Geometric structures for statistical models in ML:

~ An overview with some recent results for ML ~

Frank Nielsen

Sony Computer Science Laboratories, Inc.

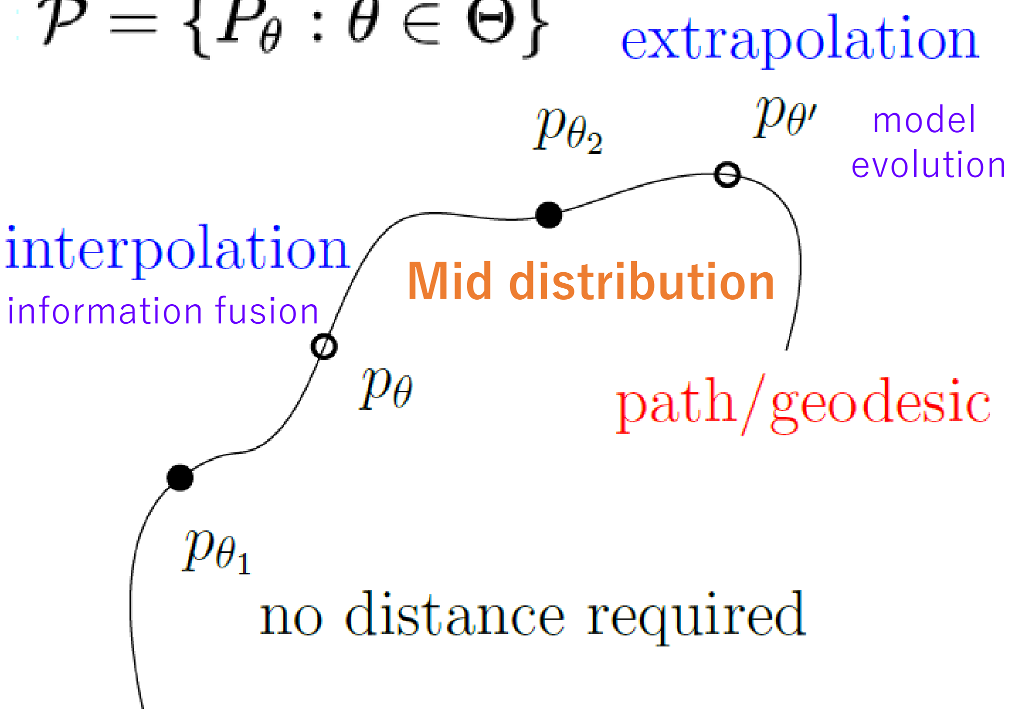
OIST ML Workshop

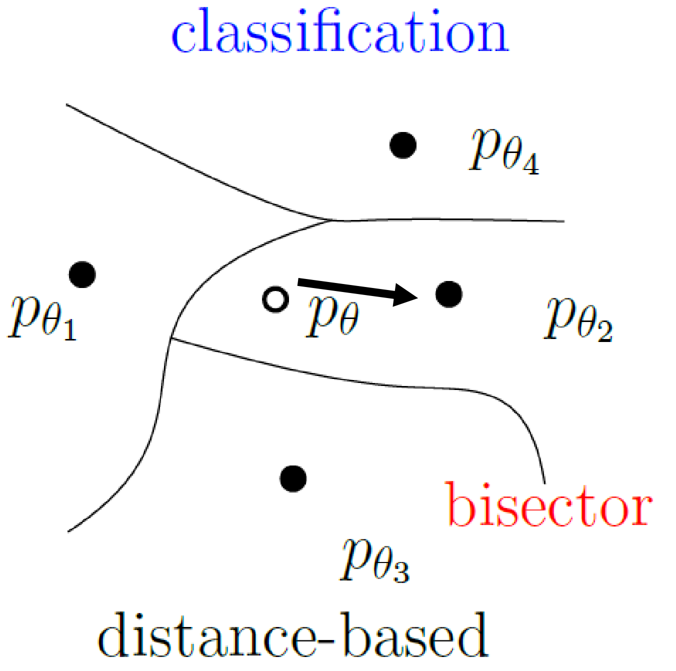
3rd March 2025

Sony CSL

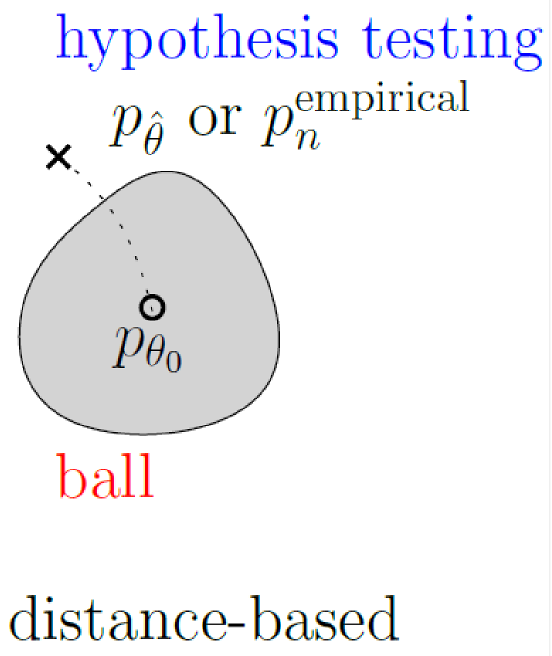
Which geometric structures for statistical models?

$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$





Voronoi diagrams



Information projection

Some benefits of the geometric approach:

① foster **geometric intuition & creativity**!

② leverage most advanced geometric calculus with **coordinate-free tensors**

③ may get **exact geometric characterization when non-closed algebraic formula**

④ obtain **new pure geometry** for mathematics: dual statistical structures

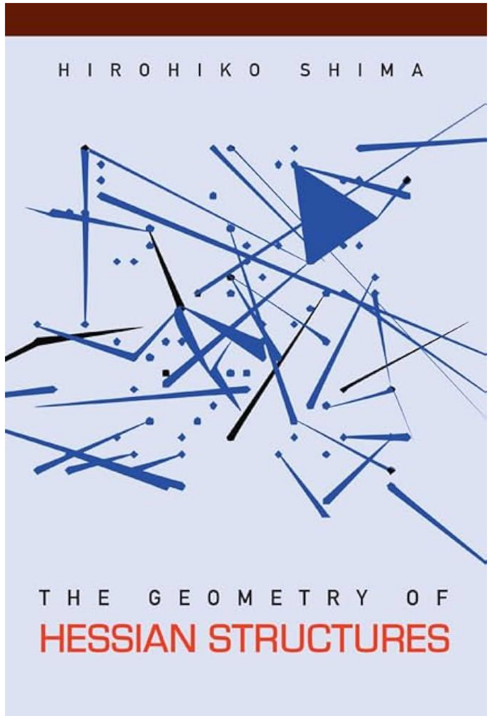
Geometric structures of statistical models & uses?

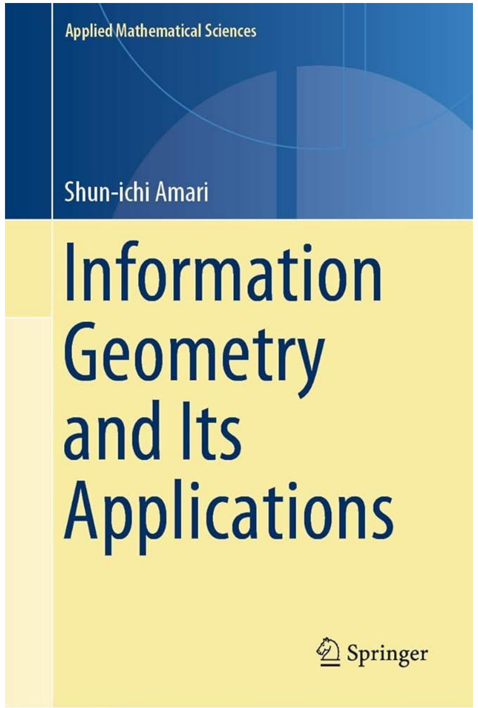
Geometry of domains Θ/manifolds

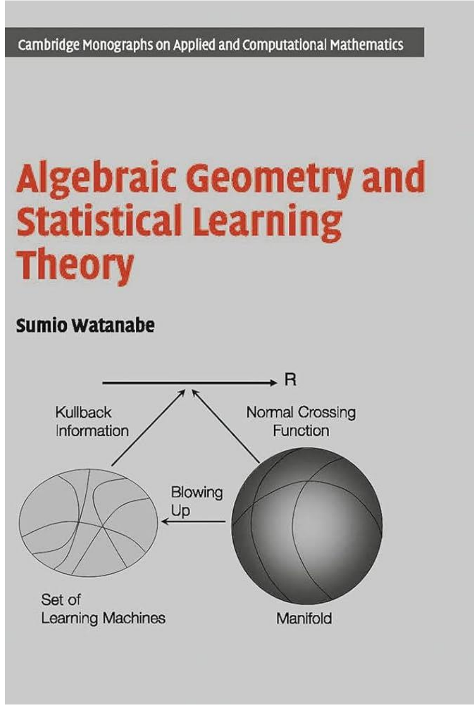
Geometry of regular statistical models statistical manifolds

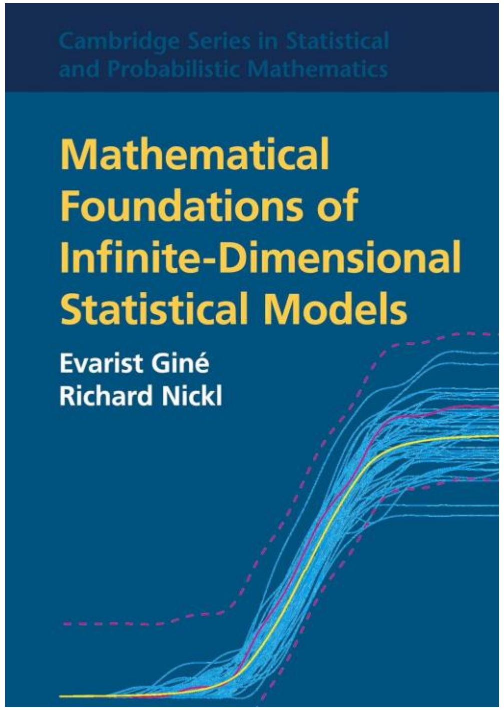
Geometry of singular hierarchical models (mixtures, DNNs)

Non-parametric statistical models









and many more!!!

Geometry of convex functions

Dual α-geometry of statistical models/divergences

Algebraic geometry resolving singularities

Function spaces (approximations)

- Model identifiability: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ $\quad P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ $\quad$ for all $\theta_1, \theta_2 \in \Theta$.

Outline

Overview of *three geometric structures* for the statistical models of <u>normal distributions</u> and <u>categorical distributions</u> with applications:

1. **Fisher-Rao manifolds** & numerical Fisher-Rao Gaussian distances

2. **Hilbert geometry** & fast distances between multivariate Gaussians

3. **Bregman manifolds** and some usages for statistical models:
   The *family of categorical distributions* view either as:
   - A **mixture family manifold** : Jensen-Shannon centroid
   - An **exponential family manifold**: Chernoff information/Chernoff point

In the beginning…

[Hotelling 1930, Rao 1945]

Fisher-Rao manifolds

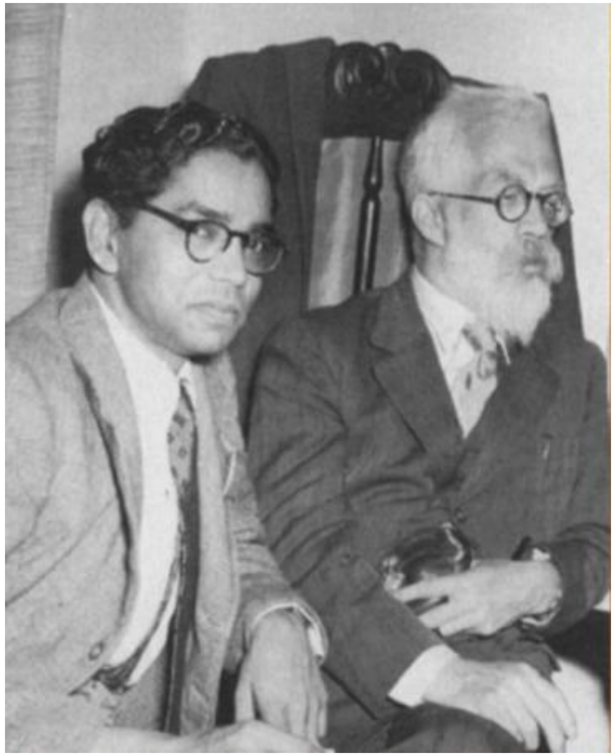
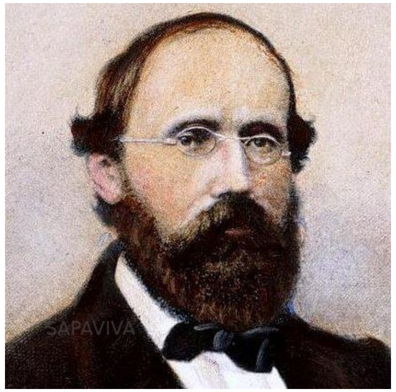


Photo 1956

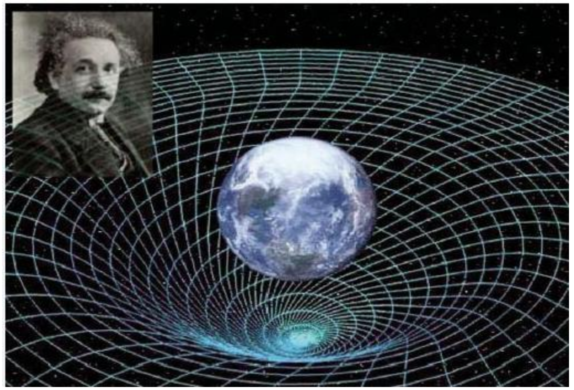
Riemannian geometry

Length element
ds

1854



``Killer''
application



1915, GR

Fisher-Rao manifolds

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

- **Manifold viewpoint**: Parameter space Θ interpreted as a global coordinate chart of a manifold M (vs general case in geometry)

$$X = (X_1, \ldots, X_m) \sim p_\theta$$

- **Fisher metric**: Consider **Fisher information matrix** of statistical models as metric tensor* field g in θ-coordinate system

$$I(\theta) = [I_{ij}(\theta)], \quad I_{ij}(\theta) = \mathrm{Cov}(X_i, X_j) = E_\theta\left[\frac{\partial}{\partial_{\theta_i}} \log p_\theta(x) \, \frac{\partial}{\partial_{\theta_j}} \log p_\theta(x)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_\theta(x)\right]$$

- Metric tensor g(.,.) provides a way to measure vector lengths, angle between vectors (orthogonality)
- **Length element** is **independent of parameterization**: $\eta(\theta) \Leftrightarrow \theta(\eta)$

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

- **Fisher-Rao distance** is length* of shortest path (= Riemannian distance): integrate length element along Riemannian geodesic

Metric distance satisfying the triangular inequality

Fisher-Rao distances: Curvatures in Statistics

- Riemannian modeling ($M=P_\Theta, g_{Fisher}$) allows to define various *curvature* notions: **sectional curvatures** (4D Riemann-Christoffel curvature tensor, Ricci curvature tensor, Ricci scalar tensors, etc.)
- *Example 1* : Family of **categorical distributions** with m+1 choices

$$\rho(p,q) = 2\arccos\left(\sum_i \sqrt{p_i q_i}\right)$$

Positive sectional curvatures: spherical geometry

- *Example 2* : Family of **univariate normal distributions** (m=2)

$$\rho_{\mathcal{N}}(N(\mu_1,\sigma_1^2), N(\mu_2,\sigma_2^2)) = \sqrt{2}\log\left(\frac{1+\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2)}{1-\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2)}\right)$$

$$\Delta(a,b;c,d) = \sqrt{\frac{(c-a)^2+2(d-b)^2}{(c-a)^2+2(d+b)^2}}$$

Negative sectional curvatures: hyperbolic geometry

- *Example 3* : **location families*** $I(\theta) = \lambda I_{m\times m}$

m-dim location parameter l $\rho(l_1,l_2) = \lambda\|l_1-l_2\|_2$

Zero sectional curvatures: Euclidean geometry

Problem: Tractability of Fisher-Rao geodesics/distances

- Need (1) to **solve** geodesic ODE equation
  and (2) **integrate** length element along the geodesic

$$\forall k \in \{1,\ldots,m\}, \quad \ddot{\gamma}_k + \sum_{i,j} \Gamma_{ij}^k \, \dot{\gamma}_i \dot{\gamma}_j = 0, \quad \dot{\gamma}(t) = \frac{d}{\mathrm{d}t}\gamma(t), \ddot{\gamma}(t) = \frac{d^2}{\mathrm{d}t^2}\gamma(t)$$

Christoffel symbols:
derived from metric tensor g

$$\Gamma_{ij}^k(\theta) = \frac{1}{2}\left(\partial_j g_{ik}(\theta) + \partial_i g_{jk}(\theta) - \partial_k g_{ij}(\theta)\right)$$

- Solve geodesic ODE either with **initial value conditions (IVC)** or with **boundary value conditions (BVC)**

$$\mathrm{IVC}: \quad \begin{cases} \ddot{\gamma}_k + \sum_{i,j} \Gamma_{ij}^k \, \dot{\gamma}_i \dot{\gamma}_j = 0 \\ \underline{\gamma(0), \dot{\gamma}(0) \in T_{\gamma(0)}} \end{cases} \qquad \mathrm{BVC}: \quad \begin{cases} \ddot{\gamma}_k + \sum_{i,j} \Gamma_{ij}^k \, \dot{\gamma}_i \dot{\gamma}_j = 0 \\ \underline{\gamma(0), \gamma(1)} \end{cases}$$

Tractability of Fisher-Rao distance: Yet the open case of the multivariate normal family!

$$I_{ij}(\theta) = \left(\frac{\partial \mu}{\partial \theta_i}\right)^\top \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial \mu}{\partial \theta_i}\Sigma^{-1}\frac{\partial \mu}{\partial \theta_j}\right)$$

Fisher length:

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

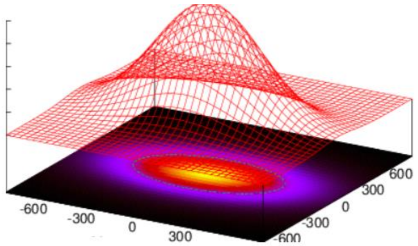
Geodesic ODE:

