processing, machine learning, etc.).

  Mainly, because Information Sciences consider asymmetric distances

What is the geometry of the Gaussian manifold?

- **Euclidean geometry/distance yields this interpretation:**



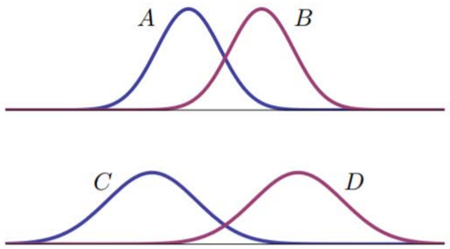


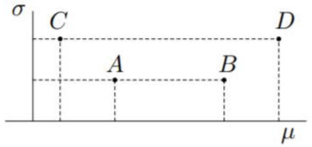
- Desiderata: Dissimilarity shall be invariant to reparameterization:

Same distance for parameterizations {N(μ, σ)} or {N(μ, σ2)}

No geometry of the sample space

Furthermore, invariant by "sufficient statistics"





Equidistant (Rao distance)

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

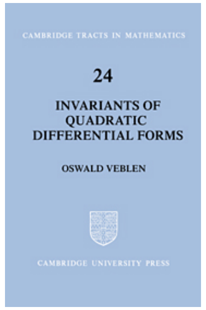
Fisher-Riemannian geometry (1930/1945)

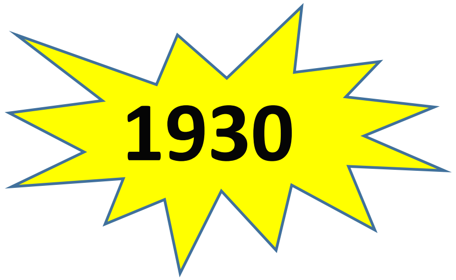
Spaces of Statistical Parameters.

By Harold Hotelling, Stanford University.

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

Oswald Veblen,
Advisor of Hotelling





Use Fisher information for the Riemannian metric tensor

Harold Hotelling
Econometrician

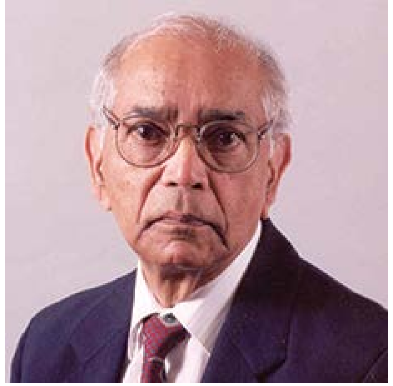
Information and the Accuracy Attainable in the Estimation of Statistical Parameters

C. Radhakrishna Rao

1. ♛ Cramer-Rao lower bound CRLB
2. ♛ Rao-Blackwellization
3. ♛ Fisher-Rao distance

Cramér-Rao Lower Bound and Information Geometry, 2013

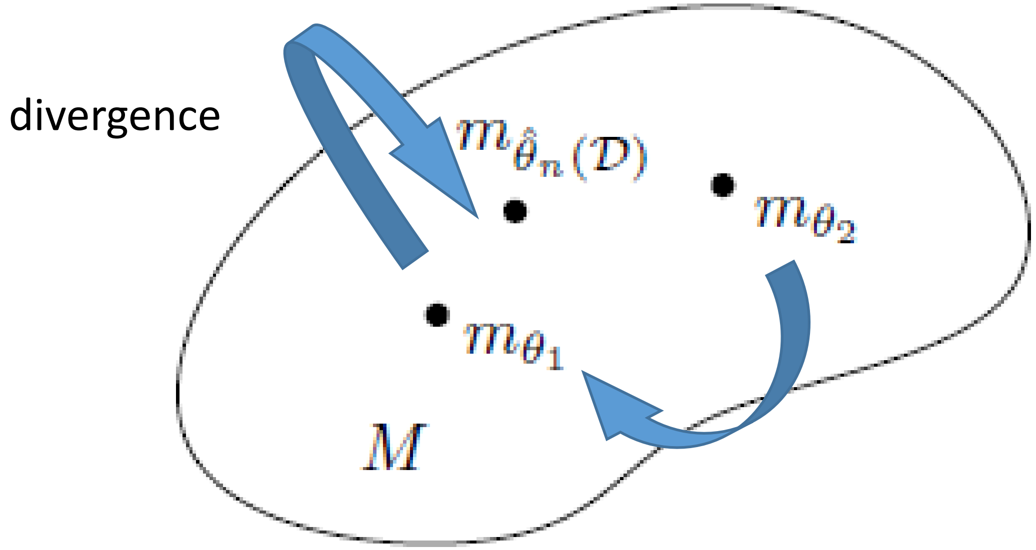
https://arxiv.org/abs/1301.3578



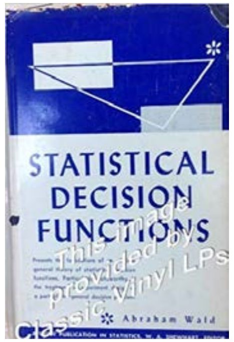
C. R. Rao
Statistician

Population space/parameter space

Example in statistical hypothesis testing: *estimate* from observations and then *classify* with respect to divergence to decide which hypothesis.

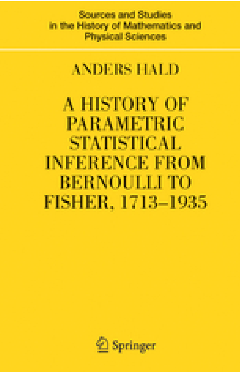


Geometry needed to build better Information Sciences:
- Deal with model and data (via *empirical distributions*)
- Deal with model and model



Wald's view: All statistical problems are decision problems…

Fisher information metric/matrix (FIM)



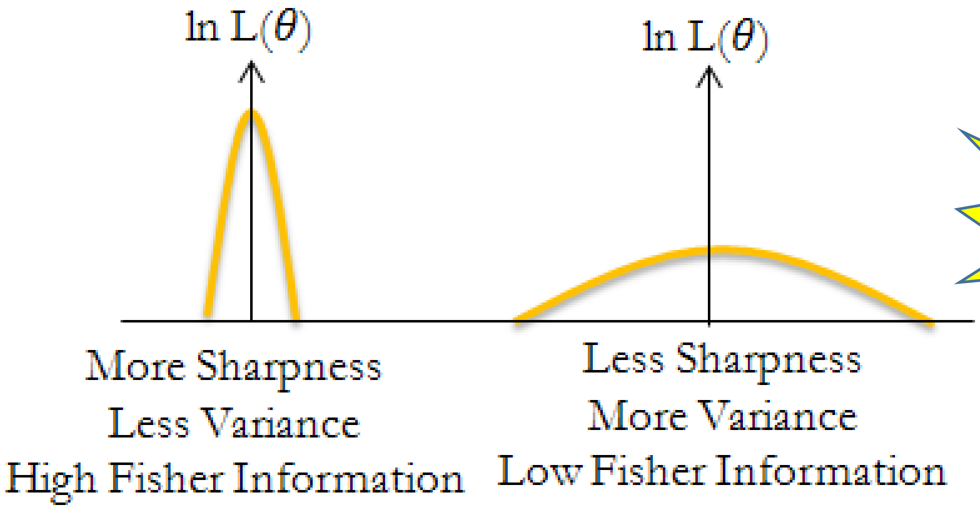
$$g(\xi)=E_\xi[\frac{\partial}{\partial\xi}\log(p_\xi)\ \frac{\partial}{\partial\xi}\log(p_\xi)]$$

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

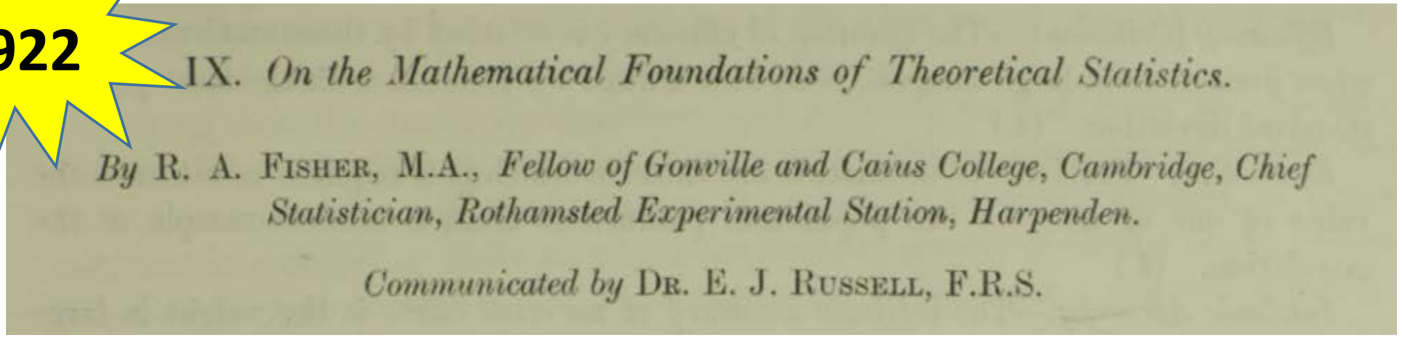
Sir Ronald Fisher

FIM is positive-semidefinite, positive-definite for <u>regular models</u>

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



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



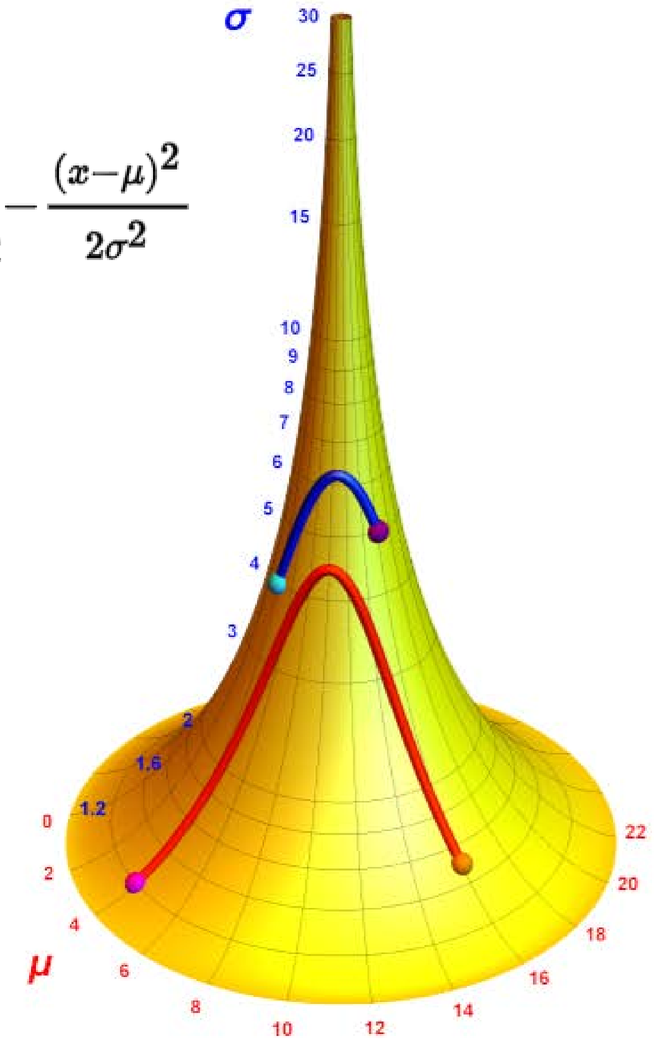
IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

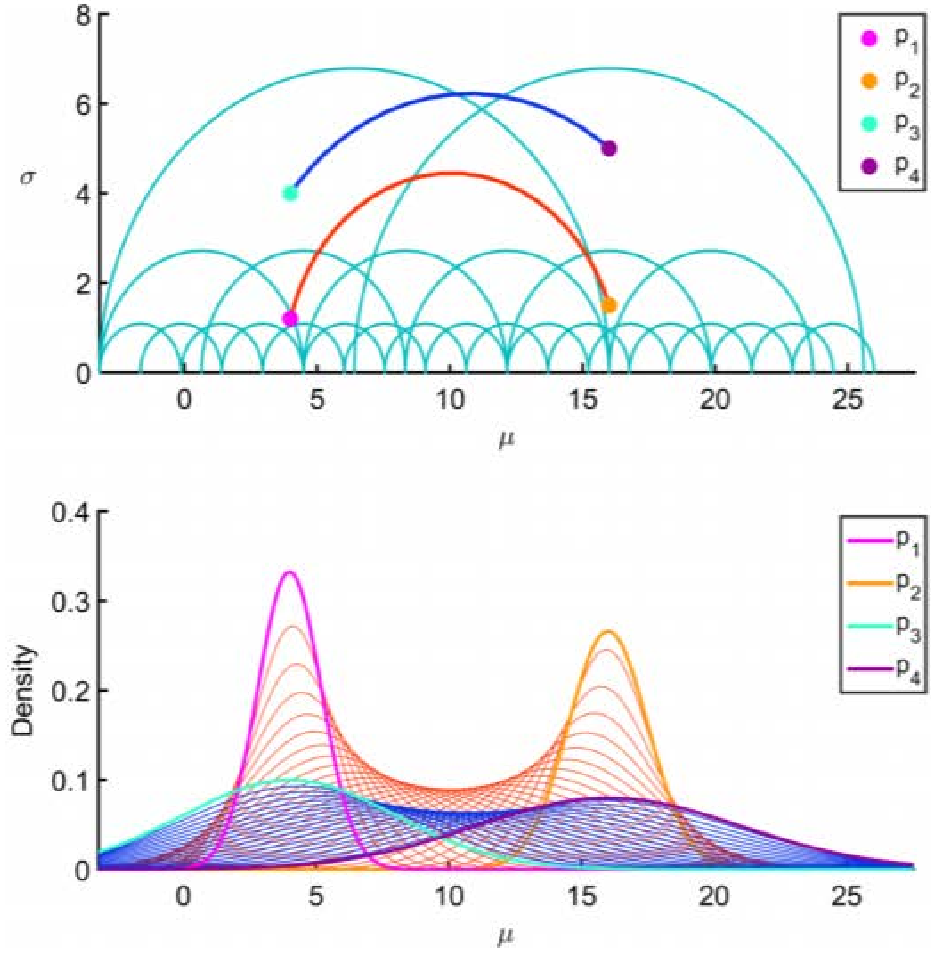
Communicated by Dr. E. J. Russell, F.R.S.

Geometry of normal distributions: hyperbolic

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



Pseudo-sphere
(negative curvature -1/2)



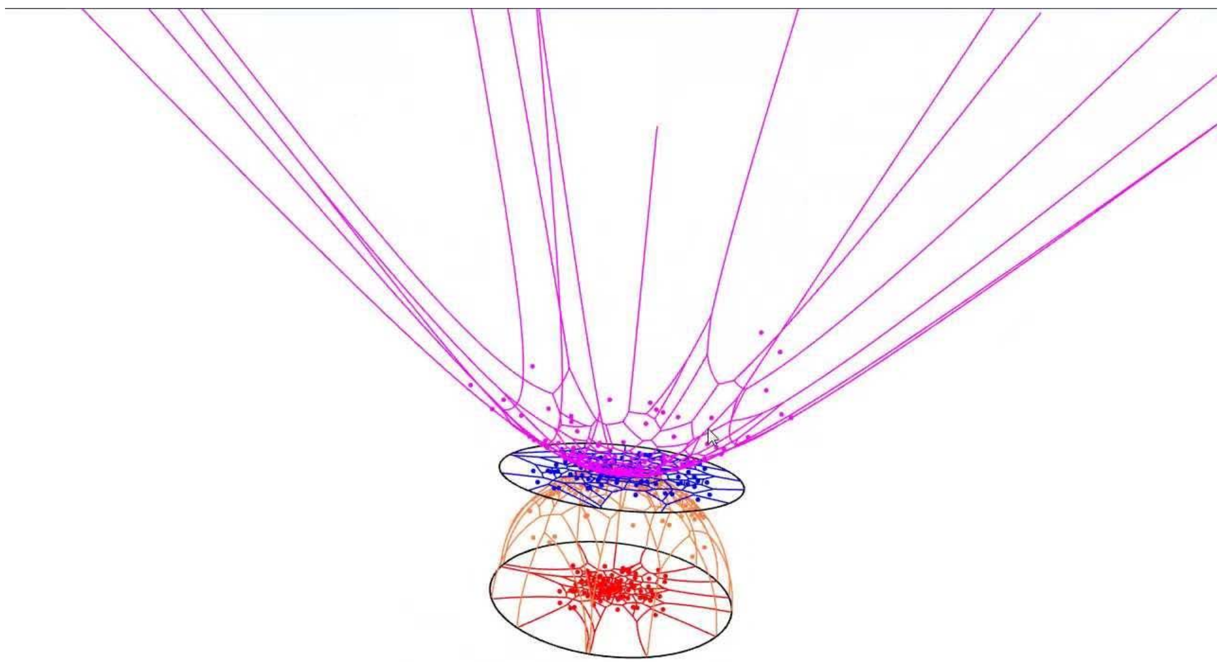
Pattern recognition in nuclear fusion data by means of geometric methods in probabilistic spaces, 2017

Hyperbolic geometry for location-scale families

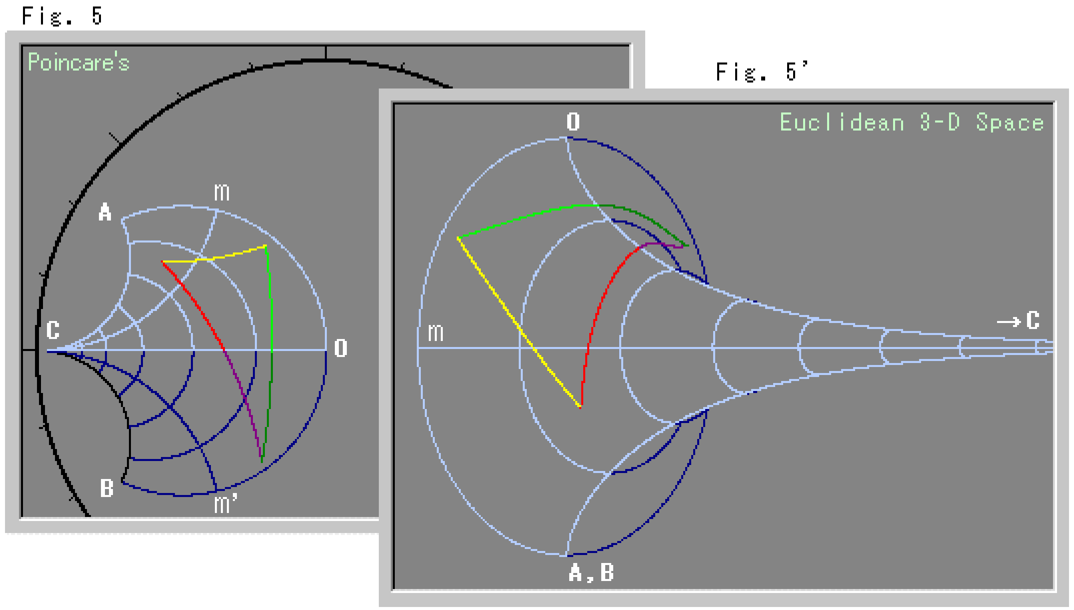
$$\mathcal{P} = \{\frac{1}{s_1} p\left(\frac{x-l_1}{s_1}\right) \ : \ (l_1, s_1) \in \mathbb{H}\}$$

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

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



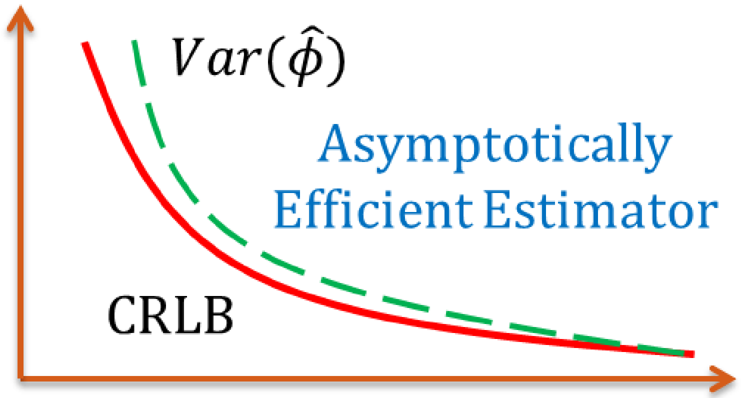
https://www.youtube.com/watch?v=i9IUzNxeH4o



Visualizing hyperbolic Voronoi diagrams. Symposium on Computational Geometry 2014

Cramer-Rao lower bound (CRLB + Frechet)

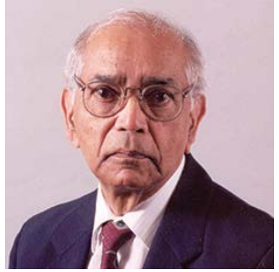
The variance of any unbiased estimator is lower bounded by the inverse of the Fisher information



Notion of efficiency!

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

René Maurice Fréchet (1878-1973)

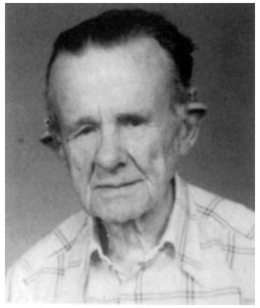


C. R. Rao

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

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

(here, positive-definite matrices, Lowner ordering)

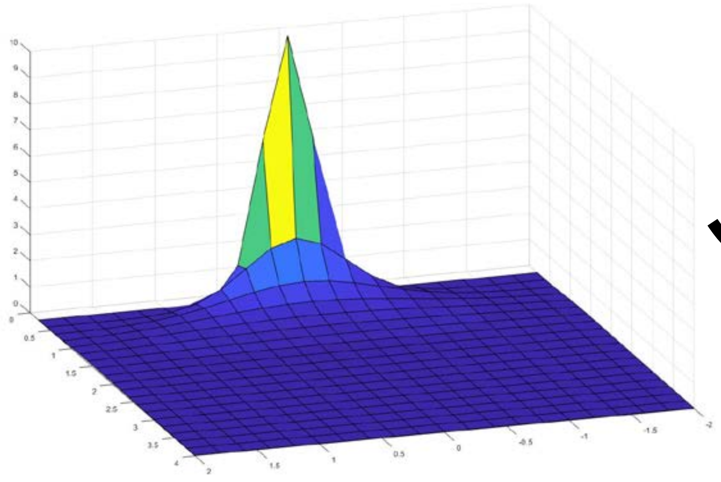


Harald Cramer

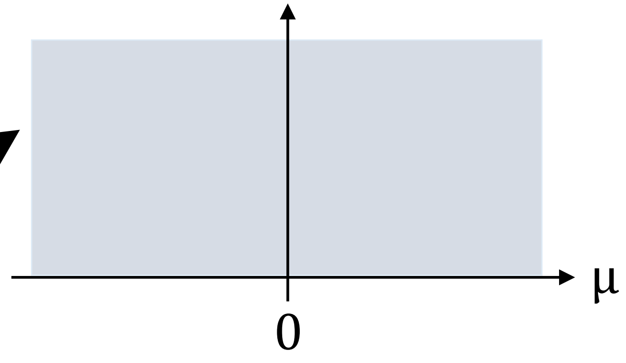
Examples of statistical models (regular/identifiable)

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

Negative curvature



$$O = \mathbb{R}$$
$$\mathbb{E} = \mathbb{R} \times \mathbb{R}^{+}$$
$$p(x,\mu,\sigma) \mapsto (\mu,\sigma)$$
$$\Phi$$

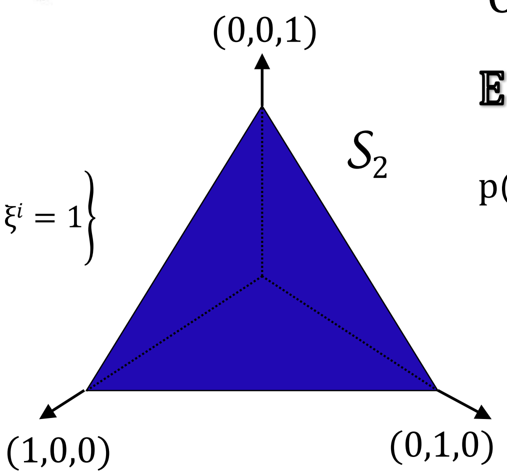


Upper half-plane

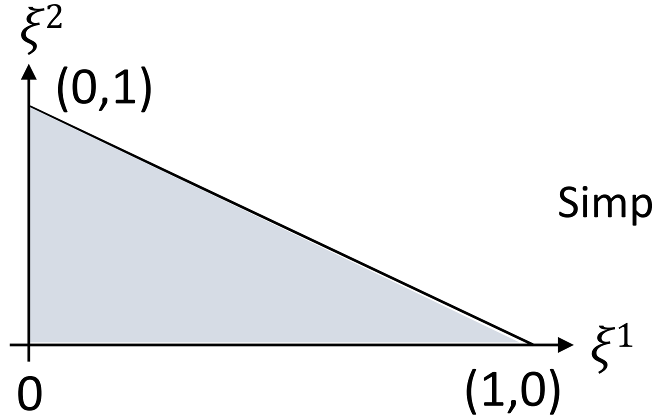
- S_n

Positive Curvature:
Multinomial
Multinoulli
Discrete dist.
Categorial dist.

$$\left\{\xi \in \mathbb{R}^{n+1} \,|\, \xi^i > 0 \sum \xi^i = 1\right\}$$



$$O = \{x_1, \ldots, xn\}$$
$$\mathbb{E} = \left\{\xi \in \mathbb{R}^n \,|\, \xi^i > 0 \sum \xi^i = 1\right\}$$
$$p(x,\xi^1,\ldots,\xi^{n+1}) \mapsto (\xi^1,\ldots,\xi^n)$$
$$\Phi$$

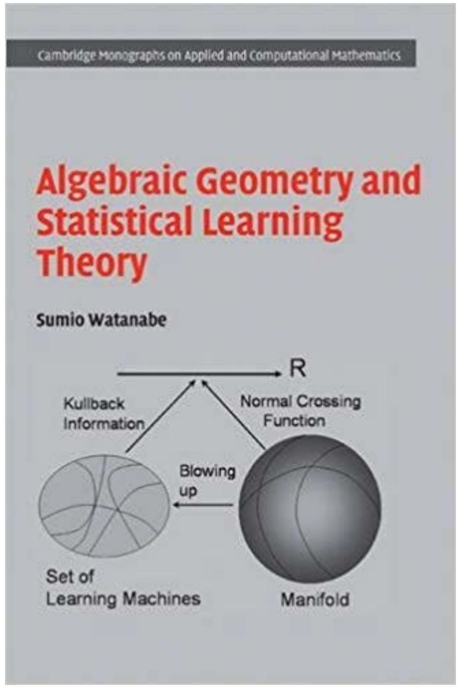


Simplex

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

Non-regular statistical models

- Not identifiable models happen often in practice…
- Usually, hierarchical models:
  - Gaussian mixture models (GMMs)
  - Multi-layer perceptrons (MLP)

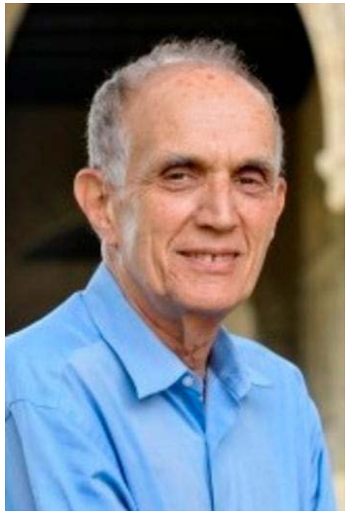


- Semi-definite matrix: Singular Semi-Riemannian manifolds
- Cramer-Rao lower bounds does not hold, need different theory for model selection (BIC, MDL), natural gradient and plateau in learning, etc.

Lightlike Neuromanifolds, Occam's Razor and Deep Learning, arXiv:1905.11027

Statistical curvature (1975)

Use of differential geometry to study the information loss in estimation



Bradley Efron

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

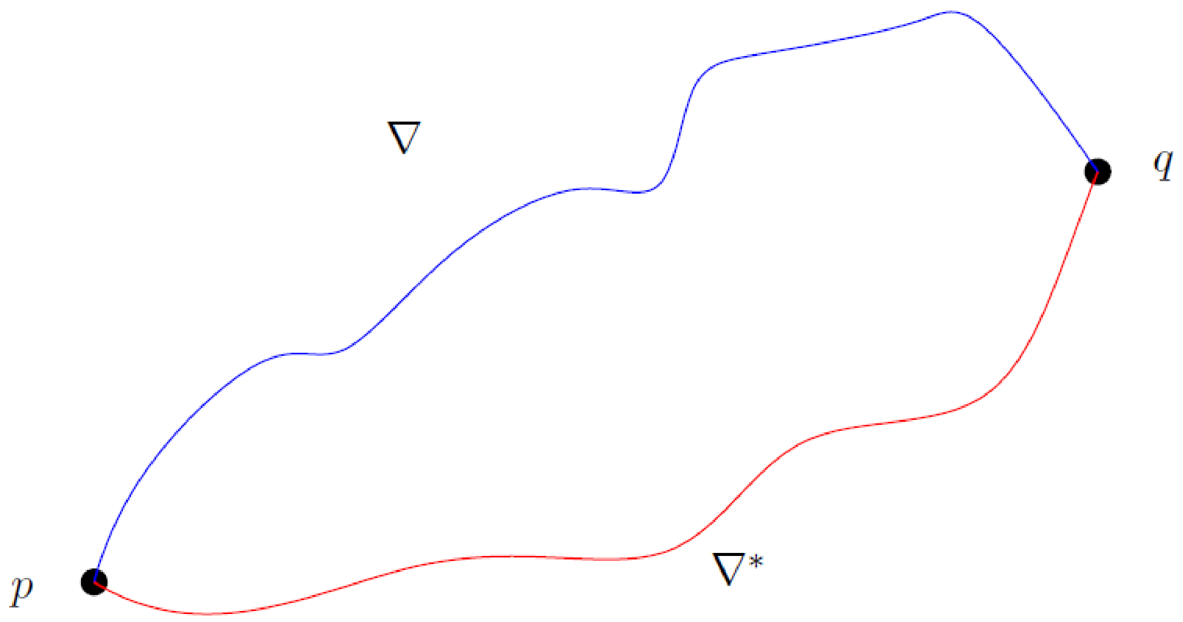
Statisticians know that one-parameter exponential families have very nice properties for estimation, testing, and other inference problems. Fundamentally this is because they can be considered to be "straight lines" through the space of all possible probability distributions on the sample space. We consider arbitrary one-parameter families $\mathscr{F}$ and try to quantify how nearly "exponential" they are. A quantity called "the statistical curvature of $\mathscr{F}$" 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.

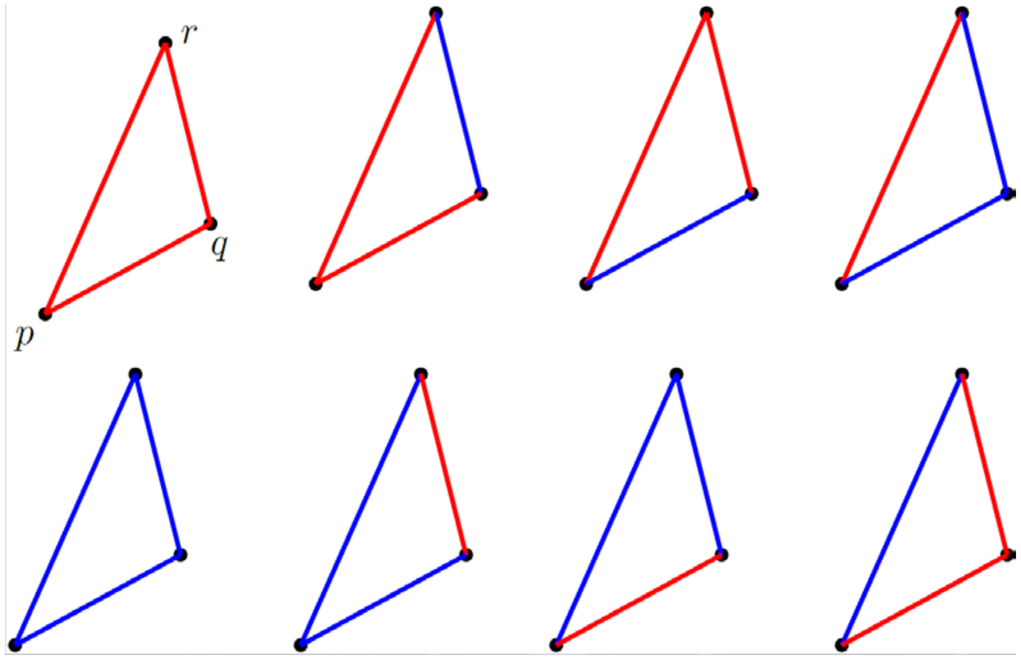
The Annals of Statistics
1978, Vol. 6, No. 2, 362–376

THE GEOMETRY OF EXPONENTIAL FAMILIES

BY BRADLEY EFRON
Stanford University

Dualistic structure of information geometry



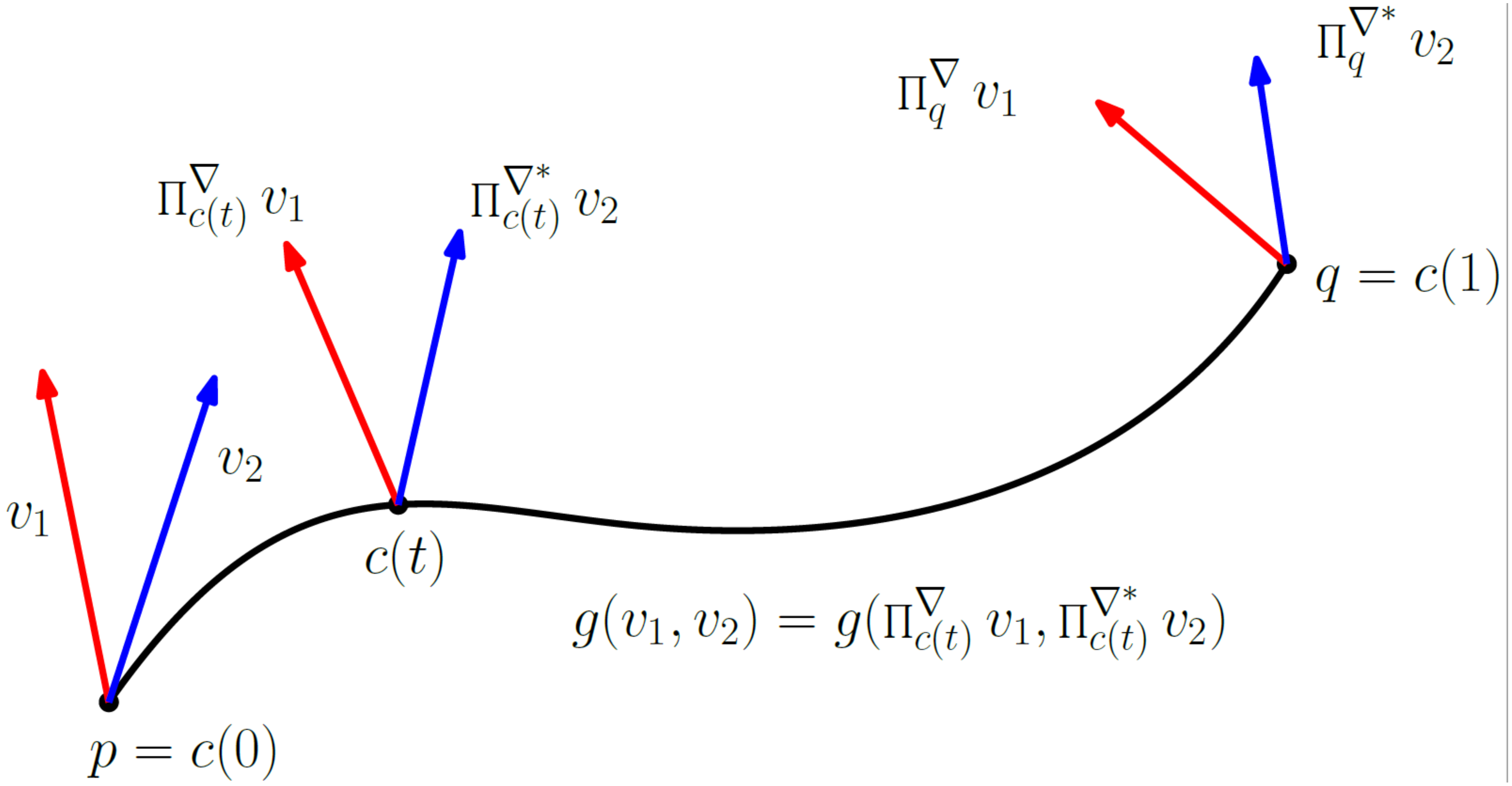


8 kinds of geodesic triangles

- Two conjugate torsion-free affine connections coupled with the metric
- Dual parallel transport is metric-compatible

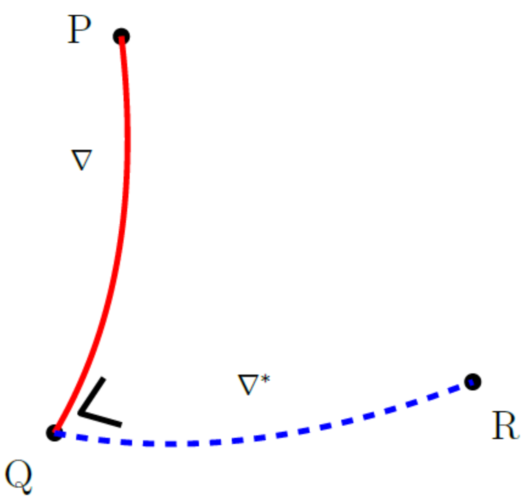
There is not necessarily a distance, 2^k types of k-gons (eg, 8 triangles)

Dual parallel transport is metric-compatible



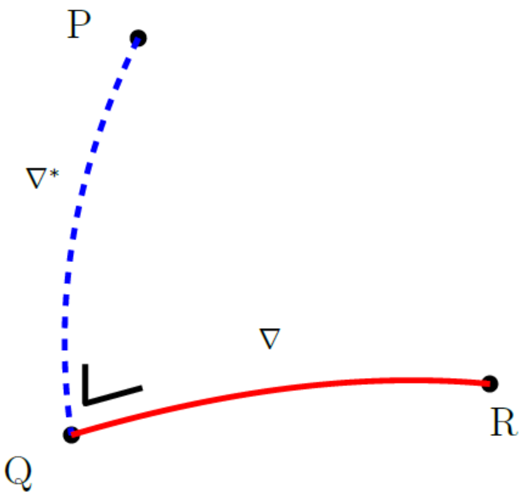
Dually flat space: Pythagoras' theorem

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



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

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



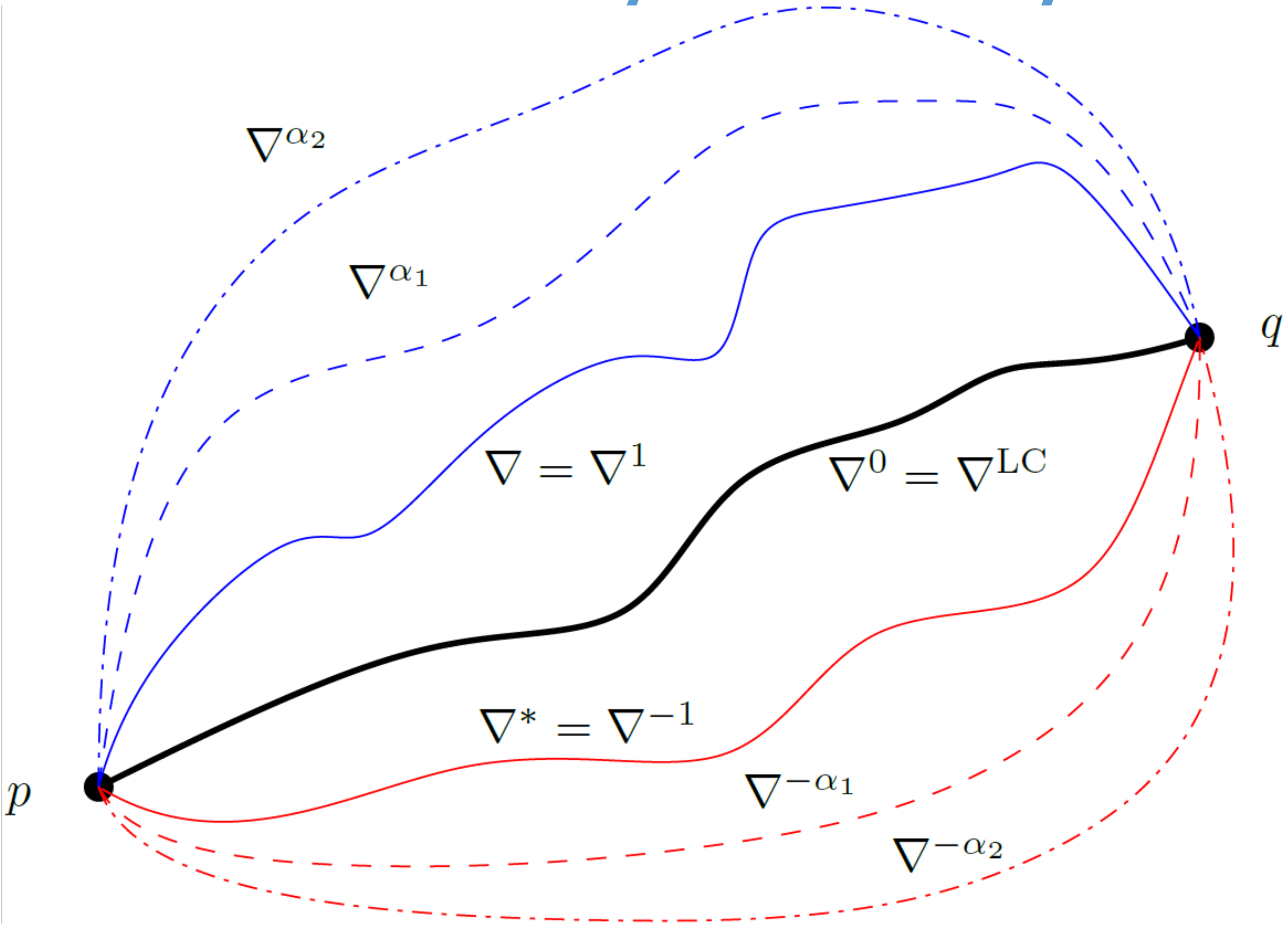
$D^*(P:R) = D^*(P:Q) + D^*(Q:R)$

Bregman manifold
induced by a convex function

Two (affine) coordinate systems coupled by Legendre-Fenchel transformation
Two dually flat connections with respect to the metric tensor
Canonical distance = Bregman divergence induced by convex generator F
Bregman manifold (a type of 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?

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

- Let $(\mathrm{M}, \mathrm{g}, \nabla, \nabla^*)$ be a dualistic structure: A dual pair of connections coupled to the metric so that dual parallel transport is metric-compatible

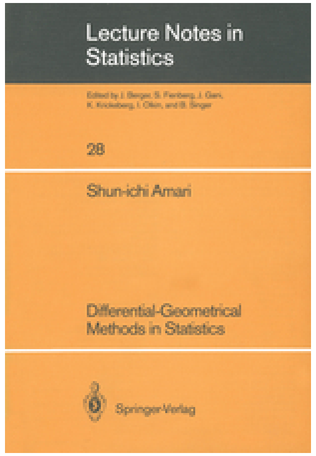
- We can build a **1-family of dualistic structures** $(\mathrm{M}, \mathrm{g}, \nabla^{-\alpha}, \nabla^{\alpha})$

so that $\frac{\nabla^{-\alpha}+\nabla^{\alpha}}{2}=\nabla^0 = \nabla^{\mathrm{LC}}$

- <u>**No distance associated with the dualistic structure.**</u>

In particular, when α =0, $(\mathrm{M}, \mathrm{g}, \nabla^0, \nabla^0) = (\mathrm{M}, \mathrm{g})$ the Riemannian geometry.
Thus information geometry generalizes (Fisher-Rao) Riemannian geometry

Amari's expected α-geometry



- Given a parametric family of distributions, consider the Fisher information matrix and a family of connections: **α connections**
- **Exponential e-mixture connection and m-mixture connection**

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

- No associated distance in the alpha-expected geometry

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

https://arxiv.org/abs/1808.08271

How to get initial dual expected connections?

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

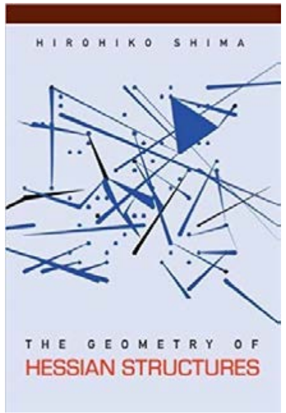
Log-likelihood $\ell(p_\xi)(x) = \ln p_\xi(x)$

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

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

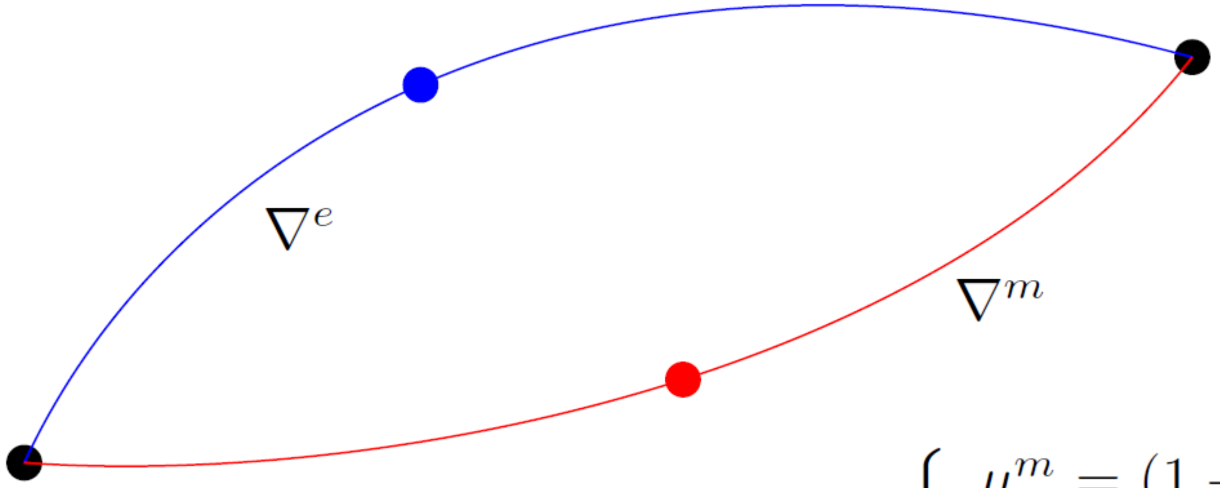
Dual connections with respect to the Fisher information (Riemannian) metric

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



$$(p_1 p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \frac{(1-\alpha)\mu_1 v_2 + \alpha \mu_2 v_1}{(1-\alpha) v_2 + \alpha v_1} \\ v^e_\alpha = \frac{v_1 v_2}{(1-\alpha) v_2 + \alpha v_1} \end{cases}$$

$p_2 = (\mu_2, v_2 = \sigma_2^2)$



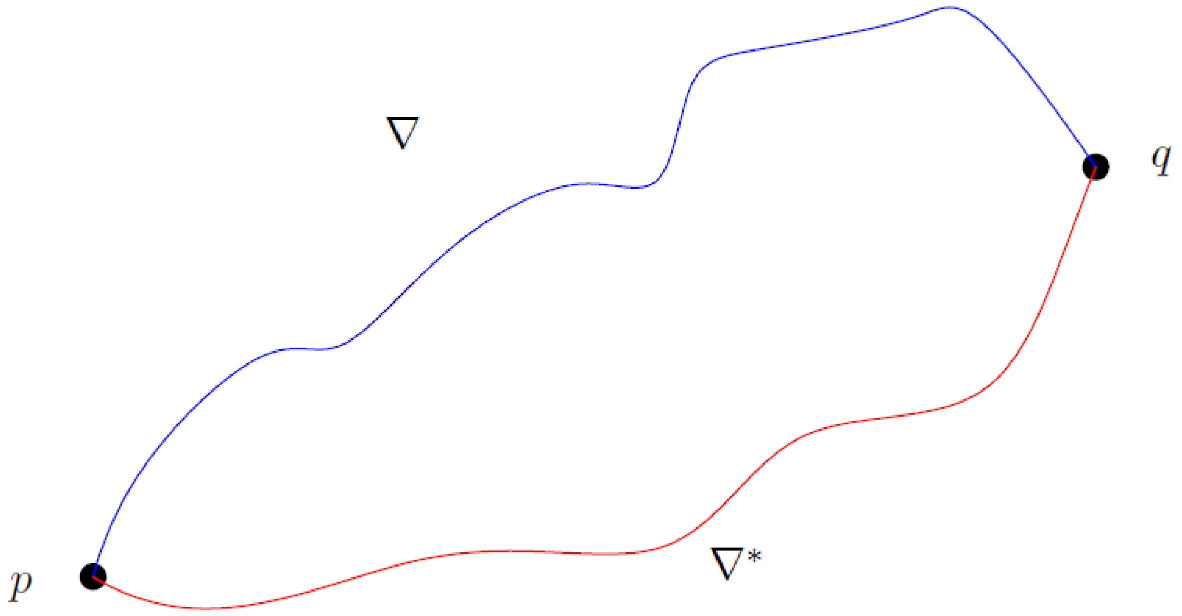
$p_1 = (\mu_1, v_1 = \sigma_1^2)$

$$(p_1 p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha \mu_2 \\ v^m_\alpha = (1-\alpha) v_1 + \alpha v_2 - \alpha(1-\alpha)(\mu_1 - \mu_2)^2 \end{cases}$$

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

Dualistic structure of the Gaussian manifold



∇: e-connection

∇^*:m-connection

$$(p_1p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ v^m_\alpha = (1-\alpha)v_1 + \alpha v_2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2 \end{cases}$$

$$(p_1p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \frac{(1-\alpha)\mu_1 v_2 + \alpha\mu_2 v_1}{(1-\alpha)v_2 + \alpha v_1} \\ v^e_\alpha = \frac{v_1 v_2}{(1-\alpha)v_2 + \alpha v_1} \end{cases}$$

$$(p_1p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ \Sigma^m_\alpha = \bar{\Sigma}_\alpha + (1-\alpha)\mu_1\mu_1^\top - \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha{}^\top \end{cases}$$

$$(p_1p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \Sigma^e_\alpha((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2) \\ \Sigma^e_\alpha = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1} \end{cases}$$

Dual connections from *any* divergence! $(M, {}_D g, {}_D\nabla, {}_D\nabla^*)$

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

- a tensor metric g: $g_{ij}(p_\xi) = \frac{\partial}{\partial \xi^i_1} \frac{\partial}{\partial \xi^j_2} D(p_{\xi_1}, p_{\xi_2})|_{\xi_1=\xi_2=\xi}$

- a torsion-less affine connection ∇:

$$\Gamma_{ijk}(p_\xi) = -\frac{\partial}{\partial \xi^i_1} \frac{\partial}{\partial \xi^j_2} \frac{\partial}{\partial \xi^k_2} D(p_{\xi_1}, p_{\xi_2})|_{\xi_1=\xi_2=\xi}$$

Dual divergences
and dual connections

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

Symmetric divergences yields the same connection:

The Levi-Civita connection

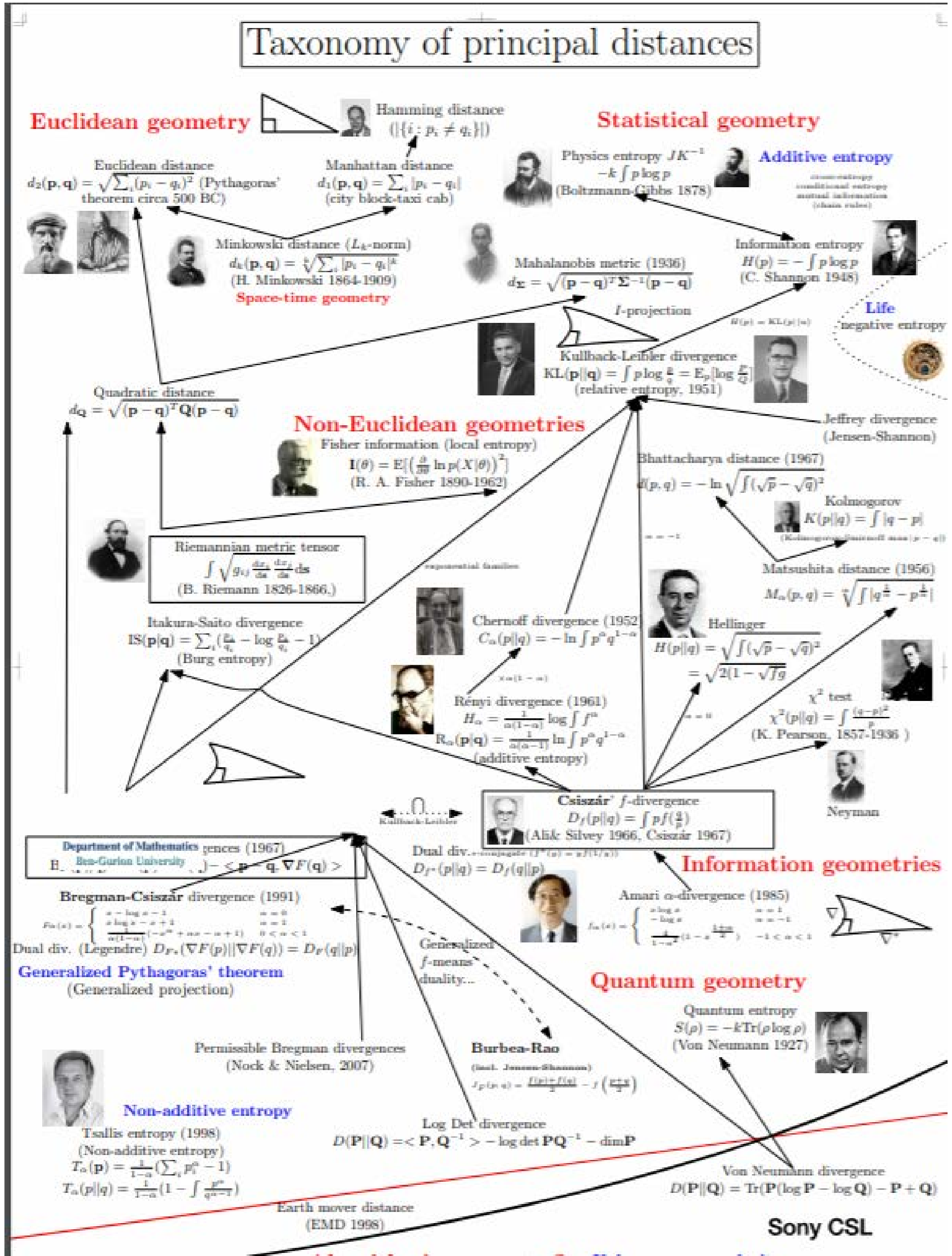
Many distances/divergences in information sciences

Divergence= discrepancy, dissimilarity, deviance between two probability distributions

Also nowadays, smooth parametric dissimilarities (contrast function)

Distance is often thought as a metric distance:

$$
\begin{aligned}
&(a) \quad d(p,q) > 0 \text{ if } p \neq q;\ d(p,p) = 0;\\
&(b) \quad d(p,q) = d(q,p);\\
&(c) \quad d(p,q) \leq d(p,r) + d(r,q),
\end{aligned}
$$



Divergences: Statistical distances

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

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

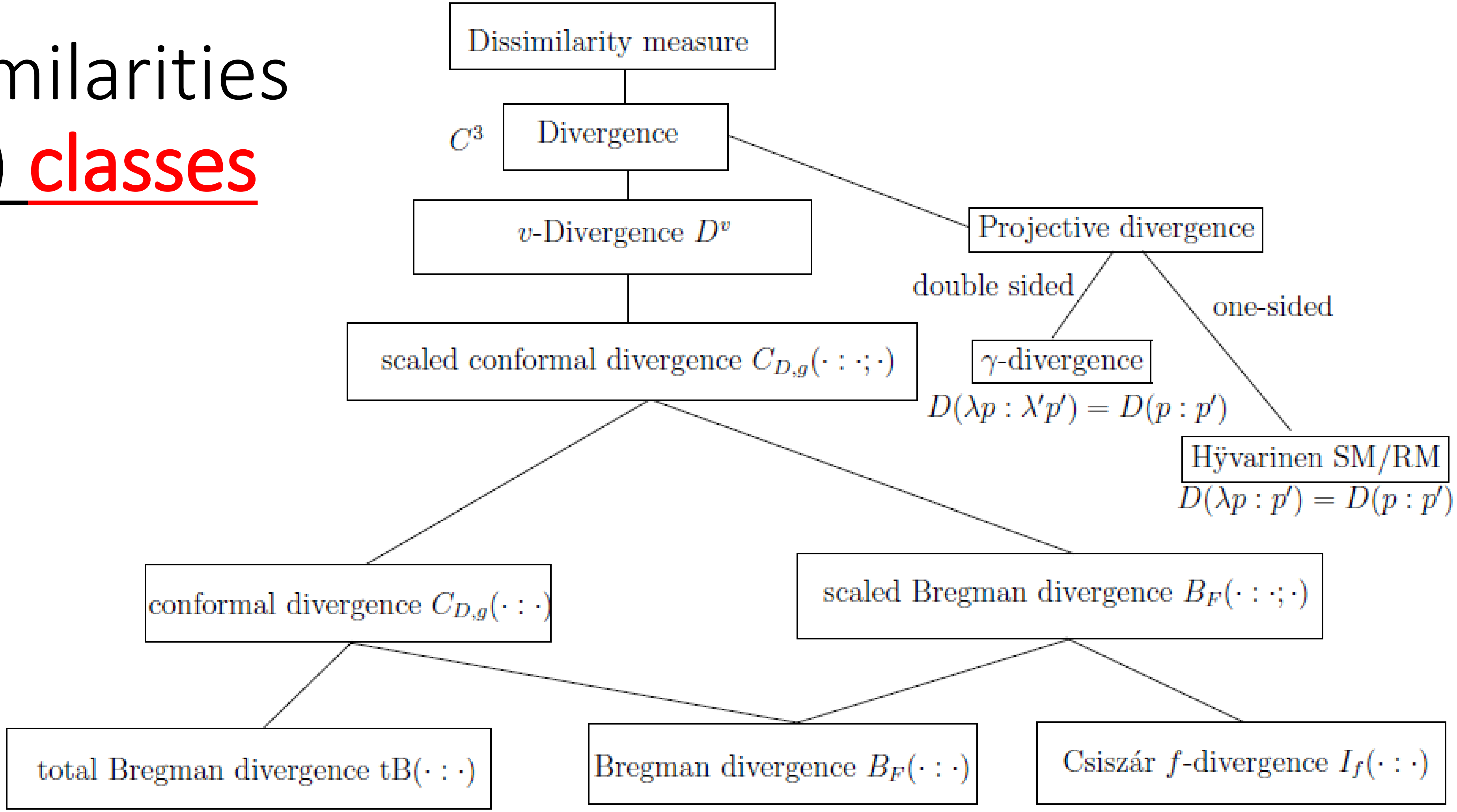
$$D_{KL}(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- Can be extended to f-divergences

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

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

Organize dissimilarities in (exhaustive) classes



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

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

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

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

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

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

Invariant divergence = f-divergences

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

information monotonicity

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

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p

coarse graining

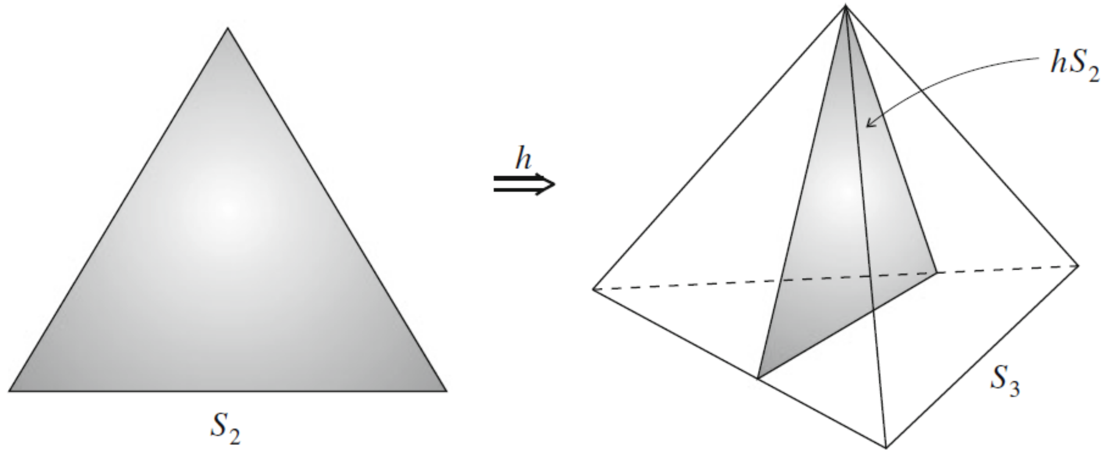
$p_1 + p_2$	$p_3 + p_4 + p_5$	p_6	$p_7 + p_8$	$p_{\mathcal{A}}$

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

f-divergences are invariant by diffeormorphisms of the sample space

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

Statistical invariance



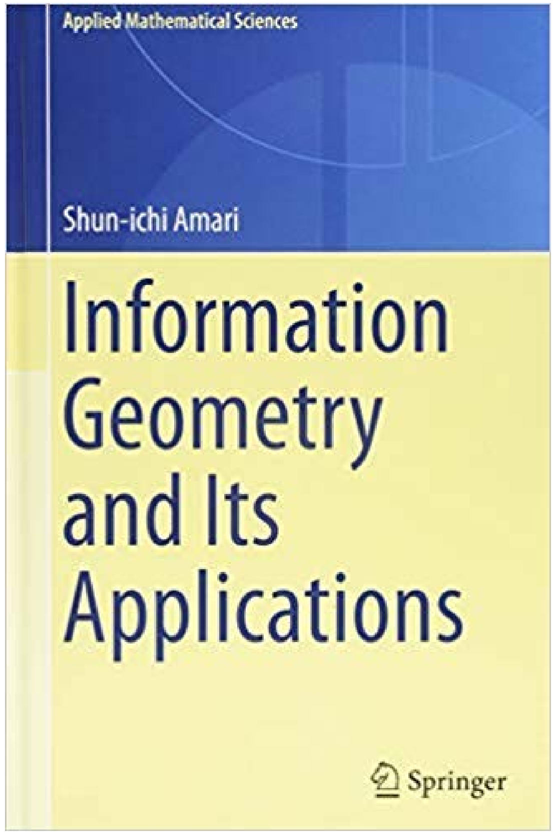
- **Fisher-Rao distance** is independent of parameterization (but FIM is covariant!)
  Same Fisher-Rao distance for parameterizations {N(μ, σ)} or {N(μ, σ2)}

- 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

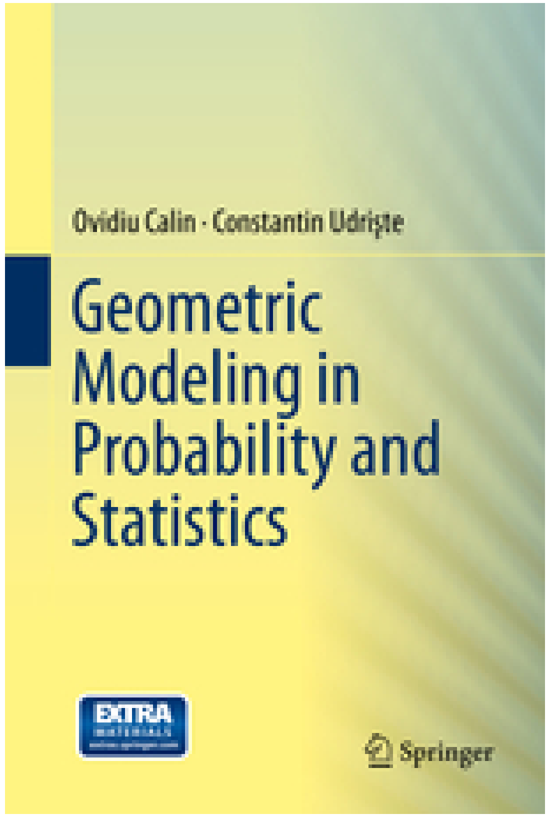
- Dual connections induced by any f-divergence yield expected alpha-connections

Recommended textbooks + overview survey



2016

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



2014

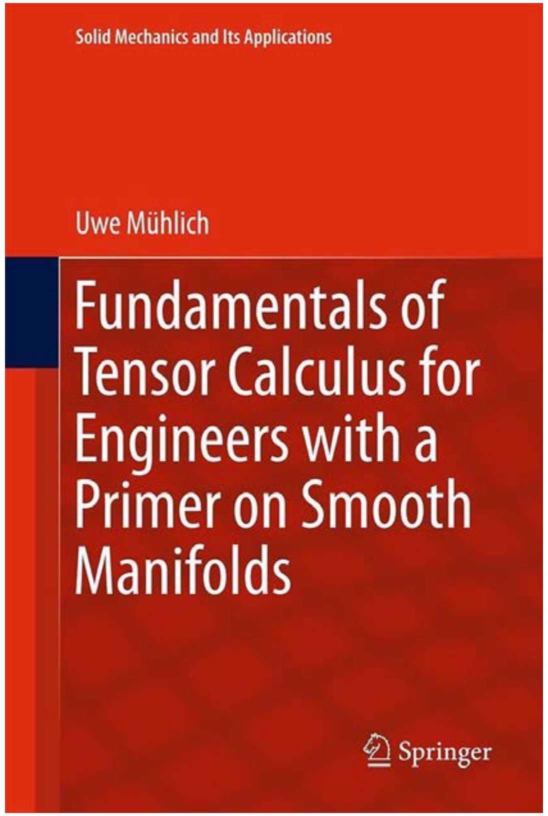
More details on differential geometry with exercices

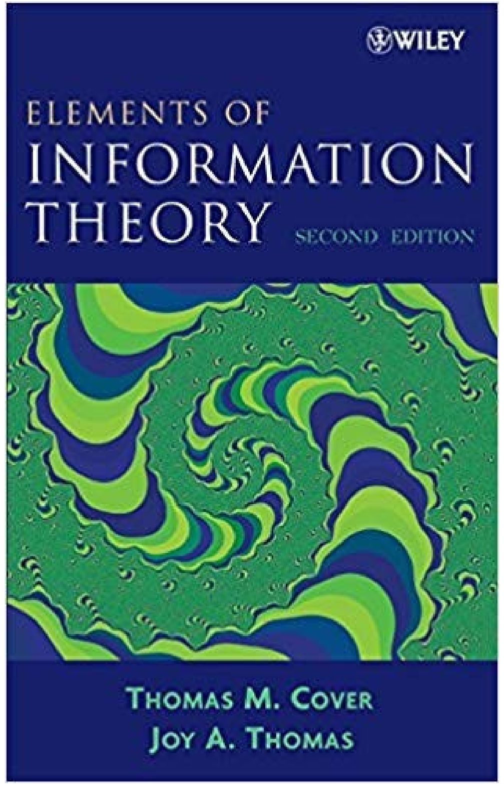
An elementary introduction to information geometry

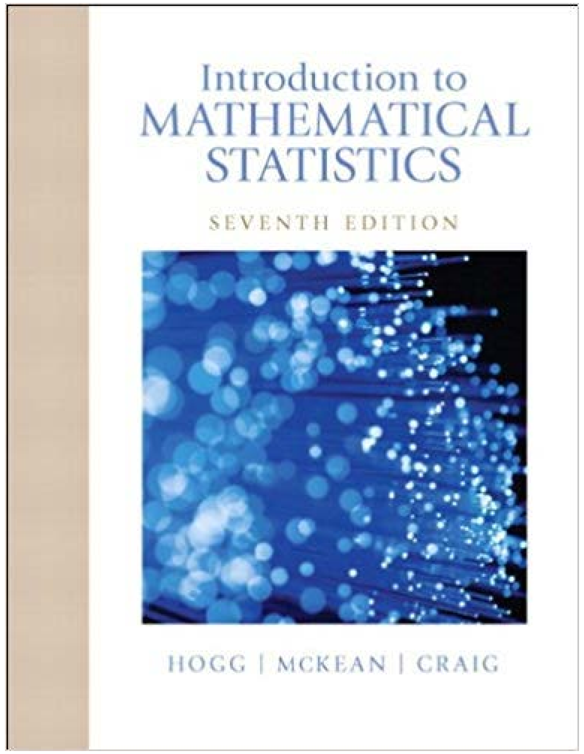
https://arxiv.org/abs/1808.08271

Prerequisite:
Information sciences + Differential geometry

- Tensors + Manifolds
- Statistics + Information theory



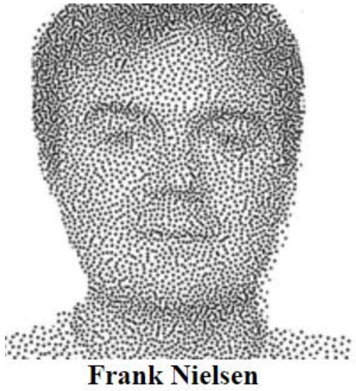


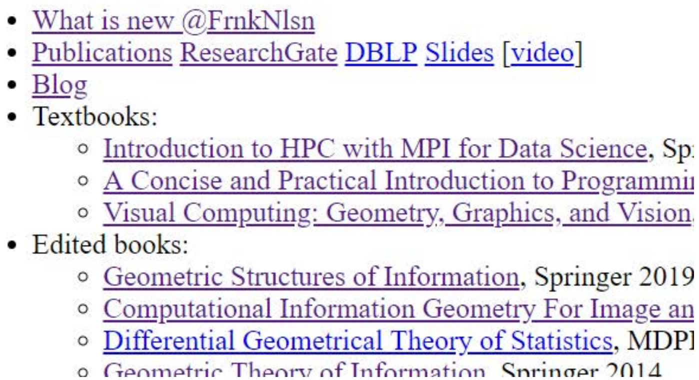


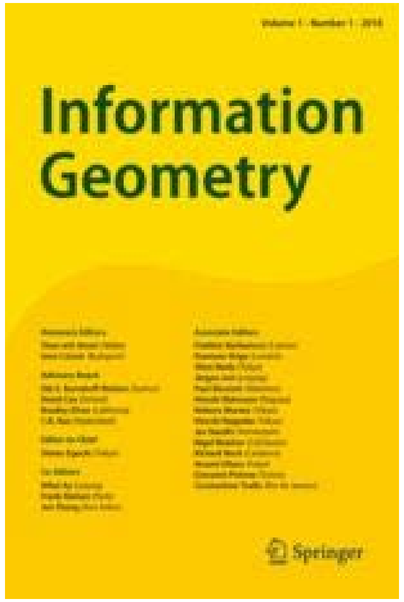
Outline of the lectures:

- Introduction and overview of the dualistic structures (these slides)
- Background:
  - Probability and statistics
  - Information theory
  - Differential geometry
  - Distances
- Information-geometric manifolds
  - Fisher-Rao Riemannian manifolds
  - Manifolds with dual connections coupled to the metric
  - Bregman manifolds
  - Geometry of mixture families with applications
- Information geometry in action:
  - Natural gradient descent methods and deep learning
  - Clustering
  - Bayesian hypothesis testing
- Advanced topics, limitations and perspectives

Thank you.

