Recent contributions to Distances and information geometry: A computational viewpoint

Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/

31st July 2020

Outline

1. **Siegel-Klein geometry (bounded complex matrix domains)**

   Hilbert geometry of the Siegel disk: The Siegel-Klein disk model
   https://arxiv.org/abs/2004.08160

2. **Information-geometric structures on the Cauchy manifold**

   On Voronoi Diagrams on the Information-Geometric Cauchy Manifolds
   Entropy 2020, 22(7), 713; https://doi.org/10.3390/e22070713
   https://www.mdpi.com/1099-4300/22/7/713

3. **Generalizations of the Jensen-Shannon divergence & JS centroids**

   On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means
   Entropy 2019, 21(5), 485; https://doi.org/10.3390/e21050485
   https://www.mdpi.com/1099-4300/21/5/485

   On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid
   Entropy 2020, 22(2), 221; https://doi.org/10.3390/e22020221
   https://www.mdpi.com/1099-4300/22/2/221

Hilbert geometry of the Siegel disk: The Siegel-Klein disk model

Frank Nielsen

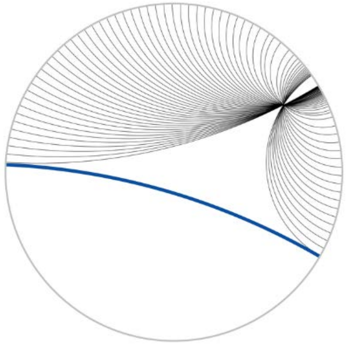
Sony Computer Science Laboratories, Inc

https://franknielsen.github.io/

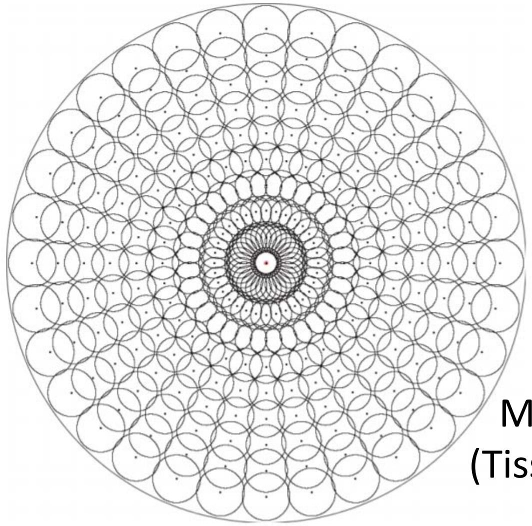
July 2020

https://arxiv.org/abs/2004.08160

Main standard models of hyperbolic geometry

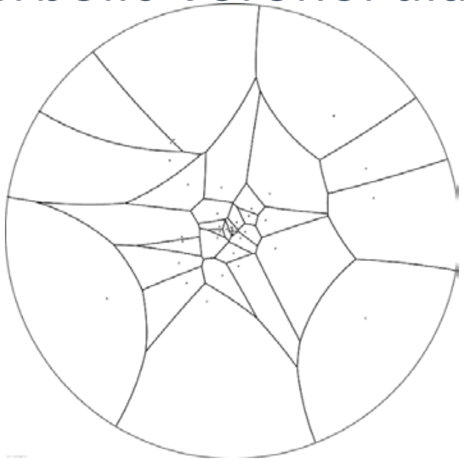


Conformal Poincaré model:

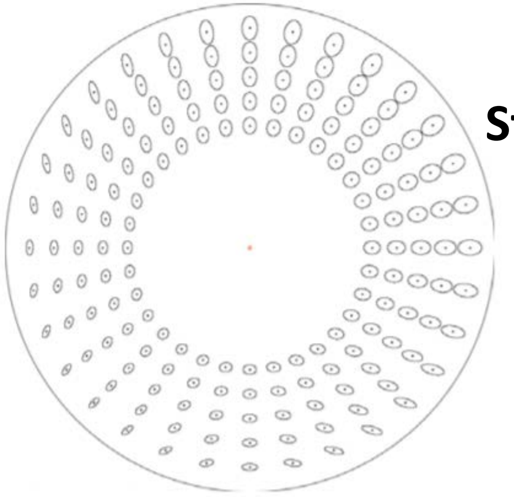


Metric tensor
(Tissot indicatrix)

Hyperbolic Voronoi diagram

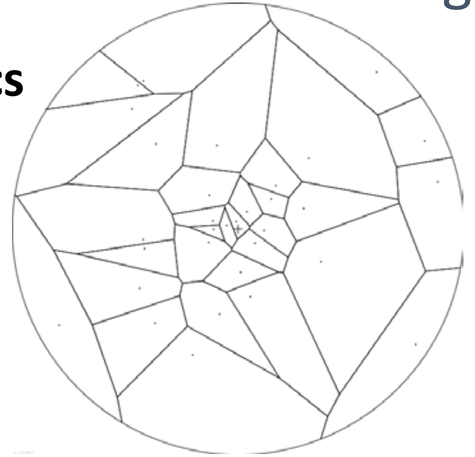


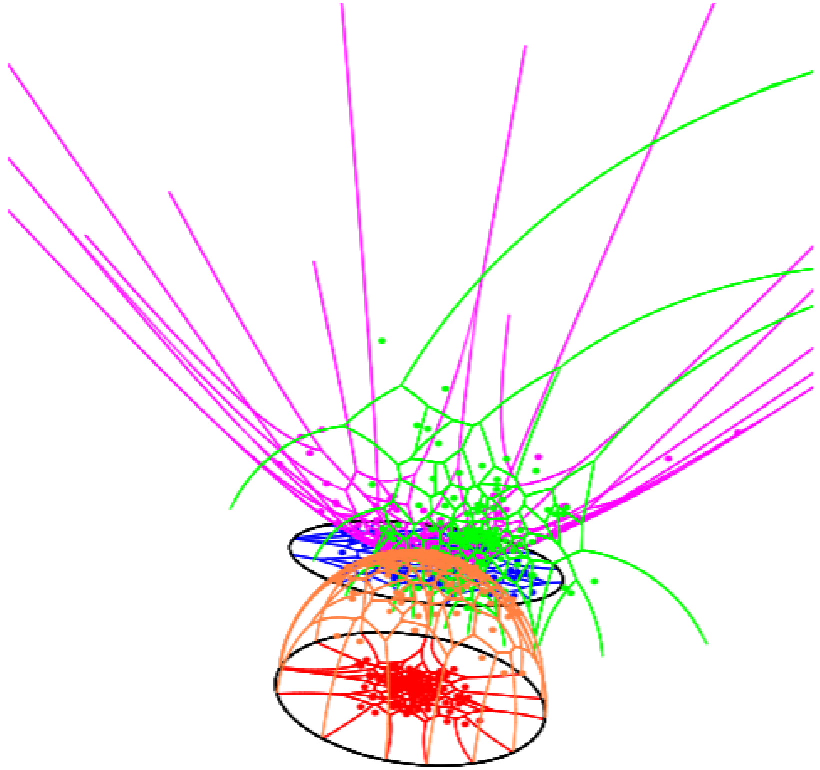
Lesser known non-conformal Klein model:



Straight geodesics

Hyperbolic Voronoi diagram





Hyperbolic Voronoi diagrams
in 5 models

https://www.youtube.com/watch?v=i9IUzNxeH4o&t=3s

Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010

Siegel upper space

Birth of symplectic geometry (complex matrix groups, Siegel & Hua, 1940's)

Generalization of the Poincaré upper plane to *complex matrix domains*:

$$\boxed{\mathbb{SH}(d) := \{Z = X + iY \ : \ X \in \mathrm{Sym}(d,\mathbb{R}), Y \in \mathrm{PD}(d,\mathbb{R})\}\,.}$$

PD: Positive-definite cone

Infinitesimal length element: $\mathrm{d}s_U^2(Z) = 2\mathrm{tr}\left(Y^{-1}\mathrm{d}Z\ Y^{-1}\mathrm{d}\bar{Z}\right)$

Geodesic length distance: $$\boxed{\rho_U(Z_1, Z_2) = \sqrt{\sum_{i=1}^d \log^2\left(\frac{1+\sqrt{r_i}}{1-\sqrt{r_i}}\right)},}$$

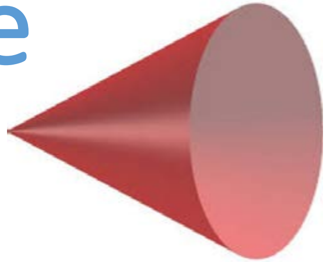
Spectral decomposition

with the i-th **real** eigenvalue $r_i = \lambda_i\left(R(Z_1, Z_2)\right)$

Matrix cross-ratio: $R(Z_1, Z_2) := (Z_1 - Z_2)(Z_1 - \overline{Z}_2)^{-1}(\overline{Z}_1 - \overline{Z}_2)(\overline{Z}_1 - Z_2)^{-1}$

R: Not Hermitian, but all real eigenvalues!

Siegel upper space: Generalize PD matrix cone



PD: Positive-definite cone

$$\mathbb{SH}(d) := \{Z = X + iY \;:\; X \in \mathrm{Sym}(d,\mathbb{R}), Y \in \mathrm{PD}(d,\mathbb{R})\}\,.$$

$$\mathrm{d}s_U^2(Z) = 2\mathrm{tr}\left(Y^{-1}\mathrm{d}Z\; Y^{-1}\mathrm{d}\bar{Z}\right) \Rightarrow \boxed{\mathrm{d}s_U^2(Z) = \mathrm{tr}\left((Y^{-1}\mathrm{d}Y)^2\right) = \mathrm{d}s_{\mathrm{PD}}(Y)}$$

$$\boxed{\begin{aligned}\rho_{\mathrm{PD}}(Y_1,Y_2) &= \|\mathrm{Log}(Y_1Y_2^{-1})\|_F\\ &= \sqrt{\sum_{i=1}^{d}\log^2\left(\lambda_i(Y_1Y_2^{-1})\right)}.\end{aligned}}$$

$$\rho_{\mathrm{PD}}(C^\top Y_1 C, C^\top Y_2 C) = \rho_{\mathrm{PD}}(Y_1,Y_2) \quad C \in \mathrm{GL}(d,\mathbb{R})$$

$$\rho_{\mathrm{PD}}(Y_1^{-1},Y_2^{-1}) = \rho_{\mathrm{PD}}(Y_1,Y_2)$$

Siegel upper space: Generalize Poincaré upper plane

When complex dimension is 1, recover the Poincaré upper plane

$$\rho_U(Z_1,Z_2) = \rho_U(z_1,z_2),$$

$$\rho_U(z_1,z_2) := \log\frac{|z_1-\bar{z}_2| + |z_1-z_2|}{|z_1-\bar{z}_2| - |z_1-z_2|}$$

several equivalent formulas…

Generalized linear fractional transformations

Siegel upper space metric is invariant under generalized Moebius transformations called (biholomorphic) **symplectic maps**:

$$\boxed{\phi_S(Z) := (AZ+B)(CZ+D)^{-1}, \qquad S = \begin{bmatrix} A & B \\ C & D \end{bmatrix}}$$

(matrix group representation)

Real symplectic group Sp(d,R):

$$\mathrm{Sp}(d,\mathbb{R}) = \left\{ \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad A,B,C,D \in M(d,\mathbb{R}) : AB^\top = BA^\top, \quad CD^\top = DC^\top, \quad AD^\top - BC^\top = I \right\}.$$

Group inverse: $S^{(-1)} =: \begin{bmatrix} D^\top & -B^\top \\ -C^\top & A^\top \end{bmatrix}$

Group action is **transitive**:

($\rightarrow$ homogeneous space)

$$\phi_{S(Z)}(iI) = Z$$

(translation Z=A+iB) $S(Z) = \begin{bmatrix} B^{-\frac{1}{2}} & 0 \\ AB^{-\frac{1}{2}} & B^{\frac{1}{2}} \end{bmatrix}$

$$\phi_{S^{(-1)}(Z)}(Z) = iI$$

$$S^{(-1)}(Z) = \begin{bmatrix} (B^{\frac{1}{2}})^\top & 0 \\ -(AB^{-\frac{1}{2}})^\top & (B^{-\frac{1}{2}})^\top \end{bmatrix}$$

Orientation-preserving isometry in the Siegel upper space

Stabilizer group of Z=iI: The symplectic orthogonal matrices:
(informally, play the role of "rotations" in the Siegel geometry)

$$\mathrm{SpO}(2d,\mathbb{R}) = \left\{ \begin{bmatrix} A & B \\ -B & A \end{bmatrix} \ : \ A^\top A + B^\top B = I, A^\top B \in \mathrm{Sym}(d,\mathbb{R}) \right\}$$

$$\mathrm{SpO}(2d,\mathbb{R}) = \mathrm{Sp}(2d,\mathbb{R}) \cap O(2d) \qquad O(2d) := \left\{ R \in M(2d,\mathbb{R}) \ : \ RR^\top = R^\top R = I \right\}$$

Orientation preserving isometry: $\boxed{\mathrm{PSp}(d,\mathbb{F}) = \mathrm{Sp}(d,\mathbb{F}) / \{I_{2d}\}}$

When complex dimension is 1 (Poincaré upper plane), recover **PSL(2,R)**

Siegel disk domain

Partial Loewner ordering

Disk domain: $\mathbb{SD}(d) \quad := \quad \{W \in \mathrm{Sym}(d, \mathbb{C}) \ : \ I - \overline{W}W \succ 0\}$

Or equivalently $\mathbb{SD}(d) := \{W \in \mathrm{Sym}(d, \mathbb{C}) \ : \ I - W\overline{W} \succ 0\}$

A generalization of Poincaré conformal disk: $\mathbb{SD}(1) = \mathbb{D}$

Spectral/operator norm:

$$\begin{aligned} \|M\|_O &= \max_{x \neq 0} \frac{\|Mx\|_2}{\|x\|_2}, \\ &= \sqrt{\lambda_{\max}(M^H M)}, \\ &= \sigma_{\max}(M). \end{aligned}$$

(= Maximum singular value >=0)

Siegel disk domain:

Shilov boundary

Stratified space (by matrix rank)

$\mathbb{SD}(d) = \{W \in \mathrm{Sym}(d, \mathbb{C}) \ : \ \|W\|_O < 1\}$

Distance in the Siegel disk domain

Siegel metric in the disk domain:

$$\boxed{\mathrm{d}s_D^2 = \mathrm{tr}\left((I - W\overline{W})^{-1}\mathrm{d}W(I - W\overline{W})^{-1}\mathrm{d}\bar{W}\right)}$$

When complex dimension is 1, recover the Poincaré disk metric:

$$\mathrm{d}s_D^2 = \frac{1}{(1-|w|^2)^2}\mathrm{d}w\mathrm{d}\bar{w}$$

Siegel disk distance:

$$\boxed{\rho_D(W_1, W_2) = \log\left(\frac{1 + \|\Phi_{W_1}(W_2)\|_O}{1 - \|\Phi_{W_1}(W_2)\|_O}\right)}$$

Siegel translation of W1 to the origin matrix 0 (= Siegel translation):

$$\Phi_{W_1}(W_2) = (I - W_1\overline{W}_1)^{-\frac{1}{2}}(W_2 - W_1)(I - \overline{W}_1W_2)^{-1}(I - \overline{W}_1W_1)^{\frac{1}{2}}$$

Costly to calculate because we need **square root and inverse matrices**

Complex symplectic group (for Siegel disk)

$$\mathrm{Sp}(d,\mathbb{C}) = \left\{ M^\top J M = J, M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in M(2d,\mathbb{C}) \right\} \qquad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$$

Equivalent to $\quad AB^\top = BA^\top, \quad CD^\top = DC^\top, \quad AD^\top - BC^\top = I.$

$$\mathrm{Sp}(d,\mathbb{C}) = \left\{ M = \begin{bmatrix} A & B \\ \bar{B} & \bar{A} \end{bmatrix} \in M(2d,\mathbb{C}) \right\}, \qquad \begin{array}{rcl} A^\top \bar{B} - B^H A & = & 0, \\ A^\top \bar{A} - B^H B & = & I. \end{array}$$

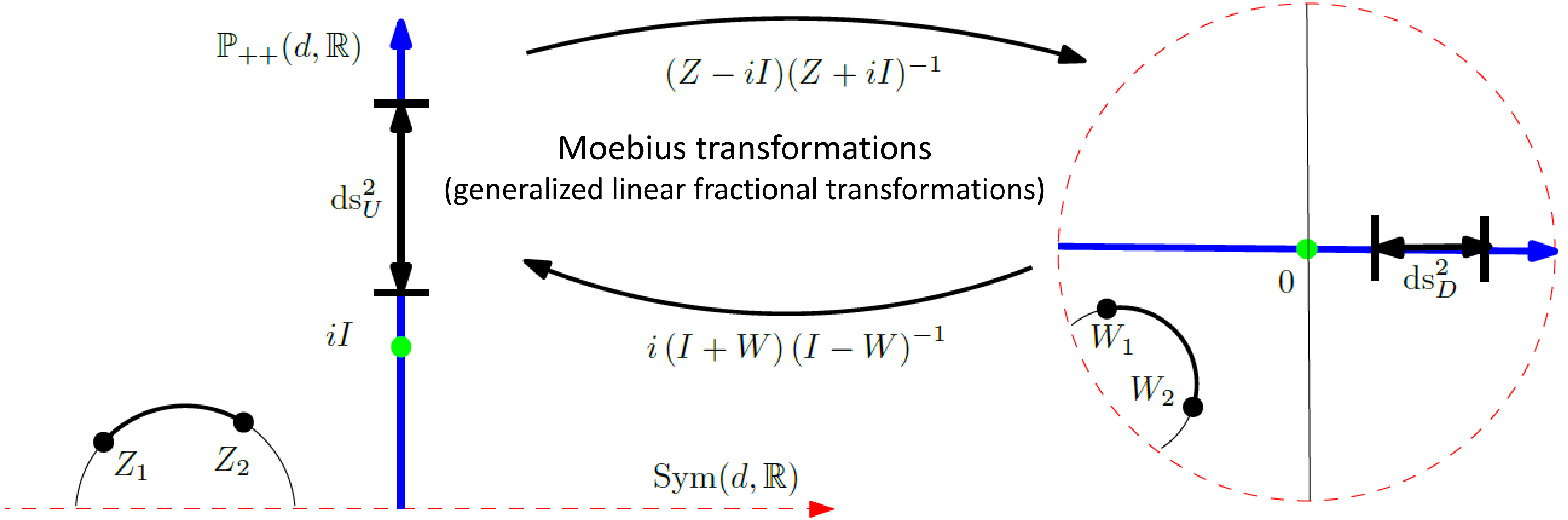
Orientation-preserving isometry in the Siegel disk:

$$\boxed{\mathrm{PSp}(d,\mathbb{C}) = \mathrm{Sp}(d,\mathbb{C})/\{\pm I_{2d}\}}$$