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

Solve ODE with initial values (IV) or boundary values (BV)

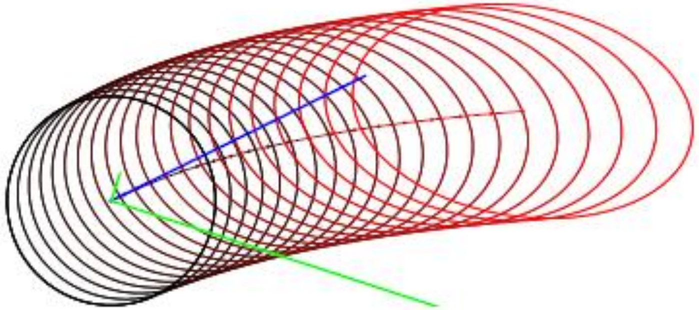
Non-constant sectional curvatures which can also be positive!
(geodesics are always unique when negative sectional curvatures)

Bivariate normal (represented by ellipsoids)



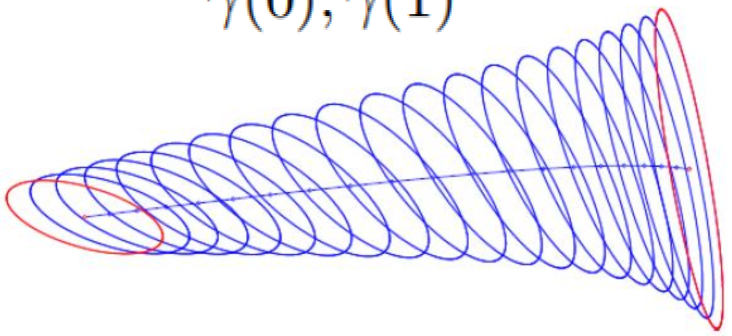
[IV: Eriksen 1987]

$\gamma(0), \dot{\gamma}(0) \in T_{\gamma(0)}$

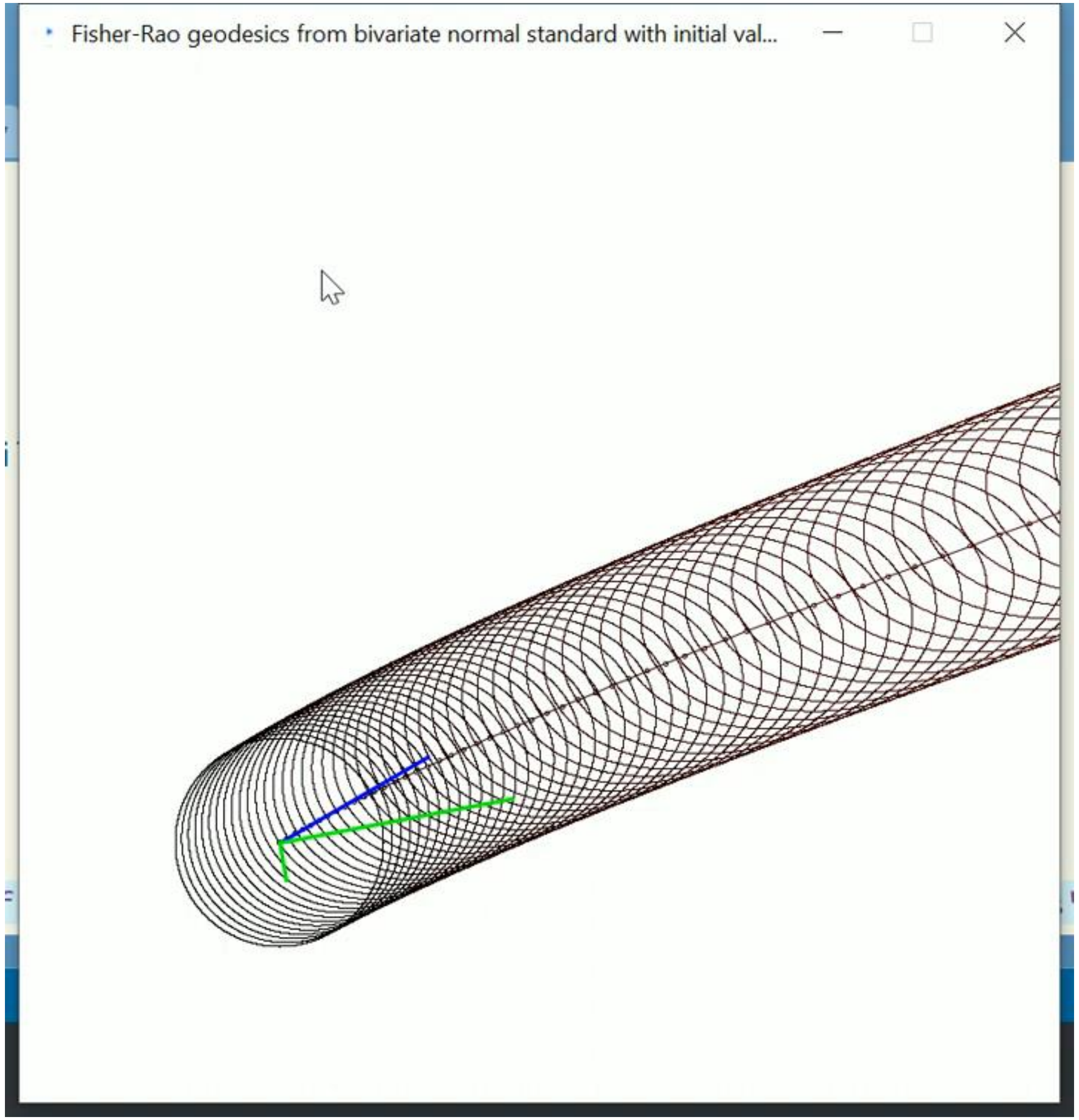


[BV: Kobayashi 2023]

$\gamma(0), \gamma(1)$



Fisher-Rao geodesics with initial values emanating from the standard bivariate Gaussian



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

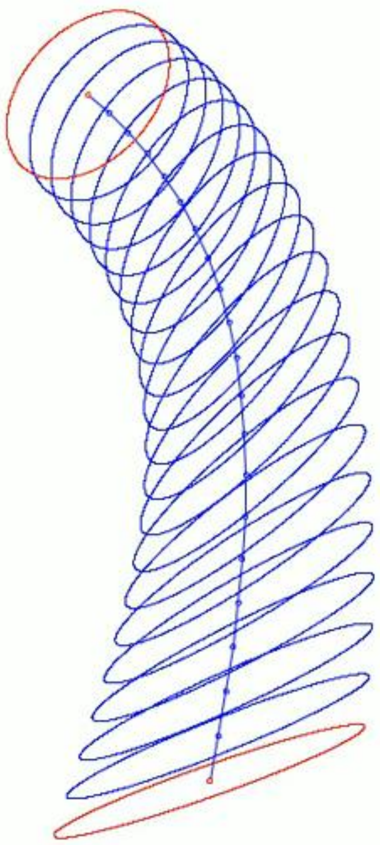
$$\gamma(0), \dot{\gamma}(0) \in T_{\gamma(0)}$$

Blue vector is initial tangent vector for μ_0

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

[IV: Eriksen 1987]

Fisher-Rao geodesics with boundary



$\gamma(0), \gamma(1)$

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

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

[BV: Kobayashi 2023]

Technically, MVN Fisher-Rao geodesic:
Riemannian submersion of a horizontal geodesic
of a Riemannian symmetric space in 2d+1 dimension

No known closed-form for Fisher-Rao between multivariate normal distributions

- May always consider distance to **standard normal distribution** because of the **invariance under action of the positive affine group**:

$$\mathrm{Aff}_+(d,\mathbb{R}) := \{(a, A) \; : \; a \in \mathbb{R}^d, A \in \mathrm{GL}_+(d,\mathbb{R})\}$$

$$\rho_{\mathcal{N}}(N(A\mu_1 + a, A\Sigma_1 A^\top), N(A\mu_2 + a, A\Sigma_2 A^\top)) = \rho_{\mathcal{N}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)).$$

Hence, we have

$$\begin{aligned}\rho_{\mathcal{N}}(N(\mu_1,\Sigma_1), N(\mu_2,\Sigma_2)) &= \rho_{\mathcal{N}}\left(\boxed{N_{\mathrm{std}}}, N\left(\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1), \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\right), \\ &= \rho_{\mathcal{N}}\left(N\left(\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2), \Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}\right), \boxed{N_{\mathrm{std}}}\right),\end{aligned}$$

- In general, hard to prove uniqueness of geodesics when some sectional curvatures are positive: The case of MVN Fisher-Rao manifold!!!

Special case: Centered multivariate normals Closed form geodesics and Fisher-Rao distances

- Submanifold of MVNs with constant mean is **totally geodesic**

- **Fisher-Rao geodesics**: $\gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t) = N(\mu, \Sigma_t)$

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

[James 1973]
[Siegel 1964]

- **Fisher-Rao distance**:

$$\rho_{\mathcal{N}_\mu}(N_0, N_1) = \sqrt{\frac{1}{2} \sum_{i=1}^{d} \log^2 \lambda_i\left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}\right)}$$

- Require to **compute all eigenvalues** (costly)
- Because of sum of $\log^2$, we have

$\rho(P_1,P_2) = \rho(P_1^{-1},P_2^{-1})$: **invariance to matrix inversion**

Embedding manifold Gaussian/MVN(d) onto SPD(d+1)

[Calvo & Oller 1990]

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 1/2 trace metric in P(d+1), get the following **metrics on MVN(d)**:

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

$\beta = 1$

When $\beta = 1$ (constant), get **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** but non-totally geodesic:

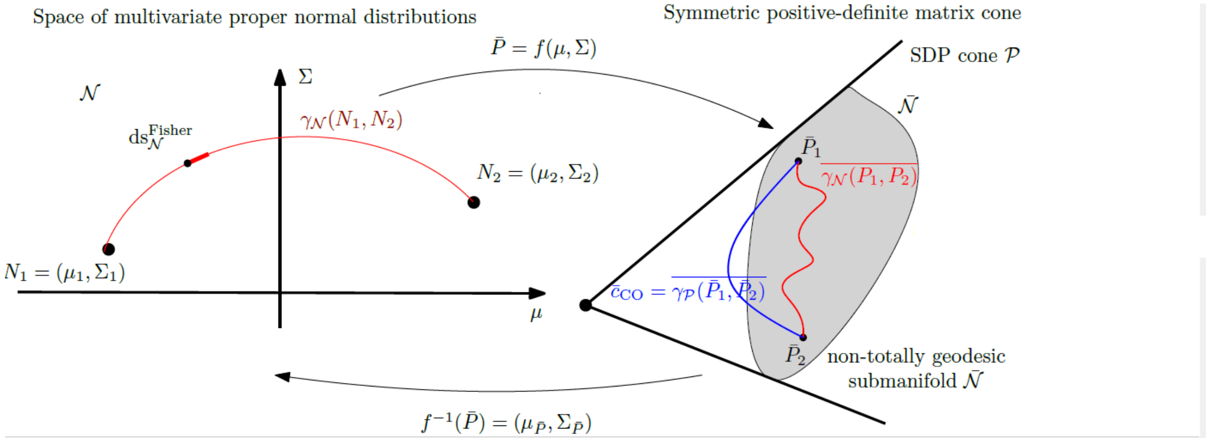
$$f(N)=f(\mu,\Sigma)=\begin{bmatrix}\Sigma+\mu\mu^\top & \mu\\ \mu^\top & 1\end{bmatrix}$$

[Calvo & Oller 1990]

- Use 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:

$$\boxed{\rho_{\mathcal{N}}(N_1,N_2)\geq\rho_{\mathrm{CO}}(\underbrace{f(\mu_1,\Sigma_1)}_{\bar P_1},\underbrace{f(\mu_2,\Sigma_2)}_{\bar P_2})=\sqrt{\frac{1}{2}\sum_{i=1}^{d+1}\log^2\lambda_i(\bar P_1^{-1}\bar P_2)}.}$$



- Cut MVN geodesics and apply lower bound piecewise : Get **fine lower bound!**

Fisher-Rao MVN distance: An upper bound

Property (Fisher–Rao upper bound). *The Fisher–Rao distance between normal distributions is upper bounded by the square root of the Jeffreys divergence:* $\boxed{\rho_{\mathcal{N}}(N_1, N_2) \leq \sqrt{D_J(N_1, N_2)}.}$

$$D_J[p,q] = \int (p-q)\log\frac{p}{q}\mathrm{d}\mu \qquad D_J[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] = \mathrm{tr}\left(\frac{\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2}{2} - I\right) + \Delta\mu^\top \frac{\Sigma_1^{-1}+\Sigma_2^{-1}}{2}\Delta\mu.$$

- Geodesics are **1d totally geodesic submanifolds**
- Cut the geodesics in many small parts using T+1 geodesic points

$$\tilde{\rho}^c_{\mathcal{N}}(N_1, N_2) := \frac{1}{T}\sum_{i=1}^{T-1}\sqrt{D_J\left[c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right]}. \qquad \boxed{D_J[p_\theta : p_{\theta+\mathrm{d}\theta}] = \mathrm{d}s^2(\theta, \mathrm{d}\theta)}$$

- **Upper bound** for nearby points Fisher-Rao distance by the square root of Jeffreys divergence
- Fine upper bound!

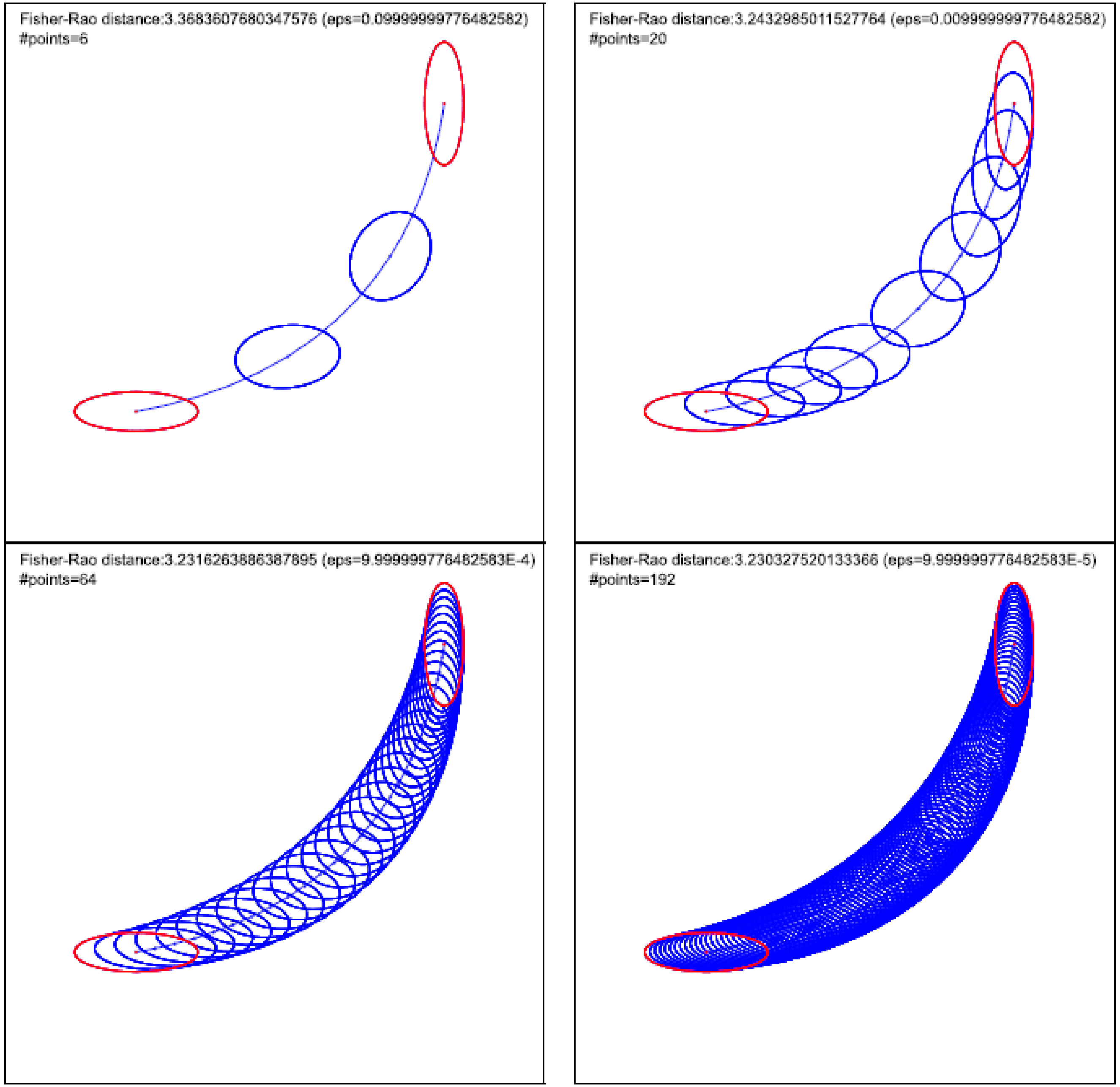
(1+ ε)-approximation of Fisher-Rao distance between multivariate normal distributions

ApproxRaoDistMVN(N0,N1, ε >0): // multiplicative factor **2403.10089**

```
LB=CalvoOllerLowerBound(N0,N1);
UB=SqrtJeffreysUpperBound(N0,N1);

if (UB/LB>1+ ε )
          {/* N is midpoint geodesic */
           N=GeodesicMVNMidpoint(N0,N1);
           return ApproxRaoDistMVN(N0,N, ε )+ApproxRaoDistMVN(N,N1, ε );}
    else
        return UB;
```

Then we can convert multiplicative approximation factor by an additive approximation factor



Precision $\boldsymbol{\varepsilon} = \mathbf{10^{-6}}$ with
192 geodesic discretization steps

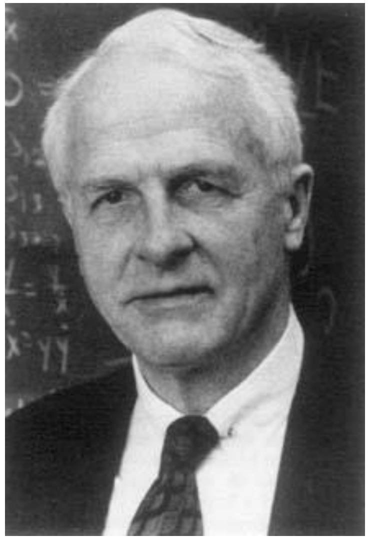
Implemented
in library **pyBregMan**
python **Breg**man **Man**ifold

Thus MVN Fisher-Rao distance can be finely approximated with guarantees but slow…

Other MVN fast distances?