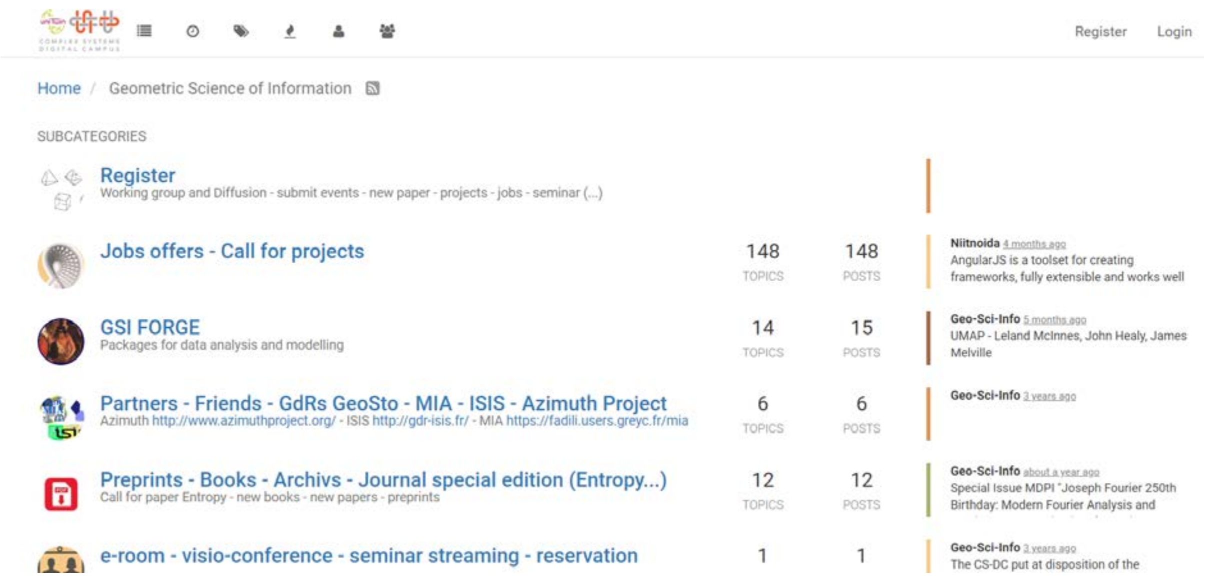


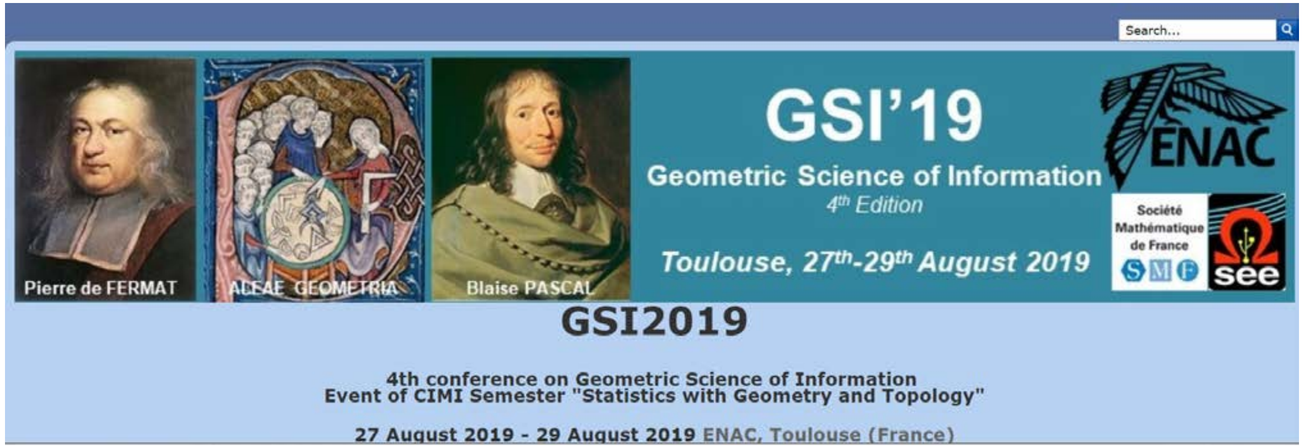

- What is new @FrnkNlsn
- Publications ResearchGate DBLP Slides [video]
- Blog
- Textbooks:
  - Introduction to HPC with MPI for Data Science, Sp
  - A Concise and Practical Introduction to Programmi
  - Visual Computing: Geometry, Graphics, and Vision
- Edited books:
  - Geometric Structures of Information, Springer 2019
  - Computational Information Geometry For Image an
  - Differential Geometrical Theory of Statistics, MDPI
  - Geometric Theory of Information, Springer 2014



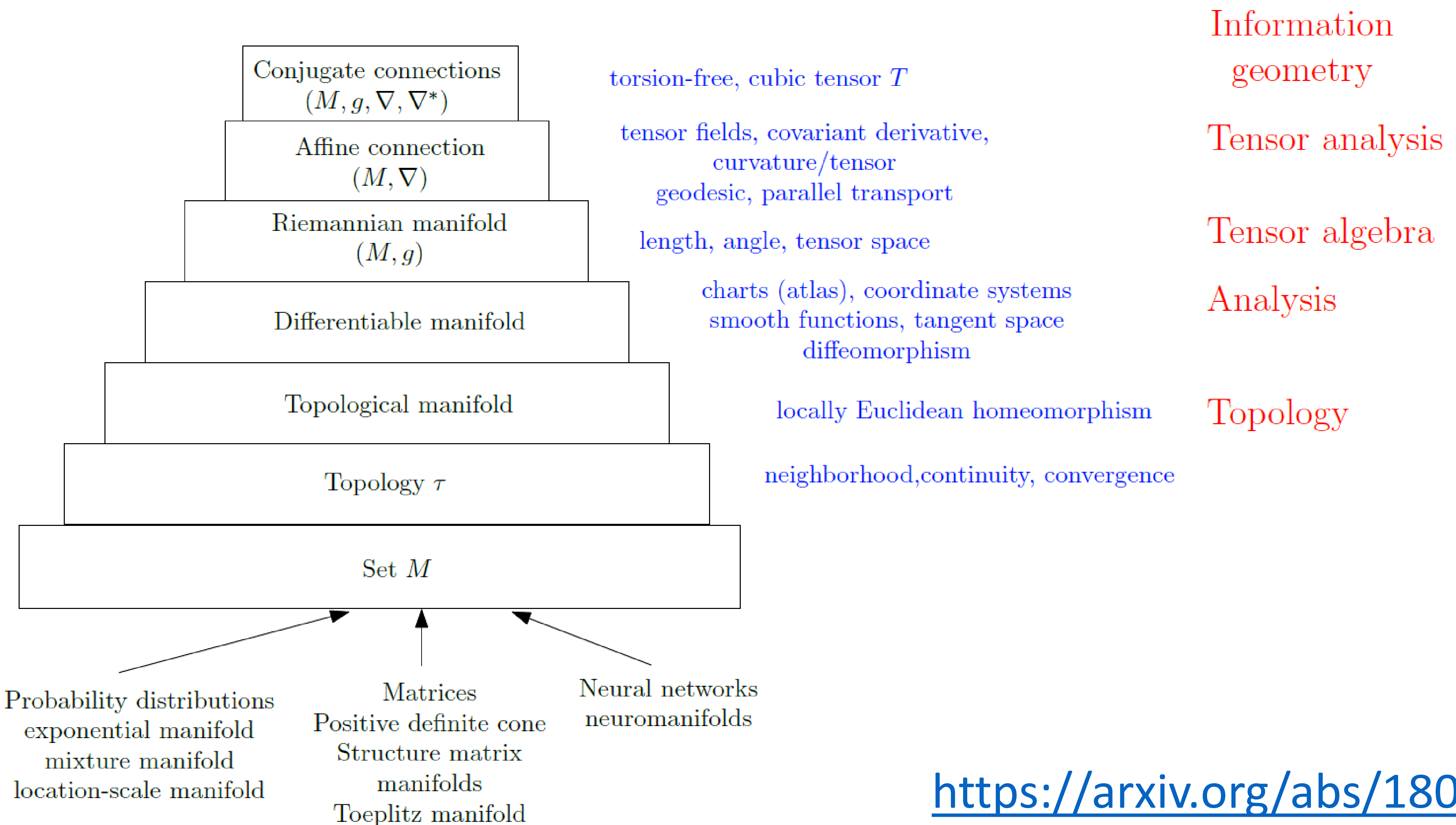
https://franknielsen.github.io/

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





Genesis of an information-geometric structure



https://arxiv.org/abs/1808.08271

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)

Applications

WHAT IS...

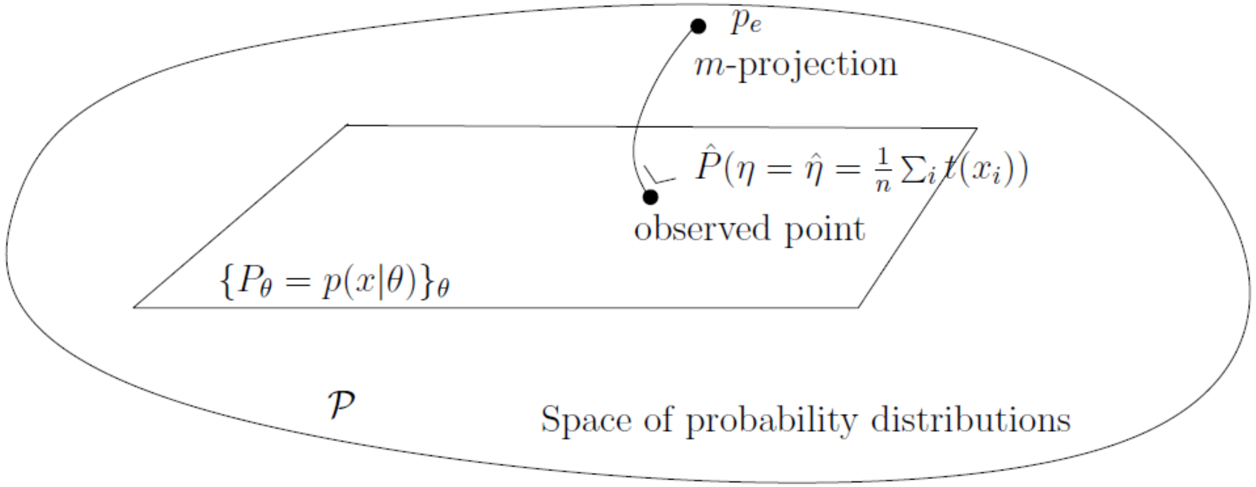
an Information Projection?

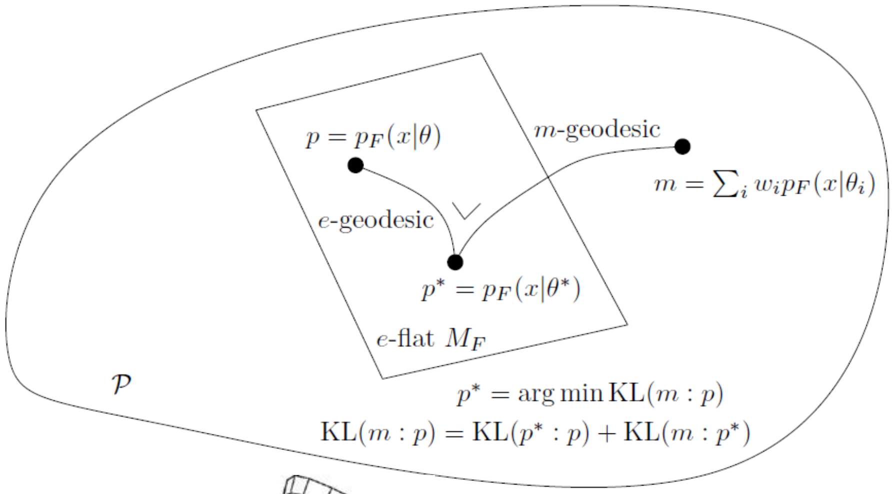
Frank Nielsen

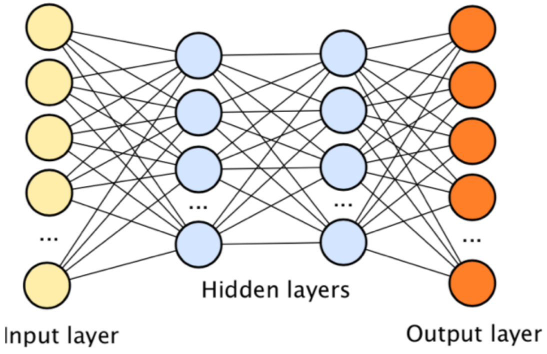
Communicated by Cesar E. Silva

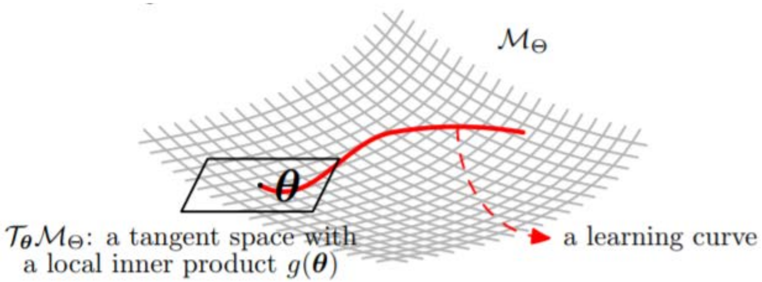
Empirical distribution : $p_e(X) = \frac{1}{n}\sum_{i=1}^{n} \delta(X - X(i))$

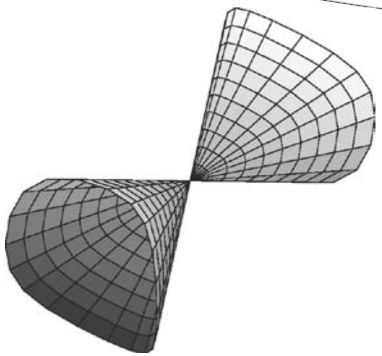
MLE $=$ m-projection from p_e to the model submanifold











Singularities in neuromanifolds

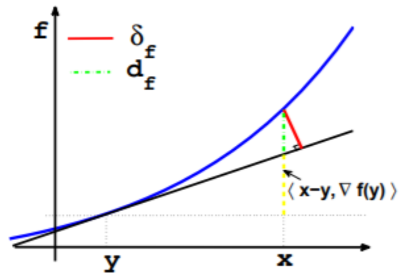
Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

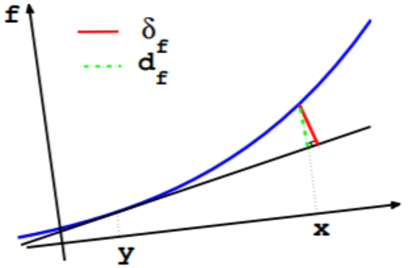
Definition *The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,*

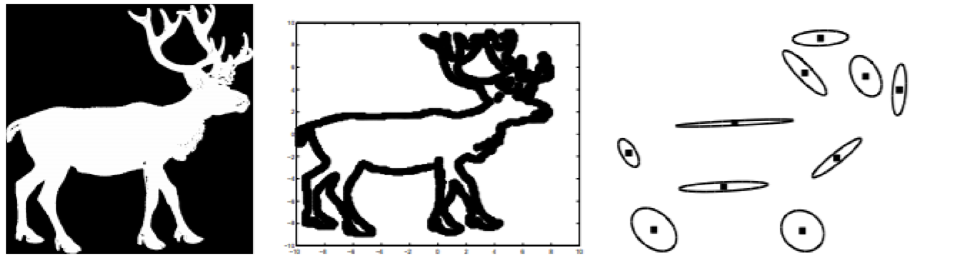
$$\delta_f(x,y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}},$$

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

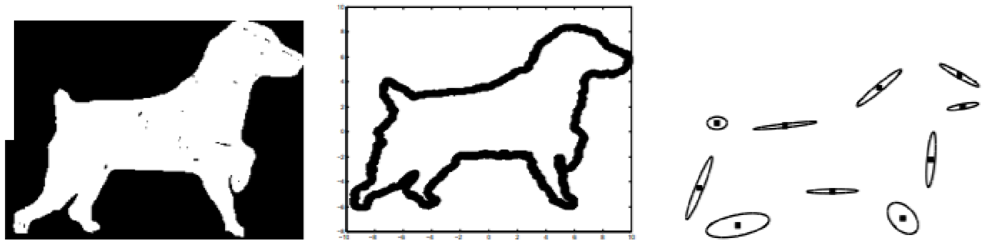
X	$f(x)$	$\delta_f(x,y)$	t-center	ℓ_1-norm BD center	Remark
$\mathbb{R}$	x^2	$\frac{(x-y)^2}{\sqrt{1+4y^2}}$	$\sum_i w_i x_i$	$\sum_i x_i$	total square loss (tSL)
$\mathbb{R} - \mathbb{R}_-$	$x \log x$	$\frac{x \log \frac{x}{y} + \bar{x} \log \frac{\bar{x}}{\bar{y}}}{\sqrt{1+y(1+\log y)^2+\bar{y}(1+\log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







(m) (n) (o)



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

IEEE TPAMI 34, 2012

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,*

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

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

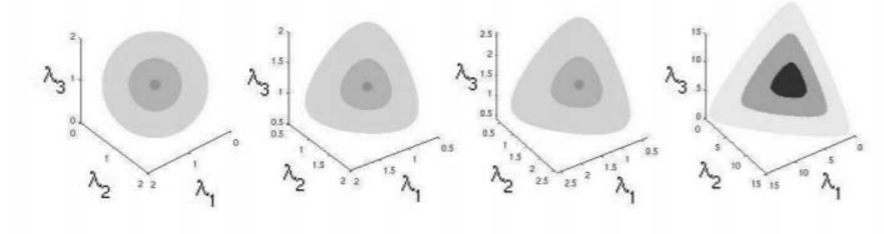
$$tKL(P,Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1 + \int (1 + \log q)^2 q dx}}$$

$$= \frac{\log(\det(P^{-1}Q)) + tr(Q^{-1}P) - n}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{n(1+\log 2\pi)}{2}\log(\det Q)}}$$

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

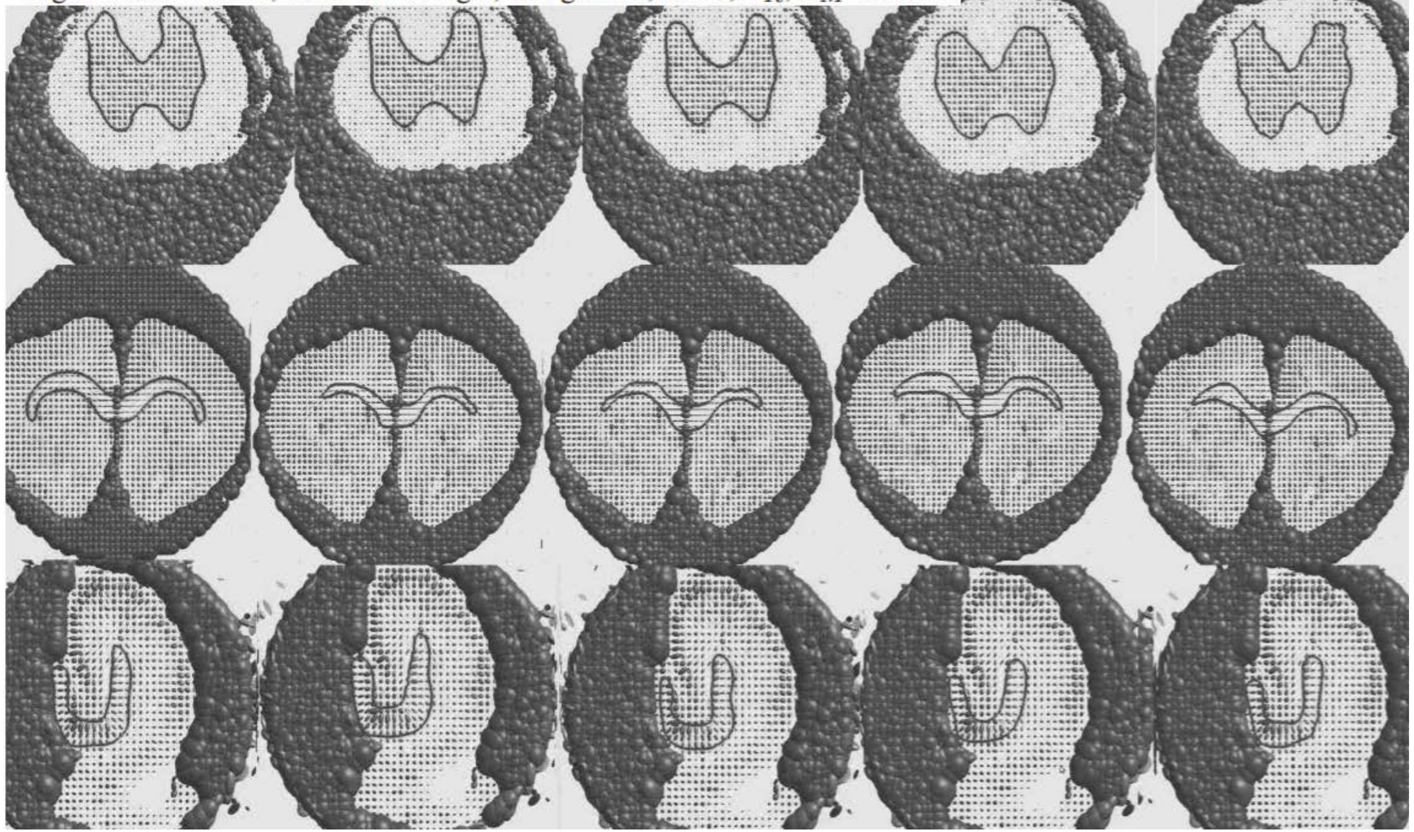
$$tSL(P,Q) = \frac{\int (p-q)^2 dx}{\sqrt{1 + \int (2q)^2 q dx}} =$$

$$\frac{1/\sqrt{\det(2P)} + 1/\sqrt{\det(2Q)} - 2/\sqrt{\det(P+Q)}}{(2\pi)^n + 4\sqrt{(2\pi)^n}/\sqrt{\det(3Q)}}$$



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)

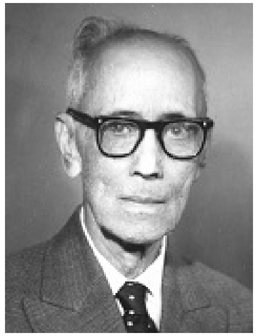
- Rabindra Nath Sen (1896-1974), “Senian geometry”

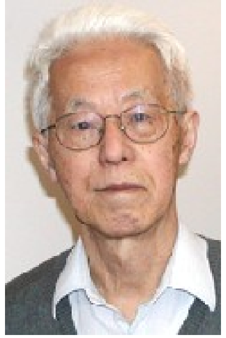
- Nomizu and Sasaki’s Affine differential geometry (geometry of immersions)

- Information geometry (Chentsov’s category approach and Amari)

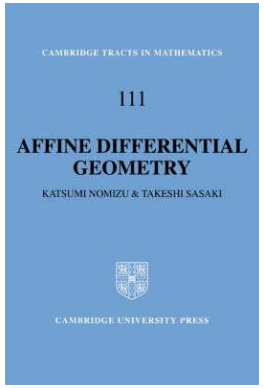
- Wong’s optimal transport and c-divergences

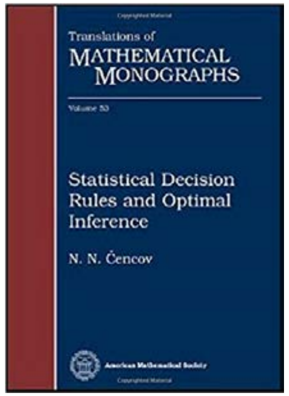

Norden


Sen


Nomizu


Amari





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)

Acknowledgements

- My collaborators (incl. Jean-Daniel Boissonnat, Gaetan Hadjeres, Richard Nock, Ke Sun, Olivier Schwander, and all my co-authors!)

- Images of these slides were mostly courtesy of the Internet.

- Some figures were drawn in PowerPoint by Joffrey Poitevin

Background for Information Geometry

- **Probability and statistics**
- **Information theory**
- **Elements of differential geometry**
- **Distances, divergences and entropies**

Frank Nielsen

Sony CSL

An elementary introduction to information geometry

https://arxiv.org/abs/1808.08271

Probability and statistics

Background

Frank Nielsen

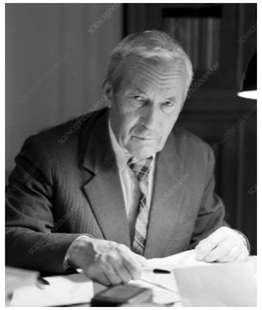
Sony CSL

Outline

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

Jacob Bernoulli

Pierre de Fermat



Kolmogorov

Sir Ronald Fisher

Barndorff-Nielsen

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** and probability of events:

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

- **Probability mass function** (pmf)

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

- **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**
- **Variance**

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

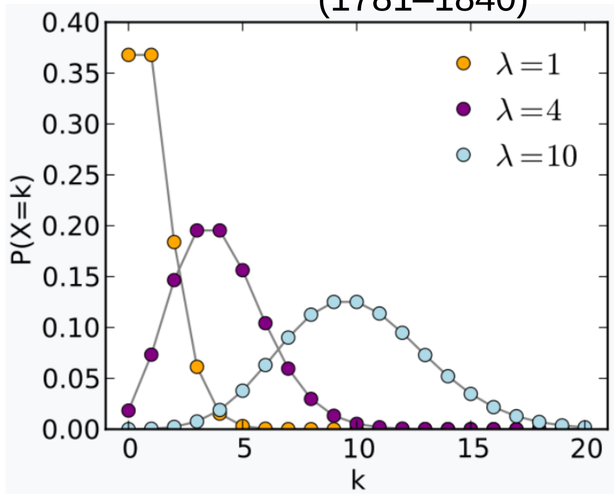
$$\mathrm{Var}[X]=\mathrm{E}[X^2]-\mathrm{E}[X]^2=p-p^2=p(1-p)=pq$$

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

Siméon Denis Poisson
(1781–1840)

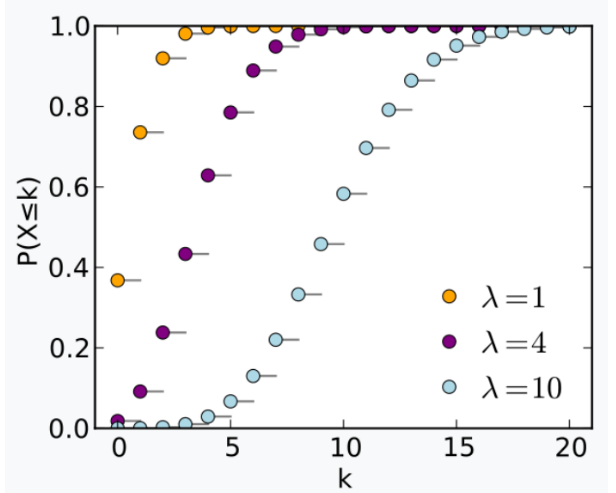
- Poisson distribution with support 0, 1, 2, 3, …
- Probability mass function:

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



- Cumulative distribution function

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



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

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

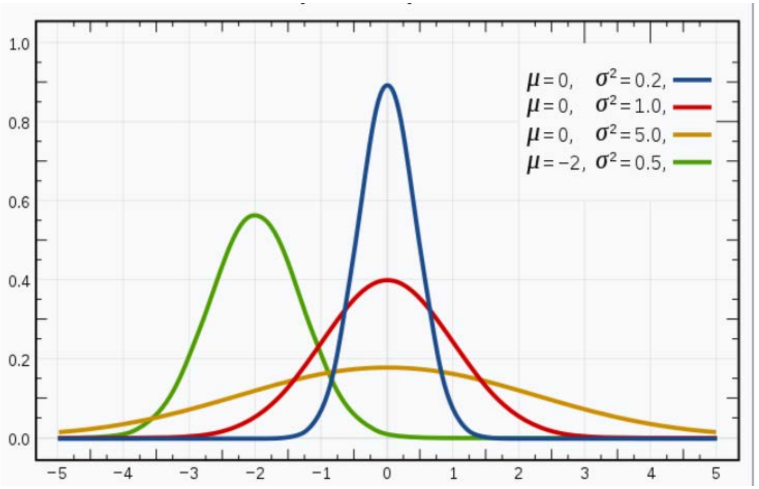
1777-1855

- **Probability density function** (PDF)
- Normal or Gaussian distribution

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

- A location-scale distribution: $X = \sigma Z + \mu \quad Z = (X-\mu)/\sigma$

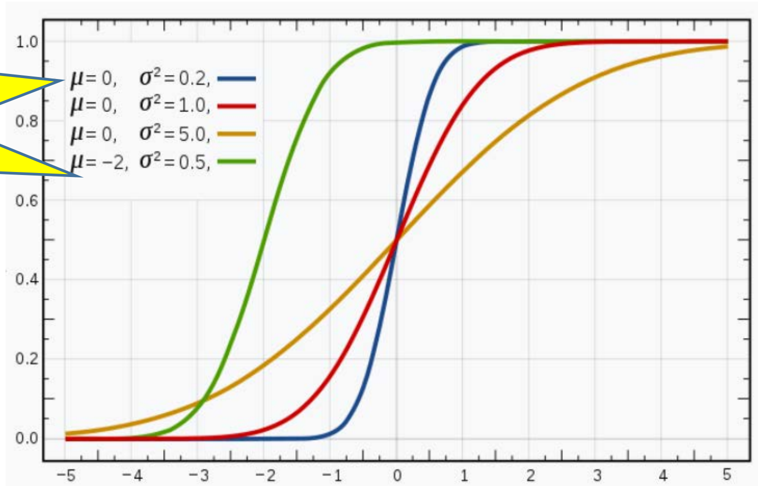
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) \qquad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



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

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

Riemann integral



- Expectation and moments

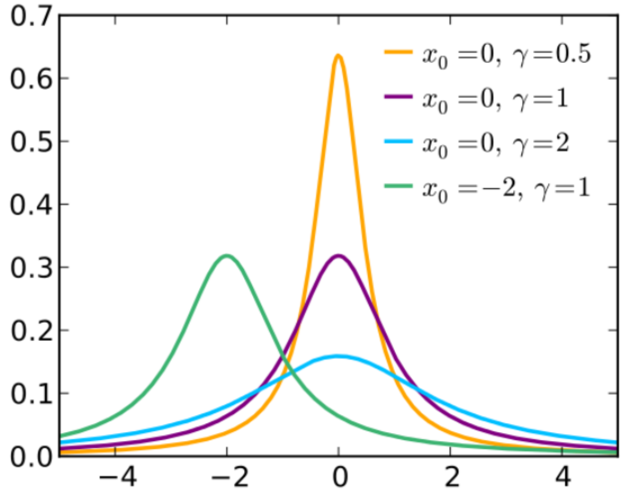
$$\mathrm{E}[X] = \int x f(x)\, dx. \qquad \mathrm{E}[X^p] = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even.} \end{cases}$$

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

Augustin-Louis Cauchy
(1789-1857)

- Lorentzian/Cauchy PDF:

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

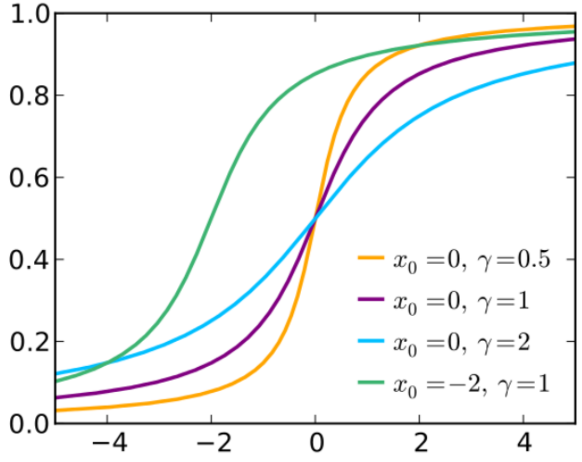


- CDF:

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

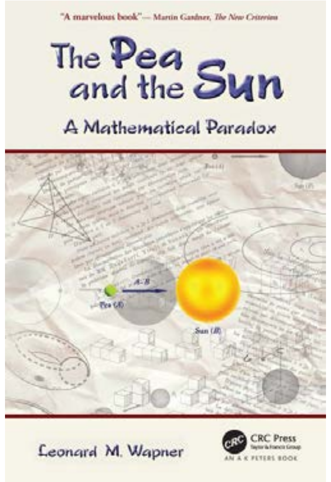
- Cauchy distributions do not have finite moments of any order! No expectation (bcs of improper integral)
- Location-scale family, standard Cauchy $\psi(x) = \frac{1}{\pi}\frac{1}{1+x^2}$

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



Probability measures

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



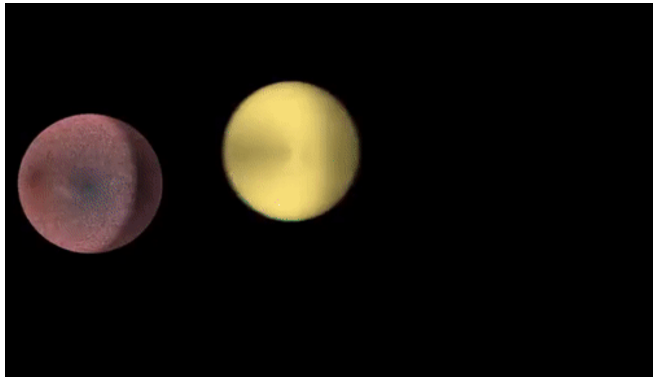
$$\Pr(\cup_i E_i) = \sum_i \Pr(E_i)$$

- 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 (of events)

- Pb: Cannot consider the full power set for continuous sample spaces

- Let us define an algebra of measurable events: the **σ-algebra**

1. $\mathbb{X} \in \mathcal{A}$,
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. $\forall i \in \mathbb{N}, A_i \in \mathcal{A} \Rightarrow \cup_{i \in \mathbb{N}} A_i \in \mathcal{A}$.

- σ-algebra generated/induced 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 $\mu : \mathcal{A} \to [0, \infty]$ 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

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

$(\mathbb{X}, \mathcal{B}(\mathbb{R}^d), \mu_L)$

Measurable function and simple functions

- 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:

$$X(\omega) = \sum_{i=1}^{k} x_i I_{A_i}(\omega), \text{ where } x_i \in [0, \infty), \ A_i \in \mathcal{A} \text{ with } A_i \cap A_j = \emptyset$$

Lebesgue integration

Henri Léon Lebesgue
(1875-1941)

- Riemann integral (signed area under the curve) not enough!
  (compact, problem with limits, etc.)
- **Integral of a simple function:**

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

- Other notations: $\int X \mathrm{d}\mu(\omega)$ $\int X \mathrm{d}\mu$

- **Integral of positive measurable functions:**

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

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

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

Random variables and expectations

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

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

- Probability:

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

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

$$\begin{aligned} E[X_1] &= \int X_1(\omega)\mu_c(\mathrm{d}\omega), \\ E[X_2] &= \int X_2(\omega)\mu_L(\mathrm{d}\omega). \end{aligned}$$

Density and dominating measure

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

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

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

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

ν is dominated by μ

$$\Downarrow$$

Let $\lambda = \frac{\mu+\nu}{2}$ then $\mu, \nu \ll \lambda$ $\qquad \mathbf{supp}(\nu) \subseteq \mathbf{supp}(\mu)$

Radon-Nikodym theorem and RN 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*

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

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

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

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

Statistical inference: Estimators

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

  Simply 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

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

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

Solving these equations gives

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

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

Pafnuty Chebyshev
(1821-1894)

- Infinitely many (point) estimators! **Which one is best?**

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:

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

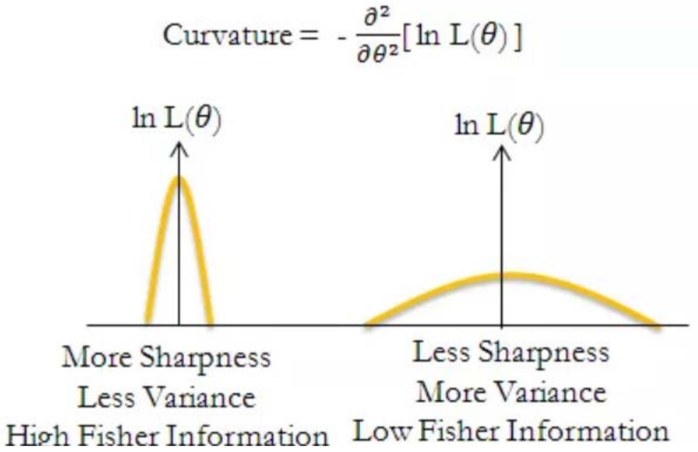
$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

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

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

- Consistent method: converge in probability to the true value $\widehat{\theta}_{\mathrm{mle}} \xrightarrow{\mathrm{p}} \theta_0$

Fisher information



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

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

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

- 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): Univariate case

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

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

- Fisher information:

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

Cramer-Rao lower bound and information geometry. Connected at Infinity II, 2013.

Cramer-Rao lower bound: Multivariate case

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

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

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

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 of the FIM $$[I(\theta)]_{ij} = 4\int_{x\in\mathcal{X}}\frac{\partial}{\partial\theta_i}\sqrt{p_\theta(x)}\frac{\partial}{\partial\theta_j}\sqrt{p_\theta(x)}\mathrm{d}x.$$

Cramer-Rao lower bound and information geometry. Connected at Infinity II, 2013.

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 σ2 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

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

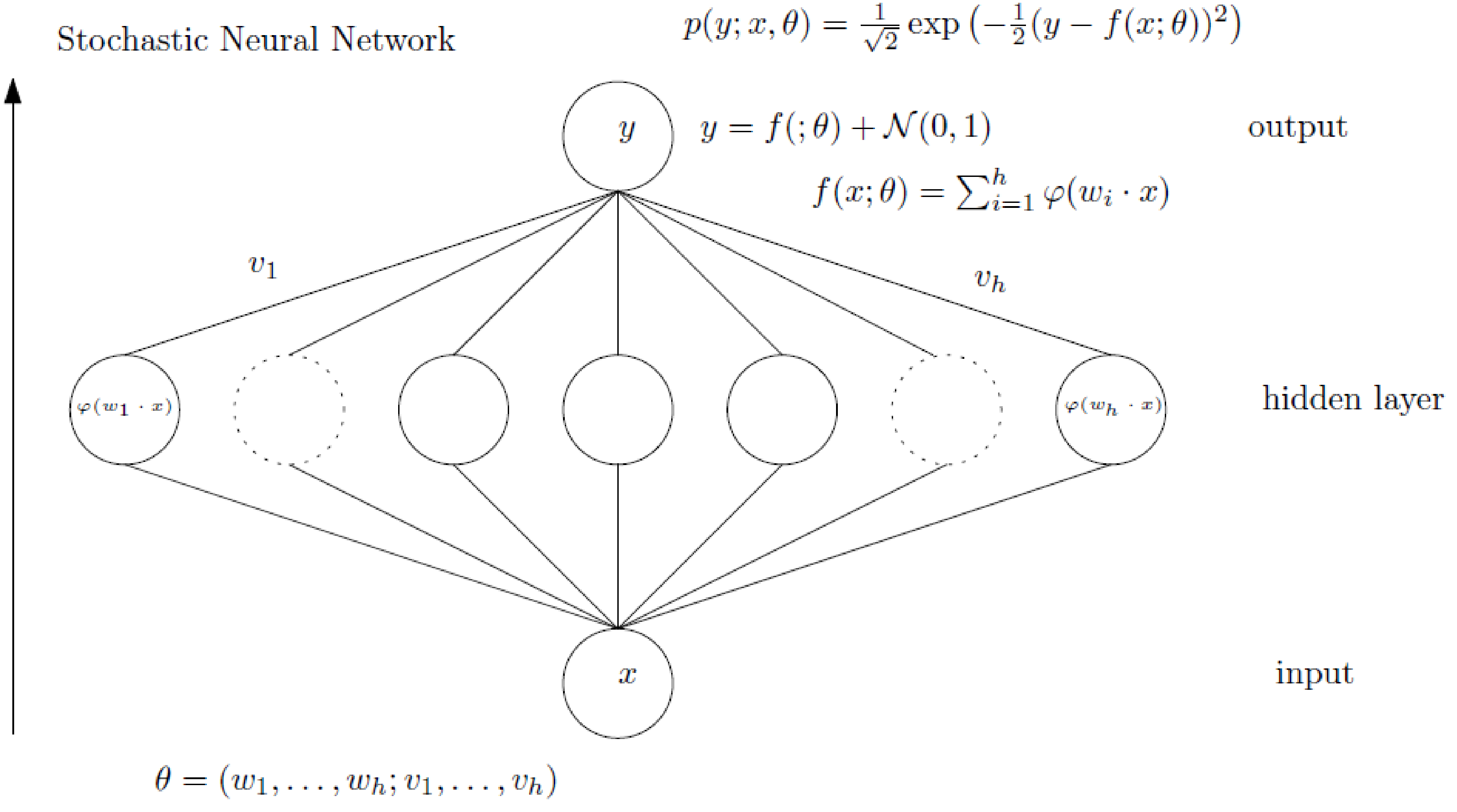
$$g_{ij}(\xi) = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ell(\xi)] = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ln p_\xi].$$

- Positive semi-definite FIM
- Positive-definite FIM for regular models (=identifiable)
- FIM is invariant under reparametrizations of the sample space X.
- 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)



Multiple Layer Perceptron (MLP)

Key concept: 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.

  Informally, a statistical lossless compression scheme…

- 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:

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

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

Example of sufficient statistics:

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

For Poisson distributions of intensity λ:

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

$\sum_{i=1}^n x_i$ *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**:

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

Log-Laplace transform

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

F: Log-normalizer (also known as partition function, cumulant function, etc.)

Exponential families (from Natural EFs to EFs)

- Consider a **(sufficient) statistic** t(x)
- Consider an **additional carrier measure term** k(x)
- Consider an **inner product** between t(x) and θ

(usual scalar/dot product)

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

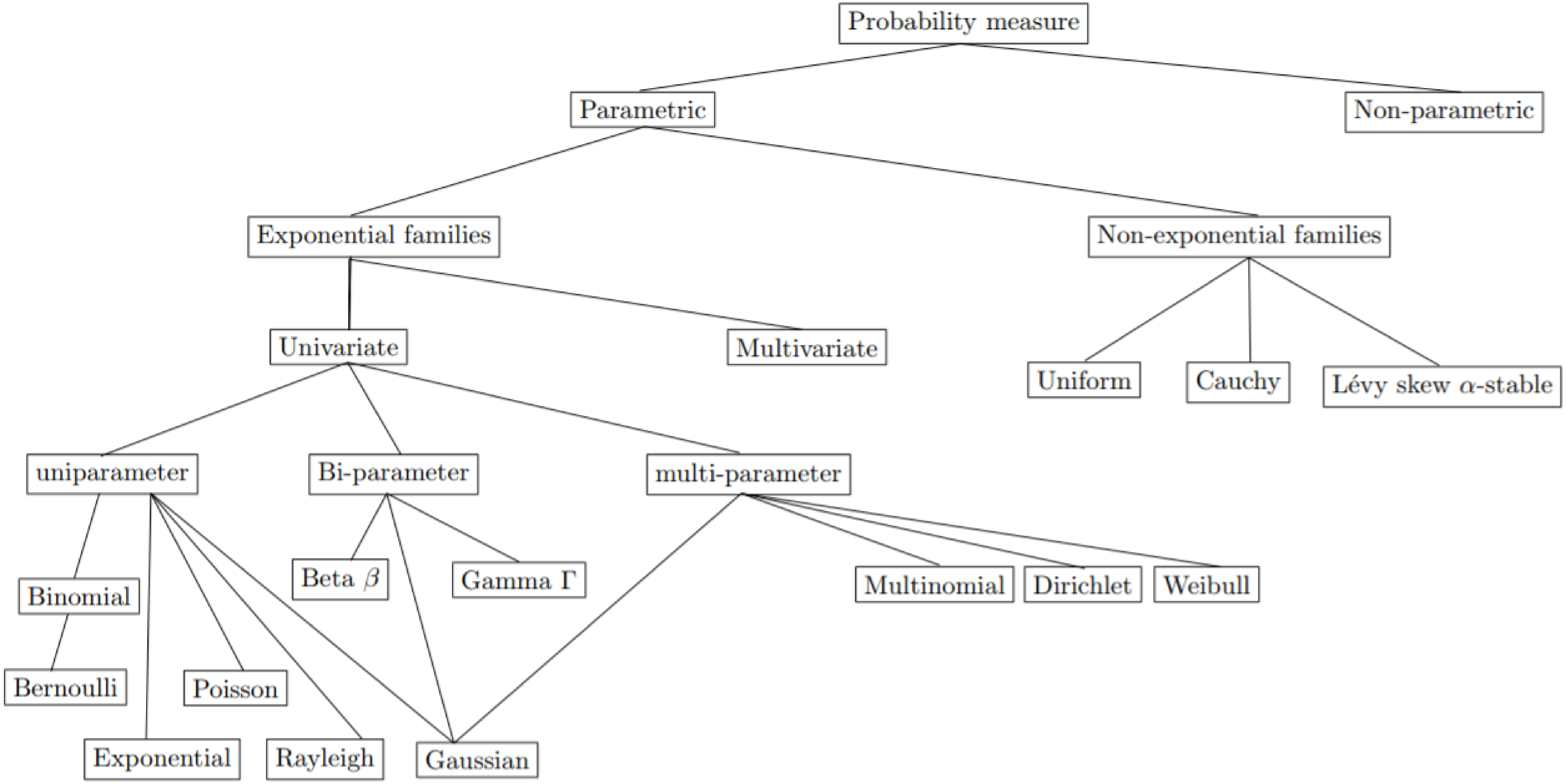
Properties:

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

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

Exponential families have finite moments of any order

Many common distributions are exponential families in disguise



Maximum likelihood estimator for exponential families

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

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

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

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

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

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

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

Regular EFs and steepness of exponential families

- An exponential family is **regular** when the natural parameter space is open

$$\Theta = \mathrm{int}(\Theta)$$

- Closed convex hull of {t(x)}: $$\mathcal{C} = \overline{\mathrm{co}(\mathcal{S})}$$

- 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 $$H = \mathrm{int}(\mathcal{C})$$
- MLE exists and is unique for regular and steep EFs when

$$\bar{t} = \sum_{i=1}^n t(x_i) \in \mathcal{C}$$

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

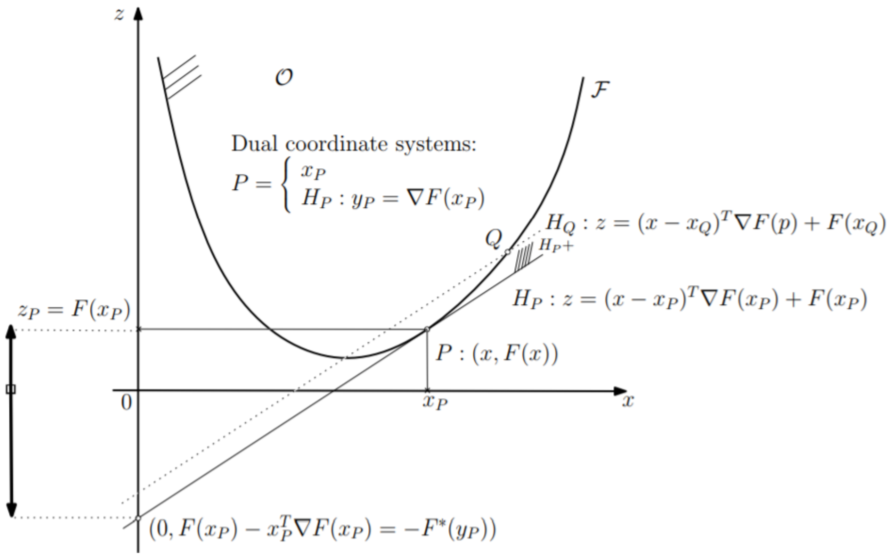
Example of non-steep family: Singly-truncated Gaussian family

Dual moment/expectation parameterization

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

- Related to the Legendre-Fenchel convex conjugate:

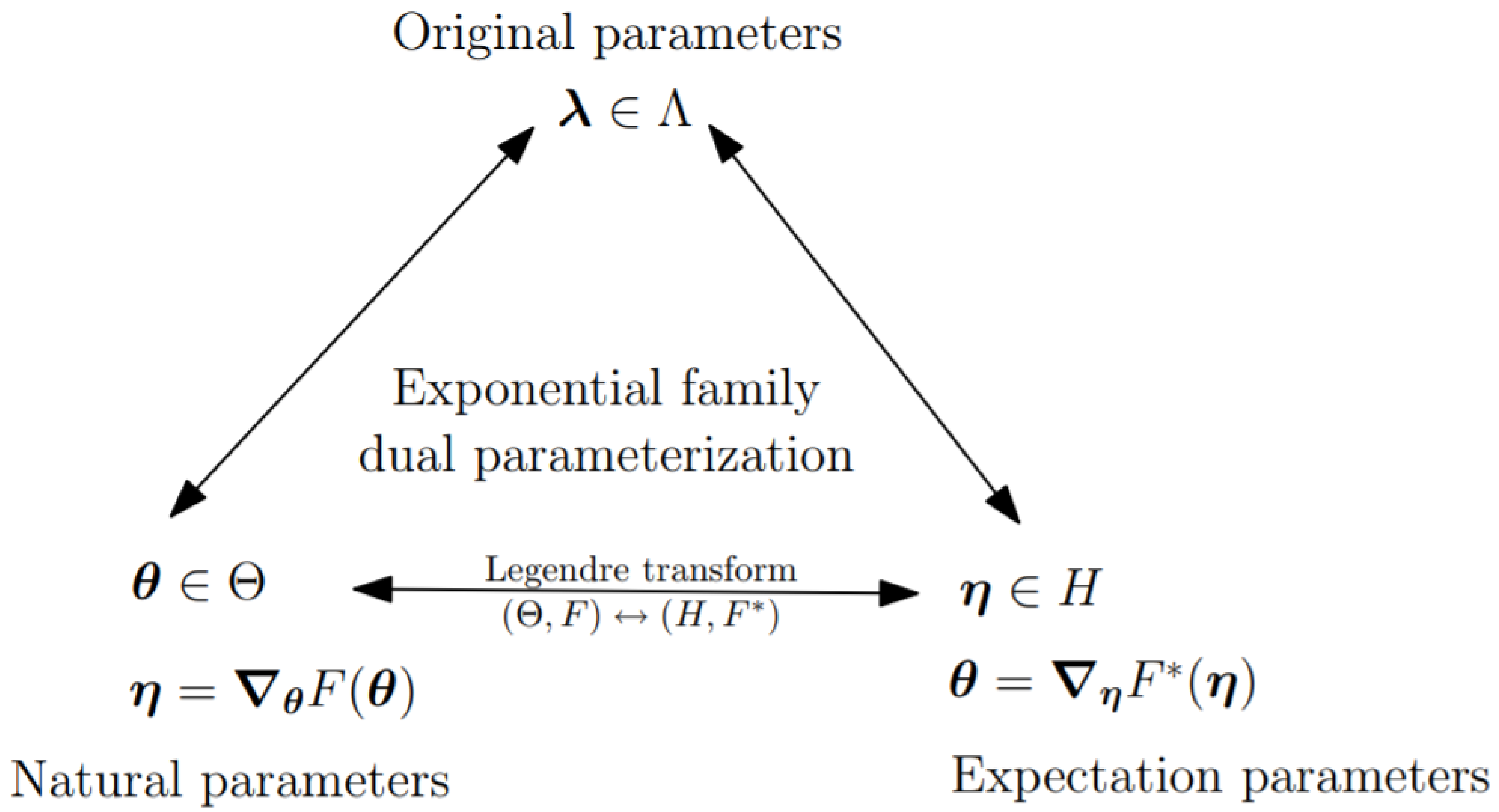
$$F^*(\eta) = \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\}$$



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

Legendre transformation and information geometry, 2010.

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:

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

- 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	$D_{\mathrm{KL}}(f_1 \| f_2) = \log\left(\frac{1-p_1}{1-p_2}\right) - p_1 \log\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)$
MLE	$\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i$
Source parameters	$\mathbf{\Lambda} = p \in [0,1]$
Natural parameters	$\mathbf{\Theta} = \theta \in \mathbb{R}^+$
Expectation parameters	$\mathbf{H} = \eta \in [0,1]$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \log\left(\frac{p}{1-p}\right)$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \frac{\exp\theta}{1+\exp\theta}$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = p$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \log(1 + \exp\theta)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \frac{\exp\theta}{1+\exp\theta}$
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

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

Univariate Gaussian family
Order 2

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	$D_{\mathrm{KL}}(f_P \\| f_Q) = \frac{1}{2}\left(2\log\frac{\sigma_Q}{\sigma_P} + \frac{\sigma_P^2}{\sigma_Q^2} + \frac{(\mu_Q-\mu_P)^2}{\sigma_Q^2} - 1\right)$
MLE	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})^2}$
Source parameters	$\mathbf{\Lambda} = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}^+$
Natural parameters	$\mathbf{\Theta} = (\theta_1,\theta_2) \in \mathbb{R}\times\mathbb{R}^-$
Expectation parameters	$\mathbf{H} = (\eta_1,\eta_2) \in \mathbb{R}\times\mathbb{R}^+$
$\mathbf{\Lambda}\to\mathbf{\Theta}$	$\mathbf{\Theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$
$\mathbf{\Theta}\to\mathbf{\Lambda}$	$\mathbf{\Lambda} = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2}\right)$
$\mathbf{\Lambda}\to\mathbf{H}$	$\mathbf{H} = (\mu, \sigma^2+\mu^2)$
$\mathbf{H}\to\mathbf{\Lambda}$	$\mathbf{\Lambda} = (\eta_1, \eta_2-\eta_1^2)$
$\mathbf{\Theta}\to\mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H}\to\mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2}\right)$
G	$G(\mathbf{H}) = -\frac{1}{2}\log\left(\eta_1^2-\eta_2\right) + C$
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	$k(x) = 0$

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

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

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

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	$D_{\mathrm{KL}}(f_P\|f_Q) = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_Q}{\det\Sigma_P}\right)+\mathrm{tr}\left(\Sigma_Q^{-1}\Sigma_P\right)\right) +\frac{1}{2}\left((\mu_Q-\mu_P)^\top\Sigma_Q^{-1}(\mu_Q-\mu_P)-d\right)$
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	$\mathbf{\Lambda}=(\mu,\Sigma)$ with $\mu\in\mathbb{R}^d$ and $\Sigma\succ 0$
Natural parameters	$\mathbf{\Theta}=(\theta,\Theta)$
Expectation parameters	$\mathbf{H}=(\eta,H)$
$\mathbf{\Lambda}\to\mathbf{\Theta}$	$\mathbf{\Theta}=\left(\Sigma^{-1}\mu,\frac{1}{2}\Sigma^{-1}\right)$
$\mathbf{\Theta}\to\mathbf{\Lambda}$	$\mathbf{\Lambda}=\left(\frac{1}{2}\Theta^{-1}\theta,\frac{1}{2}\Theta^{-1}\right)$
$\mathbf{\Lambda}\to\mathbf{H}$	$\mathbf{H}=(\mu,-(\Sigma+\mu\mu^T))$
$\mathbf{H}\to\mathbf{\Lambda}$	$\mathbf{\Lambda}=(\eta,-(H+\eta\eta^T))$
$\mathbf{\Theta}\to\mathbf{H}$	$\mathbf{H}=\nabla F(\mathbf{\Theta})$
$\mathbf{H}\to\mathbf{\Theta}$	$\mathbf{\Theta}=\nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta})=\frac{1}{4}\mathrm{tr}(\Theta^{-1}\theta\theta^T)-\frac{1}{2}\log\det\Theta+\frac{d}{2}log\pi$
Gradient log normalizer	$\nabla F(\mathbf{\Theta})=\left(\frac{1}{2}\Theta^{-1}\theta,-\frac{1}{2}\Theta^{-1}-\frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T\right)$
G	$G(\mathbf{H})=-\frac{1}{2}\log\left(1+\eta^TH^{-1}\eta\right)-\frac{1}{2}\log\det(-H)-\frac{d}{2}\log(2\pi e)$
Gradient G	$\nabla G(\mathbf{H})=\left(-(H+\eta\eta^T)^{-1}\eta,-\frac{1}{2}(H+\eta\eta^T)^{-1}\right)$
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:

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

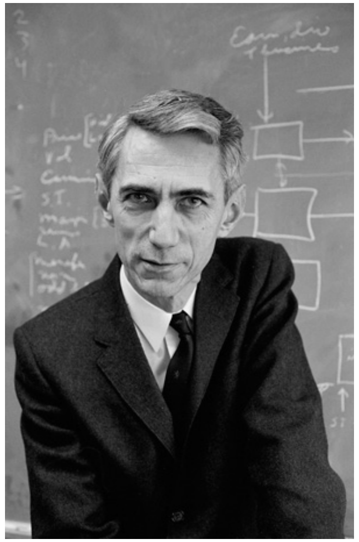
Statistical exponential families: A digest with flash cards. arXiv:0911.4863 (2009)

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

Summary

- **Probability measure** bypasses the Banach-Tarsky paradox by fixing a σ-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 in closed-form in dual moment parameterization

Information Theory



Background

Frank Nielsen

Sony CSL

Outline

- Shannon 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
- Kullback-Leibler divergence of location-scale families

Shannon's 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 $\mathcal{S}_n : \Delta_n \to \mathbb{R}^+$ 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, $U_n = (1/n, \ldots, 1/n) \in \Delta_n$, i.e. $\mathcal{S}_n(P) \leq \mathcal{S}_n(U_n)$, for any $P \in \Delta_n$;

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

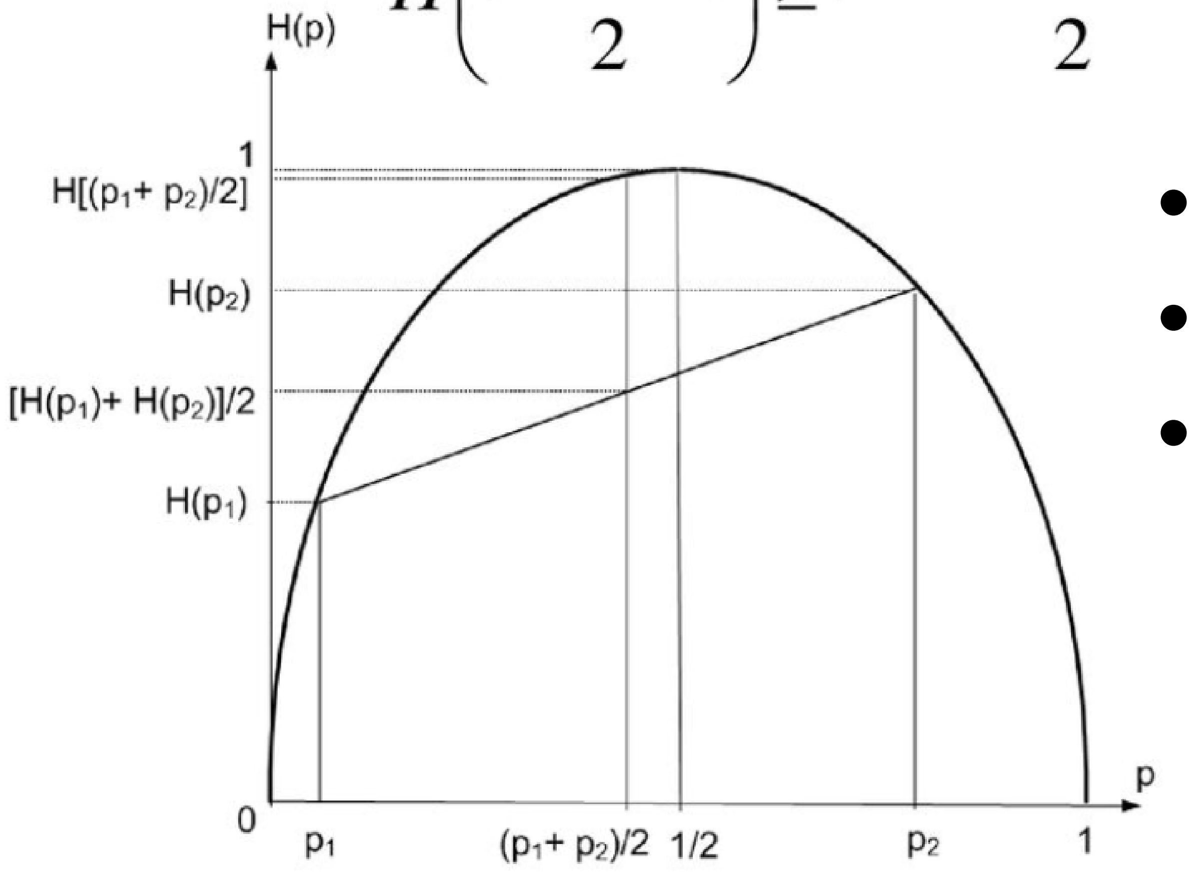
[SA4] Let $P = (p_1, \ldots, p_n) \in \Delta_n$, $PQ = (r_{11}, r_{12}, \ldots, r_{nm}) \in \Delta_{nm}$, $n, m \in \mathbb{N}$, $n, m > 1$ such that $p_i = \sum_{j=1}^m r_{ij}$, and $Q_{|k} = (q_{1|k}, \ldots, q_{m|k}) \in \Delta_m$, where $q_{i|k} = r_{ik}/p_k$. Then,

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

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

Shannon's entropy is a concave function

$$H\left(\frac{p_1+p_2}{2}\right) \geq \frac{H(p_1)+H(p_2)}{2}$$



- 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

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

- Can be negative : e.g., Gaussian distributions $\frac{1}{2}\log(2\pi e\sigma^2)$
- Can be infinite when the integral diverges $h(X) = +\infty$

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

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

NB: 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)

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

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

With Radon-Nikodym derivative with respect to to base measure μ:

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

Unifies:

- discrete entropy (counting measure)
- differential entropy (Lebesgue measure)

Relative entropy: Kullback-Leibler divergence (KLD)

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

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

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

KLD = Relative entropy with respect to a reference distribution P

Not a metric distance because (1) asymmetric and (2) failing the triangle inequality

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

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

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

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

KLD is an oriented distance!