PSL(2,C) in 1D

Conversions Siegel upper space <-> Siegel disk

$$\mathbb{SD}(d) := \{W \in \mathrm{Sym}(d,\mathbb{C}) \ : \ I - \bar{W}W \succ 0\}$$



$$\mathrm{ds}^2_U(Z) = 2\mathrm{tr}\left(Y^{-1}\mathrm{d}Z\, Y^{-1}\mathrm{d}\bar{Z}\right)$$

$$\rho_U(Z_1,Z_2) = \sqrt{\sum_{i=1}^d \log^2\left(\frac{1+\sqrt{r_i}}{1-\sqrt{r_i}}\right)}$$

$$r_i = \lambda_i\left(R(Z_1,Z_2)\right)$$

$$R(Z_1,Z_2) := (Z_1-Z_2)(Z_1-\bar{Z}_2)^{-1}(\bar{Z}_1-\bar{Z}_2)(\bar{Z}_1-\bar{Z}_2)^{-1}$$

$$\mathrm{ds}^2_D = \mathrm{tr}\left((I-W\bar{W})^{-1}\mathrm{d}W(I-W\bar{W})^{-1}\mathrm{d}\bar{W}\right)$$

$$\rho_D(W_1,W_2) = \log\left(\frac{1+\|\Phi_{W_1}(W_2)\|_O}{1-\|\Phi_{W_1}(W_2)\|_O}\right)$$

$$\Phi_{W_1}(W_2) = (I-W_1\bar{W}_1)^{-\frac{1}{2}}(W_2-W_1)(I-\bar{W}_1W_2)^{-1}(I-\bar{W}_1W_1)^{\frac{1}{2}}$$

Some applications of Siegel symplectic geometry

- <u>Radar signal processing</u>:

- Frederic Barbaresco. **Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger bration and Frechet median**.

In Matrix information geometry, pages 199-255. Springer, 2013.

- Ben Jeuris and Raf Vandebril. **The Kahler mean of block-Toeplitz matrices with Toeplitz structured blocks.**

SIAM Journal on Matrix Analysis and Applications, 37(3):1151-1175, 2016.

- Congwen Liu and Jiajia Si. **Positive Toeplitz operators on the Bergman spaces of the Siegel upper half-space.** Communications in Mathematics and Statistics, pages 1-22, 2019.

- <u>Image processing</u>:

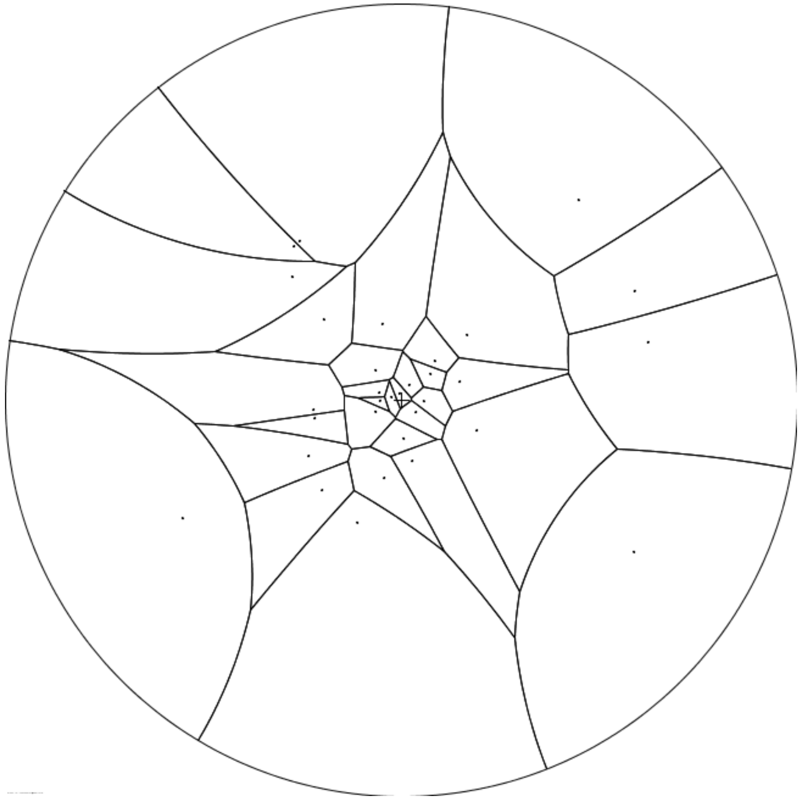
Reiner Lenz. **Siegel descriptors for image processing**. IEEE Signal Processing Letters, 23(5):625-628, 2016.

- <u>Statistics</u>:

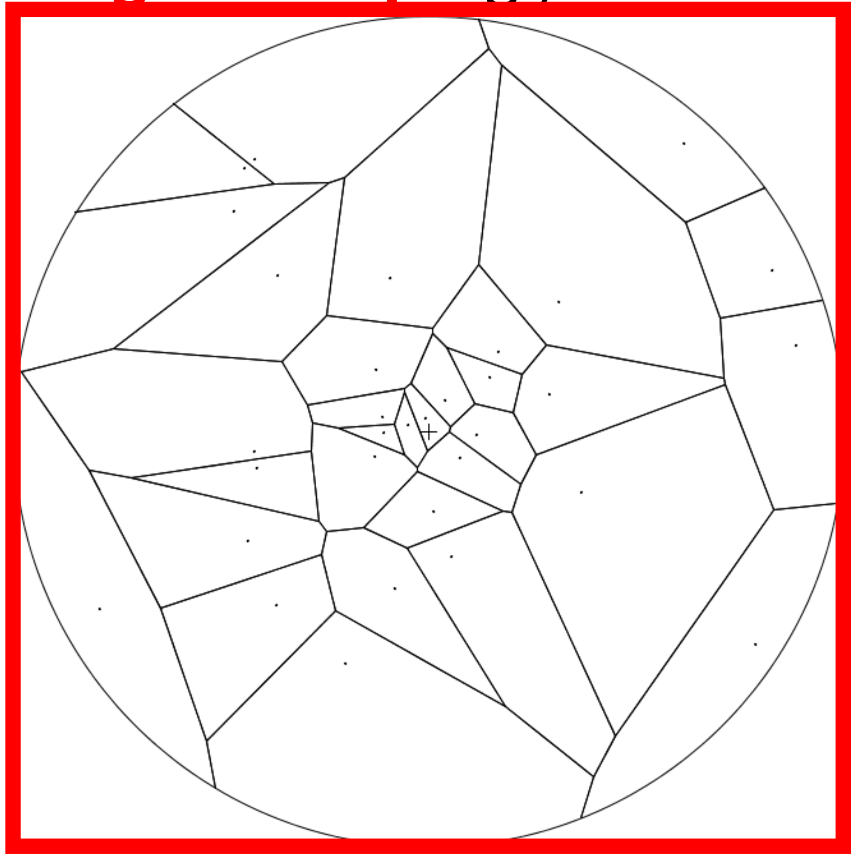
- Miquel Calvo and Josep M Oller. **A distance between elliptical distributions based in an embedding into the Siegel group**. Journal of Computational and Applied Mathematics, 145(2):319-334, 2002.
- Emmanuel Chevallier, Thibault Forget, Frederic Barbaresco, and Jesus Angulo. **Kernel density estimation on the Siegel space with an application to radar processing**. Entropy, 18(11):396, 2016.

Poincaré conformal disk vs Klein non-conformal disk

- Klein disk is **non-conformal** with **geodesics straight** Euclidean lines
- Klein mode well-suited for **computational geometry**: Eg., Voronoi diagram



Hyperbolic Voronoi diagram



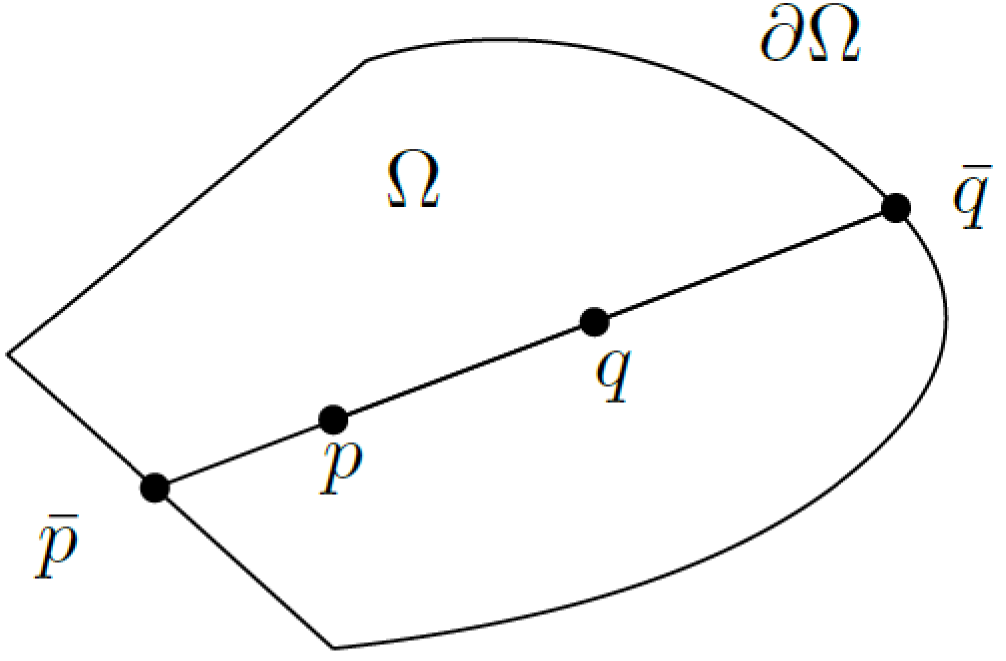
Clipped
affine diagram
(power diagram)

Q: What is the equivalent of Klein geometry for the Siegel disk domain?

Hilbert (projective) geometry

Normed vector space $(V, \|\cdot\|)$

Bounded open convex domain Ω



Define **Hilbert distance**:

$$H_{\Omega,\kappa}(p,q) := \begin{cases} \kappa \log |\mathrm{CR}(\bar{p},p;q,\bar{q})|, & p \neq q, \\ 0 & p = q. \end{cases}$$

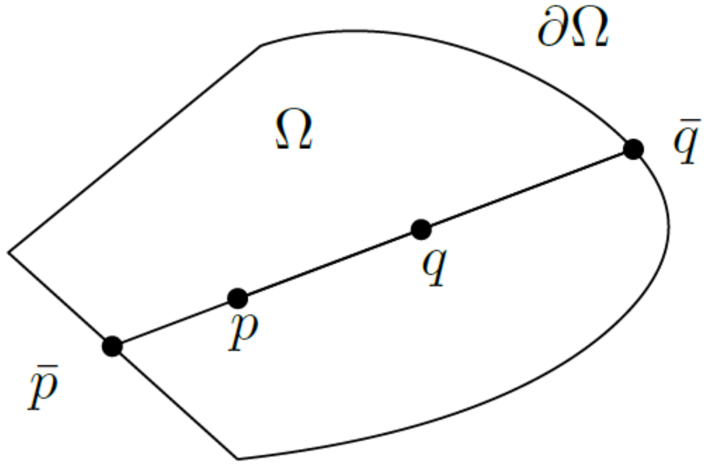
$$\boxed{H_{\Omega,\kappa}(p,q) := \kappa \log \frac{\|\bar{q}-p\|\,\|\bar{p}-q\|}{\|\bar{q}-q\|\|\bar{p}-p\|}}$$

Cross-ratio:

$$\mathrm{CR}(a,b;c,d) = \frac{\|a-c\|\|b-d\|}{\|a-d\|\|b-c\|}.$$

Related to Birkhoff geometry on (d+1)-dimensional cones

Rewriting the Hilbert distance



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

$$H_{\Omega,\kappa}(p,q) = \begin{cases} \kappa \log \left| \frac{\alpha_+(1-\alpha_-)}{\alpha_-(\alpha_+-1)} \right|, & p \neq q, \\ 0 & p = q. \end{cases} \qquad \begin{array}{l} \bar{p} = p + \alpha^-(q-p) \\ \bar{q} = p + \alpha^+(q-p) \end{array}$$

Or equivalently (p,q expressed from linear interpolations of boundary points):

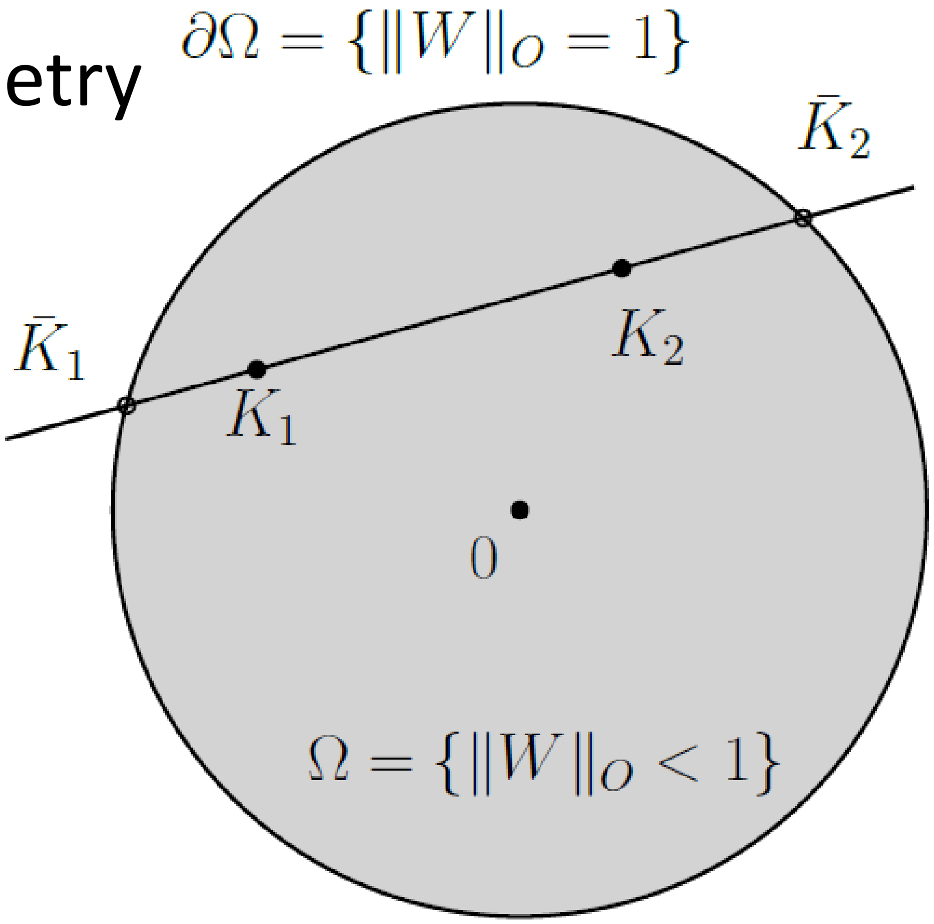
$$H_{\Omega,\kappa}(p,q) = \begin{cases} \kappa \log \left(\frac{1-\alpha_p}{\alpha_p} \frac{\alpha_q}{1-\alpha_q} \right) & \alpha_p \neq \alpha_q, \\ 0 & \alpha_p = \alpha_q. \end{cases} \qquad \begin{array}{l} p = (1-\alpha_p)\bar{p} + \alpha_p \bar{q} \\ q = (1-\alpha_q)\bar{p} + \alpha_q \bar{q} \end{array}$$

Siegel-Klein disk model

$$\mathbb{SD}(d) = \{W \in \mathrm{Sym}(d,\mathbb{C}) \ : \ \|W\|_O < 1\}$$

Definition 2 (Siegel-Klein geometry) *The Siegel-Klein disk model is the Hilbert geometry for the open bounded convex domain $\Omega = \mathbb{SD}(d)$ with constant $\kappa = \frac{1}{2}$. The Siegel-Klein distance is $\rho_K(K_1, K_2) := H_{\mathbb{SD}(d),\frac{1}{2}}(K_1, K_2)$.*