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

Hilbert geometry
&
Birkhoff cone geometry

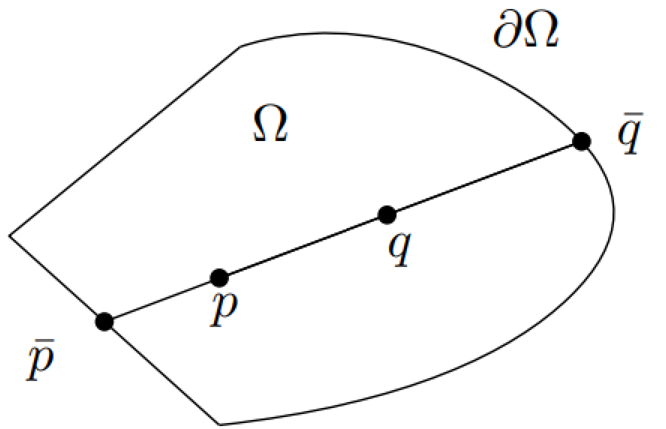


Projective/Finsler geometry of convex domains

2203.11434
ICML TAG-ML 2023

Hilbert distance: The log cross-ratio metric

Consider an **open bounded convex set** Ω



$$\rho^{\Omega}_{\mathrm{HG}}(p,q) = \begin{cases} \log \mathrm{CR}(\bar{p}, p; q, \bar{q}), & p \neq q \\ 0 & p = q \end{cases}$$

Cross-ratio $\mathrm{CR}(p, q; P, Q) = \frac{(p-P)(q-Q)}{(p-Q)(q-P)}$

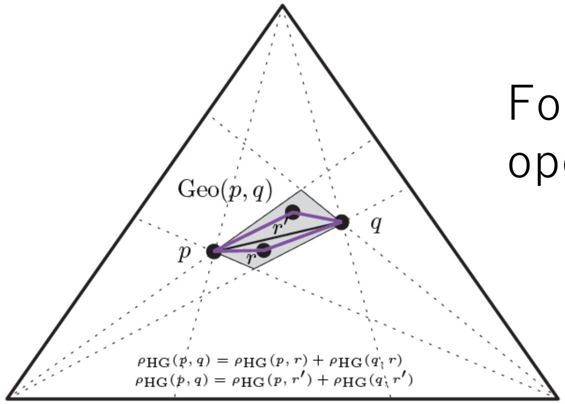
ρ^{Ω} is a **metric distance** which satisfies the **triangle inequality**:

$$\forall r \in [pq], \quad \rho^{\Omega}_{\mathrm{HG}}(p,q) = \rho^{\Omega}_{\mathrm{HG}}(p,r) + \rho^{\Omega}_{\mathrm{HG}}(r,q)$$

Straight lines are geodesics

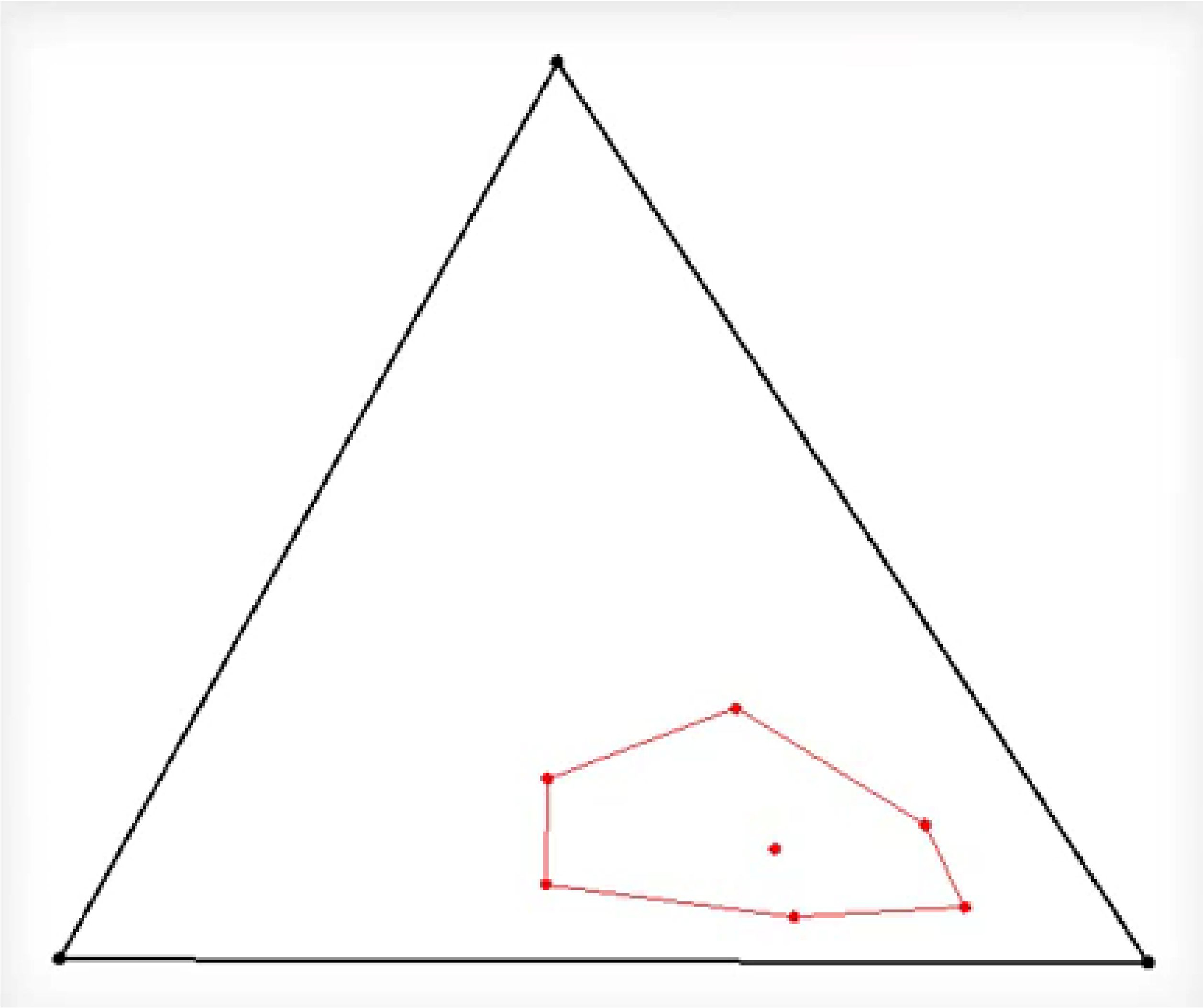
=satisfying triangle equality

but geodesics are **not unique**



For example:
open standard simplex

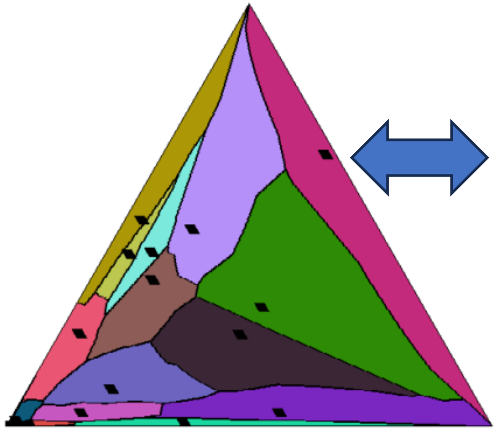
Hilbert geometry on the probability simplex: Balls have hexagonal Euclidean shapes



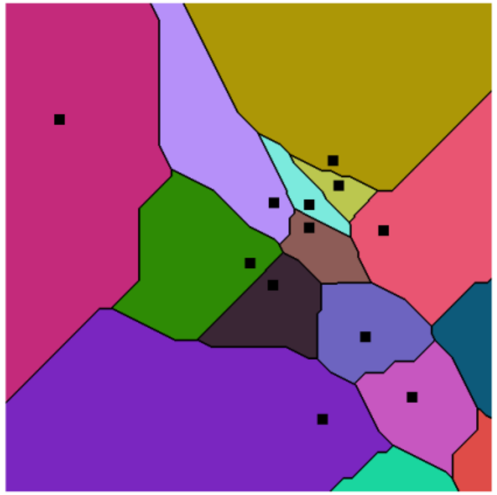
Fast to compute

$$\rho_{\mathrm{HG}}(p,q) = \log \frac{\max_{i\in\{1,\ldots,d\}} \frac{p_i}{q_i}}{\min_{i\in\{1,\ldots,d\}} \frac{p_i}{q_i}}.$$

Only when domain is a simplex,
Hilbert geometry amounts
to a **normed vector space**
with polyhedral norm (hexagonal metric)



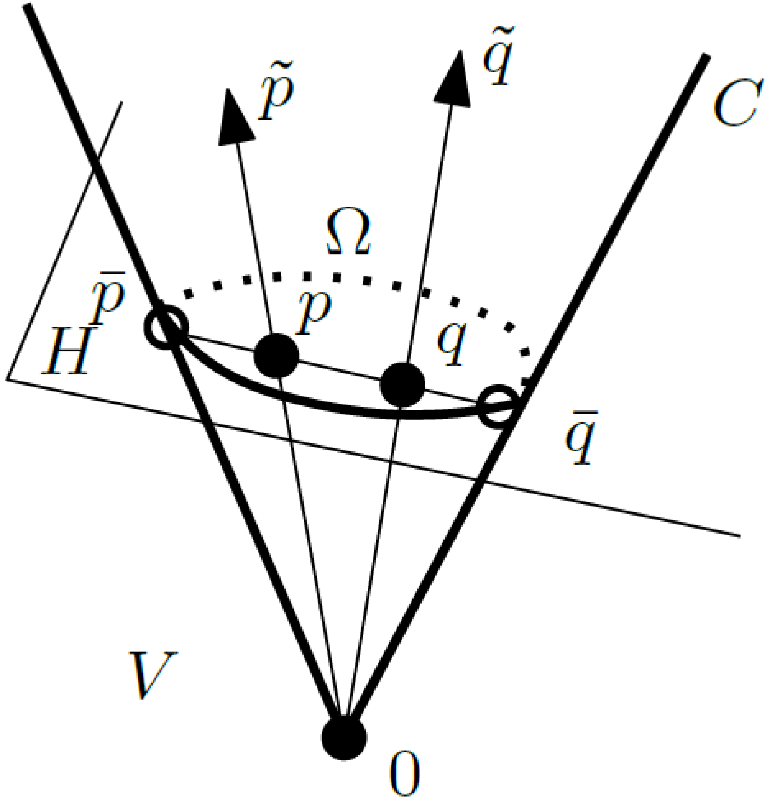
Hilbert simplex geometry



Poly. norm vector space

2203.11434

Birkhoff: Hilbert projective distance in a cone



Cone: $C = \{(\lambda, x) \ : \ \lambda \in \mathbb{R}_{>0}, x \in \Omega\}$

Cone defines a **partial ordering**:

$$p \preceq_C q \Leftrightarrow q - p \in C$$

Birkhoff distance: $\rho^C_{\mathrm{HG}}(p,q) = \log \dfrac{M(p,q)}{m(p,q)}$

where $M(p,q) = \inf\{\lambda \in \mathbb{R}_{>0} \ : \ p \preceq_C \lambda q\}$

$m(p,q) = \sup\{\lambda \in \mathbb{R}_{>0} \ : \ \lambda q \preceq_C p\}$

Birkhoff projective distance: $\rho^C_{\mathrm{HG}}(p,q) = \rho^C_{\mathrm{HG}}(\alpha p, \beta q), \quad \alpha, \beta > 0$

which becomes the **Hilbert metric distance** on Ω: $\rho^\Omega_{\mathrm{HG}}(p,q) = \rho^C_{\mathrm{HG}}(p,q), \quad \forall p, q \in \Omega$

New fast distances between multivariate normals

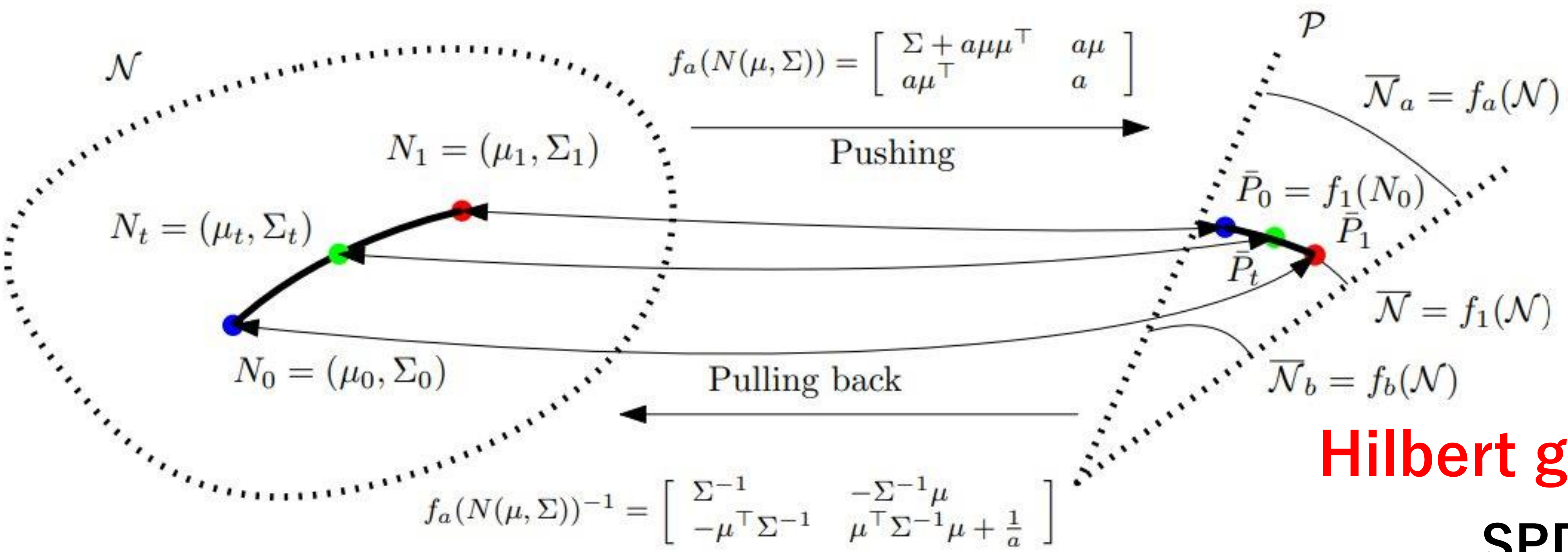
Use Calvo & Oller's diffeometric/isometric cone embeddings f(μ , Σ)

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

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

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

Gaussian(d) manifold



Hilbert geometry

SPD(d+1) cone

New fast distance between multivariate normals

- Use Calvo & Oller isometric cone embedding f(μ , Σ)

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

- In SPD cone, **Hilbert projective metric distance**

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

Projective metric on SPD

$\rho_{\mathrm{Hilbert}}(P_0, P_1) = 0$ if and only if $P_0 = \lambda P_1$

But proper metric on f(N)

LERP pregeodesics
Straight line edge!

$$\boxed{\gamma_{\mathrm{Hilbert}}(P_0, P_1; t) := \left(\frac{\beta\alpha^t - \alpha\beta^t}{\beta - \alpha}\right) P_0 + \left(\frac{\beta^t - \alpha^t}{\beta - \alpha}\right) P_1,}$$

$\alpha = \lambda_{\min}(P_1^{-1} P_0)$ and $\beta = \lambda_{\max}(P_1^{-1} P_0)$

- **Pullback** the geodesics and distance into the Gaussian manifold

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

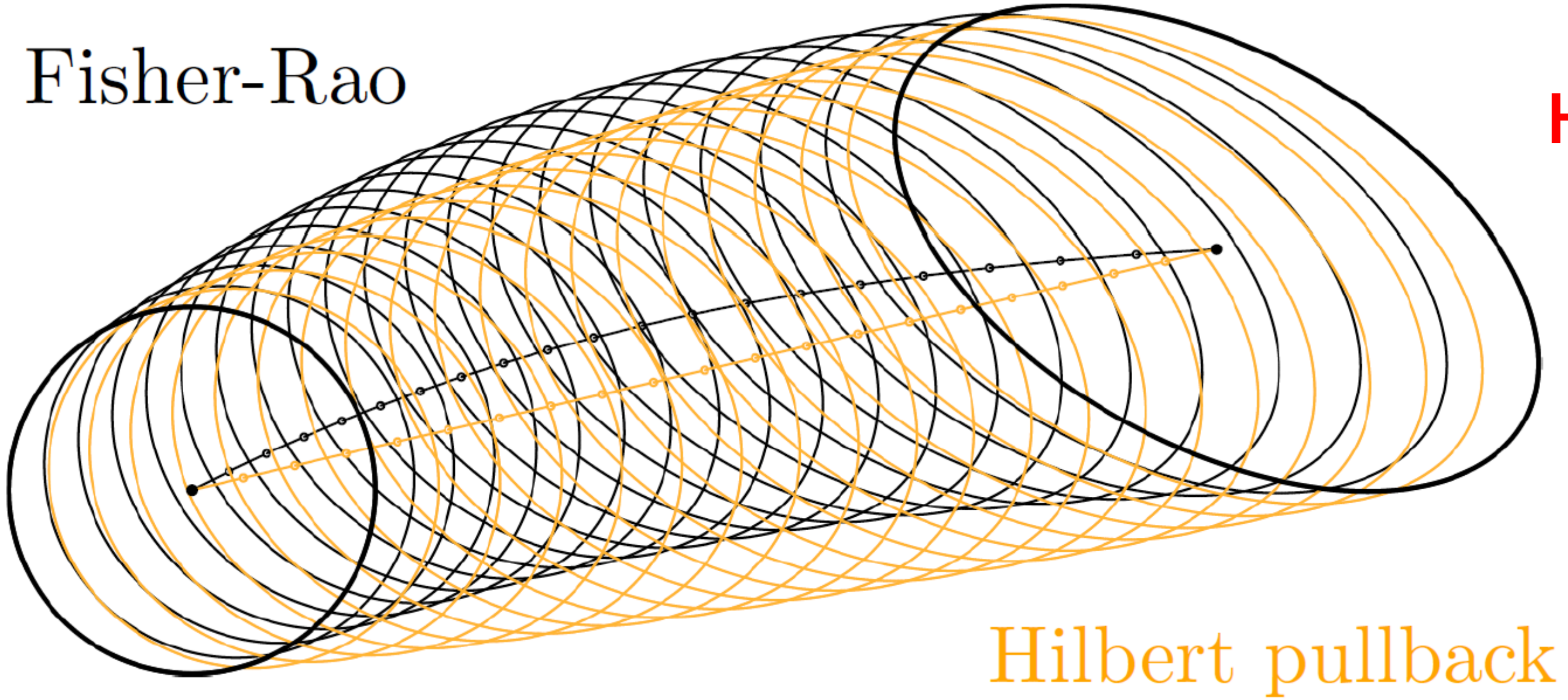
Pullback Hilbert distance/geodesics between MVNs