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 η:

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

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 $\gamma = 0.57721566...$ 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:

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

Entropies and cross-entropies of exponential families, IEEE ICIP 2010

Kullback-Leibler divergence for exponential families Fenchel-Young divergence for exponential families

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

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

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

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

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

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

Edwin Thompson Jaynes
(1922–1998)

Maximizing a concave function subject to linear constraints
(or equivalently convex mininimization optimization problem).

MaxEnt with Kullback-Leibler divergence and with a prior constraint distribution q

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

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

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

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

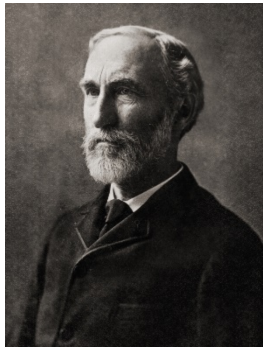
MaxEnt is KL left-sided minimization

Maximum entropy distribution is the uniform prior:

$$q(x) = \frac{1}{n}$$

MaxEnt distributions (Boltzmann-Gibbs)

Ludwig Boltzmann



Josiah Willard Gibbs

Solving the 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, etc.

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

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

Free enery
log-partition
cumulant function

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

Example: Fixed mean and fixed variance MaxEnt distribution

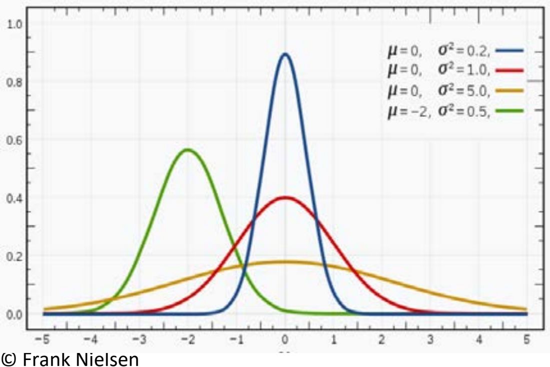
- Find the MaxEnt distributions with support the full real line and the first two moments prescribed

$$E[X] = m_1 \qquad E[X^2] = m_2$$

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

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

Gaussian family



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

MLE as a right-sided KLD minimization

Recall that MaxEnt is **KL left-sided minimization**:

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

Empirical distribution:

$$p_e(x) = \frac{1}{m} \sum_{i=1}^{m} \delta_{s_i}(x)$$

$$\begin{aligned}
\min \quad & \mathrm{KL}(p_e(x) : \boxed{p_\theta(x)}) \\
&= \int p_e(x) \log p_e(x) \mathrm{d}x - \int p_e(x) \log p_\theta(x) \mathrm{d}x \\
&= \min -H(p_e) - \underbrace{E_{p_e}[\log p_\theta(x)]}_{} \\
&\equiv \max \frac{1}{n} \sum \delta(x - x_i) \log p_\theta(x) \\
&= \max \frac{1}{n} \sum_i \log p_\theta(x_i) = \mathrm{MLE}
\end{aligned}$$

MLE is KL right-sided minimization

Upper bounding the differential entropy of mixtures (1/2)

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 with the same moment expectations has less entropy. In particular, this observation applies to statistical mixtures.

$$H(X) = \int_{\mathcal{X}} p(x) \log \frac{1}{p(x)} \mathrm{d}x = - \int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}x \qquad H(p(x;\theta)) = -F^*(\eta(\theta))$$

Absolute Monomial Exponential Family (AMEF): $p_l(x;\theta) = \exp\left(\theta |x|^l - F_l(\theta)\right)$

with log-normalizer $F_l(\theta) = \log 2 + \log \Gamma\left(\frac{1}{l}\right) - \log l - \frac{1}{l} \log(-\theta)$

$$\Gamma(u) = \int_0^\infty x^{u-1} \exp(-x) \mathrm{d}x$$
$$\Gamma(n) = (n-1)! \text{ for } n \in \mathbb{N}$$

MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions, IEEE SPL 2017, arxiv:1612.02954

https://www.lix.polytechnique.fr/~nielsen/MEUB/

Upper bounding the differential entropy of mixtures (2/2)

$$p_l(x;\theta) = \exp\left(\theta|x|^l - F_l(\theta)\right)$$

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

$$H_l(\eta) = \log 2 + \log\Gamma\left(\frac{1}{l}\right) - \log l + \frac{1}{l}(1 + \log l + \log\eta)$$
$$H_l(\theta) = \log 2 + \log\Gamma\left(\frac{1}{l}\right) - \log l + \frac{1}{l}(1 - \log(-\theta)),$$

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

$$H(X) \leq U_1(X)$$

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

MaxEnt distribution is Laplacian distribution

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

$$\bar{\mu} = \sum_{c=1}^{k} w_c\mu_c$$

MaxEnt distribution is Gaussian distribution

A series of upper bounds for h(GMMs)

Zero-centered Gaussian Mixture Models:

$$H(X) \leq H_l^\eta(A_l(X)) = b_l + \frac{1}{l}\log z_l + \log \bar{\sigma}_l,$$

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

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

MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions, IEEE SPL 2017, arxiv:1612.02954

Computing non-central absolute geometric moments of Gaussians and GMMs

Even l	$A_l = E\left[\lvert X\rvert^l\right] = E\left[X^l\right] = \sum_{i=0}^{\lfloor \frac{l}{2} \rfloor} \binom{l}{2i}(2i-1)!!\mu^{l-2i}\sigma^{2i}$
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 l	$A_l = E\left[\lvert X\rvert^l\right] = C_l(\mu,\sigma)\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2}) + D_l(\mu,\sigma)\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$
1	$\sigma\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2}) + \mu\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$
3	$(2\sigma^3+\mu^2\sigma)\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2}) + (\mu^3+3\mu\sigma^2)\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$
5	$(8\sigma^5+9\mu^2\sigma^3+\mu^4\sigma)\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2}) + (\mu^5+10\mu^3\sigma^2+15\mu\sigma^4)\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$
7	$(48\sigma^7+87\mu^2\sigma^5+20\mu^4\sigma^3+\mu^6\sigma)\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2})+$ $(\mu^7+21\mu^5\sigma^2+105\mu^3\sigma^4+105\mu\sigma^6)\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$
9	$(384\sigma^9+975\mu^2\sigma^7+345\mu^4\sigma^5+35\mu^6\sigma^3+\mu^8\sigma)\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu^2}{2\sigma^2})+$ $(\mu^9+36\mu^7\sigma^2+378\mu^5\sigma^4+1260\mu^3\sigma^6+945\sigma^8)\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma})$

Computing the Kullback-Leibler divergence…

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

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

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

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

Symmetric KL

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

Kullback-Leibler divergence: Location-scale families

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

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

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

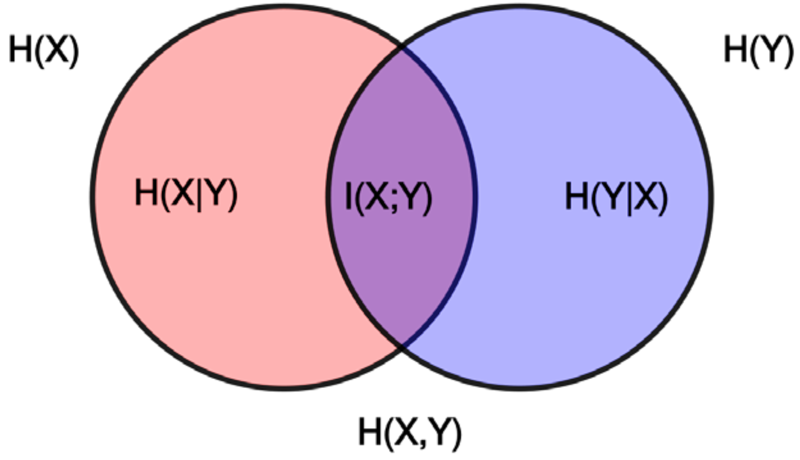
$$\begin{aligned} \mathrm{KL}(p_{l_1,s_1} : q_{l_2,s_2}) &= h^\times\left(p : q_{\frac{l_2-l_1}{s_1},\frac{s_2}{s_1}}\right) - h(p) = \mathrm{KL}\left(p : q_{\frac{l_2-l_1}{s_1},\frac{s_2}{s_1}}\right), \\ &= h^\times\left(p_{\frac{l_1-l_2}{s_1},\frac{s_1}{s_2}} : q\right) - h(p) + \log\frac{s_2}{s_1}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

Interesting properties for the KL minimization:

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

On the Kullback-Leibler divergence between location-scale densities, arXiv:1904.10428

Mutual information of RVs (MI)



- Consider two **random variables** X and Y.
- There are independent if and only if

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

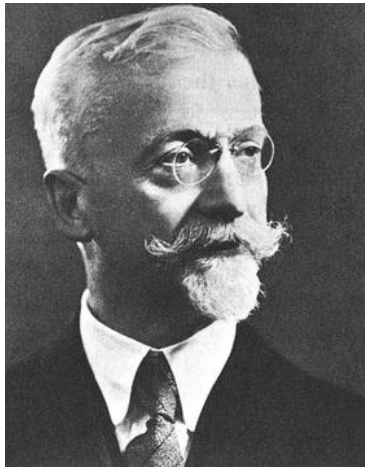
- Amount of mutual information quantified as the KL divergence between the joint distribution and the product of distributions

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

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

MI is not a metric distance but a symmetric distance between random variables

Elements of differential geometry



Elie Cartan
1869-1951

∇

Frank Nielsen

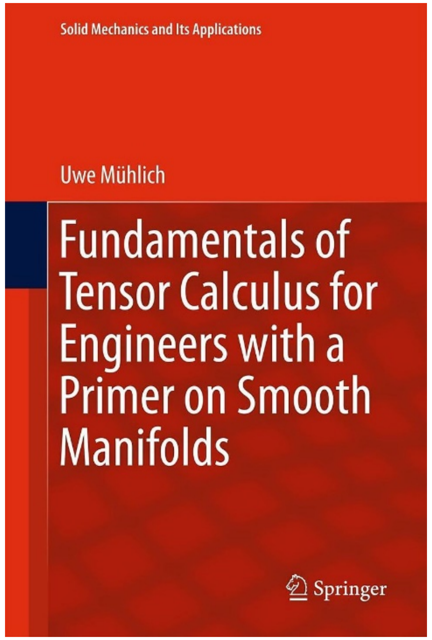
Jean-Louis Koszul
(1921-2018)

Charles Ehresmann
(1905-1979)

Sony CSL

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 $B = \{b_1, \ldots, b_D\}$ is **linearly independent** iff

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

- 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) $B = \{e_1, \ldots, e_d\}$
- Vector v written in a basis B using **coefficients/components:**

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

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: $\dim(V) = \dim(V^*)$
- Isomophism $V \simeq V^*$

- **Dual covector basis**: We have $\omega(v) = v^i \omega(e_i)$
- Choose covector basis which reads vector components: $\underline{e}^i(v) = v^i$

$$\omega(v) = \omega_i \underline{e}^i(v), \ \omega_i = \underline{\omega}(e_i) \Rightarrow \underline{e}^i(e_j) = \delta^i_j$$

Pairing product of a covector with a vector

Basis in vector space $B = \{e_1, \ldots, e_d\}$

Basis in covector space $B^* = \{e^1, \ldots, e^d\}$

Pairing product

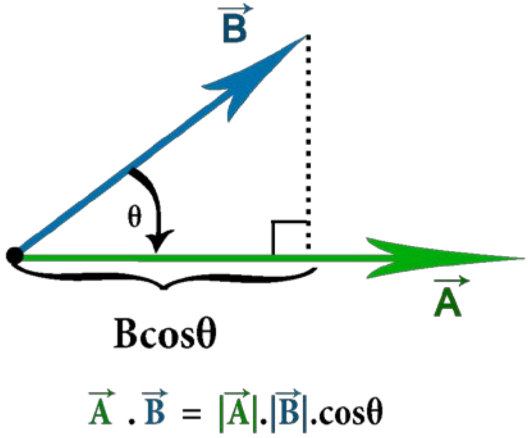
- By **notational definition**: $(\omega, v\rangle := \omega(v)$
- Vector components: $v^i = (e^i, v\rangle$
- Covector components $\omega_i = (\omega, e_i\rangle$

$$e^i(e_j) = (e^i, e_j\rangle = \delta^i_j$$

Inner product space:
notion of lengths/angles/orthogonality of vectors

Definition (*Inner product*) A mapping

$$\cdot : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$
$$(\mathbf{a}, \mathbf{b}) \mapsto \mathbf{a} \cdot \mathbf{b}$$



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 $||.||$ 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 v is its **norm**

Distance (metric) induced by a norm:

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

Reciprocal basis is a basis of vectors

- Given an inner product <.,.>, we can define a **reciprocal basis** of V

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

primal and reciprocal basis are **mutually orthogonal**

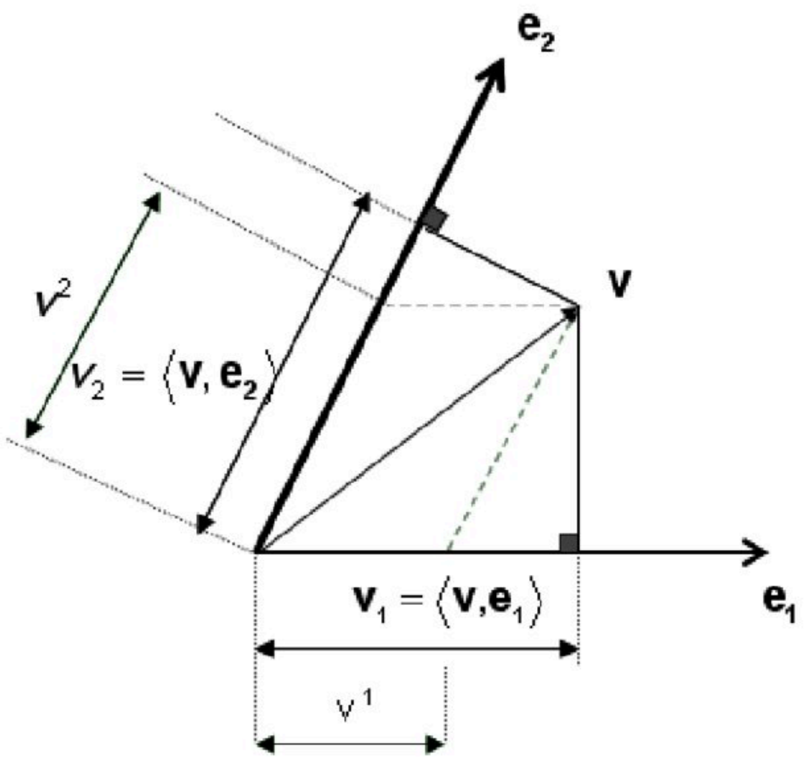
- The coefficients of a vector v in the **primal basis** are called the **contravariant coefficients**:

$$v = v^i e_i$$

- 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/components of a vector



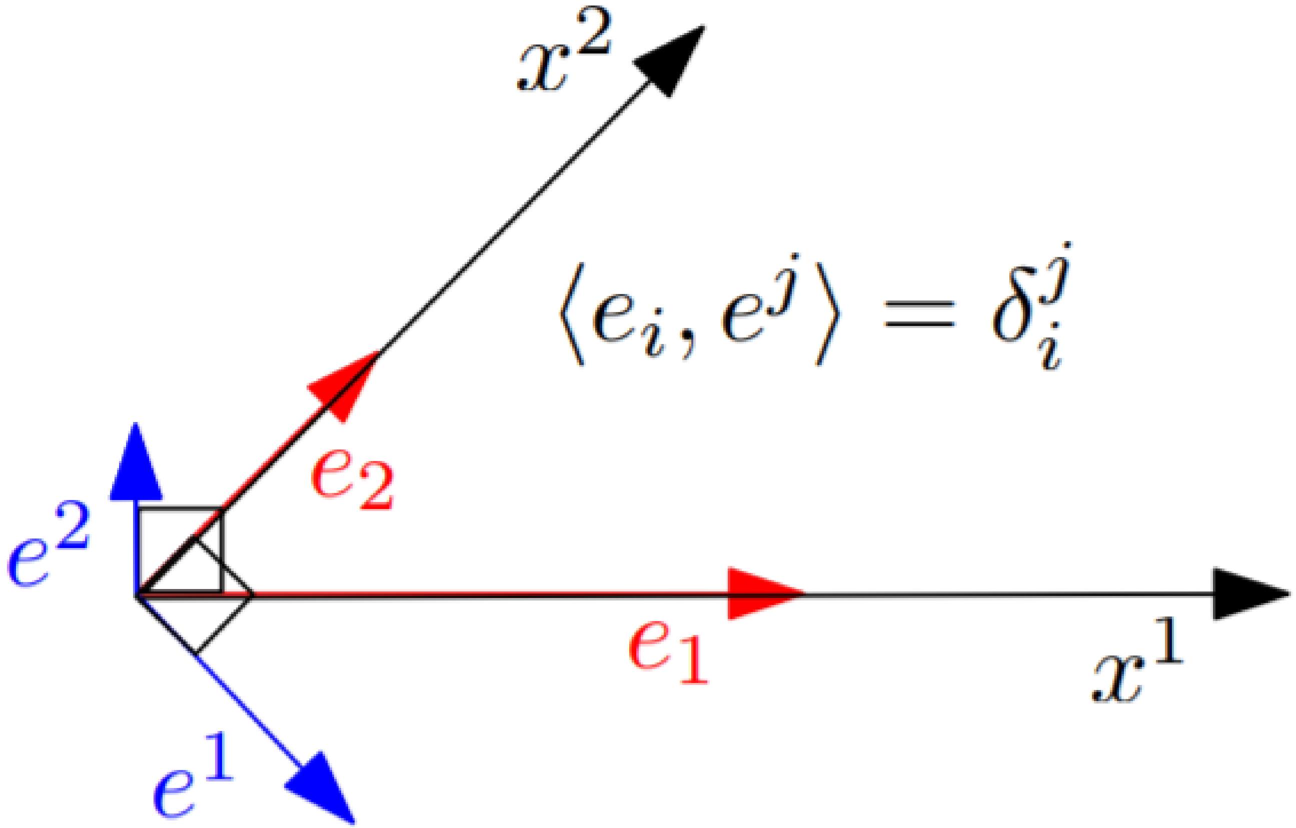
contravariant $v^i = \langle v, e_i \rangle$

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

Abide rules of change of basis

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

Primal and reciprocal basis are mutually orthogonal



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

Scalar product and dual metric tensors

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

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

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

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

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

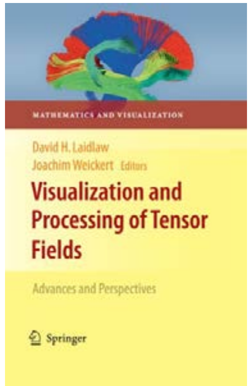
Converting covariant ↔ contravariant components

Raising and lowering indices

$$e^i = {g^*}^{ij} e_j \qquad v_i = g_{ij} v^i$$

$$e_i = g_{ij} e^j \qquad v^i = g^{ij} v_i$$

Geometric tensors and tensor algebra



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

A tensor of type (r,s) $$T : \underbrace{V^* \ldots V^*}_{r} \times \underbrace{V \times \ldots V}_{s} \to \mathbb{R}$$

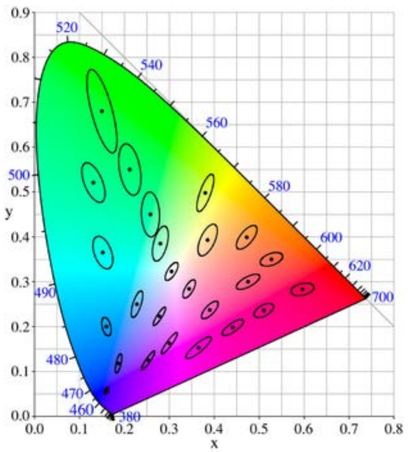
Components/coefficients $$T^{j_1 \cdots j_r}_{i_1 \ldots i_s}$$ with respect to a basis

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…

Coordinate-free description

- Bilinear positive-definite

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

- symmetric

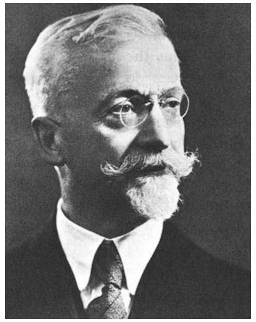
  $g(V,W) = g(W,V)$

- nondegenerate

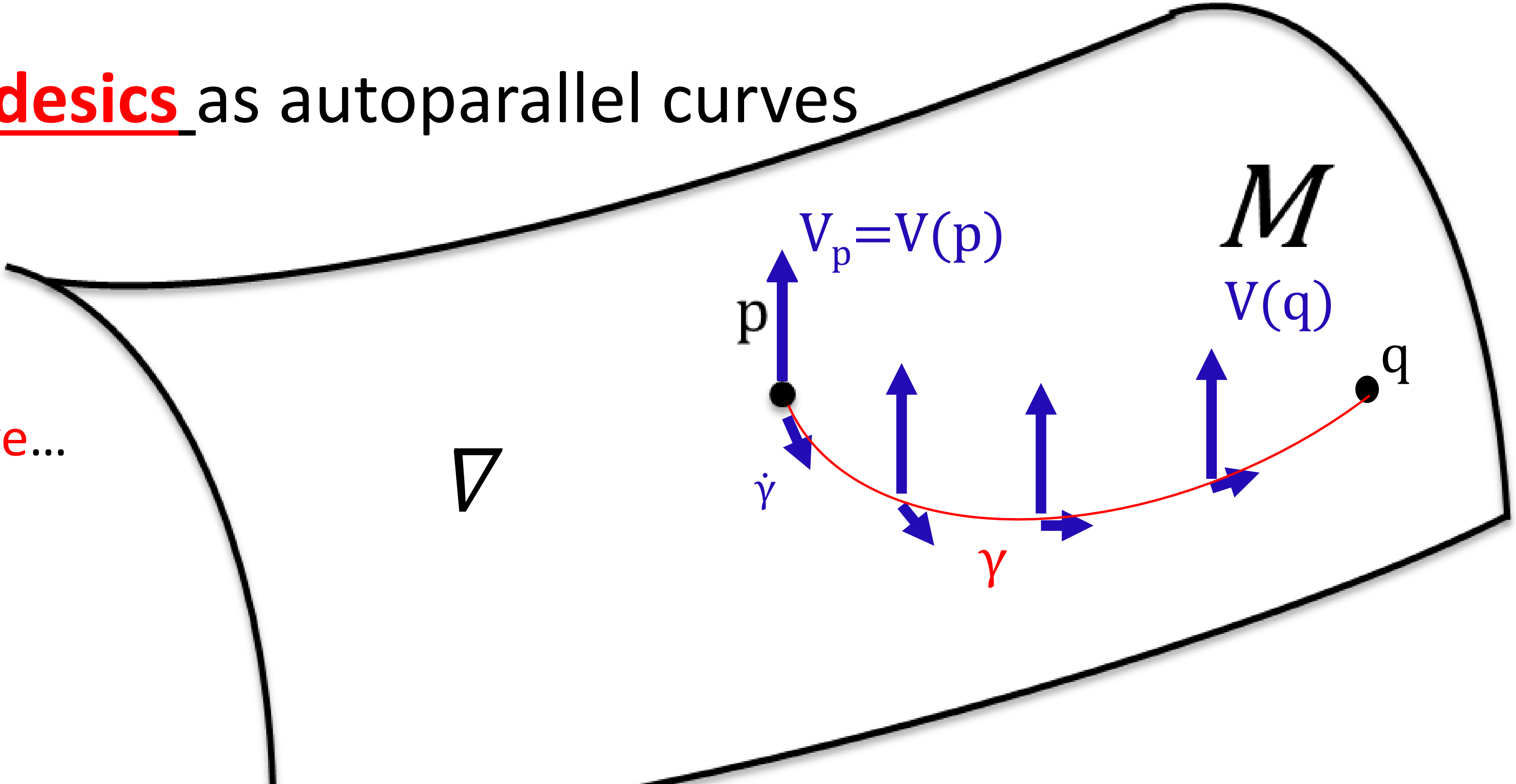
  $\forall p, \forall V \not\equiv 0\ \exists\ W, g_p(V,W) \neq 0$

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

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…

How to define an affine connection

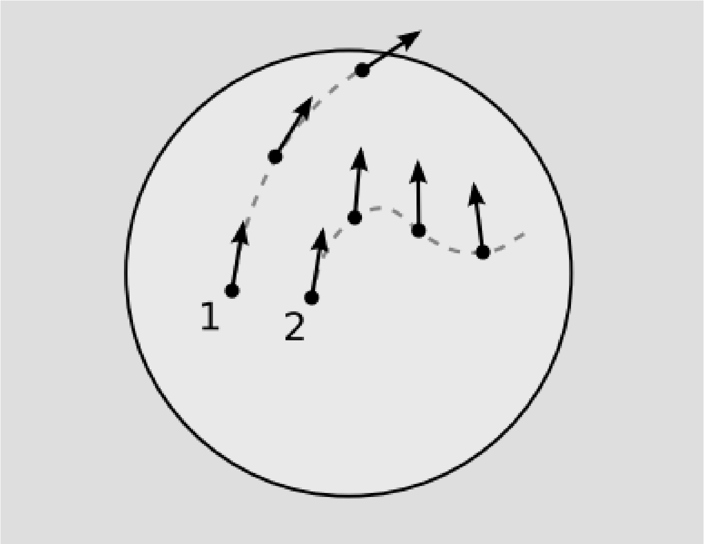
Elwin Bruno Christoffel
(1829-1900)

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

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

- 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)

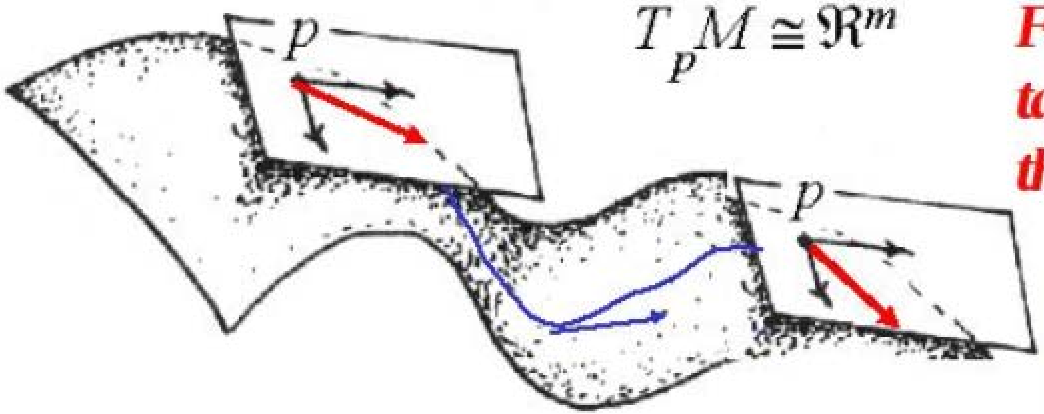
$$\ddot\gamma(t) + \Gamma^k_{ij}\dot\gamma(t)\dot\gamma(t) = 0, \quad \gamma^l(t) = x^l \circ \gamma(t)$$

Connection and covariant derivative

A connection is a map

$$\nabla : TM \times TM \rightarrow TM$$

$T_pM \cong \Re^m$



From the product of the tangent bundle with itself to the tangent bundle

with defining properties:

1. $\nabla_X(Y+Z) = \nabla_X Y + \nabla_X Z$
2. $\nabla_{(X+Y)} Z = \nabla_X Z + \nabla_Y Z$
3. $\nabla_{(fX)} Y = f\nabla_X Y$
4. $\nabla_X(fY) = X[f]Y + f\nabla_X Y$

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, this amount to check that

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

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

PT preserves metric

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

Fundamental theorem of Riemannian geometry

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

$$\nabla^{\mathrm{LC}}$$

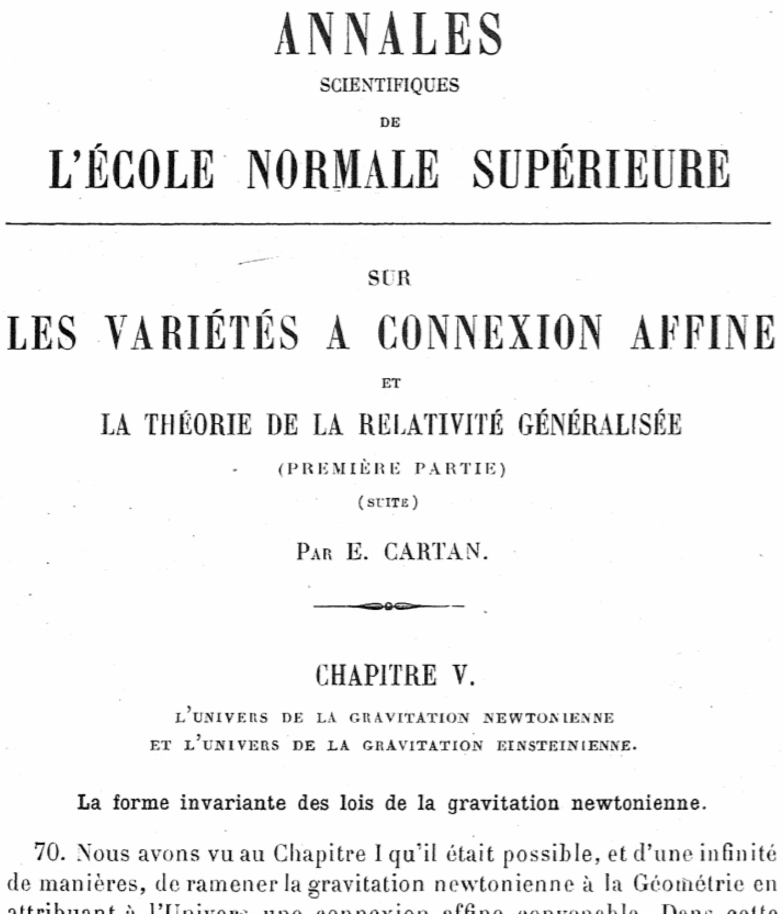
- The Christoffel symbols of the Levi-Civita connection are calculated from the metric tensor in local coordinates :

$${}^{\mathrm{LC}}\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij} \right)$$

- Or in coordinate-free equation by the **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


ANNALES
SCIENTIFIQUES
DE
L’ÉCOLE NORMALE SUPÉRIEURE

SUR
LES VARIÉTÉS A CONNEXION AFFINE
ET
LA THÉORIE DE LA RELATIVITÉ GÉNÉRALISÉE
(PREMIÈRE PARTIE)
(SUITE)

PAR E. CARTAN.

CHAPITRE V.

L’UNIVERS DE LA GRAVITATION NEWTONIENNE
ET L’UNIVERS DE LA GRAVITATION EINSTEINIENNE.

La forme invariante des lois de la gravitation newtonienne.

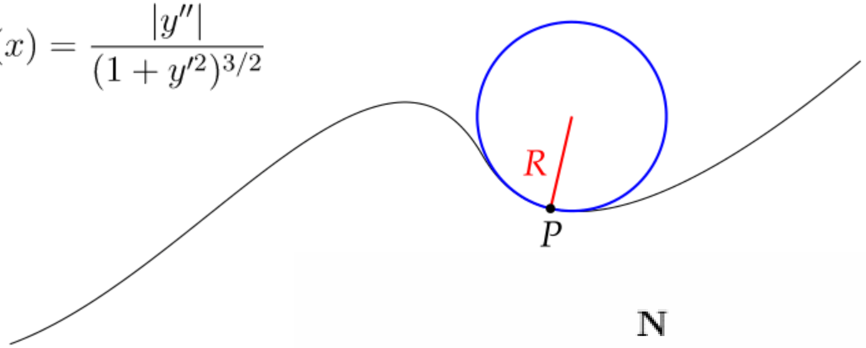
70. Nous avons vu au Chapitre I qu’il était possible, et d’une infinité de manières, de ramener la gravitation newtonienne à la Géométrie en

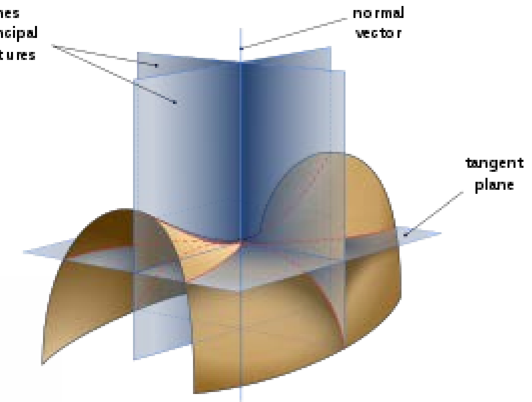
Cartan-Einstein manifold

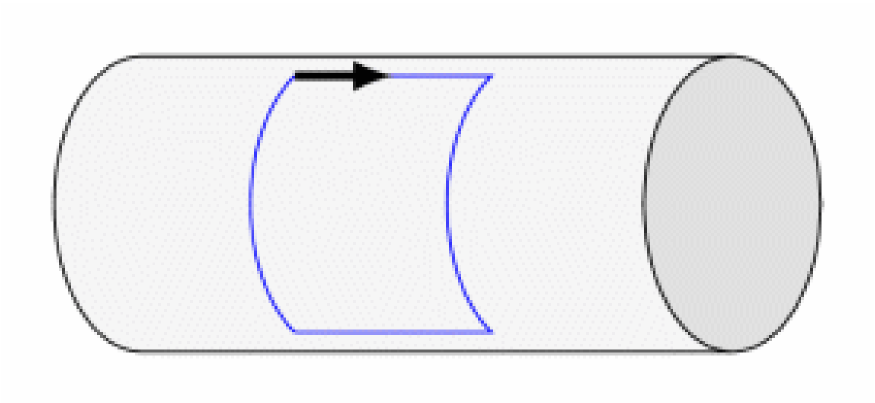
E. Cartan, Sur les variétés à connexion affine, et la théorie de la relativité généralisée , Ann. Ec. Norm. Sup. 40 (1923)

Curvature of ∇

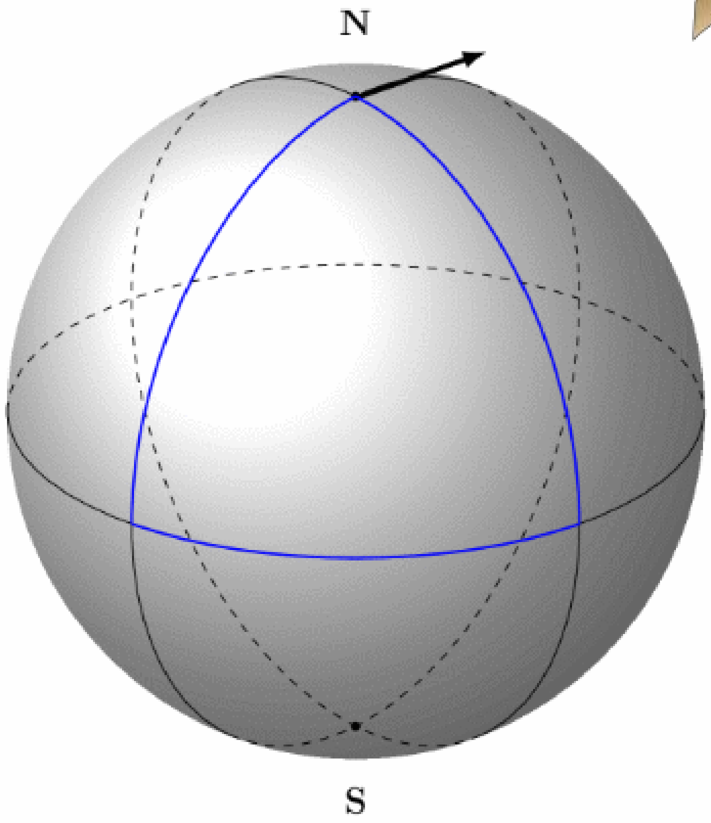
$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$







Cylinder is flat
Parallel transport is independent of path



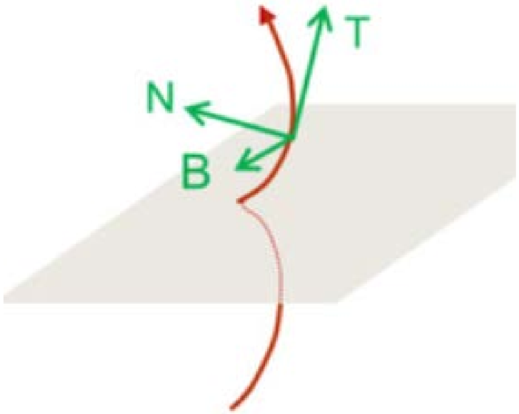
Sphere has constant curvature
Parallel transport is path-dependent

Torsion of a connection ∇

Torsion measures the speed of rotation of the binormal vector

parallel transport "twists" vectors.

- For connections:



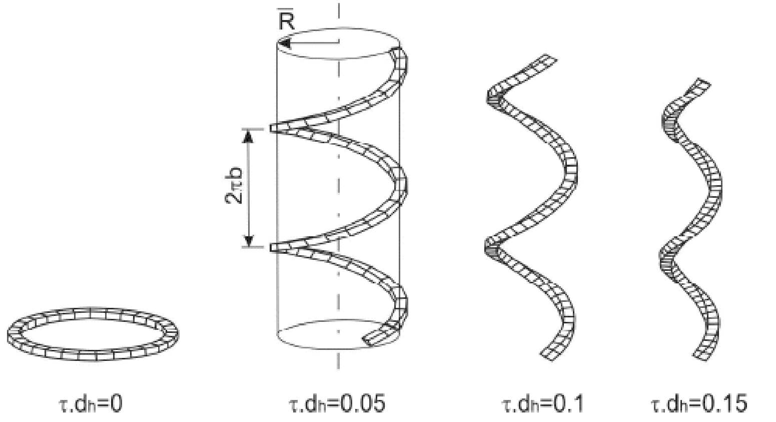


Figure 1. Helical channels with square cross section, constant curvature $\kappa.d_h = 1$ and torsion $\tau.d_h$ spanning from 0 to 0.15.

Torsion in geometry and in field theory 3

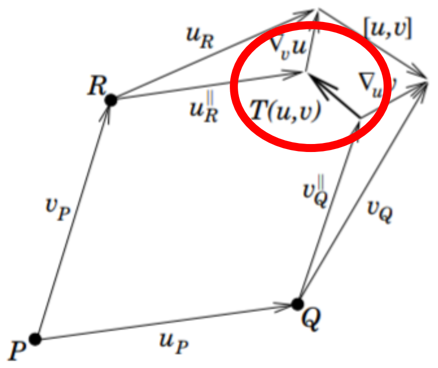
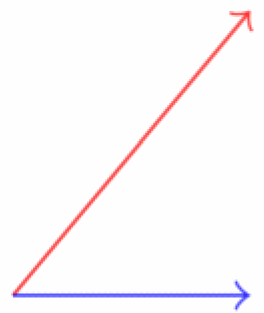


Figure 1: *On the geometrical interpretation of torsion*, see [39]: Two vector fields u and v are given. At a point P, we transport parallelly u and v along v or u, respectively. They become $u_R^{\|}$ and $v_Q^{\|}$. If a torsion is present, they don't close, that is, a *closure failure* $T(u,v)$ emerges. 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

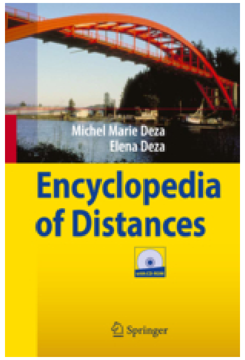
- **Manifold with an affine connection**: tensor fields, parallel transport, geodesics, curvature and torsion

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:

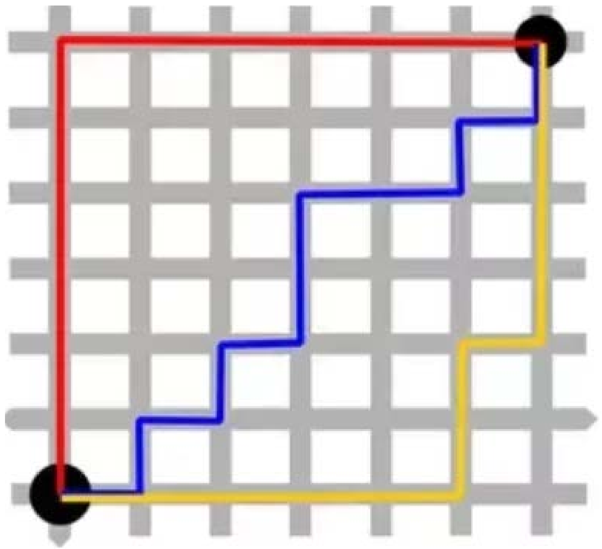
- M1. (Non-negativity) $D(p_1, p_2) \geq 0$
- M2. (Identity of the indiscernibles) $D(p_1, p_2) = 0 \Leftrightarrow p_1 = p_2$
- M3. (Symmetry) $D(p_1, p_2) = D(p_2, p_1)$
- M4. (Triangle inequality/subadditivity)

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

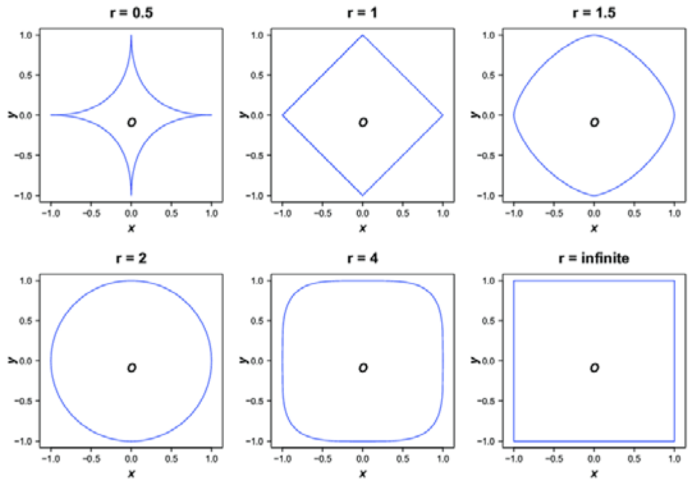
Examples of metric spaces

- 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

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



L1 is not geodesic



Non-metric (not convex) and metric balls (convex)

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 an its dot/scalar product

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

- Example with Minkowski norms

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

Distances and some notational conventions

- Typing 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 $H(X,Y)$

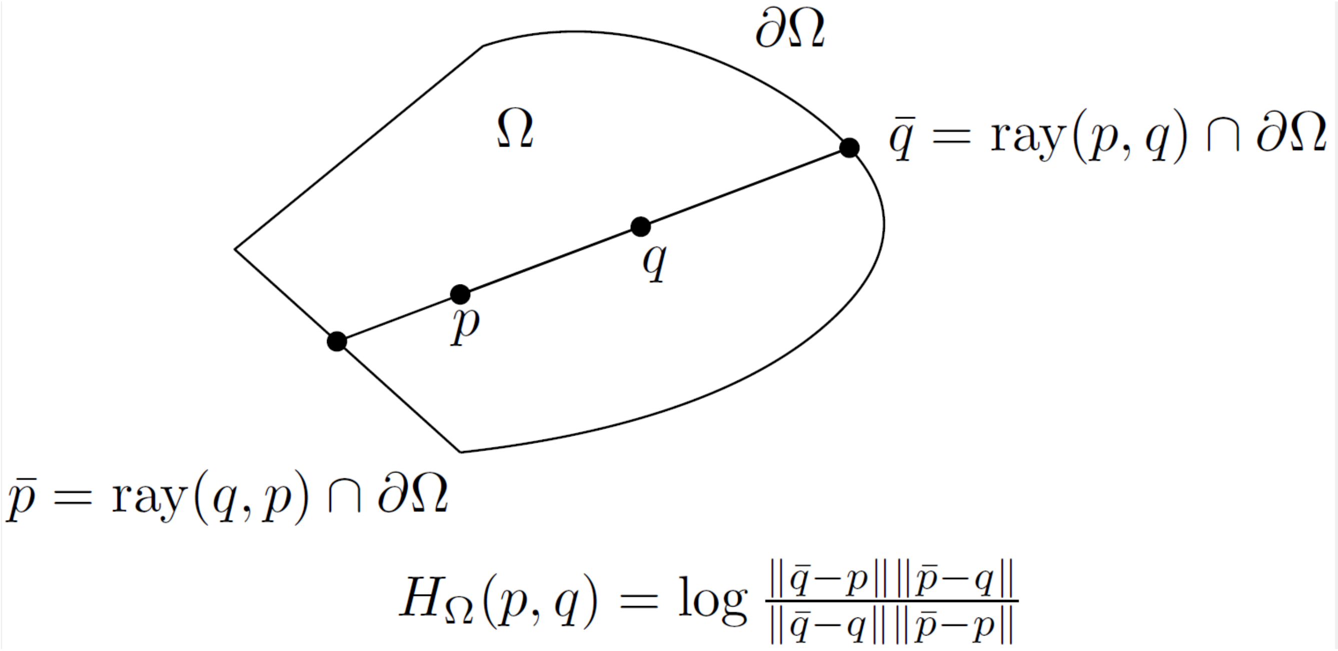
- ; to indicate a symmetric but non-metric distance: $D(p;q)$

Example: Mutual information

- **Bracket []** to indicate a statistical distance: $D[p:q]$
- For a parametric family P, 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)

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

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

Clustering in Hilbert simplex geometry, arXiv:1704.00454 (2017)

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

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

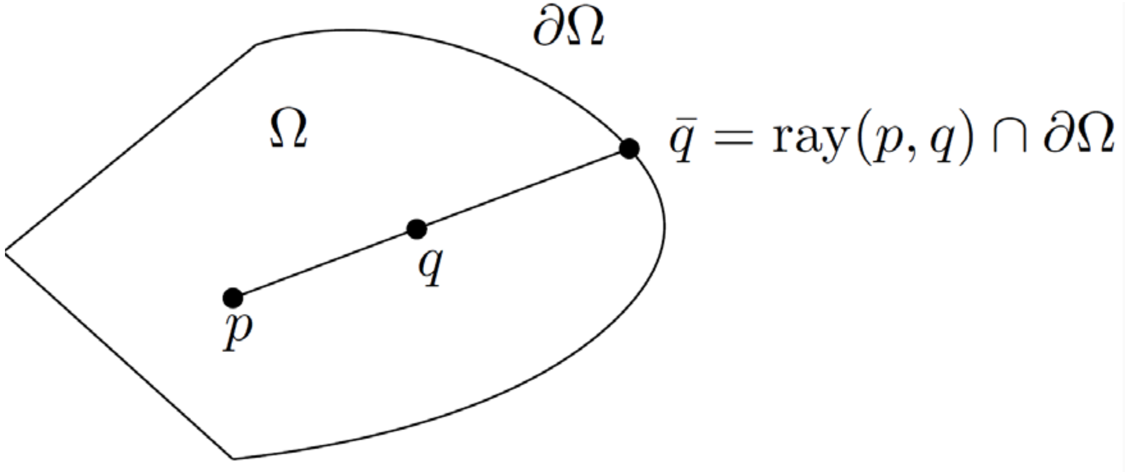
where $\theta_i(S_1, S_2) = \arccos \lambda_i(S_1^\top S_2)$ is the i-th principal angle and $\lambda_i(X)$ denotes the i-th largest eigenvalue of matrix X. We have $\delta_S(S_1, S_2) = 0$ whenever S_1 is a subspace of S_2 (an asymmetric property).

Schubert varieties and distances between subspaces of different dimensions

Failing symmetry: E.g., Funk oriented distance

Hilbert cross-ratio metric is the arithmetic symmetrization of **Funk distances**

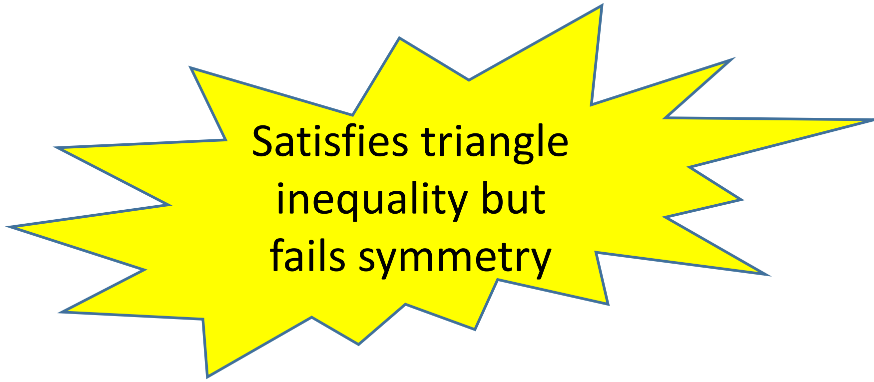
$$H_\Omega(p_1, p_2) = \frac{F_\Omega(p_1, p_2) + F_\Omega^r(p_1, p_2)}{2}$$



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

Reverse distance or **dual distance** (reference duality)

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

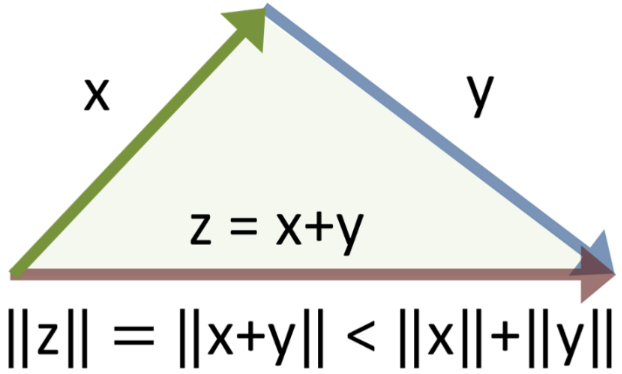


Related to **Finsler geometry** that extends Riemannian geometry with a Finsler metric (norm)

Medians and means in Finsler geometry, arXiv:1011.6076

A family of statistical symmetric divergences based on Jensen's inequality, arXiv:1009.4004

Failing triangle inequality/subadditivity:



- Example: Kullback-Leibler divergence between two pmfs:

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

- Notice that the squared Euclidean distance fails the triangle inequality

Clustering in Hilbert simplex geometry, arXiv:1704.00454

Scale-invariant distances

Fumitada Itakura

- Itakura-Saito divergence:

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

- **Scale-invariance** property:

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

- Often used in music applications (spectrum)

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

- Distance *independent* of both argument scaling factors
- C 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 projective distance**:

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

On Hölder projective divergences, Entropy 19 (3), 2017

Statistical distance: Total Variation (TV) metric

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

- 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

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

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

- Synonyms: city block distance, overlap distance, etc.

Kolmogorov metric distance

- A distance between **distribution functions**, less than TV:

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

Related to Kolmogorov–Smirnov test

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

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

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)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int(\sqrt{p(x)}-\sqrt{q(x)})^2\mathrm{d}\nu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-\lambda p(x))^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-\lambda p(x)\rvert^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\nu(x)$	$u\log u$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x)q^{1+\alpha}(x)\mathrm{d}\nu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 2013

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 $(\mathscr{X}, \mathscr{F})$ onto a measure space $(\mathscr{Y}, \mathscr{G})$. Then we should have

$$d(P_1, P_2) \geqslant d(P_1 t^{-1}, P_2 t^{-1}).$$

Coarser sigma-algebra

More distinguishability of stochastic processes

Here $P_i t^{-1}$ denotes the induced measure on $\mathscr{Y}$ corresponding to P_i.

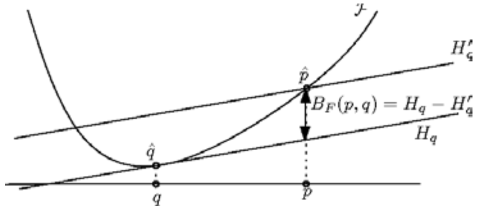
$$d(P_1^{(m)}, P_2^{(m)}) \leqslant d(P_1^{(n)}, P_2^{(n)}) \quad \text{for} \quad m < n. \qquad t(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_m).$$

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

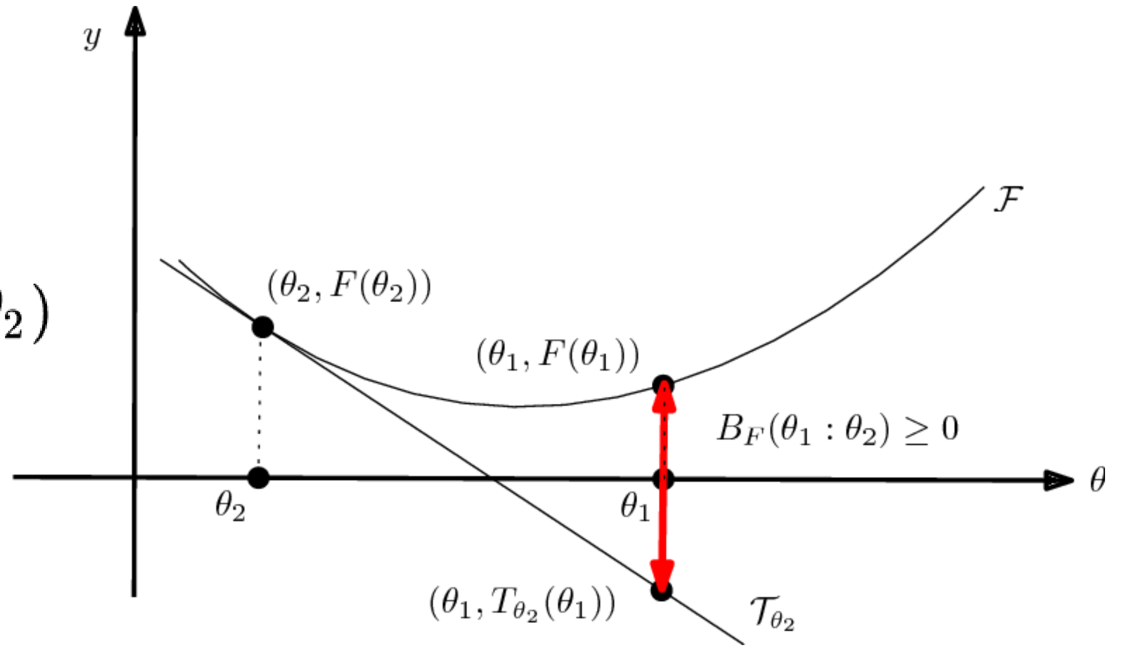
$$d(P_{\theta_1}, P_{\theta_2}) \leqslant d(P_{\theta_1}, P_{\theta_3}).$$

Classes of distances: Bregman divergence



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

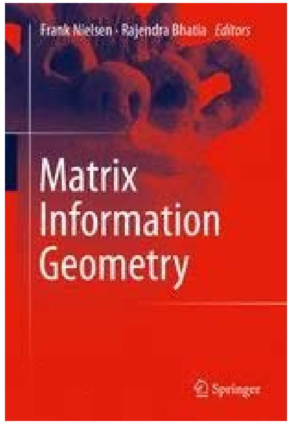
$$B_F\left(\theta_1 : \theta_2\right) := F\left(\theta_1\right) - F\left(\theta_2\right) - \left(\theta_1 - \theta_2\right)^\top \nabla F\left(\theta_2\right)$$



- 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.

Matrix Bregman divergences



For <u>real symmetric matrices</u>:

$$B_F(L:N) = F(L) - F(N) - \mathrm{tr}\left((L-N)\nabla_F^\top(N)\right)$$

where F is a strictly convex and differentiable generator $F : \mathrm{Sym}(d,d) \rightarrow \mathbb{R}$

- Squared Froebenius distance for $F(X) = \|X\|_F^2$
- von Neumann divergence for $F(X) = \mathrm{tr}(X \log X - X)$

$$D_{\mathrm{vN}}(X, : Y) = \mathrm{tr}(X \log X - X \log Y - X + Y)$$

- Log-det divergence for $F(X) = -\log\det(X)$

$$D_{\mathrm{ld}}(X:Y) = \mathrm{tr}\left(XY^{-1}\right) - \log\det\left(XY^{-1}\right) - n$$

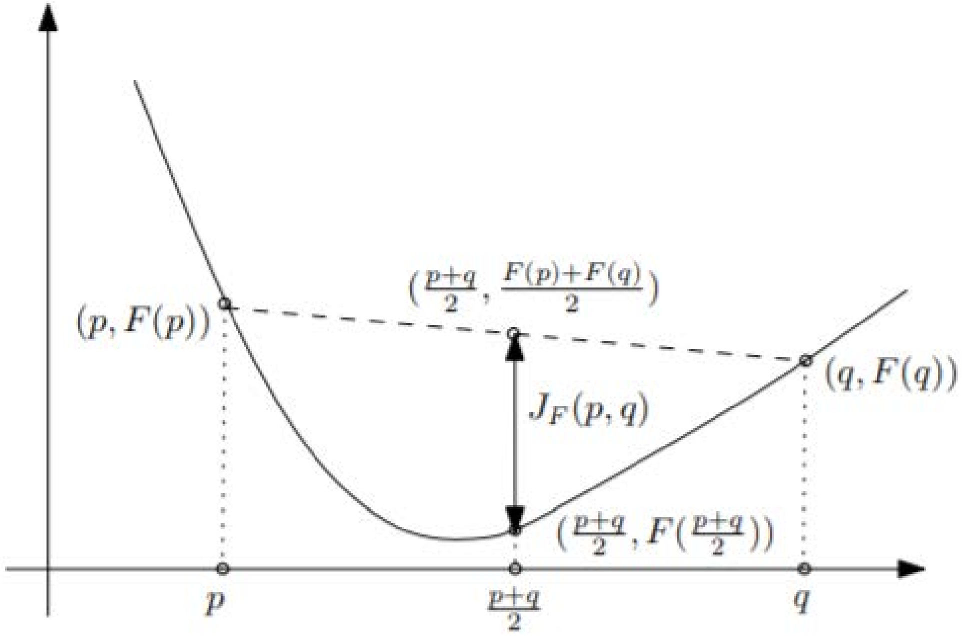
Bregman–Schatten p-divergences…

Mining Matrix Data with Bregman Matrix Divergences for Portfolio Selection, 2013

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

A family of statistical symmetric divergences based on Jensen's inequality, arXiv:1009.4004

Statistical divergences amount to parameter divergences for exponential families:

Statistical distances | Parameter divergences

$$\mathrm{Bhat}_\alpha(p:q) = -\log \int p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}x \longleftrightarrow J_F^\alpha(\theta_p:\theta_q) = (F(\theta_p)F(\theta_q))_\alpha - F((\theta_p\theta_q)_\alpha)$$

$$\lim_{\alpha\rightarrow 0+} \frac{1}{\alpha}\mathrm{Bhat}_\alpha(p:q) \qquad \lim_{\alpha\rightarrow 0+} \frac{1}{\alpha}\mathrm{J}_F^\alpha(p:q)$$

$$p(x) = p(x;\theta_p)$$

$$q(x) = p(x;\theta_q)$$

$$\mathrm{KL}(p:q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x \longleftrightarrow B_F(\theta_q:\theta_p) = F(\theta_q) - F(\theta_p) - (\theta_q-\theta_p)^\top \nabla F(\theta_q)$$

Generic distributions | Exponential families

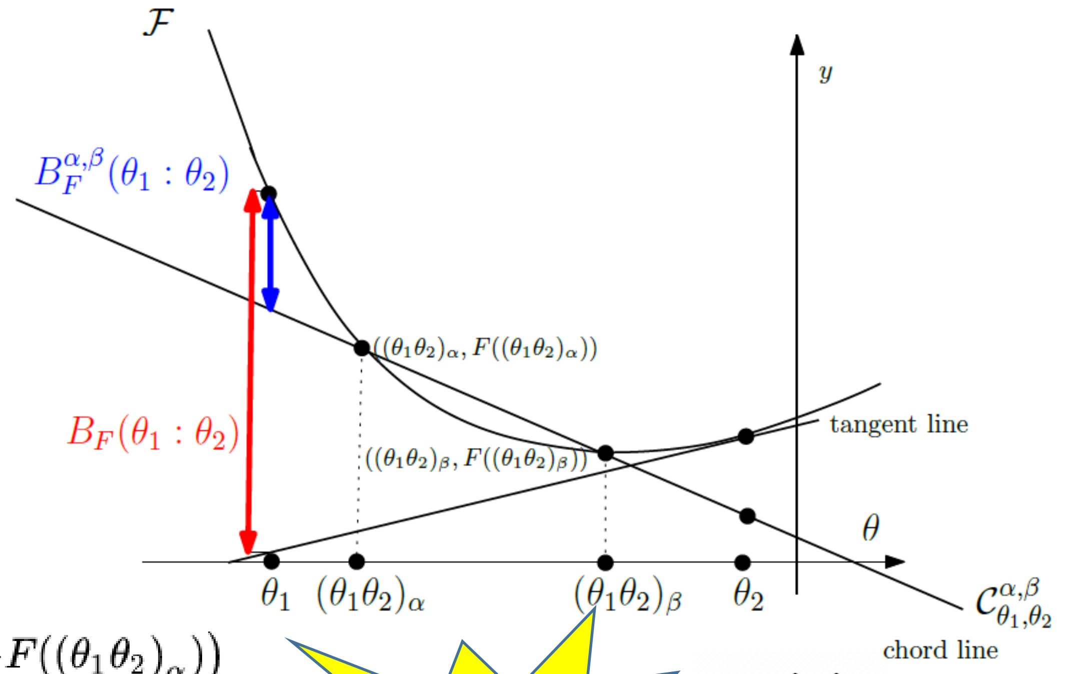
$$p(x;\theta) = \exp(\theta^\top x - F(\theta))$$

The Burbea-Bao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57(8), 2011

Bregman chord divergence: Free of gradient!

