Some recent results on statistical distances

Principles and geometry

Frank Nielsen

Sony Computer Science Laboratories Inc

Tokyo, Japan

Sony CSL

May 2024

Outline

I. Background:

- Divergences are *everywhere* in information sciences for different purposes!
- Principles of
  - ① f-divergences
  - ② Bregman divergences
  - ③ Rao distances

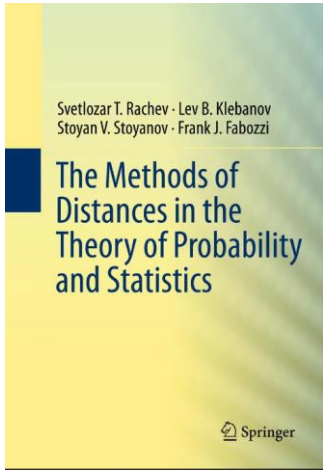
II. In this talk, three recent advances and concepts:

1. Fisher-Rao distances and **projective distances**
2. Bregman divergences and **comparative convexity**
3. f-divergence analysis via **maximal invariant**

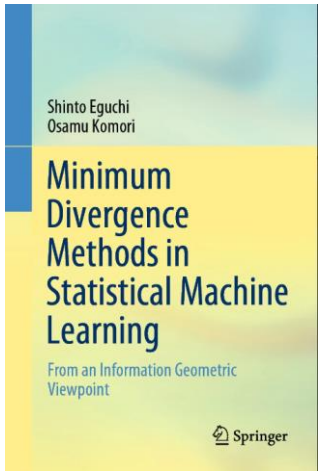
Why statistical distances in information sciences?

- **Probability theory**: convergence theorems wrt probability metrics
- **Statistics:**
  - Divergence-based estimators: dissimilarities between empirical distributions and models
  - Scoring rules: evaluates forecasts, probabilistic predictions
- **Information theory**: Mutual information of random variables
- **Signal processing**: Decompositions, approximate matrix factorization: NMF via β-divergence in sound processing, etc.
- **Machine learning, pattern recognition**: Loss functions for training models, optimal transport, Integral probability metrics (MMDs)
- **Information geometry**: canonical divergences of geometric structures, geometry of divergences

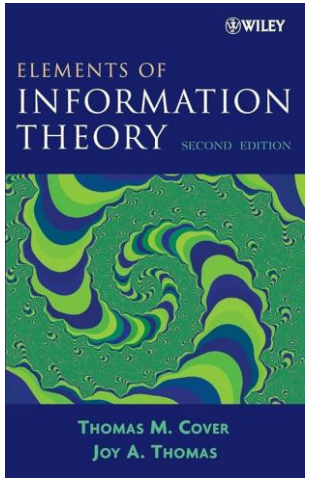
Statistical distances in information sciences



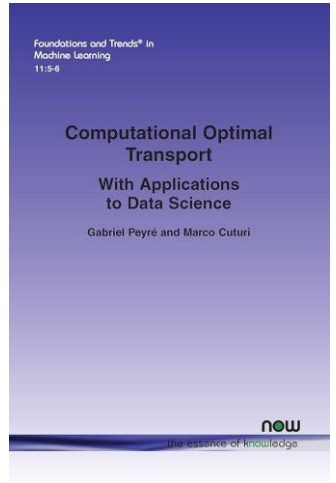
Probability theory



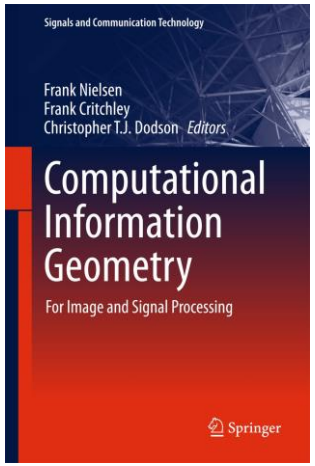
Statistics



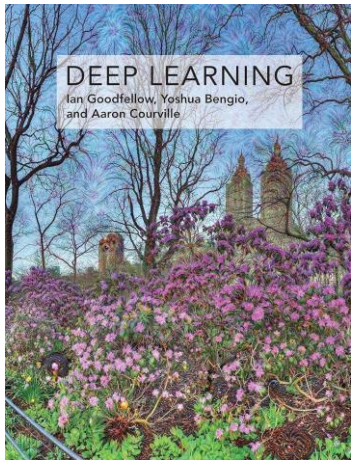
Information theory

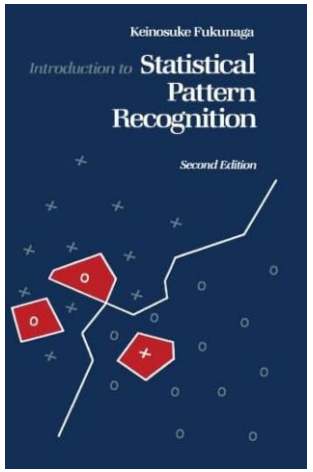


Optimal transport

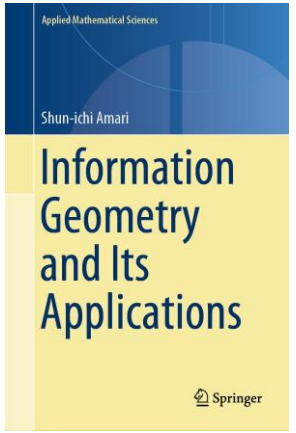


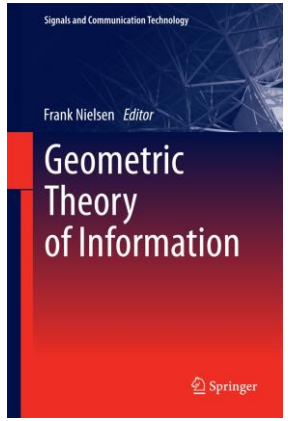
Signal processing





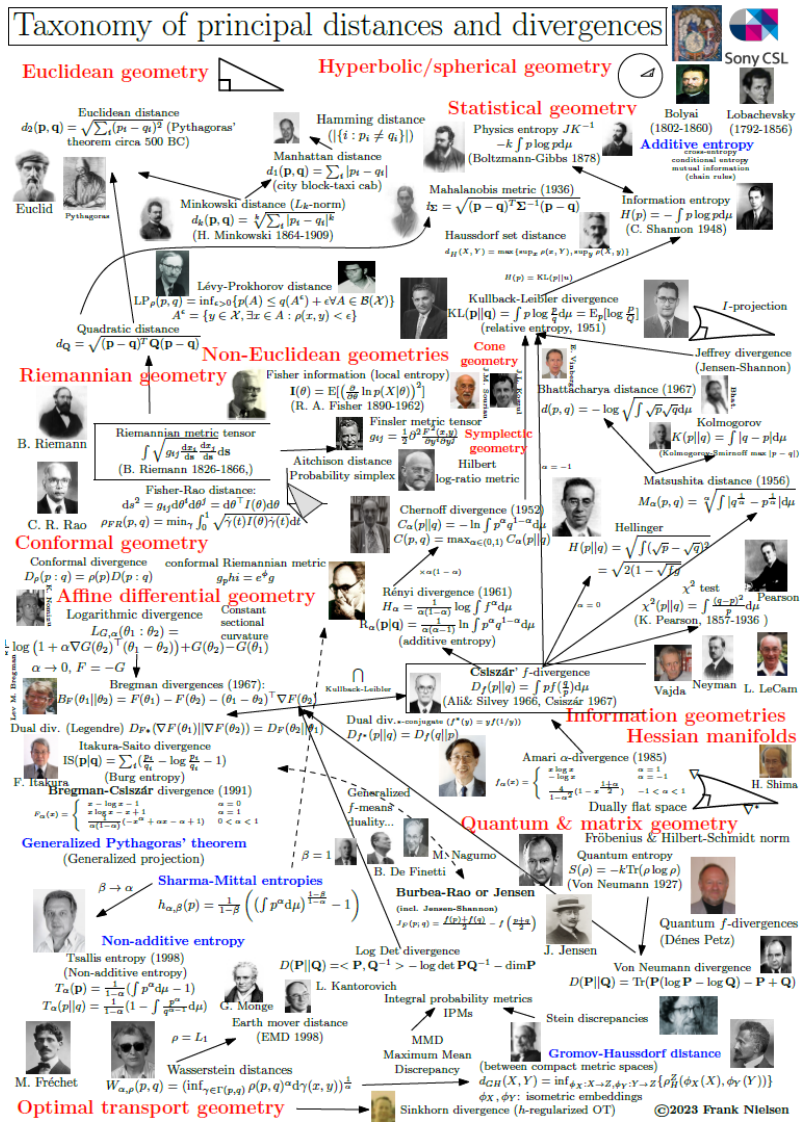
Machine learning
Pattern recognition

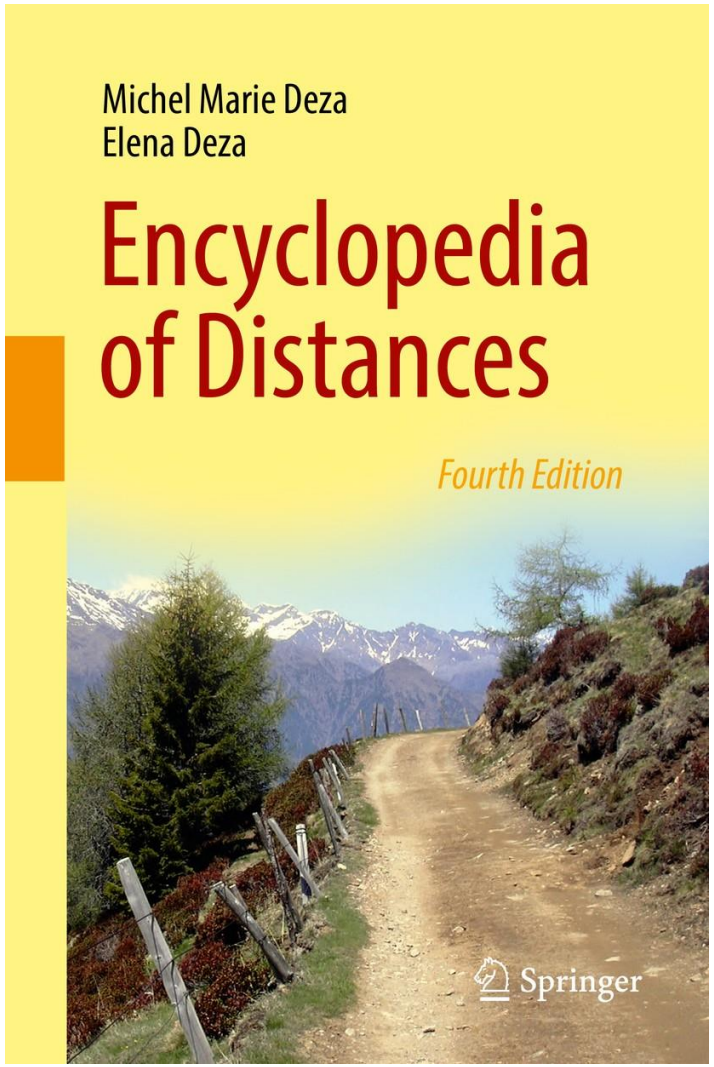




Information geometry

Historically, many distances with purposes…





https://franknielsen.github.io/Divergence/index.html

Seeking principles and properties of distances

- **f-divergences (Ali-Silvey & Csiszár)** f convex, strictly convex at 1

$$I_f(p:q) = \int p f(q/p) \mathrm{d}\mu \qquad I_f(p:q) = \int q f(p/q) \mathrm{d}\mu$$