Choose **constant ½** to match Klein disk geometry



In complex dimension 1,
recover the Klein disk:

$$\rho_K(k_1, k_2) = \mathrm{arccosh}\left(\frac{1 - (\mathrm{Re}(k_1)\mathrm{Re}(k_2) + \mathrm{Im}(k_1)\mathrm{Im}(k_2))}{\sqrt{(1-|k_1|)(1-|k_2|)}}\right)$$

$$\mathrm{arccosh}(x) = \log\left(x + \sqrt{x^2 - 1}\right)$$

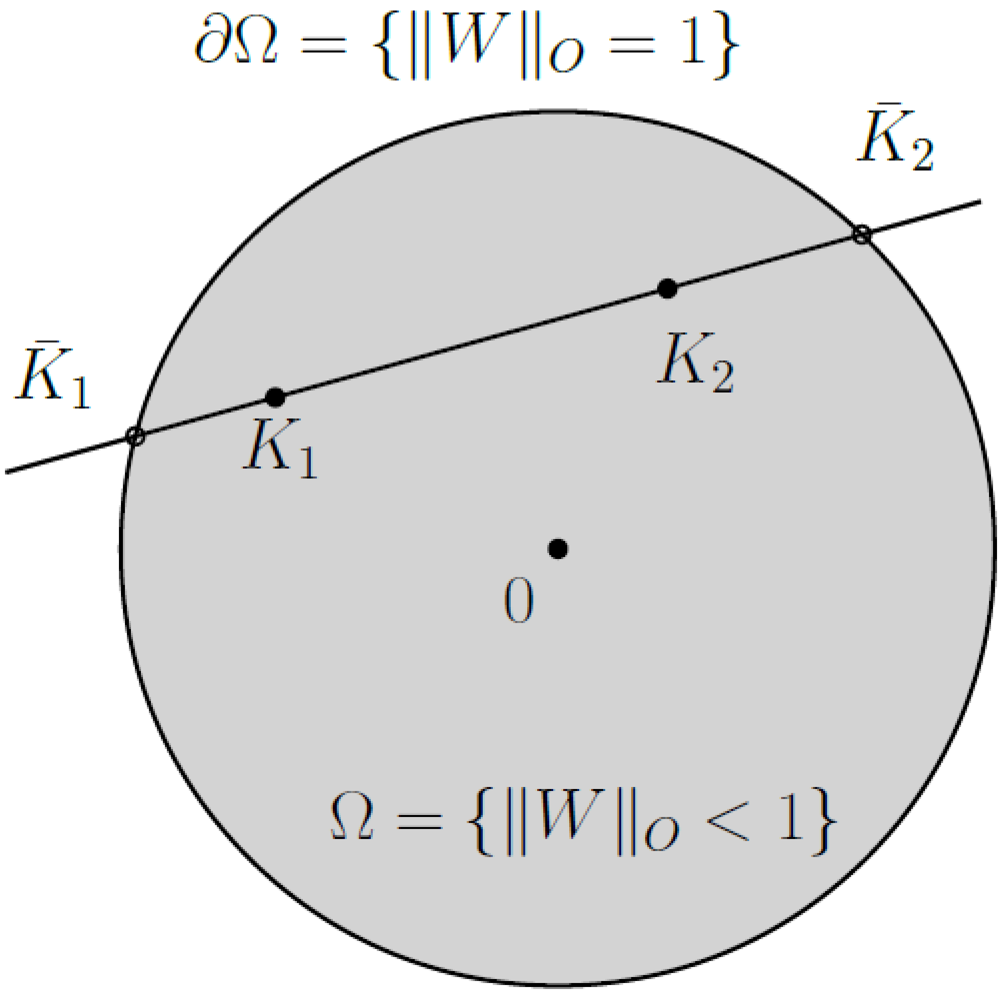
Calculating the Siegel-Klein distance

Line passing through two matrix points:

$$\{K_1 + \alpha(K_2 - K_1), \alpha \in \mathbb{R}\}$$

Calculate the **two α values** on Shilov boundary

$$\|K_1 + \alpha(K_2 - K_1)\|_O = 1.$$



Siegel-Klein distance:

$$\boxed{\begin{aligned} &\rho_K(K_1, K_2) = \frac{1}{2}\log\left(\frac{\alpha_+(1-\alpha_-)}{|\alpha_-|(\alpha_+ - 1)}\right) \\ &\bar{K}_1 = K_1 + \alpha_-(K_2 - K_1) \qquad \alpha_+ > 1 \\ &\bar{K}_2 = K_1 + \alpha_+(K_2 - K_1) \qquad \alpha_- < 0 \end{aligned}}$$

In practice, perform **bisection search** for the α values…

Siegel-Klein distance to the origin (zero matrix 0)

Special case I

Solve for $\|\alpha K\|_O = 1$

$$\alpha_+ = \frac{1}{\|K\|_O} > 1 \quad \text{and} \quad \alpha_- = -\frac{1}{\|K\|_O} < 0$$

$$\begin{aligned}\rho_K(0,K) &= \log\left(\frac{1+\frac{1}{\|K\|_O}}{\frac{1}{\|K\|_O}-1}\right),\\ &= \frac{1}{2}\log\left(\frac{1+\|K\|_O}{1-\|K\|_O}\right)\\ &= 2\,\rho_D(0,K),\end{aligned}$$

Siegel disk distance:

$$\rho_D(0,W) = \log\left(\frac{1+\|W\|_O}{1-\|W\|_O}\right)$$

Theorem 1 (Siegel-Klein distance to the origin) *The Siegel-Klein distance of matrix $K \in \mathbb{SD}(d)$ to the origin O is*

$$\boxed{\rho_K(0,K) = \frac{1}{2}\log\left(\frac{1+\|K\|_O}{1-\|K\|_O}\right)} \qquad (123)$$

Exact

Siegel-Klein distance: Line passing through the origin

Special case II

Line (K1K2) passing through the origin: $K_2 = \lambda K_1$ $\boxed{\lambda = \frac{\mathrm{tr}(K_2)}{\mathrm{tr}(K_1)}}$

$$\begin{aligned} \|K_1 + \alpha(K_2 - K_1)\|_O &= 1, \\ |1 + \alpha(\lambda - 1)| &= \frac{1}{\|K_1\|_O} \end{aligned} \Longrightarrow \begin{aligned} \alpha' &= \frac{1}{\lambda - 1}\left(\frac{1}{\|K_1\|_O} - 1\right) \\ \alpha'' &= \frac{1}{1-\lambda}\left(1 + \frac{1}{\|K_1\|_O}\right) \end{aligned}$$

Exact

Siegel-Klein distance:

$$\boxed{\begin{aligned} \rho_K(K_1, K_2) &= \frac{1}{2}\left|\log\left(\frac{\alpha'(1-\alpha'')}{\alpha''(\alpha'-1)}\right)\right|, \\ &= \frac{1}{2}\left|\log \frac{1 - \|K_1\|_O}{1 + \|K_1\|_O} \frac{\|K_1\|_O(1-\lambda) - (1 + \|K_1\|_O)}{\|K_1\|_O(\lambda - 1) - (1 - \|K_1\|_O)}\right| \end{aligned}}$$

Siegel-Klein distance between diagonal matrices

Special case III

Theorem 4 (Siegel-Klein distance for diagonal matrices) *The Siegel-Klein distance between two diagonal matrices in the Siegel-Klein disk can be calculated exactly in linear time.*

Solve **d quadratic systems** for getting two α values:

$$\alpha^2\left(\bar{k}_i' - \bar{k}_i\right)\left(k_i' - k_i\right) + \alpha\left(\bar{k}_i(k_i' - k_i) + k_i(\bar{k}_i' - \bar{k}_i)\right) + \bar{k}_i k_i - 1 \leq 0, \forall i \in \{1, \ldots, d\}.$$

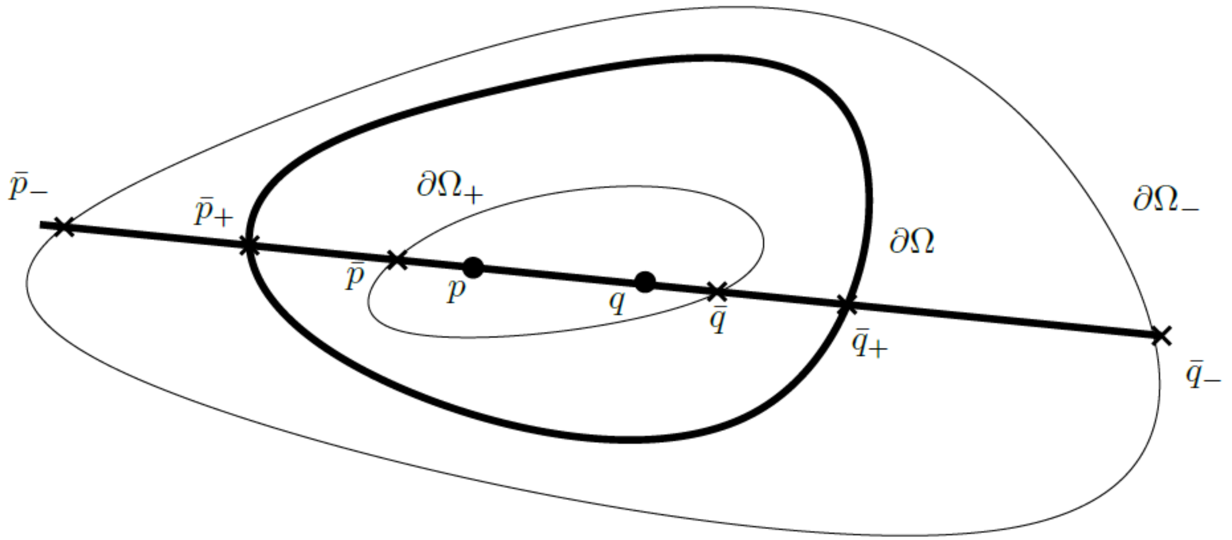
Siegel-Klein distance:

$$\boxed{\begin{aligned}\rho_K(K_1, K_2) &= \frac{1}{2}\log\left(\frac{\alpha_+(1-\alpha_-)}{|\alpha_-|(\alpha_+ - 1)}\right)\\ \alpha_- &= \max_{i\in\{1,\ldots,d\}} \alpha_i^-,\\ \alpha_+ &= \min_{i\in\{1,\ldots,d\}} \alpha_i^+,\end{aligned}}$$

Exact

Approximating Hilbert geometry with nested domains

$H_{\Omega_+,\kappa}(p,q) \geq H_{\Omega,\kappa}(p,q) \geq H_{\Omega_-,\kappa}(p,q)$

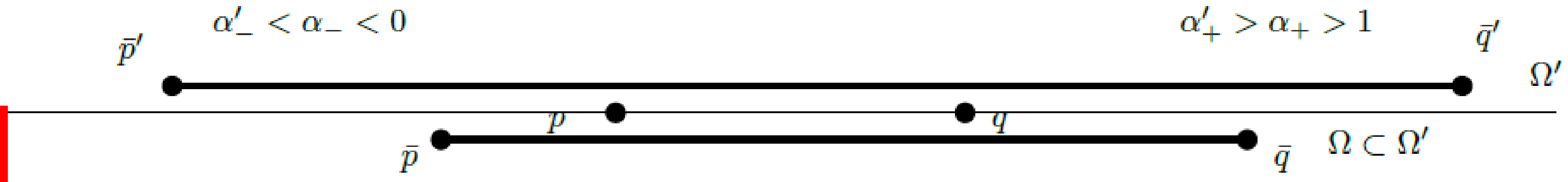


Property 1 (Bounding Hilbert distance) *Let $\Omega_+ \subset \Omega \subset \Omega_-$ be strictly nested open convex bounded domains. Then we have the following inequality for the corresponding Hilbert distances:*

$$\boxed{H_{\Omega_+,\kappa}(p,q) \geq H_{\Omega,\kappa}(p,q) \geq H_{\Omega_-,\kappa}(p,q).} \tag{151}$$

Enough to check in 1D:

$$\boxed{H_\Omega(p,q) = H_{\Omega\cap(pq)}(p,q)}$$



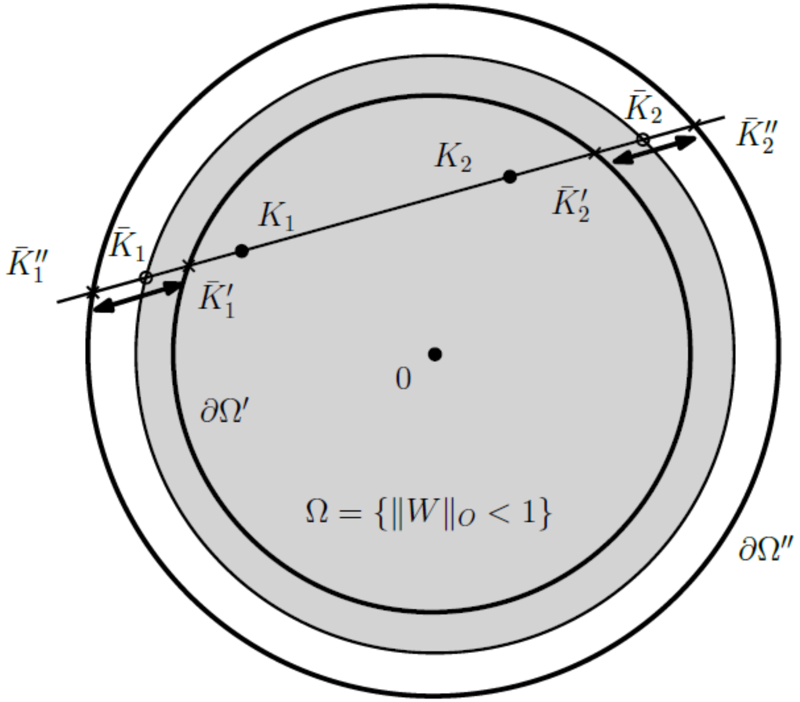
Guaranteed approximation of the Siegel-Klein distance

Theorem 5 (Lower and upper bounds on the Siegel-Klein distance) *The Siegel-Klein distance between two matrices K_1 and K_2 of the Siegel disk is bounded as follows:*

$$\boxed{\rho_K(l_-, u_+) \leq \rho_K(K_1, K_2) \leq \rho_K(u_-, l_+),} \tag{152}$$

where

$$\rho_K(\alpha_m, \alpha_M) := \frac{1}{2} \log \left(\frac{\alpha_M (1 - \alpha_m)}{|\alpha_m| (\alpha_M - 1)} \right). \tag{153}$$



$$\boxed{\begin{aligned} \bar{K}_1^+ &= K_1 + u_-(K_2 - K_1) \\ \bar{K}_1''+ &= K_1 + l_-(K_2 - K_1) \\ \bar{K}_2^+ &= K_1 + l_-(K_2 - K_1) \\ \bar{K}_2''+ &= K_1 + u_+(K_2 - K_1) \end{aligned}}$$

Converting Siegel-Poincaré (W) to/from Siegel-Klein (K)

Radial contraction to the origin

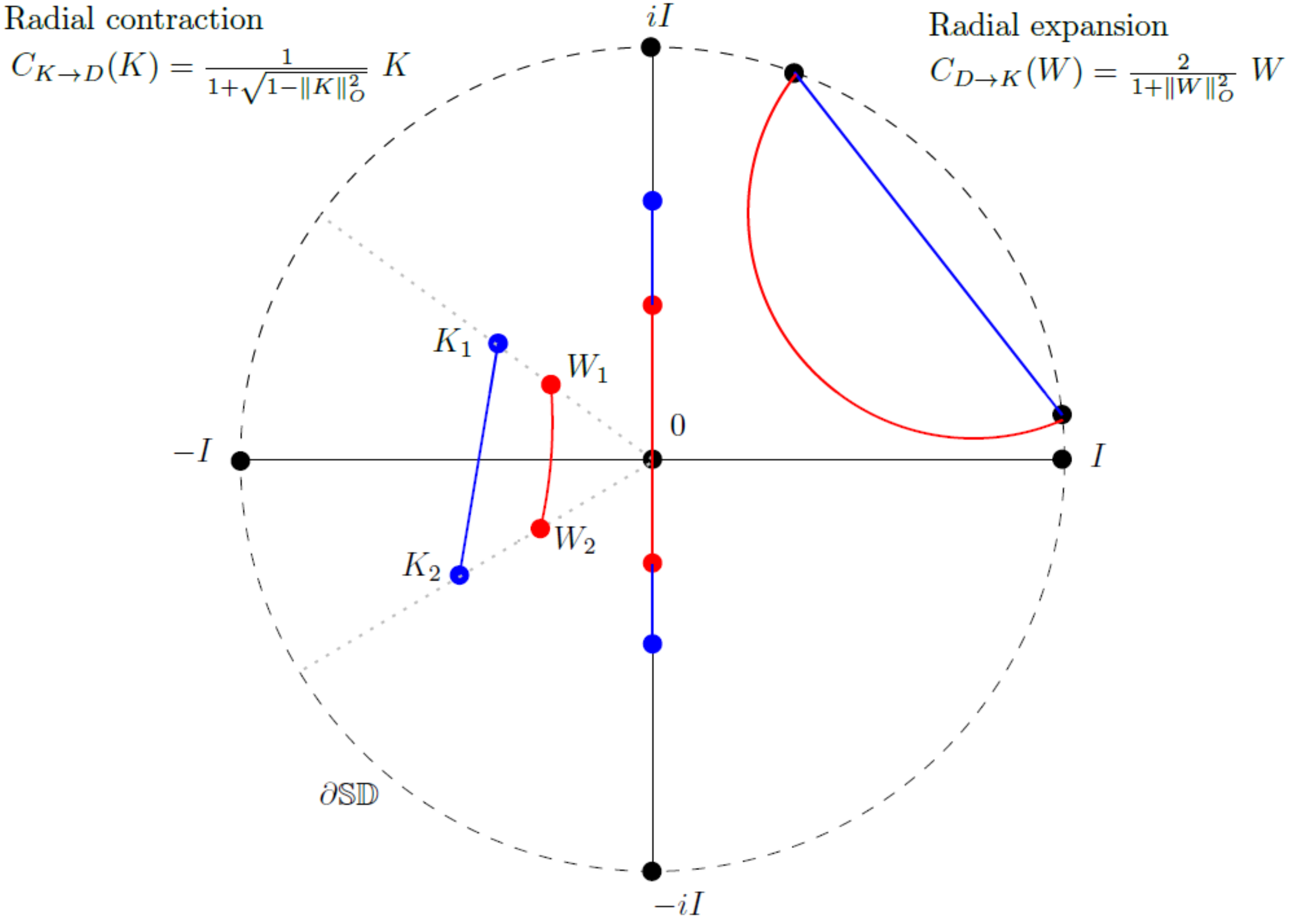
Siegel-Klein-> Siegel-Poincaré

$$C_{K\to D}(K) = \frac{1}{1+\sqrt{1-\|K\|_O^2}}\, K$$

Radial expansion to the origin:

Siegel-Poincaré-> Siegel-Klein-

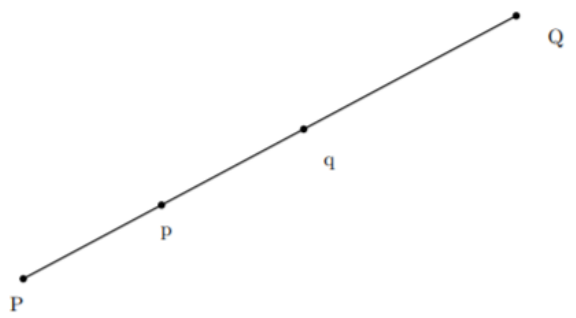
$$C_{D\to K}(W) = \frac{2}{1+\|W\|_O^2}\, W.$$



Siegel-Klein geodesics are unique Euclidean straight

$$\boxed{\gamma_{K_1,K_2}(\alpha) = (1-\alpha)K_1 + \alpha K_2 = K_1 + \alpha(K_2 - K_1).}$$

Follow from the definition of the **Hilbert distance** and the **cross-ratio properties**:



$$\boxed{(p,q;P,Q) = (p,r;P,Q) \times (r,q;P,Q) \text{ when } r \text{ is collinear with } p,q,P,Q}$$

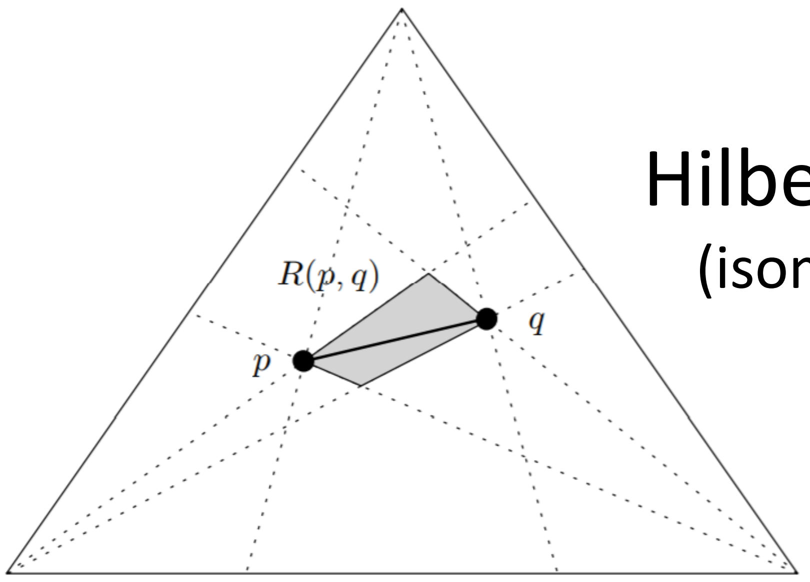
$$(p,q;P,Q) = \frac{(p-P)(q-Q)}{(p-Q)(q-P)}$$

Main advantage of the Siegel-Klein model is that **geodesics are straight**

Many <u>*computational geometric techniques*</u> *thus* apply:

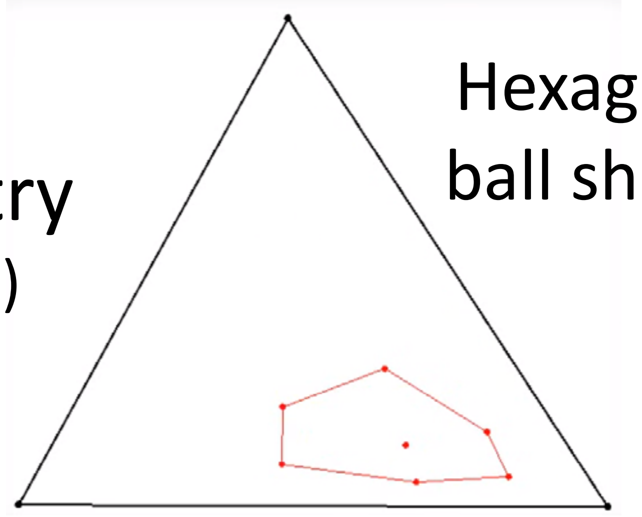
For example: Smallest Enclosing Balls, etc.

Geodesics in Hilbert geometry may not be unique



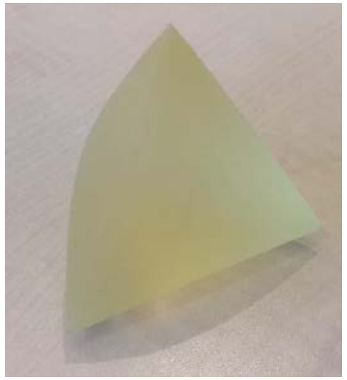
$$\rho_{\mathrm{HG}}(p,q) = \rho_{\mathrm{HG}}(p,r) + \rho_{\mathrm{HG}}(q,r)$$

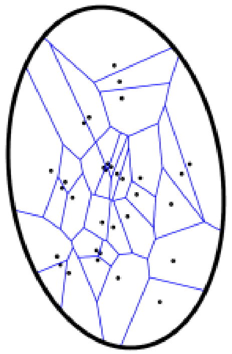
Hilbert **simplex** geometry
(isometric to a normed space)



Hexagonal ball shapes

https://www.youtube.com/watch?v=Gz0Vjk5quQE





Geodesics in Cayley-Klein geometry are unique.
(= Hilbert geometry for **ellipsoidal domains**)

Hilbert geometry of **elliptope**
(space of correlation matrices)
https://franknielsen.github.io/elliptope/index.html

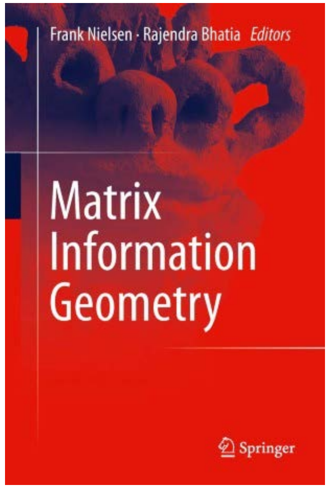
Summary of Siegel-Klein geometry:

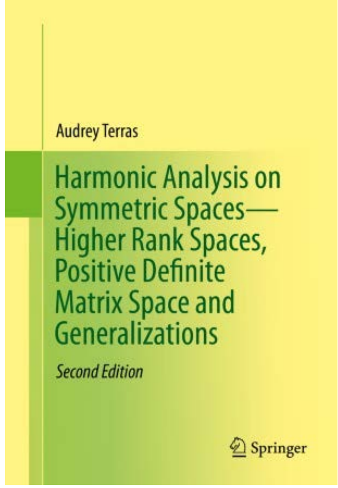
https://arxiv.org/abs/2004.08160

- Siegel and Hua studied in the 1940's the geometry of bounded complex matrix domains (= birth of *symplectic geometry* not directly related to symplectic manifolds equipped with a closed non-degenerate 2-form)
- The Siegel upper space *generalizes* the Poincaré upper plane, and the Siegel disk generalizes the Poincaré disk. Siegel upper space further *includes* in the cone of symmetric positive definite (SPD) matrices on the imaginary i-axis
- Orientation-preserving isometry group of the Siegel upper space is the projective real symplectic group. PSL(2,R) when complex dimension is 1. Orientation-preserving isometry group of the Siegel disk is the projective complex symplectic group. PSL(2,C) when complex dimension is 1.
- Hilbert geometry on the Siegel disk ensures **straight line geodesics**. Well-suited to computational geometry in the Siegel-Klein disk (eg, smallest enclosing ball)
- **Siegel-Klein distance** between two matrices can be calculated *exactly* when the line passing through the two matrices goes through the origin, or for diagonal matrices. Otherwise, **guaranteed approximations** of the Siegel-Klein distance by considering nested Hilbert geometries (require maximum singular values only).

Thank you!

https://arxiv.org/abs/2004.08160



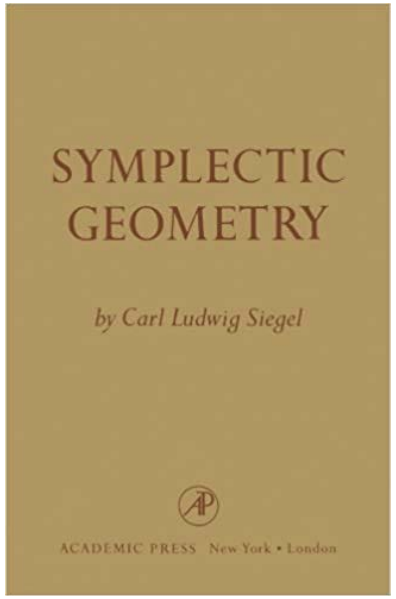


Henri Poincaré
1854–1912

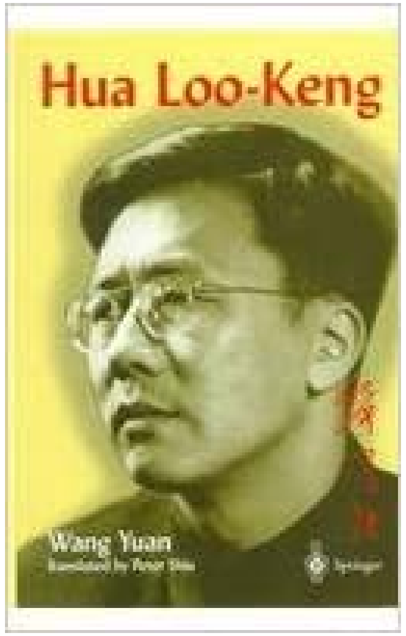
Felix Klein
1849 – 1925

David Hilbert
1862–1943

Carl Ludwig Siegel
1896 - 1981



Hua Luogeng Hua Loo-Keng
华罗庚
1910-1985

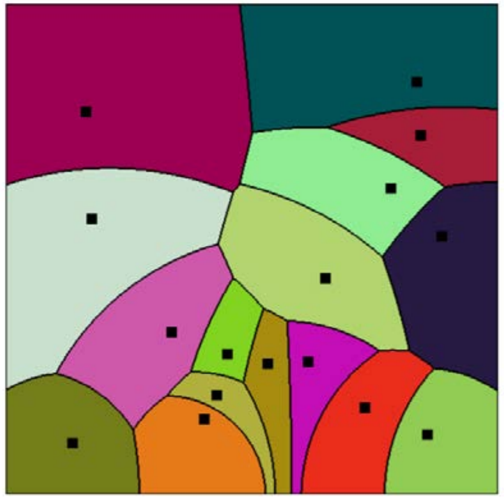


Some references:

Siegel-Klein geometry: https://arxiv.org/abs/2004.08160

- Carl Ludwig Siegel. *Symplectic geometry*. American Journal of Mathematics, 65(1):1-86, 1943.
- Loo-Keng Hua. *On the theory of automorphic functions of a matrix variable I: Geometrical basis*. American Journal of Mathematics, 66(3):470-488, 1944.
- Loo-Keng Hua. *Geometries of matrices. II. study of involutions in the geometry of symmetric matrices.* Transactions of the American Mathematical Society, 61(2):193-228, 1947.
- Frédéric Barbaresco. *Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Frechet median*. In Matrix information geometry, pages 199-255. Springer, 2013.
- Giovanni Bassanelli. *On horospheres and holomorphic endomorfisms of the Siegel disc.* Rendiconti del Seminario Matematico della Universita di Padova, 70:147-165, 1983.
- Pedro Jorge Freitas. *On the action of the symplectic group on the Siegel upper half plane*. PhD thesis, University of Illinois at Chicago, 1999.
- Nielsen, Frank, and Ke Sun. *Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices*. Geometric Structures of Information. Springer, Cham, 2019. 297-331.

On Voronoi Diagrams on the Information-Geometric Cauchy Manifolds

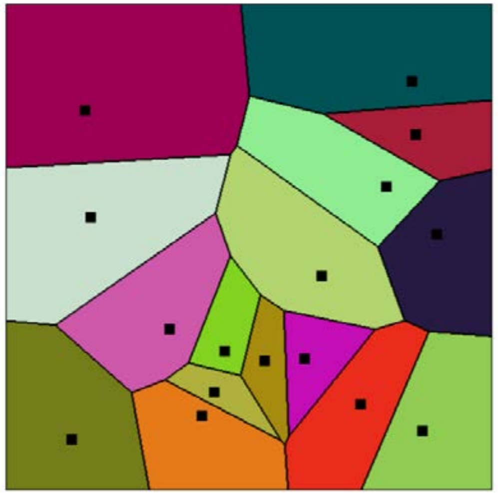


Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/



July 2020

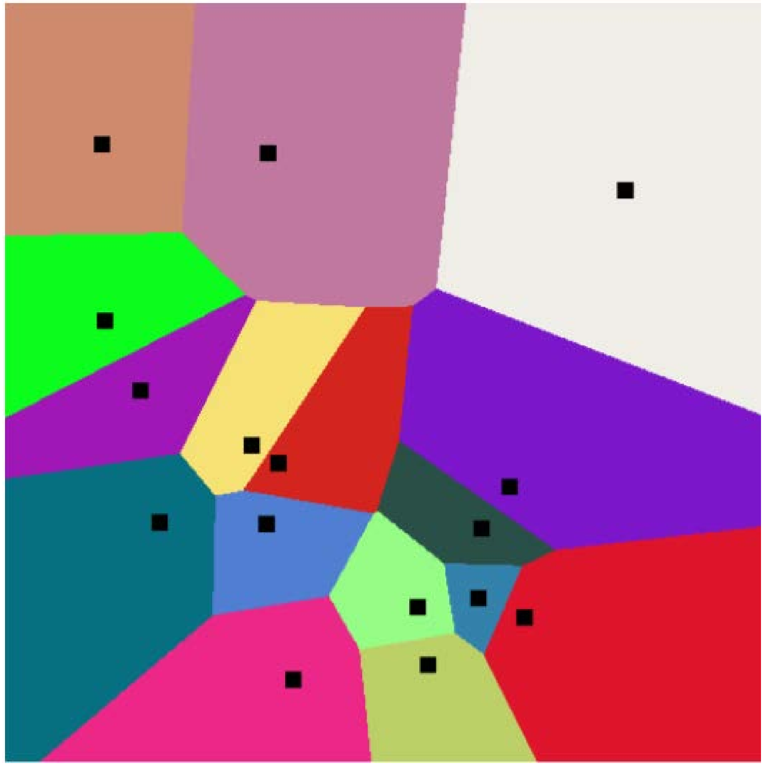
On Voronoi Diagrams on the Information-Geometric Cauchy Manifolds
Entropy 2020, 22(7), 713; https://doi.org/10.3390/e22070713
https://www.mdpi.com/1099-4300/22/7/713

Voronoi diagrams: Voronoi proximity cells

Given a finite point set $\mathcal{P} = \{P_1, \ldots, P_n\}$

Voronoi cell: $\boxed{\mathrm{Vor}_D(P_i) := \{X \in \mathbb{X}, \quad D(P_i, X) \leq D(P_j, X), \quad \forall j \in \{1, \ldots, n\}\}}$

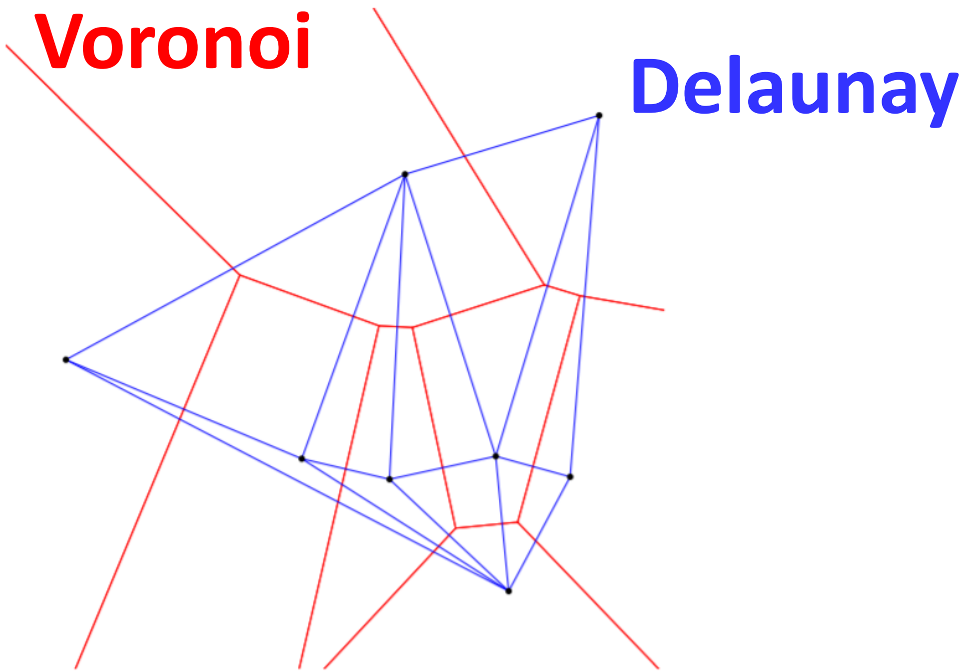
The Voronoi diagram partitions the space into Voronoi cells