Only require to calculate 2 **extreme eigenvalues** (power method iteration)

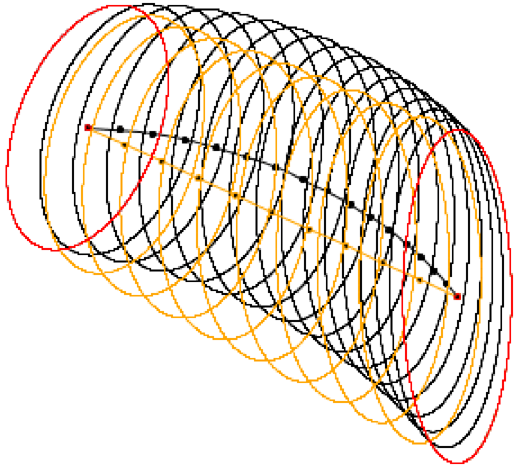
$$\rho_{\mathrm{Hilbert}}(P_0, P_1) = \log\left(\frac{\lambda_{\max}(P_0^{-1}P_1)}{\lambda_{\min}(P_0^{-1}P_1)}\right)$$

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

Hilbert-Fisher-Rao distance



Comparisons Fisher-Rao vs Fisher-Rao-Hilbert geodesics



Boundary conditions
Fast Fisher-Rao-Hilbert distance (extreme SPD matrix eigenvalues)
Slow guaranteed **Fisher-Rao distance**

Bregman manifolds:
Geometry of convex conjugates

Dual Hessian geometry

[Koszul'64, Shima'70's, Amari&Nagaoka'80's]

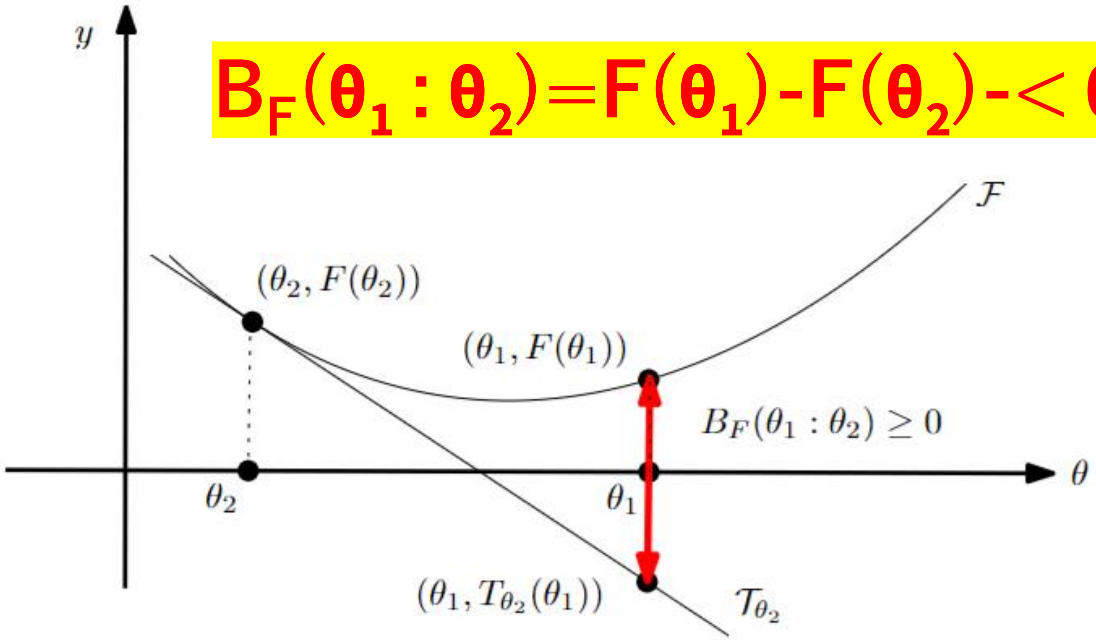
Bregman divergence (1960's)

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

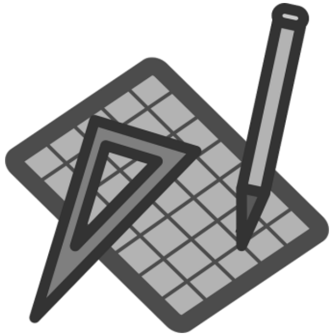
Bregman divergence B_F: $\Theta \times \mathrm{Int}(\Theta) \rightarrow \mathbb{R}$

Lev M. Bregman
(1941 - 2023)
Photo: courtesy of
Alexander Fradkov

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle$$



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



- Unify *squared Euclidean divergence* with *Kullback-Leibler divergence* $F(\theta) = \sum_i \theta_i \log(\theta_i)$ and *Itakura-Saito divergence* $F(\theta) = \sum_i -\log(\theta_i)$.
- The L22, KLD and ISD belong to a *single family* of **β-divergences**, learn β

Bregman divergences in machine learning…

- Kullback-Leibler divergence between two probability densities:

  $$D_{KL}[p(x):q(x)] = \int p(x) \log (p(x)/q(x))\, d\mu(x)$$

  difficult to calculate in closed form because of the integral ∫ …

- But the Kullback-Leibler divergence between two probability densities of an **exponential family** like Gaussian, Poisson, Dirichlet, Gamma/Beta, Wishart

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

$p(x|\theta) \propto \exp(\langle x, \theta \rangle)$

amount to a **reverse Bregman divergence** $B_F^{rev}(\theta_1 : \theta_2) := B_F(\theta_2 : \theta_1)$

$$D_{KL}[p(x|\theta_1) : p(x|\theta_2)] = B_F^{rev}(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$$

⇒ Easy calculations

Bypass the ∫, ∇F easy!

- Notice divergence between parameters B_F vs divergence between functions KL

Azoury, Katy S., and Manfred K. Warmuth. "Relative loss bounds for on-line density estimation with the exponential family of distributions." *Machine learning* 43 (2001)

Convex duality via Legendre-Fenchel transform

- Legendre-Fenchel transform of a convex function F:

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

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

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

Then we get:

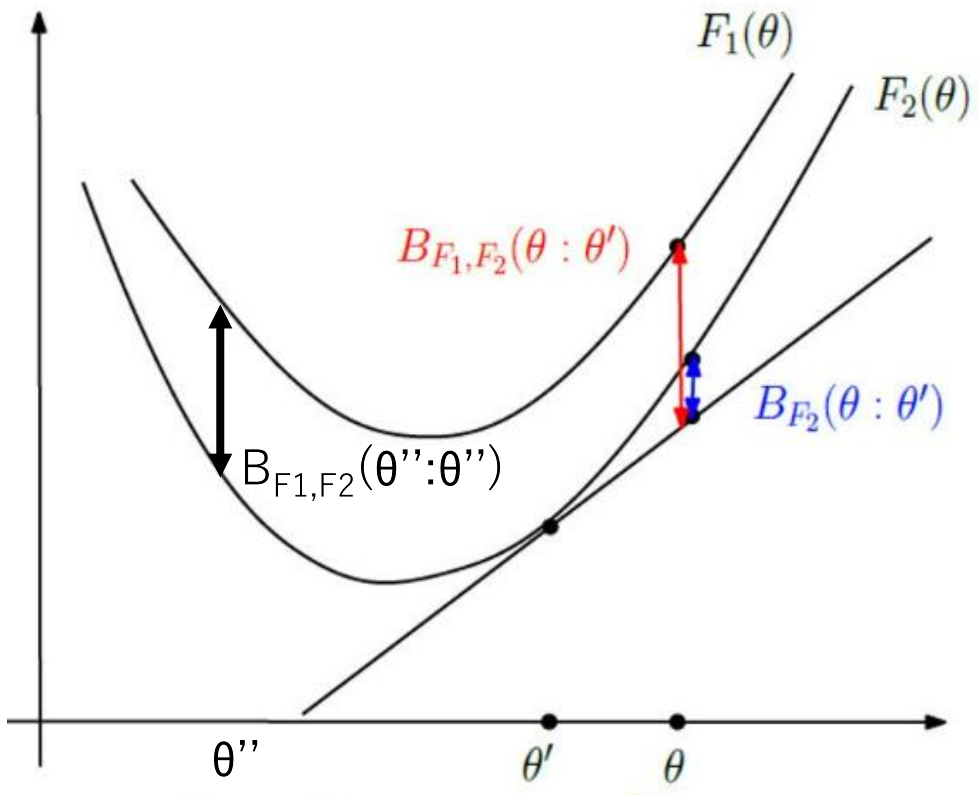
❶ **reciprocal gradient maps** $\eta=\nabla F(\theta)$ and $\theta=\nabla F^*(\eta)$, $\nabla F^*=(\nabla F)^{-1}$

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

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

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

Duo Bregman divergences:
Generalize BDs with a pair of generators



One generator majorizes the other one:

$$F_1(\theta) \geq F_2(\theta)$$

Then

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

$$\geq B_F(\theta:\theta')$$

- Recover Bregman divergence when $F_1(\theta) = F_2(\theta) = F(\theta)$

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

- Only **pseudo-divergence** because $B_{F1,F2}(\theta'':\theta'')$ positive, not zero

KLD between nested exponential families amount to duo Bregman pseudo divergences

$q(x|\theta)$» $p(x|\theta)$

$p(x|\theta)$ X_1

X_2

$q(x|\theta)$

$D_{KL}[p(x):q(x)] = \int p(x) \log (p(x)/q(x))\, d\mu(x)$

$0 \log(0/0)=0$

- Consider an exponential family on support X_1:

$$p(x|\theta) = \exp(\langle x, \theta\rangle - F_1(\theta))\, d\mu(x)$$

with cumulant function $F_1(\theta)=\log \int_{X1} \exp(\langle x, \theta\rangle)\, d\mu(x)$

- Another exponential family with **nested supports: $X_1 \subseteq X_2$**

$$q(x|\theta) = \exp(\langle x, \theta\rangle - F_2(\theta))\, d\mu(x)$$

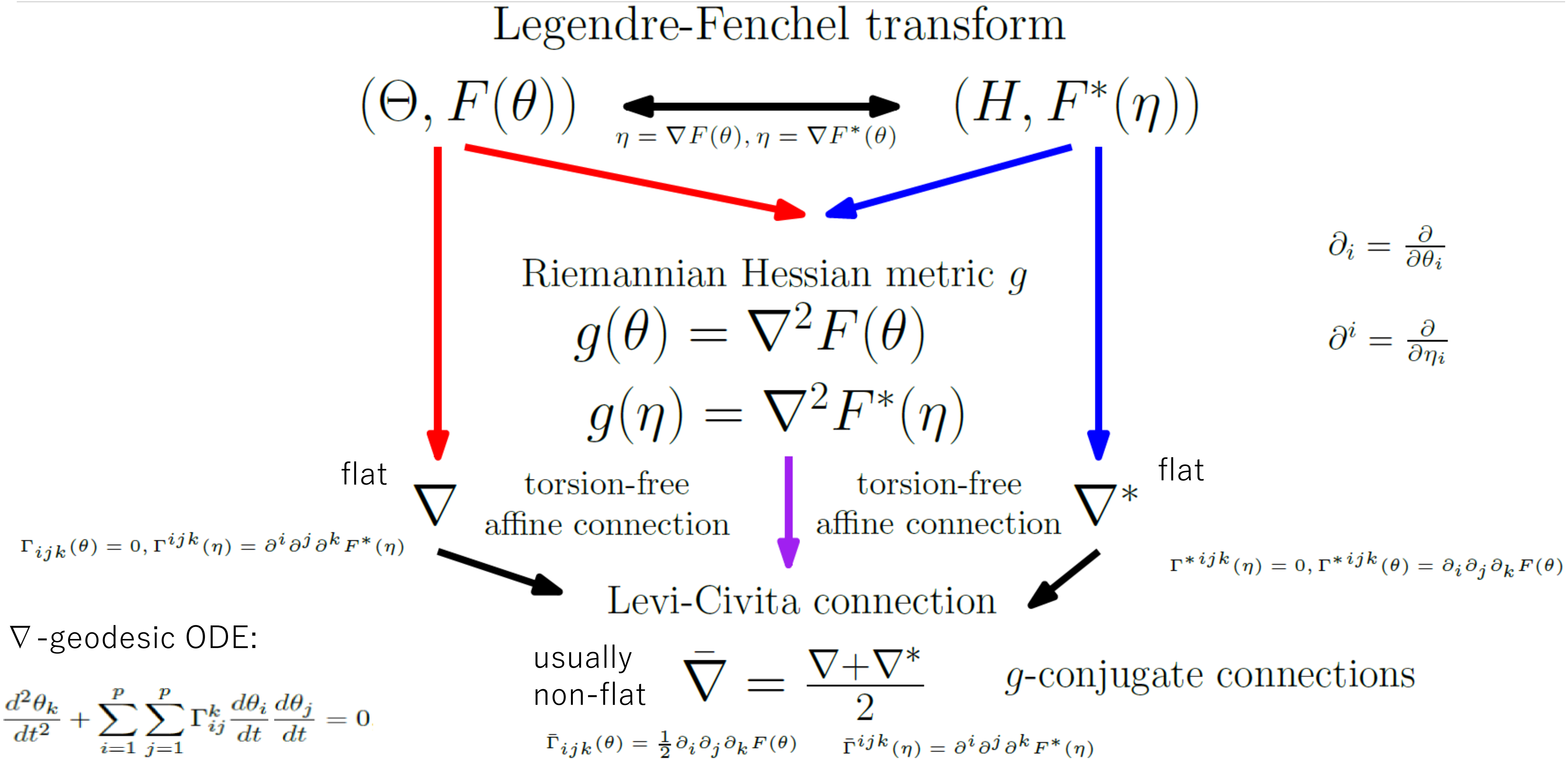
is an exponential family with $F_2(\theta)=\log \int_{X2} \exp(\langle x, \theta\rangle)\, d\mu(x) \geq F_1(\theta)$

- Then KLD amounts to a **reverse duo Bregman pseudo-divergence**:

$$\mathbf{D_{KL}[p(x|\theta_1) : q(x|\theta_2)] = B_{F2,F1}^{rev}(\theta_1 : \theta_2) = B_{F2,F1}(\theta_2 : \theta_1)}$$

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

Dual geometry of smooth Legendre-type functions



$$\partial_i = \frac{\partial}{\partial \theta_i}$$

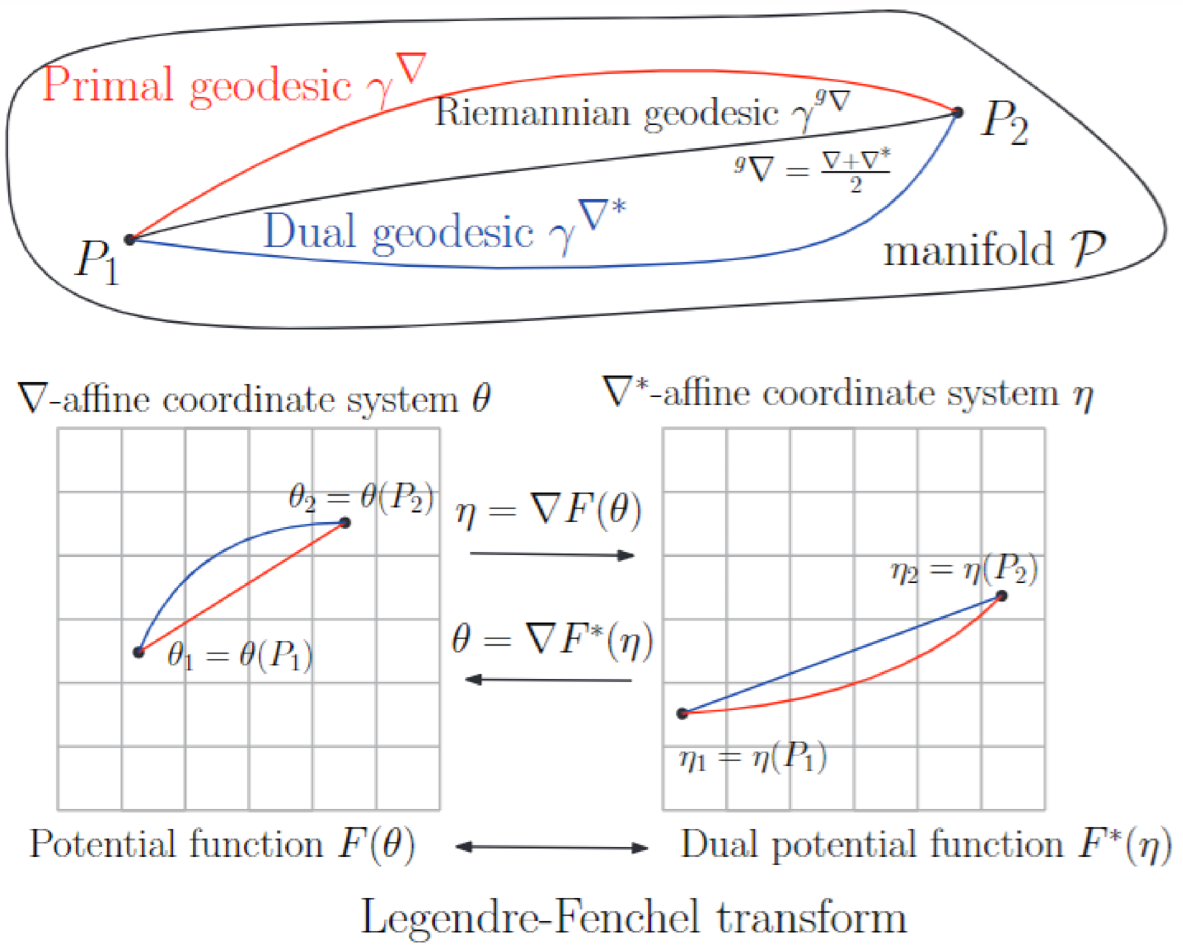
$$\partial^i = \frac{\partial}{\partial \eta_i}$$