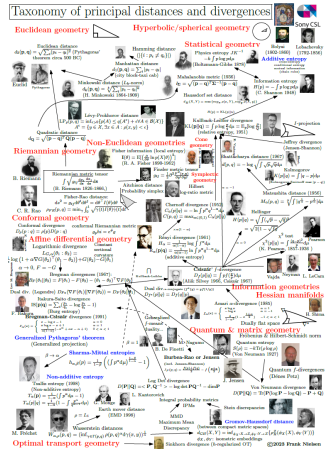
  - Only separable distances which are **monotone** under Markov kernels
  - **Invariant** under ``sufficiency''

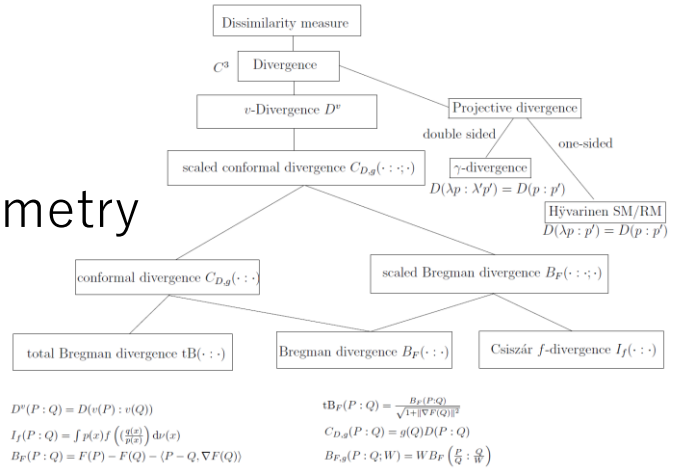
- **Bregman divergences** F strictly convex and smooth

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

  - Only distances with **right-sided centroids = centers of masses**
  - **Canonical divergences of dually flat spaces** in information geometry

- **Transport distances including Wasserstein distances**
  - Minimize **transport cost** wrt a **ground distance**
  - Fast regularized OT with dense plan, fast sparse reg. OT





- Csiszár, "On information-type measure of difference of probability distributions and indirect observations," Studia Sci. Math. Hungar. 2 (1967).
- Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming", USSR computational mathematics and mathematical physics 7.3 (1967)
- Kantorovich, "Mathematical methods of organizing and planning production," Management science 6.4 (1960)
- Nielsen, "An elementary introduction to information geometry", Entropy 22.10 (2020)

Fisher-Rao geodesic metric distances

- **Statistical model** $\{p_\lambda : \lambda \in \Lambda\}$ with model dimension $\dim(\Lambda)=m$
- Use **Riemannian geodesic distance** for the **Fisher metric** expressed in chart (λ, Λ) using **Fisher information matrix:** $g(\lambda) = -E[\nabla_\lambda^2 \log p_\lambda(x)]$

Fisher length of a curve: $$\text{Length}(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} \mathrm{d}t = \int_0^1 \mathrm{d}s_{\mathcal{N}}(t)\mathrm{d}t = \int_0^1 \|\dot{c}(t)\|_{c(t)} \mathrm{d}t,$$

Geodesic is locally length minimizing curve: $$\rho_{\mathcal{N}}(N(\lambda_1), N(\lambda_2)) = \inf_{\substack{c(t) \\ c(0)=p_{\lambda_1} \\ c(1)=p_{\lambda_2}}} \{\text{Length}(c)\},$$

Fisher-Rao distance = Fisher length of geodesic

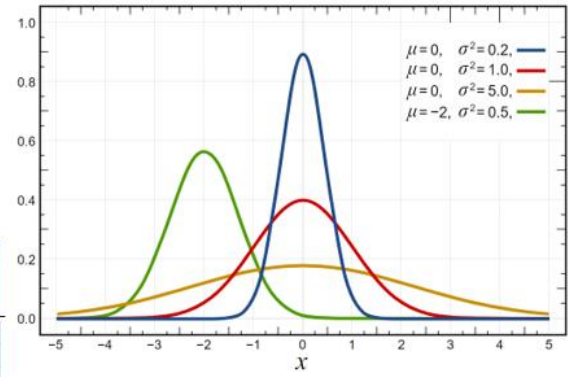
Geodesic wrt Fisher Levi-Civita connection: $$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \Gamma_{ij}^k \frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0$$

- Pro:
  - Invariant to reparamerization: a geometric distance!
  - Role of Riemann-Christoffel, Ricci, sectional **curvatures** in statistics
- Con: Can be difficult to calculate: eg, no formula for multivariate Normals, autodiff

Solve ODE with:
- Initial value problem or
- Boundary value problem

Hyperbolic Fisher-Rao Gaussian manifold and partial isometric embedding on the 3D pseudo-sphere

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



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

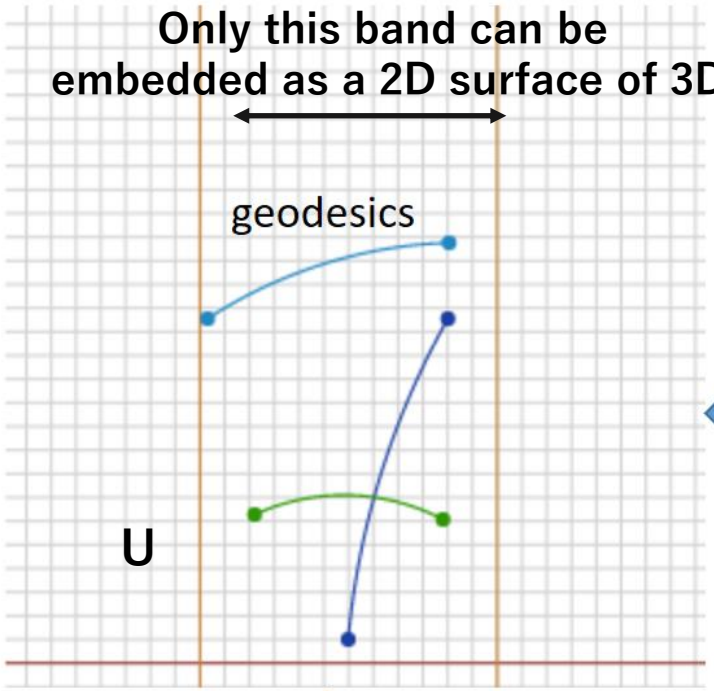
Fisher-Rao geodesic distance:

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

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

Constant curvature -1/2
(= hyperbolic manifold)

Only this band can be embedded as a 2D surface of 3D



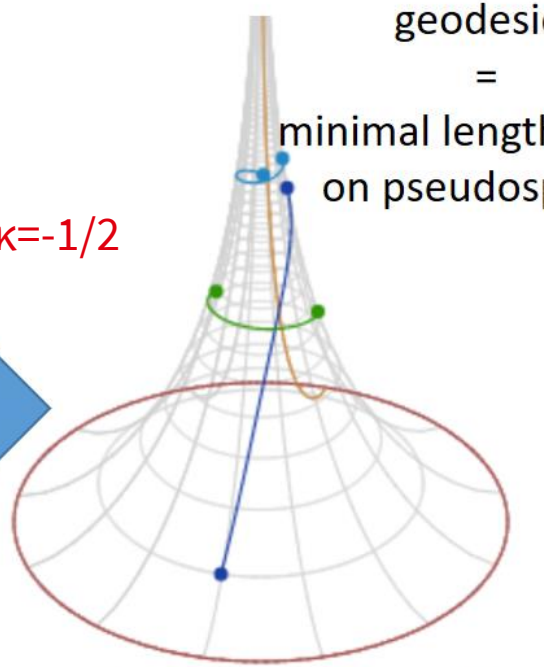
Poincaré upper half-plane

Constant Gaussian negative curvature

Isometric embedding (partial/periodic)

κ=-1/2

geodesics
=
minimal length curves on pseudosphere



Pseudosphere generated by tractrix

Precursors of statistical manifolds/information geometry

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 co-variances of efficient estimates of these parameters. Such a space, for the ordinary types of distributions, is always curved. For the two parameters of the normal law the manifold may be represented inpart as a surface of revolution of negative curvature, with a sharp circular edge. On this surface variation of the dispersion is represented by moving along a generator. For a Pearson Type III curve of any given shape the same surface occurs. For the unrestricted Type III curve there are three parameters; their space is investigated. Certain metrical properties ~~for~~ which hold in general for spaces of statistical parameters are given.