Ordinary Bregman divergence requires gradient calculation:

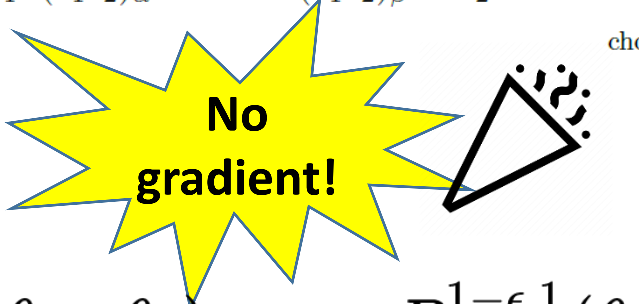
$$B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)$$



Bregman chord divergence uses two extra scalars α and β:

$$B_F^{\alpha,\beta}(\theta_1 : \theta_2) = F(\theta_1) - F((\theta_1\theta_2)_\alpha) - \frac{\alpha(F((\theta_1\theta_2)_\beta) - F((\theta_1\theta_2)_\alpha))}{\beta-\alpha}$$

Using linear interpolation notation $(\theta_1\theta_2)_\alpha = (1-\alpha)\theta_1 + \alpha\theta_2$

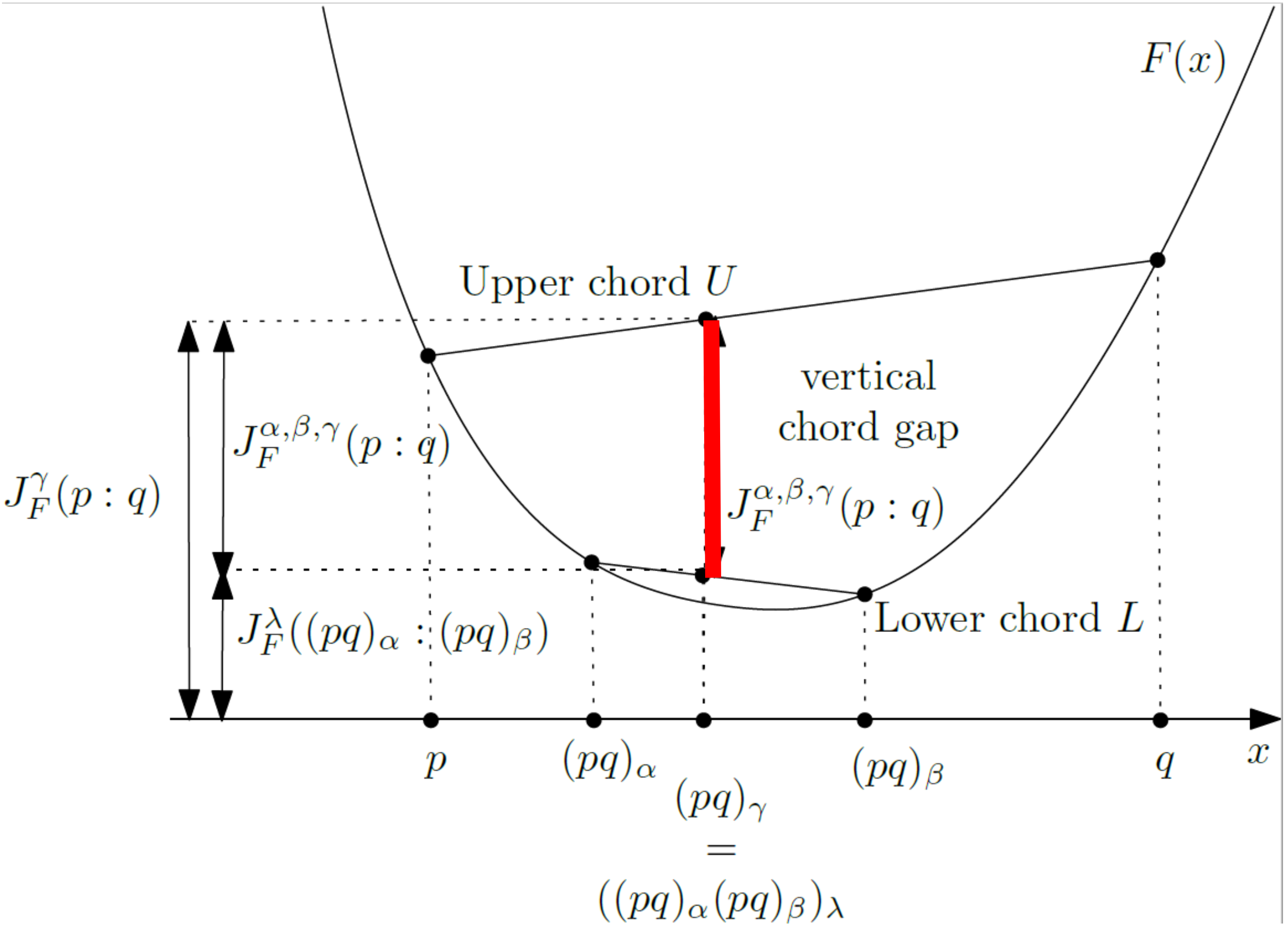


$$\lim_{\beta\to\alpha} B_F^{\alpha,\beta}(\theta_1 : \theta_2) = B_F^{\alpha}(\theta_1 : \theta_2) \quad \text{and} \quad B_F(\theta_1 : \theta_2) \simeq_{\epsilon\to 0} B_F^{1-\epsilon,1}(\theta_1 : \theta_2)$$

Subfamily of **Bregman tangent divergences**: $B_F^{\alpha}(\theta_1 : \theta_2) = F(\theta_1) - F((\theta_1\theta_2)_\alpha) - \alpha(\theta_1 - \theta_2)^\top \nabla F((\theta_1\theta_2)_\alpha)$

The Bregman chord divergence, arXiv:1810.09113

The Jensen chord divergence: Truncated skew Jensen divergences



Linear interpolation (LERP):

$$(pq)_\lambda := (1-\lambda)p + \lambda q$$

$$J_F^{\alpha,\beta,\gamma}(p:q) = (F(p)F(q))_\gamma - (F((pq)_\alpha)F((pq)_\beta))_\lambda$$

$$((pq)_\alpha(pq)_\beta)_\lambda = (pq)_\gamma \text{ with } \gamma \in (\alpha,\beta)$$

$$\boxed{J_F^{\alpha,\beta,\gamma}(p:q) = (F(p)F(q))_\gamma - (F((pq)_\alpha)F((pq)_\beta))_{\frac{\gamma-\alpha}{\beta-\alpha}}}$$

$$J_F^{\beta,\gamma}(p:q) = \gamma\left(\left(\frac{1}{\beta}-1\right)F(p) + F(q) - \frac{1}{\beta}F((pq)_\beta)\right)$$

A property: $J_F^{\alpha,\beta,\gamma}(p:q) = J_F^{\gamma}(p:q) - J_F^{\lambda}((pq)_\alpha:(pq)_\beta)$

(truncated skew Jensen divergence)

The chord gap divergence and a generalization of the Bhattacharyya distance, ICASSP 2018

Summary

- Distance measures the **separation of (same type) entities**

  (vectors, probability measures, probability densities,

  cumulative distribution functions, random variables, matrices, functions, etc.)
- A **metric (distance)** is a symmetric non-negative distance (dissimilarity) that satisfies both the law of the indiscernibles and the triangle inequality
- A **divergence** originally meant a *statistical distance* (eg., probability metric), and also means a *smooth parametric distance* in information geometry
- Statistical divergences between densities of a same parametric family amount to parameter divergences
- Three classes of **non-mutually exclusive parametric distances**:

  The Csiszar f-divergences, Bregman divergences, and Jensen divergences, that are **non-mutually exclusive**
- But also **Wasserstein distance** in optimal transport (ground distance?), etc.

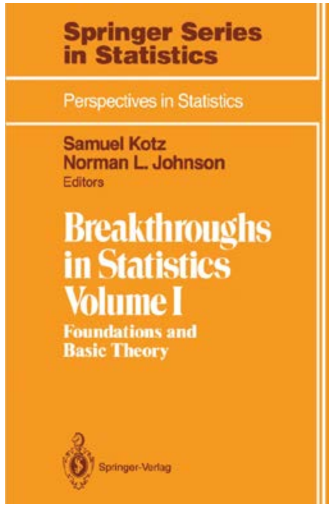
Information-geometric structures:

- Fisher-Rao geometry
- Dualistic information-geometric structures
- Bregman manifolds and information projections
- Mixture family manifolds and exponential family manifolds

Frank Nielsen

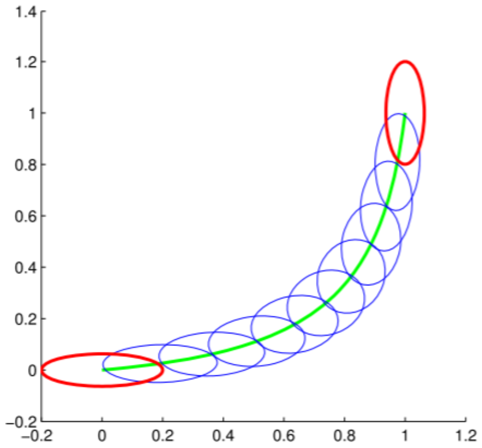
Sony CSL

Fisher-Rao Riemannian geometry



Frank Nielsen

Sony CSL



Recalling the Fisher information metric…

covariant to reparameterization of θ

- *Fisher Information Metric* (FIM):

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

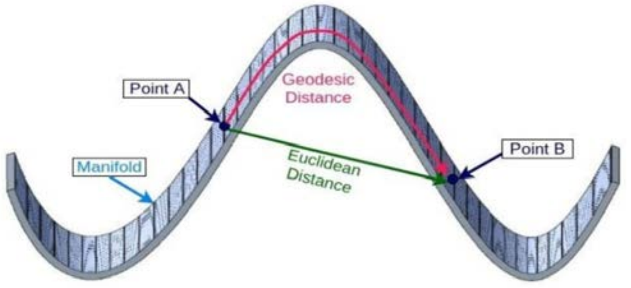
- Infinitesimally, the KLD is related to the FIM via:

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

This is a **squared Mahalanobis distance**

This Taylor' expansion holds for any **standard f-divergence** **(f''(1)=1)**

Rao distance is Riemannian geodesic distance



independent to reparameterization of θ

- Infinitesimal length element :

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

- Geodesic and distance are hard to explicitly calculate :

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

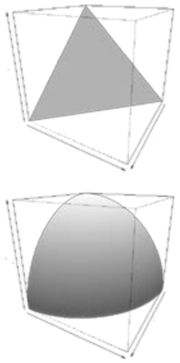
Riemannian geodesics locally minimize lengths

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

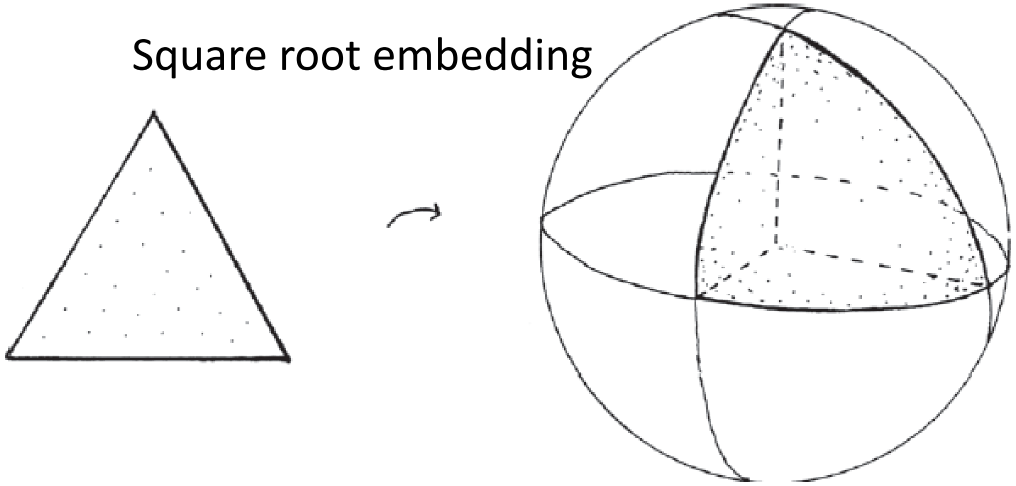
C. R. Rao with Sir R. Fisher in 1956

STATISTICAL DATA ANALYSIS AND INFERENCE edited by Yadolah DODGE, 1989

Fisher-Rao geometry: Standard simplex (categorical distribution)



- Trinomial (trinoulli)



Embedding to the sphere positive orthant

Fisher information metric:

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

(Hotelling)-Fisher-Rao distance:

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

Pattern Learning and Recognition on Statistical Manifolds: An Information-Geometric Review, SIMBAD 2013
Clustering in Hilbert simplex geometry, arXiv:1704.00454

In practice, calculating Rao's distance is difficult

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

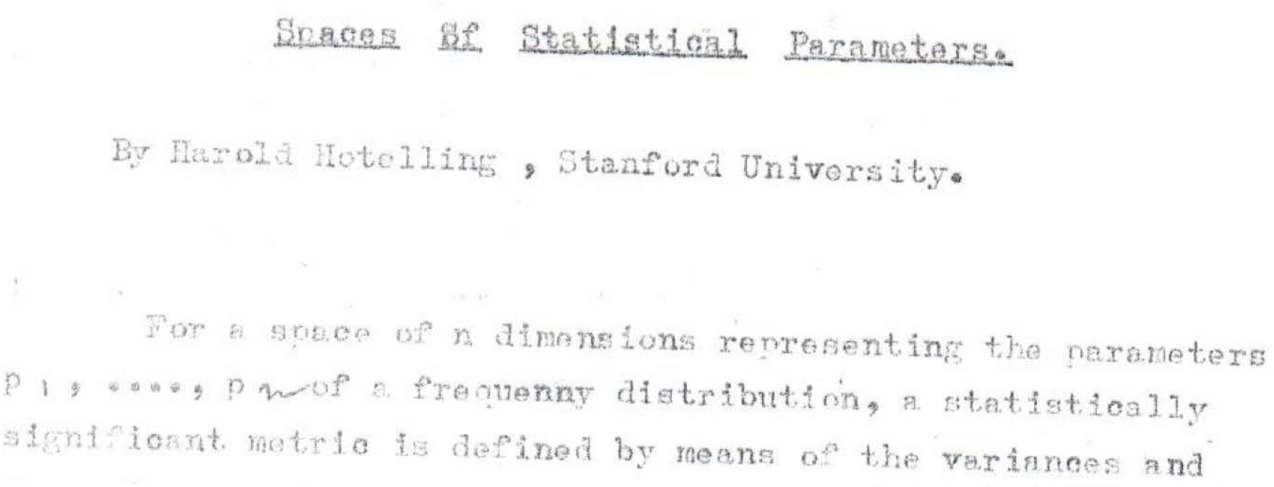
1. Need to solve the Ordinary Differential Equation (ODE) for find the geodesic:

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

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

2. Need to integrate the infinitesimal length elements along the geodesics…

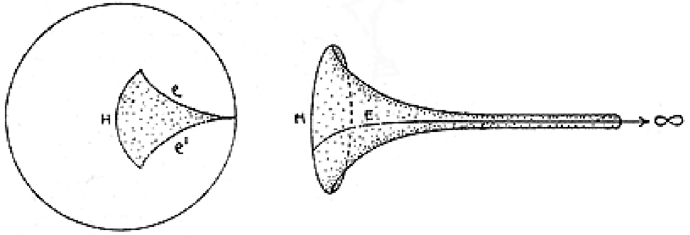
Hotelling's 1930 paper considered location-scale families!


Spaces of Statistical Parameters.

By Harold Hotelling, Stanford University.

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

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



Harold Hotelling

- 2D FIM
- **Constant (non-positive) curvature**, isometric to hyperbolic geometry of curvature

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

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

Some common Fisher-Rao geodesic distances

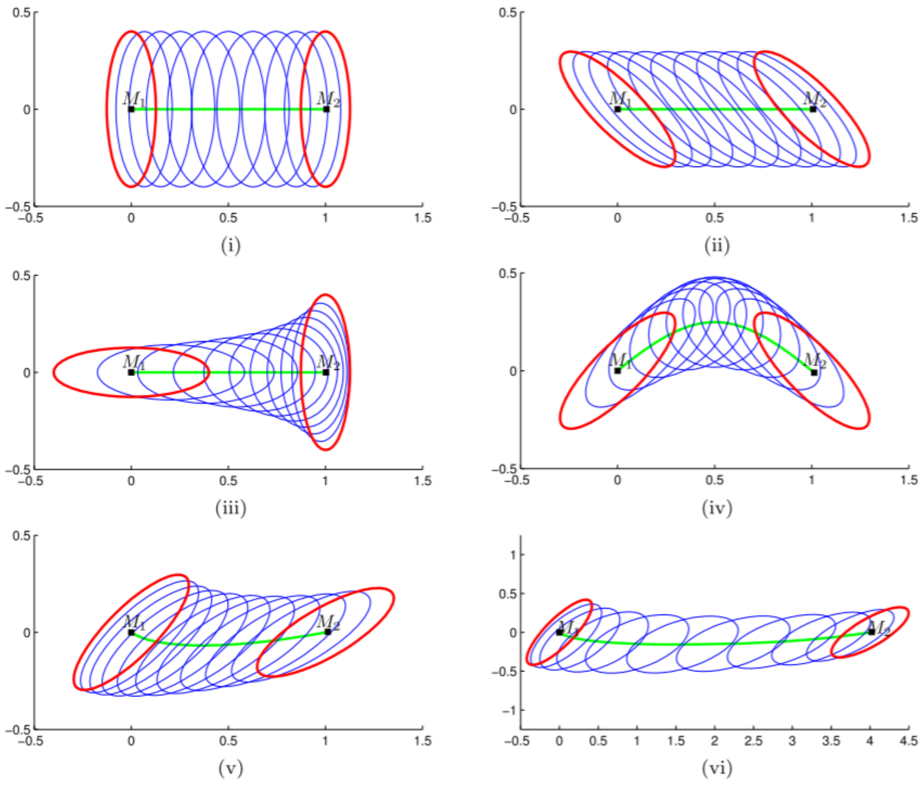
Distribution	Density	Geodesic Distance
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$2\sqrt{n}\,\lvert \arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2})\rvert$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$2\,\lvert\sqrt{\lambda_1} - \sqrt{\lambda_2}\rvert$
Geometric	$(1-p)\,p^x$	$2\log\frac{1-\sqrt{p_1p_2}+\lvert\sqrt{p_1}-\sqrt{p_2}\rvert}{\sqrt{(1-p_1)(1-p_2)}}$
Gamma	$\frac{e^{-\theta x}\theta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}$	$\sqrt{\alpha}\,\lvert\log\theta_1 - \log\theta_2\rvert$
Normal (fixed variance)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\frac{\lvert\mu_1-\mu_2\rvert}{\sigma}$
Normal (fixed mean)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\sqrt{2}\,\lvert\log\sigma_1 - \log\sigma_2\rvert$
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)	$\frac{e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}}{(2\pi)^{p/2}\lvert\Sigma\rvert^{1/2}}$	$(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$
p-Variate Normal (μ fixed)	$\frac{e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}}{(2\pi)^{p/2}\lvert\Sigma\rvert^{1/2}}$	$\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{p}\log\lambda_i^2}$
	(here, $\{\lambda_i\}$ are the	roots of $\lvert\Sigma_2 - \lambda\Sigma_1\rvert = 0$)
Multinomial	$\frac{n!}{\prod_{i=1}^k n_i!}p_i^{n_i}$	$2\sqrt{\pi}\arccos(\sum_{i=1}^k\sqrt{p_i\theta_i})$

Anirban DasGupta, Probability for Statistics and Machine Learning

Approximating geodesics for multivariate normal via geodesic shooting

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

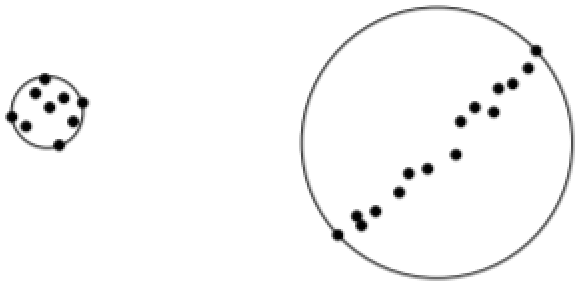
Given: Initial point $P_0 = (\mu_0, \Sigma_0)$, final point $P_1 = (\mu_1, \Sigma_1)$.
Output: Minimal geodesic $P(t) = (\mu(t), \Sigma(t))$, $t \in [0, 1]$, such that $P(1) = (\mu_1, \Sigma_1)$.
Initialization: Choose initial velocities $V(0) = (\dot\mu(0), \dot\Sigma(0))$ (e.g., zeroes), initial values for ϵ (10^-5), error $= 10^6$.
while 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 $= \|W(1)\|_{P_1} = \sqrt{W_\mu(1)^T \Sigma_1^{-1} W_\mu(1) + \frac{1}{2}\mathrm{tr}((\Sigma_1^{-1} W_\Sigma(1))^2)}$;
  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),
  $J(1) = \frac{\exp_{P_0}(V(0)+\alpha W(0)) - \exp_{P_0}(V(0))}{\alpha}$, where α is sufficiently small value and we use $\frac{\epsilon}{\|W(0)\|_{P_0}}$
  Determine proper update size s:
  $s_1 = \frac{\langle W(1), J(1)\rangle_{P(1)}}{\|J(1)\|^2_{P(1)}}$
  if $\|W(1)\|_{P(1)} > 0.05$ **then**
    $s = 0.05/\|W(1)\|_{P(1)} s_1$;
  else
    $s = s_1$;
  end if
  $V(0) \leftarrow V(0) + sW(0)$;
end while



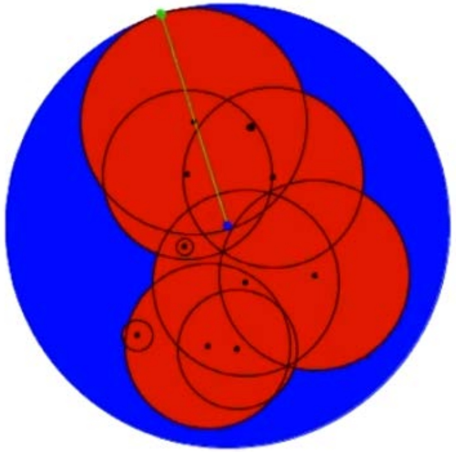
Minyeon Han · F.C. Park, DTI Segmentation and Fiber Tracking Using Metrics on Multivariate Normal Distributions, 2014
Calvo, Miquel, and Josep Maria Oller. "An explicit solution of information geodesic equations for the multivariate normal model." *Statistics & Risk Modeling* 9.1-2 (1991): 119-138.

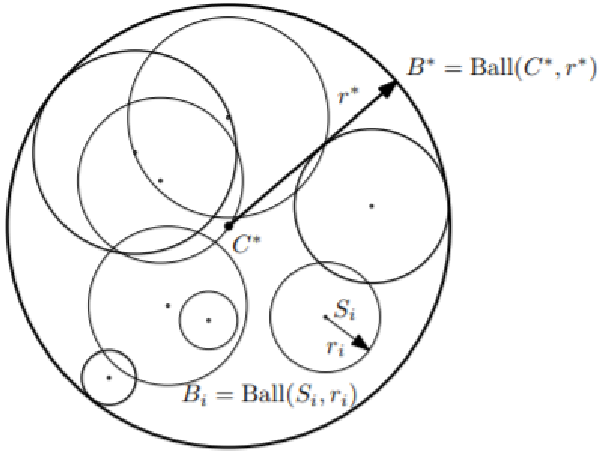
Approximating the smallest enclosing ball

- Iterative algorithm that yields a **core-set**
- **Extends to balls, etc.**
- Useful for **k-center clustering**.



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





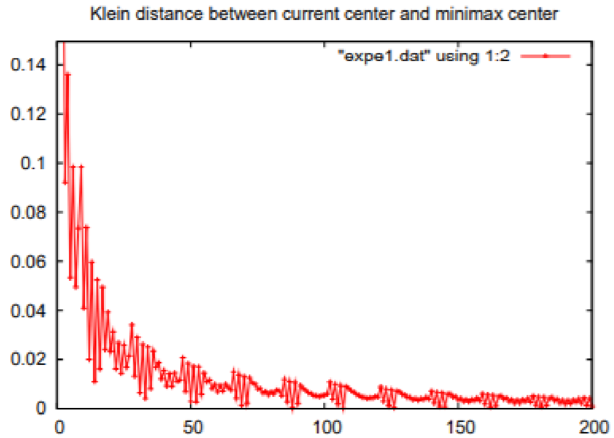
Approximating smallest enclosing balls with applications to machine learning, IJCGA, 2009

Riemannian minimum enclosing ball

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

Algorithm GeoA

$c_1 \leftarrow$ 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` $[c_i, p_{s_i}]$
  $c_{i+1} \leftarrow c_i \#^M_{\frac{1}{i+1}} p_{s_i}$;
end
`// Return the SEB approximation`
return $\mathrm{Ball}(c_l, r_l = \rho(c_l, \mathcal{P}))$;



Hyperbolic geometry:

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

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

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

Positive-definite matrices:

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

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

On Approximating the Riemannian 1-Center, Comp. Geom. 2013
Approximating Covering and Minimum Enclosing Balls in Hyperbolic Geometry, GSI, 2015

f-divergence between isotropic Gaussians: = monotic increasing function of Mahalanobis Smallest enclosing ball same for <u>all f-divergences</u>…

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 $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Mahalanobis's generalized distance is α^2, where

$$\alpha^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

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

$$y = (\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/\alpha$$

and so reducing the problem to that of the divergence of a $N(\alpha, 1)$ distribution from a $N(0, 1)$. The family $\{N(\alpha, 1): \alpha \geqslant 0\}$ 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.

From Ali and Silvey'66

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:

$$\sum_1^n \sum_1^n K(x_i, x_j) a_i a_j \leq 0, \qquad \text{for all } x_1, \cdots, x_n \in \mathscr{X}$$
$$a_1 + \ldots + a_n = 0$$

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

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

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: Hotelling-Fisher-Rao geometry

- 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
- FIM properties: statistical invariance by a 1-to-1 transformation of the sample space X
- Geodesic length invariant by reparameterization of the parameter space θ
- The Fisher-Rao distance is the Riemannian metric distance

  = **geodesic distance**
- Difficult to calculate/approximate, even for the multivariate normal family:
  a. Explicit geodesic calculation
  b. Integration of infinitesimal length elements on the geodesics

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

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.

C.R. Rao

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 human body. All of us had all education from primary school to college in India. None of us has any foreign

Dualistic structures of information geometry

Shun-ichi Amari

Frank Nielsen

Sony Computer Science Laboratories, Inc

Shinto Eguchi

Sony CSL

An elementary introduction to information geometry https://arxiv.org/abs/1808.08271

Covariant derivative ∇

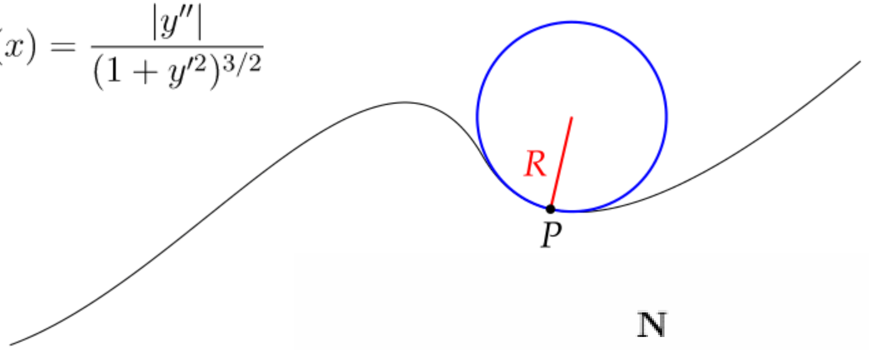
$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$$

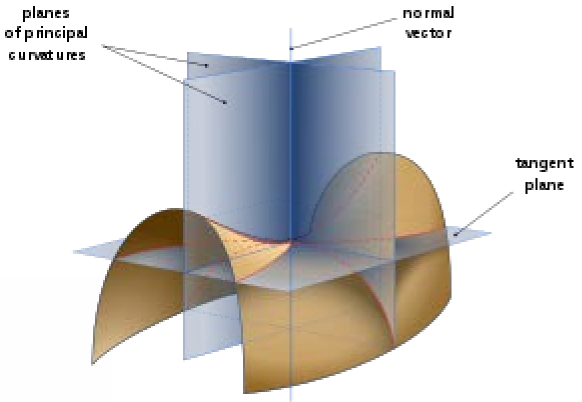
- calculate **differentials** of a vector field Y with respect to another vector field X: Namely, the covariant derivative $\nabla_X Y := \nabla(X, Y)$
- Defined by prescribing a dimension cubic number of smooth functions: The **Christoffel symbols** $\Gamma_{ij}^k = \Gamma_{ij}^k(p)$
- In local coordinates 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

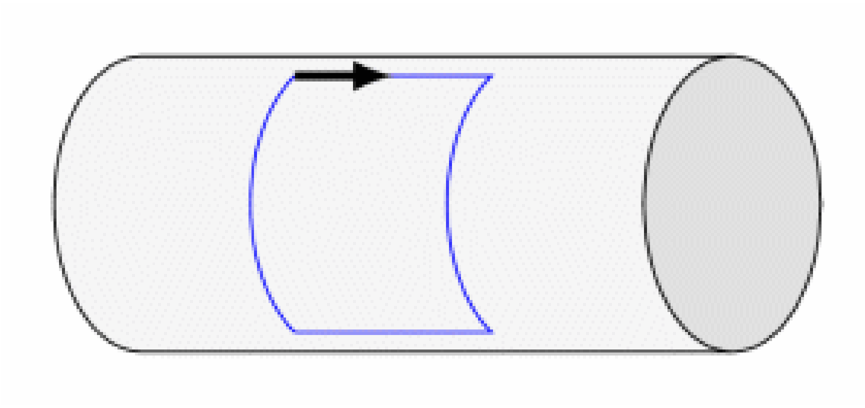
$$(\nabla_X Y)^k \stackrel{\Sigma}{=} X^i (\nabla_i Y)^k \stackrel{\Sigma}{=} X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma_{ij}^k Y^j \right)$$

Curvature of ∇

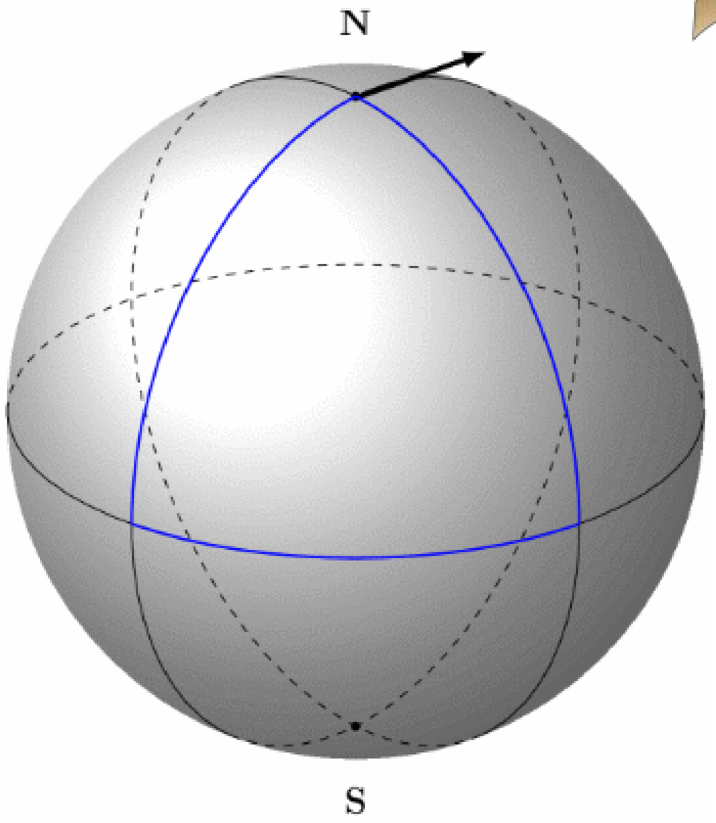
$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$







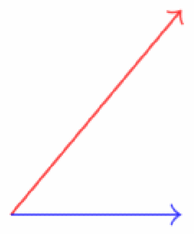
Cylinder is flat
Parallel transport is independent of path



Sphere has constant curvature
Parallel transport is path-dependent

Curvature/torsion of an affine connection ∇

parallel transport "twists" vectors.



- **Curvature tensor** (or Riemann-Christoffel RC curvature)

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

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

- Connection is said **flat** when R=0
- Symmetric connection: $\nabla_X Y - \nabla_Y X = [X,Y]$

  In local coordinates: $\Gamma^k_{ij} = \Gamma^k_{ji}$

- (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

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

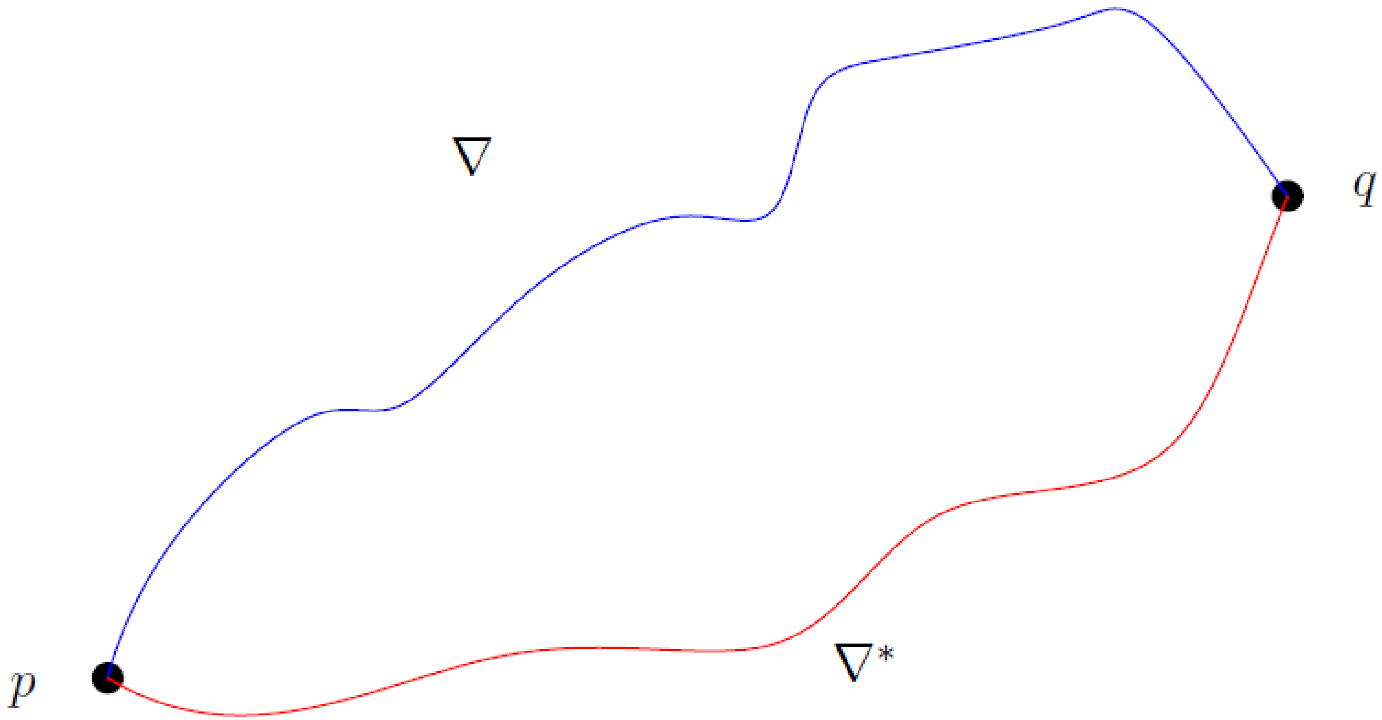
$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$$

- 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. <u>Unique dual torsion-free affine connection ∇*</u>

- **Involution**: $(\nabla^*)^* = \nabla$ $\Rightarrow$ (M, g, ∇, ∇^*)

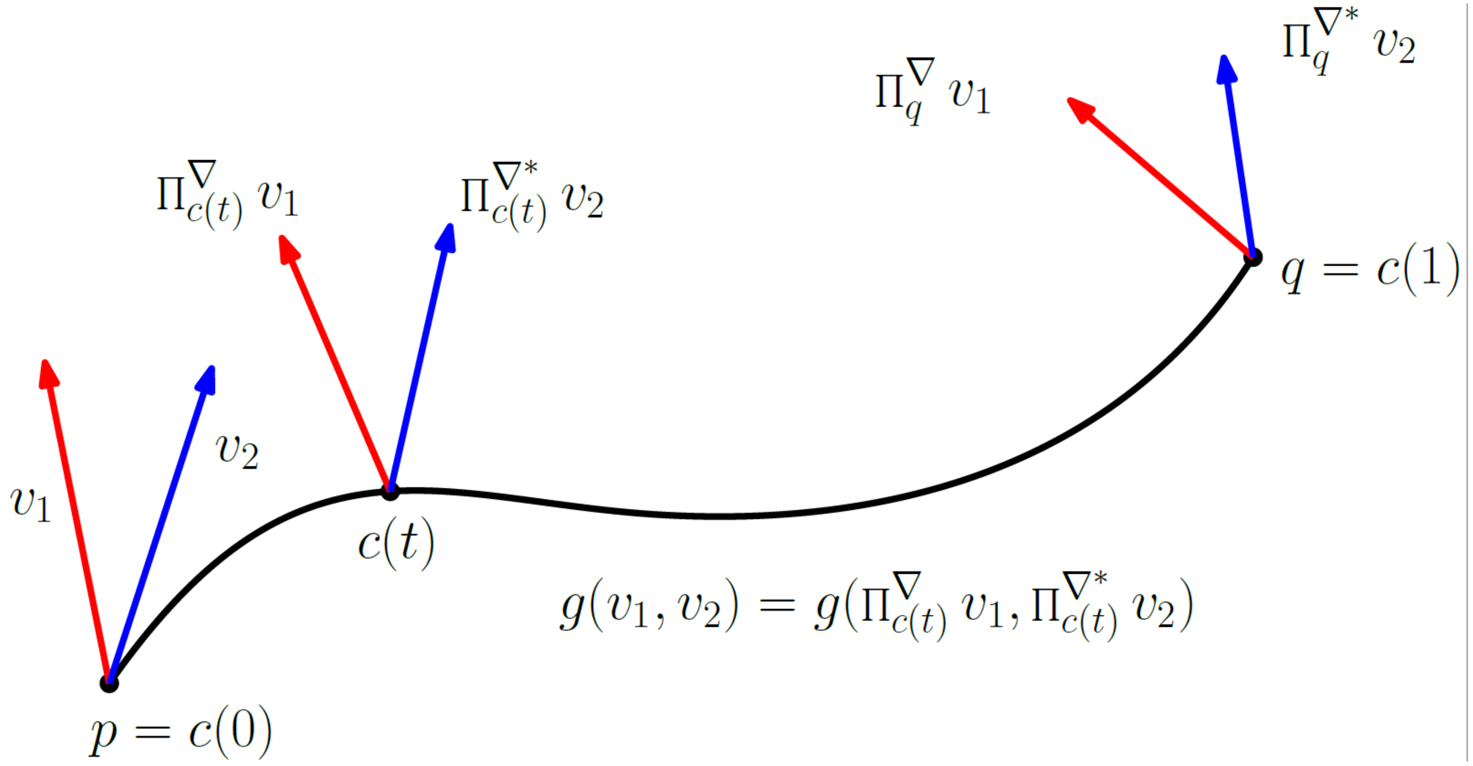
Dual ∇-geodesic and ∇*-geodesic

With respect to the metric tensor



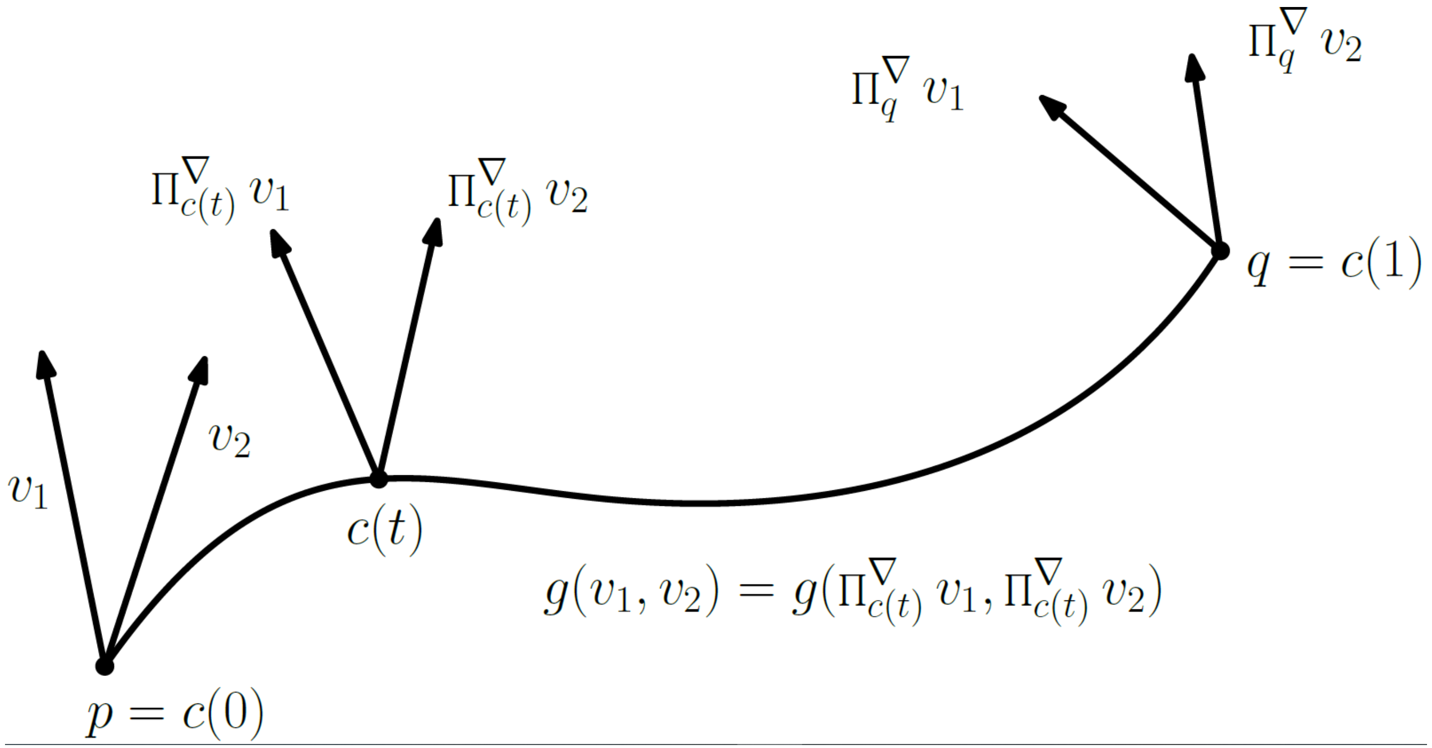
Property: Dual parallel transport of vectors preserves the metric

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



Metric Levi-Civita connection from averaging dual connections

$$\bar{\nabla} = \frac{\nabla + \nabla^*}{2} \qquad \bar{\nabla} = {}^{\mathrm{LC}}\nabla \qquad \bar{\Gamma}$$



Statistical manifolds: Cubic tensor (M, g, C)

Apply also to non-statistical contexts!

Dualistic structure with metric tensor g and cubic tensor C

Steffen Lauritzen (1987)

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

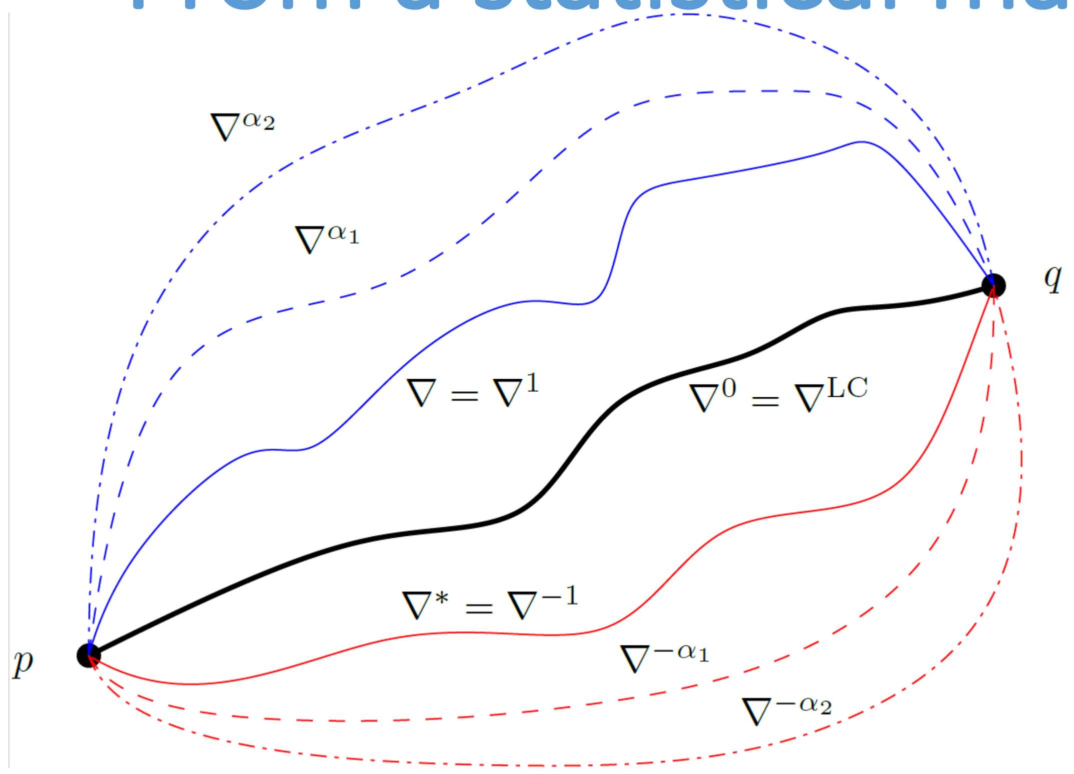
$$C_{ijk} := \Gamma_{ij}^k - {\Gamma^*}_{ij}^k$$ (local coordinates)

In a local basis:

$$C_{ijk} = C(\partial_i, \partial_j, \partial_k) = \langle \nabla_{\partial_i} \partial_j - \nabla^*_{\partial_i} \partial_j, \partial_k \rangle$$

… totally symmetric (=components invariant by index permutation)

From a statistical manifold to a 1-family of structures



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

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

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

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

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

$$C \Rightarrow \alpha C \quad \Rightarrow \quad (M, g, \alpha C)$$

The fundamental theorem of information geometry

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

Case K=0

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

Case K=0

A manifold $(M, g, \nabla, \nabla^{*})$ 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

  Linked to parametric family of densities/manifolds

- Then Eguchi showed how to define dual connections from any smooth parameter distances called divergences (originally, called contrast functions). From that, we get a 1-family of α-connections

Definition of a parameter divergence

Definition (Divergence) *A divergence* $D : M \times M \to [0, \infty)$ *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.*

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

Statistical divergence (deviance) like the Kullback-Leibler divergence

versus

Parameter divergence as a synonym for a contrast function

Statistical manifolds from divergences

- Reverse/dual parameter divergence (reference duality)

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

- Statistical manifold structures:

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

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

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

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

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

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

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

Statistical manifolds from Bregman divergences

Bregman divergence (1967, on Operations research):

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

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

Dual Bregman divergence and Legendre-Fenchel transformation F*

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

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

Described later on,
In Bregman Hessian
manifolds

Expected α-geometry for a parametric model

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

- Use Fisher information metric (FIM)
- Define the expected α-connections:
- **Amari-Chentsov cubic tensor**

$$C_{ijk} := E_\theta\left[\partial_i l \partial_j l \partial_k l\right]$$

$$l(\theta;x) := \log L(\theta;x) = \log p_\theta(x)$$

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

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 $t(x) = (t_1(x), \ldots, t_D(x))$, 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:

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

Example 2 (FIM of a mixture family $\mathcal{M}$) *A mixture family is defined for $D+1$ functions $F_1, \ldots, F_D$ 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 $\int F_i(x)\mathrm{d}\mu(x) = 0$. Function C is such that $\int C(x)\mathrm{d}\mu(x) = 1$. Mixture families include statistical mixtures with prescribed component distributions and the probability simplex Δ. The FIM of a mixture family is given by:

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

Monte Carlo information geometry: The dually flat case, arXiv:1803.07225

Exponential e-connection and mixture m-connection: An example of dually flat connections wrt. FIM

- 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 too.

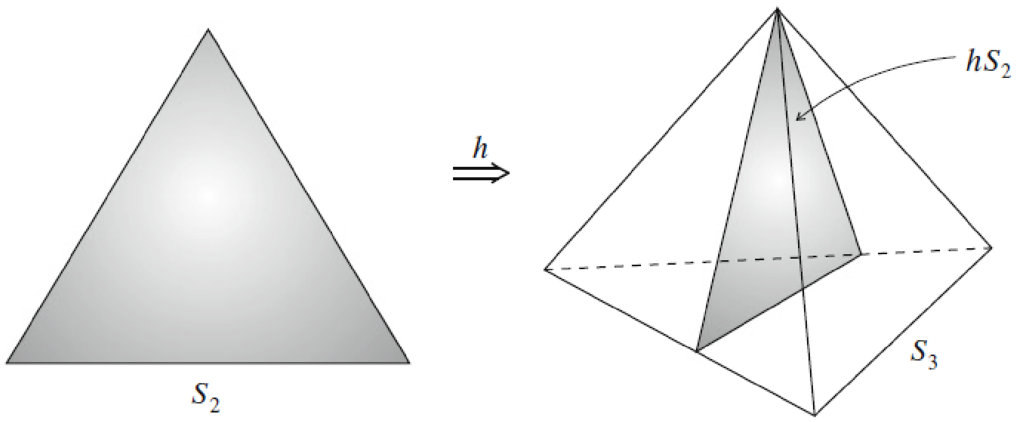
- For a **mixture family, the m-connection is flat.** Then by using the fundamental theorem of information geometry, we have the dual e-connection flat too.

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).



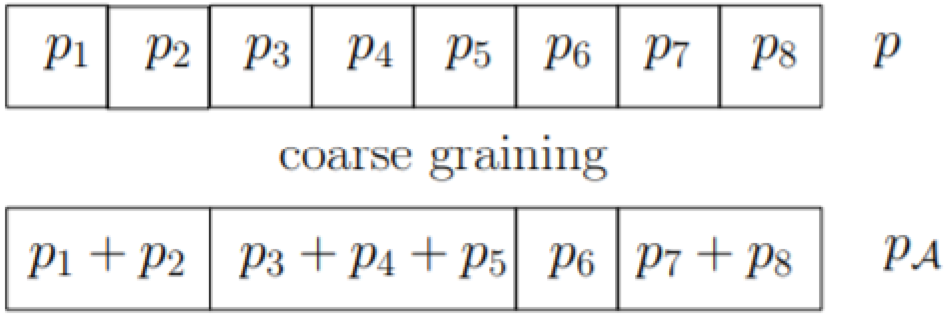
Embedding of S_2 in S_3 (m = 2, n = 3)

- L. Lorne Campbell. An extended Cencov characterization of the information metric. ˇ 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 divergences

- **Information monotonicity** of parameter divergences:

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



Invariant divergence

Markov embeddings, Markov kernels, etc.

Statistical invariance: Csiszar/Ali-Silvey 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)
- Theorem: **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 ensuring the law of the indiscernibles)
- Choose f(1)=0 (for lower bound of f-divergence 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 any standard f-divergence corresponds to the expected α-connections for**

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

Summary

- Geometry of parametric families of distributions:
  - Fisher Riemannian geometry (Levi-Civita connections)
  - α-expected geometry (Conjugate/dual connections)
  - Statistical invariance

- Expected α-geometry vs 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

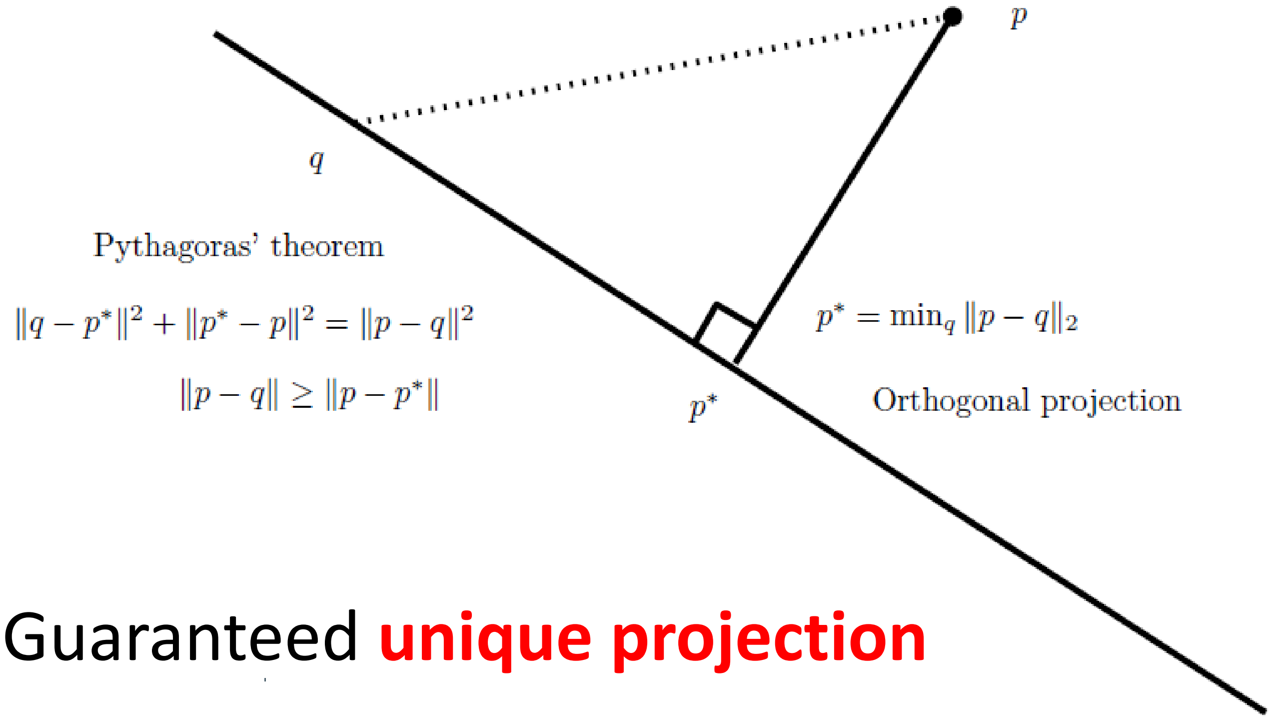
Frank Nielsen

Sony CSL

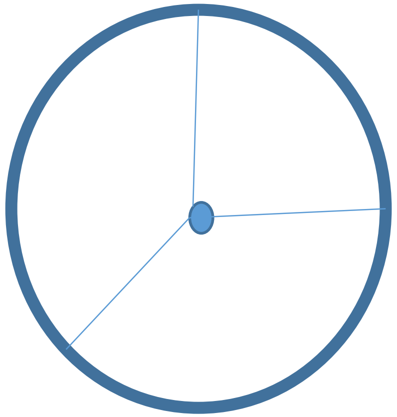
Recalling Euclidean geometry….

Distance, geodesic, orthogonality, uniqueness of projection

Projection, orthogonality and Pythagoras' theorem



Guaranteed **unique projection**



Non-unique projection

Goal: Provide geometric interpretations of MLE/MaxEnt of KL divergence minimizations as information projections

MaxEnt (with prior q)

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

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

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

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

Maximum Likelihood Estimate

$$\begin{aligned}
\min \quad & \mathrm{KL}(p_e(x) : \boxed{p_\theta(x)}) \\
&= \int p_e(x) \log p_e(x) \mathrm{d}x - \int p_e(x) \log p_\theta(x) \mathrm{d}x \\
&= \min -H(p_e) - \underbrace{E_{p_e}[\log p_\theta(x)]} \\
&\equiv \max \frac{1}{n} \sum \delta(x - x_i) \log p_\theta(x) \\
&= \max \frac{1}{n} \sum_i \log p_\theta(x_i) = \mathrm{MLE}
\end{aligned}$$

Bregman manifolds in a nutshell

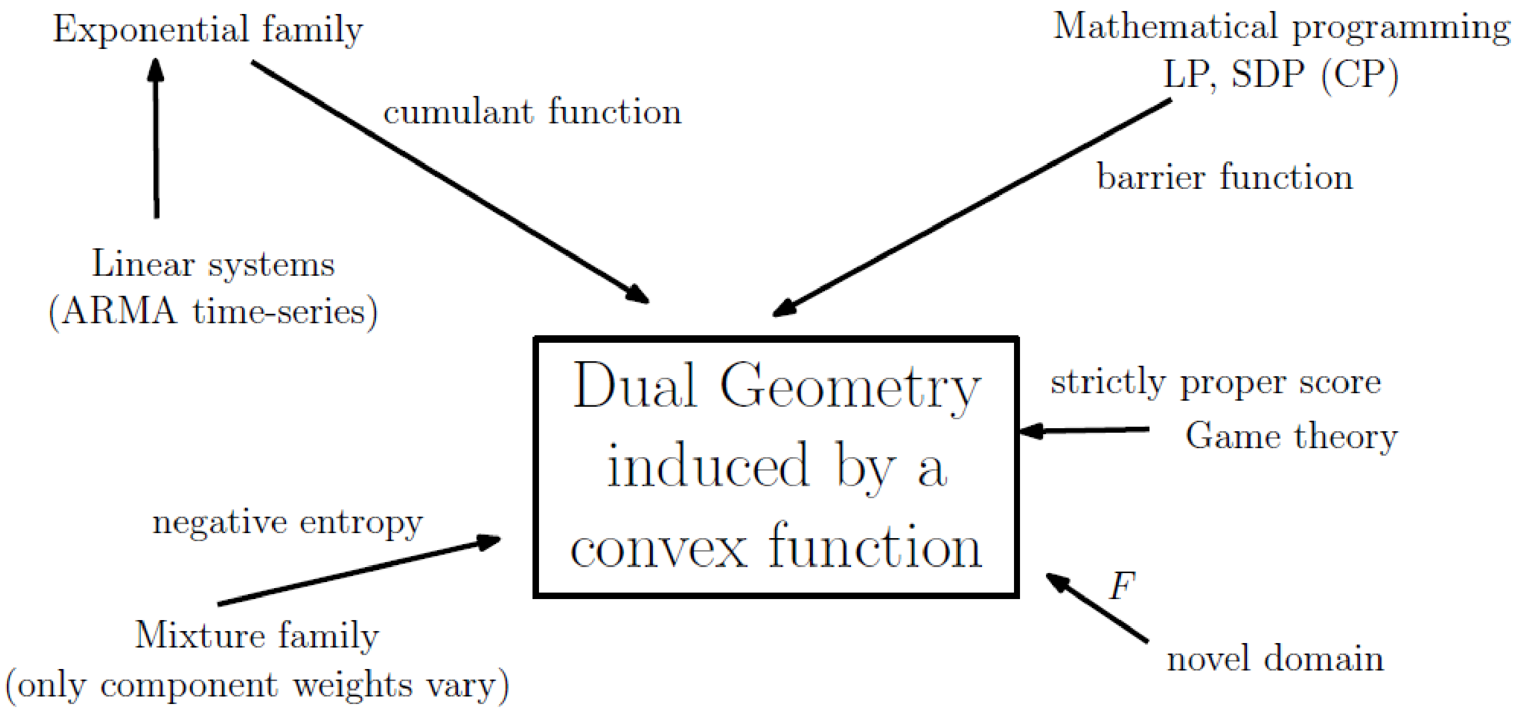
- From *any* smooth (C3) convex function F, we can build a dualistic information-geometric structure called a dually flat manifold.
- Duality emanates from Legendre-Fenchel conjugation
- There are two global (affine) coordinate systems: primal θ and dual η

- We can associate a canonical divergence to dually flat manifolds: Bregman divergences or Fenchel-Young divergences (mixed coordinates)

- There are two dual Pythagoras theorems (and generalized laws 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



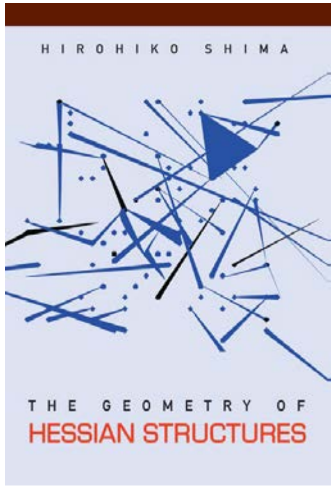
Historically, the dualistic structure of information geometry was called by Lauritzen (1987) a **statistical manifold**.