∇-geodesic ODE:

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

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

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



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

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

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

Example: Bregman manifold of multivariate Gaussians

(M, g, ∇, ∇^*)

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

Cumulant function is convex:

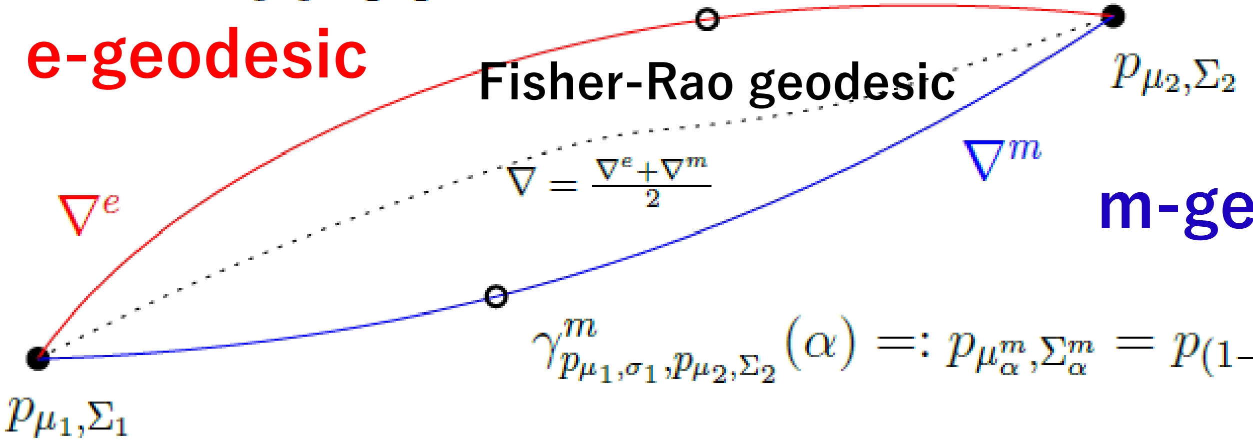
$$F_\theta(\theta) = \frac{1}{2}\left(d \log \pi - \log|\theta_M| + \frac{1}{2}\theta_v^\top \theta_M^{-1}\theta_v\right)$$

with respect to natural parameters:

$$\gamma^e_{p_{\mu_1,\sigma_1}, p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^e, \Sigma_\alpha^e} = p_{(1-\alpha)\theta_1 + \alpha\theta_2}$$

$$\theta = (\Sigma^{-1}\mu, \tfrac{1}{2}\Sigma^{-1})$$

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



…but not convex wrt (μ Σ) parameters

m-geodesic beware not mixture of Gaussians!

$$\gamma^m_{p_{\mu_1,\sigma_1}, p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^m, \Sigma_\alpha^m} = p_{(1-\alpha)\eta_1 + \alpha\eta_2}$$

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

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

Bregman divergence = reverse Kullback-Leibler divergence

$$\frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1)\right)$$

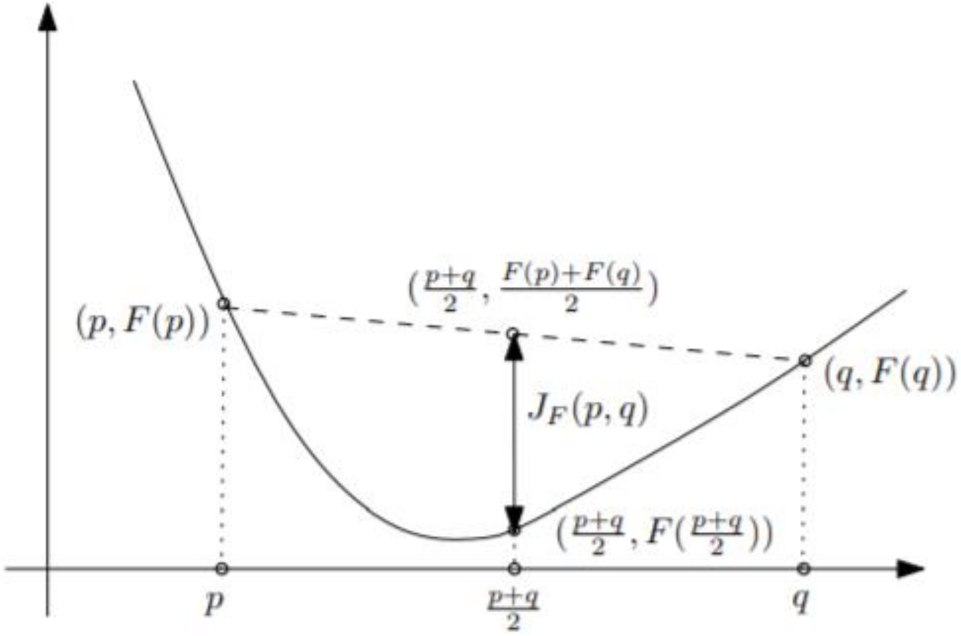
Jensen difference/Jensen divergence
(also called Burbea-Rao divergences)

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

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

n-point Jensen diversity index:

$$J_F(\theta_1, \ldots, \theta_n; w) = \sum_i w_i F(\theta_i) - F(\sum_i w_i \theta_i) \geq 0$$



skewed Jensen divergence $J_\alpha^F(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2)$

Asymptotic scaled skew Jensen divergences amount to forward/reverse Bregman divergences

The Burbea-Rao and Bhattacharyya centroids." *IEEE Transactions on Information Theory* (2011)
A family of statistical symmetric divergences based on Jensen's inequality, arXiv:1009.4004

Right Bregman centroid: Bregman/Jensen decomposition

Right Bregman centroid minimizes
$$\min_\theta \sum_i w_i B_F(\theta_i : \theta)$$

From **Bregman information-bias decomposition**

$$\min_\theta \underbrace{J_F(\theta_1, \ldots, \theta_n; w)}_{\text{independent of } \theta} + \underbrace{B_F(\bar\theta : \theta)}_{\geq 0 \text{ with equality iff } \theta = \bar\theta}$$

We get $\theta = \bar\theta$ wi $\bar\theta = \sum_i w_i \theta_i$. Right Bregman centroid is **unique**.

Furthermore, a **D-centroid** minimizing $\min_\theta \sum_i w_i D(\theta_i : \theta)$ at the center of mass of parameters is necessarily a Bregman divergence: exhaustive property

Banerjee and Wang: On the optimality of conditional expectation as a Bregman predictor
IEEE Transactions on Information Theory 51.7 (2005)

Bregman information-bias decomposition

The weighted average right Bregman divergence (BD)

$$I(\theta_1, \ldots, \theta_n; w : \underline{\theta}) := \sum_i w_i \, B_F(\theta_i : \underline{\theta})$$

decomposes into the sum of a **Bregman information** (aka Jensen diversity index) and a **bias divergence** term:

$$\underline{I(\theta_1, \ldots, \theta_n; w : \theta) = J_F(\theta_1, \ldots, \theta_n; w) + B_F(\bar{\theta} : \theta)}$$

where $\bar{\theta} = \sum_i w_i \theta_i$ is a *right Bregman centroid* and the Bregman information generalizing variance when BD is squared Euclidean distance is:

$$J_F(\theta_1, \ldots, \theta_n; w) := \sum_i w_i \, B_F(\theta_i : \bar{\theta}) = \left(\sum_i w_i F(\theta_i) \right) - F(\bar{\theta})$$

Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory* 55.6 (2009)

Jensen-Bregman divergences = Jensen div.

- **Jensen-Bregman divergence** is Jensen-Shannon symmetrization of Bregman divergence:

$$\begin{aligned}
\mathrm{JB}_F(\theta:\theta') &:= \frac{1}{2}\left(B_F\left(\theta:\frac{\theta+\theta'}{2}\right)+B_F\left(\theta':\frac{\theta+\theta'}{2}\right)\right)\\
&= \frac{F(\theta)+F(\theta')}{2}-F\left(\frac{\theta+\theta'}{2}\right)=: J_F(\theta:\theta')
\end{aligned}$$

amounts to a **Jensen divergence** also called **Burbea-Rao divergence**.

Skew Jensen-Bregman Voronoi diagrams." *Transactions Voronoi Diagrams and Delaunay Triangulation* (2011)
On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* 21.5 (2019)

Categorical model as a mixture family

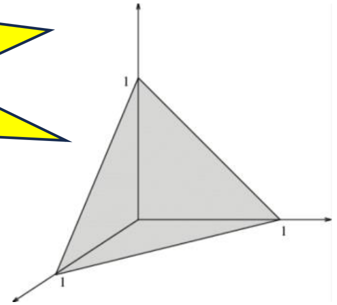
- Set of categorical distributions form a **mixture family M**, a **Bregman manifold for the negentropy generator**

$$\mathcal{M} = \left\{ m_\theta(x) = \sum_{i=1}^{D} \theta_i \delta(x - x_i) + \left(1 - \sum_{i=1}^{D} \theta_i\right) \delta(x - x_0) \right\}$$

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

Legendre Convex generator

A mixture family is closed under mixture operations



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

$$\mathrm{JS}(p,q) := \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right)$$

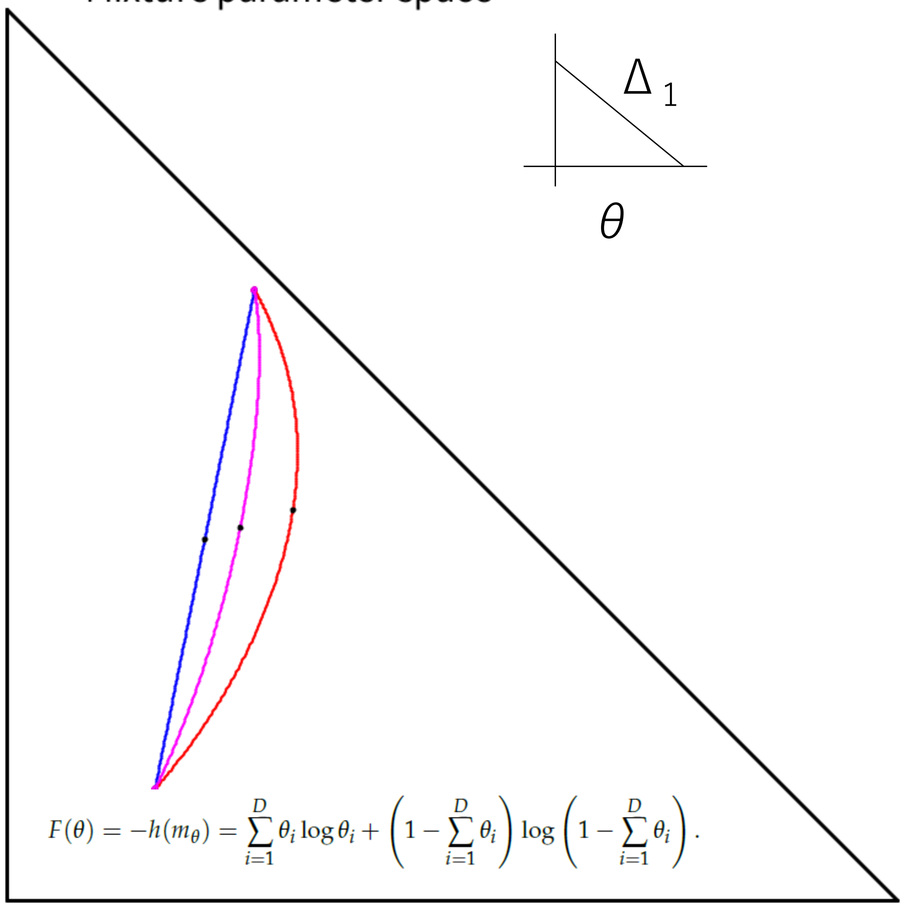
$$\mathrm{JS}(p,q) = h\left(\frac{p+q}{2}\right) - \frac{h(p)+h(q)}{2}$$

$$h(p) = -\int p \log p \, \mathrm{d}\mu$$

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

Dual geodesics and Fisher-Rao geodesics on the categorical manifold

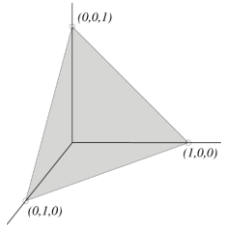
Mixture parameter space

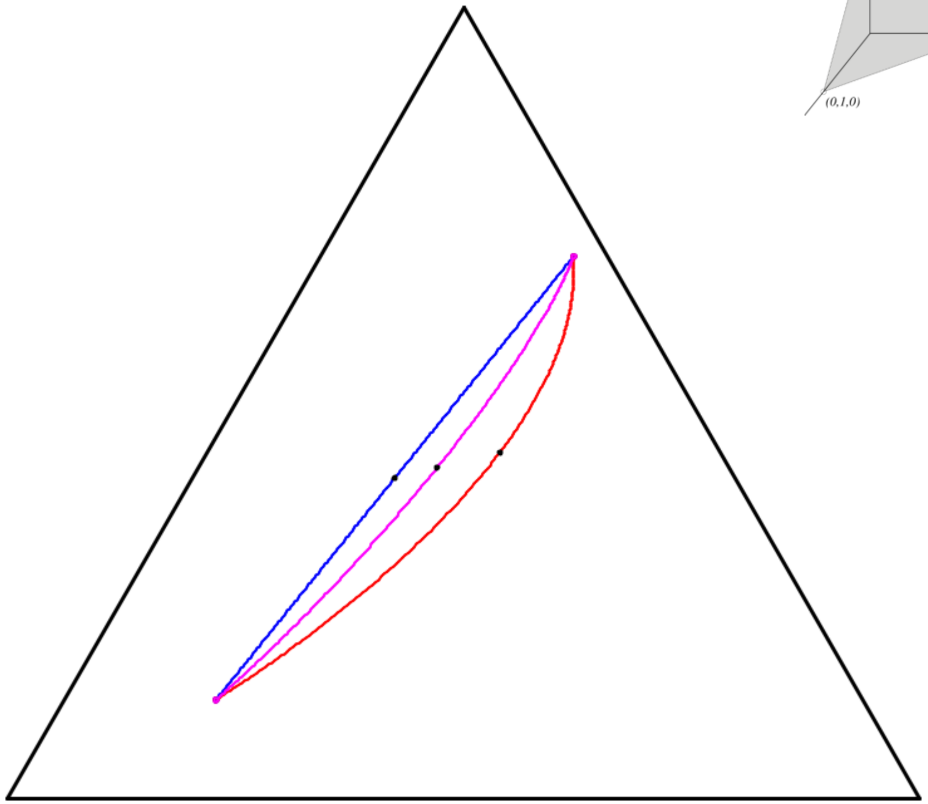


$$F(\theta) = -h(m_\theta) = \sum_{i=1}^{D} \theta_i \log \theta_i + \left(1 - \sum_{i=1}^{D} \theta_i\right) \log \left(1 - \sum_{i=1}^{D} \theta_i\right).$$

Coordinate chart

Probability simplex/Categorical manifold





Embedded manifold

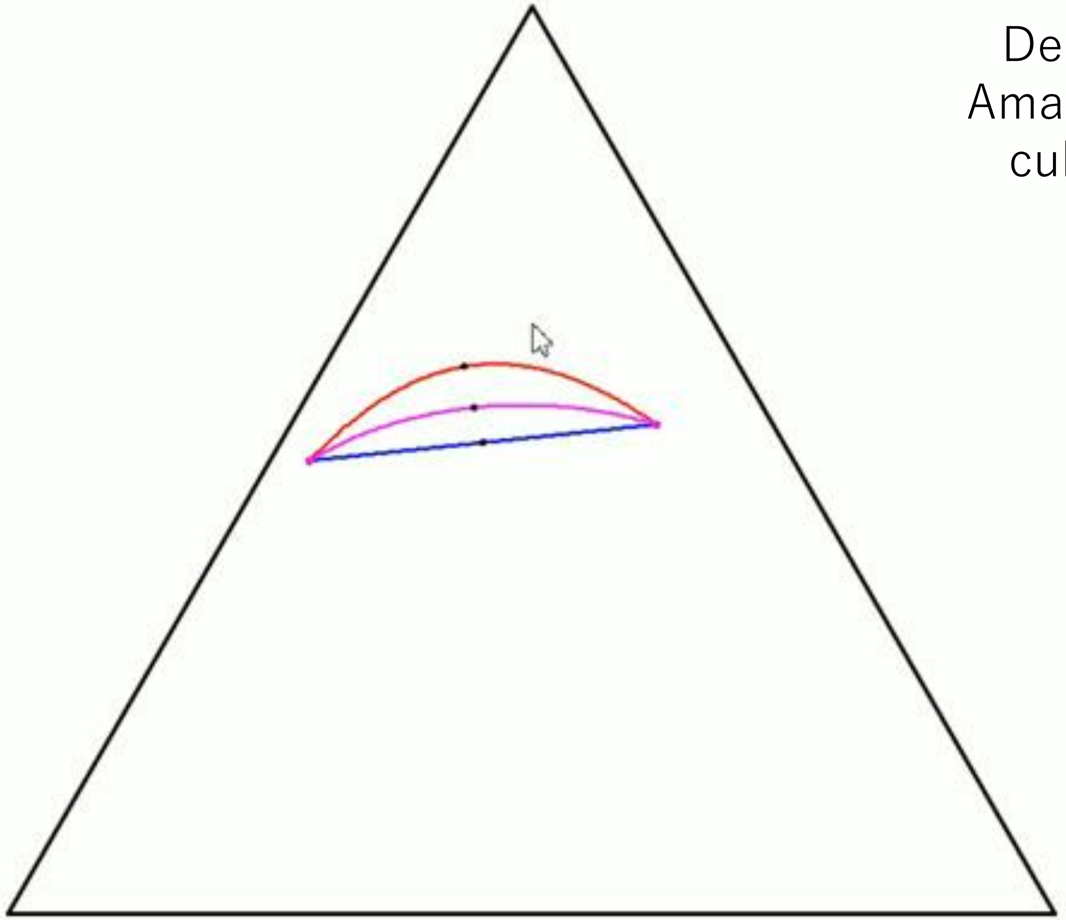
Exponential ∇-geodesic

Mixture ∇^*-geodesic

Fisher-Rao ∇^g-geodesic (Levi-Civita)