[Hotelling 1930]
Space of statistical parameter

ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

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

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

Comparing with the formula for ds^2

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

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

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

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

[Mahalanobis 1936]
Statistical field

Information and the Accuracy Attainable in the Estimation of Statistical Parameters

C. Radhakrishna Rao

The Population Space

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

$$\phi(x, \theta_1, \ldots, \theta_q)\, dv. \qquad (6.1)$$

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

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

$$d\phi(\theta_1, \theta_2, \ldots, \theta_q) \qquad (6.2)$$

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

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

where

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

[Rao 1945]
Population space

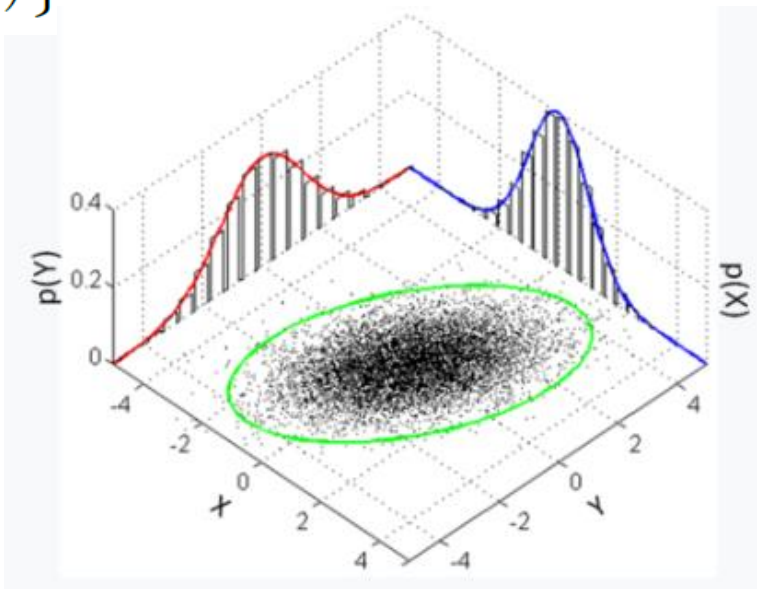
Fisher-Rao geometry: multivariate normals (MVNs)

$$N(\mu,\Sigma) \sim p_{\mu,\Sigma}(x) = \frac{(2\pi)^{-\frac{d}{2}}}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right)$$

$$\mathcal{N}(d) = \left\{ N(\lambda) \ : \ \lambda = (\mu,\Sigma) \in \Lambda(d) = \mathbb{R}^d \times \mathrm{Sym}_+(d,\mathbb{R}) \right\}$$

Fisher information matrix (vector, matrix):

$$g_{\mathcal{N}}^{\mathrm{Fisher}}(\mu,\Sigma) = \mathrm{Cov}[\nabla \log p_{(\mu,\Sigma)}(x)]$$



Fisher metric tensor:

$$\begin{aligned} g_{(\mu,\Sigma)}^{\mathrm{Fisher}}((v_1,V_1),(v_2,V_2)) &= \langle (v_1,V_1),(v_2,V_2)\rangle_{(\mu,\Sigma)}, \\ &= [v_1]^\top \Sigma^{-1}[v_2] + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}[V_1]\Sigma^{-1}[V_2]\right). \end{aligned}$$

Length element:

$$\begin{aligned} \mathrm{d}s^2_{\mathcal{N}}(\mu,\Sigma) &= \begin{bmatrix} \mathrm{d}\mu \\ \mathrm{d}\Sigma \end{bmatrix}^\top I(\mu,\Sigma) \begin{bmatrix} \mathrm{d}\mu \\ \mathrm{d}\Sigma \end{bmatrix}, \\ &= \mathrm{d}\mu^\top \Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right). \end{aligned}$$

v= vector space R^d
V=Symmetric matrix vector space

[Skovgaard 1984]

⚠ **Non-constant sectional curvatures which can also be positive, not a NPC space (d>1)**

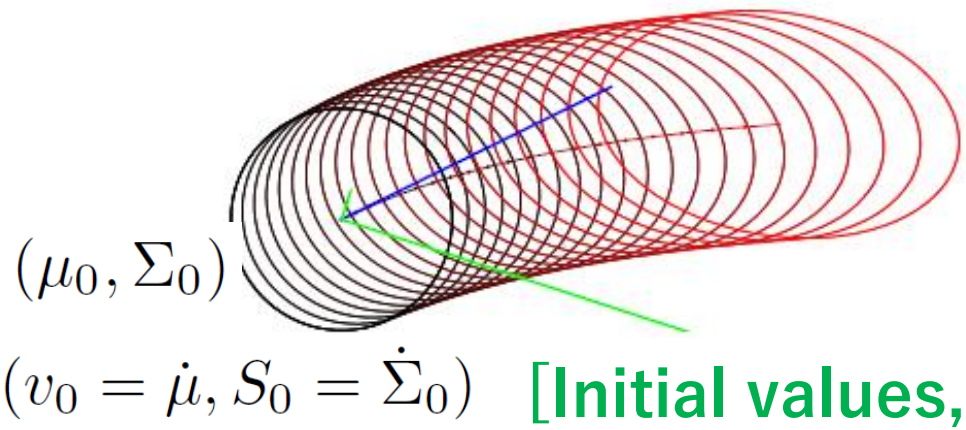
Fisher-Rao geodesic equation for MVNs:

Using (vector, Matrix) parameterization:

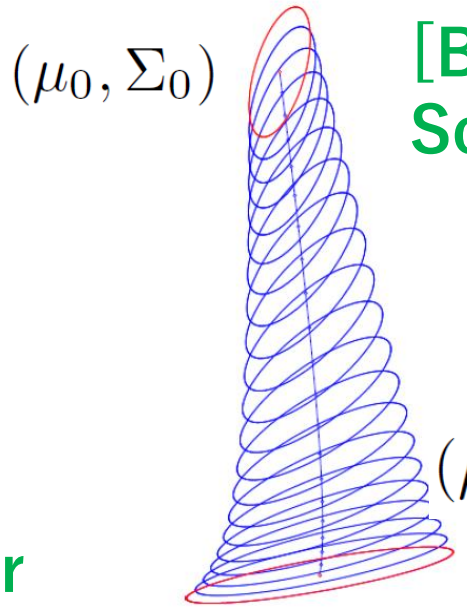
Second-order ODE:

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

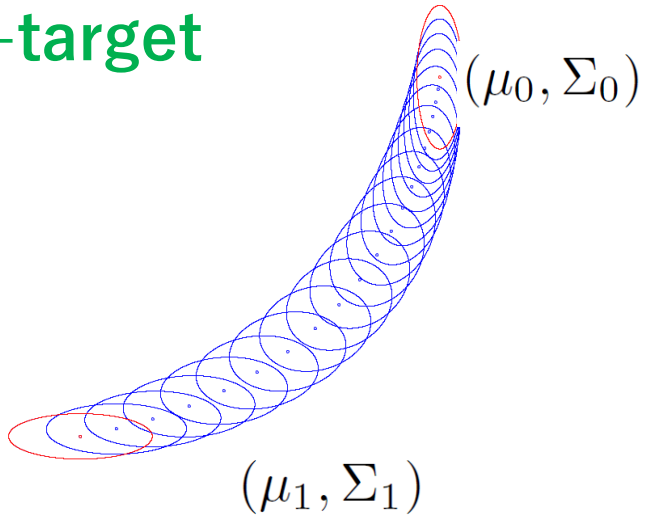
- Consider either **initial value conditions** or **boundary value conditions** of ODE



[Initial values, 1987]
Source+tangent vector



[Boundary values, 2023]
Source+target

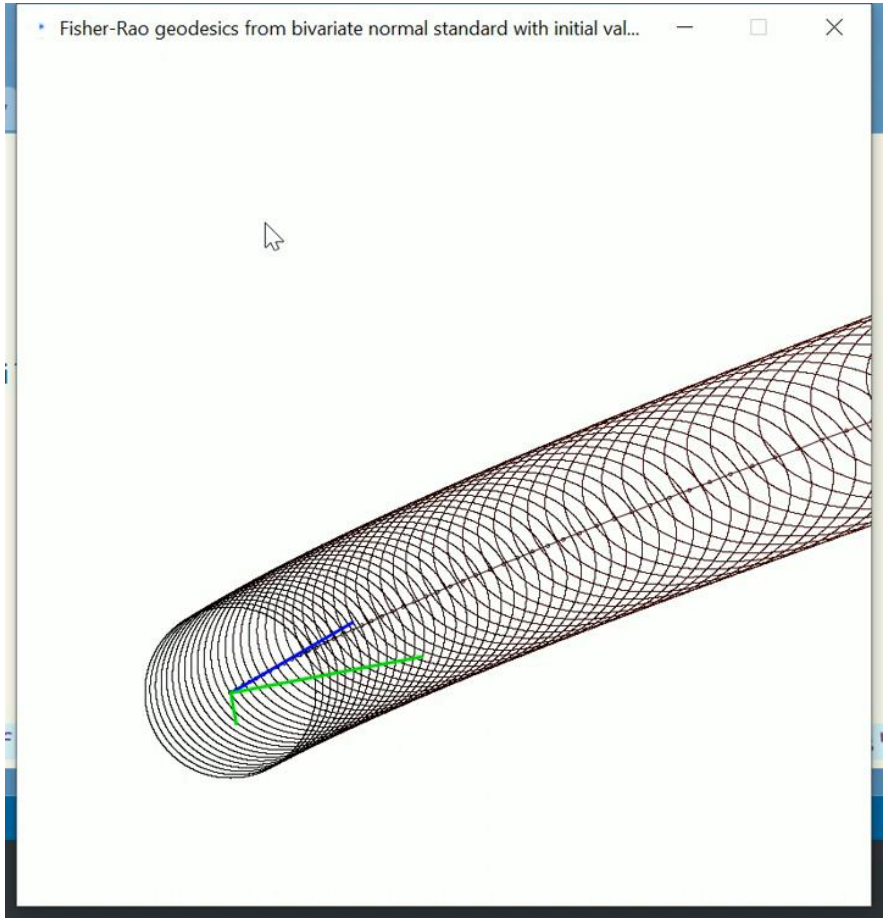


- Yet, once Fisher-Rao geodesics are known, integrate **Fisher-Rao length elements to get Rao distance.**

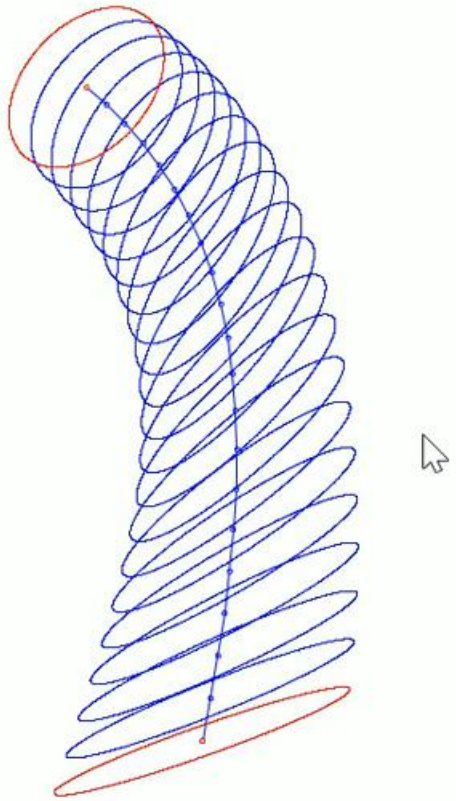
$$\mathrm{Length}(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} \mathrm{d}t = \int_0^1 \mathrm{d}s_{\mathcal{N}}(t) \mathrm{d}t = \int_0^1 \|\dot{c}(t)\|_{c(t)} \mathrm{d}t.$$

MVN: No closed form yet…

Demos: MVN Fisher-Rao geodesics on bivariate normal Fisher-Rao manifold



Geodesics with initial values

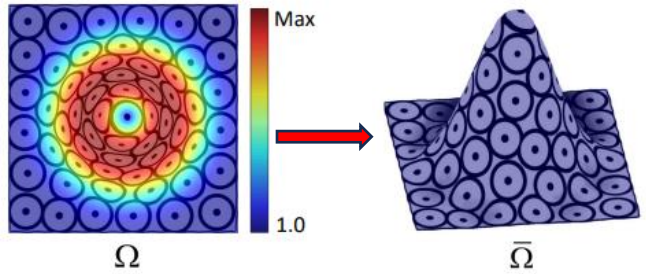


Geodesics with boundary values

Strategy to get lower bounds on Fisher-Rao distances

- **Nash embedding theorem**: A Riemannian manifold (M,g) of dimension m can always be **embedded** in Euclidean space $(E,g_{Euclidean})$ of dimension $O(m^3)$.