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

Not necessarily related to statistical models, but can always associate a regular statistical model

Vân Lê, Hông. "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** g^*

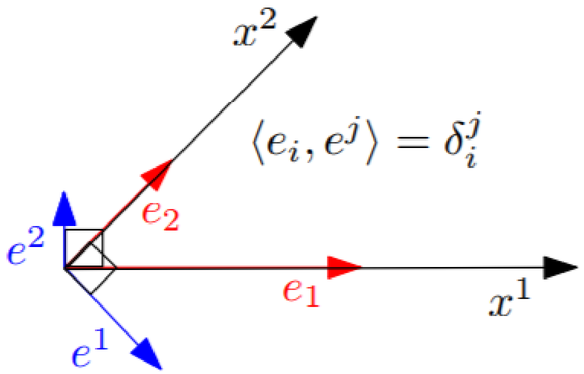
- Expressed in the dual coordinate system η : $\;{}^F g^* = \nabla^2 F^*(\eta)$
- Coordinate-free notation: ${}^F g^* = {}^{F^*} g$
- **∇*-geodesic** of the connection ∇* are straight lines since

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

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 of Bregman manifolds

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

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

Amari-Chentsov cubic tensor: $${}^F C_{ijk} = \partial_i \partial_j \partial_k F(\theta)$$

$$\nabla^1 = \nabla \qquad \nabla^{-1} = \nabla^* \qquad \Gamma^0 = \Gamma^{\mathrm{LC}}$$

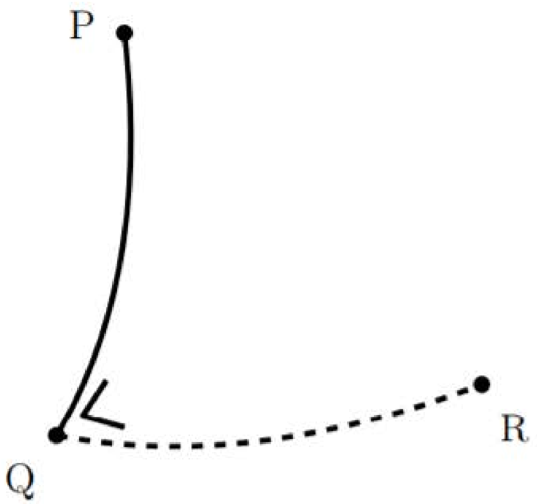
Get the **α-connections**:

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

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

Dual Pythagoras' theorem

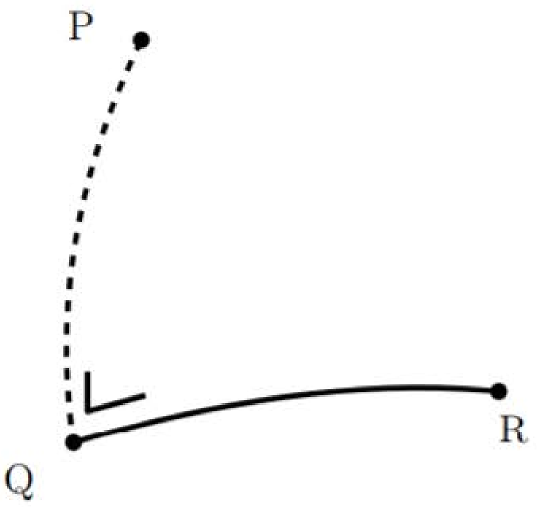
$\gamma^*(P,Q) \perp_F \gamma(Q,R)$



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

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

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



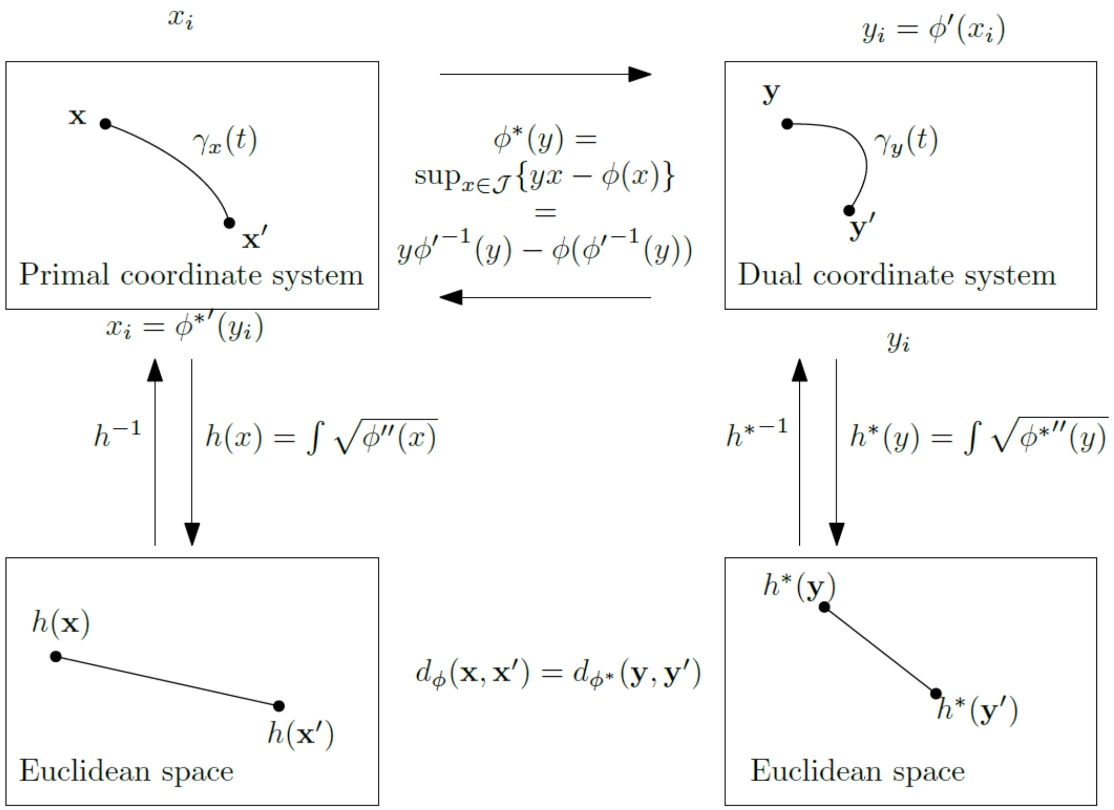
$$D^*(P:R) = D^*(P:Q) + D^*(Q:R)$$

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

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

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

Dual Riemann geodesic distances induced by a separable Bregman divergence



$$\boxed{\rho_\phi(x, x') = \rho_{\phi^*}(y, y') = \rho_{\phi^*}(\nabla\Phi(x), \nabla\Phi(x'))}$$

Legendre conjugate: $\phi^*(y) = y\phi'^{-1}(y) - \phi(\phi'^{-1}(y))$

Bregman divergence:

$$B_\Phi(x, x') := \Phi(x) - \Phi(x') - (x - x')^\top \nabla\Phi(x')$$

Separable Bregman generator:

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

Riemannian metric tensor:

$$g_{ij}(x) = \phi''(x_i)\delta_{ij}$$

Geodesics:

$$\gamma_i(t) = h^{-1}\left((1-t)h(x_i) + th(x_i')\right), \quad t \in [0, 1].$$

Riemannian distance (metric):

$$\rho_\phi(x, x') = \sqrt{\sum_{j=1}^K \left(h(x_j) - h(x_j')\right)^2}$$

where $\boxed{h(x) := \int \sqrt{\phi''(x)}}$

Geometry and clustering with metrics derived from separable Bregman divergences, arXiv:1810.10770

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* $D(\theta(P):\theta(Q))$*:*

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

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

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

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

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

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

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

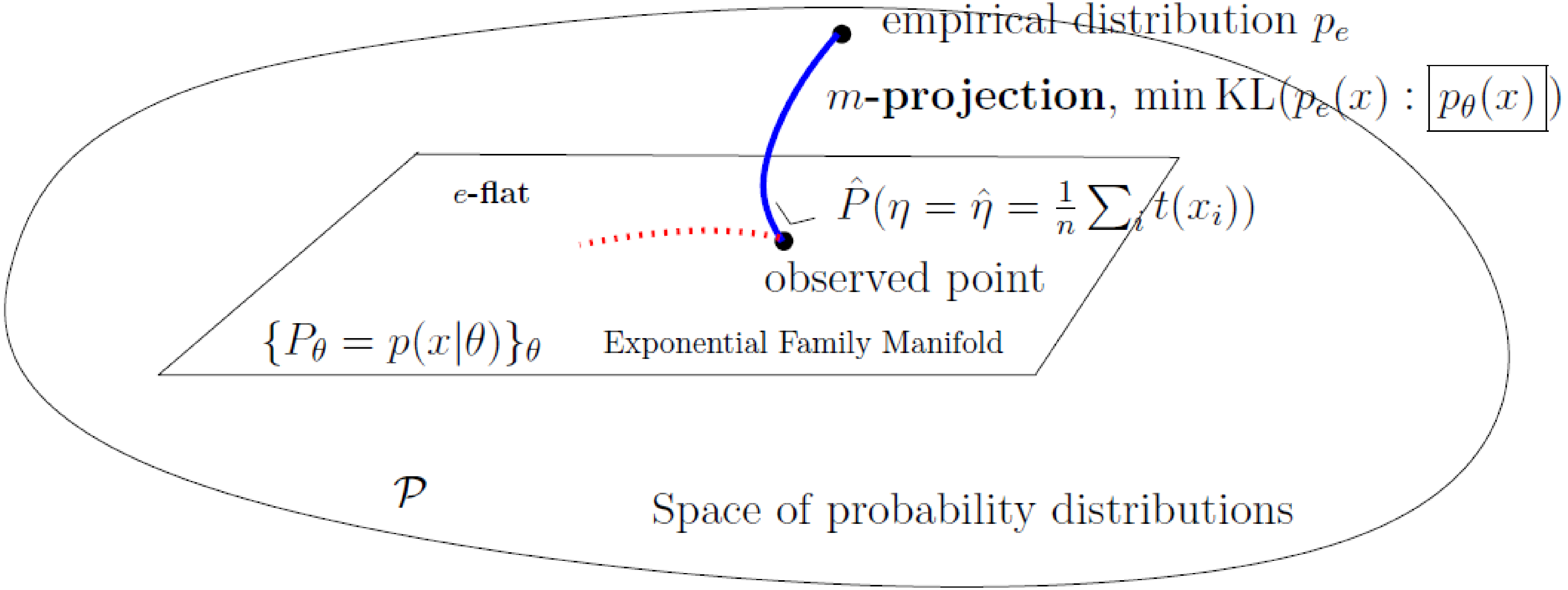
m-projection: $$\boxed{q_m^* = \arg\min_q \mathrm{KL}(p : \boxed{q})}$$

I−projection, rI−projection, KL−projection

MLE for an exponential family as an information projection

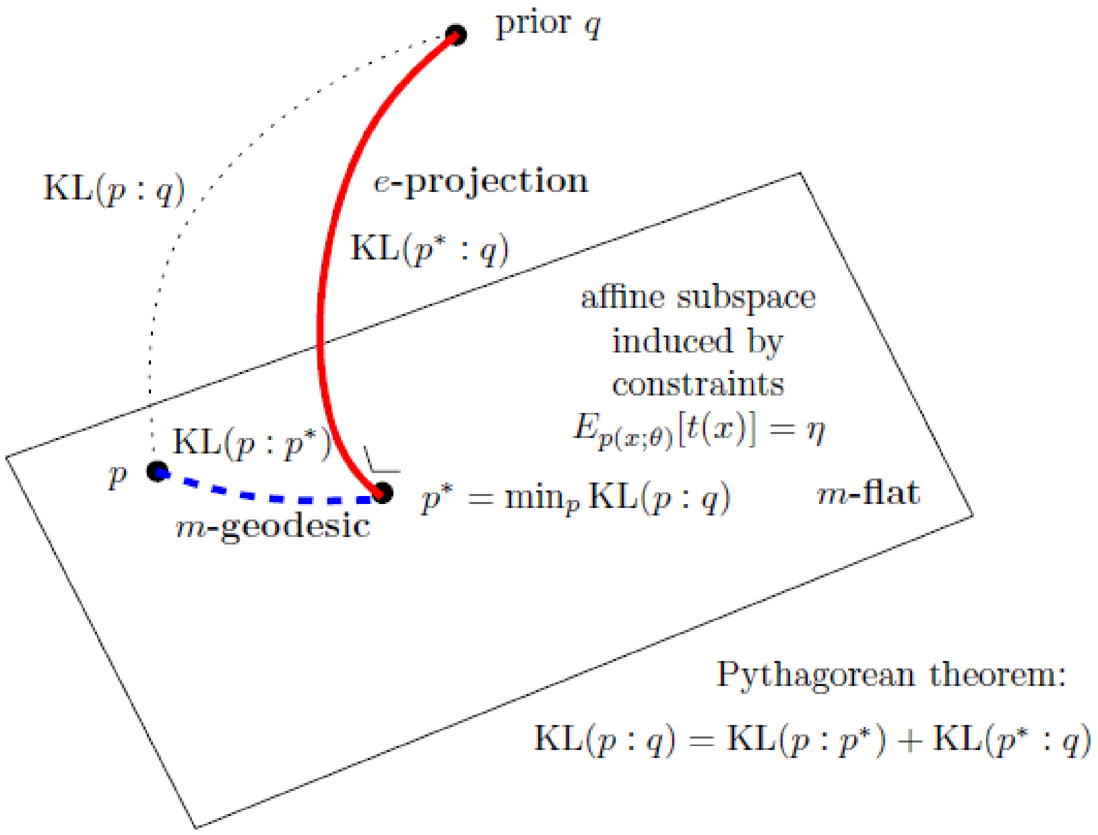
Exponential Family Manifold (EFM) is **e-flat**

Observed point



MaxEnt as an information projection

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



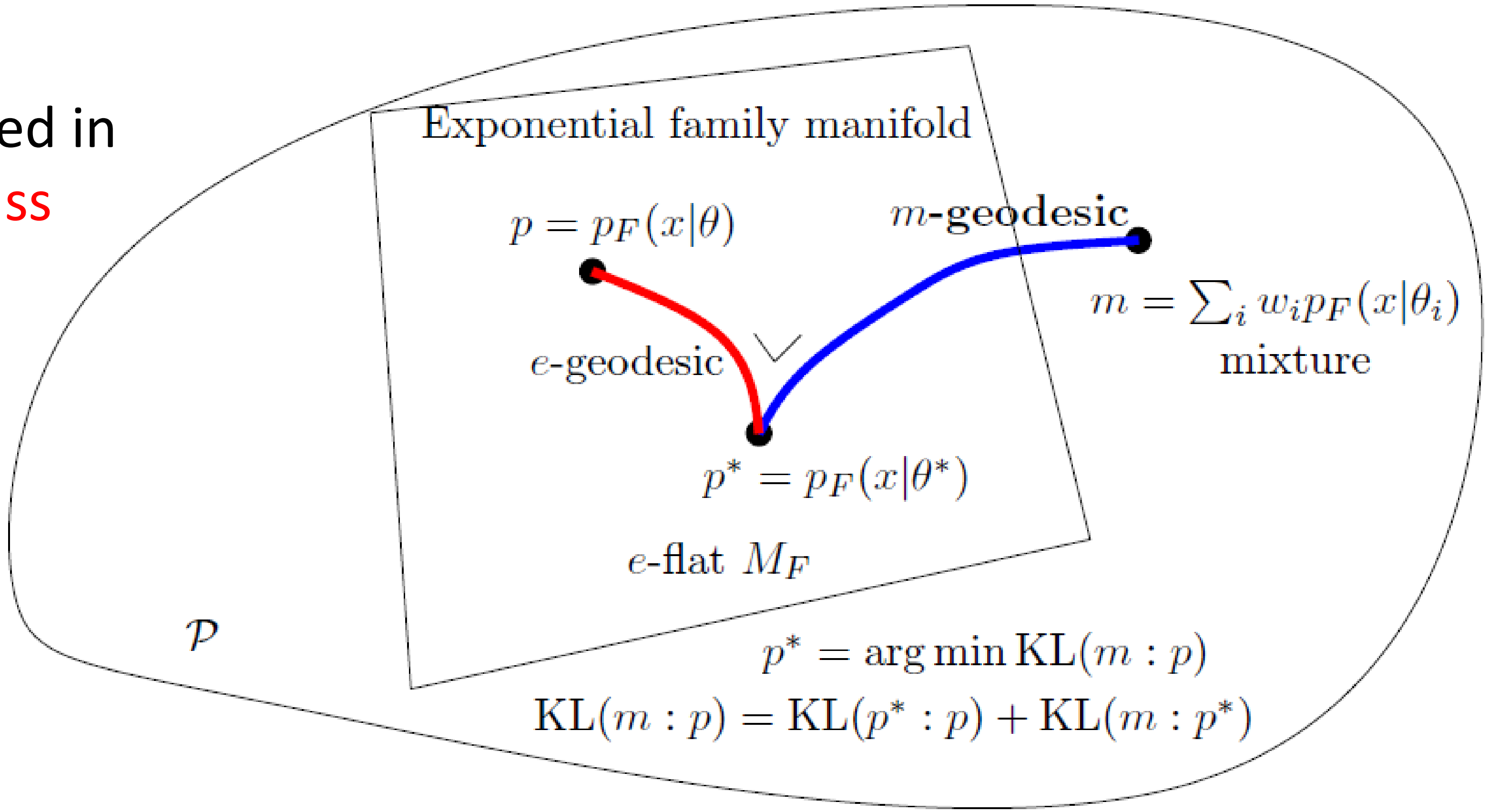
Pythagoras' theorem (Fisher orthogonality) $\gamma_m(p, p^*) \perp_{\mathrm{FIM}} \gamma_e(p^*, q)$

Simplifying a mixture model to a single component

KL right-sided minimization problem for simplifying a mixture of EFs

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

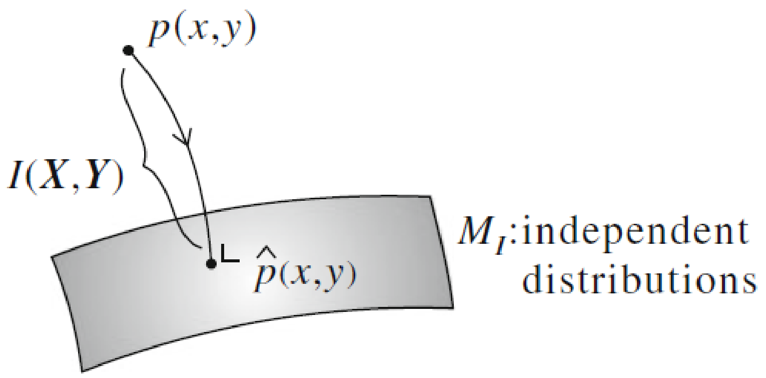
Model centroids for the simplification of kernel density estimators, ICASSP 2012

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

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



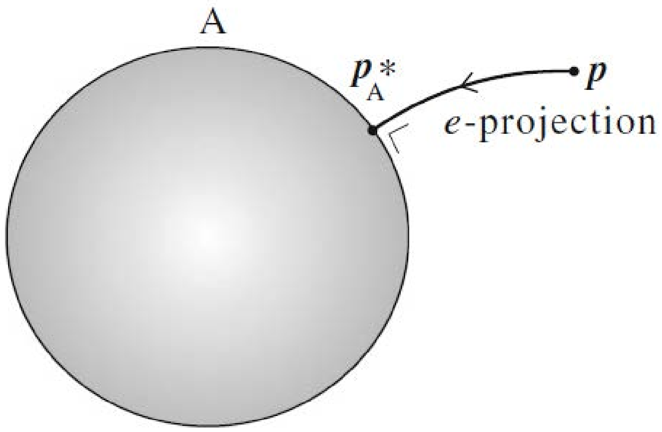
e-geodesic of two independent distributions is family of independent distributions

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

Sanov's theorem (large deviation theory)

Empirical distribution from iid observations is MLE of categorical distributions

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

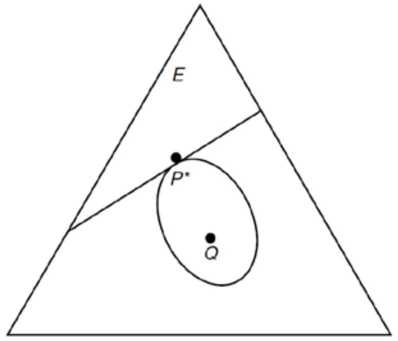


Large Deviation Theorem The probability that $\hat{\boldsymbol{p}}$ is included in A is given asymptotically by

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

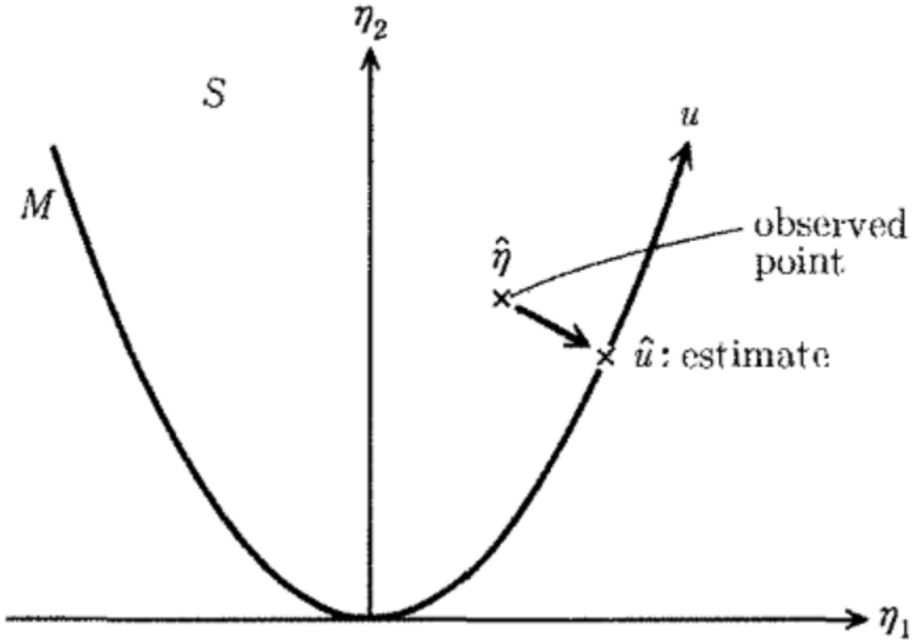
where

$$\boldsymbol{p}_A^* = \arg\min_{\boldsymbol{q}\in A} D_{KL}\left[\boldsymbol{q} : \boldsymbol{p}\right].$$



When A is a closed set having a boundary, $\boldsymbol{p}_A^*$ is given by e-projecting $\boldsymbol{p}$ to the boundary of A.

MLE on a curved exponential family



$$p(x;u) = \exp\{\theta^i(u)x_i - \psi(\theta(u))\}$$

$$\hat\eta = \bar{x}. \qquad \bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t.$$

$$\mu = u \quad \textit{and} \quad \sigma^2 = u^2.$$

$$\hat u = f(\hat\eta)$$

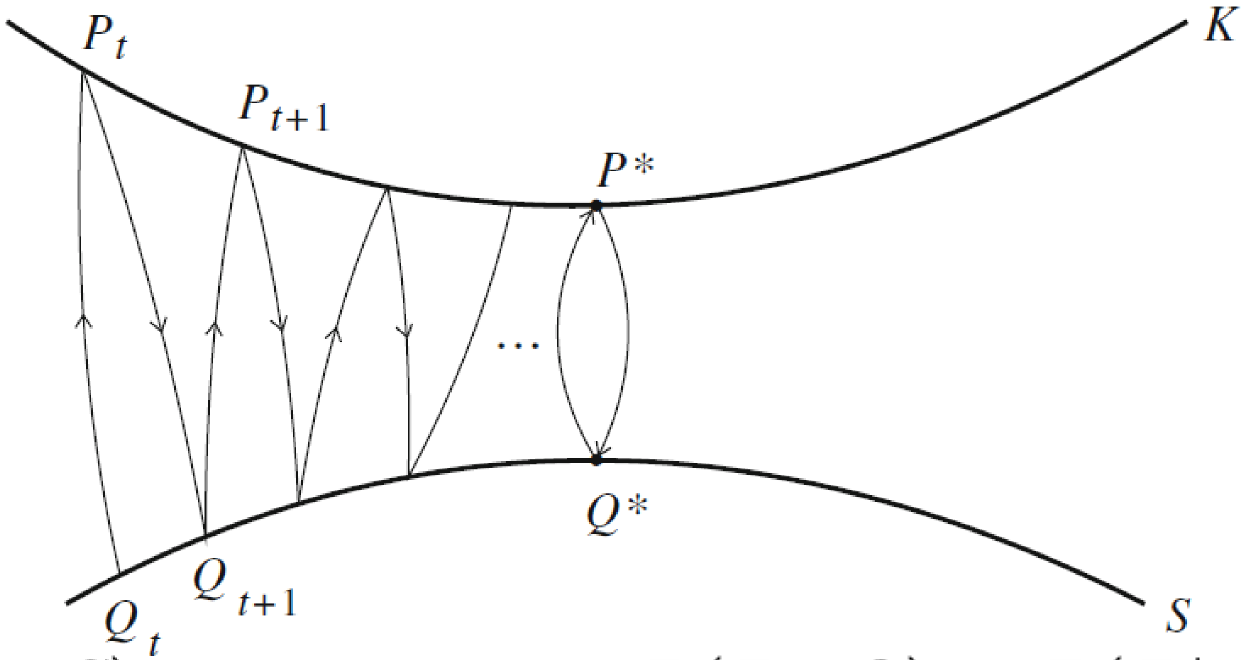
m-geodesic

m-projection

$$\begin{aligned} D(\hat\eta \,\|\, \eta(u)) &= \psi(\theta(u)) + \varphi(\hat\eta) - \theta^i(u)\hat\eta_i \\ &= \varphi(\hat\eta) - \frac{1}{N}\log p_N(x^N; u). \end{aligned}$$

Divergence between two submanifolds

Alternating minimization algorithm



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

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

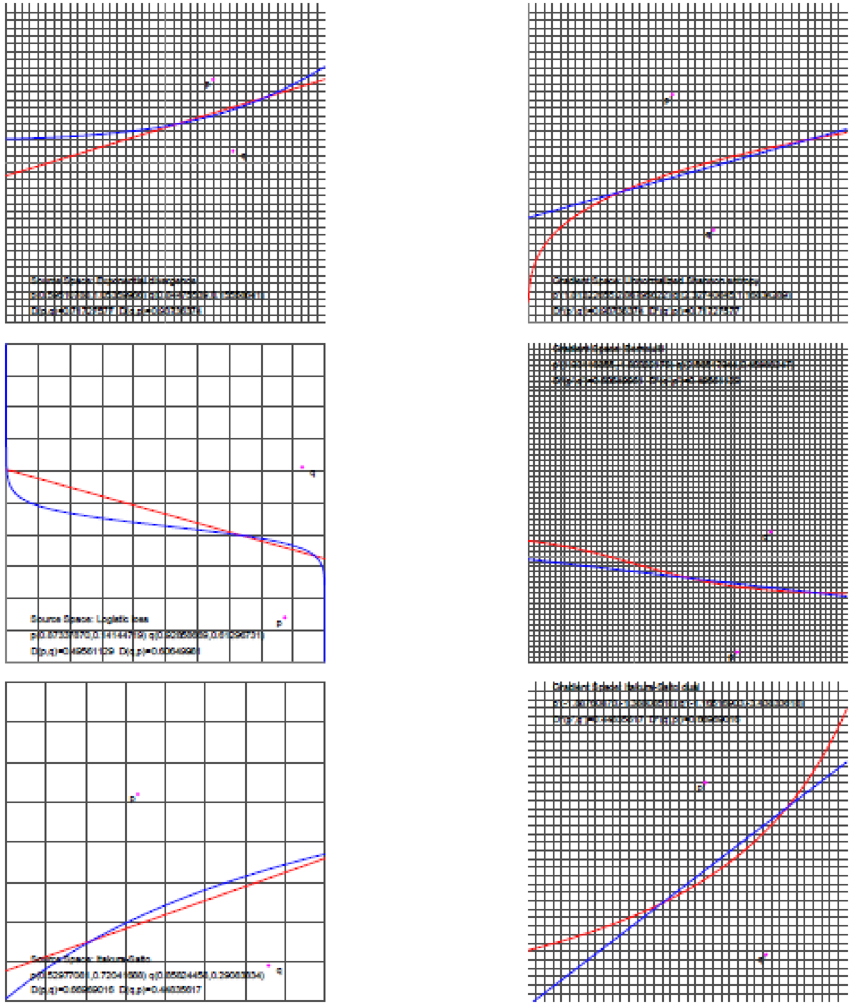
Unique when *S* is flat and *K* is dually flat.

Otherwise, converging point not necessarily unique.

Bregman bisectors

Primal coordinates θ natural parameters

Dual coordinates η expectation parameters



Right-sided bisector: $\rightarrow$ Hyperplane

$$H_F(p,q) = \{x \in \mathcal{X} \mid B_F(x||p) = B_F(x||q)\}.$$

$$H_F : (\nabla F(p) - \nabla F(q))x + (F(p) - F(q) + \langle q, \nabla F(q)\rangle - \langle p, \nabla F(p)\rangle) = 0$$

Left-sided bisector: $\rightarrow$ Hypersurface

$$H'_F(p,q) = \{x \in \mathcal{X} \mid B_F(p||x) = B_F(q||x)\}.$$

$$H'_F : \langle \nabla F(x), q - p\rangle + F(p) - F(q) = 0$$

(hyperplane in the "gradient space" $\nabla\mathcal{X}$ = dual coordinate system)

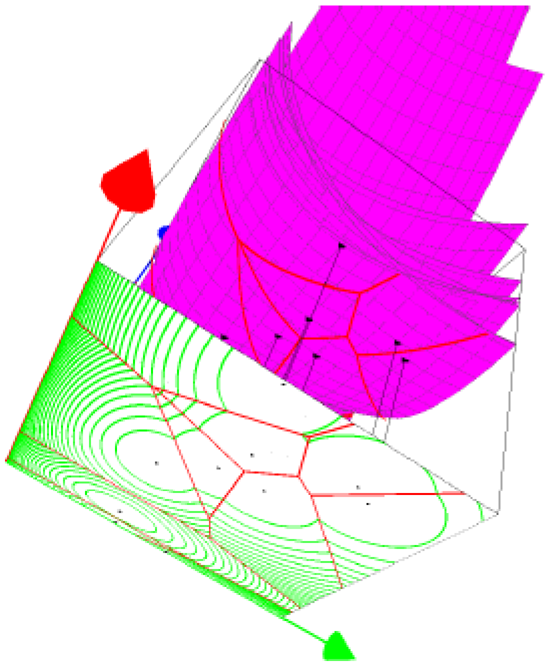
Bregman voronoi diagrams, Discrete & Computational Geometry, 2010

Bregman Voronoi diagrams from lower envelopes

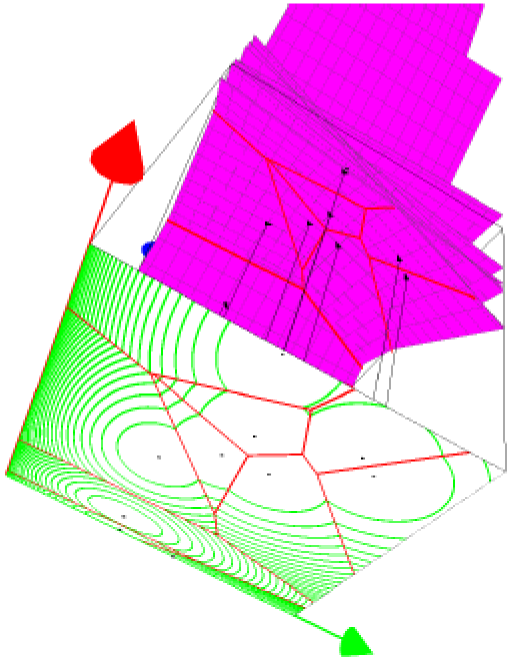
A subclass of **affine diagrams** which have all cells non-empty.
Extend Euclidean Voronoi to Voronoi diagrams in dually flat spaces.
Minimization diagram of the n functions

$D_i(x) = B_F(x||p_i) = F(x) - F(p_i) - \langle x - p_i, \nabla F(p_i)\rangle$.

$\equiv$ minimization of n linear functions: $H_i(x) = (p_i - x)^T \nabla F(q_i) - F(p_i)$.



$\Longleftrightarrow$



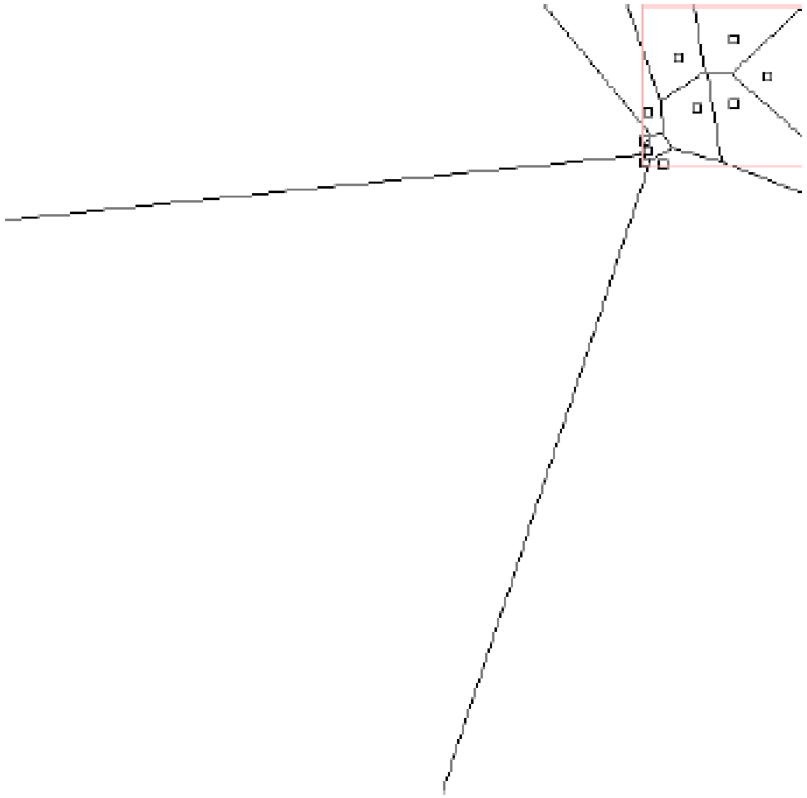
The sided Bregman Voronoi diagrams of n d-dimensional points have complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$

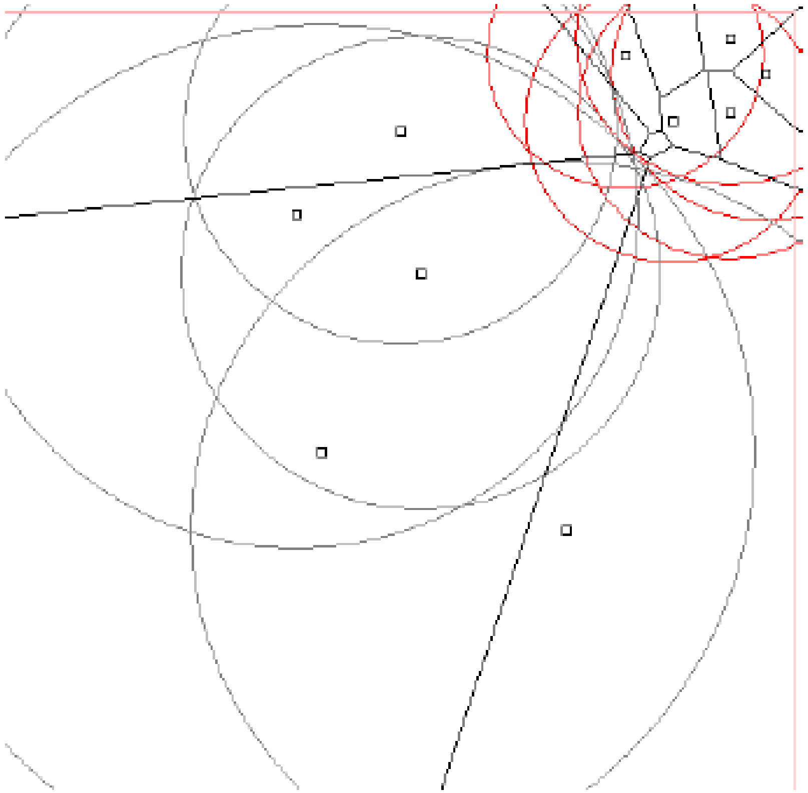
Bregman voronoi diagrams, Discrete & Computational Geometry, 2010

Bregman Voronoi diagrams from power diagrams

Equivalence: $B(\nabla F(p_i), r_i)$ with

$$r_i^2 = \langle \nabla F(p_i), \nabla F(p_i) \rangle + 2(F(p_i) - \langle p_i, \nabla F(p_i) \rangle)$$

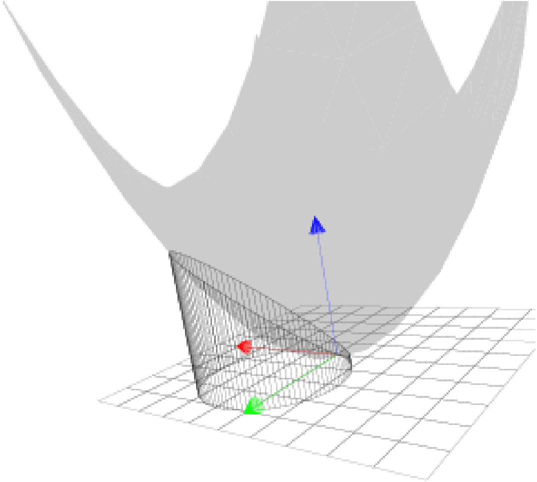


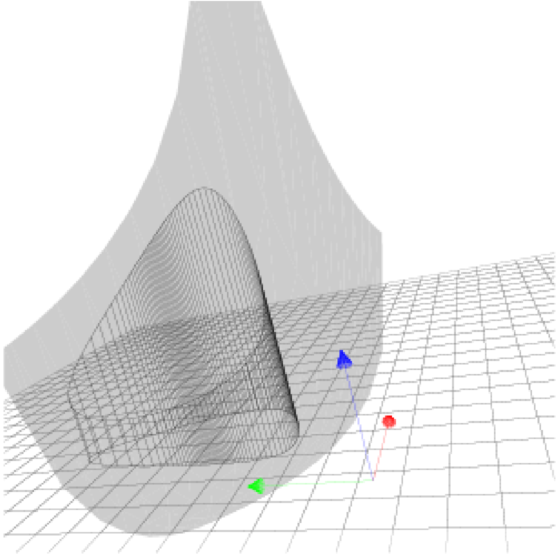


Bregman voronoi diagrams, Discrete & Computational Geometry, 2010

Space of Bregman spheres

$\mathcal{F} : x \mapsto \hat{x} = (x, F(x))$, hypersurface in $\mathbb{R}^{d+1}$.
H_p: Tangent hyperplane at $\hat{p}$, $z = H_p(x) = \langle x - p, \nabla F(p) \rangle + F(p)$
Bregman sphere $\sigma \longrightarrow \hat{\sigma}$ with supporting hyperplane
$H_\sigma : z = \langle x - c, \nabla F(c) \rangle + F(c) + r$. (// to H_c and shifted vertically by r)
$\hat{\sigma} = \mathcal{F} \cap H_\sigma$.
Conversely, the intersection of any hyperplane H with $\mathcal{F}$ projects onto $\mathcal{X}$ as a Bregman sphere:
$H : z = \langle x, a \rangle + b \rightarrow \sigma : \mathrm{Ball}_F(c = (\nabla F)^{-1}(a), r = \langle a, c \rangle - F(c) + b)$





$$\mathtt{InSphere}(x; p_0, ..., p_d) = \begin{vmatrix} 1 & ... & 1 & 1 \\ p_0 & ... & p_d & x \\ F(p_0) & ... & F(p_d) & F(x) \end{vmatrix}$$

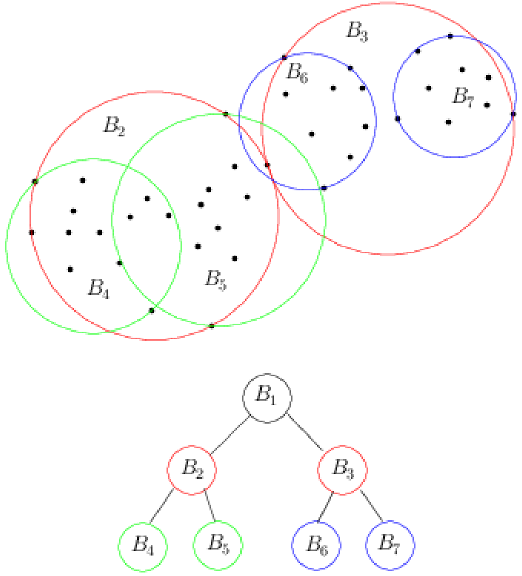
Bregman voronoi diagrams, Discrete & Computational Geometry, 2010

Fast Proximity queries for Bregman divergences (incl. KL)

Fast <u>Nearest Neighbour Queries</u> for Bregman divergences

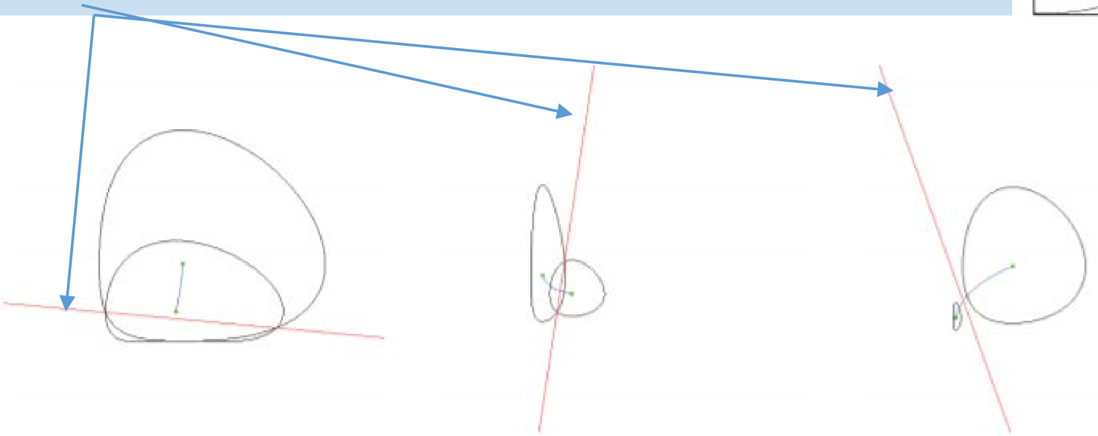
Space partition induced by

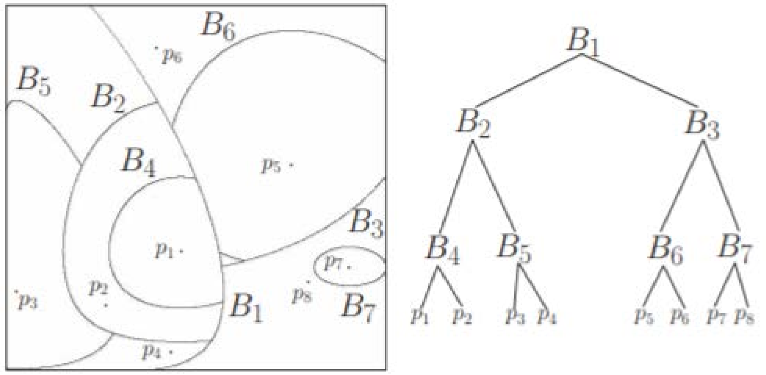
<u>Bregman vantage point trees</u>

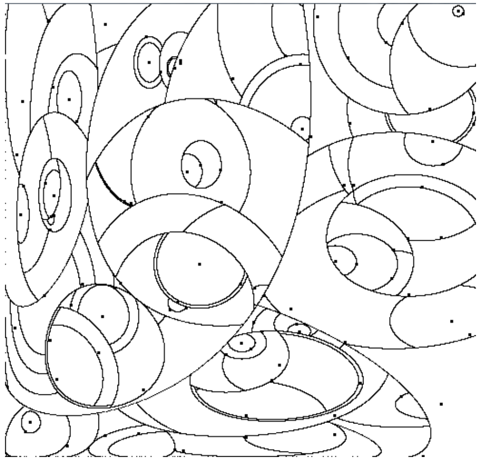


<u>Bregman ball trees</u>

<u>Key property</u>:
Check whether two Bregman spheres
Intersect or not easily
(radical hyperplane, space of spheres)





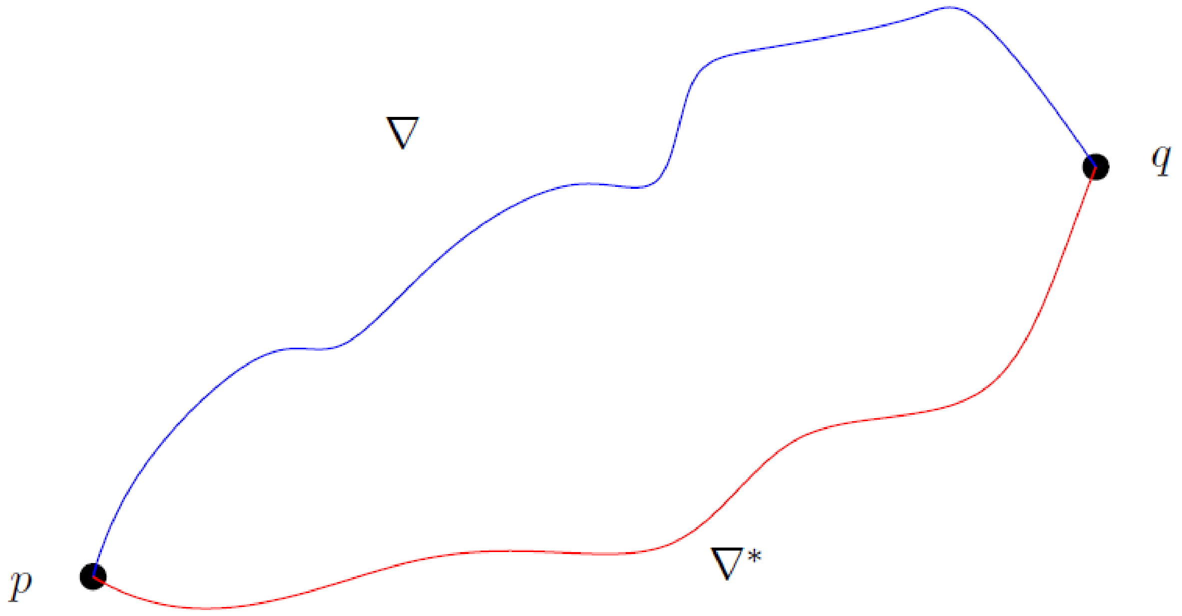


E.g., Extended Kullback-Leibler

C++ source code https://www.lix.polytechnique.fr/~nielsen/BregmanProximity/

Bregman vantage point trees for efficient nearest Neighbor Queries, ICME 2009
Tailored Bregman ball trees for effective nearest neighbors, EuroCG 2009

Dualistic structure of the Gaussian manifold



∇: e-connection

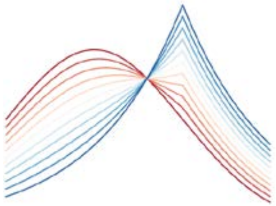
∇^*:m-connection

$$(p_1p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ v^m_\alpha = (1-\alpha)v_1 + \alpha v_2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2 \end{cases}$$

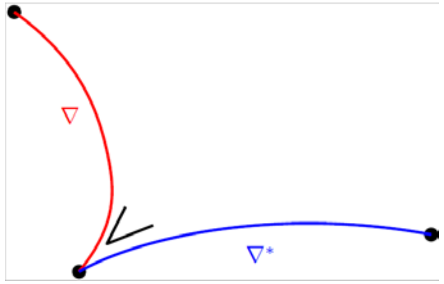
$$(p_1p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \frac{(1-\alpha)\mu_1 v_2 + \alpha\mu_2 v_1}{(1-\alpha)v_2 + \alpha v_1} \\ v^e_\alpha = \frac{v_1 v_2}{(1-\alpha)v_2 + \alpha v_1} \end{cases}$$

$$(p_1p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ \Sigma^m_\alpha = \bar{\Sigma}_\alpha + (1-\alpha)\mu_1\mu_1^\top - \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha{}^\top \end{cases}$$

$$(p_1p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \Sigma^e_\alpha((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2) \\ \Sigma^e_\alpha = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1} \end{cases}$$



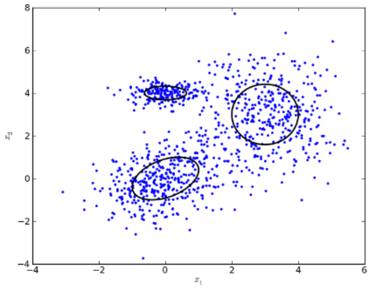
Distances and information geometry of finite statistical mixtures

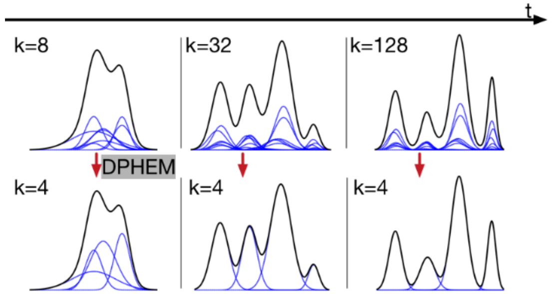


Frank Nielsen

Sony CSL

Finite statistical mixtures





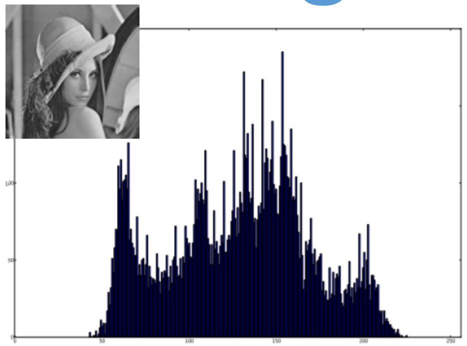
- **Semi-parametric** models, universal estimators of smooth densities
- Gaussian mixture models (GMMs), Exponential family mixture models (EFMMs), etc.

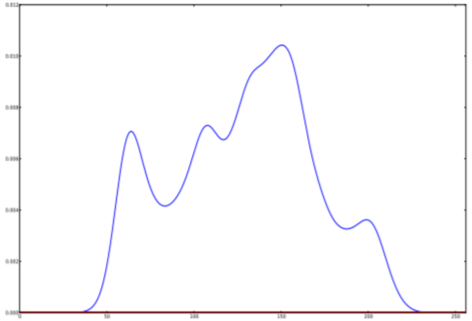
$$f(x) = \sum_{i=1}^{n} \omega_i \, g(x;\, \mu_i, \sigma_i^2) \qquad \text{with} \qquad g(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

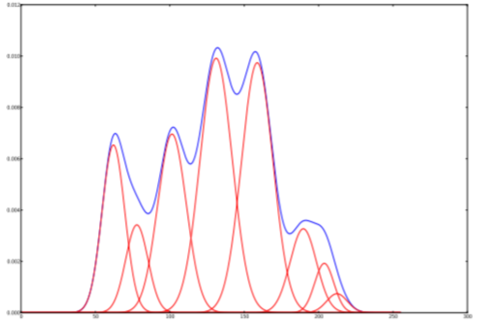
- But non-identifiable/non-regular !!! (not 1-to-1 parameter/density)
- Usually learn GMMs by Expectation-Maximization (EM, local optimum)
- But also can learn mixtures by simplifying a Kernel Density Estimator

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

Learning a mixture by simplifying a kernel density estimator







Original histogram
raw KDE (14400 components)
simplified mixture (8 components)

$$f(x) = \sum_{i=1}^{n} \omega_i \, g(x;\, \mu_i, \sigma_i^2) \qquad g(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Usual **centroids** based on Kullback-Leibler sided/symmetrized

$$\arg\min_c \textstyle\sum_i \omega_i KLD(c, x_i)$$

$$\arg\min_c \textstyle\sum_i \omega_i KLD(x_i, c)$$

$$\arg\min_c \textstyle\sum_i \omega_i \mathrm{SKL}(x_i, c)$$

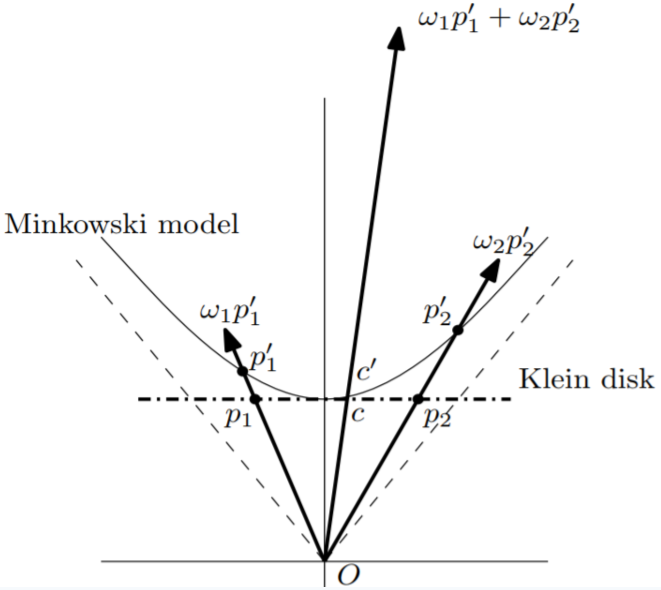
$$\mathrm{KLD}(f_p, f_q) = \frac{1}{2}\log\left(\frac{\det \Sigma_p}{\det \Sigma_q}\right) + \frac{1}{2}\mathrm{tr}(\Sigma_q^{-1}\Sigma_p) + \frac{1}{2}(\mu_q - \mu_p)^T \Sigma_q^{-1}(\mu_q - \mu_p) - \frac{d}{2}$$

or Fisher-Rao distance (hyperbolic distance)

$$\mathrm{FRD}(f_p, f_q) = \sqrt{2}\ln \frac{|(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)| + |(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)|}{|(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)| - |(\frac{\mu_p}{\sqrt{2}}, \sigma_p) - (\frac{\mu_q}{\sqrt{2}}, \sigma_q)|}$$

Problem:
No closed-form FR/SKL centroids!!!

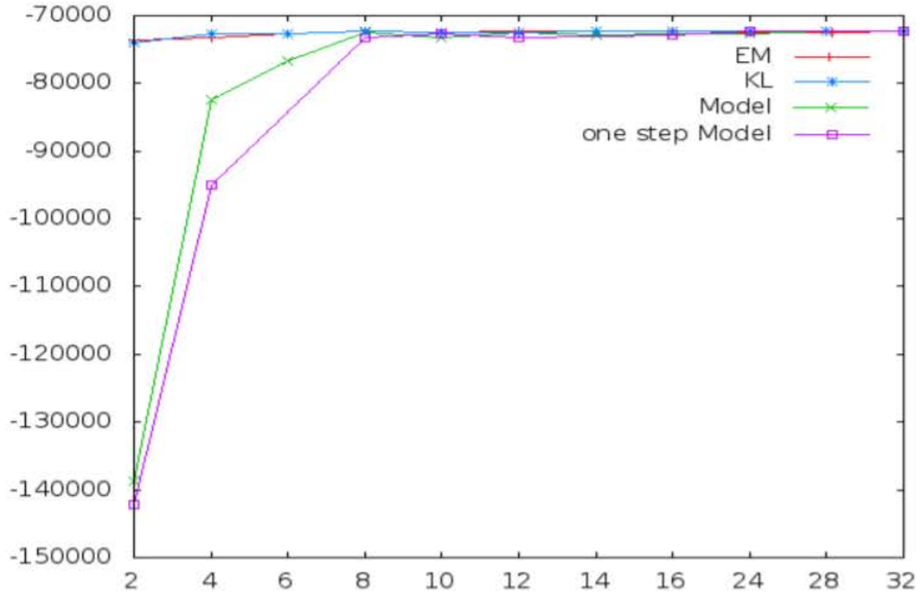
Galperin's model centroid (HG)

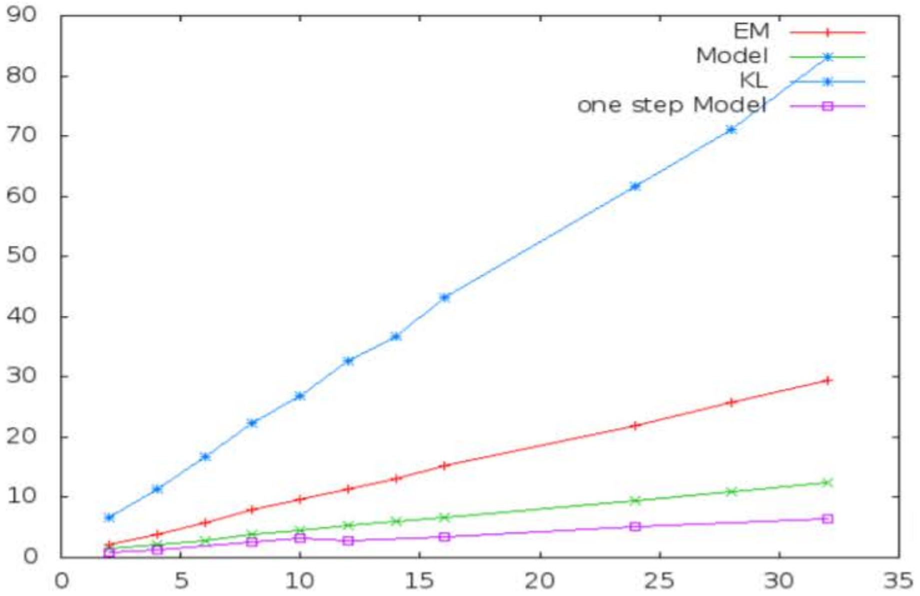


Simple model centroid algorithm:
Embed Klein points to points of the Minkowski hyperboloid
Centroid = center of mass c, scaled back to c' of the hyperboloid
Map back c' to Klein disk

Model centroids for the simplification of Kernel Density estimators. ICASSP 2012

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.

Model centroids for the simplification of Kernel Density estimators. ICASSP 2012

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

Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities. Entropy 18(12) (2016)

Batch learning of mixtures and lightspeed distance calculations

$$m(x)=\sum_{i=1}^{k_1}\omega_i p_F(x;\eta_i) \quad m'(x)=\sum_{i=1}^{k_2}\omega'_i p_F(x;\eta'_i)$$

$$\mathrm{KL}_{\mathrm{MC}}(m\|m')=\frac{1}{n}\sum_{i=1}^{n}\log\frac{m(x_i)}{m'(x_i)}$$

Kullback-Leibler divergence

$$\mathrm{KL}(p\|q)=\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x = H(p,q)-H(p)$$

Monte-Carlo stochastic estimation (iid sampling from m)

- Hungarian best bipartite matching

of components (Goldberger)

$$\mathrm{KL}_{\mathrm{Gold}}(m\|m')=\arg\min_{\sigma}\mathrm{KL}(\omega\|\sigma(\omega')) + \sum\omega_i \mathrm{KL}\left(p_F(\cdot\|\eta_i)\|p_F\left(\cdot\|\eta'_{\sigma(i)}\right)\right)$$

- Variational approximation of KL for mixtures:

$$\boxed{\mathrm{KL}_{\mathrm{var}}(m\|m')=\sum_i \omega_i\log\frac{\sum_j\omega_j e^{-\mathrm{KL}(p_F(\cdot\|\eta_i)\|p_F(\cdot\|\eta_j))}}{\sum_j\omega'_j e^{-\mathrm{KL}(p_F(\cdot\|\eta_i)\|p_F(\cdot\|\eta'_j))}}}$$

Definition A co-mixture of exponential families (a *comix*) with K components is a set of S statistical mixture models of the form:

$$\begin{cases} m_1(x;\omega_i^{(1)}\ldots\omega_K^{(1)})=\sum_{i=1}^K\omega_i^{(1)}p_F(x;\eta_i)\\ m_2(x;\omega_i^{(2)}\ldots\omega_K^{(2)})=\sum_{i=1}^K\omega_i^{(2)}p_F(x;\eta_i)\\ \ldots\\ m_S(x;\omega_i^{(S)}\ldots\omega_K^{(S)})=\sum_{i=1}^K\omega_i^{(S)}p_F(x;\eta_i)\end{cases}$$

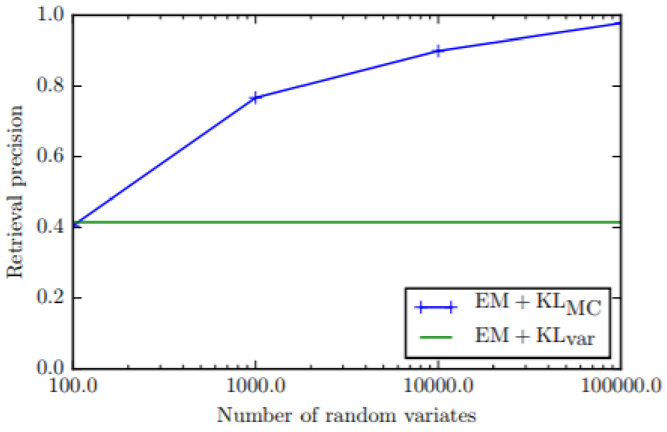
Extend Expectation-Maximization algorithms for batch learning of co-mixtures (co-EM, adapt Bregman soft clustering)

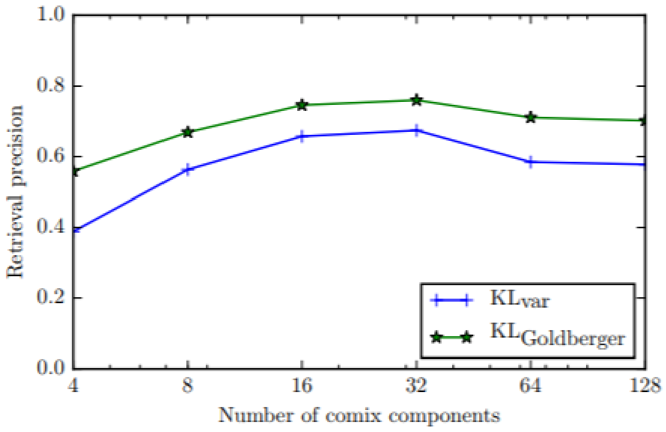
Precompute the matrix: $D_{ij}=\mathrm{KL}\left(p_F(\cdot\|\eta_i)\|p_F(\cdot\|\eta_j)\right).$

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 co-mixturess





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	$\times 1.5$	0.00020s	$\times 180$	$\times$ 20	0.00015s
8	99s	$\times 0.77$	0.00044s	$\times 84$	$\times$ 5.8	0.00030s
16	48s	$\times 1.6$	0.0012s	$\times 28$	$\times$ 1.6	0.00059s
32	150s	$\times 0.49$	0.0040s	$\times 9.1$	$\times$ 0.41	0.0012s
64	450s	$\times 0.17$	0.014s	$\times 2.5$	$\times$ 0.10	0.0024s
128	600s	$\times 0.12$	0.046s	$\times 0.80$	$\times 0.026$	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

$$m_1(x) = \sum_{i=1}^{k_1} \alpha_i p_i(x) \qquad m_2(x) = \sum_{i=1}^{k_2} \beta_i q_i(x) \qquad p_{i,j} = p_i$$

$$m_1 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j} \qquad m_2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} q_{i,j} \qquad q_{i,j} = q_j$$

Solve the **Linear Program**: $H_\delta(p,q) = \sum^{k_1} \sum^{k_2} w_{ij} \delta(p_i, q_j)$

with the following constraints:

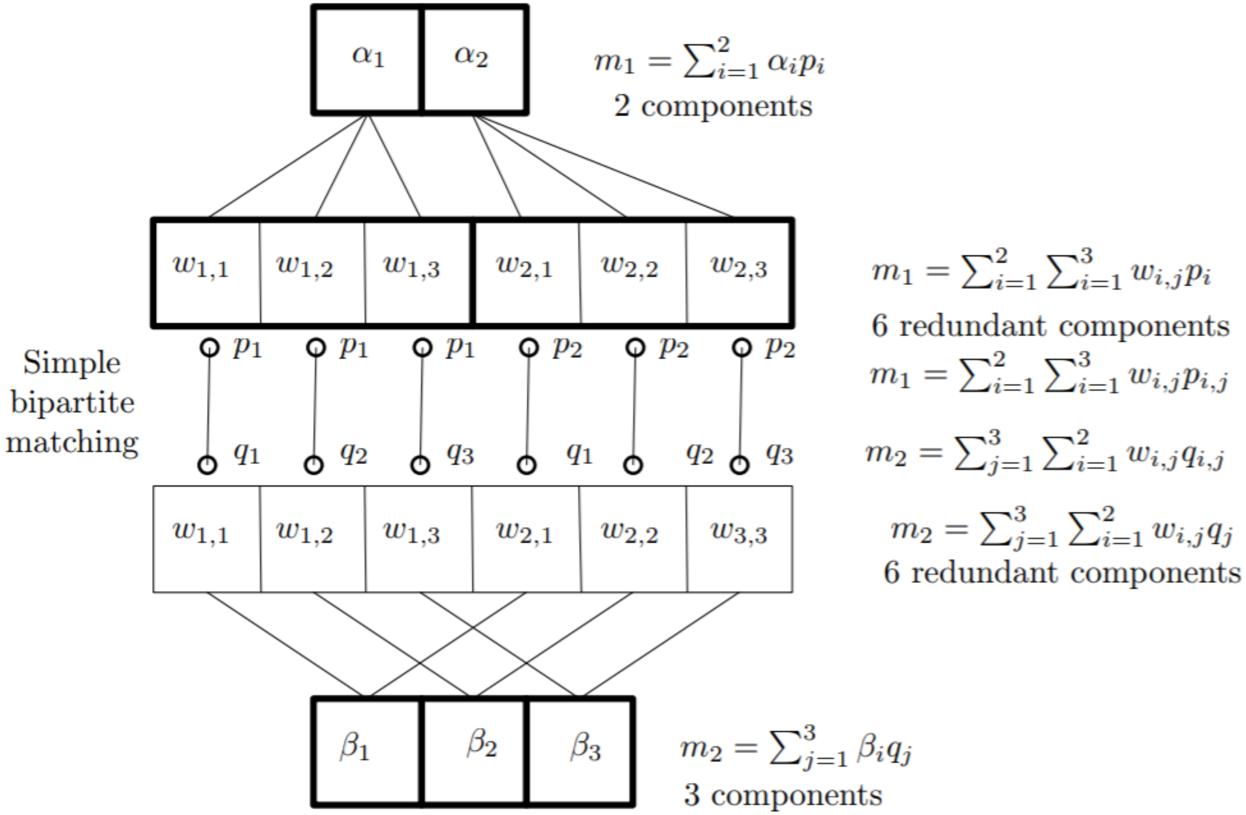
$$w_{ij} \geq 0, \quad \forall i \in [k_1], j \in [k_2]$$

$$\sum_{l=1}^{k_2} w_{il} = \alpha_i, \quad \forall i \in [k_1]$$

$$\sum_{l=1}^{k_1} w_{lj} = \beta_j, \quad \forall j \in [k_2].$$

Equivalent **optimal transport** problem:

$$H_\delta(m_1 : m_2) = \min_{W \in U(\alpha,\beta)} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{ij} \delta(p_i, q_j).$$



On The Chain Rule Optimal Transport Distance. CoRR abs/1812.08113 (2018)

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**

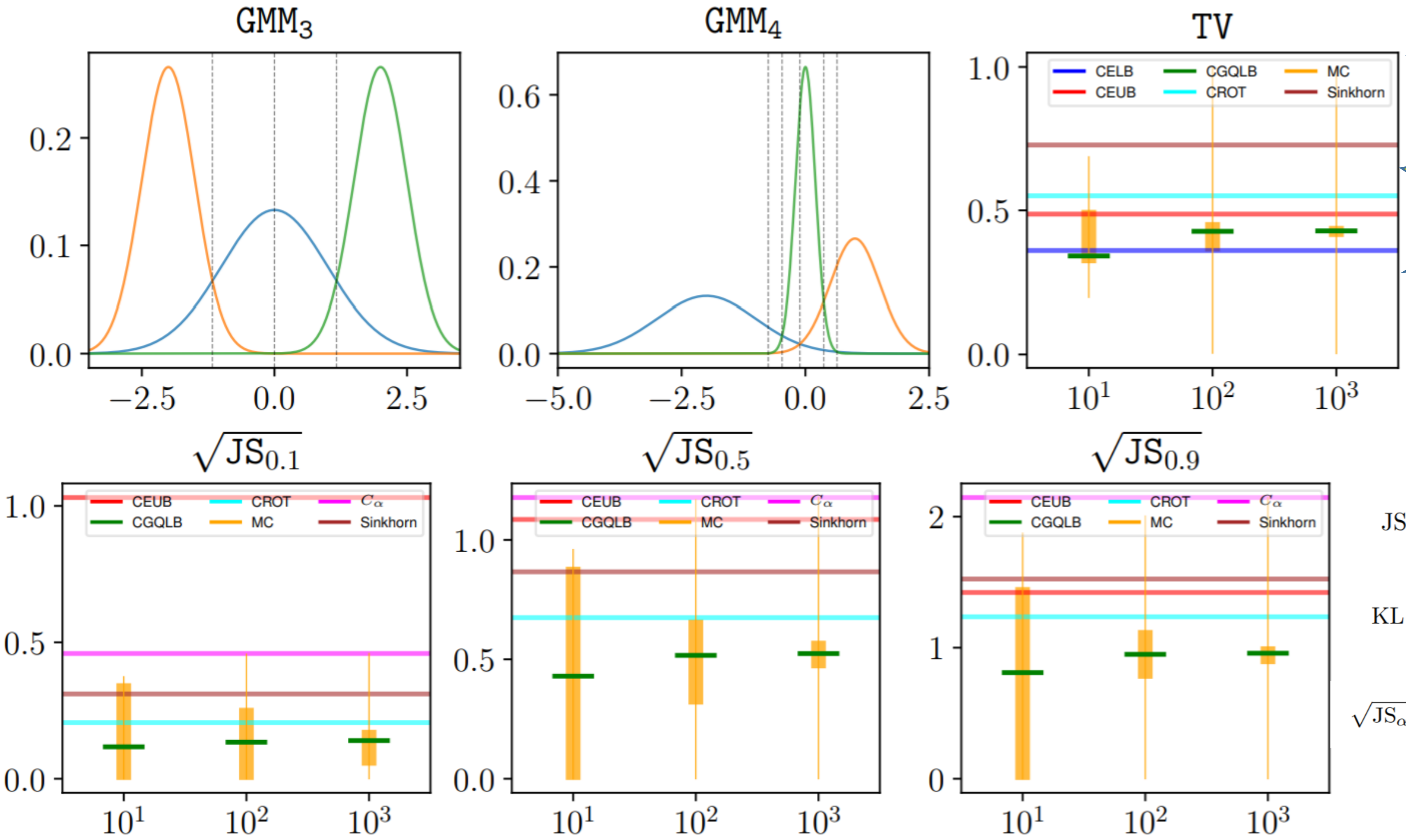
$$
\begin{aligned}
\delta(m_1 : m_2) &= \delta\left(\sum_{i=1}^{k_1} \alpha_i p_i, \sum_{j=1}^{k_2} \beta_j q_j\right) \\
&= \delta\left(\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{i,j} p_{i,j} : \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{i,j} q_{i,j}\right) \\
&\leq \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{i,j} \delta(p_{i,j} : q_{i,j}), \\
&\leq \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{i,j} \delta(p_i : q_j) =: H_\delta(m_1, m_2).
\end{aligned}
$$

f-divergences (incl. KL) are joint convex

But also the p-powered Wasserstein distances, Etc.