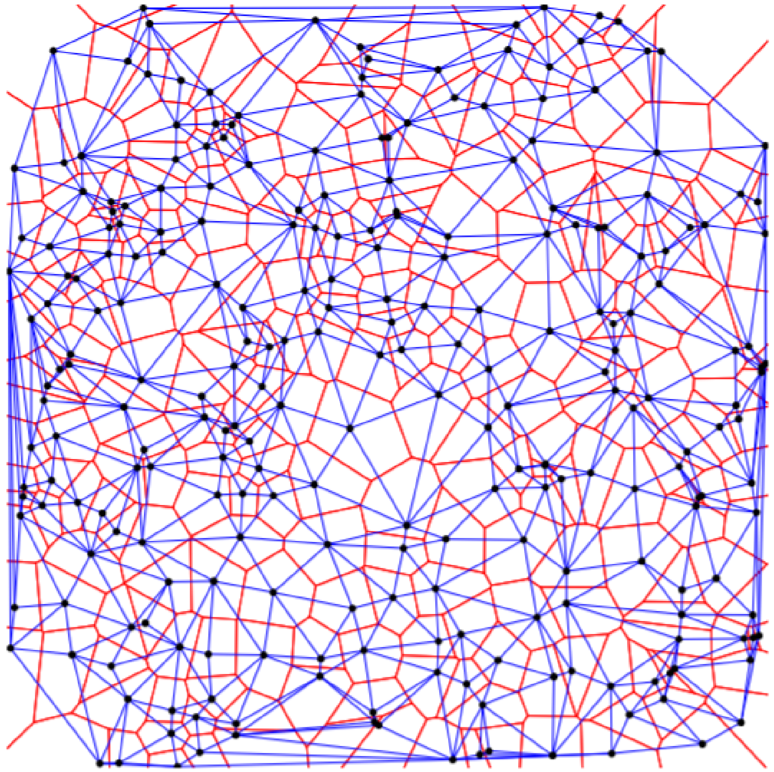
Euclidean distance (norm-induced): $\rho_E(P, Q) = \|p - q\|_2$

Dual Voronoi structure is the Delaunay complex

Link adjacent Voronoi generators by a straight (geodesic) edge:



Dual orthogonal structures



Delaunay complex yields the Delaunay triangulation

when no *d+2 cocircular* : nice meshing properties

Voronoi diagrams for asymmetric dissimilarities

Asymmetric (oriented) distance: $D(P,Q) \neq D(Q,P)$

Dual distance: $D^*(P,Q) := D(Q,P)$ Involution: $(D^*)^*(P,Q) = D(P,Q)$

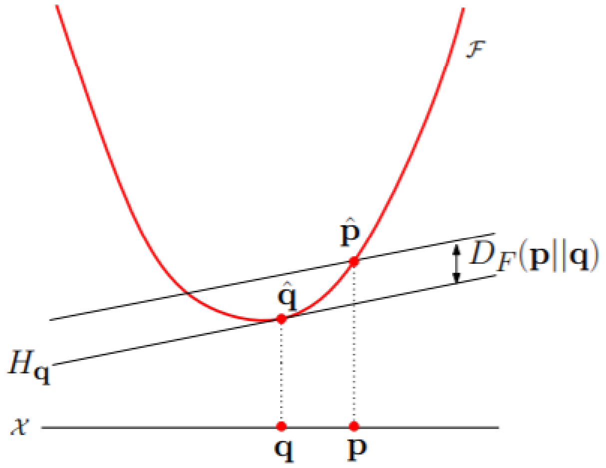
Dual Voronoi cells:

$$\mathrm{Vor}_D(P_i) := \{X \in \mathbb{X}, \quad D(P_i : X) \leq D(P_j : X), \quad \forall j \in \{1, \ldots, n\}\}$$

$$\begin{aligned}
\mathrm{Vor}^*_D(P_i) &:= \{X \in \mathbb{X} \quad D(X : P_i) \leq D(X : P_j), \quad \forall j \in \{1, \ldots, n\}\}, \\
&= \{X \in \mathbb{X} \quad D^*(P_i : X) \leq D^*(P_j : X), \quad \forall j \in \{1, \ldots, n\}\}, \\
&= \boxed{\mathrm{Vor}^*_D(P_i) = \mathrm{Vor}_{D^*}(P_i)}
\end{aligned}$$

= Dual bisector is primal bisector for dual dissimilarity

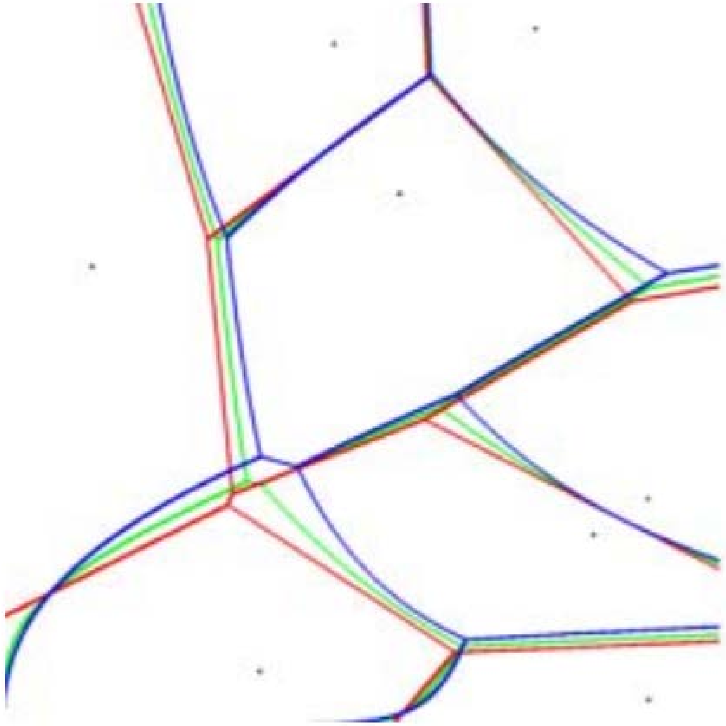
Example: Bregman Voronoi diagrams



Bregman divergence for a convex C2 generator F:

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

Recover the ordinary Euclidean Voronoi diagram when $F_{\mathrm{Eucl}}(\theta) = \frac{1}{2}\theta^\top \theta$



Three types of Voronoi diagrams:

Primal (curved)

Dual (always affine)

Symmetrized (curved)

Boissonnat, N, Nock. "Bregman Voronoi diagrams." *Discrete & Computational Geometry* 44.2 (2010): 281-307.

The Cauchy manifold

Manifold of the Cauchy distributions (Lorentzian distributions):

$$\mathcal{C} := \left\{ p_\lambda(x) := \frac{s}{\pi(s^2 + (x-l)^2)}, \quad \lambda := (l,s) \in \mathbb{H} := \mathbb{R} \times \mathbb{R}_+ \right\}$$

Location-scale family (l,s) with base *standard Cauchy distribution*:

$$p_{l,s}(x) := \frac{1}{s} p\left(\frac{x-l}{s}\right) \qquad p(x) := \frac{1}{\pi(1+x^2)} =: p_{0,1}(x)$$

Several **kinds of manifold information-geometric structures** induced by:

1. **Fisher-Rao geometry**: Fisher information metric (+ Levi-Civita metric connection)
2. **α-geometry**: Dualistic structure (Amari-Chentsov cubic tensor T), alpha connections
3. **D-geometry**: Dualistic geometry from divergence (e.g., Kullback-Leibler divergence)
4. **Hessian geometry** from Hessian metrics (smooth flat divergence + conformal flattening)

Cauchy manifold: Fisher-Rao Riemannian geometry

Fisher information matrix (FIM) yielding **Fisher Riemannian metric (FIm)**:

$$\boxed{g_{\mathrm{FR}}(\lambda) = [g_{ij}^{\mathrm{FR}}(\lambda)], \quad g_{ij}^{\mathrm{FR}}(\lambda) := E_{p_\lambda}\left[\partial_i l_\lambda(x) \partial_j l_\lambda(x)\right]}$$

$$g_{\mathrm{FR}}(\lambda) = g_{\mathrm{FR}}(l,s) = \frac{1}{2s^2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Scaled hyperbolic Poincaré upper plane metric

$$g_P(x,y) = \frac{1}{y^2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathrm{d}s_{\mathrm{FR}} = \frac{1}{\sqrt{2}}\mathrm{d}s_P$$

Fisher-Rao distance is a geodesic length and metric distance:

$$\rho_{\mathrm{FR}}\left(p_{\lambda_1}(x), p_{\lambda_2}(x)\right) = \min_{\substack{\lambda(s) \\ \text{such that} \\ \lambda(0)=\lambda_1, \lambda(1)=\lambda_2}} \int_0^1 \sqrt{\left(\frac{\mathrm{d}\lambda(t)}{\mathrm{d}t}\right)^T g_{\mathrm{FR}}(\lambda(s)) \frac{\mathrm{d}\lambda(t)}{\mathrm{d}t}} \mathrm{d}t$$

$$\rho_{\mathrm{FR}}[p_{l_1,s_1}, p_{l_2,s_2}] = \frac{1}{\sqrt{2}}\rho_P(l_1,s_1;l_2,s_2) \quad \text{where} \quad \rho_P(l_1,s_1;l_2,s_2) := \operatorname{arccosh}\left(1 + \delta(l_1,s_1,l_2,s_2)\right)$$

$$\delta(l_1,s_1;l_2,s_2) := \frac{(l_2-l_1)^2 + (s_2-s_1)^2}{2s_1s_2} \qquad \operatorname{arccosh}(x) := \log\left(x + \sqrt{x^2-1}\right), \quad x > 1$$

Cauchy manifold: Rao's distance

Fisher-Rao distance between Cauchy distributions:

$$\rho_{\mathrm{FR}}[p_{l_1,s_1}, p_{l_2,s_2}] = \begin{cases} \frac{1}{\sqrt{2}} \left|\log \frac{s_1}{s_2}\right| & \textit{when } l_1 = l_2, \\ \frac{1}{\sqrt{2}} \operatorname{arccosh}\left(1 + \frac{(l_2-l_1)^2+(s_2-s_1)^2}{2s_1 s_2}\right) & \textit{when } l_1 \neq l_2. \end{cases}$$

Extended to *multidimensional "isotropic" location-scale families*:

$$\lambda = (l, s) \in \mathbb{R}^d \times \mathbb{R},$$

$$\rho_{\mathrm{FR}}[p_{l_1,s_1}, p_{l_2,s_2}] = \frac{1}{\sqrt{2}} \operatorname{arccosh}\left(1 + \Delta(l_1, s_1, l_2, s_2)\right)$$

$$\Delta(l_1, s_1, l_2, s_2) := \frac{\|l_2 - l_1\|_2^2 + (s_2 - s_1)^2}{2 s_1 s_2}$$

Cauchy manifold: Always curved self-dual structures!

Skewness cubic tensor (Amari-Chentsov totally symmetric tensor):

$$T_{ijk}(\theta) := E_{p_\lambda}\left[\partial_i l_\lambda(x) \partial_j l\lambda(x) \partial_k l\lambda(x)\right]$$

$$T_{\sigma(i)\sigma(j)\sigma(k)} = T_{ijk}$$

α-geometry: $(M, g_{\mathrm{FR}}, \nabla^{-\alpha}, \nabla^{\alpha})$

$$g_{\mathrm{FR}}(\lambda) = g_{\mathrm{FR}}(l,s) = \frac{1}{2s^2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

All α-geometries coincide with the Fisher-Rao geometry for the Cauchy manifold:

$$\begin{aligned} {}^{\alpha}\Gamma^{1}_{12} &= {}^{\alpha}\Gamma^{1}_{21} = {}^{\alpha}\Gamma^{2}_{22} = -\frac{1}{s}, \\ {}^{\alpha}\Gamma^{2}_{11} &= \frac{1}{s}. \end{aligned}$$

Scalar curvature: $\kappa = -2.$

Fisher-Rao geometry is 0-geometry : $(\mathcal{C}, g_{\mathrm{FR}}) = (\mathcal{C}, g_{\mathrm{FR}}, \nabla^0, \nabla^0)$

No way to choose α so that the α-geometry becomes dually flat

- For the Gaussian distributions, we can choose α=1 or α=-1
- For the t-Student distributions, we can choose: $\alpha = \pm\frac{k+5}{k-1}$

Cauchy manifold: q-Gaussians for q=2

q-Gaussians are maximum entropy distributions wrt Tsallis' q-entropy:

Tsallis' q-entropy: $$\boxed{T_q(p) := \frac{1}{q-1}\left(1 - \int_{-\infty}^{\infty} p^q(x)\mathrm{d}x\right), \quad q \neq 1.}$$

$$\lim_{q\to 1} T_q(p) = S(p) := -\int p(x)\log p(x)\mathrm{d}x,$$ **Shannon entropy**

Cauchy distributions are q-Gaussians for **q=2**:

MaxEnt distributions for Tsallis' quadratic entropy:

$$T_2(p) := 1 - \int_{-\infty}^{\infty} p^2(x)\mathrm{d}x.$$

Related to Onicescu's informational energy: $E(p) := \int_{-\infty}^{\infty} p^2(x)\mathrm{d}x$

Deformed q=2-exponential families

Deformed exponential function: $\exp_C(u) := \frac{1}{1-u}, \quad u \neq 1,$

Deformed reciprocal logarithm function: $\log_C(u) := 1 - \frac{1}{u}, \quad u \neq 0,$

Deformed 2-exponential families (= Cauchy family):

$$\boxed{p_\theta(x) = \exp_C(\theta^\top x - F(\theta))}$$

For Cauchy distributions,

we find:

$$\begin{aligned}
\log_C(p_\theta(x)) &= 1 - \frac{1}{s}\pi(s^2 + (x-l)^2) = 1 - \pi\left(s + \frac{(x-l)^2}{s}\right), \\
&=: \theta^\top t(x) - F(\theta), \\
&= \underbrace{\left(2\pi\frac{l}{s}\right)x + \left(-\frac{\pi}{s}\right)x^2}_{\theta^\top t(x)} - \underbrace{\left(\pi s + \pi\frac{l^2}{s} - 1\right)}_{F(\theta)}.
\end{aligned}$$

Cauchy 2-Gaussians: Canonical factorization

Natural parameters:

$$\theta(l,s) = \boxed{(\theta_1,\theta_2) = \left(2\pi\frac{l}{s}, -\frac{\pi}{s}\right)} \in \Theta = \mathbb{R}\times\mathbb{R}_-$$

Natural-to ordinary parameter conversion: $\lambda(\theta) = (l,s) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{\pi}{\theta_2}\right)$

Log-normalizer: $F(\theta(\lambda)) = \pi s + \pi\frac{l^2}{s} - 1 =: F_\lambda(\lambda),$

$$\boxed{F(\theta) = -\frac{\pi^2}{\theta_2} - \frac{\theta_1^2}{4\theta_2} - 1.}$$

Gradient of the log-normalizer:

yields **dual coordinate system** eta

$$\nabla F(\theta) = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{\pi^2}{\theta_2^2} + \frac{\theta_1^2}{4\theta_2^2} \end{bmatrix}$$

Cauchy manifold: Dually flat manifold

$$\boxed{D_{\mathrm{flat}}[p_{\lambda_1} : p_{\lambda_2}] := \frac{1}{\int p_{\lambda_2}^2(x)\mathrm{d}x}\left(\int \frac{p_{\lambda_2}^2(x)}{p_{\lambda_1}(x)}\mathrm{d}x - 1\right)}$$

$$= 2\pi s_2 \left(\frac{s_1^2 + s_2^2 + (l_1 - l_2)^2}{2 s_1 s_2} - 1\right),$$

$$= 2\pi s_2 \frac{(s_1 - s_2)^2 + (l_1 - l_2)^2}{2 s_1 s_2},$$

$$= 2\pi s_2 \delta(l_1, s_1, l_2, s_2),$$

$$\boxed{D_{\mathrm{flat}}[p_{\lambda_1} : p_{\lambda_2}] = B_F(\theta_1 : \theta_2)}$$

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

called the **<u>Bregman-Tsallis q=2-divergence</u>**

Dual potential functions of the Hessian structure

Dual to primal conversion: $\boxed{\theta(\eta) = \begin{bmatrix} \frac{2\pi\eta_1}{\sqrt{\eta_2-\eta_1^2}} \\ \frac{-\pi-}{\sqrt{\eta_2-\eta_1^2}} \end{bmatrix} := \nabla F^*(\eta)}$

Dual potential function:

$$F^*(\eta) := \theta(\eta)^\top \eta - F(\theta(\eta)) \qquad \boxed{F^*(\eta) = 1 - 2\pi\sqrt{\eta_2 - \eta_1^2}.}$$

Dual-to-ordinary parameter conversion:

$$\eta(\lambda) = \eta(\theta(\lambda)) = (\lambda_1, \lambda_1^2 + \lambda_2^2) = (l, l^2 + s^2).$$

$$F^*_\lambda(\lambda) := F^*(\eta(\lambda)) = 1 - 2\pi\sqrt{l^2 + s^2 - l^2} = 1 - 2\pi s$$

$$F^*_\lambda(\lambda) := 1 - \frac{1}{\int p^2(x)dx} = 1 - \frac{1}{\frac{1}{2\pi s}} = 1 - 2\pi s.$$

Dual-to-ordinary parameter conversion: $\lambda(\eta) = (l, s) = (\eta_1, \sqrt{\eta_2 - \eta_1^2}).$

Dually flat divergence (=Bregman divergence)

$$\boxed{D_{\mathrm{flat}}[p_{\lambda_1} : p_{\lambda_2}] = B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1) = A_F(\theta_1 : \eta_2) = A_{F^*}(\eta_2 : \theta_1)}$$

with the **Legendre-Fenchel divergence**:

(non-negativity from Young's inequality)

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

Dual Hessians of the potential functions:

Dual Hessian metrics

$$\nabla^2 F(\theta) = \begin{bmatrix} -\frac{1}{2\theta_2} & \frac{\theta_1}{2\theta_2^2} \\ \frac{\theta_1}{2\theta_2^2} & -\frac{\theta_1^2}{2\theta_2^2} - \frac{2\pi^2}{\theta_2^2} \end{bmatrix} =: g_F(\theta),$$

$$\nabla^2 F^*(\eta) = \begin{bmatrix} \frac{2}{\sqrt{\eta_2-\eta_1^2}} + \frac{2\eta_1^2}{(\eta_2-\eta_1^2)^{\frac{3}{2}}} & -\frac{\eta_1}{(\eta_2-\eta_1^2)^{\frac{3}{2}}} \\ -\frac{\eta_1}{(\eta_2-\eta_1^2)^{\frac{3}{2}}} & \frac{1}{2}(\eta_2-\eta_1^2)^{\frac{3}{2}} \end{bmatrix} =: g_F^*(\eta).$$