Intrinsic
2D manifold



2D manifold embedded in 3D

Figure from
[Zhong et al' 2018]

- Let $S=\{f(p), p \in M\}$ be the submanifold. If geodesics between f(p) and f(q) stay in S wrt $g_{Euclidean}$ then the manifold is **totally geodesic**
- Riemannian distance in $(E,g_{Euclidean})$ is $||\mathbf{f(p)-f(q)}||_2$

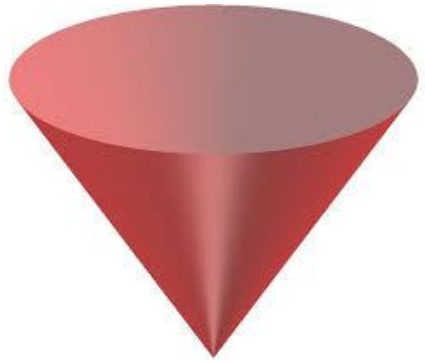
- Rao distance in (M,g) **lower bounded** by $||f(p)-f(q)||_2$: ⟹ $\rho(p,q) \geq ||f(p)-f(q)||_2$

- We may embed (M,g) into high-dimensional non-Euclidean manifolds (M',g') too (next slide!) and get

$$\rho_M(p,q) \geq \rho_{M'}(f(p),f(q))$$

Diffeomorphic embeddings of MVN(d) onto SPD(d+1)

The **diffeomorphisms** $\{f_\beta\}$ foliates the SPD cone P(d+1)



$$f_\beta(N) = f_\beta(\mu, \Sigma) = \begin{bmatrix} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{bmatrix} \in \mathcal{P}(d+1)$$

Using **half trace metric** in P(d+1), get the following **metrics on MVN(d)**:

$$\begin{aligned} \mathrm{d}s^2_{\mathrm{CO}} &= \frac{1}{2}\mathrm{tr}\left(\left(f^{-1}(\mu,\Sigma)\mathrm{d}f(\mu,\Sigma)\right)^2\right), \\ &= \frac{1}{2}\left(\frac{\mathrm{d}\beta}{\beta}\right)^2 + \beta\mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right). \end{aligned}$$

[Calvo & Oller 1990]

$\beta = 1$

When $\beta = 1$ (constant), we thus get a **Fisher isometric embedding** of MVN(d) into SPD(d+1):

$$\mathrm{d}s^2_{\mathrm{Fisher}} = \mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right)$$

Fisher-Rao MVN distance: A lower bound

- Embed isometrically the Gaussian manifold N(d) into a **submanifold of codimension 1 into the SPD cone of dimension d+1: non-totally geodesic submanifold** {f(N)}:

[Calvo & Oller 1990]

$$f(N) = f(\mu, \Sigma) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix} \qquad \beta = 1$$

- Use closed-form **SPD geodesic** in the (d+1)-dimensional cone:

$$\Sigma_t = \Sigma_0^{\frac{1}{2}} \left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}\right)^t \Sigma_0^{\frac{1}{2}}$$

- SPD path is of length necessarily smaller than the MVN geodesic in submanifold f(N). Thus get a **lower bound** on Rao distance:

$$\rho_{\mathcal{N}}(N_1, N_2) \geq \rho_{\mathrm{CO}}(\underbrace{f(\mu_1, \Sigma_1)}_{\bar{P}_1}, \underbrace{f(\mu_2, \Sigma_2)}_{\bar{P}_2}) = \sqrt{\frac{1}{2} \sum_{i=1}^{d+1} \log^2 \lambda_i(\bar{P}_1^{-1} \bar{P}_2)}.$$

New fast distances between MV Normals

$$\rho_{\mathrm{Hilbert}}(N_0, N_1) := \rho_{\mathrm{Hilbert}}(f(N_0), f(N_1))$$

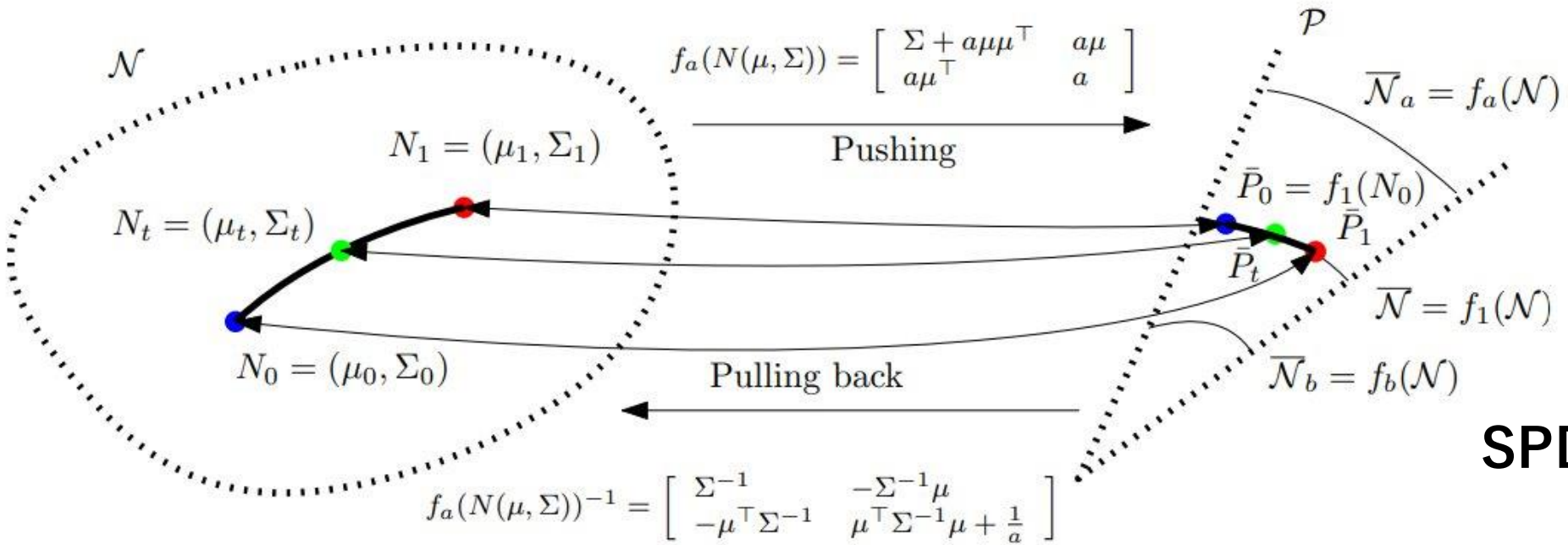
$$f(N) = f(\mu, \Sigma) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$$

Gaussian(d) manifold

Projective Hilbert SPD distance:

$$\rho_{\mathrm{Hilbert}}(P_0, P_1) = \log\left(\frac{\lambda_{\max}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}{\lambda_{\min}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}\right)$$

$$= \log\left(\frac{\lambda_{\max}(P_0^{-1} P_1)}{\lambda_{\min}(P_0^{-1} P_1)}\right)$$



SPD(d+1) cone

"Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures", ICML TAG 2023

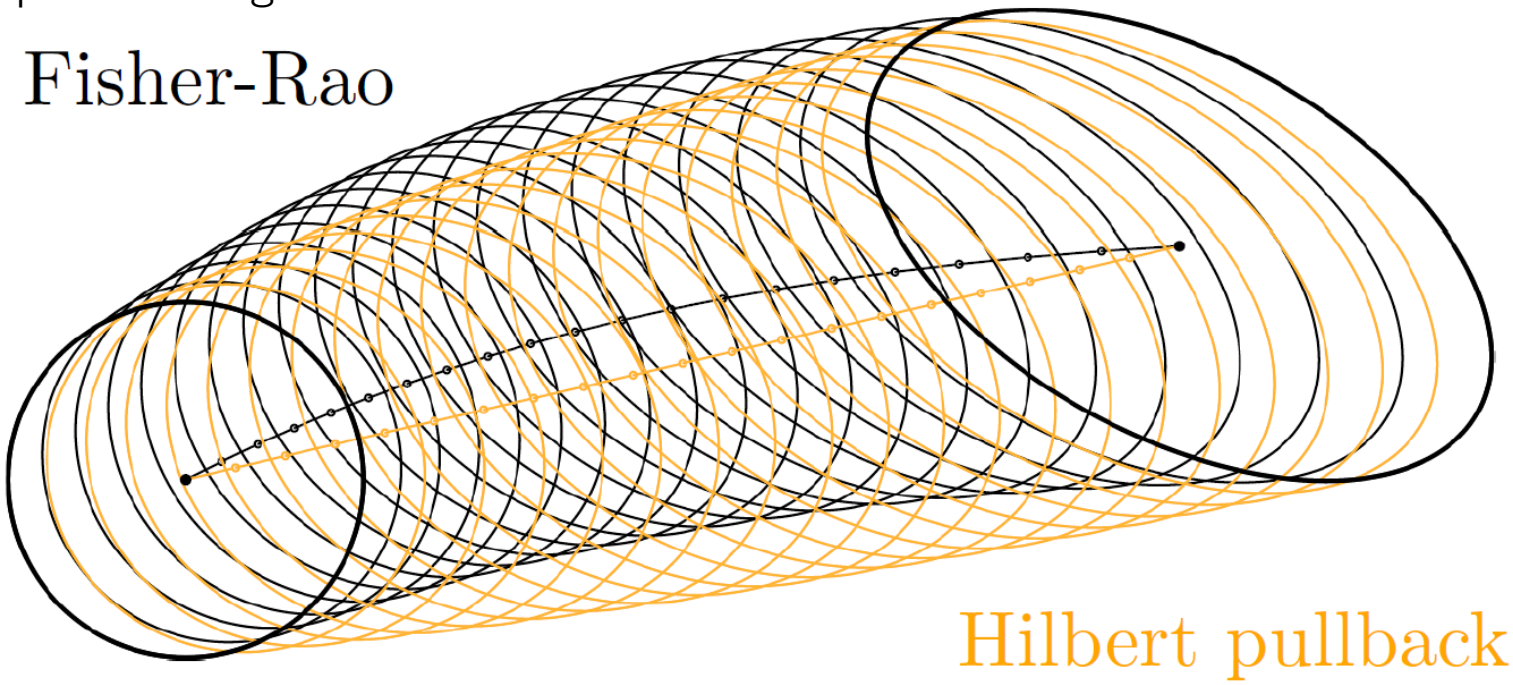
- MVN Fisher-Rao distance needs approximations

by sampling geodesics, require **all eigenspectrum** of SPD matrices.