On The Chain Rule Optimal Transport Distance. arXiv:1812.08113 (2018)

Chain Rule Optimal Transport (CROT) distance



Fast Sinkhorn calculations

$$\mathrm{JS}_\alpha(p:q) := \frac{1}{2}\mathrm{KL}(p:(pq)_\alpha) + \frac{1}{2}\mathrm{KL}(q:(pq)_\alpha)$$

$$\mathrm{KL}(p:(pq)_\alpha) \leq \int p \log \frac{p}{(1-\alpha)p} \leq -\log(1-\alpha)$$

$$\sqrt{\mathrm{JS}_\alpha(p:q)} \leq C_\alpha = \sqrt{-\frac{1}{2}\log(1-\alpha) - \frac{1}{2}\log\alpha}.$$

On The Chain Rule Optimal Transport Distance. CoRR abs/1812.08113 (2018)

Statistical mixtures versus mixture families

- In statistics, finite statistical mixtures are **irregular models** (non-identifiable)

$$m(x;w) := \sum_{i=0}^{k-1} w_i p_i(x),$$

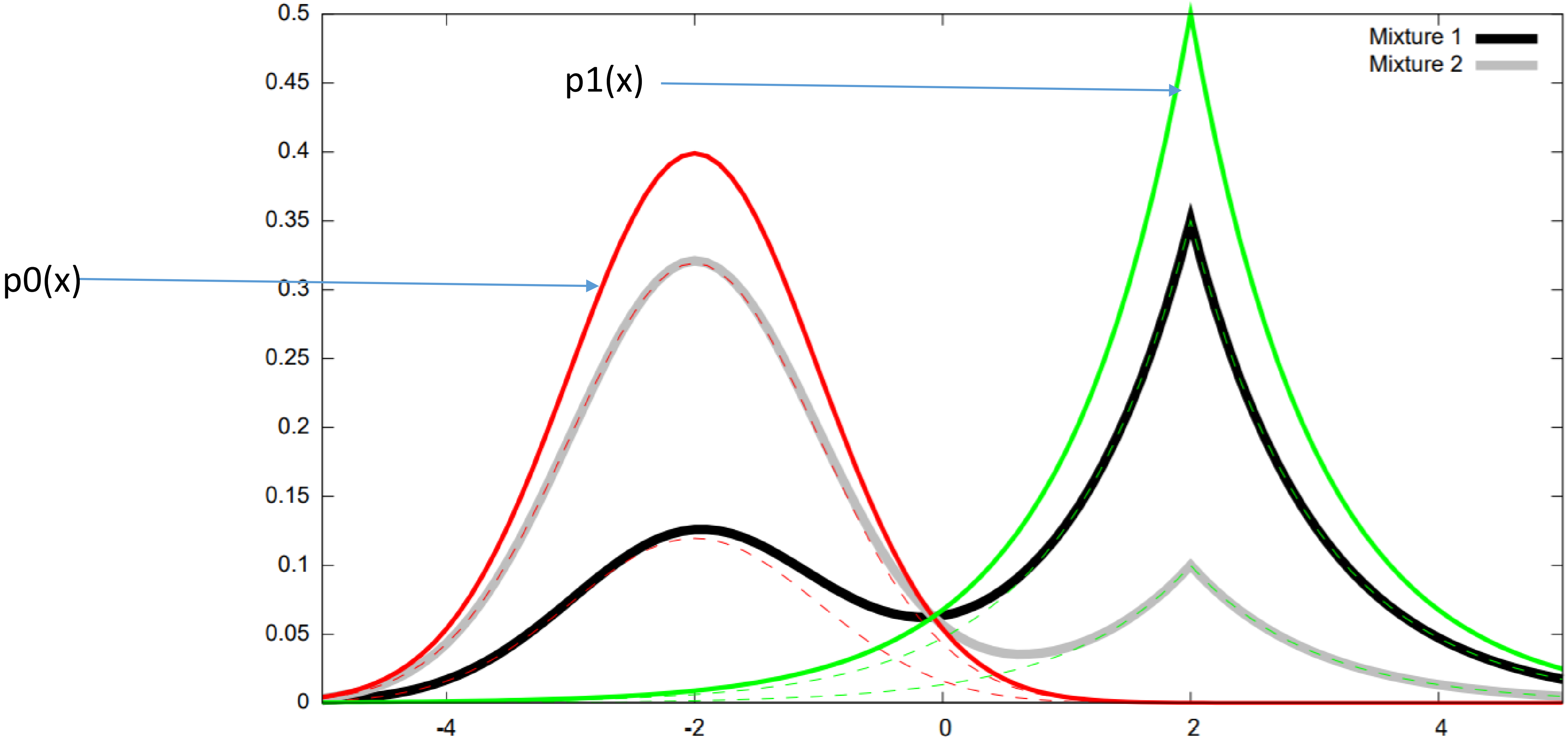
- Information geometry primarily considers regular models
- In information geometry, **mixture families are regular parametric** models $\mathcal{M} := \{m(x;w)\ ,\ w \in \Delta_{k-1}^{\circ}\}$ $\quad f_i(x) = p_i(x) - p_0(x) \quad c(x) = p_0(x)$

$$\mathcal{M} = \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x), \eta \in \mathbb{R}_{++}^{k-1} \right\} \quad \mathcal{M} = \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x), \quad \eta \in H^{\circ} \right\}$$

- Statistical mixtures with prescribed distinct component distributions form mixture families

On the Geometry of Mixtures of Prescribed Distributions. ICASSP 2018

A mixture family of order 1 (=2 fixed 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 fixed components)

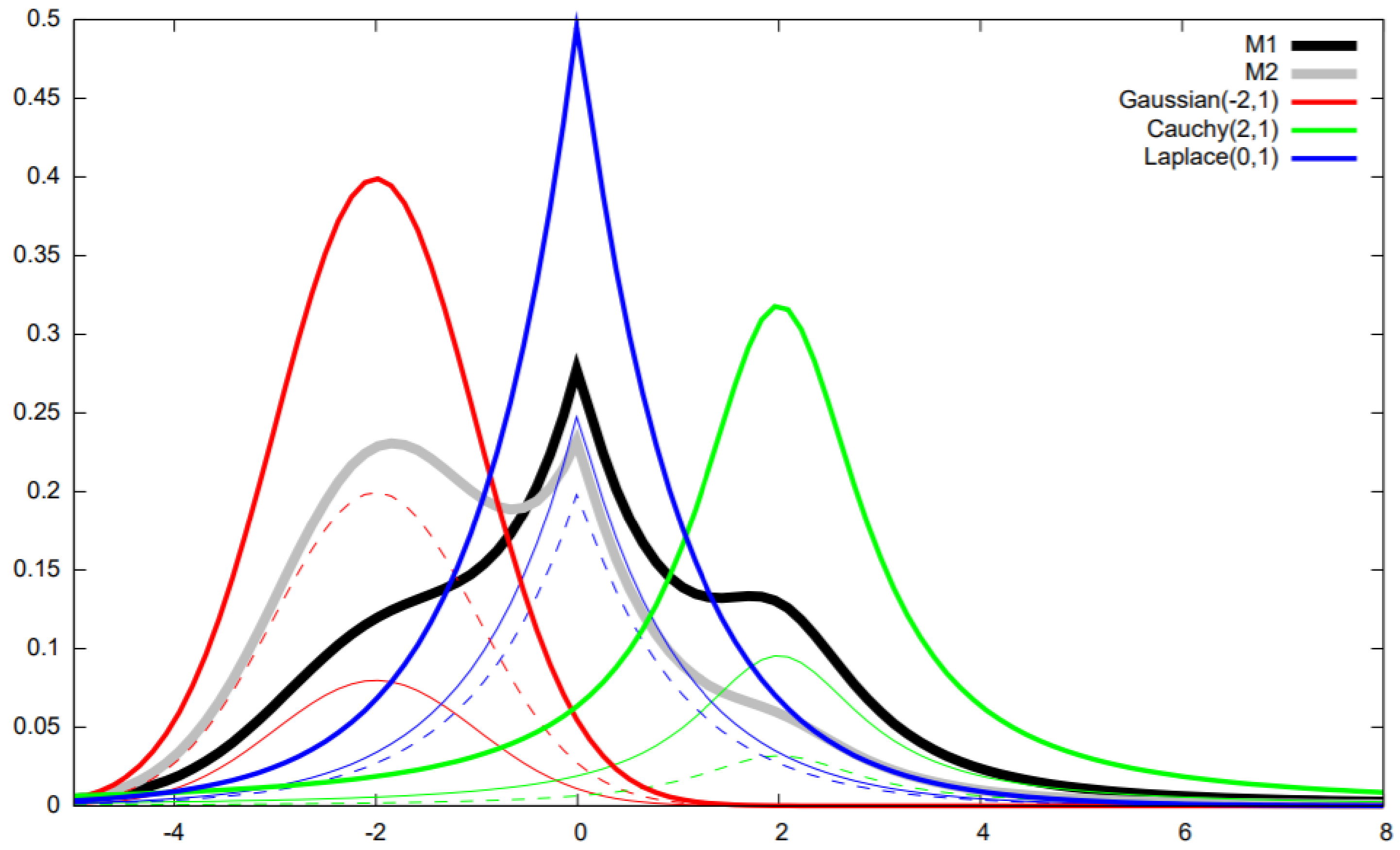


Figure : Example of a mixture family of order $D = 2$ ($k = 3$): $p_0(x) \sim$ Gaussian$(-2, 1)$ (red), $p_1(x) \sim$ Laplace$(0, 1)$ (blue) and $p_2(x) \sim$ Cauchy$(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = (0.3, 0.5)$ and $m_2(x) = m(x; \eta)$ (gray) with $\eta = (0.1, 0.4)$.

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	$\theta(\cdot)$	$\eta(\cdot)$
Conversion $\theta \leftrightarrow \eta$	$\theta(\eta) = \nabla F^*(\eta)$	$\eta(\theta) = \nabla F(\theta)$
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	$G(\theta) = \nabla^2 F(\theta)$	$G^*(\eta) = \nabla^2 F^*(\eta)$
	$g_{ij} = \partial_i \partial_j F(\theta)$	$g^{ij} = \partial^i \partial^j F^*(\eta)$
Geodesic $(\lambda \in [0,1])$	$\gamma(P,Q) = \{(PQ)_\lambda = (1-\lambda)\theta(P) + \lambda\theta(Q)\}_\lambda$	$\gamma^*(P,Q) = \{(PQ)^*_\lambda = (1-\lambda)\eta(P) + \lambda\eta(Q)\}_\lambda$

Monte Carlo Information-Geometric Structures, Geometric Structures of Information, 2019

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	$\mathcal{M}=\{p(x;\theta)\ :\ \theta\in\Theta^\circ\}$	$\mathcal{M}=\{m(x;\eta)\ :\ \eta\in H^\circ\}$
Convex function ($\equiv ax+b$)	F: cumulant	F^*: negative entropy
Dual coordinates	moment $\eta=E[t(x)]$	$\theta^i=h^\times(p_0:m)-h^\times(p_i:m)$
Fisher Information $g=(g_{ij})_{ij}$	$g_{ij}(\theta)=\partial_i\partial_j F(\theta)$ $g=\mathrm{Var}[t(X)]$	$g_{ij}(\eta)=\int_{\mathcal{X}}\frac{f_i(x)f_j(x)}{m(x;\eta)}\mathrm{d}\mu(x)$
Christoffel symbol	$\Gamma_{ij,k}=\frac{1}{2}\partial_i\partial_j\partial_k F(\theta)$	$g_{ij}(\eta)=-\partial_i\partial_j h(\eta)$ $\Gamma_{ij,k}=-\frac{1}{2}\int_{\mathcal{X}}\frac{f_i(x)f_j(x)f_k(x)}{m^2(x;\eta)}\mathrm{d}\mu(x)$
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

$$\boxed{\mathrm{KL}(m(x;\eta):m(x;\eta')) = B_G(\eta:\eta')}$$

The KL between two mixtures with prescribed components amounts to a Bregman divergence

- Strictly convex and differential convex generator:

$$\boxed{G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta)\log m(x;\eta)\mathrm{d}\mu(x)}$$

- 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 (but not on the closure).

Computational 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: Monte Carlo

- 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**

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

This can be negative
(because "positive" measures)

- 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

Monte Carlo Information-Geometric Structures, Geometric Structures of Information, 2019

Random 1D mixture manifolds

q is a proposal distribution

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, $\tilde{G}_{\mathcal{S}}(\eta)$ is a Bregman generator

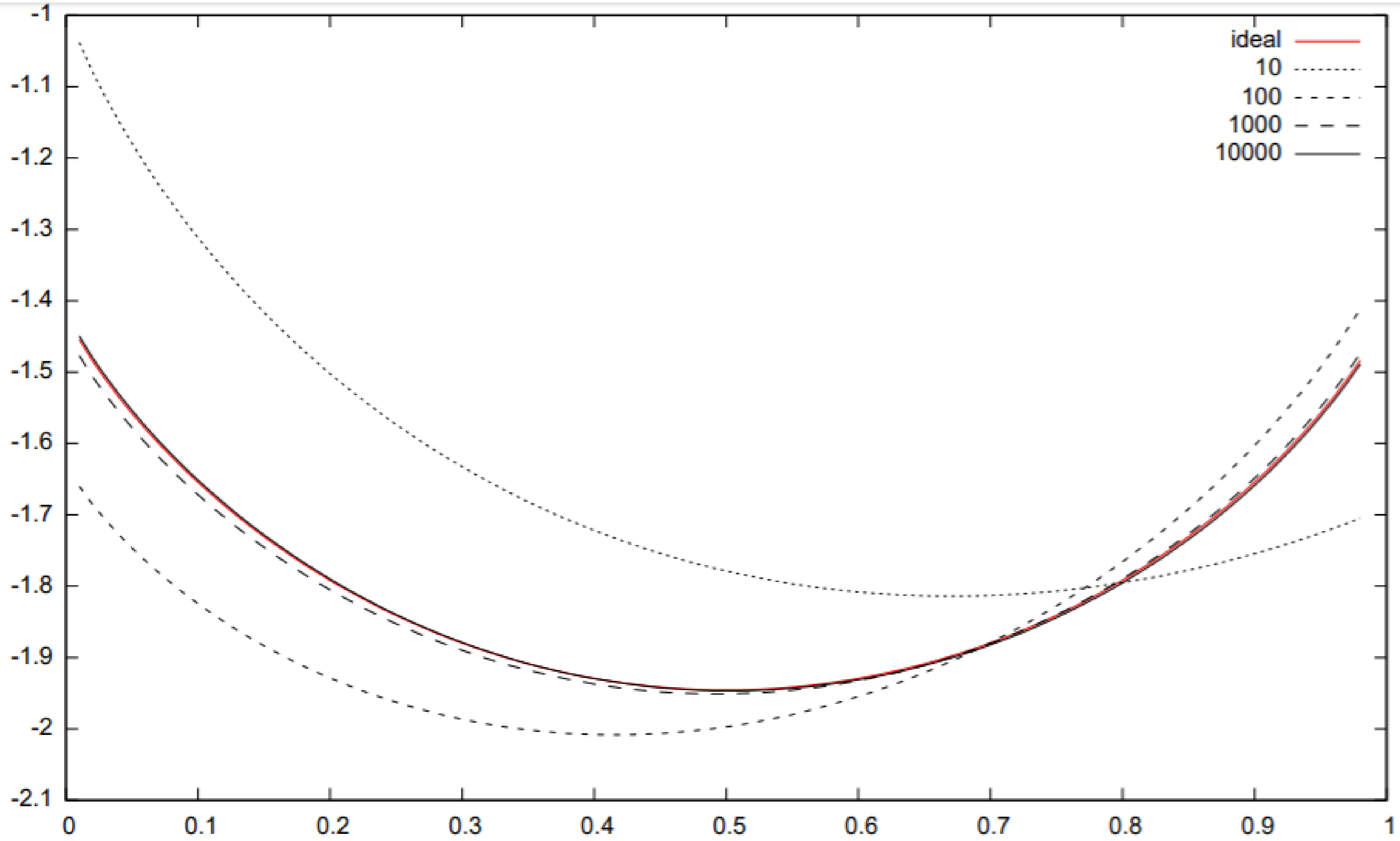


Figure 2: A series $G_{\mathcal{S}}(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x;\eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x;\eta)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

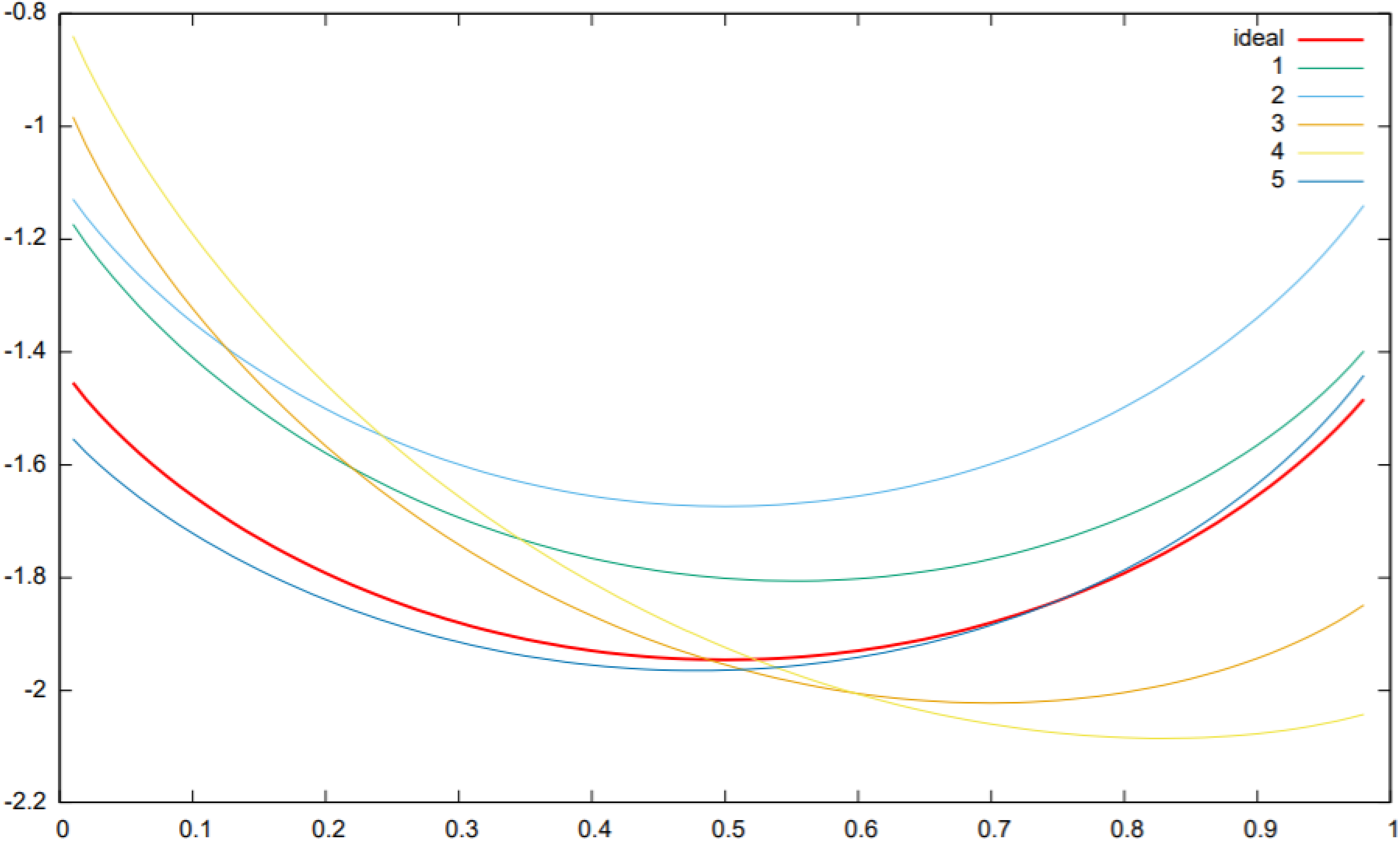


Figure : The Monte Carlo Mixture Family Generator $\hat{G}_{10}$ (MCMFG) considered as a random variable: Here, we show five realizations (i.e., $\mathcal{S}_1, \ldots, \mathcal{S}_5$) of the randomized generator for $m = 5$. The ideal generator is plot in thick red.

Application to clustering Gaussian mixtures (with prescribed Gaussian components)

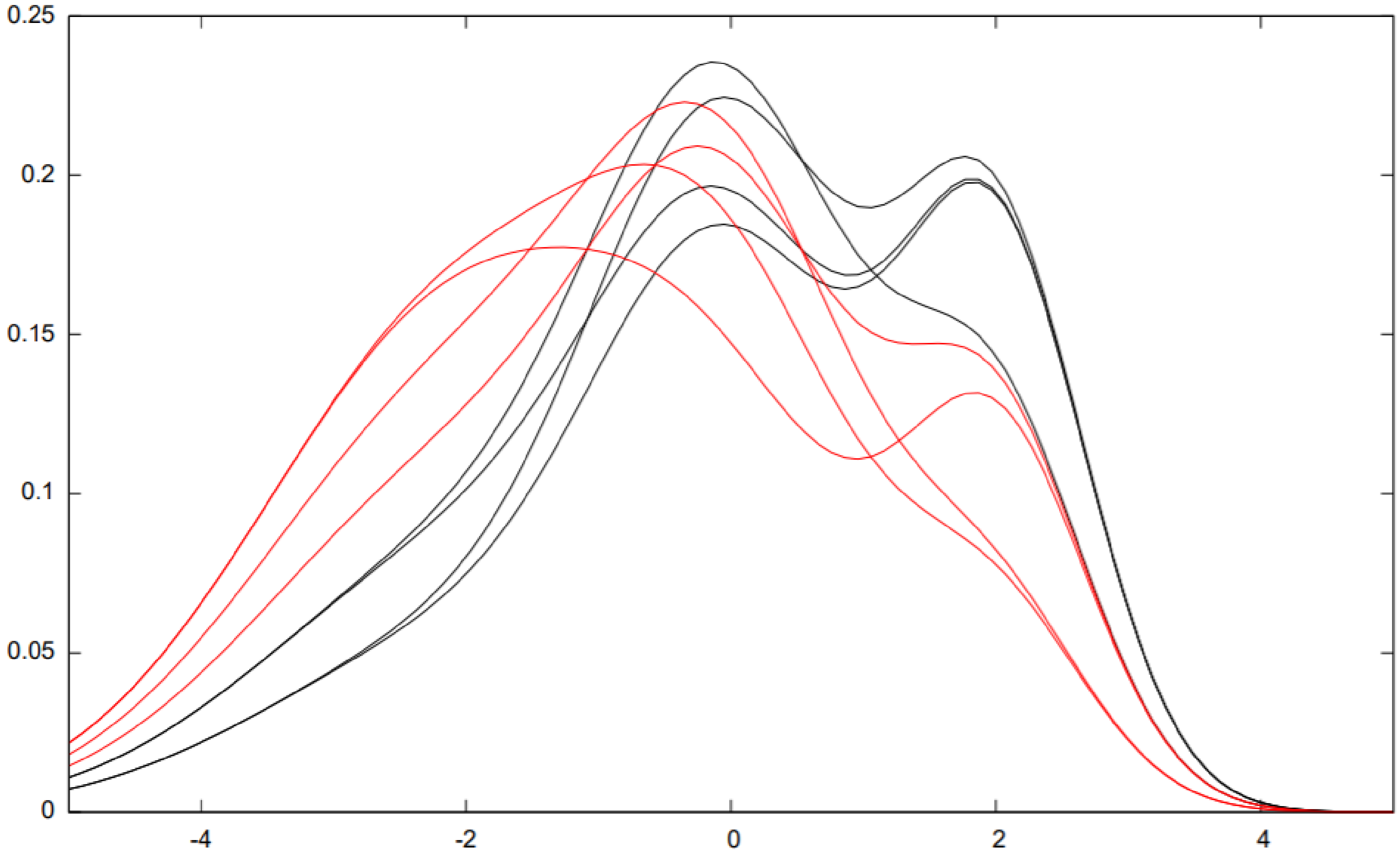


Figure 6: Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point on the mixture family manifold. The Kullback-Leibler divergence is equivalent to an integral-based Bregman divergence that is computationally untractable: The Bregman generator is stochastically approximated by Monte Carlo sampling.

Random d-dimensional mixture manifolds

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)} m(x_i;\eta)\log m(x_i;\eta).$$

$$\partial^i\partial^j \tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{q(x_l)}\frac{(p_i(x_l)-p_0(x_l))(p_j(x_l)-p_0(x_l))}{m(x_l;\eta)}.$$

Theorem (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo multivariate function $\tilde{G}_{\mathcal{S}}(\eta)$ is always convex and twice continuously differentiable, and strictly convex almost surely.*

Monte Carlo Information-Geometric Structures, Geometric Structures of Information, 2019

Random Exponential Family Manifolds

$$\mathcal{E} := \{p(x;\theta) = \exp(t(x)\theta - F(\theta) + k(x)) : \theta \in \Theta\}$$

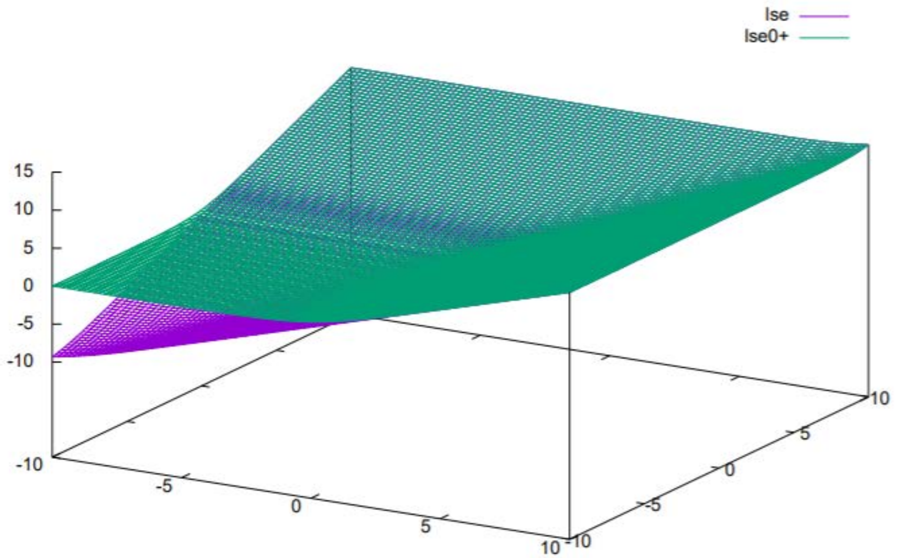
$$F(\theta) = \log\left(\int \exp(t(x)\theta + k(x))\mathrm{d}\mu(x)\right)$$

Computationally intractable

$$F(\theta) \simeq \tilde{F}^{\dagger}_{\mathcal{S}}(\theta) := \log\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\exp(t(x_i)\theta + k(x_i))\right)$$

$$\begin{aligned}
\tilde{F}^{\dagger}_{\mathcal{S}}(\theta) &\equiv \tilde{F}_{\mathcal{S}}(\theta),\\
\tilde{F}_{\mathcal{S}}(\theta) &= \log\left(1 + \sum_{i=2}^{m}\exp((t(x_i) - t(x_1))\theta + k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1))\right)\\
&= \log\left(1 + \sum_{i=2}^{m}\exp(a_i\theta + b_i)\right),\\
&:= \mathrm{lse}_0^{+}(a_2\theta + b_2, \ldots, a_m\theta + b_m),
\end{aligned}$$

Log-sum-exp modified function to ensure always strict convexity



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)

$$D_\gamma(p,q) = \frac{1}{\gamma(1+\gamma)} \log I_\gamma(p,p) - \frac{1}{\gamma} \log I_\gamma(p,q) + \frac{1}{1+\gamma} \log I_\gamma(q,q),$$

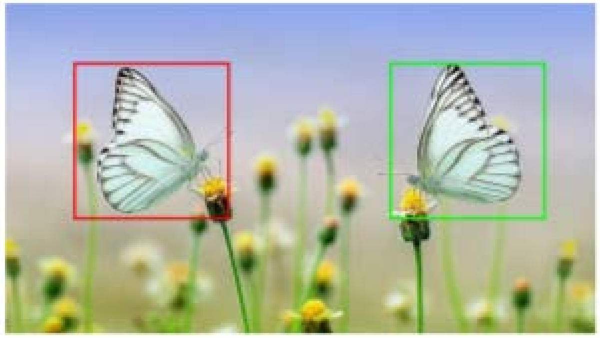
where

$$I_\gamma(p,q) = \int_{x\in\mathcal{X}} p(x)q(x)^\gamma \mathrm{d}x.$$

When $\gamma \to 0$, $D_\gamma(p,q) \to \mathrm{KL}(p,q)$.

$$I_\gamma(\theta_p,\theta_q) = \exp\left(F(\theta_p + \gamma\theta_q) - F(\theta_p) - \gamma F(\theta_q)\right).$$

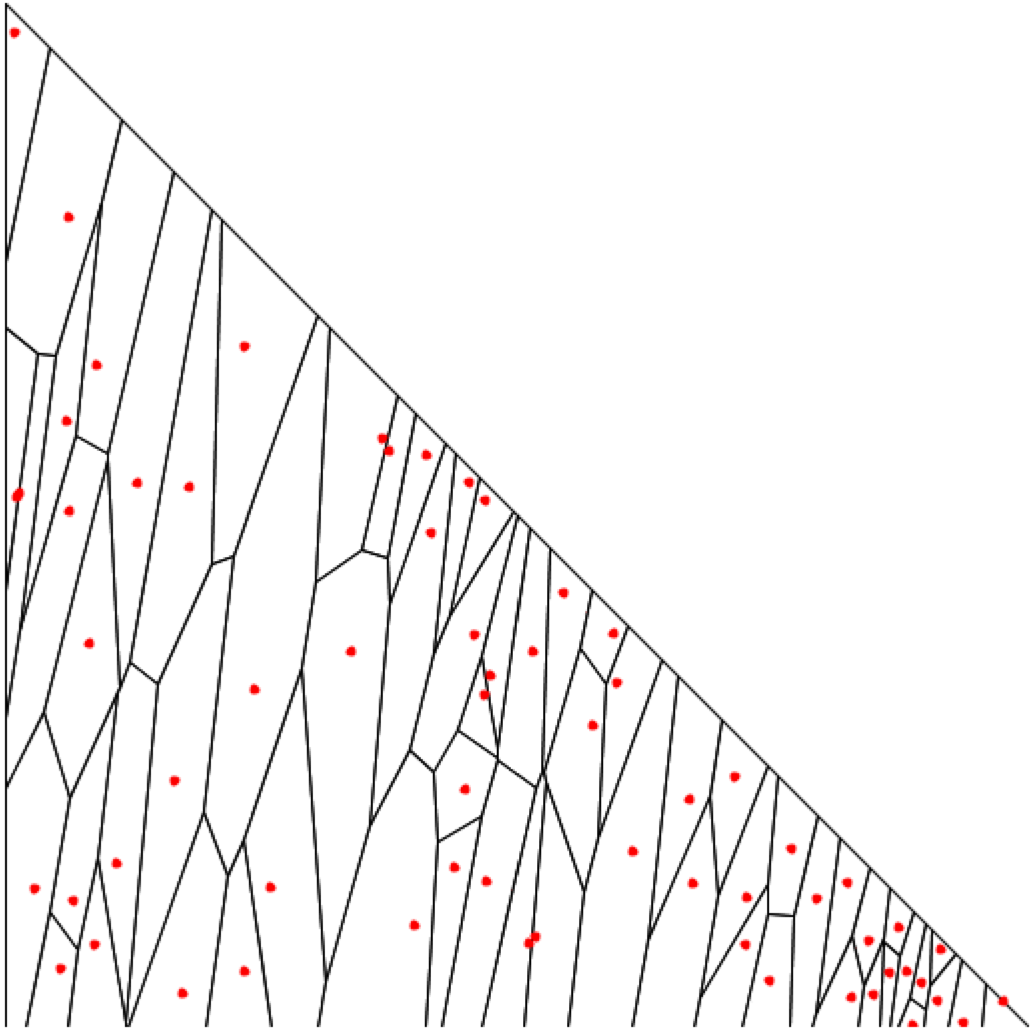
$$I_\gamma(p,q) = \int_{x\in\mathcal{X}} p(x)q(x)^\gamma \mathrm{d}x \simeq \frac{1}{m}\sum_{i=1}^m q(x_i)^\gamma$$

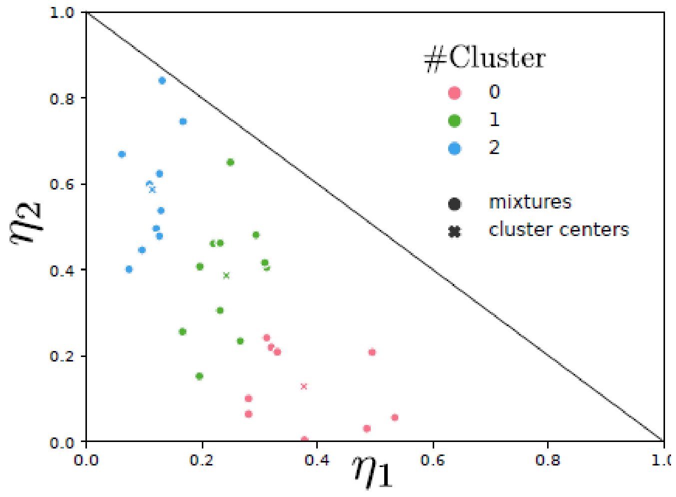


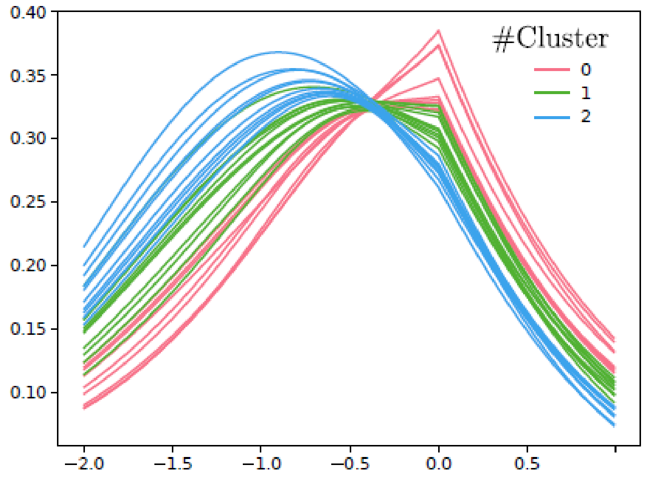
Patch matching with polynomial exponential families and projective divergences.
International Conference on Similarity Search and Applications (SISAP). 2016

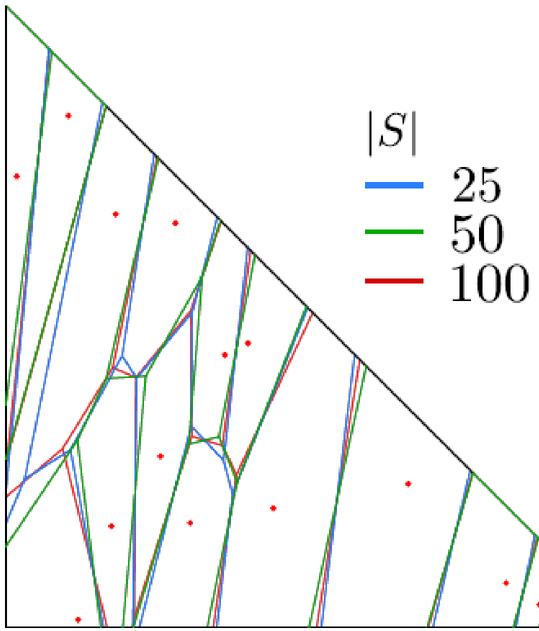
Random/Monte Carlo Bregman Voronoi diagrams

$$p_1 = \mathrm{Laplace}(0,1), p_2 = \mathcal{N}(-1,1), p_0 = \mathrm{Cauchy}(-0.5,1).$$









Some statistical distances with closed-form expressions for statistical mixtures

- **Cauchy-Schwarz divergence**: $\mathrm{CS}(P:Q) = -\log \frac{\int p(x)q(x)\mathrm{d}x}{\sqrt{\int p(x)^2\mathrm{d}x \int q(x)^2\mathrm{d}x}},$
- For mixtures of exponential families with conic natural parameter space:

$$\int m(x)m'(x)\mathrm{d}x = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j \int p_F(x;\theta_i)p_F(x;\theta'_j)\mathrm{d}x$$

$$\int p_F(x;\theta_i)p_F(x;\theta'_j)\mathrm{d}x = e^{F(\theta_i+\theta'_j)-(F(\theta_i)+F(\theta'_j))} \underbrace{\int e^{\langle t(x),\theta_i+\theta'_j\rangle - F(\theta_i+\theta'_j)}\mathrm{d}x}_{=1},$$

When natural parameter space
Is a cone

$$\int m(x)m'(x)\mathrm{d}x = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j e^{F(\theta_i+\theta'_j)-(F(\theta_i)+F(\theta'_j))}$$

Closed-form information-theoretic divergences for statistical mixtures, ICPR 2012.

Examples of conic exponential families (CEFs)

$$\int m(x)m'(x)\mathrm{d}x = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j e^{\Delta_F(\theta_i,\theta'_j)}, \qquad \Delta_F(\theta_i,\theta'_j) = F(\theta_i+\theta'_j) - (F(\theta_i)+F(\theta'_j)).$$

Bernoulli. $p(x;\lambda) = \lambda^x(1-\lambda)^{1-x}$ (with $\lambda \in (0,1)$), $\theta = \log\frac{\lambda}{1-\lambda}$, $\Theta = \mathbb{R}$, $F(\theta) = \log(1+e^\theta)$.

$$\Delta_{\mathrm{Bernoulli}}(\lambda_i,\lambda_j) = \log\frac{1+\frac{\lambda_i+\lambda_j}{1-\lambda_i-\lambda_j}}{(1+\frac{\lambda_i}{1-\lambda_i})(1+\frac{\lambda_j}{1-\lambda_j})}$$

Zero-centered Laplacian. $p(x;\sigma) = \frac{1}{2\sigma}e^{-\frac{|x|}{\sigma}}$, $\theta = -\frac{1}{\sigma}$, $\Theta = (-\infty,0)$, $F(\theta) = \log(\frac{2}{-\theta})$.

$$\Delta_{\mathrm{Laplacian}}(\sigma_i,\sigma_j) = \log\frac{1}{2(\sigma_i+\sigma_j)}$$

Gaussian. $p(x;\mu,\Sigma) =$

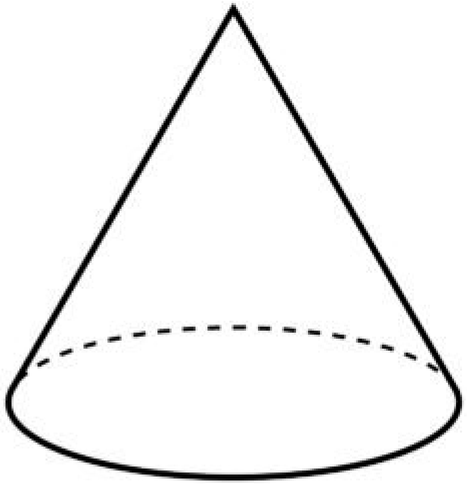
$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right),$$

$\theta = (\theta_v,\theta_M) = (\Sigma^{-1}\mu,\Sigma^{-1})$, $\Theta = \mathbb{R}^d \times S^d_{++}$ where S^d_{++} denotes the cone of positive definite matrices of dimension $d\times d$,

$$F(\theta) = \frac{1}{2}\theta_v^T\theta_M^{-1}\theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log 2\pi.$$

$$\Delta_{\mathrm{Gaussian}}((\mu_i,\Sigma_i),(\mu_j,\Sigma_j)) = \frac{1}{2}\Big(\mu_{ij}^T\Sigma_{ij}^{-1}\mu_{ij} - (\mu_i^T\Sigma_i^{-1}\mu_i + \mu_j^T\Sigma_j^{-1}\mu_j) - \log\frac{|\Sigma_i^{-1}+\Sigma_j^{-1}|}{|\Sigma_i^{-1}||\Sigma_j^{-1}|} - d\log 2\pi \Big)$$

Wishart $p(x;n,S) = \frac{|X|^{\frac{n-d-1}{2}}e^{-\frac{1}{2}\mathrm{tr}(S^{-1}X)}}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})}$, 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. $F(\theta) = \frac{(2\theta_s+d+1)d}{2}\log 2 + (\theta_s + \frac{d+1}{2})\log|\theta_M| + \log\Gamma_d(\theta_s+\frac{d+1}{2})$.



Some applications of information geometry:

- Natural gradient and deep learning
- Bayesian hypothesis testing

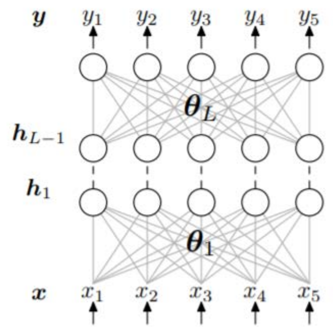
  geometry of the error exponent
- Clustering

  partition-based, soft mixtures and hierarchical

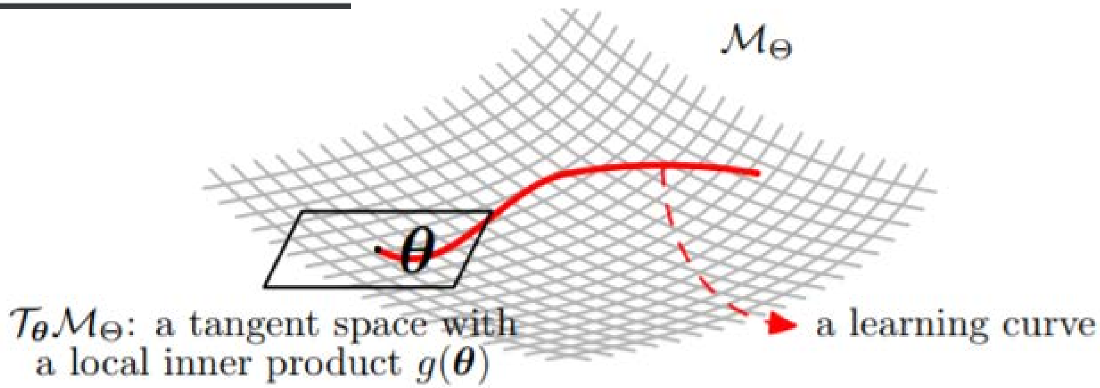
Frank Nielsen

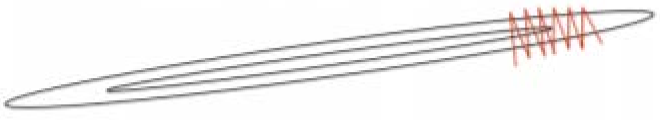
Sony CSL

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \Theta) = \sum_{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{L-1}} p(\boldsymbol{y} \mid \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_L) \cdots p(\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \boldsymbol{\theta}_2) p(\boldsymbol{h}_1 \mid \boldsymbol{x}, \boldsymbol{\theta}_1),$$

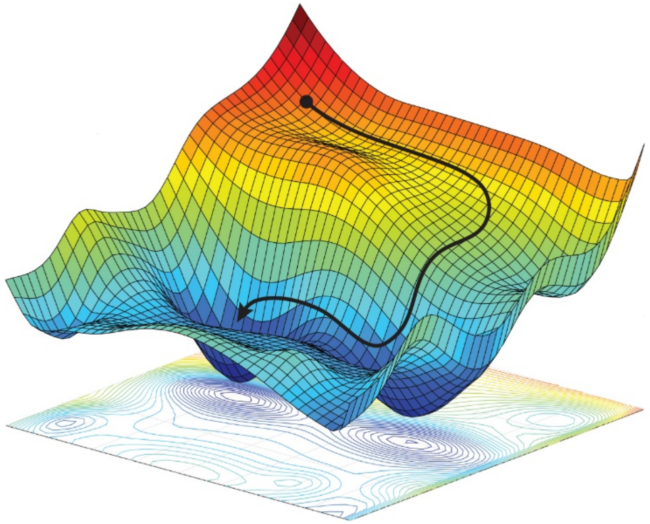


Natural gradient and mirror descent



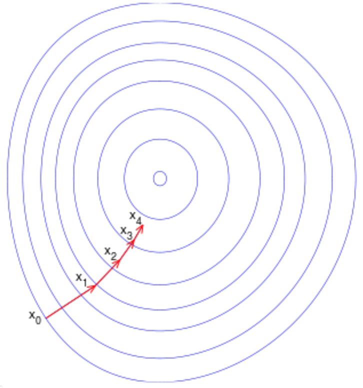


Frank Nielsen



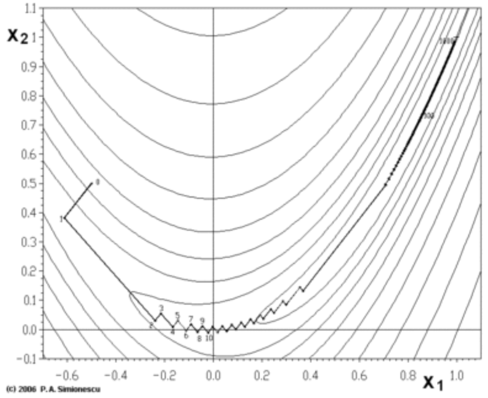
Sony CSL

Steepest gradient descent method

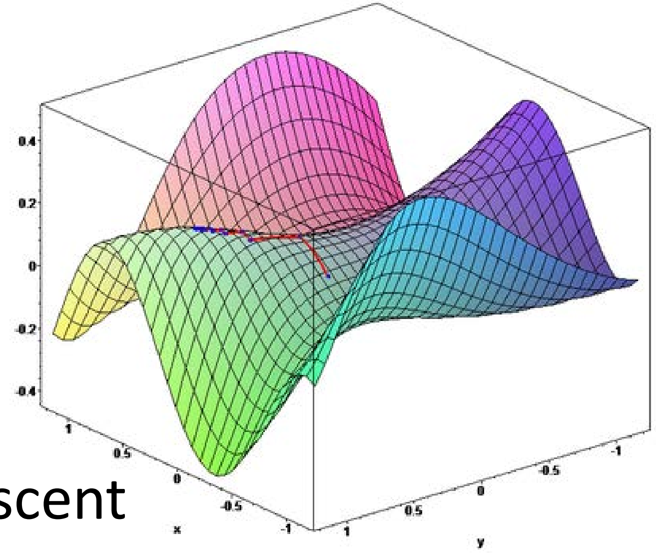


- Iterative optimization algorithm
- Start from an initial parameter value θ_0
- **Update iteratively** the current parameter using a learning rate α (step size) and the gradient of the energy function:

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



- First-order optimization method
- Zig-zag local minimum convergence
- Stopping criterion



Similarly, maximization with hill climbing, steepest ascent

Steepest descent in a Riemannian space

- The steepest descent direction of E(θ) 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

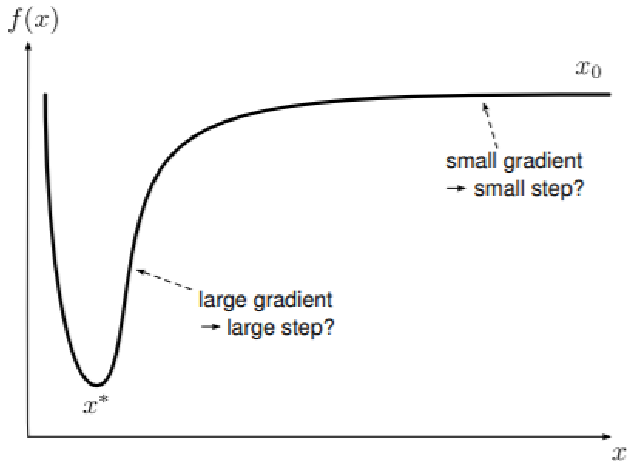
Learning rate

Computing the inverse of the Fisher information matrix is tricky

Amari, Shun-Ichi. "Natural gradient works efficiently in learning." *Neural computation* 10.2 (1998): 251-276.

Pros and cons of natural gradient

- **Pros:**
  - Invariant (intrinsic) gradient (at infinitesimal scale/ODE)
  - Not trapped in plateaus
  - Achieve Fisher efficiency in online learning



- **Cons:**
  - Too expensive to compute (no closed-form FIM; need matrix inversion; numerical stability)
  - Degenerate for *irregular models* (e.g., hierarchical models, Deep learning)
  - Need to adapt step size

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

In a dually flat space (Hessian manifold), 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 inference (VI)

Zhang, Guodong, et al. "Noisy natural gradient as variational inference." ***arXiv:1712.02390*** **(2017).**

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 <u>any Bregman divergence:</u>

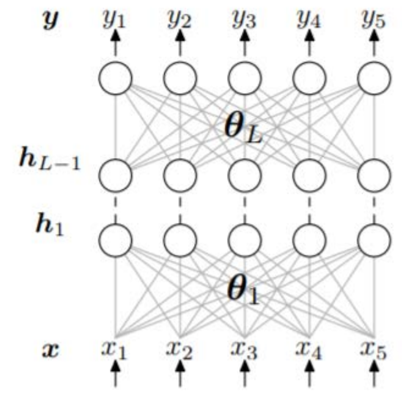
$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k} B_F(x : x_k) \right\}$$

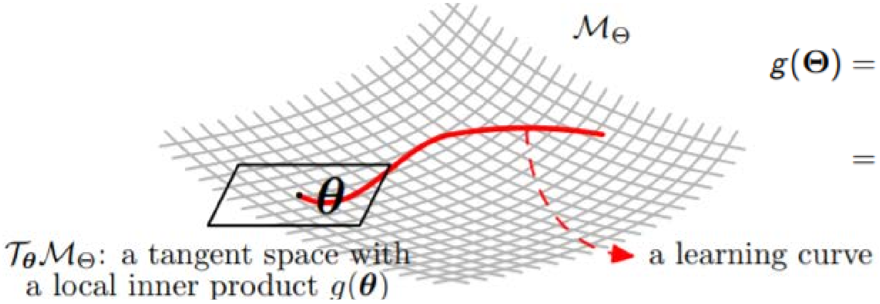
Thus mirror descent for the Bregman divergence on the primal parameter amounts to natural gradient for the dual parameter

Garvesh Raskutti, Sayan Mukherjee: The Information Geometry of Mirror Descent. IEEE Trans. Information Theory 61(3): 1451-1457 (2015)

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

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \Theta) = \sum_{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{L-1}} p(\boldsymbol{y} \mid \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_L) \cdots p(\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \boldsymbol{\theta}_2) p(\boldsymbol{h}_1 \mid \boldsymbol{x}, \boldsymbol{\theta}_1),$$

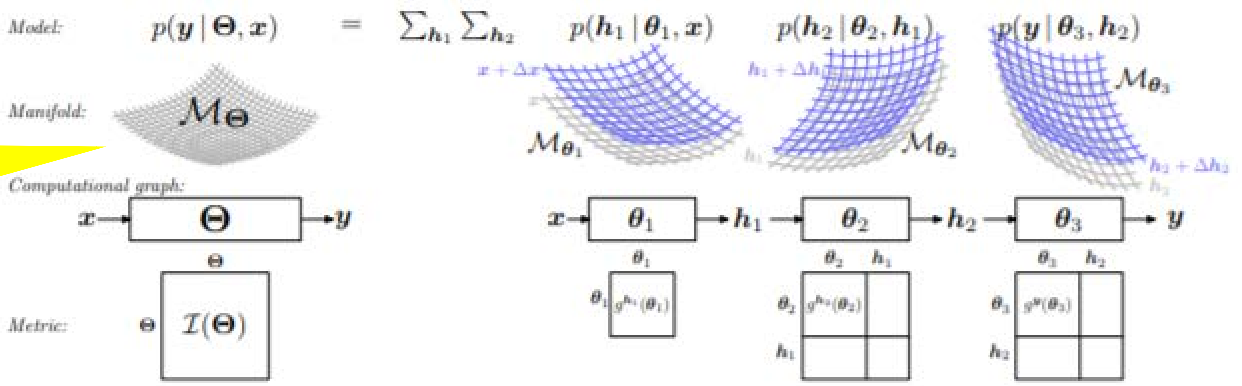




$$g(\Theta) = E_{\boldsymbol{x} \sim \hat{p}(X_n), \boldsymbol{y} \sim p(\boldsymbol{y} \mid \boldsymbol{x}, \Theta)} \left[\frac{\partial l}{\partial \Theta} \frac{\partial l}{\partial \Theta^\intercal} \right] = \frac{1}{n} \sum_{i=1}^{n} E_{p(\boldsymbol{y} \mid \boldsymbol{x}_i, \Theta)} \left[\frac{\partial l_i}{\partial \Theta} \frac{\partial l_i}{\partial \Theta^\intercal} \right]$$

Relative Fisher IM: $$g^{\boldsymbol{h}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_f) = E_{p(\boldsymbol{h} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_f)} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{h} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_f) \frac{\partial}{\partial \boldsymbol{\theta}^\intercal} \ln p(\boldsymbol{h} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_f) \right]$$

Dynamic geometry



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

Relative Fisher Information and Natural Gradient for Learning Large Modular Models (ICML'17)

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)

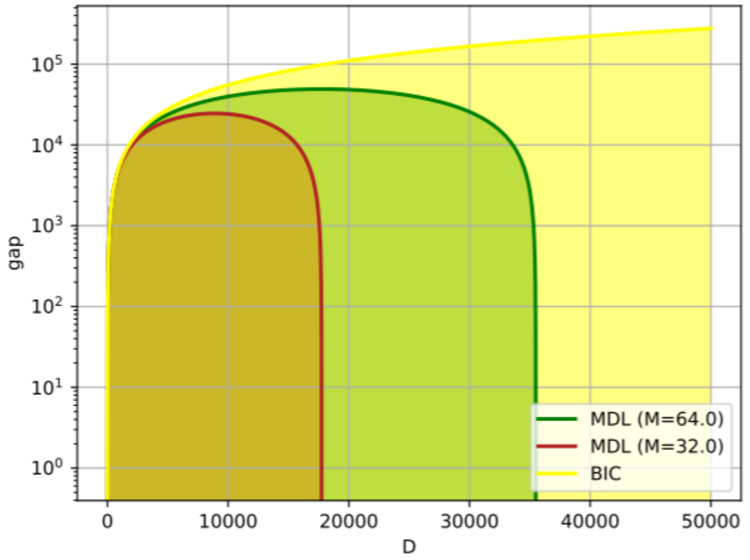
Occam's razor for Deep Neural Networks (DNNs):
(uniform width M, L layers, N #observations, d: dimension of screen distributions in lightlike neuromanifold)
Θ: parameters of the DNN, $\hat{\Theta}$: estimated parameters

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

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

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

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



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

https://arxiv.org/abs/1905.11027

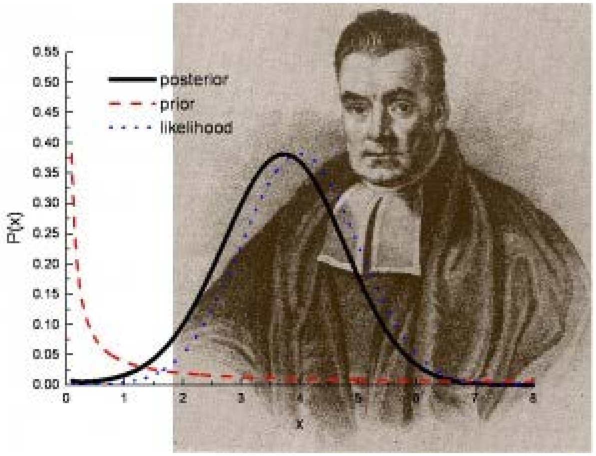
Summary

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

Thomas, Philip, et al. "Energetic natural gradient descent." International Conference on Machine Learning. 2016.