Amari-Nagaoka dual $\pm\alpha$-geometry

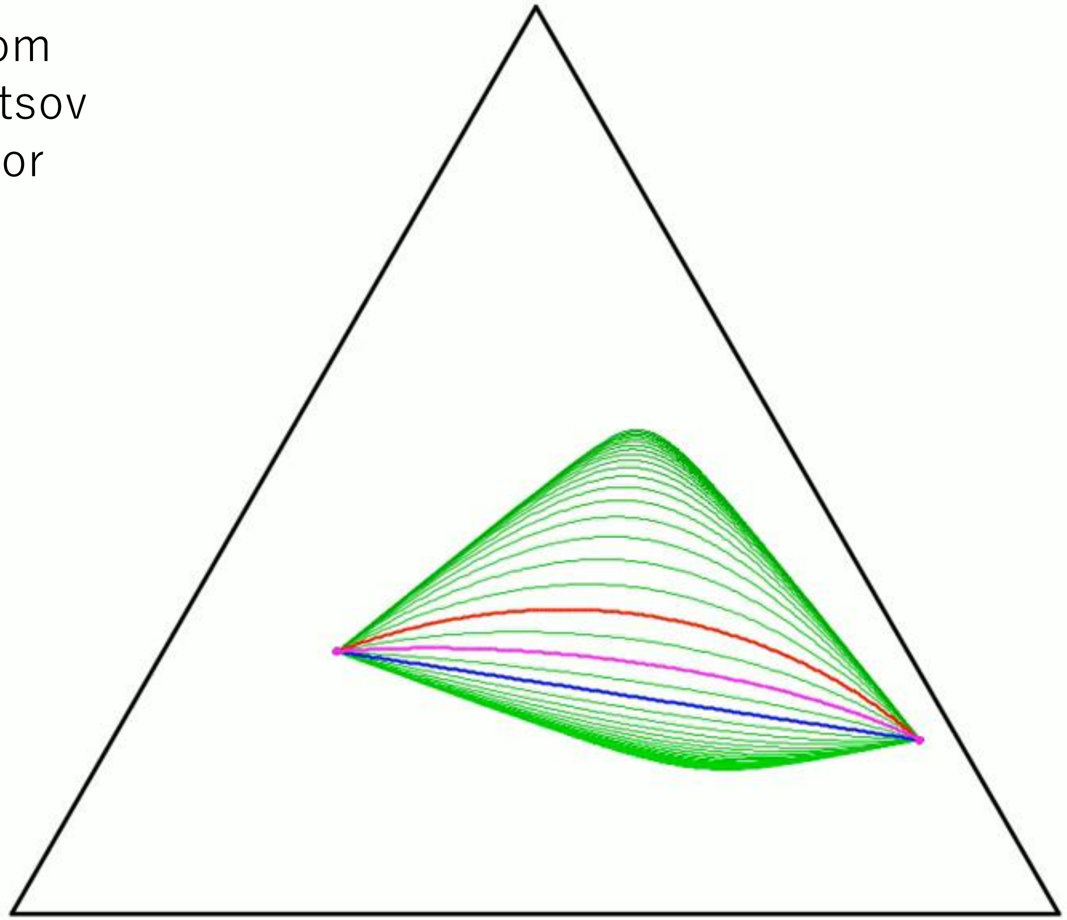
Probability simplex/Categorical manifold



$\pm$ **1-geometry or em-geometry**

Derived from
Amari-Chentsov
cubic tensor

Embedded probability simplex



$\pm$ **α-geometry**

Jensen-Shannon centroid for mixture families

- **Jensen-Shannon divergence between two mixtures amounts to a Jensen divergence**: $\mathrm{JS}(p_1, p_2) = J_F(\theta_1, \theta_2)$ for $p_1 = m_{\theta_1}$ and $p_2 = m_{\theta_2}$, where

$$\mathrm{JS}(p,q) := \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right) \qquad J_F(\theta_1 : \theta_2) = \frac{F(\theta_1)+F(\theta_2)}{2} - F\left(\frac{\theta_1+\theta_2}{2}\right).$$

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

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

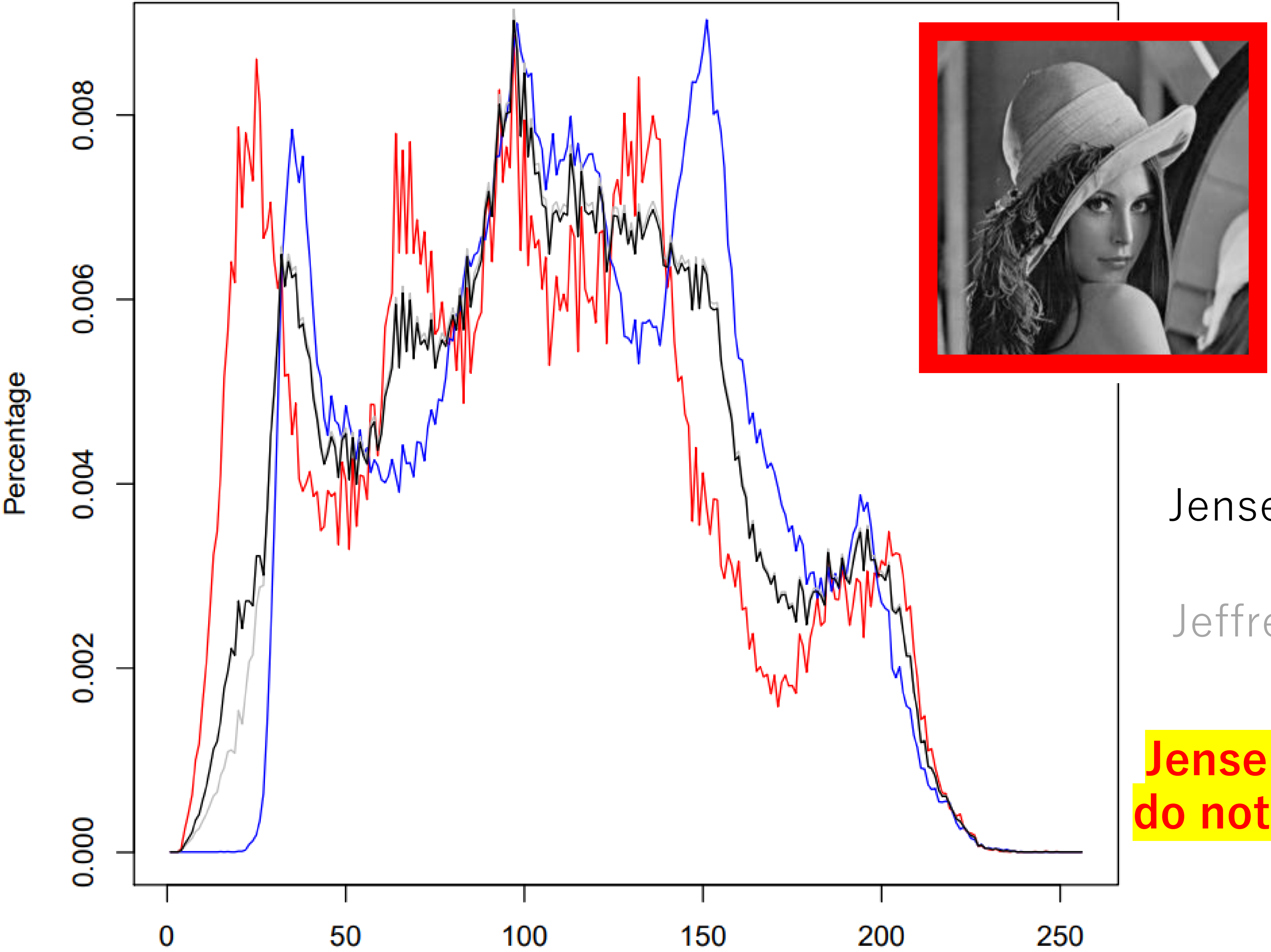
$$\min_\theta \sum_i J_F(\theta_i, \theta),$$

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

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

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

F is negentropy



Jensen–Shannon centroid

Jeffreys/SKL centroid

Jensen–Shannon centroid do not require same support

Family of categorical distributions is both an exponential family and a mixture family!

	Exponential Family $\xleftrightarrow{*}$	**Mixture Family**
pdf	$p_\theta(x) = \prod_{i=1}^{d} p_i^{t_i(x)}, p_i = \Pr(x = e_i), t_i(x) \in \{0,1\}, \sum_{i=1}^{d} t_i(x) = 1$	$m_\theta(x) = \sum_{i=1}^{d} p_i \delta_{e_i}(x)$
primal θ	$\theta_i = \log \frac{p_i}{p_d}$	$\theta_i = p_i$
$F(\theta)$	$\log(1 + \sum_{i=1}^{D} \exp(\theta_i))$	$\theta_i \log \theta_i + (1 - \sum_{i=1}^{D} \theta_i) \log(1 - \sum_{i=1}^{D} \theta_i)$
dual $\eta = \nabla F(\theta)$	$\frac{e^{\theta_i}}{1+\sum_{j=1}^{D} \exp(\theta_j)}$	$\log \frac{\theta_i}{1-\sum_{j=1}^{D} \theta_j}$
primal $\theta = \nabla F^*(\eta)$	$\log \frac{\eta_i}{1-\sum_{j=1}^{D} \eta_j}$	$\frac{e^{\theta_i}}{1+\sum_{j=1}^{D} \exp(\theta_j)}$
$F^*(\eta)$	$\sum_{i=1}^{D} \eta_i \log \eta_i + (1 - \sum_{j=1}^{D} \eta_j) \log(1 - \sum_{j=1}^{D} \eta_j)$	$\log(1 + \sum_{i=1}^{D} \exp(\eta_i))$
Bregman divergence	$B_F(\theta : \theta') = \mathrm{KL}^*(p_\theta : p_{\theta'})$ $= \mathrm{KL}(p_{\theta'} : p_\theta)$	$B_F(\theta : \theta') = \mathrm{KL}(m_\theta : m_{\theta'})$

Dual of a categorical exponential family is a categorical mixture family, and vice versa

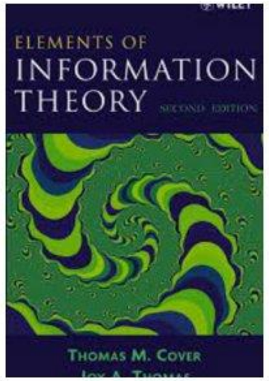
Chernoff information: A geometric characterization

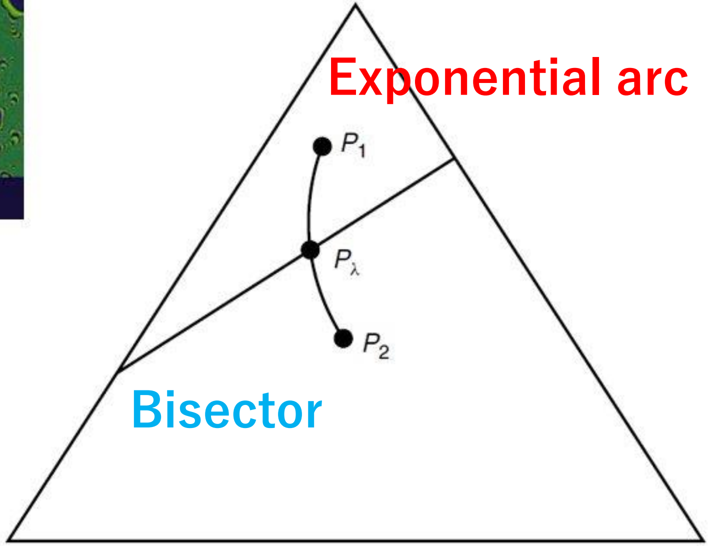
Exponential arc

$$C(P_1, P_2) \triangleq -\min_{0\le\lambda\le 1} \log\left(\sum_x P_1^\lambda(x) P_2^{1-\lambda}(x)\right)$$

Bisector

$$= D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2)$$





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

Probability simplex

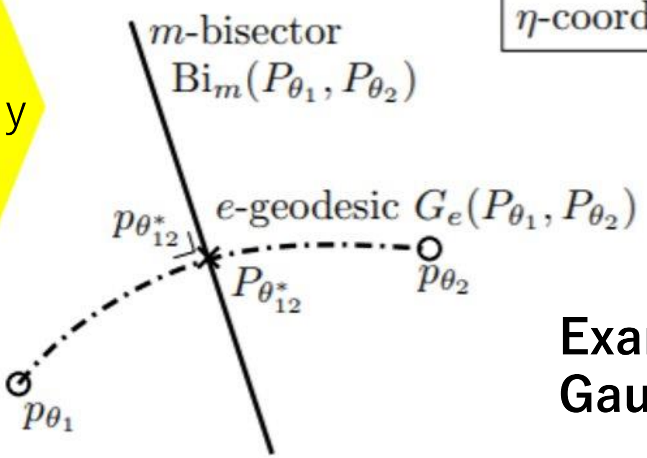
Generalized
to
exponential family
manifold

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

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

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

$\boxed{\eta\text{-coordinate system}}$



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

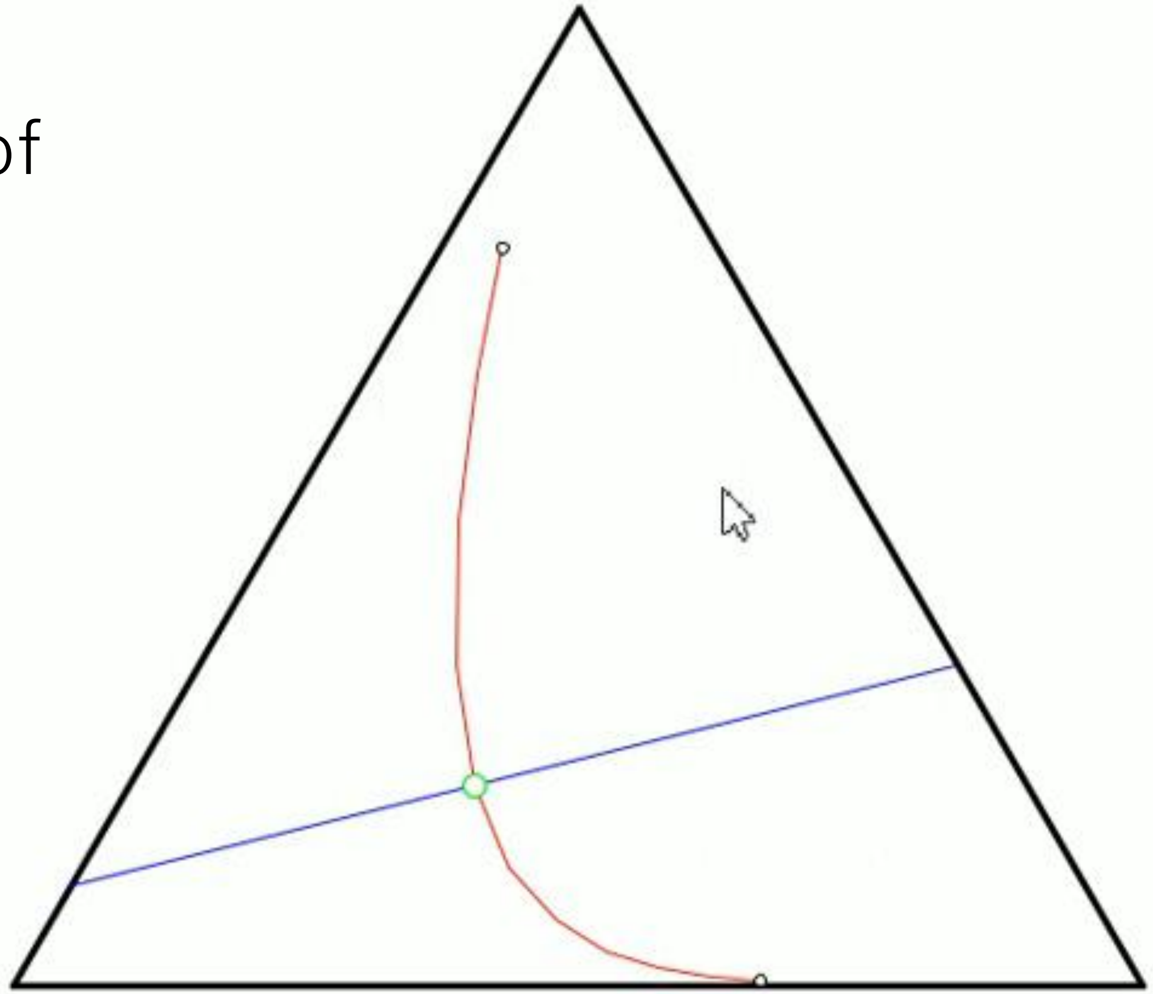
Example:
Gaussian manifold

$p(x|\theta) \propto \exp(<x, \theta>)$

Exponential family manifold

Chernoff point

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



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

Summary: Geometries of statistical models in ML

- **Fisher-Rao manifolds** = Riemannian manifolds wrt Fisher metrics.