- Hilbert SPD distance only requires to calculate **extreme eigenvalues**

(eg., power method iterations), + geodesics are in simple closed form

Comparison of geodesics for two bivariate normal distributions shown by ellipses centered at means

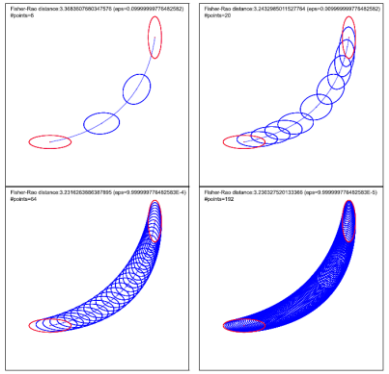


A simple approximation method for the Fisher–Rao distance between multivariate normal distributions, Entropy 25.4 (2023): 654

Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures, ICML TAG 2023.

Fisher-Rao distances for MultiVariate Normals

- Geodesic equation solved for boundary values recently [Kobayashi 2023]



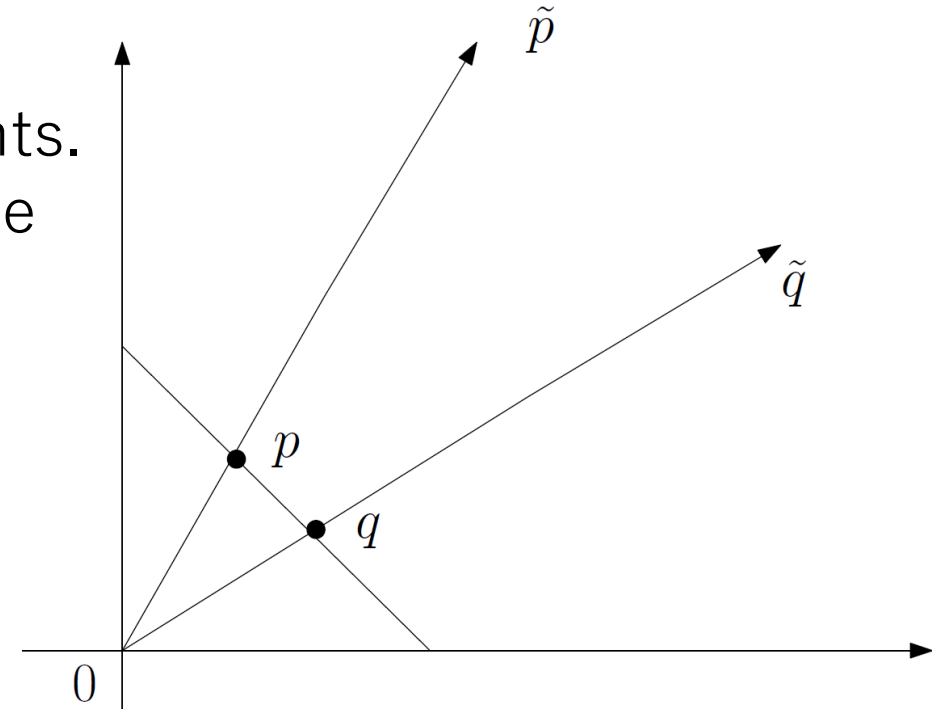
- Guaranteed **1+ ε approximation algorithms** [N1]
- New fast distances based on Hilbert projective geometry of the symmetric positive-definite cone [N2]
- General principles for approximating and bounding Fisher-Rao distances, specially when Fisher metric is Hessian: [N3] $\exists F(\theta) \ : \ g(\theta) = \nabla^2 F(\theta) \succ 0$

- [K] Kobayashi, Shimpei. "Geodesics of multivariate normal distributions and a Toda lattice type Lax pair", Physica Scripta 98.11 (2023)
- [N1] "A simple approximation method for the Fisher–Rao distance between multivariate normal distributions", Entropy 25.4 (2023): 654
- [N2] “Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures", Topological, Algebraic and Geometric Learning Workshops 2023. PMLR, 2023.
- [N3] "Approximation and bounding techniques for the Fisher-Rao distances", arXiv:2403.10089 (2024)

Projective divergences beyond Hilbert/Birkhoff

Projective divergences are pseudo-divergences ☺ which are invariant under rescaling of their arguments.
= Divergences between rays on the positive measure orthant cone

$$D(\lambda p : \lambda' q) = D(p : q), \quad \forall \lambda, \lambda' > 0$$



For example, **Cauchy-Schwarz divergence:**

$$\mathrm{CSD}\,(\tilde{p}, \tilde{q}) = -\log \frac{\int \tilde{p}(x)\tilde{q}(x)\mathrm{d}\nu(x)}{\sqrt{\int \tilde{p}(x)^2\mathrm{d}\nu(x)) \int \tilde{q}(x)^2\mathrm{d}\nu(x)}}$$

and **Hölder divergences**:

$$\mathrm{HD}_{\alpha,\rho,\tau}\,(\tilde{p} : \tilde{q}) = -\log \left(\frac{\int \rho(\tilde{p}(x))\tau(\tilde{q}(x))\mathrm{d}\nu(x)}{\left(\int \rho(\tilde{p}(x))^\alpha \mathrm{d}\nu(x)\right)^{\frac{1}{\alpha}} \left(\int \tau(\tilde{q}(x))^\beta \mathrm{d}\nu(x)\right)^{\frac{1}{\beta}}} \right)$$

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1 \; (\beta = \frac{\alpha}{\alpha - 1} > 1)$$

N, Sun, and Marchand-Maillet, "On Hölder projective divergences." Entropy 19.3 (2017)

Projective divergences for statistical inference

- **Half-sided projective divergences** are useful for estimating parameters of densities which have intractable normalizers:

For example, **Hyvarinen/Fisher divergence** in score matching

$$D_{\mathrm{Hyv}}[p:q] := \frac{1}{2}\int \left\| \nabla_x \log \frac{p(x)}{q(x)} \right\|^2 p(x)\mathrm{d}\mu(x)$$

Empirical distribution $p_e(x) = \frac{1}{n}\sum_i \delta_{x_i}(x)$

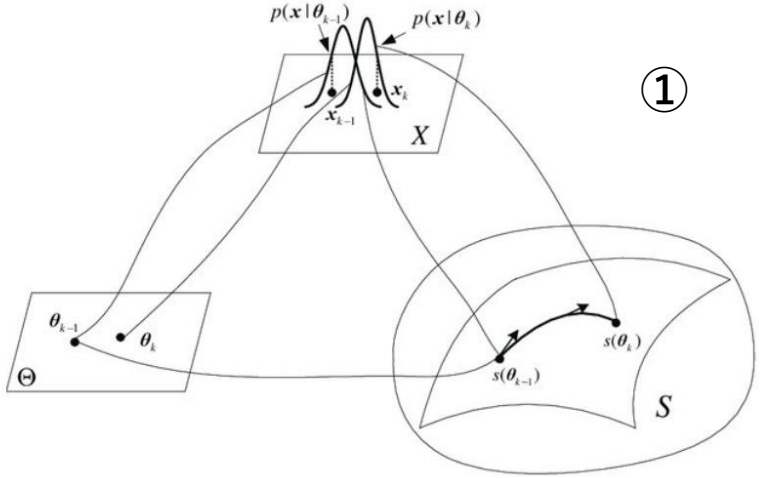
Intractable model because of Z $q_\theta(x) = \frac{\tilde{q}_\theta(x)}{Z(\theta)}$

Since we have:

$$\nabla_x \log q_\theta(x) = \nabla_x \log \tilde{q}_\theta(x) \Rightarrow \min_\theta D_{\mathrm{Hyv}}[p_e : q_\theta] = \min_\theta D_{\mathrm{Hyv}}[p_e : \tilde{q}_\theta]$$

Estimate unnormalized models

What is a “Statistical manifold”?



①

Two meanings in the literature:

① **Typed statistical model manifold** versus

② **manifold with a dualistic structure**

Figure from
[Cheng et al. 2017]

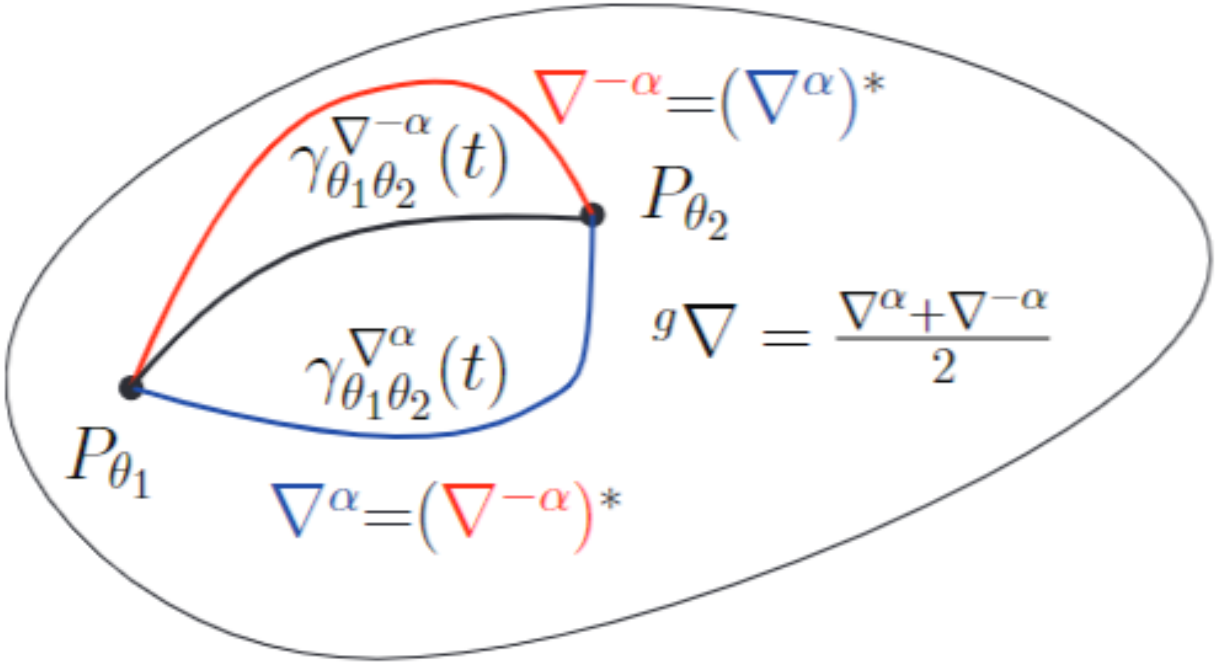
- ① **Manifold of statistical models** (points are typed as statistical models):

  For example, the Fisher-Rao manifold of Gaussian distributions

- ②**Manifold with a ``statistical structure”** (differential geometry): Lauritzen coined the term “statistical manifold” for manifolds equipped with **structure (g,C)**, a Riemannian metric tensor g and a totally symmetric cubic tensor C (next slide!)