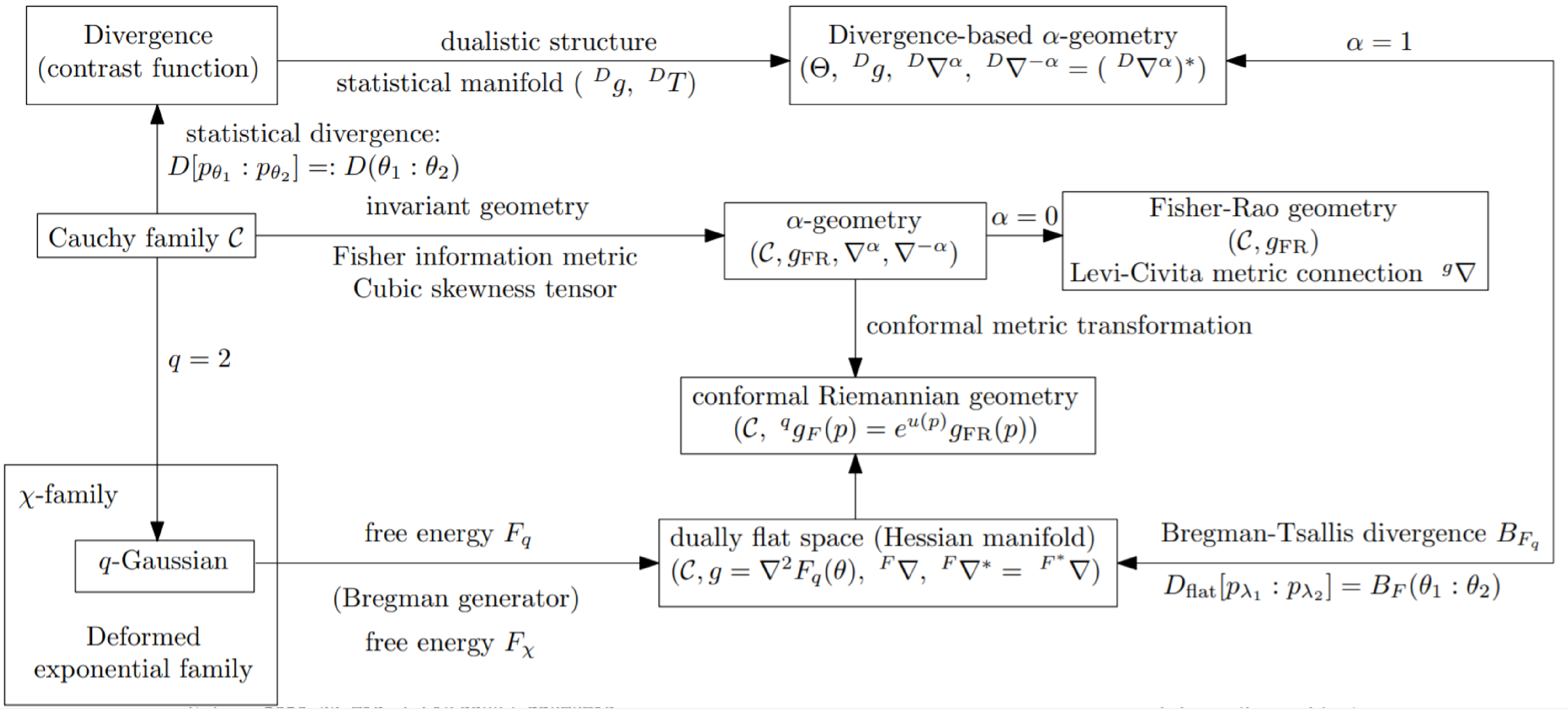
Crouzeix identity: $\boxed{\nabla^2 F(\theta)\nabla^2 F^*(\eta(\theta)) = \nabla^2 F(\theta(\eta))\nabla^2 F^*(\eta) = I,}$

Hessian metrics are conformal to the Fisher information metric:

$$g_F^\theta(\theta) = -\frac{2\theta_2}{\pi^2} g_{\mathrm{FR}}^\theta(\theta),$$

$$g_F^\lambda(\lambda) = \frac{2}{\pi\sigma} g_{\mathrm{FR}}^\lambda(\lambda).$$

Summary: Cauchy information-geometric structures:



Invariant f-divergences and α-divergences:

f-divergences:

$$\boxed{I_f[p:q] := \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x}$$

f convex, f(1)=0

Standard f-divergence: f'(1)=0, f''(1)=1

- Invariant because its satisfies the **information monotonicity,** and
- Infinitesimal small f-divergence is related to the **Fisher information** $^{I_f}g = g_{\mathrm{FR}}$

α-divergences:

$$\boxed{\begin{aligned} I_\alpha[p:q] &:= \frac{1}{\alpha(1-\alpha)}(1 - C_\alpha[p:q]), \quad \alpha \notin \{0,1\} \\ I_\alpha[p:q] &= I_{1-\alpha}[q:p] = I_\alpha{}^*[p:q]. \end{aligned}}$$

Chernoff α-coefficient: $C_\alpha[p:q] \quad := \quad \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}x$

α-divergences are f-divergences: $I_f[p:q] := \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x,$

$$f_\alpha(u) = \begin{cases} \frac{u^{1-\alpha}-u}{\alpha(\alpha-1)}, & \text{if } \alpha \neq 0, \alpha \neq 1 \\ u \log(u), & \text{if } \alpha = 0 \quad \text{(reverse Kullback-Leibler divergence)}, \\ -\log(u), & \text{if } \alpha = 1 \quad \text{(Kullback-Leibler divergence)}. \end{cases}$$

Kullback-Leibler divergence:
(relative entropy)

$$D_{\mathrm{KL}}[p:q] := \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$

Kullback-Leibler divergence between Cauchy distributions is **symmetric**:

$$\boxed{D_{\mathrm{KL}}[p_{l_1,s_1} : p_{l_2,s_2}] = \log\left(1 + \frac{(s_1 - s_2)^2 + (l_1 - l_2)^2}{4 s_1 s_2}\right)}$$

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

Fisher-Rao distance and chi-squared divergences:

$$
\begin{aligned}
D_{\chi^2_P}[p:q] &:= \int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}x,\\
D_{\chi^2_N}[p:q] &:= \int \frac{(q(x)-p(x))^2}{q(x)}\mathrm{d}x = D^*_{\chi^2_P}[p:q] = D_{\chi^2_P}[q:p]\\
D_{\chi^2_P}[p_{l_1,s_1}:p_{l_2,s_2}] &= D_{\chi^2_N}[p_{l_1,s_1}:p_{l_2,s_2}],\\
&= \frac{(s_1-s_2)^2+(l_2-l_1)^2}{2s_1s_2},\\
&=: \delta(l_1,s_1;l_2,s_2).
\end{aligned}
$$

$$
\boxed{\rho_{\mathrm{FR}}[p_{l_1,s_1},p_{l_2,s_2}] = \frac{1}{\sqrt{2}}\operatorname{arccosh}\left(1+D_{\chi^2}[p_{l_1,s_1}:p_{l_2,s_2}]\right)}
$$

Fisher-Rao distance is a metric distance

Square-root metrization of the KL divergence

Theorem 3. *The square root of the Kullback-Leibler divergence between two Cauchy density p_{l_1,s_1} and p_{l_2,s_2} is a metric distance:*

$$\rho_{\mathrm{KL}}[p_{l_1,s_1}, p_{l_2,s_2}] := \sqrt{D_{\mathrm{KL}}[p_{l_1,s_1} : p_{l_2,s_2}]} = \sqrt{\log\left(1 + \frac{(s_1 - s_2)^2 + (l_1 - l_2)^2}{4 s_1 s_2}\right)}. \tag{112}$$

The following function is a **metric transform** (and FR is metric distance):

$$\boxed{t_{\mathrm{FR}\to\mathrm{KL}}(u) := \log\left(\frac{1}{2} + \frac{1}{2}\cosh(\sqrt{2}u)\right)}$$

$$\cosh(x) := \frac{e^x + e^{-x}}{2}.$$

Scale family case: Hilbertian metric distance

Theorem 4. *The square root of the KL divergence between to Cauchy densities of the same scale family is a Hilbertian distance.*

$$D_{\mathrm{KL}}[p_{l,s_1} : p_{l,s_2}] = \log\left(\frac{(s_1+s_2)^2}{4s_1s_2}\right).$$

$$D_{\mathrm{KL}}[p_{l,s_1} : p_{l,s_2}] = 2\log\left(\frac{A(s_1,s_2)}{G(s_1,s_2)}\right)$$

$$= \|\phi(p)-\phi(q)\|_{H}.$$

Hilbertian norm

Arithmetic mean:

$$A(s_1,s_2) = \frac{s_1+s_2}{2}$$

Geometric mean:

$$G(s_1,s_2) = \sqrt{s_1s_2}$$

A-G inequality: A>=G

Cauchy hyperbolic Voronoi diagrams

Theorem 5. *The Cauchy Voronoi diagrams under the Fisher-Rao distance, the the chi-square divergence and the Kullback-Leibler divergence all coincide, and amount to a hyperbolic Voronoi diagram on the corresponding location-scale parameters.*

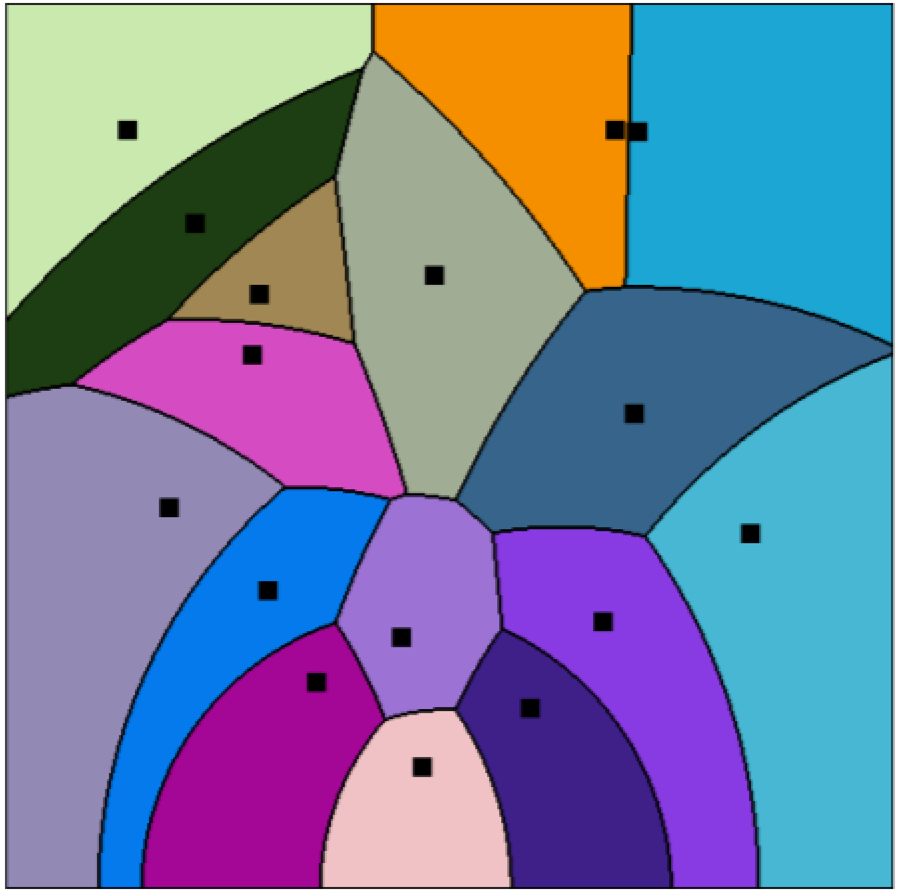
Voronoi bisectors are invariant under strictly monotonically increasing functions

Voronoi bisectors (dual bisectors coincide for symmetric distances):

$$\begin{aligned}\mathrm{Bi}_D(p_{\lambda_1} : p_{\lambda_2}) &= \{\lambda \in \mathbb{H} \,:\, \delta(\lambda, \lambda_1) = \delta(\lambda, \lambda_2)\},\\ \mathrm{Bi}_D(p_{l_1,s_1} : p_{l_2,s_2}) &= \{(l,s) \in \mathbb{H} \,:\, \delta(l,s,l_1,s_1) = \delta(l,s,l_2,s_2)\}.\end{aligned}$$

$$D \in \{\rho_{\mathrm{FR}}, D_{\mathrm{KL}}, \sqrt{D_{\mathrm{KL}}}, D_{\chi^2}\}$$

Cauchy hyperbolic Voronoi diagrams

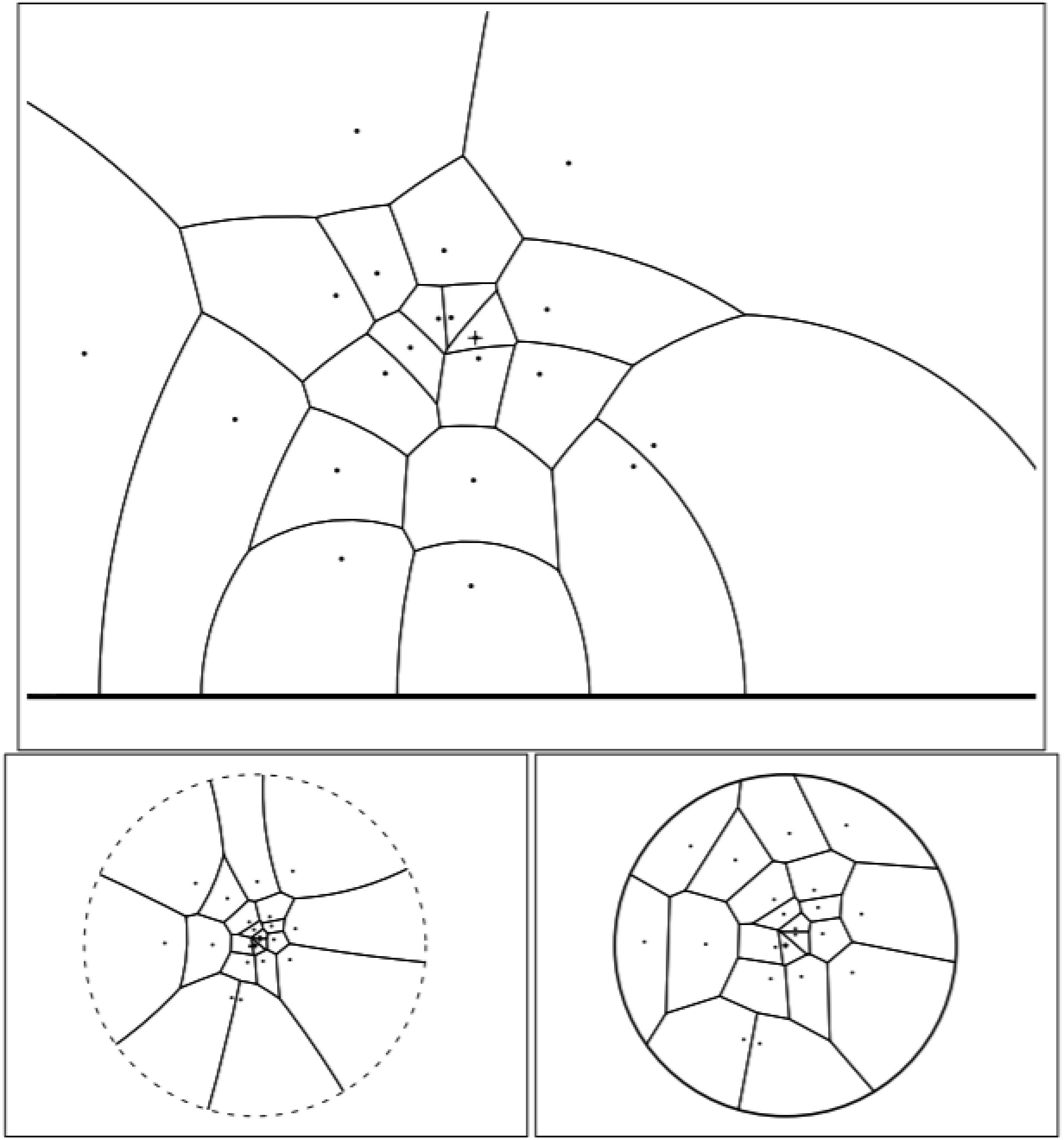


Poincaré conformal upper plane

Cauchy hyperbolic Voronoi diagrams

Several **models** of hyperbolic geometry:

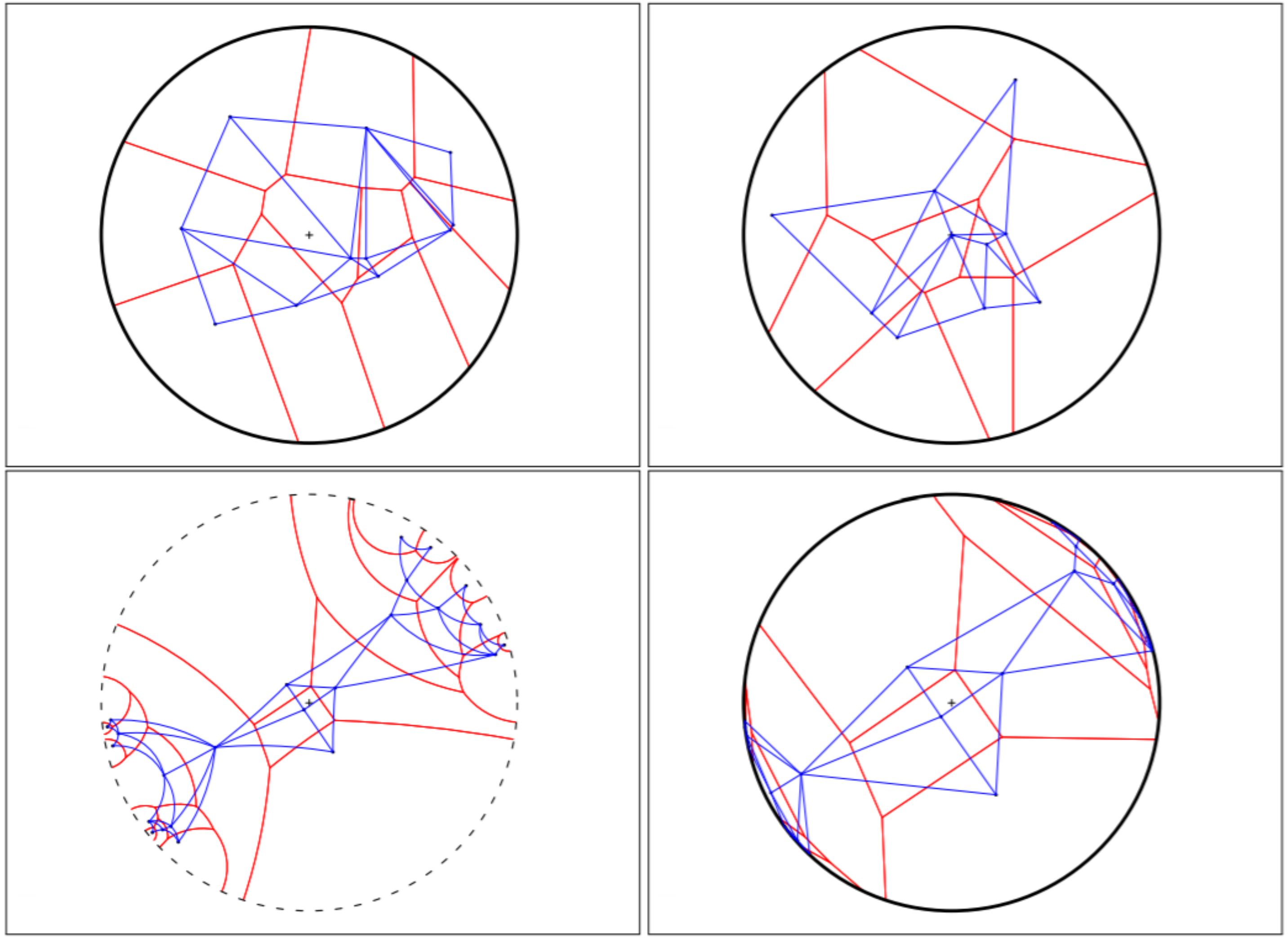
1. Poincaré conformal upper plane
2. Poincaré conformal disk
3. Klein **non-conformal** disk:



Cauchy hyperbolic Delaunay complex

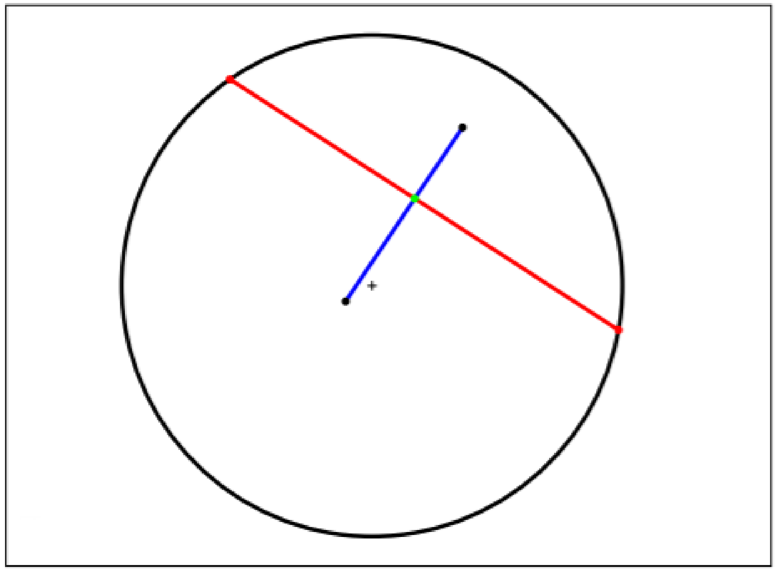
Dual Delaunay complex by **geodesically** linking adjacent Voronoi cells

Not necessarily a triangulation but a **simplicial complex**!

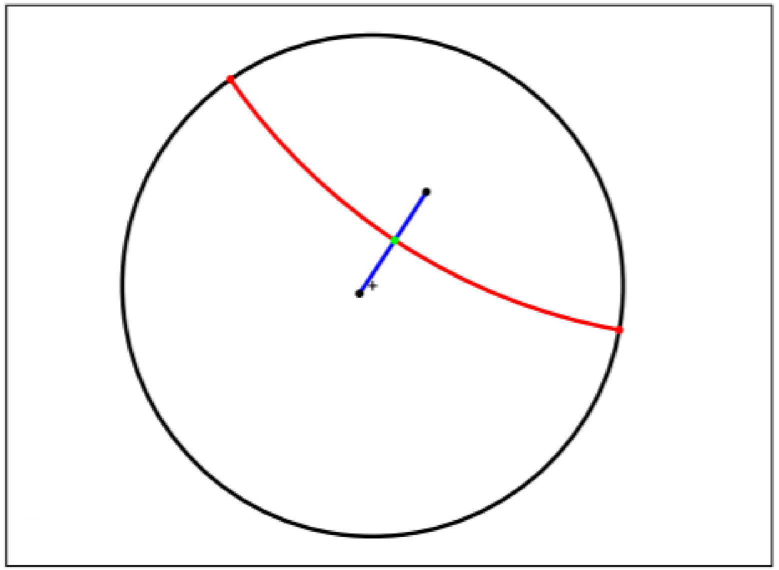


Hyperbolic geometry is often used in ML for embedding hierarchical structures

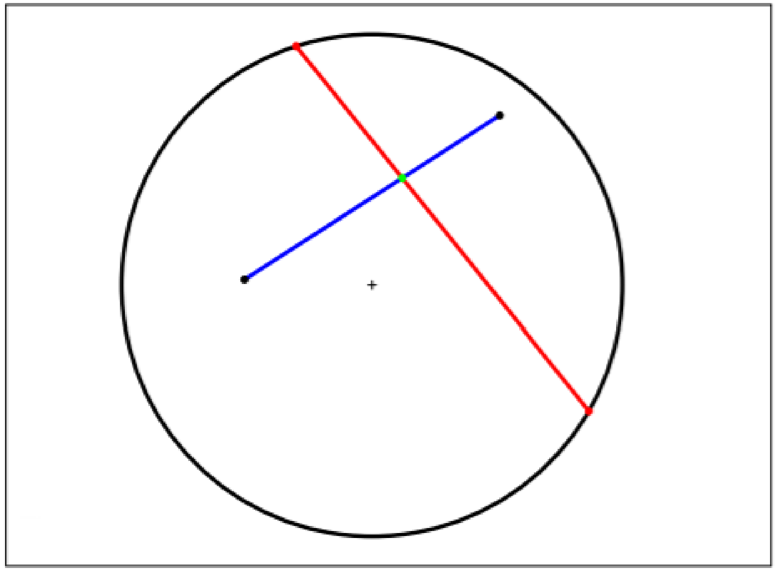
Hyperbolic Delaunay edges are orthogonal to Voronoi bisectors



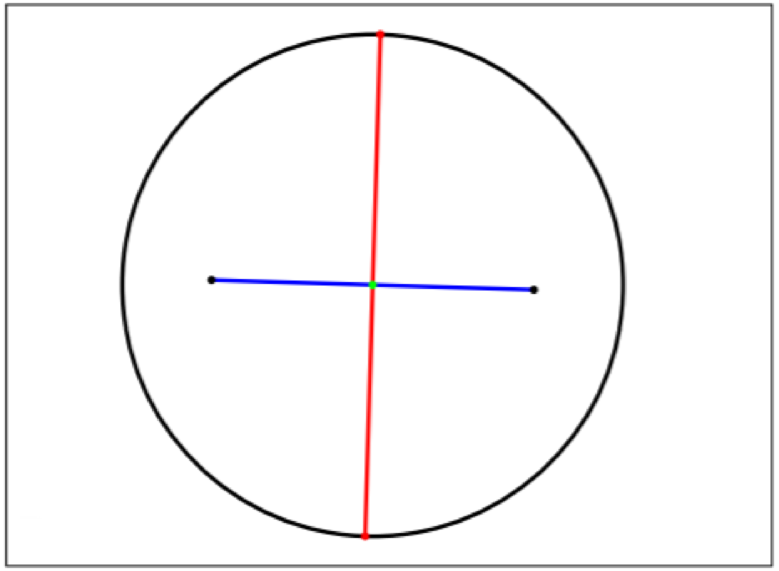
non-conformal (Klein)



conformal (Poincaré)

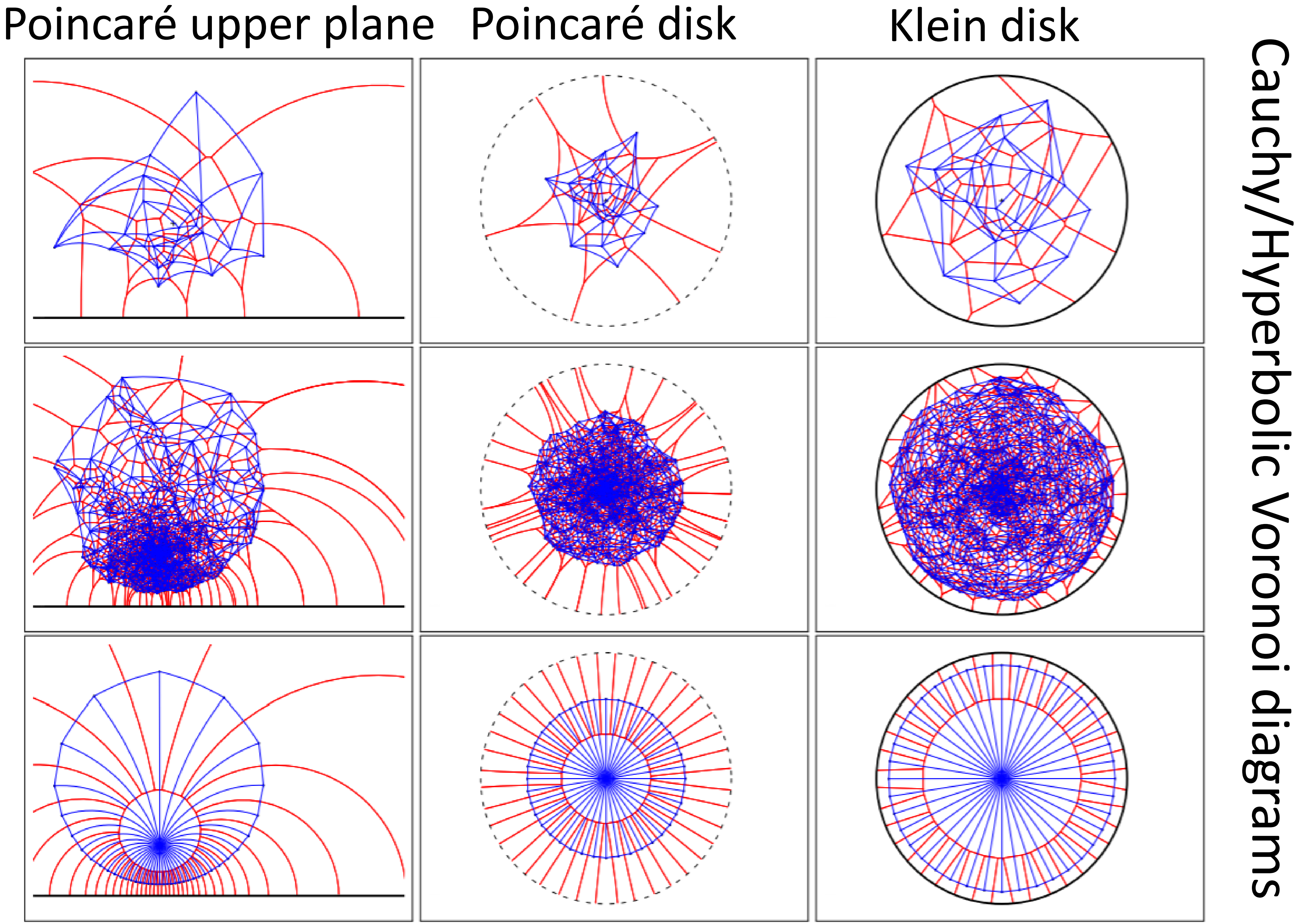


non-conformal (Klein)

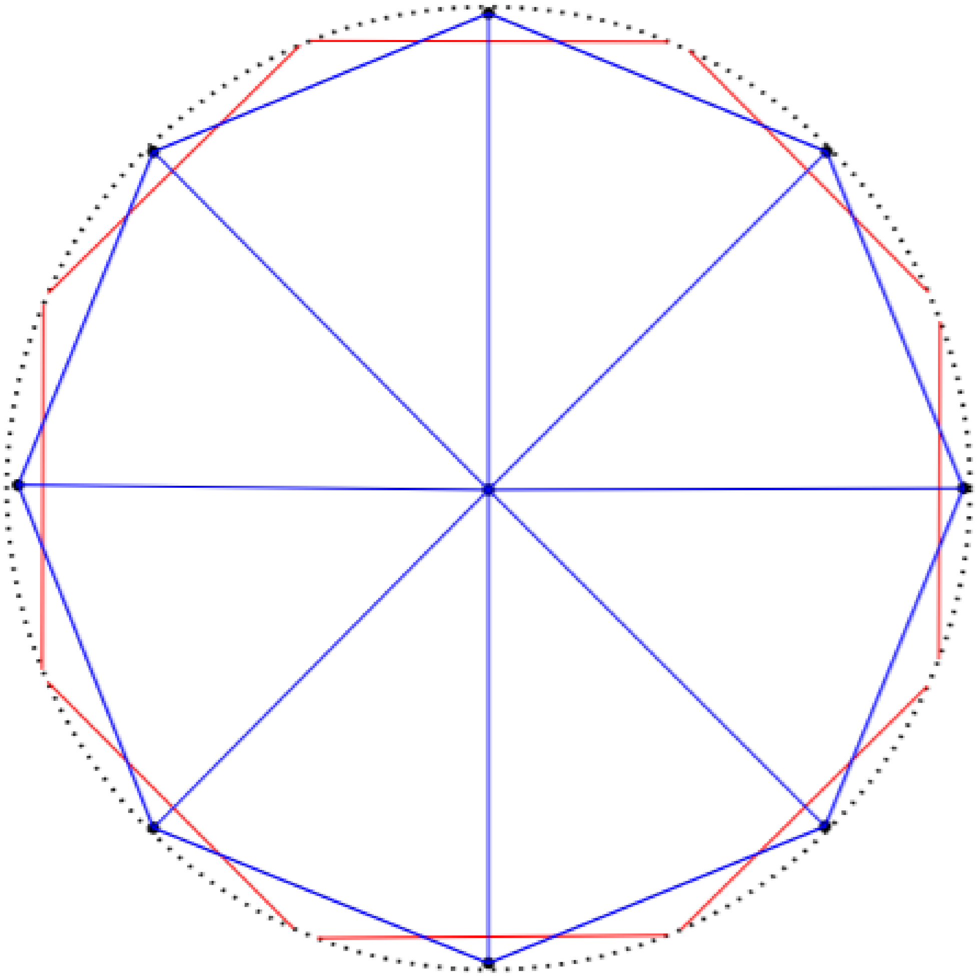


conformal at the origin (Klein)