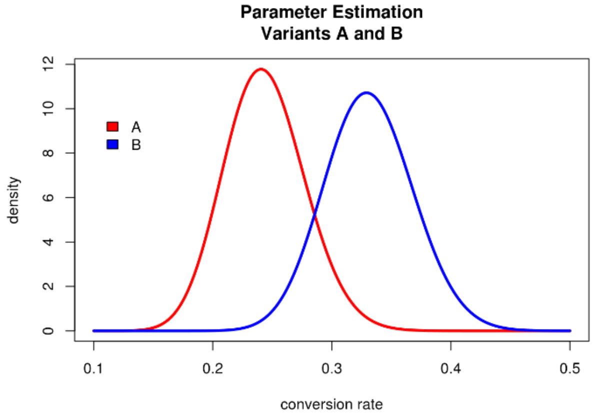
Information geometry of Bayesian binary/multiple hypothesis testing

Detecting signal from noise



Frank Nielsen

Sony CSL



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)

Recalling Bayes' rule

Using probability's chain rule:

$$P(X,Y) = P(X)P(Y|X) = P(Y)P(X|Y)$$

Reverend Thomas Bayes (1701-1761)

Get **Bayes' rule**:

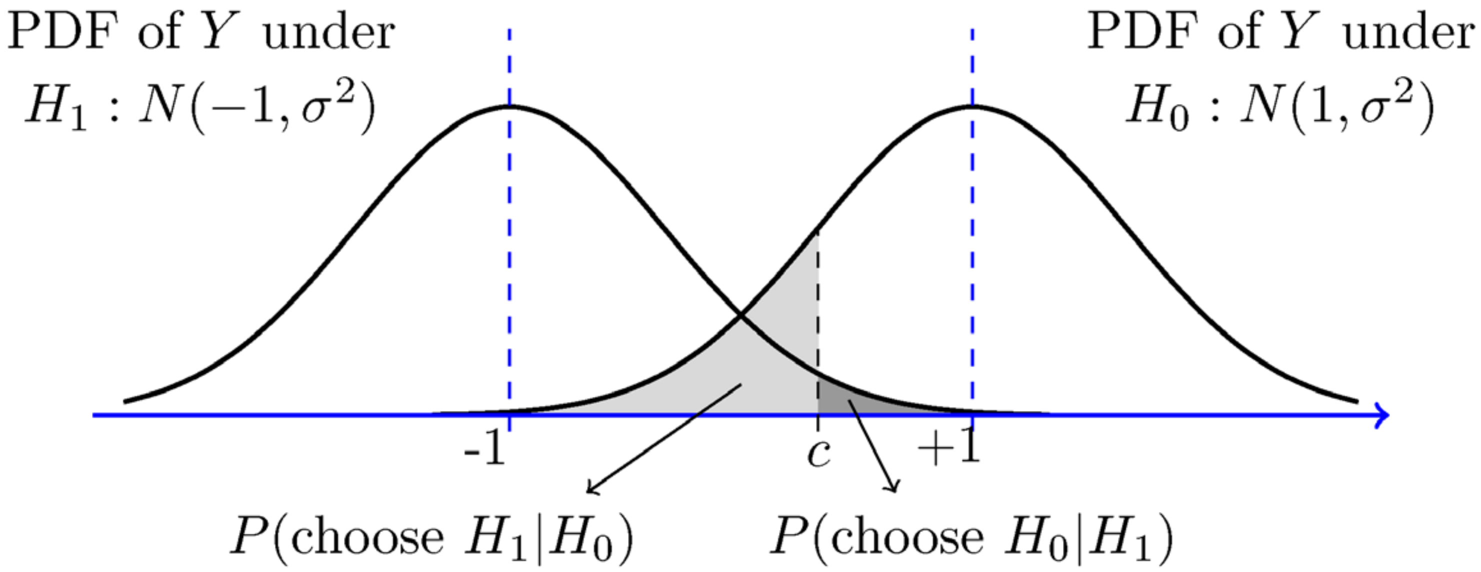
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Interpreted as:

- P(A|B): **conditional probability** = likelihood of event A occurring given that B is true.
- P(B|A): **conditional probability** = likelihood of event B occurring given that A is true.
- P(A) and P(B) are the probabilities of observing A and B independently of each other = **marginal probability** of A and B

Setting for the Bayesian binary hypothesis testing

- Given an iid sample set, **decide** whether it emanates from the distribution of the null hypothesis H0 or the alternative hypothesis H1 -> unavoidable **probability of error**

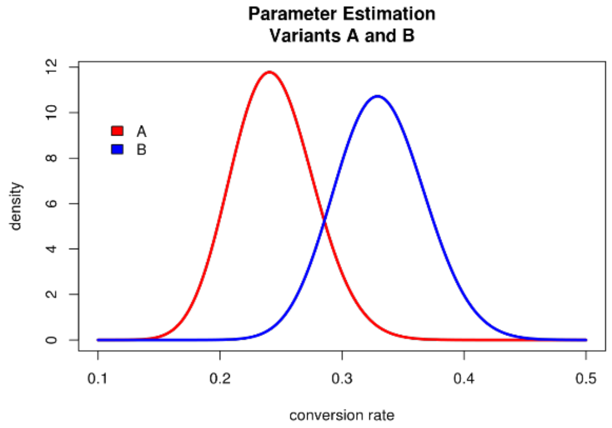


Among the many decision rules, the best rule is provably the **Maximum A Posteriori** (MAP) rule:

$$P(H_0 | X = x) \geq P(H_1 | X = x)$$

Probability of error

(Bayes' error for diagonal cost matrix)



- Confusion matrix
- Cost design matrix, where errors uniformly account (diagonal matrix)
- **Probability of error:**

$$P_{\text{error}} = P\left(\text{ choose } H_1 | H_0\right) P\left(H_0\right) + P\left(\text{ choose } H_0 | H_1\right) P\left(H_1\right)$$

- **A priori probabilities** of classes: w0=P(H0) and w1=P(H1)
- **Theorem: MAP rule minimizes the probability of error among all decision rules:** $\mathrm{MAP}(x) = \mathrm{argmax}_{i \in \{1,\ldots,n\}} w_i p_i(x)$

Class conditional probabilities

Probability of error with equal priors (w1=w2=1/2)

$$P_{error} = \int_{x\in\mathcal{X}} p(x) \min\left(\Pr\left(H_1|x\right), \Pr\left(H_2|x\right)\right) \mathrm{d}\nu(x)$$

From Bayes' rule: $\Pr\left(H_i|X = x\right) = \frac{\Pr(H_i)\Pr(X=x|H_i)}{\Pr(X=x)} = \frac{w_i p_i(x)}{p(x)}$

It follows that we have:

$$P_{\mathrm{error}} = \frac{1}{2}\int_{x\in\mathcal{X}} \min\left(p_1(x), p_2(x)\right) \mathrm{d}\nu(x)$$

This is also called **histogram intersection similarity** in computer vision

Bounding the probability of error

Trick: $\boxed{\min(a,b) \leq \min_{\alpha\in(0,1)} a^\alpha b^{1-\alpha}}$ for $a, b > 0$, *upper bound* P_e:

$$
\begin{aligned}
P_e &= \frac{1}{2}\int_{x\in\mathcal{X}} \min(p_1(x), p_2(x))\mathrm{d}\nu(x) \\
&\leq \frac{1}{2}\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x)\mathrm{d}\nu(x).
\end{aligned}
$$

Define **Chernoff information** :

$$
\boxed{C(P_1,P_2) = -\log \min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x)\mathrm{d}\nu(x) \geq 0,}
$$

For alpha=1/2, we get the Bhattacharyya distance, skewed Bhattacharyya distance: $B_\alpha(p,q) = -\ln \int_x p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x$

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

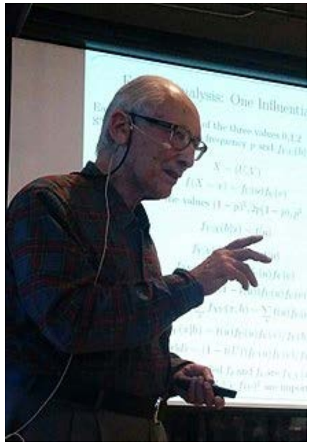
- For m iid samples $$P^m_{\text{correct}} = 1 - P^m_{\text{error}} = 1 - P^m_e$$

- **Asymptotic regime** when m->oo $$\alpha = -\frac{1}{m}\log P^m_e$$

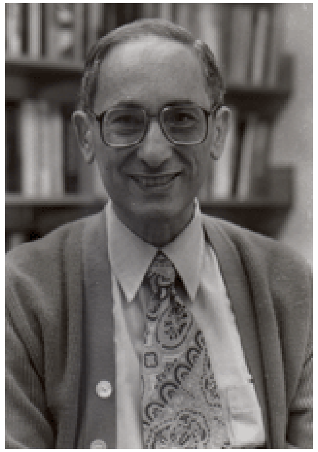
- **Best error exponent**:

$$P_e \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C(P_1,P_2)} \leq e^{-C(P_1,P_2)}$$

$$C(P_1, P_2) = -\log \min_{\alpha\in(0,1)} \int_{x\in\mathcal{X}} p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}\nu(x) \geq 0,$$



Herman Chernoff
(1923, 95 yo)
© photo 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**

$$\begin{aligned} \mathrm{MAP}(x) &= \mathrm{argmax}_{i\in\{1,\ldots,n\}} w_i p_i(x) \\ &= \arg\min_{i\in\{1,\ldots,n\}} B^*(t(x) : \eta_i) - \log w_i \end{aligned}$$

Geometry of the best error exponent

$$\boxed{c_\alpha(P_{\theta_1}:P_{\theta_2}) = \int p_{\theta_1}^{\alpha}(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1:\theta_2)),}$$

Jensen divergence:

$$\boxed{J_F^{(\alpha)}(\theta_1:\theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\theta_{12}^{(\alpha)}),}$$

Theorem: At best exponent, the Chernoff information amounts to an <u>equivalent Bregman divergence</u>:

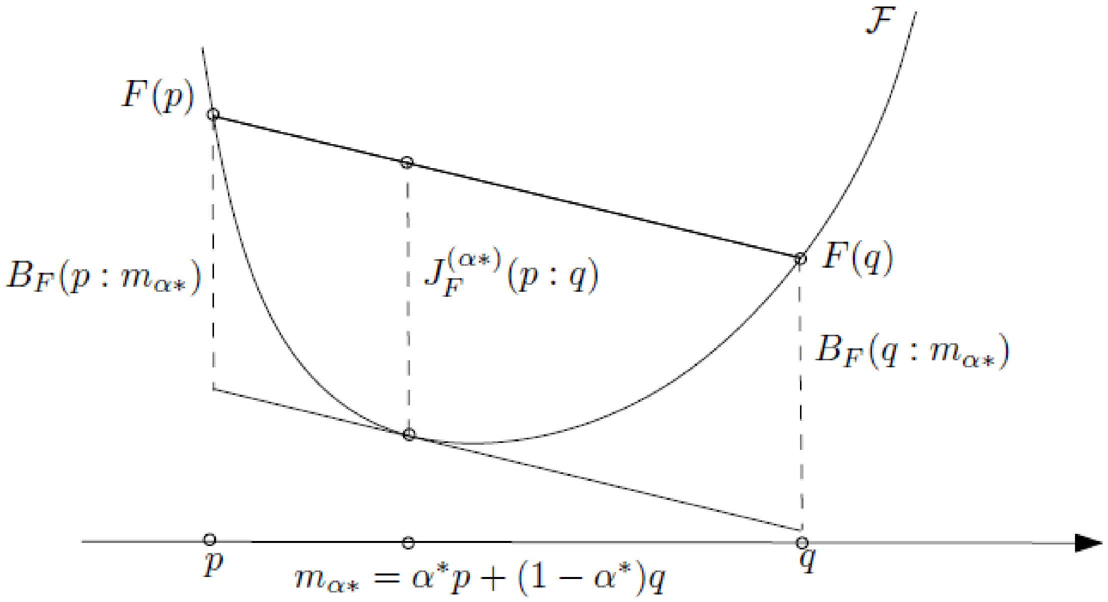
$$C(P_{\theta_1}:P_{\theta_2}) = B(\theta_1:\theta_{12}^{(\alpha^*)}) = B(\theta_2:\theta_{12}^{(\alpha^*)})$$

Visualizing that maximizing skew Jensen divergence yields a Bregman divergence

$$\boxed{\alpha^* = \arg \max_{0<\alpha<1} J_F^{(\alpha)}(p : q)}$$

$$\boxed{J_F^{(\alpha^*)}(p : q) = B_F(p : m_{\alpha^*}) = B_F(q : m_{\alpha^*})}$$

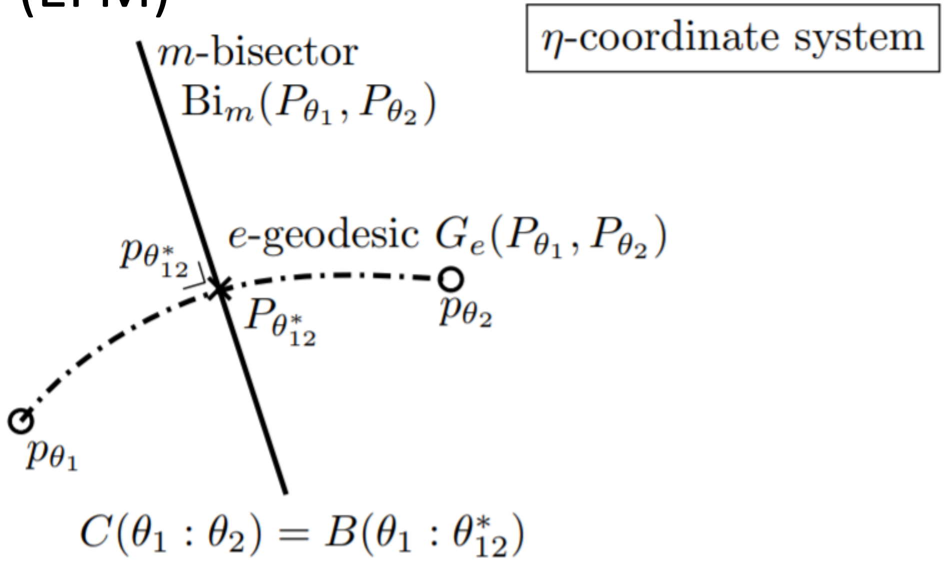
$m_\alpha = \alpha p + (1-\alpha)q$: α-mixing of p and q.



Bayesian hypothesis testing:
Geometric characterization of the best error exponent

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

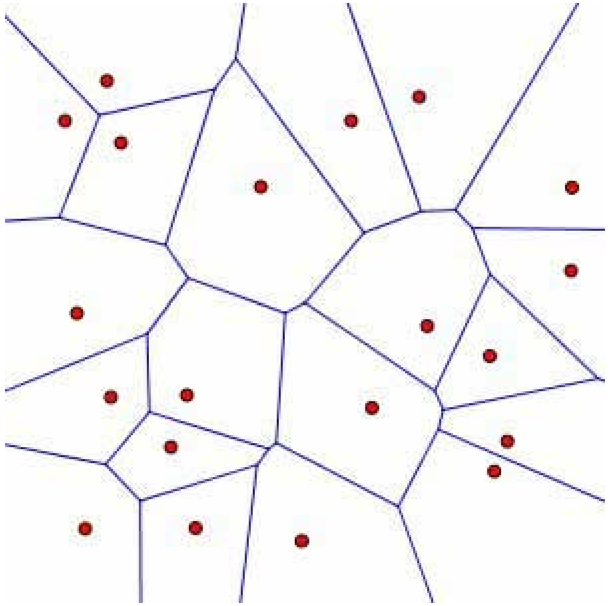
Dually flat Exponential Family Manifold (EFM)



This characterization yields to an <u>exact closed-form solution in 1D</u> EFs, and a <u>simple geodesic bisection</u> search for arbitrary dimension

An Information-Geometric Characterization of Chernoff Information, IEEE SPL, 2013 (arXiv:1102.2684)

Multiple hypothesis testing

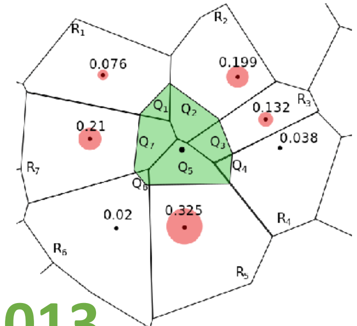


- Minimum **pairwise** Chernoff information distance

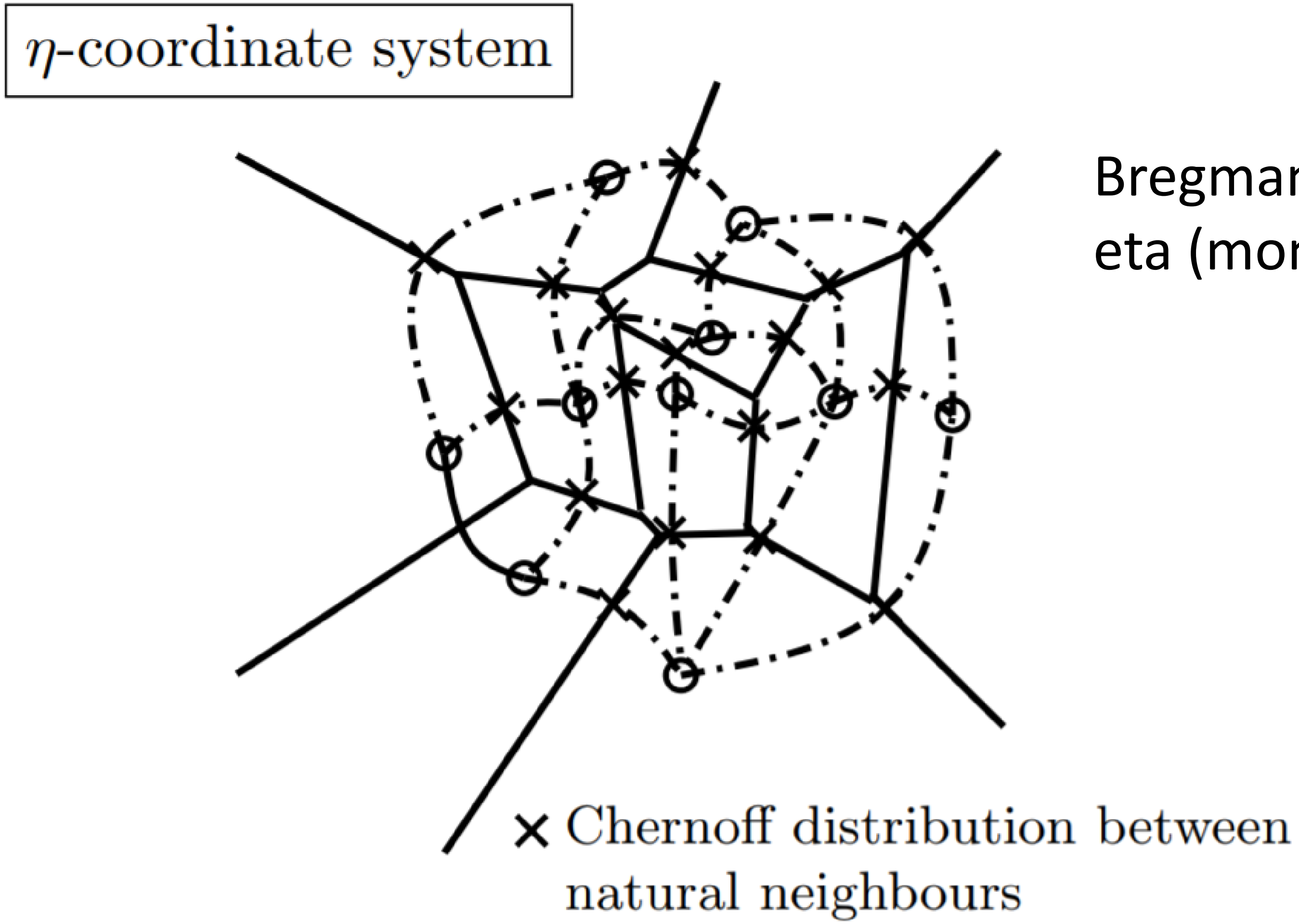
$$\boxed{C(P_1, \ldots, P_n) = \min_{i,j \neq i} C(P_i, P_j)}$$

$$P_e^m \leq e^{-mC(P_{i^*}, P_{j^*})}, \quad (i^*, j^*) = \arg\min_{i,j \neq i} C(P_i, P_j)$$

- In the (additive) **Bregman Voronoi diagram**, check only the **natural neighbors** (with Voronoi cells sharing a common facet)



Multiple hypothesis testing on EFM



Bregman Voronoi diagram is affine in the eta (moment/expectation) coordinate system

Natural neighbors

Hypothesis testing, information divergence and computational geometry, GSI 2013

Link between the Probability of error and the Total Variation (TV) distance:

Use the trick $\min(a,b) = \frac{a+b}{2} - \frac{1}{2}|b-a|,$

$$P_{\mathrm{error}} = \frac{1}{2} \int_{x \in \mathcal{X}} \min\left(p_1(x), p_2(x)\right) \mathrm{d}\nu(x)$$

$$P_e = \frac{1}{2} - \mathrm{TV}(w_1 p_1, w_2 p_2).$$

$$P_e = \frac{1}{2}(1 - \mathrm{TV}(p_1, p_2)).$$ (same weights here)

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. Pattern Recognition Letters (2014)

Computing Total Variation can be difficult…

Pe between two multivariate Gaussians with same positive semi-definite covariance matrix

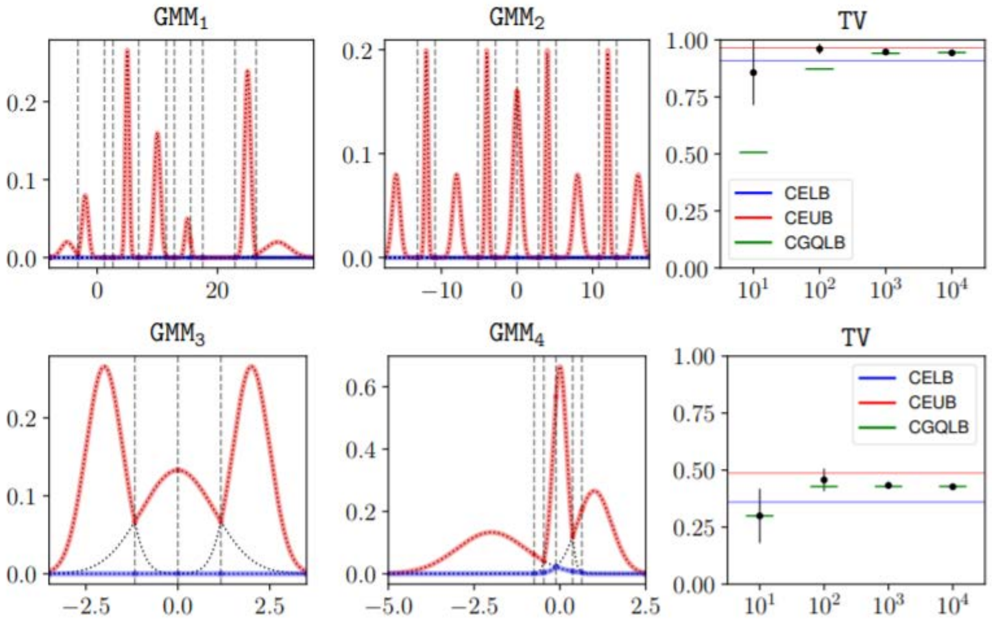
$$\mathrm{TV}(p_1,p_2) = \frac{1}{2}\left|\mathrm{erf}\left(\frac{x_1-\mu_1}{\sigma_1\sqrt{2}}\right) - \mathrm{erf}\left(\frac{x_1-\mu_2}{\sigma_2\sqrt{2}}\right)\right| + \frac{1}{2}\left|\mathrm{erf}\left(\frac{x_2-\mu_1}{\sigma_1\sqrt{2}}\right) - \mathrm{erf}\left(\frac{x_2-\mu_2}{\sigma_2\sqrt{2}}\right)\right|$$

$$P_e = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{1}{2\sqrt{2}}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|\right)$$

$$\frac{1}{2}|a-b| = \frac{a+b}{2} - \min(a,b) = \max(a,b) - \frac{a+b}{2},$$

$$\mathrm{TV}(p,q) = \int_{\mathcal{X}} \left(\frac{p(x)+q(x)}{2} - \min(p(x),q(x))\right)\mathrm{d}\mu(x),$$

$$= 1 - \int_{\mathcal{X}} \min(p(x),q(x))\mathrm{d}\mu(x) = \int_{\mathcal{X}} \max(p(x),q(x))\mathrm{d}\mu(x) - 1.$$



Guaranteed Deterministic Bounds on the total variation Distance between univariate mixtures, MLSP 2019

From geometric mean to other abstract means

Remember the trick:
Geometric weighted mean
is greater than the minimum

$$\begin{aligned} P_e &= \frac{1}{2}\int_{x\in\mathcal{X}} \min(p_1(x), p_2(x))\mathrm{d}\nu(x) \\ &\leq \frac{1}{2}\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x). \end{aligned}$$

Internness property of any mean abstract M:

$$\boxed{\min(a,b) \leqslant M(a,b) \leqslant \max(a,b)}$$

Consider **quasi arithmetic means** for a strictly monotone function f (with well-defined inverse function)

$$M_f(a,b;\alpha) = f^{-1}\left(\alpha f(a) + (1-\alpha) f(b)\right)$$

Abstract weighted means: f-means (quasi-arithmetic)

$$\inf\{x,y\} \leq M(x,y) \leq \sup\{x,y\}, \quad \forall x,y \in I.$$

$$M_\alpha^h(x,y) := h^{-1}\left((1-\alpha)h(x) + \alpha h(y)\right)$$

Weighted arithmetic mean: $A_\alpha(x,y) = (1-\alpha)x + \alpha y$

Weighted geometric mean: $G_\alpha(x,y) = x^{1-\alpha} y^\alpha$

Weighted harmonic mean: $H_\alpha(x,y) = \frac{xy}{(1-\alpha)y+\alpha x}$

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)

Chernoff information with quasi-arithmetic means

$$M_f(a,b;\alpha) = f^{-1}\left(\alpha f(a) + (1-\alpha) f(b)\right)$$

Definition 2. The Chernoff-type information for a strictly monotonous function *f* is defined by:

$$\boxed{\begin{aligned} C_f(p_1,p_2) &= -\log \rho_*^f(p_1,p_2) \\ &= \max_{\alpha\in[0,1]} -\log \int M_f(p_1(x),p_2(x);\alpha)\mathrm{d}x \geqslant 0. \end{aligned}}$$

Andrey Nikolaevich Kolmogorov, Sur la notion de la moyenne (1930)
Mitio Nagumo, Über eine Klasse der Mittelwerte (1930)
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)

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:

$$M_f(w_1p_1(x), w_2p_2(x); \alpha) = e^{\alpha \log w_1 p_1(x) + (1-\alpha)\log w_2 p_2(x)}, \tag{60}$$

$$= w_1^\alpha w_2^{1-\alpha} p_1^\alpha(x) p_2^{1-\alpha}(x). \tag{61}$$

$$\begin{aligned} f^{-1}(m_\alpha(x;\theta_1,\theta_2)) &= e^{F(\alpha\theta_1+(1-\alpha)\theta_2)-\alpha F(\theta_1)-(1-\alpha)F(\theta_2)} \\ &\quad \times p(x;\alpha\theta_1+(1-\alpha)\theta_2), \\ &= e^{-J_F^{(\alpha)}(\theta_1,\theta_2)} p(x; \underbrace{\alpha\theta_1+(1-\alpha)\theta_2}_{\theta_{12}^{(\alpha)}}) \end{aligned}$$

1 since natural parameter space is convex

Thus

$$P_e \leqslant w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1,\theta_2)} \int p(x;\alpha\theta_1+(1-\alpha)\theta_2)\mathrm{d}x.$$

$$P_e \leqslant \min_{\alpha\in[0,1]} w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1,\theta_2)}.$$

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

$$f(x) = f^{-1}(x) = \frac{1}{x},$$

$$\begin{aligned}
P_e &\leqslant \int M_H\left(\frac{1}{2}p_1(x), \frac{1}{2}p_2(x); \alpha\right) \mathrm{d}x, \\
&\leqslant \frac{1}{2}\int \frac{p_1(x)p_2(x)}{(1-\alpha)p_1(x)+\alpha p_2(x)} \mathrm{d}x, \\
&\leqslant \frac{1}{2}\int \frac{\frac{s_1}{\pi(x^2+s_1^2)}\frac{s_2}{\pi(x^2+s_2^2)}}{(1-\alpha)\frac{s_1}{\pi(x^2+s_1^2)}+\alpha\frac{s_2}{\pi(x^2+s_2^2)}} \mathrm{d}x, \\
&\leqslant \frac{1}{2}\int \frac{s_1 s_2}{\pi((1-\alpha)s_1(x^2+s_2^2)+\alpha s_2(x^2+s_1^2))} \mathrm{d}x, \\
&\leqslant \frac{1}{2}\int \frac{s_1 s_2}{\pi(((1-\alpha)s_1+\alpha s_2)x^2+(1-\alpha)s_1 s_2^2+\alpha s_2 s_1^2)} \mathrm{d}x, \\
&\leqslant \frac{1}{2}\frac{s_1 s_2}{((1-\alpha)s_1+\alpha s_2)s_\alpha}\underbrace{\int \frac{1}{\pi}\frac{s_\alpha}{x^2+s_\alpha^2}\mathrm{d}x}_{=1},
\end{aligned}$$

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)

Probability of error for Cauchy hypothesis

$$\mathrm{TV}(p_1, p_2) = \frac{2}{\pi}\left(\arctan\left(\sqrt{\frac{s_2}{s_1}}\right) - \arctan\left(\sqrt{\frac{s_1}{s_2}}\right)\right).$$

$$\begin{aligned} P_e &= \frac{1}{2} - \frac{1}{\pi}\left(\arctan\left(\sqrt{\lambda}\right) - \arctan\left(\sqrt{1/\lambda}\right)\right), \\ &= 1 - \frac{2}{\pi}\arctan\left(\sqrt{\lambda}\right), \quad \lambda = \frac{s_2}{s_1}. \end{aligned}$$

$$P_e \leqslant \frac{1}{2}\frac{s_1 s_2}{((1-\alpha)s_1 + \alpha s_2)\sqrt{\frac{(1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2}{(1-\alpha)s_1 + \alpha s_2}}}.$$

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)

Hypothesis Testing: 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 $f^{-1}(x)=x^{-\lambda}$).

$$P_e \leqslant \frac{1}{2}\left(\alpha|\Sigma_1|^{\frac{1}{2\lambda}} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}}\right)^{-\lambda} |\Sigma_\alpha|^{\frac{1}{2}} \underbrace{\int p(x;\Sigma_\alpha)\mathrm{d}x}_{=1},$$

$$= \frac{1}{2}\left(\alpha|\Sigma_1|^{\frac{1}{2\lambda}} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}}\right)^{-\lambda} |\Sigma_\alpha|^{\frac{1}{2}},$$

since $\Sigma_\alpha \in \Theta$.

Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. PRL (2014)

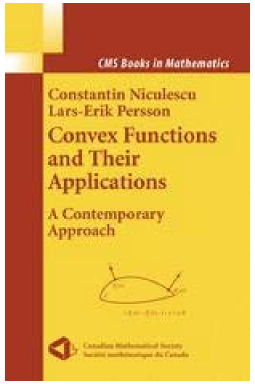
New Bregman divergences from abstract means

A function is **(M,N)-convex** (comparative convexity) if and only if

$$F(M(p,q)) \leq N(F(p),F(q)), \quad \forall p,q \in \mathcal{X}$$

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:

$$\boxed{J_F^{M,N}(p,q) = N(F(p),F(q)) - F(M(p,q))}$$

$$J_{F,\alpha}^{M,N}(p:q) \geq 0$$

Example of non-regular means: Lehmer mean (also Bajraktarevic mean)

$$L_\delta(x_1,\ldots,x_n;w_1,\ldots,w_n) = \frac{\sum_{i=1}^n w_i x_i^{\delta+1}}{\sum_{i=1}^n w_i x_i^\delta}$$

Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity, IEEE SPL 2017

(M,N)-Bregman divergences from comparative convexity

(M,N) Bregman divergences obtained in the scaled limit case of Jensen divergence:

$$\boxed{B_F^{M,N}(p:q) = \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p:q)} = \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} \left(N_\alpha(F(p),F(q)) - F(M_\alpha(p,q))\right)$$

Quasi-arithmetic Bregman divergences obtained

$$M_f(p,q) = f^{-1}\left(\frac{f(p)+f(q)}{2}\right)$$

$$B_F^{\rho,\tau}(p:q) = \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p)-\rho(q)}{\rho'(q)} F'(q).$$

$$B_F^{\rho,\tau}(p:q) = \kappa_\tau(F(q):F(p)) - \kappa_\rho(q:p)F'(q)$$

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	$\gamma(x) = \frac{1}{x}$	$x^2\left(\frac{1}{y} - \frac{1}{x}\right)$
$P_\delta, \delta \neq 0$	$\gamma_\delta(x) = x^\delta$	$\frac{y^\delta - x^\delta}{\delta x^{\delta-1}}$

For example, the **power mean Bregman divergences**:

$$B_F^{\delta_1,\delta_2}(p:q) = \frac{F^{\delta_2}(p) - F^{\delta_2}(q)}{\delta_2 F^{\delta_2-1}(q)} - \frac{p^{\delta_1} - q^{\delta_1}}{\delta_1 q^{\delta_1-1}} F'(q)$$

Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity, IEEE SPL 2017

Generalizing Jensen-Shannon divergences

$$\mathrm{JS}(p;q) := \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right),$$
$$= \frac{1}{2}\int\left(p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q}\right)\mathrm{d}\mu.$$

$$\mathrm{JS}(p;q) = h\left(\frac{p+q}{2}\right) - \frac{h(p)+h(q)}{2}.$$

Jensen-Shannon divergence is the total divergence to the average divergence

Always bounded by log 2, and the square root of JSD is a metric

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

https://www.mdpi.com/1099-4300/21/5/485

Symmetrizing the KL divergence

Jeffreys divergence:

$$J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p-q)\log\frac{p}{q}\mathrm{d}\mu = J(q;p).$$

Resistor average divergence:

$$\begin{aligned} \frac{1}{R(p;q)} &= \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right), \\ R(p;q) &= \frac{2\left(\mathrm{KL}(p:q) + \mathrm{KL}(q:p)\right)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)} = \frac{2J(p;q)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)}. \end{aligned}$$

Jensen-Bregman divergence as a Jensen divergence

$$\begin{aligned}
\mathrm{JB}_F(\theta:\theta') &:= \frac{1}{2}\left(B_F\left(\theta:\frac{\theta+\theta'}{2}\right)+B_F\left(\theta':\frac{\theta+\theta'}{2}\right)\right),\\
&= \frac{F(\theta)+F(\theta')}{2}-F\left(\frac{\theta+\theta'}{2}\right) =: J_F(\theta:\theta'),
\end{aligned}$$

$$\begin{aligned}
\mathrm{JB}_F^{\alpha}(\theta:\theta') &:= (1-\alpha)B_F\left(\theta:(\theta\theta')_\alpha\right)+\alpha B_F\left(\theta':(\theta\theta')_\alpha\right),\\
&= (F(\theta)F(\theta'))_\alpha - F\left((\theta\theta')_\alpha\right) =: J_F^{\alpha}(\theta:\theta'),
\end{aligned}$$

Skew Jensen-Bregman Voronoi diagrams, 2011

M-statistical mixture

$$(pq)_\alpha^M(x) := \frac{M_\alpha(p(x), q(x))}{Z_\alpha^M(p:q)}$$

Need to normalize M-mixtures

$$Z_\alpha^M(p:q) = \int_{t\in\mathcal{X}} M_\alpha(p(t), q(t)) \mathrm{d}\mu(t)$$

$$(p_1 \ldots p_k)_\alpha^M := \frac{p_1(x)^{\alpha_1} \times \ldots \times p_k(x)^{\alpha_k}}{Z_\alpha(p_1, \ldots, p_k)}$$

$$\boxed{\mathrm{JS}_D^{M_\alpha}(p:q) := (1-\alpha) D\left(p : (pq)_\alpha^M\right) + \alpha D\left(q : (pq)_\alpha^M\right)}$$

$$\mathrm{JS}^{M_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p : (pq)_\alpha^M\right) + \alpha \mathrm{KL}\left(q : (pq)_\alpha^M\right)$$

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

When does M-Jensen-Shannon divergence are bounded?

The M-JSD is upper bounded by $\log \frac{Z_\alpha^M(p,q)}{1-\alpha}$ *when* $M \geq A$.

A further generalization of the Jensen-Shannon divergence:

$$\boxed{\mathrm{JS}_D^{M_\alpha,N_\beta}(p:q) := N_\beta\left(D\left(p:(pq)_\alpha^M\right), D\left(q:(pq)_\alpha^M\right)\right)}$$

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

Closed-form formula for exponential families

$$
\begin{aligned}
\mathrm{KL}\left(p_\theta : (p_{\theta_1} p_{\theta_2})_\alpha^G\right) &= \mathrm{KL}\left(p_\theta : p_{(\theta_1\theta_2)_\alpha}\right)\\
&= B_F((\theta_1\theta_2)_\alpha : \theta).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{JS}_\alpha^G(p_{\theta_1} : p_{\theta_2}) &:= (1-\alpha)\mathrm{KL}(p_{\theta_1} : (p_{\theta_1} p_{\theta_2})_\alpha^G) + \alpha \mathrm{KL}(p_{\theta_2} : (p_{\theta_1} p_{\theta_2})_\alpha^G),\\
&= (1-\alpha) B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{JS}_{\mathrm{KL}^*}^{G_\alpha}(p_{\theta_1} : p_{\theta_2}) &:= (1-\alpha)\mathrm{KL}((p_{\theta_1} p_{\theta_2})_\alpha^G : p_{\theta_1}) + \alpha \mathrm{KL}((p_{\theta_1} p_{\theta_2})_\alpha^G : p_{\theta_2}),\\
&= (1-\alpha) B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha) = \mathrm{JB}_F^\alpha(\theta_1 : \theta_2),\\
&= (1-\alpha) F(\theta_1) + \alpha F(\theta_2) - F((\theta_1\theta_2)_\alpha),\\
&= J_F^\alpha(\theta_1 : \theta_2).
\end{aligned}
$$

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

Case study of multivariate Gaussians

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) = \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta_\mu^\top \Sigma_2^{-1}\Delta_\mu + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\right\}$$

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}) + \alpha\mathrm{KL}(p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}),\\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2),\\
&= \frac{1}{2}\Bigg(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \\
&\quad (1-\alpha)(\mu_\alpha - \mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha - \mu_1) + \alpha(\mu_\alpha - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha - \mu_2) - d\Bigg)\\
\mathrm{JS}_*^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}),\\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha),\\
&= J_F(\theta_1 : \theta_2),\\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right)
\end{aligned}$$

$$\Sigma_\alpha = (\Sigma_1\Sigma_2)_\alpha^\Sigma = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1} \qquad \mu_\alpha = (\mu_1\mu_2)_\alpha^\mu = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right)$$

Case study of Cauchy family: Harmonic mean

$$\mathcal{C}_\Gamma := \left\{ p_\gamma(x) = \frac{1}{\gamma} p_{\text{std}}\left(\frac{x}{\gamma}\right) = \frac{\gamma}{\pi(\gamma^2 + x^2)} : \gamma \in \Gamma = (0, \infty) \right\}$$

$$(p_{\gamma_1} p_{\gamma_2})^H_{\frac{1}{2}}(x) = \frac{H_\alpha(p_{\gamma_1}(x) : p_{\gamma_2}(x))}{Z^H_\alpha(\gamma_1, \gamma_2)} = p_{(\gamma_1\gamma_2)_\alpha}$$

$$Z^H_\alpha(\gamma_1, \gamma_2) := \sqrt{\frac{\gamma_1\gamma_2}{(\gamma_1\gamma_2)_\alpha (\gamma_1\gamma_2)_{1-\alpha}}} = \sqrt{\frac{\gamma_1\gamma_2}{(\gamma_1\gamma_2)_\alpha (\gamma_2\gamma_1)_\alpha}}$$

$$\begin{aligned} \mathrm{JS}^H(p:q) &= \frac{1}{2}\left(\mathrm{KL}\left(p : (pq)^H_{\frac{1}{2}}\right) + \mathrm{KL}\left(q : (pq)^H_{\frac{1}{2}}\right)\right), \\ \mathrm{JS}^H(p_{\gamma_1} : p_{\gamma_2}) &= \frac{1}{2}\left(\mathrm{KL}\left(p_{\gamma_1} : p_{\frac{\gamma_1+\gamma_2}{2}}\right) + \mathrm{KL}\left(p_{\gamma_2} : p_{\frac{\gamma_1+\gamma_2}{2}}\right)\right) \\ &= \log \frac{(3\gamma_1 + \gamma_2)(3\gamma_2 + \gamma_1)}{8\sqrt{\gamma_1\gamma_2}(\gamma_1 + \gamma_2)}. \end{aligned}$$

Kullback-Leibler divergence between Cauchy densities

Cauchy density $$p_{l,s}(x) = \frac{\mathrm{d}P_{l,s}}{\mathrm{d}\mu}(x) = \frac{s}{\pi(s^2+(x-l)^2)}$$

Symmetric KL

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

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 $$h(p_{l,s}) = h^\times(p_{l,s} : p_{l,s}) = \log 4\pi s,$$

Relies on this definite integral with

$$A(a,b,c;d,e,f) = \int_{-\infty}^{\infty} \frac{\log(dx^2+ex+f)}{ax^2+bx+c}\mathrm{d}x,$$

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

A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions https://arxiv.org/pdf/1905.10965.pdf

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

$$\boxed{I_f(p_{l_1,s_1} : q_{l_2,s_2}) = I_f\left(p : q_{\frac{l_2-l_1}{s_1},\frac{s_2}{s_1}}\right) = I_f\left(p_{\frac{l_1-l_2}{s_2},\frac{s_1}{s_2}} : q\right)}$$

Proof by change of variable

$$y = \frac{x-l_1}{s_1}$$

$$\mathrm{d}x = s_1 \mathrm{d}y$$

$$x = s_1 y + l_1$$

$$\frac{x-l_2}{s_2} = \frac{s_1 y + l_1 - l_2}{s_2} = \frac{y - \frac{l_2-l_1}{s_1}}{\frac{s_2}{s_1}}$$

$$\begin{aligned}
I_f(p_{l_1,s_1} : q_{l_2,s_2}) &:= \int_{\mathcal{X}} p_{l_1,s_1}(x) f\left(\frac{q_{l_2,s_2}(x)}{p_{l_1,s_1}(x)}\right) \mathrm{d}x, \\
&= \int_{\mathcal{Y}} \frac{1}{s_1} p(y) f\left(\frac{\frac{1}{s_2} q\left(\frac{y-\frac{l_2-l_1}{s_1}}{\frac{s_2}{s_1}}\right)}{\frac{1}{s_1}p(y)}\right) s_1 \mathrm{d}y, \\
&= \int p(y) f\left(\frac{q_{\frac{l_2-l_1}{s_1},\frac{s_2}{s_1}}(y)}{p(y)}\right) \mathrm{d}y, \\
&= I_f\left(p : q_{\frac{l_2-l_1}{s_1},\frac{s_2}{s_1}}\right).
\end{aligned}$$

Location-scale group

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

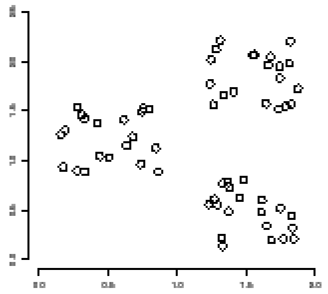
On the Kullback–Leibler divergence between location–scale densities, 2019 https://arxiv.org/abs/1904.10428

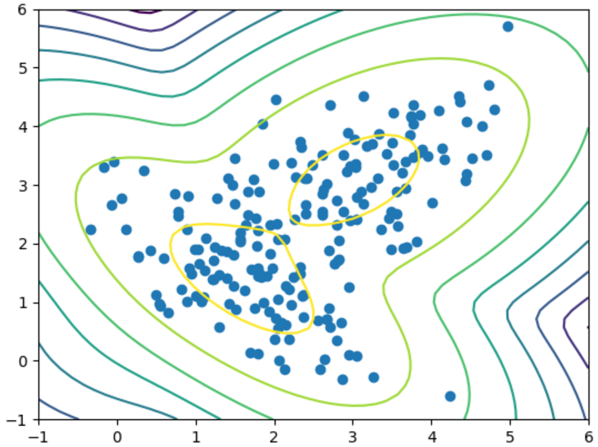
Information geometry of clustering: Hard, Soft and Hierarchical

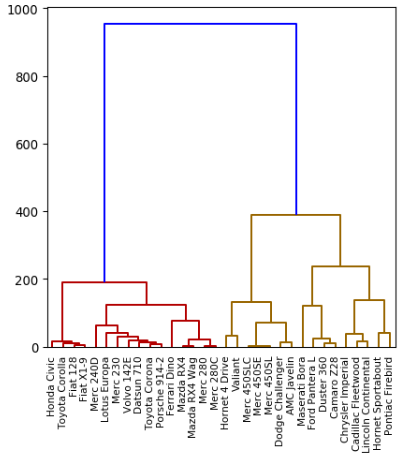
Frank Nielsen

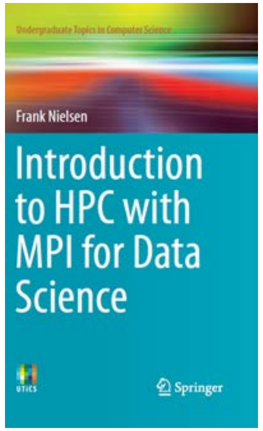
Sony CSL

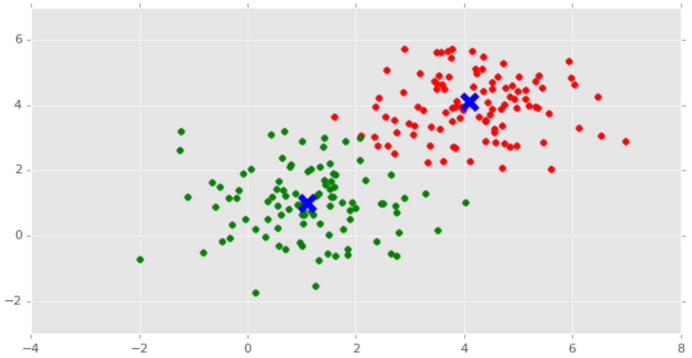
Finding structures (clusters) in datasets

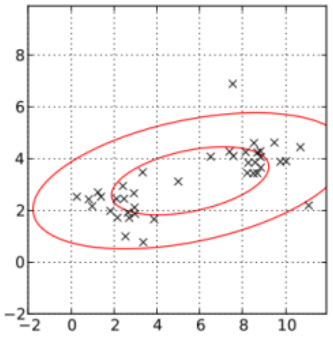


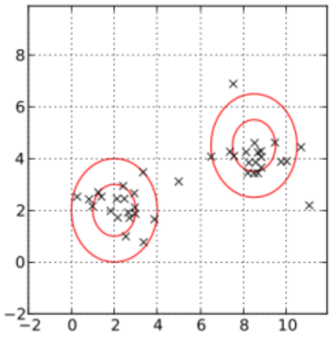


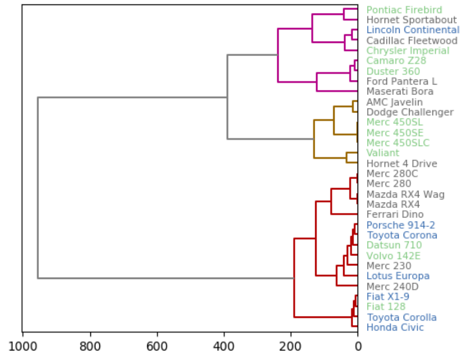












Hard membership
Flat clustering (partitions)

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, etc.

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

- 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

$$\underbrace{[x_1 \ldots x_{l_2-1}]}_{\mathcal{C}_1} \; \underbrace{[x_{l_2} \ldots x_{l_3-1}]}_{\mathcal{C}_2} \ldots \underbrace{[x_{l_k} \ldots x_n]}_{\mathcal{C}_k}$$

- 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

Optimal Interval Clustering: Application to Bregman Clustering and Statistical Mixture Learning. IEEE Signal Process. Lett. 21(10) (2014)
https://arxiv.org/abs/1403.2485

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 $d_\phi : \mathcal{S} \times \mathrm{ri}(\mathcal{S}) \mapsto \mathbb{R}$, 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 $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S})$ (one possible initialization is to choose $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S})$ at random)
repeat
  {The Assignment Step}
  Set $\mathcal{X}_h \leftarrow \emptyset,\ 1 \le h \le k$
  for $i = 1$ to n **do**
    $\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{\mathbf{x}_i\}$
    where $h = h^\dagger(\mathbf{x}_i) = \underset{h'}{\mathrm{argmin}}\ d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'})$
  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 $\mathcal{M}^\dagger \leftarrow \{\boldsymbol{\mu}_h\}_{h=1}^k$

Bregman centroids are centers of mass, independent of the generator

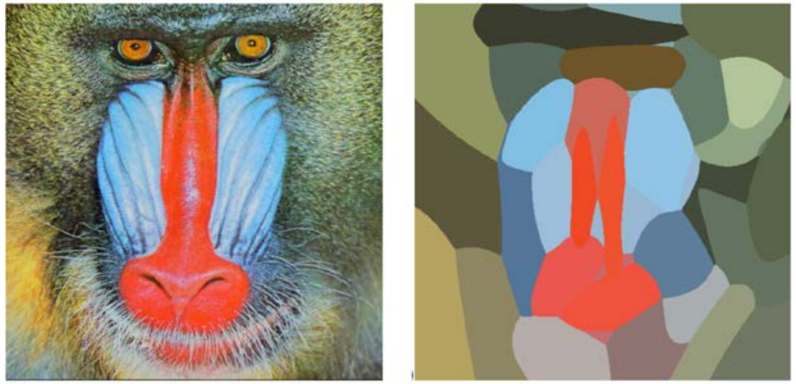
Compared to squared Euclidean k-means, <u>only the assignment step changes</u>

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

arxiv:1203.5181

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

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



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

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

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

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

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

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:

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

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

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

K-MLE: Classification Expectation-Maximization (CEM)

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

Additive Bregman Voronoi diagrams
Biased, not consistent

On learning statistical mixtures maximizing the complete likelihood, AIP 2014

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: $\Gamma^\dagger$: local maximizer of $L_\mathcal{X}(\Gamma) = \prod_{i=1}^n(\sum_{h=1}^k \pi_h p_{\psi,\boldsymbol{\theta}_h}(\mathbf{x}_i))$ where $\Gamma = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$.
Method:
Initialize $\{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\theta}_h \in \Theta$, 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**
$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p_{(\psi,\boldsymbol{\theta}_h)}(\mathbf{x}_i)}{\sum_{h'=1}^k \pi_{h'} p_{(\psi,\boldsymbol{\theta}_{h'})}(\mathbf{x}_i)}$
end for
end for
{The Maximization Step (M-step)}
for $h = 1$ to k **do**
$\pi_h \leftarrow \frac{1}{n}\sum_{i=1}^n p(h|\mathbf{x}_i)$
$\boldsymbol{\theta}_h \leftarrow \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \sum_{i=1}^n \log(p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}_i))p(h|\mathbf{x}_i)$
end for
until *convergence*
return $\Gamma^\dagger = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$

Algorithm 3 Bregman Soft Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$, Bregman divergence $d_\phi : \mathcal{S} \times \mathrm{ri}(\mathcal{S}) \mapsto \mathbb{R}$, number of clusters k.
Output: $\Gamma^\dagger$, local maximizer of $\prod_{i=1}^n(\sum_{h=1}^k \pi_h b_\phi(\mathbf{x}_i)\exp(-d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)))$ where $\Gamma = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$
Method:
Initialize $\{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S}), \pi_h \geq 0$, 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**
$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h \exp(-d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'}))}$
end for
end for
{The Maximization Step (M-step)}
for $h = 1$ to k **do**
$\pi_h \leftarrow \frac{1}{n}\sum_{i=1}^n p(h|\mathbf{x}_i)$
$\boldsymbol{\mu}_h \leftarrow \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}$
end for
until *convergence*
return $\Gamma^\dagger = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$

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:

$$\Pr(c_j = p_i) = \frac{D(p_i, \{c_1, \ldots, c_{j-1}\})}{\sum_{i=1}^{n} D(p_i, \{c_1, \ldots, c_{j-1}\})} \quad (2 \leq j \leq k).$$

K-means++ probabilistic seeding

$$E_D(\Lambda, C) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} D(p_i : c_j)$$

Theorem (Generalized k-means++ performance). *Let* κ_1 *and* κ_2 *be two constants such that* κ_1 *defines the quasi-triangular inequality property:*

$$D(x:z) \leq \kappa_1 \left(D(x:y) + D(y:z) \right), \quad \forall x, y, z$$

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:

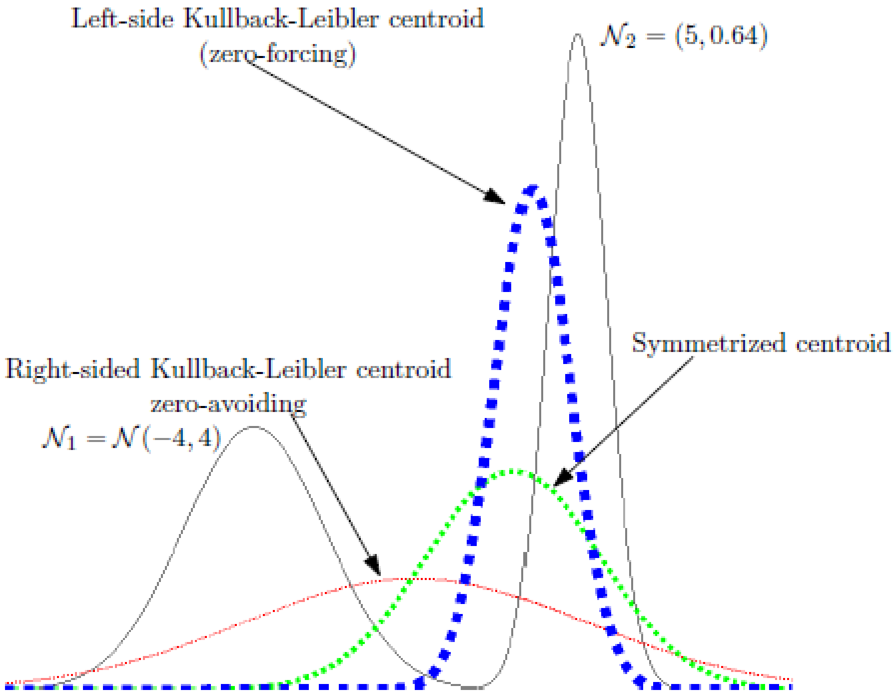
$$E_D(\Lambda, C) \leq 2\kappa_1^2 (1+\kappa_2)(2+\log k) E_D^*(\Lambda, k).$$

Total Jensen divergences: Definition, properties and clustering. ICASSP 2015

Left-sided or right-sided centroids (k-means) ?

Left/right Bregman centroids=Right/left entropic centroids (KL of exp. fam.)
Left-sided/right-sided centroids: *different* (statistical) properties:

- **Right-sided entropic centroid** : **zero-avoiding** (cover support of pdfs.)
- **Left-sided entropic centroid** : **zero-forcing** (captures highest mode).

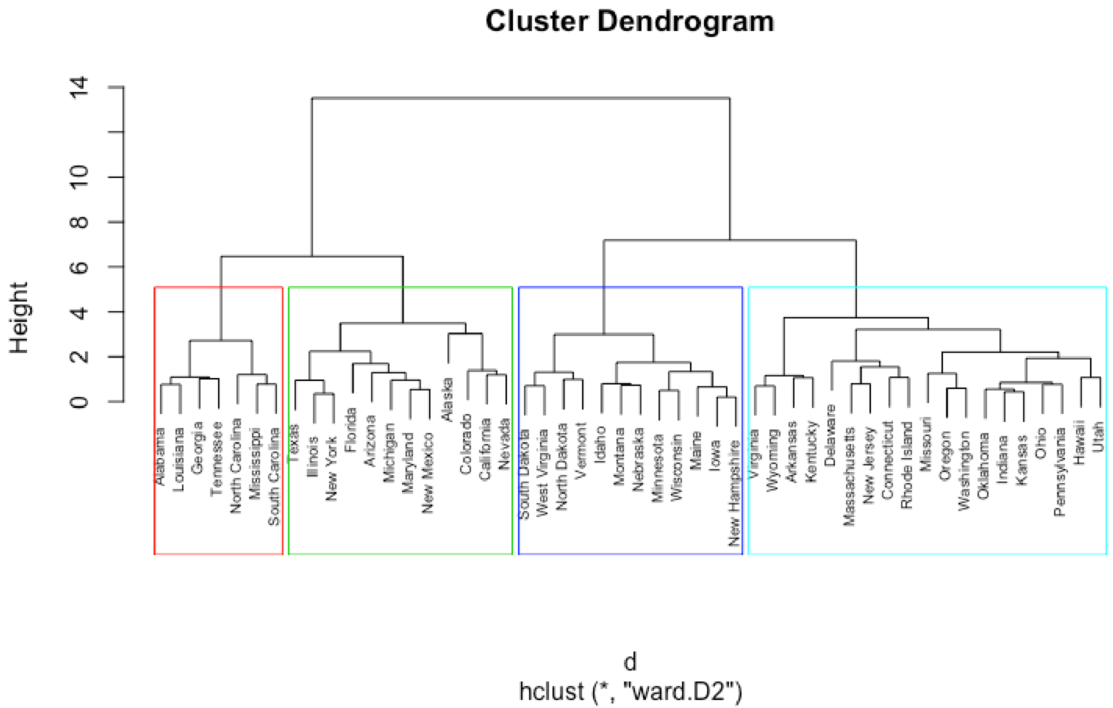


Hierarchical clustering (Ward criterion)

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.



Potential inversions…

Telgarsky, Matus, and Sanjoy Dasgupta. "Agglomerative Bregman Clustering." (2012).

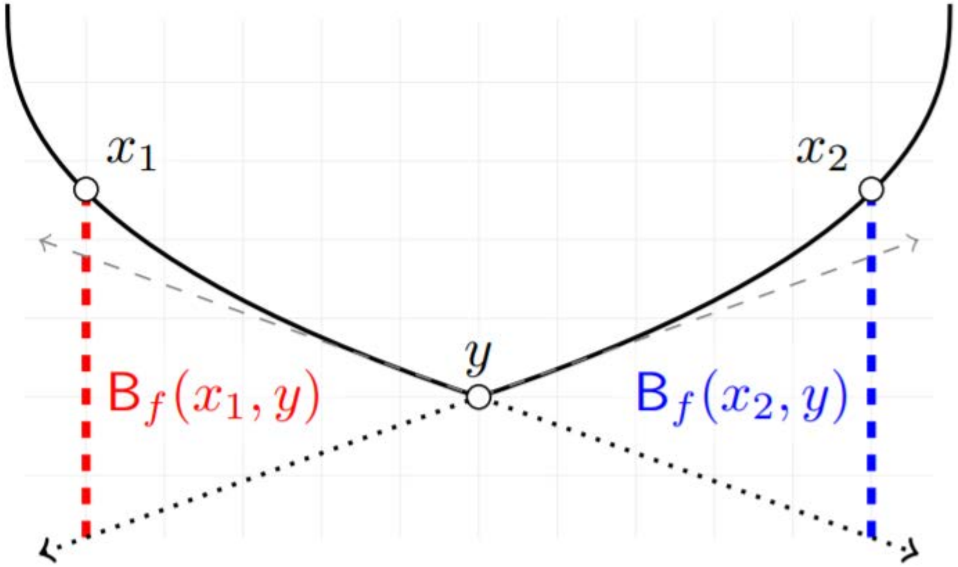
Extending to Bregman divergences

$$\mathrm{B}_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x-y \rangle.$$

Consider more general **directional derivatives**

$$\mathrm{B}_f(x,y) := f(x) - f(y) + f'(y; y-x).$$

Subgradient derivatives



Bregman
Ward
Criterion

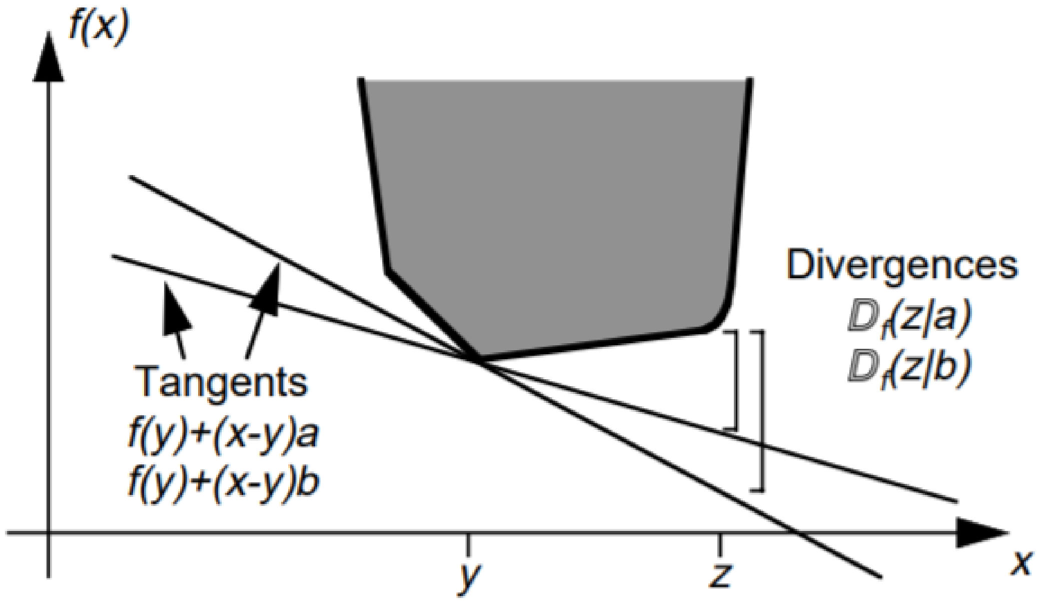
Proposition 3.8. *Let a proper convex relatively differentiable f and two finite subsets C_1, C_2 of $\mathcal{X}$ with $\tau(C_i) \in \mathrm{ri}(\mathrm{dom}(f))$ be given. Then*

$$\Delta_{f,\tau}(C_1, C_2) = \sum_{j \in \{1,2\}} |C_j| \mathrm{B}_f(\tau(C_j), \tau(C_1 \cup C_2)).$$

Another generalization of Bregman divergences

$$\mathrm{D}_f(x,g) := f(x) + f^*(g) - \langle g, x\rangle \,. \qquad g \in \partial f(y)$$

$$\mathrm{B}_f(x,y) := \max\{\mathrm{D}_f(x,g) : g \in \partial f(y)\}.$$



Gordon, Approximate Solutions to Markov Decision Processes. CMU PhD, 1999.

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; $\mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$

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 $i = 2, 3, \ldots, k$ **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 $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

//Assignment

for $i = 1, 2, \ldots, k$ **do**

$\mathcal{A}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

// Centroid relocation

for $i = 1, 2, \ldots, k$ **do**

$r_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

$l_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

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 $\lambda \leftarrow \lambda_{\mathrm{init}} \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

//Expectation

for $i = 1, 2, \ldots, m$ **do**

for $j = 1, 2, \ldots, k$ **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 $j = 1, 2, \ldots, k$ **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 $\lambda_{\mathrm{init}} \neq 0, 1$ **then**

$\lambda \leftarrow \lambda - \eta_2 \left(\sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_\alpha(l_j : h_i) - \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_\alpha(h_i : r_j)\right)$;

//for some small η_1, η_2; ensure that $\lambda \in [0,1]$.

until *convergence*;

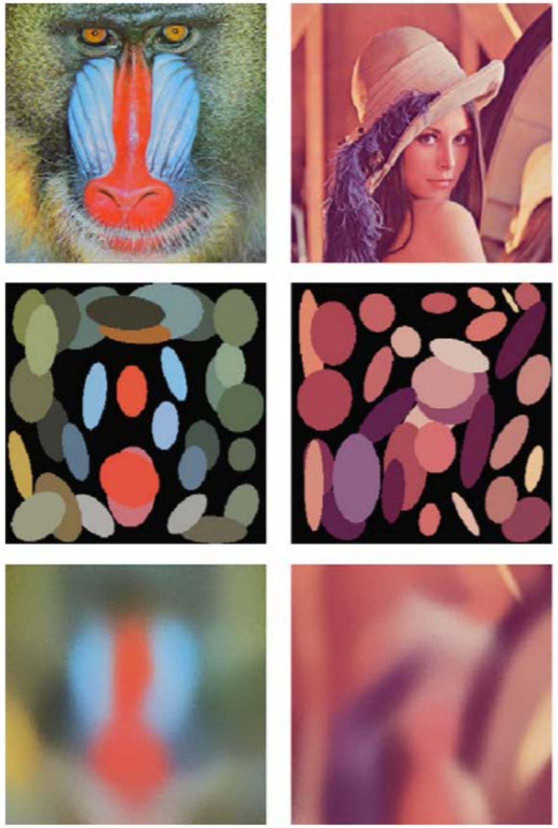
Output: Soft clustering of $\mathcal{H}$ according to k densities $p(j|.)$ following $\mathcal{C}$;

On Clustering Histograms with k-Means by Using Mixed α-Divergences. Entropy 16(6): 3273-3301 (2014)

Hierarchical mixtures of exponential families

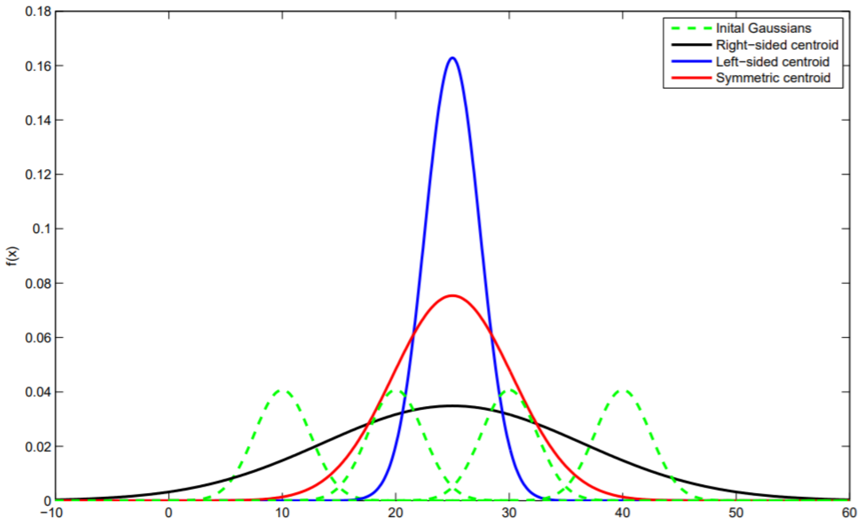
Hierarchical clustering with Bregman sided and symmetrized divergences

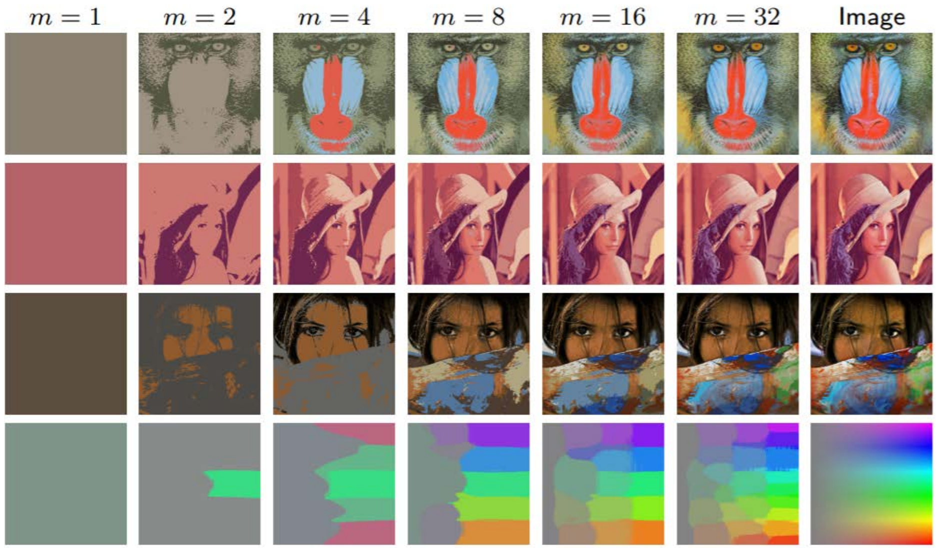
Learning & simplifying
Gaussian mixture models (GMMs)

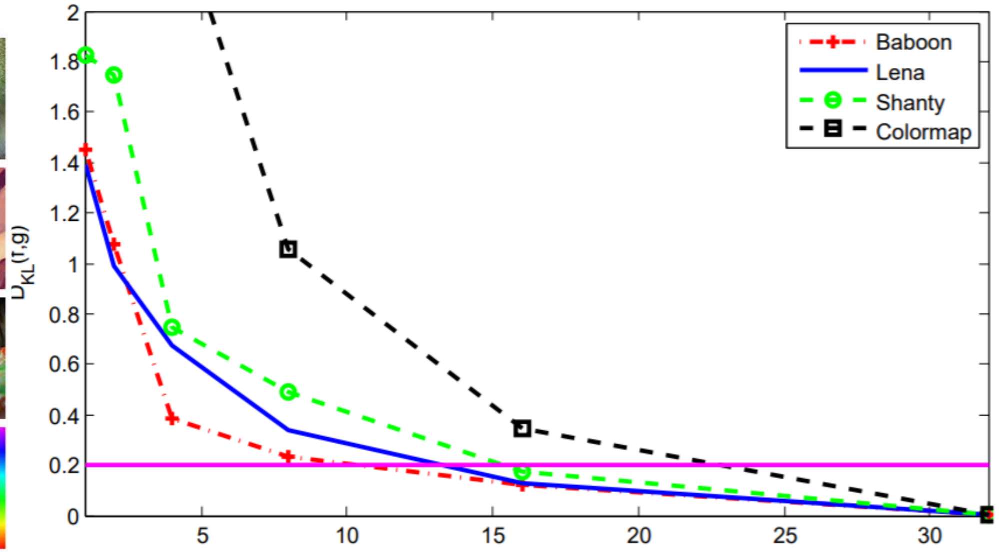


- Agglomerative method:
  1. Find the two closest subsets $\mathcal{S}_i$ and $\mathcal{S}_j$
  2. Merge the subsets $\mathcal{S}_i$ and $\mathcal{S}_j$
  3. Go back to 1. until one single set remains

Criterion	Formula
Minimum distance	$\mathrm{D}_{\min}(A,B) = \min\{d(a,b) \mid a \in A,\ b \in B\}$
Maximum distance	$\mathrm{D}_{\max}(A,B) = \max\{d(a,b) \mid a \in A,\ b \in B\}$
Average distance	$\mathrm{D}_{\mathrm{av}}(A,B) = \frac{1}{\lvert A\rvert\lvert B\rvert} \sum_{a\in A}\sum_{b\in B} d(a,b)$







Simplification and hierarchical representations of mixtures of exponential families. Signal Processing 90(12): 3197-

Conformal divergences

$$D'(p:q) = \rho(p,q) D(p:q)$$

$$\mathbf{D}_{F,\kappa}\left[\xi : \xi'\right] := \kappa(\xi) \mathbf{B}_F\left[\xi : \xi'\right]$$

Consider the right-sided centroid: Amount to reweight the points according to a positive conformal factor.
Related to conformal geometry

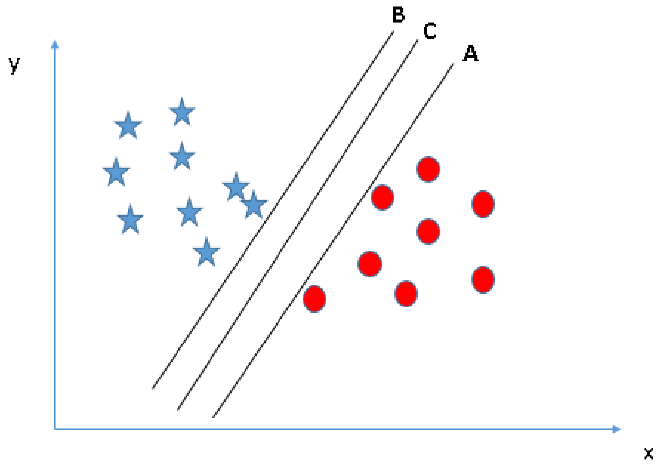
Total Bregman divergences, **total Jensen divergences**, etc.