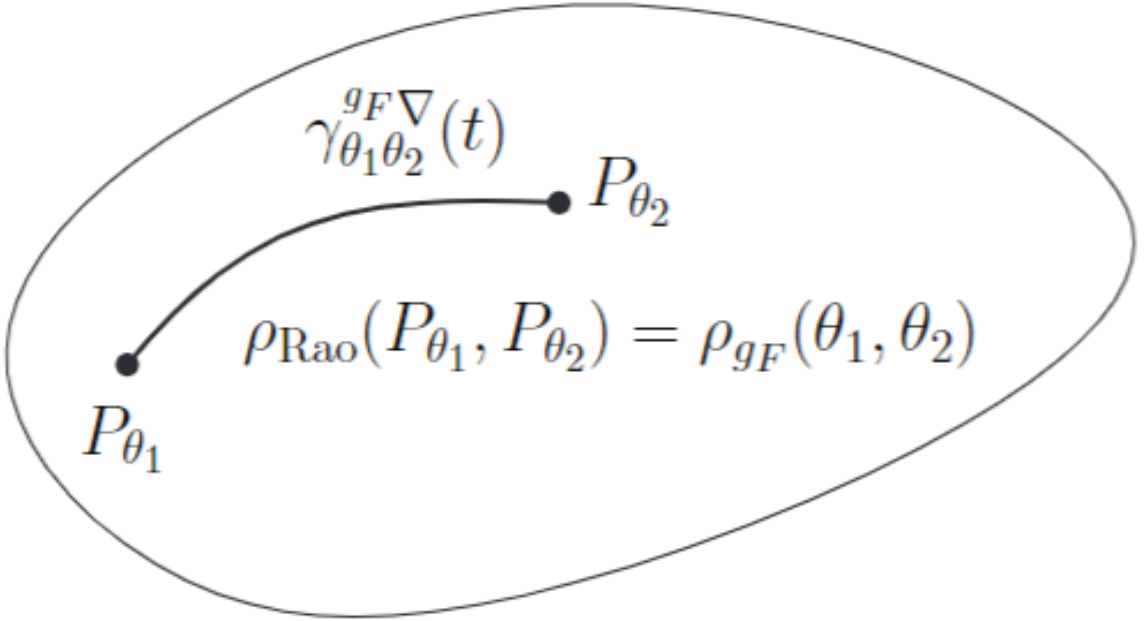
Lee et al., "A statistical manifold framework for point cloud data", Intl Conf on Machine Learning, ICML 2022.
Lauritzen, "Statistical manifolds", Differential geometry in statistical inference 10 (1987): 163-216

Dualistic structures of statistical manifolds: Pure geometric structures



Dual α-geometry
→ No default divergence

Fisher-Rao geometry
→ Fisher-Rao geodesic distance



α-manifold

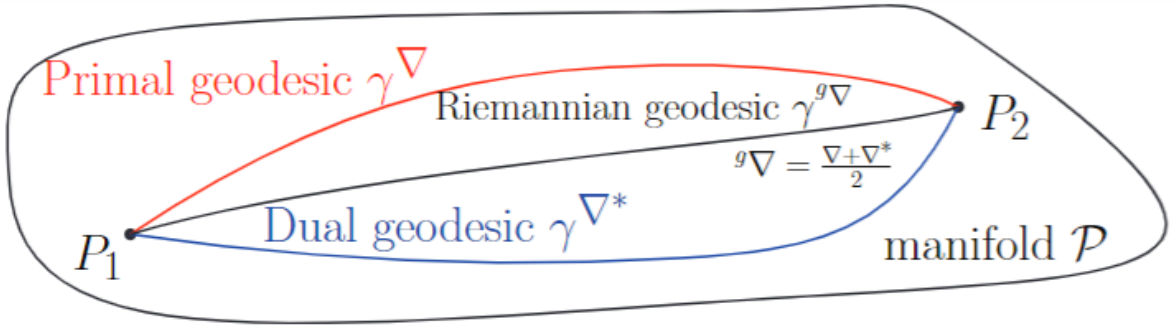
0-manifold = Riemannian manifold

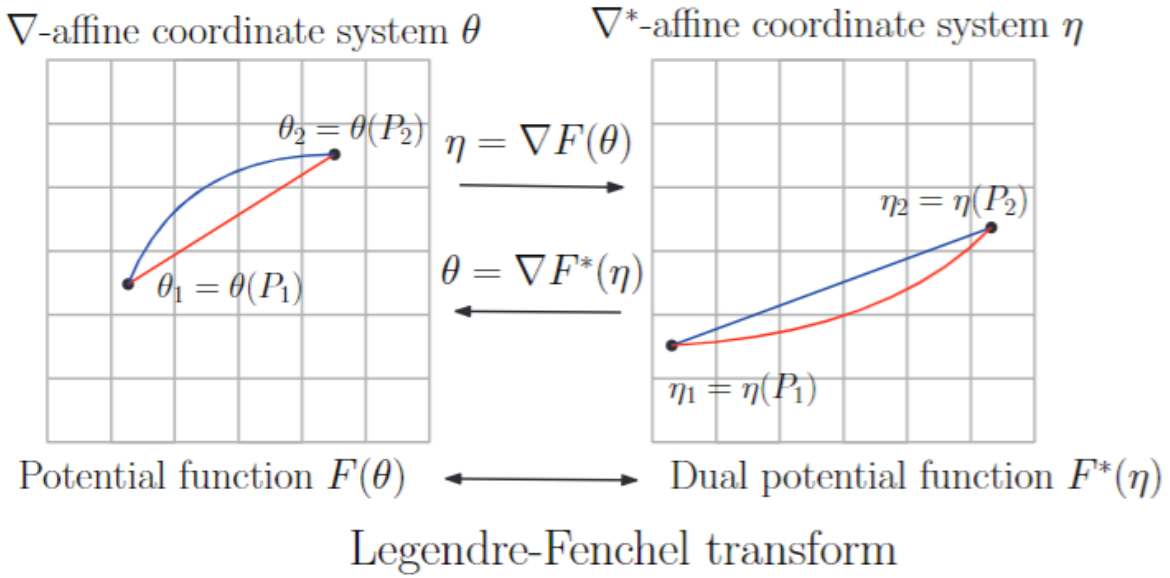
α -geometry from (g,C) structure

https://www.ams.org/journals/notices/202201/rnoti-p36.pdf

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

Dually flat spaces (M,g, ∇ , ∇ *): Bregman manifolds





Also called **Hessian manifolds**

- A connection ∇ is **flat** if there exists a coordinate system θ such that all Christoffel symbols vanish: $\Gamma(\theta)=0$.
- θ is called **∇ –affine coordinate system**
- **∇-geodesic** solves as **line segments**

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

"On geodesic triangles with right angles in a dually flat space." Progress in Information Geometry: Theory and Applications, 2021

Bregman manifolds in Statistics/ML

- Whenever you have a strictly convex smooth function, you can build a **dually flat space** (M,g, ∇ , ∇ *).
- Whenever you have a dually flat space (M,g, ∇ , ∇ *), you can **reconstruct** the convex conjugates F(θ), F*(η) , and the dual Bregman divergences
- Thm: **Bregman divergences = canonical divergences of dually flat spaces**
- In Stats/ML, we often consider **exponential families** with densities

$$p_\theta(x) \propto \tilde{p}_\theta(x) = \exp\left(\sum_{i=1}^{m} \theta_i x_i\right) \qquad p_\lambda(x) \propto \tilde{p}_\lambda(x) = \exp(\langle \theta(\lambda), t(x) \rangle)\, h(x)$$

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

- It turns out that the normalizer Z(θ) **partition function** AND the log-normalizer F(θ)=log Z(θ) **cumulant function** are both convex!
- So we get **two dually flat spaces** built from Z or F yielding Bregman divergences. Classically DFS from F was considered…

Two Bregman manifolds from partition/cumulants functions of exponential families, two pairs of BDs

$$Z(\theta) = \int \tilde{p}_\theta(x)\mathrm{d}\mu(x) \qquad \mathbf{F(\theta)=\log Z(\theta)} \quad \mathbf{Z(\theta)=\exp(F(\theta)=)}$$

① $$B_Z(\theta_1 : \theta_2) = Z(\theta_1) - Z(\theta_2) - \langle \theta_1 - \theta_2, \nabla Z(\theta_2) \rangle \geq 0,$$

② $$B_{\log Z}(\theta_1 : \theta_2) = \log\left(\frac{Z(\theta_1)}{Z(\theta_2)}\right) - \left\langle \theta_1 - \theta_2, \frac{\nabla Z(\theta_2)}{Z(\theta_2)} \right\rangle \geq 0,$$

And furthermore, we can define **skewed Jensen divergences** from the convex generators:

① $$J_{Z,\alpha}(\theta_1 : \theta_2) = \alpha Z(\theta_1) + (1-\alpha) Z(\theta_2) - Z(\alpha\theta_1 + (1-\alpha)\theta_2) \geq 0,$$

② $$J_{\log Z,\alpha}(\theta_1 : \theta_2) = \log \frac{Z(\theta_1)^\alpha Z(\theta_2)^{1-\alpha}}{Z(\alpha\theta_1 + (1-\alpha)\theta_2)} \geq 0.$$

Including the **symmetric Jensen divergence** when α=1/2:

$$J_F(\theta_1, \theta_2) = J_{F,\frac{1}{2}}(\theta_1 : \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right)$$

Statistical divergences corresponding to BDs wrt cumulant functions F of EFs

$$D^s_{B,\alpha}(p:q) = \begin{cases} -\frac{1}{\alpha(1-\alpha)} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\}, \\ D_{\mathrm{KL}}(p:q), & \alpha = 1, \\ 4\, D_B(p,q) & \alpha = \frac{1}{2}, \\ {D_{\mathrm{KL}}}^*(p:q) = D_{\mathrm{KL}}(q:p) & \alpha = 0. \end{cases} \Longleftrightarrow J^s_{F,\alpha}(\theta_1:\theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1:\theta_2), & \alpha \in \mathbb{R}\backslash\{0,1\}, \\ B_F(\theta_1:\theta_2), & \alpha = 0, \\ 4\, J_F(\theta_1,\theta_2), & \alpha = \frac{1}{2}, \\ B^*_F(\theta_1:\theta_2) = B_F(\theta_2:\theta_1), & \alpha = 1. \end{cases}$$