  Fisher-Rao distance = Riemannian distance, metric distance

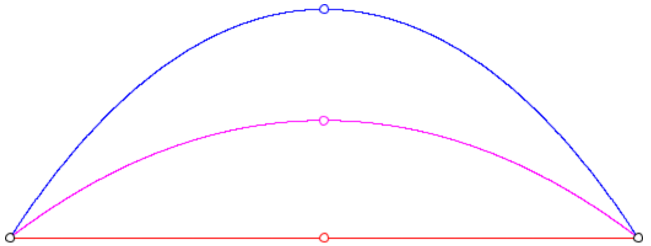
  ***Problems*: Are geodesics unique? Fisher-Rao distance in closed form?**

  → **Get fine approximations of Fisher-Rao between MVNs.**

- **Hilbert geometry** on bounded convex domains and Birkhoff geometry on cones, diffeomorphic and metric embeddings of MVN in the SPD cone:

  → **Get new fast distances between MVNs, straight line geodesics**

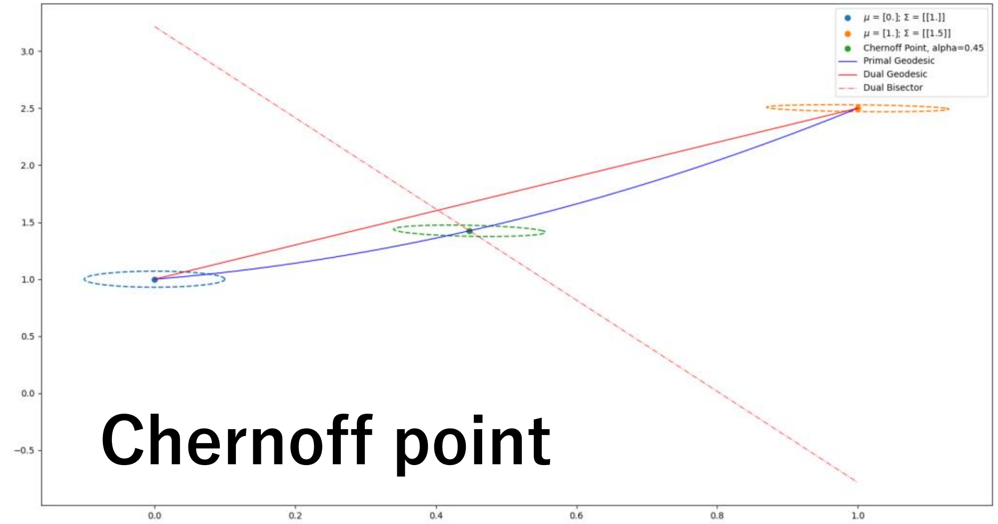
- **Bregman manifolds** = **dual geometry of convex functions**
  - Mixture families: F=negentropy, Jensen-Shannon centroid = Jensen centroid
  - Exponential families: F=cumulant function or Z=partition functions
  - **Duo Bregman divergences** and KLD between truncated exponential family densities



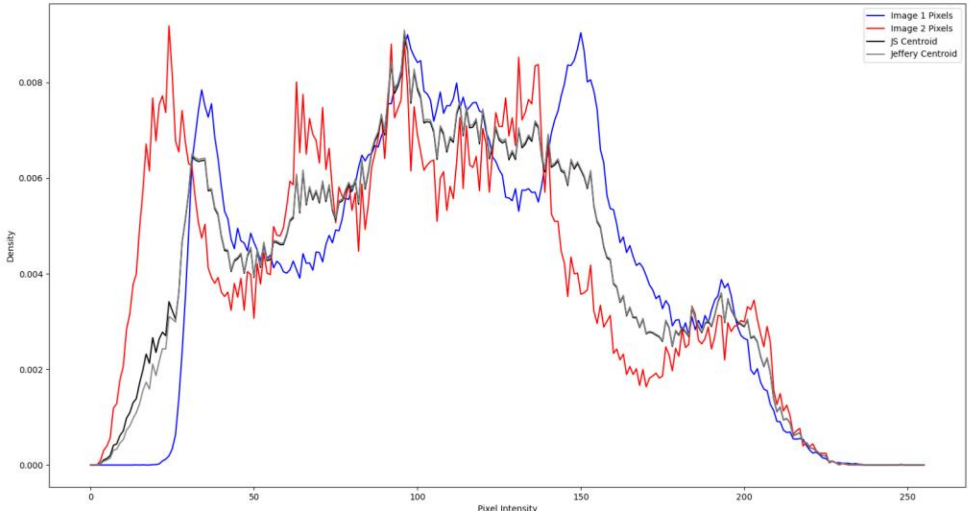
A Python library for geometric computing on Bregman Manifolds

pyBregMan

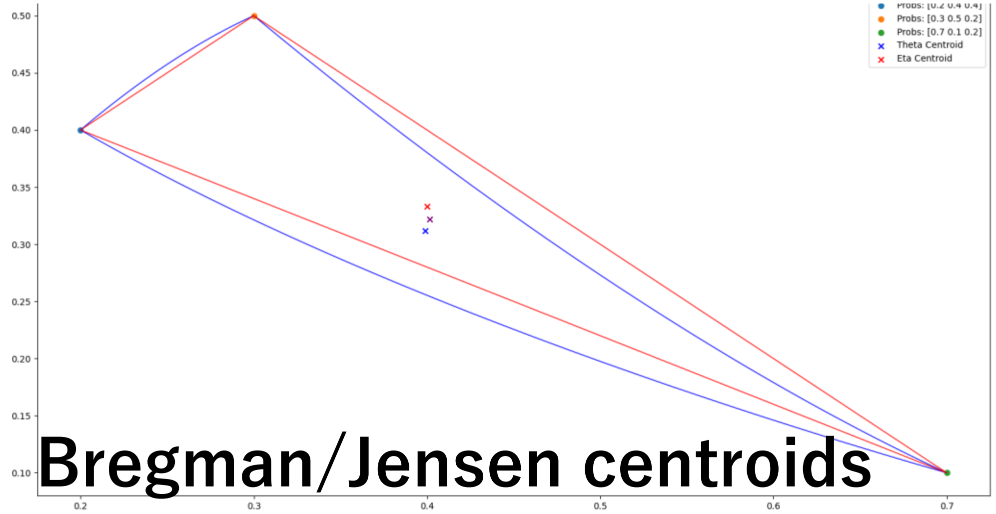
https://franknielsen.github.io/pyBregMan/



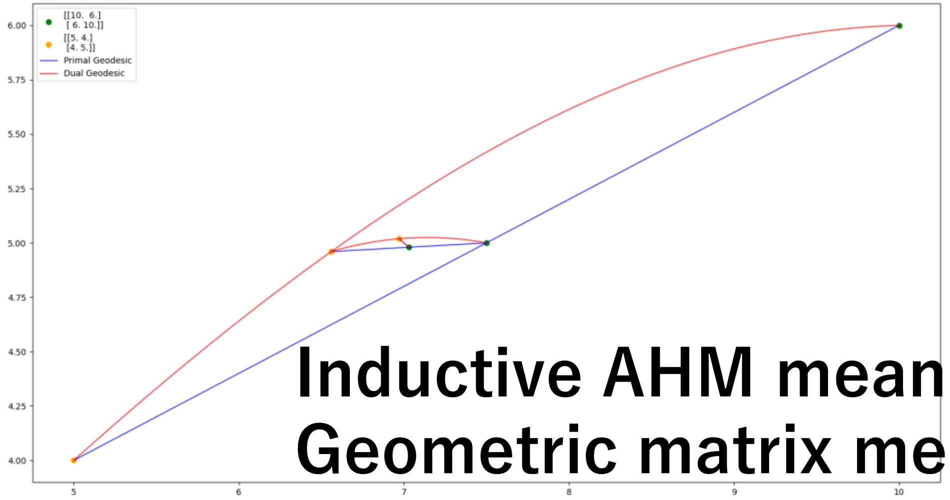
Chernoff point



Jensen-Shannon centroid

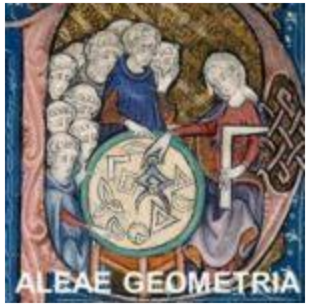


Bregman/Jensen centroids



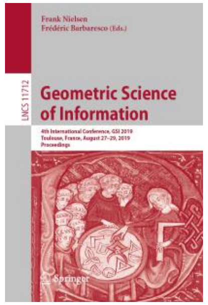
Inductive AHM mean
Geometric matrix mean

Joint work of Frank Nielsen and Alexander Soen

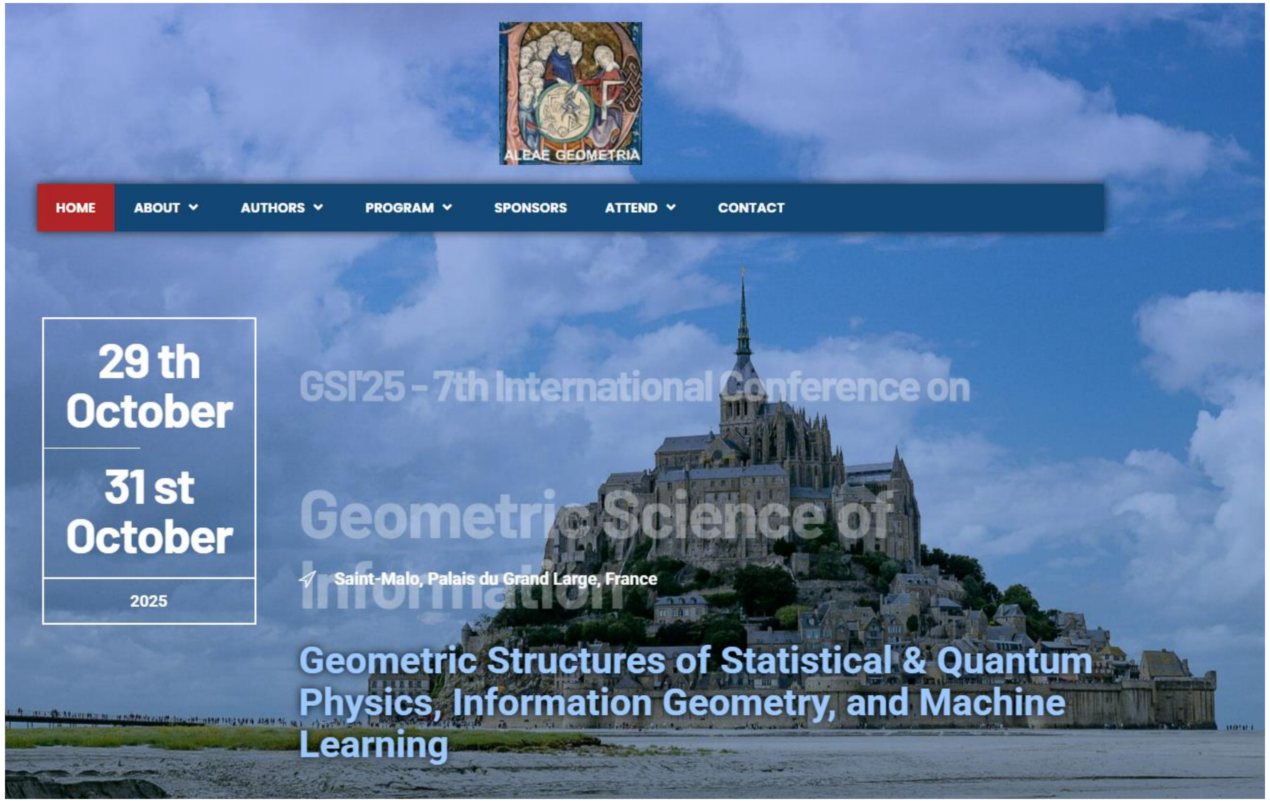


7th Geometric Science of Information (GSI)

https://conference-gsi.org/



Saint-Malo, convention center



Gala: Mont Saint Michel, France

Deadline (8-page LNCS paper): April 2nd, 2025

Topics of the 7th Geometric Science of Information: GSI'25

- Geometric Learning and Differential Invariants on Homogeneous Spaces
- Statistical Manifolds and Hessian information geometry
- Renyi Entropy & Information
- Geometric Foliation Structures of Dissipation and Machine Learning
- Geometric Structures of Quantum Information & Processing
- Applied Geometric Learning
- Probability, Information and
- Divergences in Statistics and Machine Learning
- Geometric Statistics
- Geometric Methods in Hybrid Classical/Quantum Systems
- Computational Information Geometry and Divergences
- Geometric Methods in Thermodynamics
- The Geometry of Classical & Quantum States
- Geometric Mechanics
- Stochastic Geometric Dynamics
- New trends in Nonholonomic Systems
- Learning of Dynamic Processes
- Neurogeometry
- PINN (Physics-Informed Neural Network) with Geometric Structures
- Lie Groups in Machine Learning
- Information Geometry, Toric Manifold
- A symplectic approach to differential equations
- Lie Group Based Method in Robotics & Kalman Filters
- Geometric and Analytical Aspects of Quantization and Non-Commutative Harmonic Analysis on Lie Groups
- Probability and Statistics on manifolds
- Deep learning: Methods, Analysis and Applications to Mechanical Systems
- Integrable Systems and Information Geometry
- Computing Geometry & Algebraic Statistics
- Geometric Green & Quantum Machine Learning
- Others

Thank you

Quoting Sir Michael Atiyah on **thinking geometrically**:

‘Algebra is the offer made by the devil to the mathematician. The devil says: "I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine.“’

One way to put the dichotomy in a more philosophical or literary framework is to say that algebra is to the geometer what you might call the ‘Faustian offer’. As you know, Faust in Goethe’s story was offered whatever he wanted (in his case the love of a beautiful woman), by the devil, in return for selling his soul. Algebra is the offer made by the devil to the mathematician. The devil says: ‘I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine.’ (Nowadays you can think of it as a computer!) Of course we like to have things both ways; we would probably cheat on the devil, pretend we are selling our soul, and not give it away. Nevertheless, the danger to our soul is there, because when you pass over into algebraic calculation, essentially you stop thinking; you stop thinking geometrically, you stop thinking about the meaning.

Michael Atiyah
"Mathematics in the 20th century." *Bulletin of the London Mathematical Society* 34.1 (2002): 1-15.

Some references for geometric structures

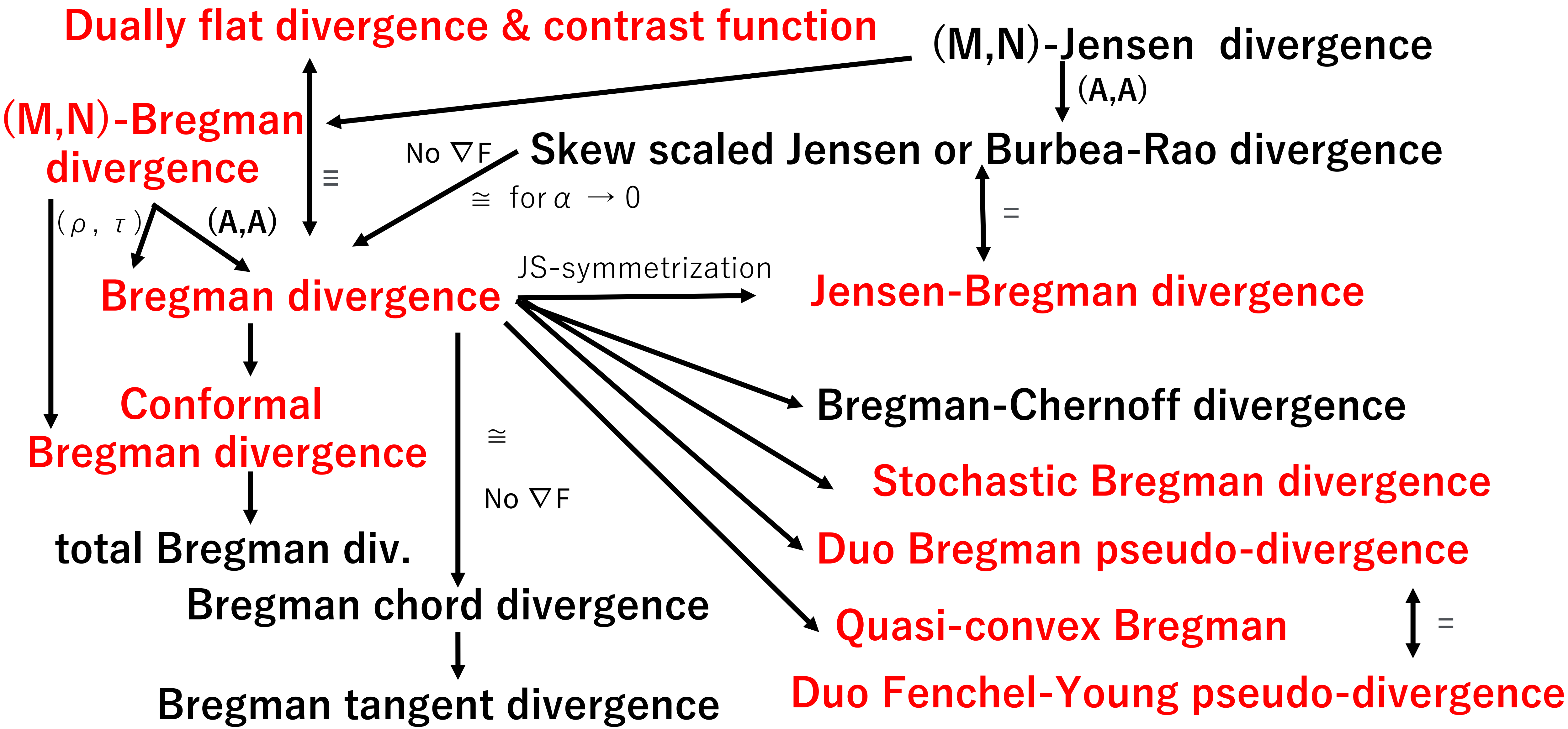
- **Fisher-Rao manifolds**: 2403.10089, 2302.08175
- **Hilbert/Birkhoff geometry**:
  - Simplex domain (categorical distributions): 2203.11434
  - Elliptope domain (correlation matrices): 1704.00454
  - Symmetric positive-definite cone (SPDs, MVNs): 2307.10644
  - Siegel domain, complex domain including SPD: 2004.08160
- **Bregman manifolds**:
  - Dual geometry of convex conjugate functions: 1910.03935
  - Applications to mixture families (Jensen-Shannon centroid): 1912.00610
  - Applications to exponential families (Chernoff information): 2207.03745
- **Dual statistical structures**:
  - "The many faces of information geometry." Not. Am. Math. Soc 69.1 (2022): 36-45.
  - "An elementary introduction to information geometry." Entropy 22.10 (2020): 1100.
- **Semi-Riemannian structures** (stochastic NNs): 1905.11027

Some references for geometric structures

- **Fisher-Rao manifolds**: 2403.10089, 2302.08175
- **Hilbert/Birkhoff geometry**:
  - Simplex domain (categorical distributions): 2203.11434
  - Elliptope domain (correlation matrices): 1704.00454
  - Symmetric positive-definite cone (SPDs, MVNs): 2307.10644
  - Siegel domain, complex domain including SPD: 2004.08160
- **Bregman manifolds**:
  - Dual geometry of convex conjugate functions: 1910.03935
  - Applications to mixture families (Jensen-Shannon centroid): 1912.00610
  - Applications to exponential families (Chernoff information): 2207.03745
- **Dual statistical structures**:
  - "The many faces of information geometry." Not. Am. Math. Soc 69.1 (2022): 36-45.
  - "An elementary introduction to information geometry." Entropy 22.10 (2020): 1100.
- **Semi-Riemannian structures** (stochastic NNs): 1905.11027

https://franknielsen.github.io/geometrymodels.html

Some generalizations of Bregman divergences



But also matrix Bregman divergence, functional Bregman divergence, submodular Bregman divergence, etc.