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

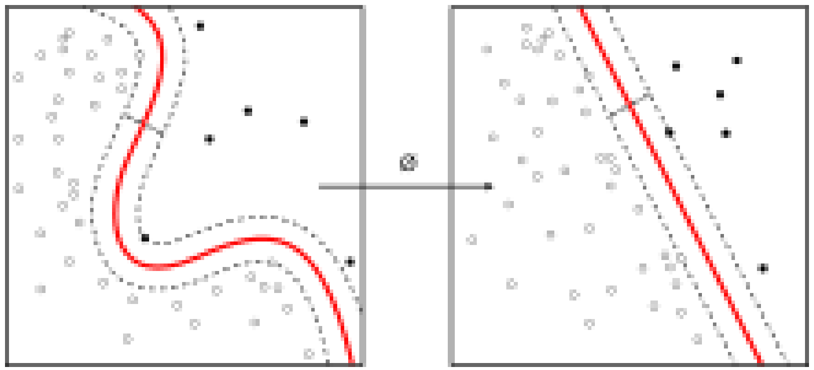
- Conformal kernel

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = D(\mathbf{x})D(\mathbf{x}')K(\mathbf{x}, \mathbf{x}'),$$



- 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.

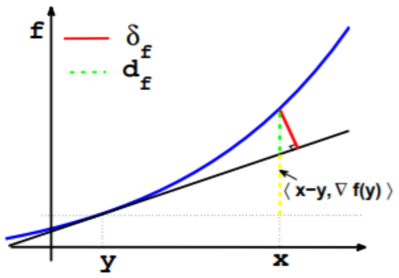
Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

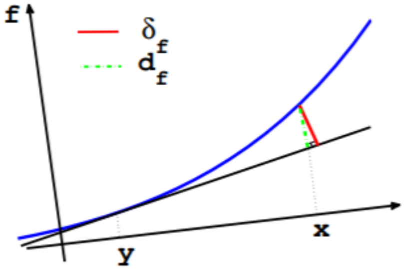
Definition *The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,*

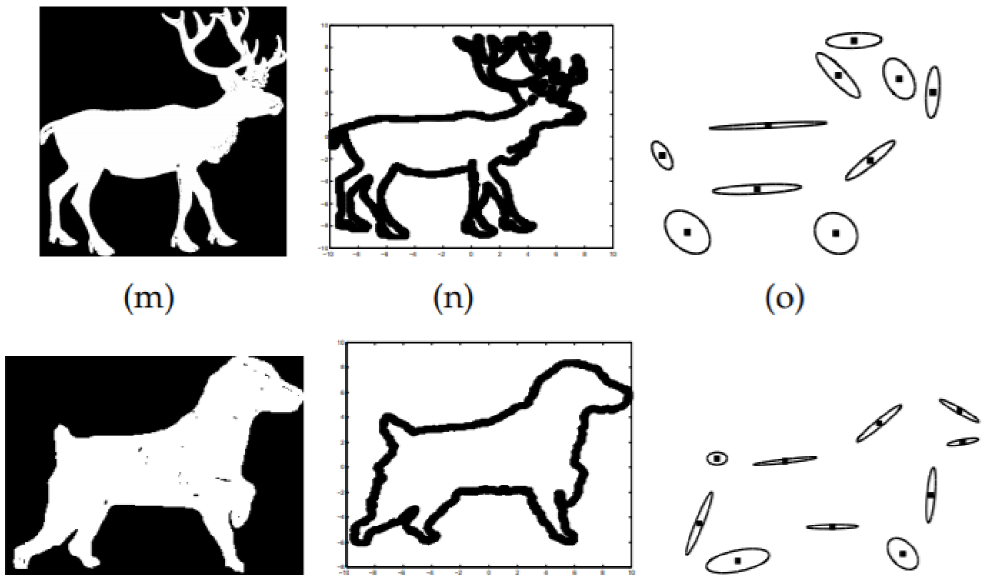
$$\delta_f(x,y) = \frac{f(x) - f(y) - \langle x-y, \nabla f(y)\rangle}{\sqrt{1+\|\nabla f(y)\|^2}},$$

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

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







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

IEEE TPAMI 34, 2012

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,*

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

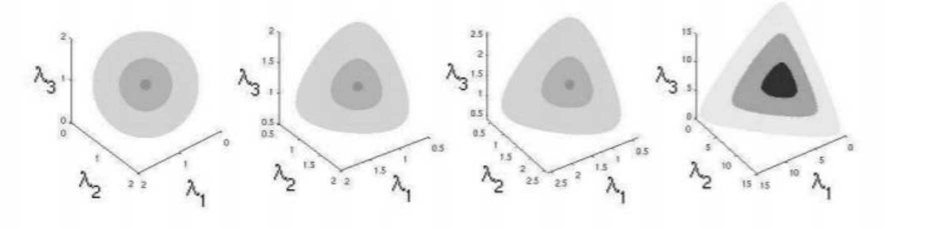
$\langle \cdot, \cdot \rangle$ is inner product as in definition II.1, and $\|\nabla f(y)\|^2 = \langle \nabla f(y), \nabla f(y) \rangle$ generally.

$$tKL(P,Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1 + \int (1 + \log q)^2 q dx}}$$

$$= \frac{\log(\det(P^{-1}Q)) + tr(Q^{-1}P) - n}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{n(1+\log 2\pi)}{2} \log(\det Q)}}$$

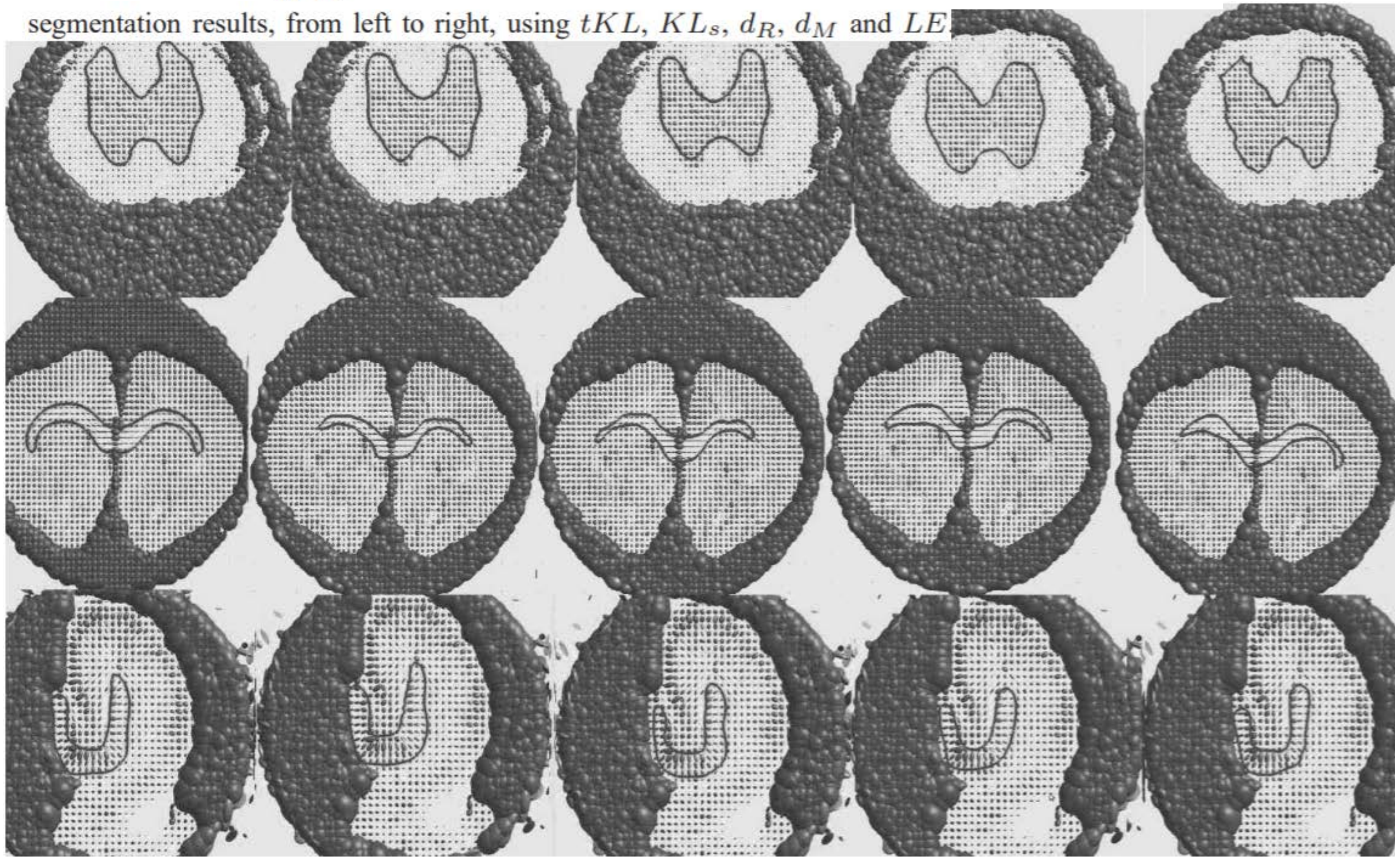
$$tKL(P,Q) = tKL(A'PA, A'QA), \quad \forall A \in SL(n),$$

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

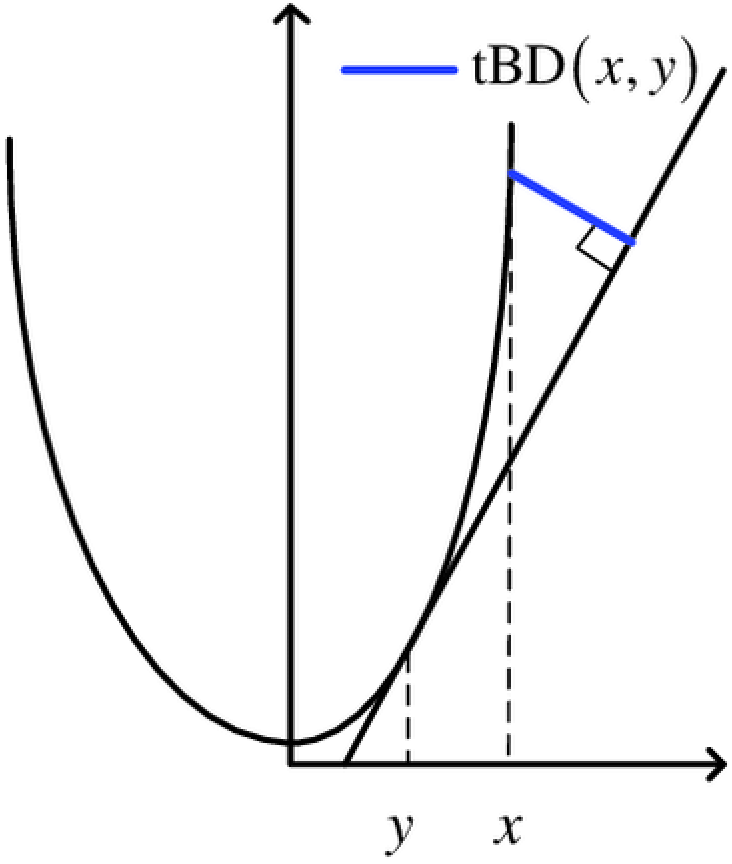


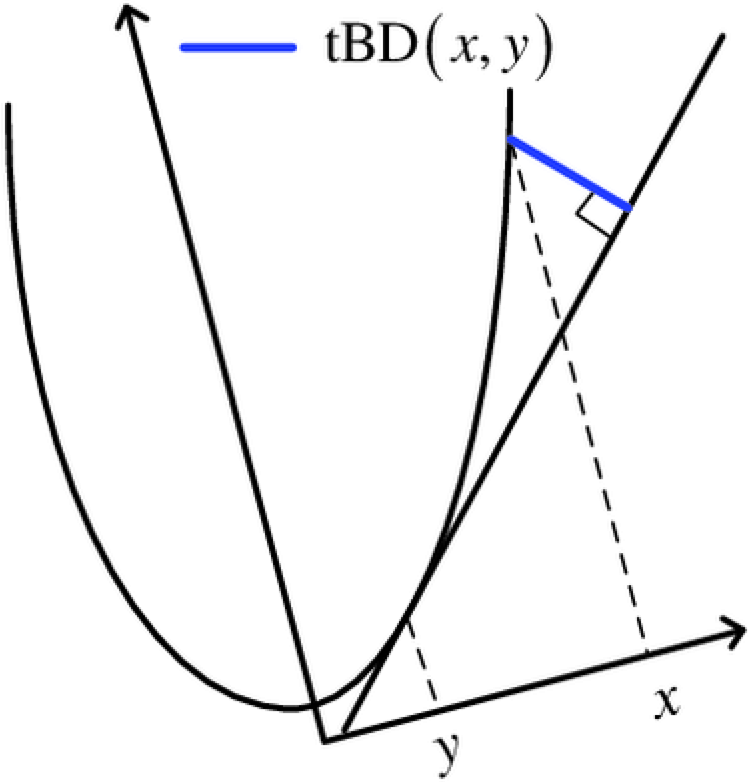
The isosurfaces of $d_F(P,I) = r$, $d_R(P,I) = r$, $KL_s(P,I) = r$ and $tKL(P,I) = r$ shown from left to right. The three axes are eigenvalues of P.

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



Total Bregman divergence

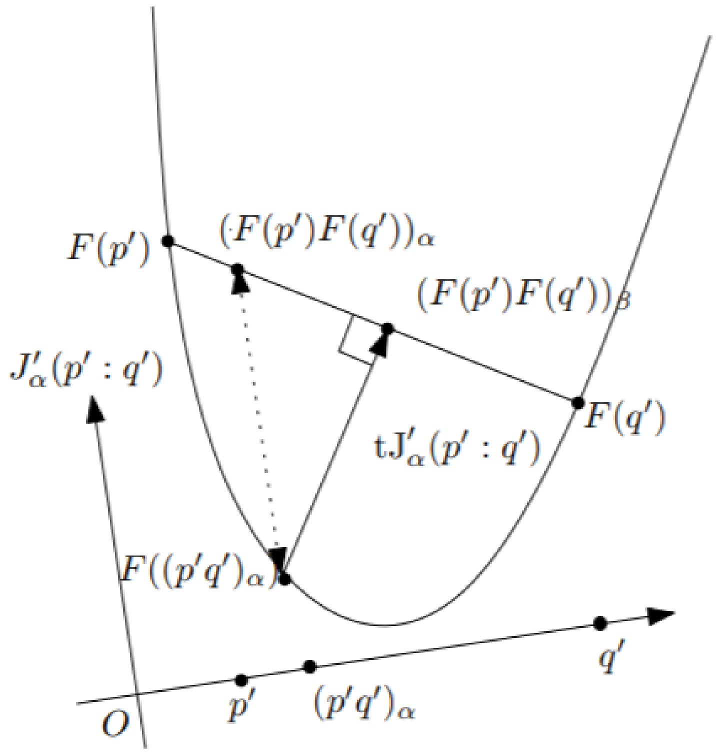




$$\mathrm{TBD}(p:q) = \frac{\varphi(p) - \varphi(q) - \nabla\varphi(q)\cdot(p-q)}{\sqrt{1+|\nabla\varphi(q)|^2}}$$

Invariant to axis rotation

Total Jensen divergence

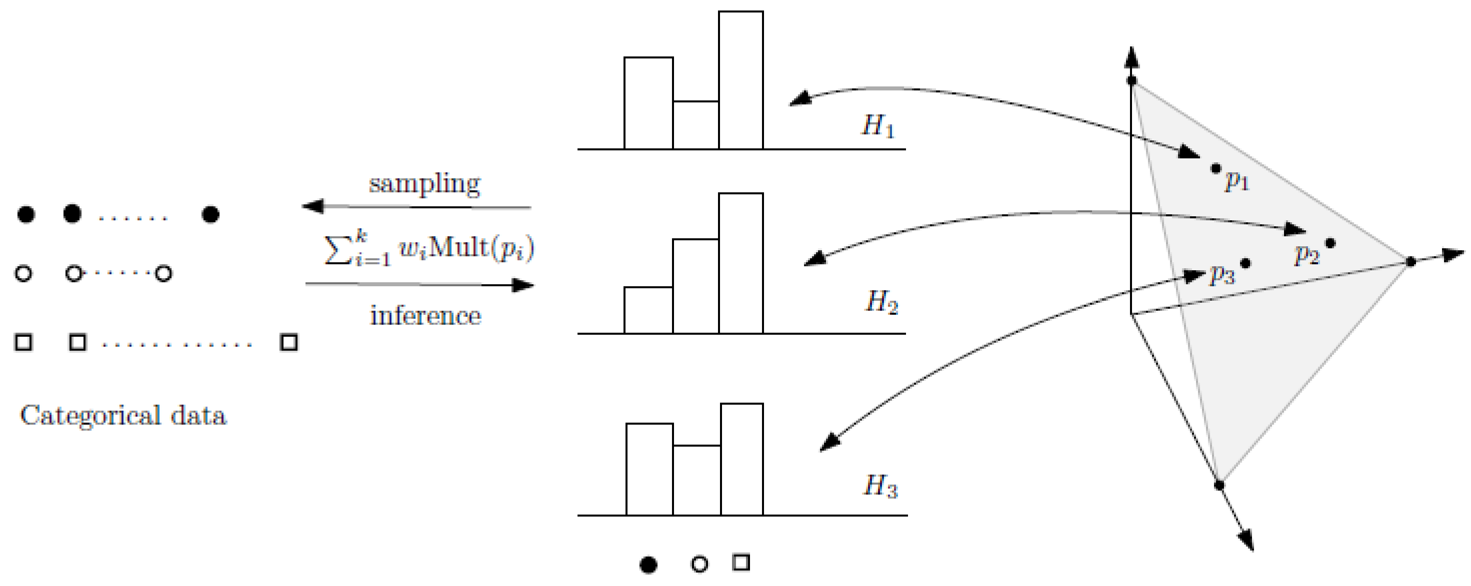


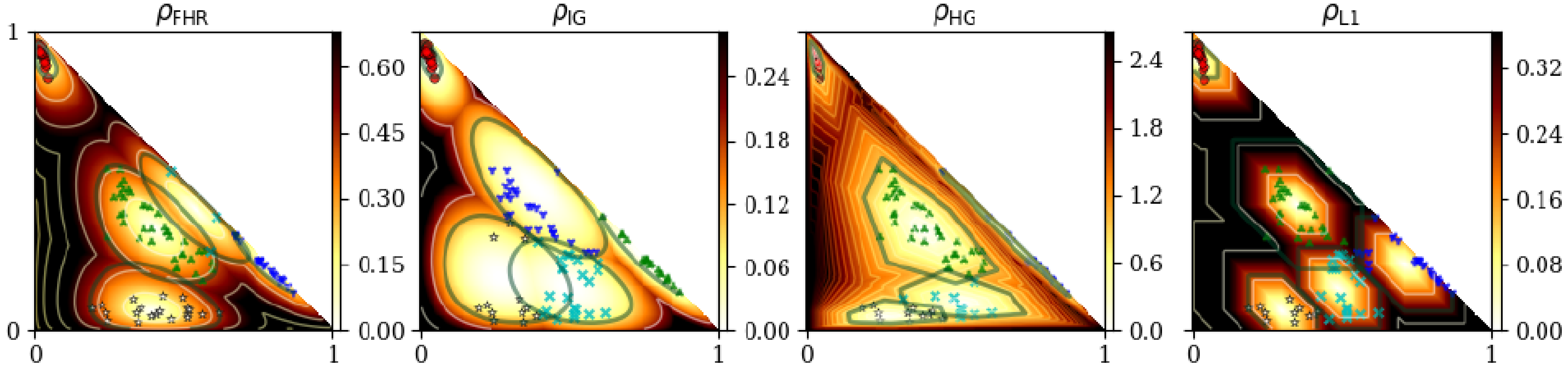
Invariant to axis rotation

$$\mathrm{tB}(p:q) = \rho_B(q)B(p:q), \quad \rho_B(q) = \sqrt{\frac{1}{1+\langle \nabla F(q), \nabla F(q)\rangle}}$$

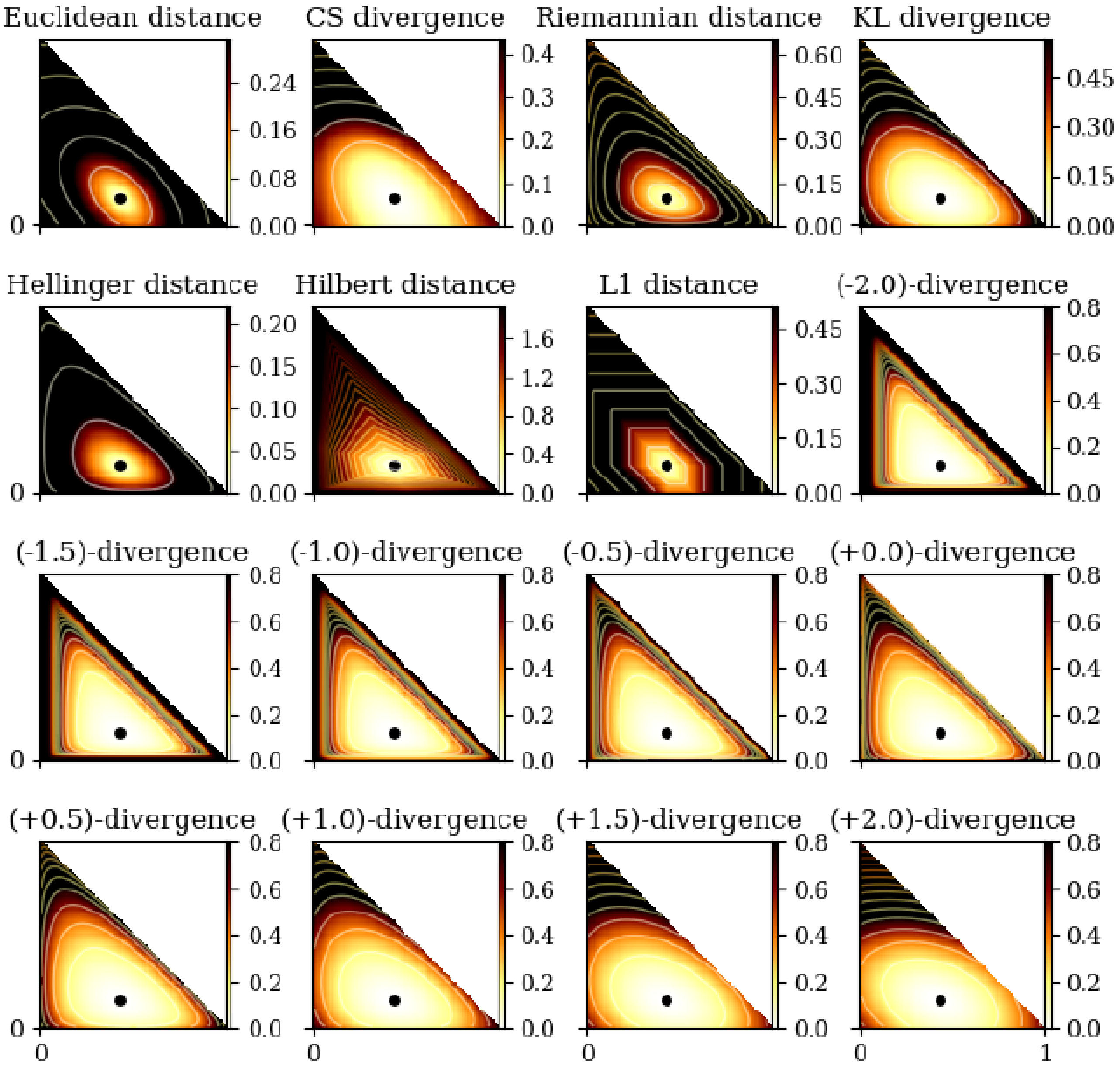
$$\mathrm{tJ}_\alpha(p:q) = \rho_J(p,q)J_\alpha(p:q), \quad \rho_J(p,q) = \sqrt{\frac{1}{1+\frac{(F(p)-F(q))^2}{\langle p-q,p-q\rangle}}}$$

Clustering categorical distributions





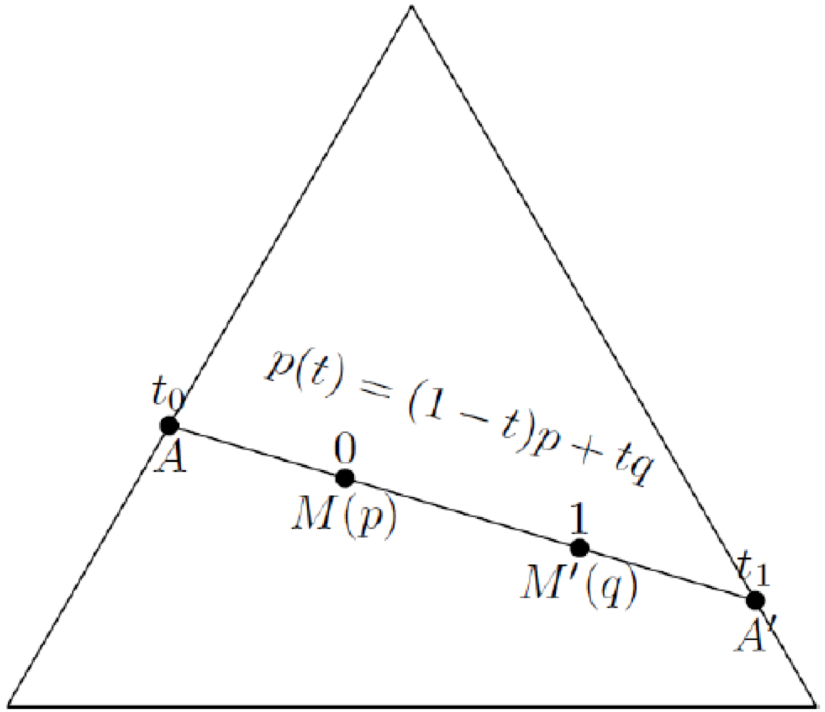
$k = 5$ clusters



Reference point (3/7,3/7,1/7)

Hilbert log cross-ratio metric

$$\rho_{\mathrm{HG}}(M, M') = \begin{cases} \left|\log \frac{|A'M||AM'|}{|A'M'||AM|}\right|, & M \neq M', \\ 0 & M = M'. \end{cases}$$



Geodesics are straight lines but not unique

Isometry of Hilbert simplex geometry with a normed vector space $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$

- $V^d = \{v \in \mathbb{R}^{d+1} \ : \ \sum_i v^i = 0\} \subset \mathbb{R}^{d+1}$
- Map $p = (\lambda^0, \ldots, \lambda^d) \in \Delta^d$ to $v(x) = (v^0, \ldots, v^d) \in V^d$:

$$v^i = \frac{1}{d+1}\left(d \log \lambda^i - \sum_{j \neq i} \log \lambda^j\right) = \log \lambda^i - \frac{1}{d+1}\sum_j \log \lambda^j.$$

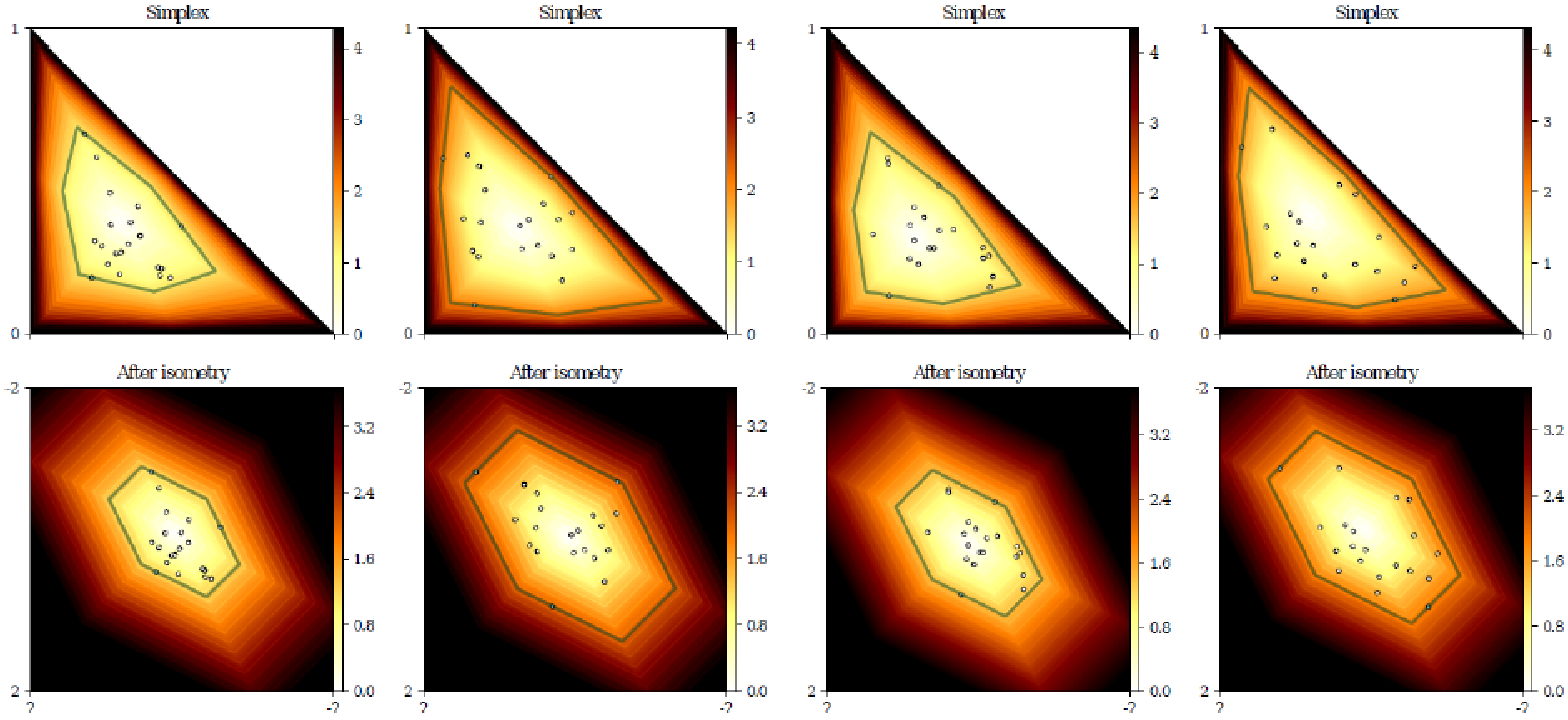
$$\lambda^i = \frac{\exp(v^i)}{\sum_j \exp(v^j)}.$$

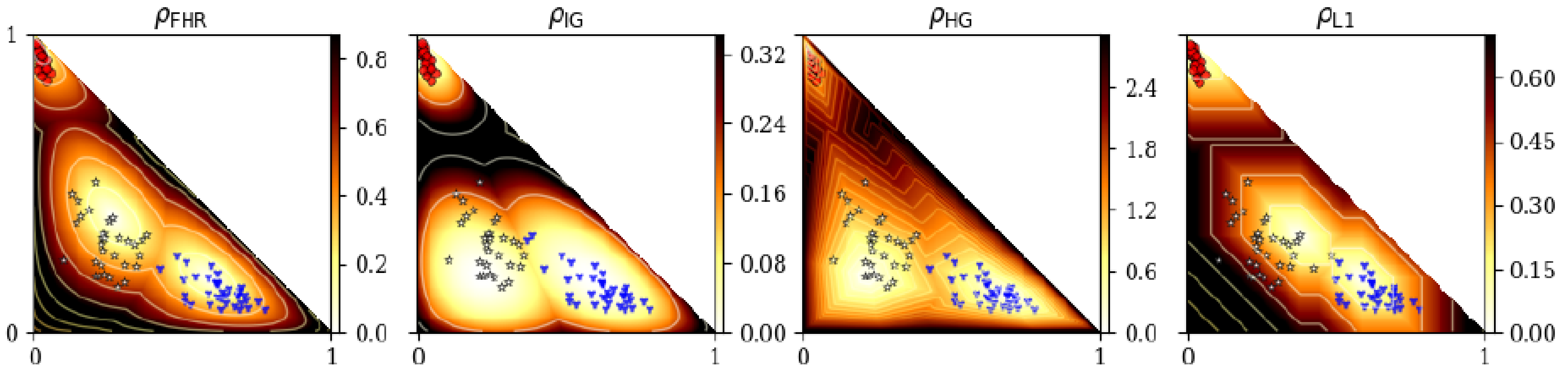
- Norm $\|\cdot\|_{\mathrm{NH}}$ in V^d defined by the shape of its unit ball $B_V = \{v \in V^d \ : \ |v^i - v^j| \leq 1, \forall i \neq j\}$.
- Polytopal norm-induced distance:

$$\rho_V(v, v') = \|v - v'\|_{\mathrm{NH}} = \inf\left\{\tau \ : \ v' \in \tau(B_V \oplus \{v\})\right\},$$

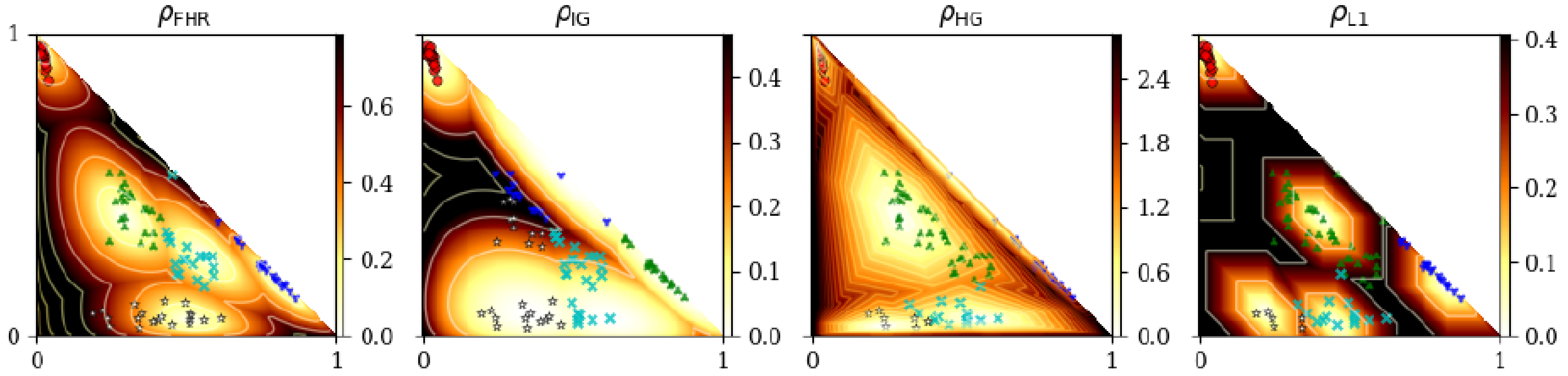
- Norm does not satisfy parallelogram law (no inner product)

Visualizing the isometry: $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$





$k = 3$ clusters



$k = 5$ clusters

K-center clustering in metric spaces

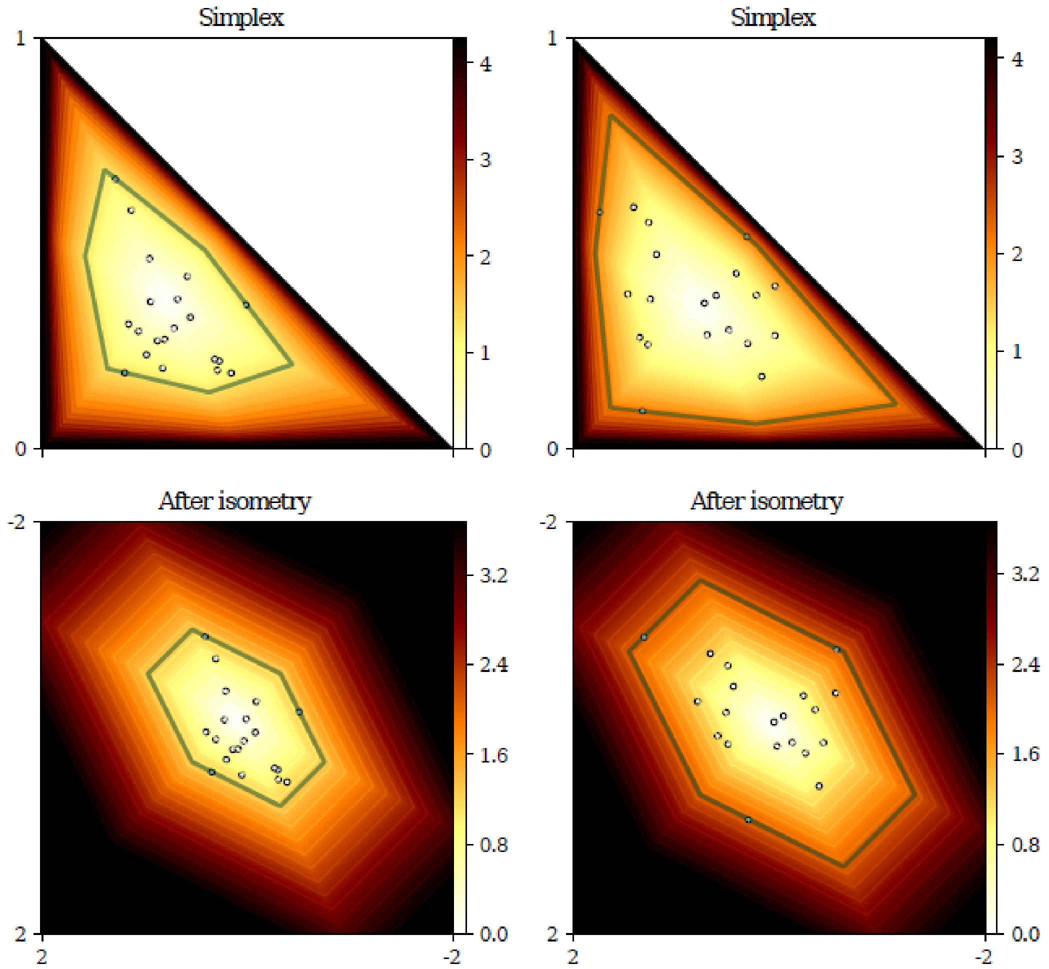
Algorithm : A 2-approximation of the k-center clustering for any metric distance ρ.

Data: A set Λ; a number k of clusters; a metric distance ρ.
Result: A 2-approximation of the k-center clustering

```
begin
    c_1 ← ARandomPointOf(Λ);
    C ← {c_1};
    for i = 2, ··· , k do
        c_i ← argmax_{p∈Λ} ρ(p, C);
        C ← C ∪ {c_i};
Output C;
```

Guaranteed performance: 2-factor for any metric

Smallest enclosing ball in the Hilbert simplex geometry



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

$c_1 \leftarrow$ choose randomly a point in $\mathcal{P}$;

for $i = 2$ **to** l **do**

```
// farthest point from c_i
```

$s_i \leftarrow \arg\max_{j=1}^{n} \rho(c_i, p_j)$;

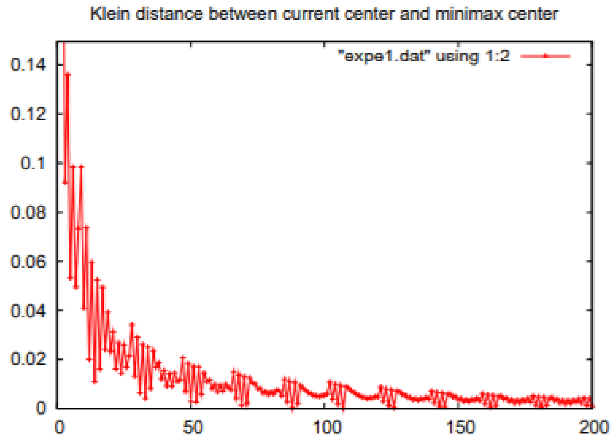
```
// update the center: walk on the geodesic line
   segment [c_i, p_{s_i}]
```

$c_{i+1} \leftarrow c_i \#^M_{\frac{1}{i+1}} p_{s_i}$;

end

```
// Return the SEB approximation
```

return $\mathrm{Ball}(c_l, r_l = \rho(c_l, \mathcal{P}))$;



Hyperbolic geometry:

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

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

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

Positive-definite matrices:

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

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

On Approximating the Riemannian 1-Center, Comp. Geom. 2013
Approximating Covering and Minimum Enclosing Balls in Hyperbolic Geometry, GSI, 2015

Approximating the smallest enclosing ball in Hilbert simplex geometry

Algorithm 4: Geodesic walk for approximating the Hilbert minimax center, generalizing [11]

Data: A set of points $p_1, \cdots, p_n \in \Delta^d$. The maximum number T of iterations.

Result: $c \approx \arg\min_c \max_i \rho_{\mathrm{HG}}(p_i, c)$

begin

$c_0 \leftarrow \mathrm{ARandomPointOf}(\{p_1, \cdots, p_n\})$;

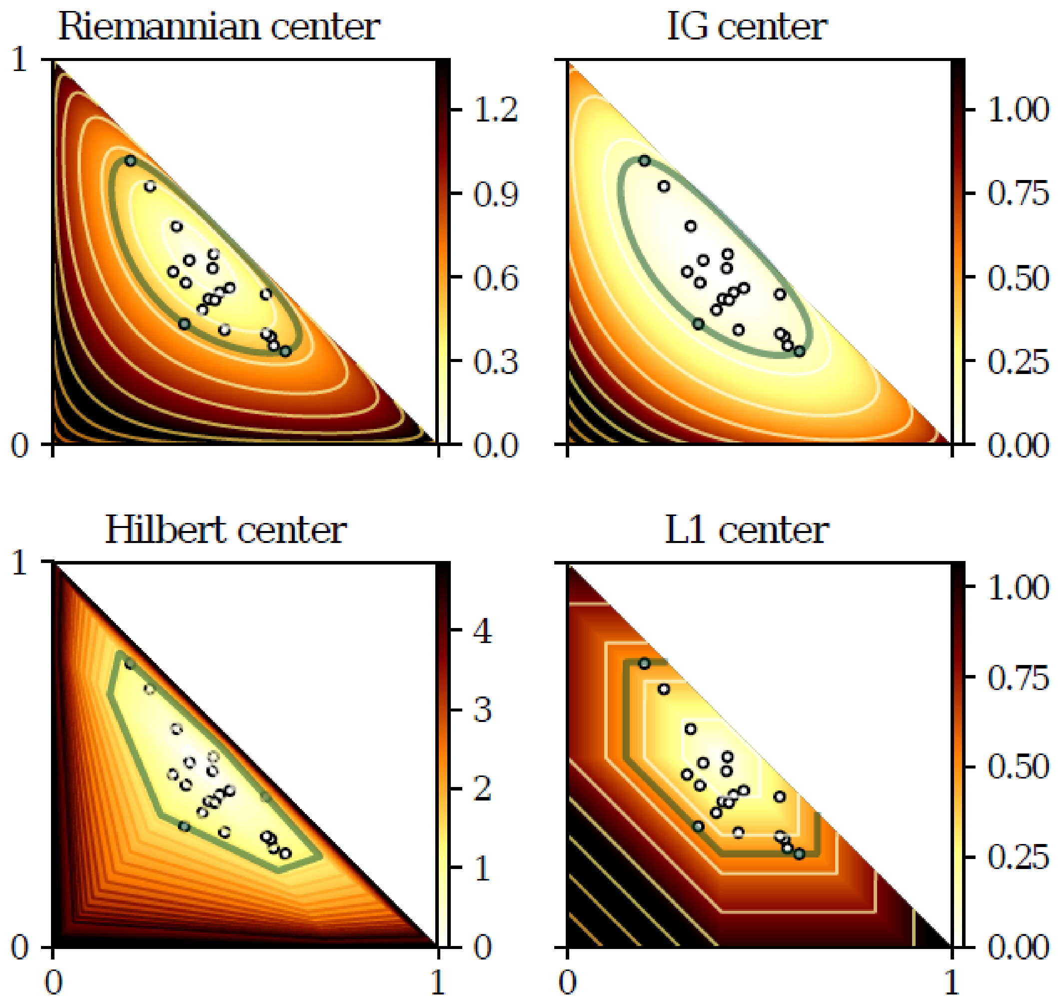
for $t = 1, \cdots, T$ **do**

$p \leftarrow \arg\max_{p_i} \rho_{\mathrm{HG}}(p_i, c_{t-1})$;

$c_t \leftarrow c_{t-1} \#^{\rho}_{1/(t+1)} p$;

Output c_T;

Some enclosing balls in the simplex



Experiments: 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	$\mathbf{0.71 \pm 0.23}$	0.45 ± 0.20	0.54 ± 0.22
			0.9	0.29 ± 0.17	0.27 ± 0.16	$\mathbf{0.39 \pm 0.19}$	0.17 ± 0.13	0.25 ± 0.15
		255	0.5	0.70 ± 0.25	0.69 ± 0.26	$\mathbf{0.74 \pm 0.25}$	0.37 ± 0.29	0.70 ± 0.26
			0.9	$\mathbf{0.42 \pm 0.25}$	0.35 ± 0.20	0.40 ± 0.19	0.03 ± 0.08	$\mathbf{0.44 \pm 0.26}$
	100	9	0.5	0.63 ± 0.22	0.61 ± 0.22	$\mathbf{0.71 \pm 0.22}$	0.46 ± 0.19	0.56 ± 0.20
			0.9	0.29 ± 0.15	0.26 ± 0.14	$\mathbf{0.38 \pm 0.20}$	0.18 ± 0.12	0.24 ± 0.14
		255	0.5	0.71 ± 0.26	0.69 ± 0.27	$\mathbf{0.75 \pm 0.25}$	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	$\mathbf{0.43 \pm 0.26}$
5	50	9	0.5	0.64 ± 0.15	0.61 ± 0.14	$\mathbf{0.70 \pm 0.14}$	0.48 ± 0.14	0.57 ± 0.15
			0.9	0.31 ± 0.12	0.29 ± 0.12	$\mathbf{0.41 \pm 0.15}$	0.20 ± 0.09	0.26 ± 0.10
		255	0.5	0.74 ± 0.17	0.72 ± 0.17	$\mathbf{0.77 \pm 0.16}$	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	$\mathbf{0.47 \pm 0.17}$
	100	9	0.5	0.62 ± 0.14	0.61 ± 0.14	$\mathbf{0.71 \pm 0.14}$	0.46 ± 0.13	0.54 ± 0.14
			0.9	0.30 ± 0.10	0.27 ± 0.11	$\mathbf{0.40 \pm 0.13}$	0.19 ± 0.08	0.25 ± 0.09
		255	0.5	0.73 ± 0.18	0.70 ± 0.18	$\mathbf{0.75 \pm 0.16}$	0.37 ± 0.20	0.73 ± 0.17
			0.9	0.43 ± 0.16	0.35 ± 0.14	0.41 ± 0.12	0.03 ± 0.06	$\mathbf{0.46 \pm 0.18}$

generator 2

Experiments: K-center

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.87 ± 0.19	0.85 ± 0.19	$\mathbf{0.92 \pm 0.16}$	0.72 ± 0.22	0.80 ± 0.20
			0.9	0.54 ± 0.21	0.51 ± 0.21	$\mathbf{0.70 \pm 0.23}$	0.36 ± 0.17	0.44 ± 0.19
		255	0.5	0.93 ± 0.16	0.92 ± 0.18	$\mathbf{0.95 \pm 0.14}$	0.89 ± 0.18	0.90 ± 0.19
			0.9	0.76 ± 0.24	0.72 ± 0.26	$\mathbf{0.82 \pm 0.24}$	0.50 ± 0.28	0.76 ± 0.25
	100	9	0.5	0.88 ± 0.17	0.86 ± 0.18	$\mathbf{0.93 \pm 0.14}$	0.70 ± 0.20	0.80 ± 0.20
			0.9	0.53 ± 0.20	0.49 ± 0.19	$\mathbf{0.70 \pm 0.22}$	0.33 ± 0.14	0.41 ± 0.18
		255	0.5	0.93 ± 0.16	0.92 ± 0.17	$\mathbf{0.95 \pm 0.13}$	0.88 ± 0.19	0.93 ± 0.16
			0.9	0.81 ± 0.22	0.75 ± 0.24	$\mathbf{0.83 \pm 0.22}$	0.47 ± 0.28	0.79 ± 0.22
5	50	9	0.5	0.82 ± 0.13	0.81 ± 0.13	$\mathbf{0.89 \pm 0.12}$	0.67 ± 0.13	0.75 ± 0.13
			0.9	0.50 ± 0.13	0.47 ± 0.13	$\mathbf{0.66 \pm 0.15}$	0.34 ± 0.11	0.40 ± 0.12
		255	0.5	$\mathbf{0.92 \pm 0.11}$	$\mathbf{0.91 \pm 0.12}$	$\mathbf{0.93 \pm 0.11}$	0.87 ± 0.13	$\mathbf{0.92 \pm 0.12}$
			0.9	0.77 ± 0.15	0.71 ± 0.17	$\mathbf{0.85 \pm 0.17}$	0.54 ± 0.19	0.74 ± 0.16
	100	9	0.5	0.83 ± 0.12	0.81 ± 0.13	$\mathbf{0.89 \pm 0.11}$	0.67 ± 0.11	0.76 ± 0.13
			0.9	0.48 ± 0.12	0.46 ± 0.12	$\mathbf{0.66 \pm 0.15}$	0.33 ± 0.09	0.39 ± 0.10
		255	0.5	$\mathbf{0.93 \pm 0.10}$	$\mathbf{0.92 \pm 0.11}$	$\mathbf{0.94 \pm 0.09}$	0.89 ± 0.11	0.92 ± 0.11
			0.9	0.81 ± 0.14	0.74 ± 0.15	$\mathbf{0.84 \pm 0.16}$	0.52 ± 0.19	0.79 ± 0.14

generator 1

Aitchison distance in the simplex

- Non-separable (g= geometric mean)

$$D_\Delta(x_i, x_j) = \left[\sum_{k=1}^{D} \left(\log\left(\frac{x_{ik}}{g(\mathbf{x}_i)}\right) - \log\left(\frac{x_{jk}}{g(\mathbf{x}_j)}\right)\right)^2\right]^{\frac{1}{2}}$$

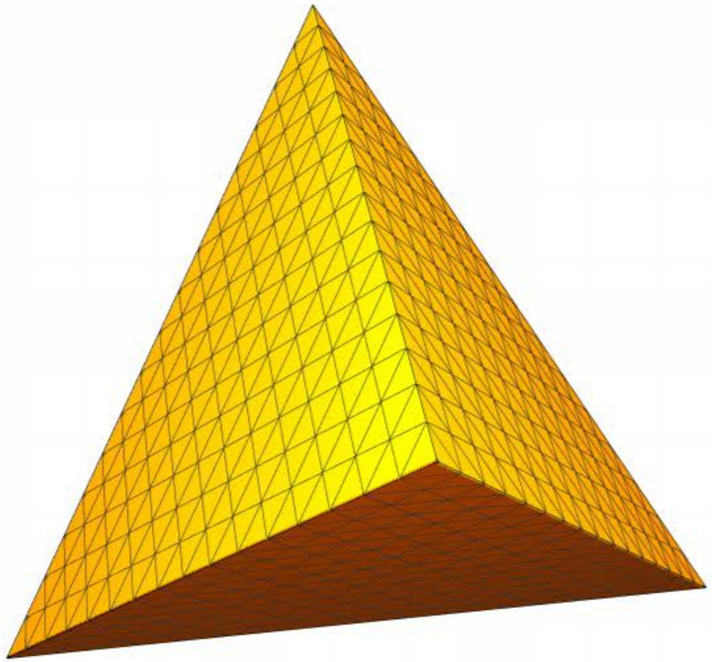
- Invariant by permutation, by scaling, by subcompositional dominance

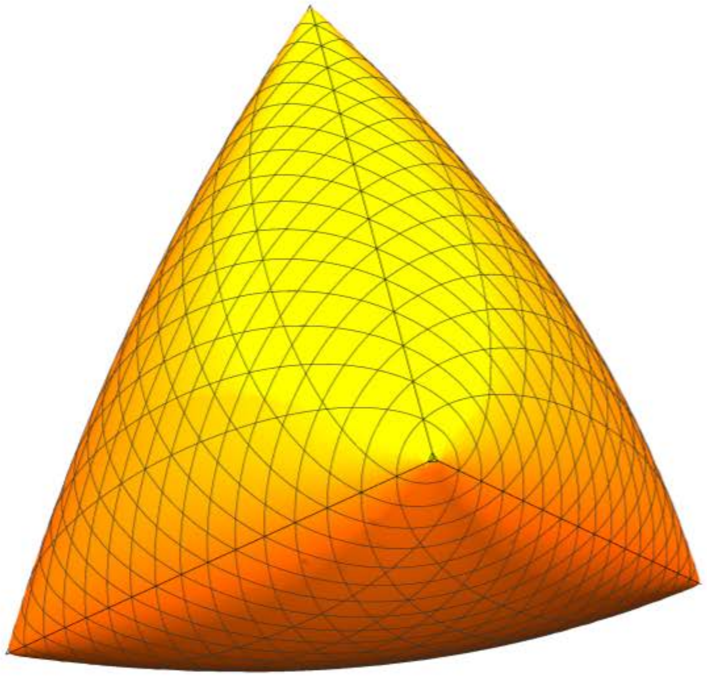
-> Compositional Data (CoDa) Analysis

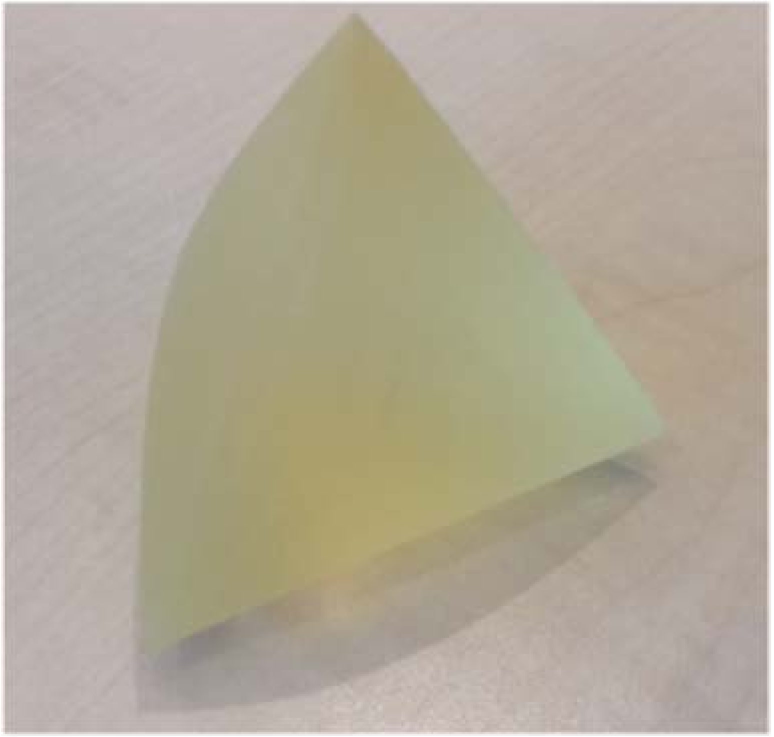
Clustering correlation matrices (elliptope)

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

$$\rho_{\mathrm{HG}}(C_1, C_2) = \left| \log \frac{\|C_1 - C_2'\| \|C_1' - C_2\|}{\|C_1 - C_1'\| \|C_2 - C_2'\|} \right| .$$

- L1-norm

- L2-norm

- Log-det divergence $\rho_{\mathrm{LD}}(C_1, C_2) = tr(C_1 C_2^{-1}) - \log |C_1 C_2^{-1}| - d.$

Experiments of clustering in the elliptope

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**

Information geometry: Advanced topics, limitations and perspectives

Frank Nielsen

Sony CSL

α-representations of the FIM

We introduced the FIM in two ways formerly

$$I(\theta) := (I_{ij}(\theta)), \quad I_{ij}(\theta) := E_{p(x;\theta)}[\partial_i l(x;\theta)\partial_j l(x;\theta)]$$

$$I'_{ij}(\theta) := 4\int \partial_i \sqrt{p(x;\theta)}\partial_j \sqrt{p(x;\theta)}\mathrm{d}\nu(x)$$

α-likelihood function $l^{(\alpha)}(x;\theta) := k_\alpha(p(x;\theta))$

α-Embedding

$$k_\alpha(u) = \begin{cases} \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1 \\ \log u, & \text{if } \alpha = 1. \end{cases}$$

α-representation of the FIM:

$$I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)}(x;\theta)\partial_j l^{(-\alpha)}(x;\theta)d\nu(x)$$

Corresponds to a basis choice in the tangent space (α-base)

α-representations of the FIM

α-Embedding

$$I^{(\alpha)}_{ij}(\theta) = \int \partial_i l^{(\alpha)}(x;\theta)\partial_j l^{(-\alpha)}(x;\theta)d\nu(x)$$

- 0-representation (square root) : $I'_{ij}(\theta) := 4 \int \partial_i \sqrt{p(x;\theta)}\partial_j \sqrt{p(x;\theta)}\mathrm{d}\nu(x)$
- 1-representation (log): $I_{ij}(\theta) := E_{p(x;\theta)}[\partial_i l(x;\theta)\partial_j l(x;\theta)]$
- Under mild regularity conditions:

$$I^{(\alpha)}_{ij}(\theta) = -\frac{2}{1+\alpha}\int p(x;\theta)^{\frac{1+\alpha}{2}}\partial_i\partial_j l^{(\alpha)}(x;\theta)d\nu(x)$$

- Coefficients of the connection: $\Gamma^{(\alpha)}_{ij,k} = \int \partial_i\partial_j l^{(\alpha)}\partial_k l^{(-\alpha)}d\nu(x)$

The α-representations of the Fisher Information Matrix, 2017

(ρ, τ)-representations of the FIM

Smooth convex function and convex conjugates: $f^*(t) = t(f')^{-1}(t) - f\left((f')^{-1}(t)\right)$

τ -representation $\tau(p) = f'(\rho(p)) = \left((f^*)'\right)^{-1}(\rho(p))$

ρ-representation $\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p))$

(ρ, τ)-FIM

$$g_{ij}(\theta) = E_\mu\left\{f''\left(\rho(p(\zeta|\theta))\frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^i}\frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^j}\right\}$$

(ρ, τ)-α-connections

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu\left\{\tfrac{1-\alpha}{2}f'''(\rho(p(\zeta|\theta)))A_{ijk} + f''(\rho(p(\zeta|\theta)))B_{ijk}\right\}$$

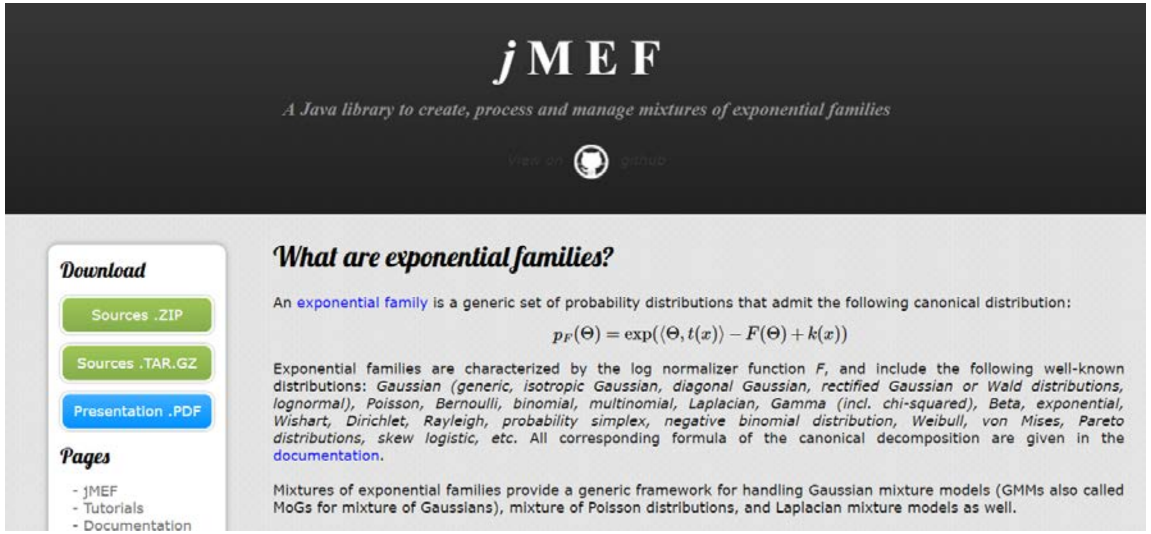
$$A_{ijk}(\zeta,\theta) = \frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^i}\frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^j}\frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^k}, \quad B_{ijk}(\zeta,\theta) = \frac{\partial^2\rho(p(\zeta|\theta))}{\partial\theta^i\partial\theta^j}\frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^k}$$

Zhang, Jun. "On monotone embedding in information geometry." *Entropy* 17.7 (2015

Libraries for Mixture of Exponential Families

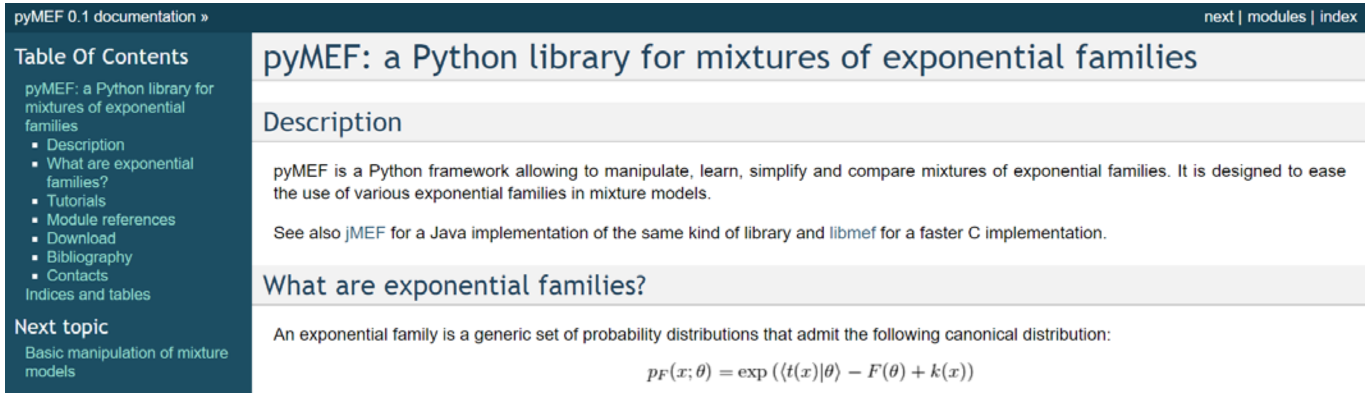
- **jMEF** in Java

http://vincentfpgarcia.github.io/jMEF/



- **pyMEF** in Python

http://www-connex.lip6.fr/~schwander/pyMEF/



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 distance** 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. Two difficulties when using OT: (1) choosing the ground distance, and (2) bad convergence rates of empirical estimators.

Topics to be covered in an extended lecture series

- Deformed exponential families
- Kernel exponential families and deep exponential families
- Non-parametric information geometry
- Wong's logarithmic-divergence and relationship IG with OT via c-divergence
- Quantum information geometry
- Many applications!
- Etc.