Scaled Bhattacharyya/Rényi distances $\Longleftrightarrow$ **Scaled skewed Jensen divergence for cumulant function F**

Proposition 4 ([32]). *The scaled α-skewed Bhattacharyya distances between two probability densities p_{θ_1} and p_{θ_2} of an exponential family amounts to the scaled α-skewed Jensen divergence between their natural parameters:*

$$D^s_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J^s_{F,\alpha}(\theta_1, \theta_2). \tag{13}$$

Statistical divergences corresponding to BDs wrt partition functions Z of EFs

$$D_\alpha(\tilde{p}:\tilde{q}) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \int \left(\alpha\tilde{p} + (1-\alpha)\tilde{q} - \tilde{p}^\alpha \tilde{q}^{1-\alpha}\right) \mathrm{d}\mu, & \alpha \notin \{0,1\} \\ D^*_{\mathrm{KL}}(\tilde{p}:\tilde{q}) = D_{\mathrm{KL}}(\tilde{q}:\tilde{p}) & \alpha = 0, \\ 4\, D^2_H(\tilde{p},\tilde{q}) & \alpha = \frac{1}{2}, \\ D_{\mathrm{KL}}(\tilde{p}:\tilde{q}) & \alpha = 1. \end{cases}$$

$$\Leftrightarrow \quad \mathrm{J}^s_{Z,\alpha}(\theta_1:\theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{Z,\alpha}(\theta_1:\theta_2), & \alpha \in \backslash\{0,1\}, \\ B_Z(\theta_1:\theta_2), & \alpha = 0, \\ 4\, J_Z(\theta_1,\theta_2), & \alpha = \frac{1}{2}, \\ B^*_Z(\theta_1:\theta_2) = B_Z(\theta_2:\theta_1), & \alpha = 1. \end{cases}$$

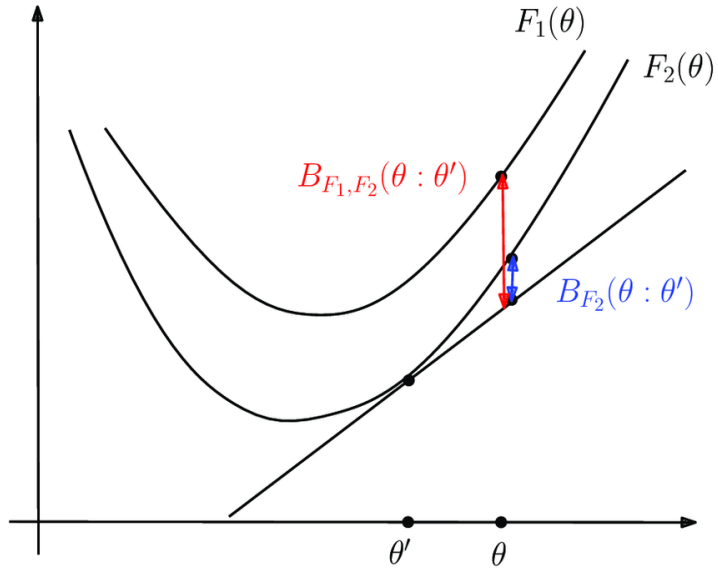
Amari α-divergences $\Leftrightarrow$ **Scaled skewed Jensen divergence for partition function Z**

Proposition 5. *The α-divergences between unnormalized densities of an exponential family amounts to scaled α-Jensen divergences between their natural parameters for the partition function:*

$$D_\alpha(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) = J^s_{Z,\alpha}(\theta_1:\theta_2).$$

KLD between normalized and unnormalized densities

$$
\begin{aligned}
D_{\mathrm{KL}}(p_{\theta_1} : \tilde{p}_{\theta_2}) &= B_F(\theta_2 : \theta_1) - \log Z(\theta_2) + Z(\theta_2) - 1, \\
&= Z(\theta_2) - 1 - F(\theta_1) - \langle \theta_2 - \theta_1, \nabla F(\theta_2) \rangle, \\
&= B_{Z-1,F}(\theta_2 - \theta_1). \qquad \text{with} \quad Z(\theta) - 1 \geq F(\theta)
\end{aligned}
$$



With generalized BDs to **duo Bregman pseudo-divergences**:

$$B_{F_1,F_2}(\theta_1 : \theta_2) = F_1(\theta_1) - F_2(\theta_2) - \langle \theta_1 - \theta_2, \nabla F_2(\theta_2) \rangle$$

with $F_1(\theta) = Z(\theta) - 1$ *and* $F_2(\theta) = F(\theta)$

Kullback-Leibler divergence between two **truncated densities** of a same exponential family amount to a duo Bregman pseudo-divergence

Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences, Entropy 24.3 (2022)

Comparative convexity: (M,N)-convexity

Ordinary convexity of a function: $f\left(tx_1 + (1-t)x_2\right) \leq tf\left(x_1\right) + (1-t)f\left(x_2\right)$
for all t in [0,1]

- *Definition*: A function Z is **(M,N)-convex** iff for in α in [0,1]:

$$Z(M(x,y;\alpha,1-\alpha)) \leq N(Z(x),Z(y);\alpha,1-\alpha)$$

- Ordinary convexity: (A,A)-convexity wrt to arithmetic weighted mean

$$A(x,y;\alpha,1-\alpha) = \alpha x + (1-\alpha)y \qquad f\left(tx_1 + (1-t)x_2\right) \leq tf\left(x_1\right) + (1-t)f\left(x_2\right)$$
for all t in [0,1]

- **Log-convexity**: **(A,G)-convexity** wrt to A/geometric weighted means:

$$G(x,y;\alpha,1-\alpha) = x^{\alpha}y^{1-\alpha} \qquad f(tx_1 + (1-t)x_2) \leq f(x_1)^t f(x_2)^{1-t}$$
for all t in [0,1]

Comparative convexity wrt quasi-arithmetic means

- Kolmogorov-Nagumo-De Finitti **quasi-arithmetic mean** for a strictly monotone generator h(u):

$$M_h(x,y;\alpha,1-\alpha) = h^{-1}(\alpha h(x) + (1-\alpha)h(x)).$$

- Includes **power means** which are *homogeneous means*:

$$M_p(x,y;\alpha,1-\alpha) = (\alpha x^p + (1-\alpha)y^p)^{\frac{1}{p}} = M_{h_p}(x,y;\alpha,1-\alpha), \quad p \neq 0$$

$$h_p(u) = \frac{u^p - 1}{p} \qquad h_p^{-1}(u) = (1+up)^{\frac{1}{p}}$$

Include the **geometric mean** when p→0

Proposition 6 ([1, 34]). *A function $Z(\theta)$ is strictly (M_ρ, M_τ)-convex with respect to two strictly increasing smooth functions ρ and τ if and only if the function $F = \tau \circ Z \circ \rho^{-1}$ is strictly convex.*

Generalizing Bregman divergences with (M,N)-convexity

- Skew Jensen divergence from (M,N) comparative convexity:

Definition:

$$J_{F,\alpha}^{M,N}(p:q) = N_\alpha(F(p),F(q)) - F(M_\alpha(p,q)).$$

Non-negative for **(M,N)-convex generators** F, provided regular means M and N (e.g. power means)

Definition 5 (Bregman Comparative Convexity Divergence, BCCD) *The Bregman Comparative Convexity Divergence (BCCD) is defined for a strictly (M,N)-convex function $F : I \to \mathbb{R}$ by*

$$B_F^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} \left(N_\alpha(F(p),F(q))) - F(M_\alpha(p,q))\right) \quad (31)$$

By analogy to limit of skewed Jensen divergences amount to forward/reverse Bregman divergences.

Generalizing Bregman divergences with quasi-arithmetic mean convexity

Theorem 1 (Quasi-arithmetic Bregman divergences, QABD) *Let $F : I \subset \mathbb{R} \to \mathbb{R}$ be a real-valued (M_ρ, M_τ)-convex function defined on an interval I for two strictly monotone and differentiable functions ρ and τ. The quasi-arithmetic Bregman divergence (QABD) induced by the comparative convexity is:*

$$B_F^{\rho,\tau}(p:q) = \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p) - \rho(q)}{\rho'(q)} F'(q). \quad (45)$$

Amounts to a **conformal Bregman divergence on monotonic representations**:

$$B_F^{\rho,\tau}(p:q) = \frac{1}{\tau'(F(q))} B_G(\rho(p):\rho(q))$$

With generator:

$$G(x) = \tau(F(\rho^{-1}(x)))$$

Conformal factor

Remark: Conformal Bregman divergences may yield **robustness** in applications

Shape retrieval using hierarchical total Bregman soft clustering, IEEE Transactions on pattern analysis and machine intelligence (2012)

Maximal invariant: Definitions and Eaton theorem

- Function f(x) is **invariant** under **group action (G,∘)** when f(g ∘x)=f(x)

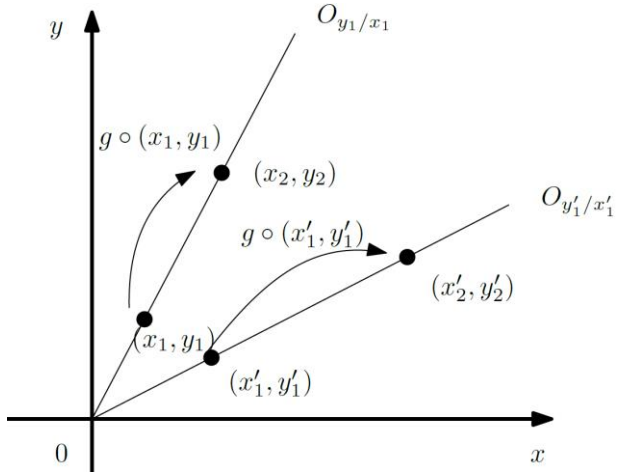
- A **maximal invariant** is a function h such that all orbits O_x={h(g ∘x) | g G} for g in G have **distinct values**

- Theorem [Eaton]:
  Any invariant function is a function of a maximal invariant

Eaton, Morris L. "Group invariance applications in statistics." IMS, 1989.