Inductive matrix arithmetic-harmonic mean (AHM)

- Consider the cone of symmetric positive-definite matrices (SPD cone), and extend the AHM to SPD matrices:

[Nakamura 2001]

$$A_{t+1} = \frac{A_t + H_t}{2} = A(A_t, H_t)$$ ←arithmetic mean

$$H_{t+1} = 2\left(A_t^{-1} + H_t^{-1}\right)^{-1} = H(A_t, H_t)$$ ←harmonic mean

- Sequence with A_0=X and H_0=Y converge quadratically to **matrix geometric mean**:

$$\mathrm{AHM}(X,Y) = \lim_{t\to+\infty} A_t = \lim_{t\to+\infty} H_t.$$

$$\boxed{\mathrm{AHM}(X,Y) = X^{\frac{1}{2}}\left(X^{-\frac{1}{2}}\, Y\, X^{-\frac{1}{2}}\right)^{\frac{1}{2}} X^{\frac{1}{2}} = G(X,Y)}$$

which is also the **Riemannian center of mass** wrt the trace metric:

$$G(X,Y) = \arg\min_{M\in\mathbb{P}(d)} \frac{1}{2}\rho^2(X,M) + \frac{1}{2}\rho^2(Y,M).$$

$$\rho(P_1,P_2) = \sqrt{\sum_{i=1}^{d} \log^2 \lambda_i\left(P_1^{-\frac{1}{2}}\, P_2\, P_1^{-\frac{1}{2}}\right)}$$ Riemannian distance

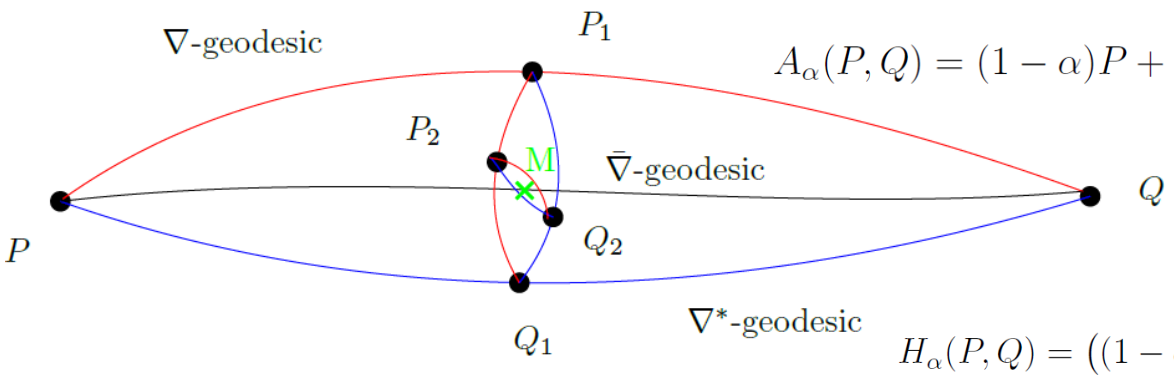
What is… an Inductive Mean? *Notices of the American Mathematical Society* 2023

Geometric interpretation of the AHM matrix mean

Repeat:
$$A_{t+1} = \frac{A_t + H_t}{2} = A(A_t, H_t) \qquad P_{t+1} = \gamma\left(P_t, Q_t : \frac{1}{2}\right)$$
$$H_{t+1} = 2\,(A_t^{-1} + H_t^{-1})^{-1} = H(A_t, H_t) \qquad Q_{t+1} = \gamma^*\left(P_t, Q_t : \frac{1}{2}\right)$$

(SPD, g^G, ∇^A, ∇^H) is a dually flat space, ∇^G is Levi-Civita connection

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



$$A_\alpha(P,Q) = (1-\alpha)P + \alpha Q$$

$$H_\alpha(P,Q) = \left((1-\alpha)P^{-1} + \alpha Q^{-1}\right)^{-1}$$

Dually flat space (SPD, g^G, ∇^A, ∇^H)

Primal geodesic midpoint is the arithmetic center wrt Euclidean metric
Dual geodesic midpoint = harmonic center wrt an isometric Eucl. metric
Levi-Civita geodesic midpoint is geometric Karcher mean

$$g_P^A(X,Y) = \mathrm{tr}(X^\top Y)$$
$$g_P^H(X,Y) = \mathrm{tr}(P^{-2}XP^{-2}Y)$$
$$g_P^G(X,Y) = \mathrm{tr}(P^{-1}XP^{-1}Y)$$

Here, all 3 connections are metric connections

[Nakamura 2001]

Symmetrized Bregman centroid

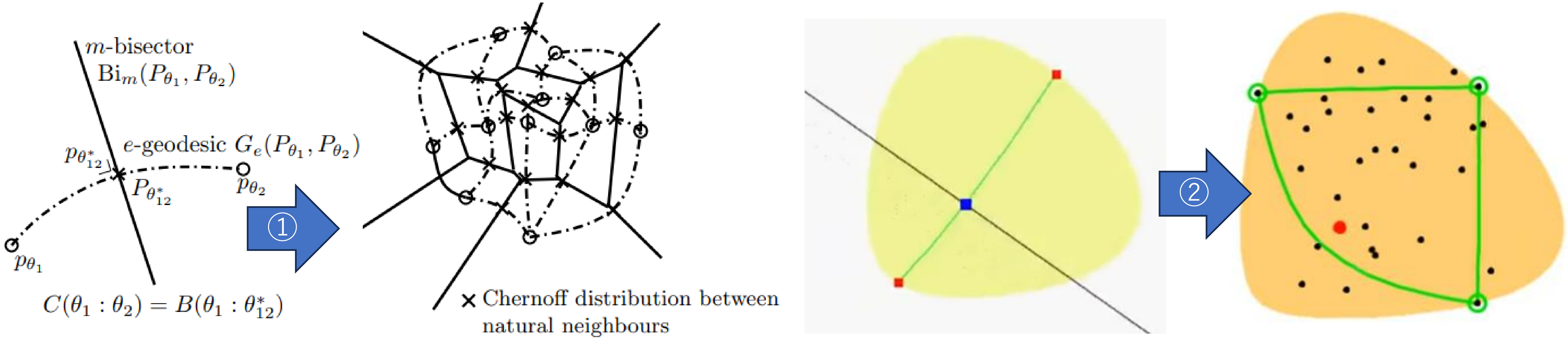
Use convex duality + dual Bregman information-bias decompositions:

$$
\begin{aligned}
& \min_\theta && \sum_i w_i S_F(\theta_i, \theta), \\
=\ & \min_\theta && \sum_i w_i B_F(\theta_i : \theta) + \sum_i w_i B_F(\theta : \theta_i), \\
=\ & \min_\theta && J_{F,w}((\theta_i)_i) + B_F(\bar\theta : \theta) + \sum_i w_i B_{F^*}(\eta_i : \nabla F(\theta)), \\
\equiv\ & \min_\theta && B_F(\bar\theta : \theta) + J_{F^*,w}((\eta_i)_i) + B_{F^*}(\bar\eta : \nabla F(\theta)), \\
\equiv\ & \min_\theta && B_F(\bar\theta : \theta) + B_F(\theta : \underline{\theta}), \\
=\ & \min_\eta && B_{F^*}(\eta : \nabla F(\bar\theta)) + B_{F^*}(\nabla F(\underline{\theta}) : \eta), \\
=\ & \min_\eta && B_{F^*}(\bar\eta : \eta) + B_{F^*}(\eta : \underline{\eta})
\end{aligned}
$$

Amounts to simpler dual optimization problems on the sided Bregman centroids (2-point optimization vs n-point optimization problems)

Two generalizations of m-Chernoff information

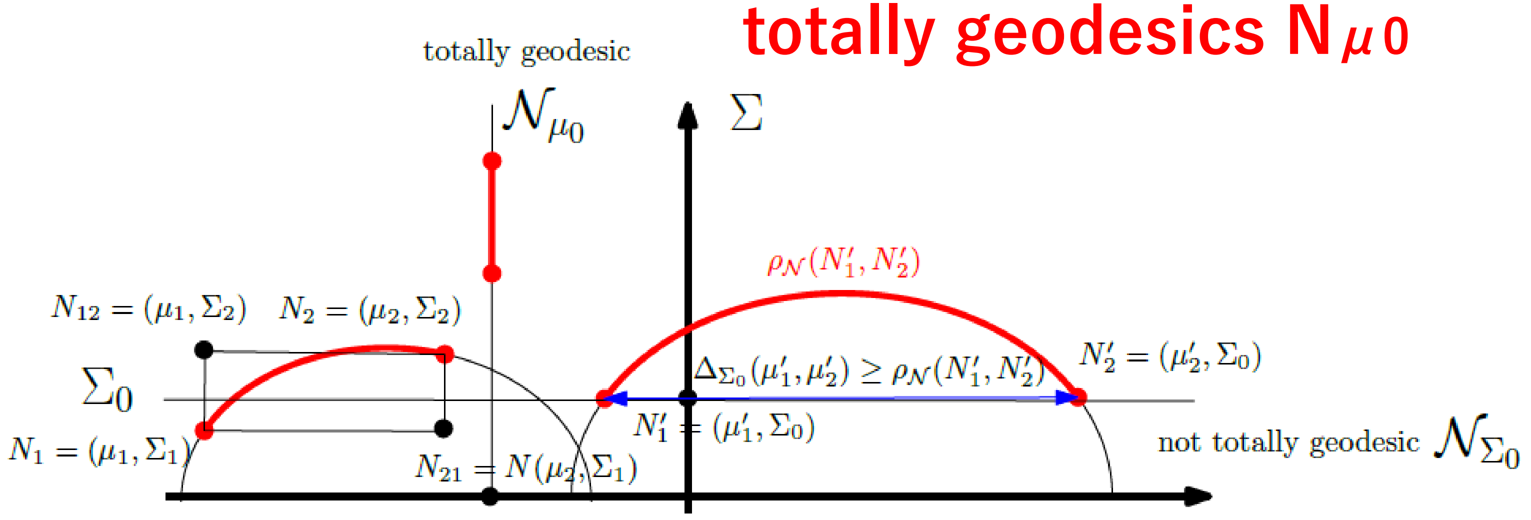
- Historically, **Chernoff information** defined to upper bound the probability of error in Bayesian hypothesis testing, found later applications in information fusion, distributed estimation, etc.
- Generalization to m hypothesis: ① **minimum pairwise Chernoff information**



- Interpret Chernoff information as ② **radius of minimum enclosing Kullback-Leibler divergence**, extend Chernoff information to m

Revisiting Chernoff information with likelihood ratio exponential families. *Entropy*, *24*(10), 2022

Special case: Submanifolds of constant covariance matrices



not totally geodesics $N_{\Sigma 0}$
(hence upper bounds Fisher-Rao)

Proposition . *The Fisher–Rao distance $\rho_{\mathcal{N}}((\mu_1, \Sigma), (\mu_2, \Sigma))$ between two MVNs with same covariance matrix is*

$$\begin{aligned}
\rho_{\mathcal{N}}((\mu_1, \Sigma), (\mu_2, \Sigma)) &= \rho_{\mathcal{N}}((0,1), (\Delta_\Sigma(\mu_1, \mu_2), 1)), \\
&= \sqrt{2}\log\left(\frac{\sqrt{8+\Delta_\Sigma^2(\mu_1,\mu_2)}+\Delta_\Sigma(\mu_1,\mu_2)}{\sqrt{8+\Delta_\Sigma^2(\mu_1,\mu_2)}-\Delta_\Sigma(\mu_1,\mu_2)}\right), \\
&= \sqrt{2}\,\mathrm{arccosh}\left(1+\frac{1}{4}\Delta_\Sigma^2(\mu_1,\mu_2)\right),
\end{aligned}$$

where $\Delta_\Sigma(\mu_1, \mu_2) = \sqrt{(\mu_2-\mu_1)^\top \Sigma^{-1}(\mu_2-\mu_1)}$ is the Mahalanobis distance.

Jensen-Shannon centroid of categorical distributions

Input: A set $\{p_i = (p_i^1, \dots, p_i^d)\}_{i\in[n]}$ of n categorical distributions belonging to the $(d-1)$-dimensional probability simplex Δ_{d-1}

Input: T: The number of CCCP iterations

Output: An approximation ${}^{(T)}\bar{p}$ of the Jensen–Shannon centroid $\bar{p}$ minimizing $\sum_i D_{\mathrm{JS}}(c, p_i)$

/* Convert the categorical distributions to their natural parameters by dropping the last coordinate */

$\theta_i^j = p_i^j$ for $j \in \{1, \dots, d-1\}$;

/* Initialize the JS centroid */

$t \leftarrow 0$;

${}^{(0)}\bar{\theta} = \frac{1}{n}\sum_{i=1} \theta_i$;

/* Convert the initial natural parameter of the JS centroid to a categorical distribution */

${}^{(0)}\bar{p}^j = {}^{(0)}\bar{\theta}^j$ for $j \in \{1, \dots, d-1\}$;

${}^{(0)}\bar{p}^d = 1 - \sum_{i=1}^d {}^{(0)}\bar{p}^j$;

/* Perform the ConCave-Convex Procedure (CCCP) */

while $t \leq T$ **do**

/* Use $\nabla F(\theta) = \left[\log \frac{\theta_i}{1-\sum_{j=1}^D \theta_j}\right]_i$ and $\nabla F^{-1}(\eta) = \frac{1}{1+\sum_{j=1}^D \exp(\eta_j)}[\exp(\eta_i)]_i$ */

$\boxed{{}^{(t+1)}\theta = (\nabla F)^{-1}\left(\frac{1}{n}\sum_i \nabla F\left(\frac{\theta_i + {}^{(t)}\theta}{2}\right)\right);}$ **ConCave Convex algorithm or DCA**

$t \leftarrow t+1$;

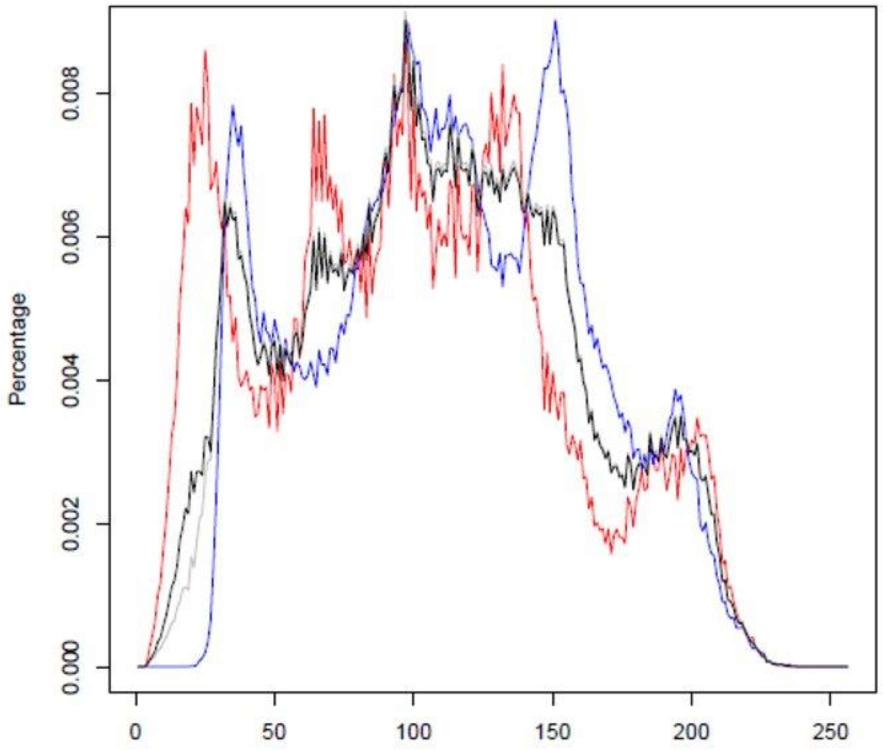
end

/* Convert back the natural parameter to the categorical distribution of the approximated Jensen-Shannon centroid */

${}^{(T)}\bar{p}^j = {}^{(T)}\bar{\theta}^j$ for $j \in \{1, \dots, d-1\}$;

${}^{(T)}\bar{p}^d = 1 - \sum_{i=1}^d {}^{(T)}\bar{p}^j$;

return ${}^{(T)}\bar{p}$;



- Use the fact that the set of categorical distributions is a **mixture family** in information geometry

JSD centroid = Jensen centroid

Unifying Jeffreys with Jensen-Shannon divergences

Kullback-Leibler divergence $\mathrm{KL}(p:q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ can be symmetrized as:

- Jeffreys divergence: $J(p,q) = \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = J(q,p) \quad = \int (p(x) - q(x)) \log \frac{p(x)}{q(x)} \mathrm{d}x.$
- Jensen-Shannon divergence: $\mathrm{JS}(p,q) = \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right) = \mathrm{JS}(q,p)$

$$\boxed{\mathrm{sKL}^{(\alpha)}(p,q) = \frac{1}{2\alpha(1-\alpha)} \left(H(\alpha p + (1-\alpha)q) + H((1-\alpha)p + \alpha q) - (H(p) + H(q))\right) \geq 0}$$

with Shannon entropy: $H(p) = \int p(x) \log \frac{1}{p(x)} \mathrm{d}x = -\int p(x) \log p(x) \mathrm{d}x.$

Unify and generalize Jeffreys divergence with Jensen-Shannon divergence:

$$\lim_{\alpha \to 0} \mathrm{sKL}^{(\alpha)}(p,q) = J(p,q) \qquad \mathrm{sKL}^{(\frac{1}{2})}(p,q) = 2\left(2H\left(\frac{p+q}{2}\right) - (H(p) + H(q))\right) = 4\mathrm{JS}(p,q).$$

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

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