Maximal invariance: A toy example

- Consider function **slope(x,y)=y/x** and **group of positive reals G=(R_+,*,1)**
- Invariant under rescaling: slope(s ∘(x,y))=slope(s*x,s*y)=y/x
- **Slope = maximal invariant**: group orbits have different values (=slopes)



- Itakura-Saito divergence is a distance used in sound processing which is also invariant to rescaling: $\mathrm{D}_\psi(y\|x) = \frac{y}{x} - \log\frac{y}{x} - 1$
- **Therefore** Itakura-Saito divergence can be expressed as a function of max invariant slope: **D(y:x)=h(slope(x,y))** with h(u)=u-log u-1

Maximal invariance: f-divergences between scale models

- Compute **relative entropy** (Kullback-Leibler divergence), Hellinger, chi^2, etc between location-scale (elliptical) distributions

$$I_f(p:q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x,$$

- Consider **scale families** of probabilities: $\left\{ p_s(x) = \frac{1}{s} p\left(\frac{x}{s}\right), \quad x \in \mathbb{R}, s \in \mathbb{R}_{>0} \right\}$

Centered normal distributions
Cauchy distributions
t-Students

- Those f-divergences are all **invariant** under rescaling:

$$I_f(s \circ (p_{s_1} : p_{s_2})) = I_f((p_{s_1} : p_{s_2}))$$

- Thus, we express them using the **maximal invariant**:

$$I_f((p_{s_1} : p_{s_2})) = h_{p,f}(\mathrm{slope}(s_1, s_2)) = h_{p,f}\left(\frac{s_2}{s_1}\right)$$

- Calculate $h_{p,f}$ symbolic regression from data obtained by Monte Carlo integrations of the f-divergences. **Useful in practice**!

N and Okamura, "On f-divergences between Cauchy distributions, " IEEE Transactions on Information Theory 69.5 (2022)

Computing with maximal invariants

- The Mahalanobis distance and **Jensen-Shannon divergence** (JSD) between two isotropic Gaussian distributions are **invariant** under translation: that is, the translation = action of group G=(R,+,0)

- Mahalanobis distance is **maximal invariant** for the group action (after dD to 1d reparameterization…)

- Thus, we have $\mathrm{JSD}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = h(\Delta_\Sigma(\mu_1, \mu_2))$ for a strictly monotone function h

- JSD between Gaussians does not admit closed-form. Costly numerical approximations in practice!

- This **formula structure result** allows to compare exactly JSDs:

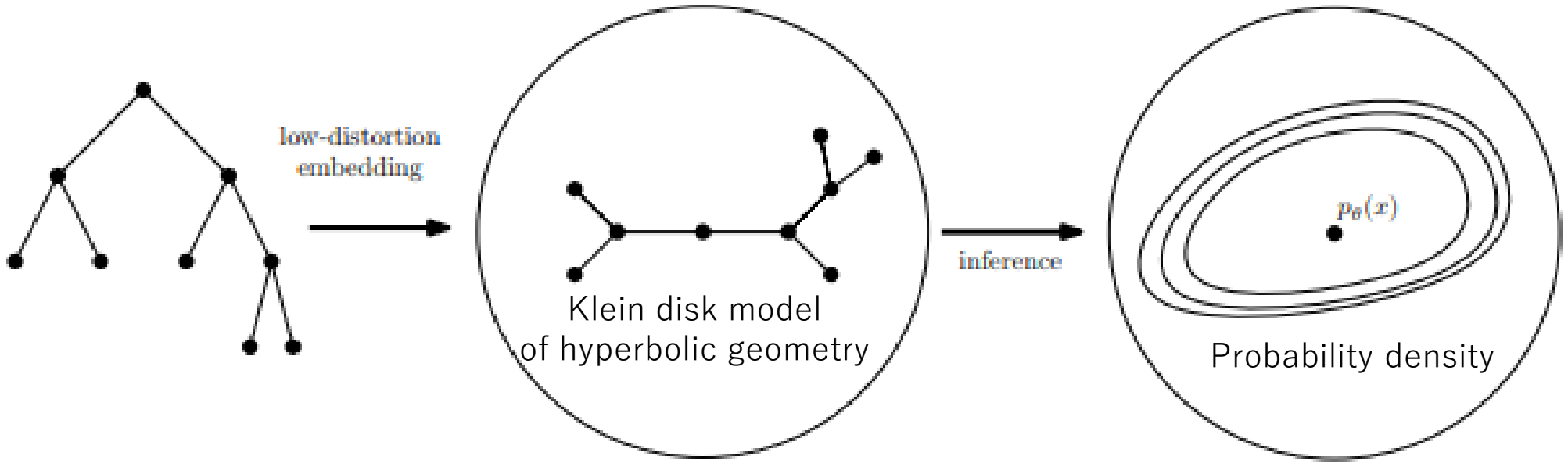
$$\mathrm{JSD}(N1,N2) < \mathrm{JSD}(N3,N4) \Leftrightarrow \Delta_\Sigma(\mu_1, \mu_2) < \Delta_\Sigma(\mu_3, \mu_4)$$

Since Mahalanobis distance Δ_Σ can be calculated exactly!

N and Okamura, "On the f-divergences between densities of a multivariate location or scale family", Statistics and Computing 34.1 (2024)

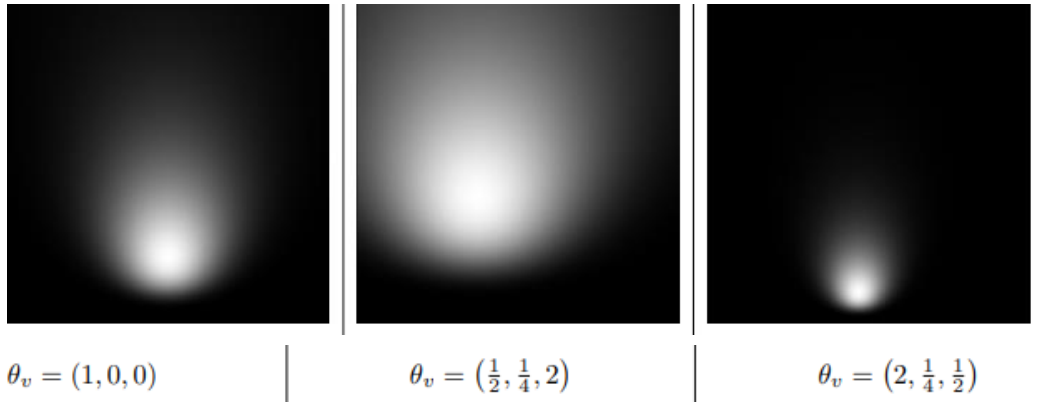
Maximal invariance and hyperbolic ML

- Hot topic in ML: **Discrete hierarchical graph structures** can be embedded in **continuous hyperbolic spaces** for downstream tasks
- We need **statistical analysis in hyperbolic spaces**
- **How to define probability distributions in hyperbolic spaces? and statistical distances between them?**



N and Okamura, "On the f-divergences between hyperboloid and Poincaré distributions," GSI (2023)

Poincaré distributions in the upper plane



Poincaré upper plane

$$p_\theta(x,y) := \frac{\sqrt{ac-b^2}\exp(2\sqrt{ac-b^2})}{\pi}\exp\left(-\frac{a(x^2+y^2)+2bx+c}{y}\right)\frac{1}{y^2}$$

Theorem 1. *Every f-divergence between two Poincaré distributions p_θ and $p_{\theta'}$ is a function of $\left(|\theta|, |\theta'|, \mathrm{tr}\left(\theta'\theta^{-1}\right)\right)$.* Triplet = Maximal invariant

(i) (Kullback-Leibler divergence) Let $f(u) = -\log u$. Then,

$$D_f\left[p_\theta : p_{\theta'}\right] = \frac{1}{2}\log\frac{|\theta|}{|\theta'|} + 2\left(\sqrt{|\theta|}-\sqrt{|\theta'|}\right) + \left(\frac{1}{2}+\sqrt{|\theta|}\right)\left(\mathrm{tr}(\theta'\theta^{-1})-2\right).$$

N and Okamura, "On the f-divergences between hyperboloid and Poincaré distributions," GSI (2023)
Tojo and Yoshino, Harmonic exponential families on homogeneous spaces, Information Geometry (2021)

f-divergences from higher-order chi divergence series

Theorem *Let X be a topological space and μ be a Borel measure on X with full support. Let $\{p_\theta(x)\}_\theta$ be a family of probability density functions on (X,μ). Assume that for each θ, $p_\theta(x)$ is positive and continuous with respect to x. We also assume that for each θ_1 and θ_2 there exists $C = C(\theta_1,\theta_2)$ such that $p_{\theta_1}(x) \leq C p_{\theta_2}(x)$ for every $x \in X$. Let $f(z) = \sum_{n=1}^{\infty} a_n (z-1)^n$ be an analytic function ($f \in C^\omega$), and denote by r_f be the convergence radius of f. Assume that $r_f \geq 1$. Let I_f be the induced f-divergence. Then,*

If $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} < 1 + r_f$ for every x, then,

$$I_f(p,q) = \int p f(q/p) \mathrm{d}\mu$$

$$I_f(p_{\theta_1} : p_{\theta_2}) = \sum_{n=2}^{\infty} a_n \int_X \left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} - 1\right)^n p_{\theta_1}(x) \mathrm{d}\mu(x) = \sum_{n=2}^{\infty} a_n D_{\chi,n}(p_{\theta_1} : p_{\theta_2}).$$

higher-order chi divergence: $$D_{\chi,k}(p:q) = \int \frac{(q(x)-p(x))^k}{p(x)^{k-1}} \mathrm{d}\mu(x).$$

On the chi square and higher-order chi distances for approximating f-divergences.
***IEEE Signal Processing Letters* 21.1 (2013) [2101.12459]**

Summary: Three concepts for distances/divergences

- Divergences are everywhere in information sciences!

- Fisher-Rao distances: lower bounds by submanifold embeddings.
  For MV normals: fast Hilbert ① projective distance using only extreme eigenvalues

- ② Maximal invariants for f-divergences and distances

- Bregman divergences:
  - canonical divergences of dually flat manifolds,
  - duo Bregman pseudo-divergences for calculating KLD between truncated EF densities
  - Generarized BDs from ③ comparative convexity: conformal Bregman divergences

Some references covering concepts in this talk

- Comparative convexity:
  - *Divergences Induced by the Cumulant and Partition Functions of Exponential Families and Their Deformations Induced by Comparative Convexity*, Entropy 26.3 (2024): 193
  - *Generalizing skew Jensen divergences and Bregman divergences with comparative convexity*, IEEE Signal Processing Letters 24.8 (2017): 1123-1127

- Maximal invariant:
  - *On the f-divergences between densities of a multivariate location or scale family,* Statistics and Computing 34.1 (2024): 60
  - *On f-divergences between Cauchy distributions*, IEEE Transactions on Information Theory 69.5 (2022): 3150-3171

- Projective distance:
  - *Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures*, ICML TAG 2023.
  - *Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices*, Geometric structures of information (2019): 297-331
  - *On Hölder projective divergences*, Entropy 19.3 (2017): 122

Thank you!