Orthogonality with respect to the Riemannian metric

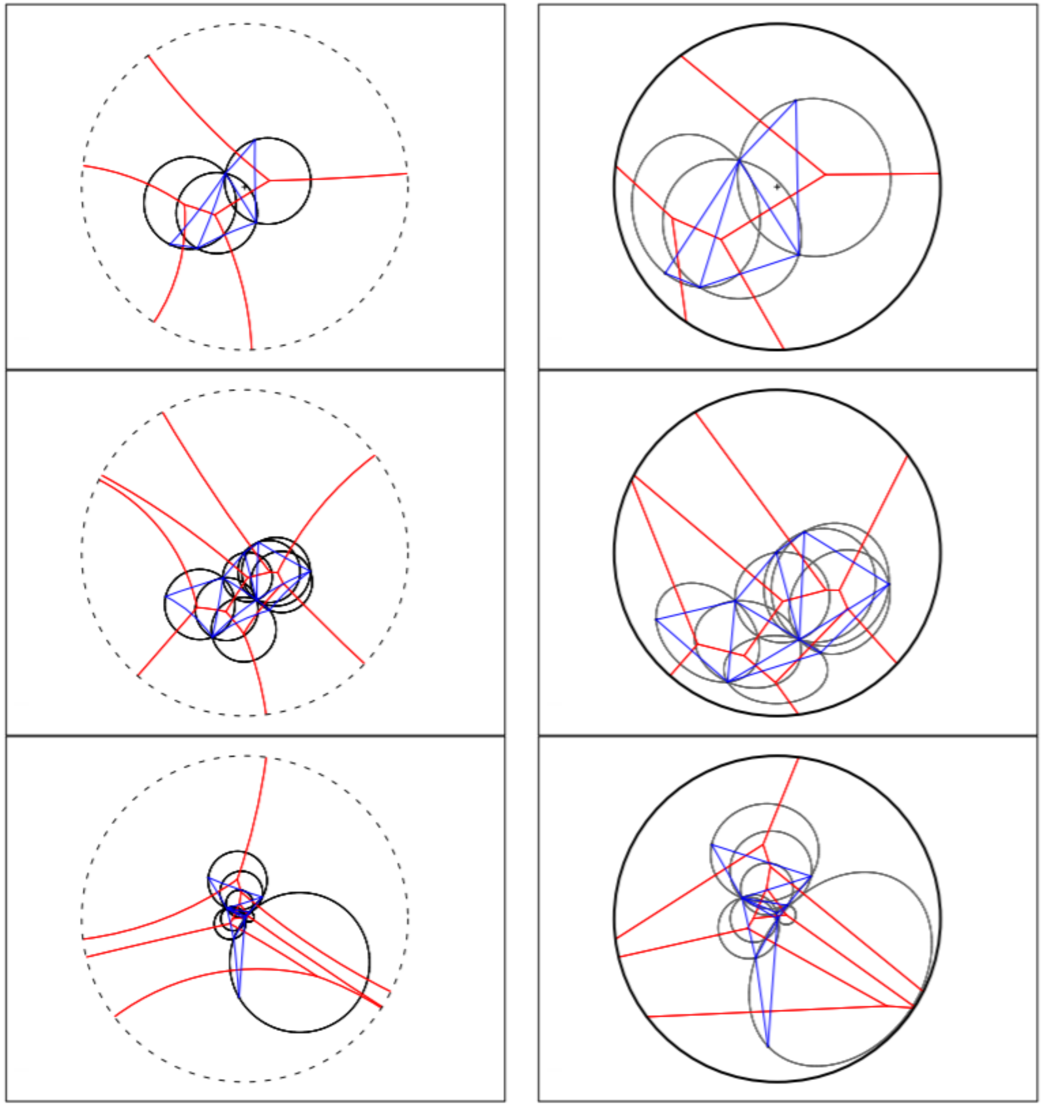

Poincaré upper plane
Poincaré disk
Klein disk
Cauchy/Hyperbolic Voronoi diagrams

Hyperbolic Voronoi diagram with all unbounded Voronoi cells



Klein disk

Hyperbolic Delaunay complex: Empty-sphere property



Empty sphere: The ball passing through d+1 sites is empty of other sites

Generalize the **empty sphere property** of the ordinary Voronoi diagram

Dually flat Cauchy Voronoi diagrams

Primal bisector: coincide with the hyperbolic bisector:

$$\begin{aligned}\mathrm{Bi}_{D_{\mathrm{flat}}}(p_{\lambda_1}:p_{\lambda_2}) &= \left\{p_\lambda \,:\, D_{\mathrm{flat}}[p_{\lambda_1}:p_\lambda] = D_{\mathrm{flat}}[p_{\lambda_2}:p_\lambda]\right\},\\ &= \left\{\lambda \,:\, \delta(l_1,s_1;l,s) = \delta(l_2,s_2;l,s)\right\}.\end{aligned}$$

$$\boxed{\mathrm{Bi}_{D_{\mathrm{flat}}}(p_{\lambda_1}:p_{\lambda_2}) = \mathrm{Bi}_{\rho_{\mathrm{FR}}}(p_{\lambda_1}:p_{\lambda_2}) = \mathrm{Bi}_{D_{\mathrm{KL}}}(p_{\lambda_1}:p_{\lambda_2}) = \mathrm{Bi}_{D_{\chi^2}}(p_{\lambda_1}:p_{\lambda_2}).}$$

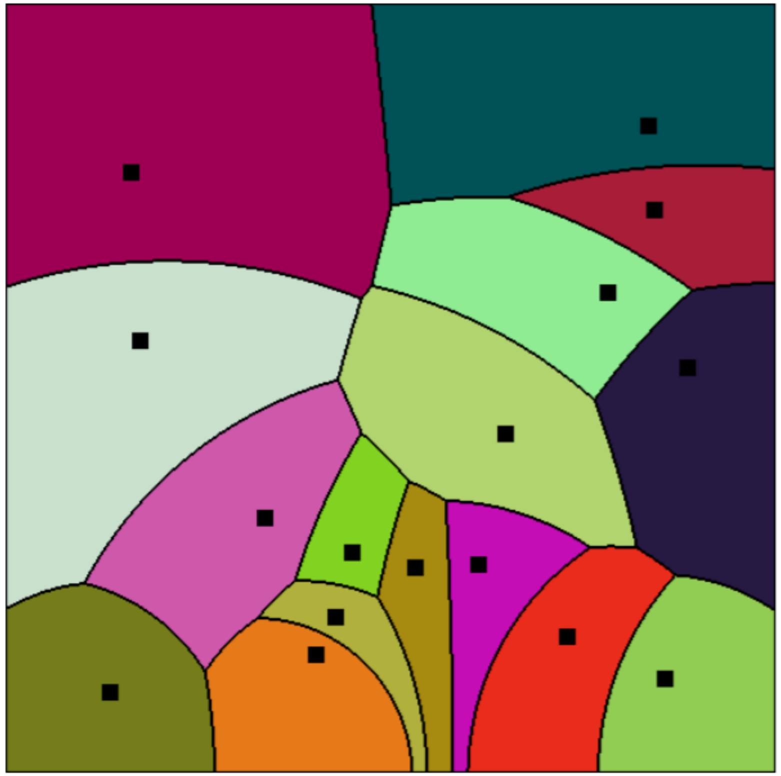
Dual bisector: coincide with the Euclidean bisector:

$$\begin{aligned}\mathrm{Bi}^*_{D_{\mathrm{flat}}}(p_{\lambda_1}:p_{\lambda_2}) &= \left\{p_\lambda \,:\, D_{\mathrm{flat}}[p_\lambda:p_{\lambda_1}] = D_{\mathrm{flat}}[p_\lambda:p_{\lambda_2}]\right\},\\ &= \left\{\lambda \,:\, \|\lambda-\lambda_1\| = \|\lambda-\lambda_2\|\right\}.\end{aligned}$$

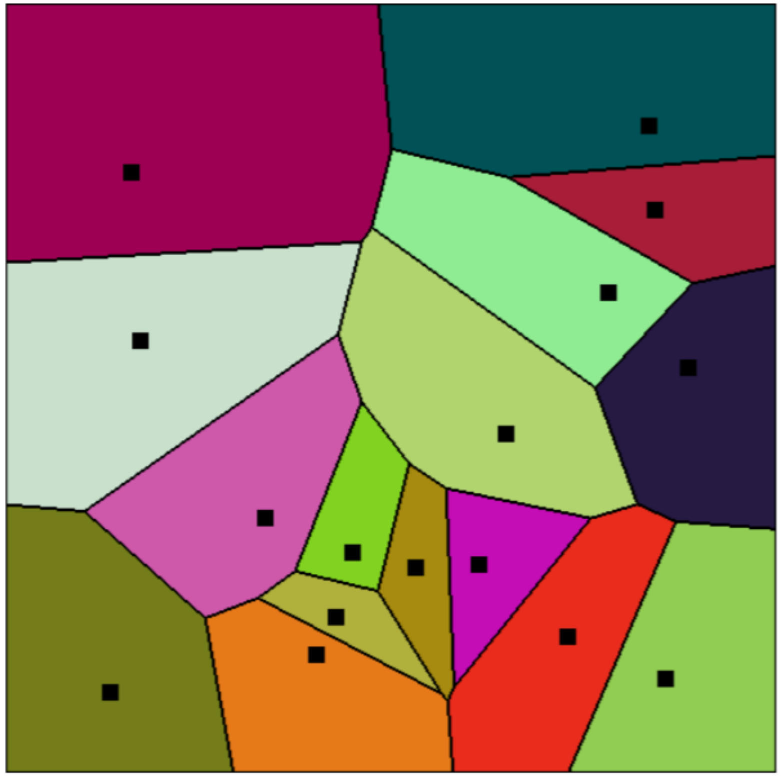
$$\boxed{\mathrm{Bi}^*_{D_{\mathrm{flat}}}(p_{\lambda_1}:p_{\lambda_2}) = \mathrm{Bi}_{\rho_E}(p_{\lambda_1},p_{\lambda_2}).}$$

Summary of Cauchy Voronoi diagrams:

Formula	Voronoi
$D_{\chi^2}[p_{l_1,s_1}, p_{l_2,s_2}] = \frac{(l_2-l_1)^2+(s_2-s_1)^2}{2s_1s_2}$	$\mathrm{Vor}_{D_{\chi^2}}$ hyperbolic Voronoi
$\rho_{\mathrm{FR}}[p_{l_1,s_1}, p_{l_2,s_2}] = \frac{1}{\sqrt{2}}\mathrm{arccosh}(1 + D_{\chi^2}[p_{l_1,s_1}, p_{l_2,s_2}])$	$\mathrm{Vor}_{\rho_{\mathrm{FR}}}$ hyperbolic Voronoi
$D_{\mathrm{KL}}[p_{l_1,s_1}, p_{l_2,s_2}] = \log\left(1 + \frac{1}{2}D_{\chi^2}[p_{l_1,s_1}, p_{l_2,s_2}]\right)$	$\mathrm{Vor}_{D_{\mathrm{KL}}}$ hyperbolic Voronoi
$\rho_{\mathrm{KL}}[p_{l_1,s_1}, p_{l_2,s_2}] = \sqrt{D_{\mathrm{KL}}[p_{l_1,s_1}, p_{l_2,s_2}]}$ (metric)	$\mathrm{Vor}_{\rho_{\mathrm{KL}}}$ hyperbolic Voronoi
$D_{\mathrm{flat}}[p_{l_1,s_1}, p_{l_2,s_2}] = 2\pi s_2 D_{\chi^2}[p_{l_1,s_1}, p_{l_2,s_2}]$	Bregman Voronoi: $\mathrm{Vor}_{D_{\mathrm{flat}}}$ hyperbolic Voronoi, $\mathrm{Vor}^*_{D_{\mathrm{flat}}}$ Euclidean Voronoi.



$\mathrm{Vor}_{\rho_{\mathrm{FR}}} = \mathrm{Vor}_{\rho_{\mathrm{KL}}} = \mathrm{Vor}_{\rho^2_{\chi}} = \mathrm{Vor}_{D_{\mathrm{flat}}}$



$\mathrm{Vor}_{D^*_{\mathrm{flat}}} = \mathrm{Vor}_{\rho_E}$.

Summary: Information-geometric Cauchy manifolds

- The **α-geometries** of the Cauchy manifolds all coincide, and yields a **hyperbolic geometry** of constant negative scalar curvature -2.
- By using Tsallis' quadratic entropy, we can realize Cauchy distributions (q-Gaussians for q=2) as **maximum entropy distributions**.
- The dual potential functions induced by deformed q=2 log-normalizer yields a conformal flattening of the curved Fisher-Rao geometry where the Riemannian metric is a **conformal metric of the Fisher information metric**.
- The Kullback-Leibler divergence between two Cauchy distributions is **symmetric**, and its square root yields a **metric distance**. For scaled Cauchy distributions, the square root of the KLD is a **Hilbertian metric**.
- The **Cauchy Voronoi diagrams** wrt to the chi-squared, KL, and Fisher-Rao distances coincide with a **hyperbolic Voronoi diagram**. The dual Voronoi diagram for the **flat divergence** coincides with the **Euclidean Voronoi diagram**.
- The hyperbolic Delaunay complex is **orthogonal** to the hyperbolic Voronoi diagram, and is often not a triangulation, hence its name **hyperbolic Delaunay complex**.

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid

Frank Nielsen

Sony Computer Science Laboratories, Inc

https://franknielsen.github.io/

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid
Entropy 2020, 22(2), 221; https://doi.org/10.3390/e22020221
https://www.mdpi.com/1099-4300/22/2/221

The Jensen-Shannon divergence in a nutshell

Kullback-Leibler divergence: (asymmetric, unbounded)

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

require same support

Jensen-Shannon divergence: (symmetric, bounded)

$$0 \leq \mathrm{JS}(p:q) \leq \log 2$$

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

$$= \frac{1}{2}\int\left(p\log\frac{2p}{p+q}+q\log\frac{2q}{p+q}\right)\mathrm{d}\mu = \mathrm{JS}(q,p).$$

Do not require same support

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

Shannon entropy: $h(p) = -\int p \log p \mathrm{d}\mu$

JSD (capacitory discrimination) = total KL divergence to the average distribution

$(\mathcal{X}, \sqrt{\mathrm{JS}})$ is a **Hilbert metric space**

The extended Jensen-Shannon divergence

Extended Kullback-Leibler divergence to **positive measures**:

$$\begin{aligned}\mathrm{KL}^{+}(\tilde{p}:\tilde{q}) &:= \mathrm{KL}(\tilde{p}:\tilde{q}) + \int \tilde{q}\mathrm{d}\mu - \int \tilde{p}\mathrm{d}\mu, \\ &= \int \left(\tilde{p}\log\frac{\tilde{p}}{\tilde{q}} + \tilde{q} - \tilde{p}\right)\mathrm{d}\mu.\end{aligned}$$

Extended Jensen-Shannon divergence to **positive measures**:

$$\boxed{\begin{aligned}\mathrm{JS}^{+}(\tilde{p},\tilde{q}) &:= \frac{1}{2}\left(\mathrm{KL}^{+}\left(\tilde{p}:\frac{\tilde{p}+\tilde{q}}{2}\right) + \mathrm{KL}^{+}\left(\tilde{q}:\frac{\tilde{p}+\tilde{q}}{2}\right)\right), \\ &= \frac{1}{2}\left(\mathrm{KL}\left(\tilde{p}:\frac{\tilde{p}+\tilde{q}}{2}\right) + \mathrm{KL}\left(\tilde{q}:\frac{\tilde{p}+\tilde{q}}{2}\right)\right) = \mathrm{JS}(\tilde{p},\tilde{q})\end{aligned}}$$

Extended Jensen-Shannon divergence upper bounded by $(\frac{1}{2}\log 2)(\int(\tilde{p}+\tilde{q})\mathrm{d}\mu)$

Skewed Jensen-Shannon divergences

Notation for *statistical mixture*: $(pq)_\alpha(x) := (1-\alpha)p(x) + \alpha q(x) \qquad \alpha \in [0,1]$

Skewed Jensen-Shannon divergence for $\alpha \in (0,1)$

$$\begin{aligned} \mathrm{JS}_a^{\alpha}(p:q) &:= (1-\alpha)\mathrm{KL}(p:(pq)_\alpha) + \alpha \mathrm{KL}(q:(pq)_\alpha), \\ &= (1-\alpha)\int p \log \frac{p}{(pq)_\alpha} \mathrm{d}\mu + \alpha \int q \log \frac{q}{(pq)_\alpha} \mathrm{d}\mu. \end{aligned}$$

By introducing the **skewed Kullback-Leibler divergence**:

$$K_\alpha(p:q) := \mathrm{KL}(p:(1-\alpha)p + \alpha q) = \mathrm{KL}(p:(pq)_\alpha)$$

Symmetric skewed Jensen-Shannon divergence:

$$\mathrm{JS}^{\alpha}(p,q) := \frac{1}{2}K_\alpha(p:q) + \frac{1}{2}K_\alpha(q:p) = \mathrm{JS}^{\alpha}(q,p).$$

… and we recover the JSD for ½:

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

Jensen-Shannon divergences are f-divergences

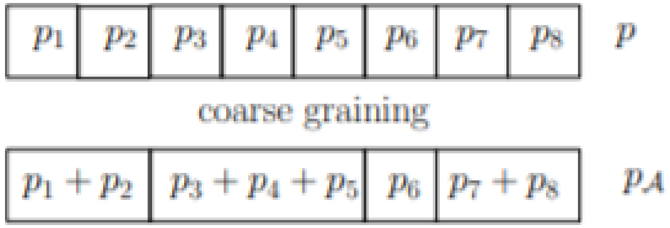
f-divergences for convex generator f, strictly convex at 1 with f(1)=0

(standard when f’(1)=0, f’’(1)=1)

$$\boxed{I_f(p:q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \geq f(1) = 0.}$$

f-divergences satisfy **information monotonicity**

(= data processing inequality) $\boxed{D(\theta_{\bar{A}} : \theta'_{\bar{A}}) \leq D(\theta : \theta')}$



coarse binning, lumping

f-divergences **upper bounded** by $f(0) + f^*(0)$

Skewed Jensen-Shannon divergences are f-divergences for the generator:

$$f_\alpha(x) = -\log((1-\alpha)+\alpha x) - x\log\left((1-\alpha)+\frac{\alpha}{x}\right)$$

Extending Jensen-Shannon divergences: Vector skewed Jensen–Bregman Divergences

Vector-skewed α-Jensen–Bregman divergence (α-JBD):

$$\boxed{\mathrm{JB}_F^{\alpha,\gamma,w}(\theta_1 : \theta_2) := \sum_{i=1}^{k} w_i B_F\left((\theta_1\theta_2)_{\alpha_i} : (\theta_1\theta_2)_{\gamma}\right) \geq 0,}$$

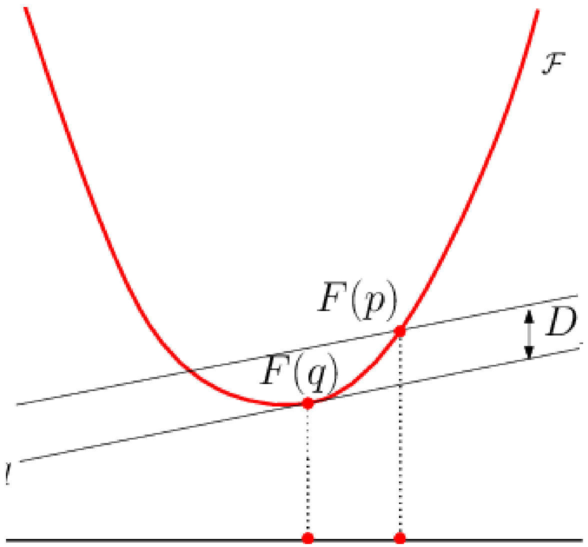
Skewing vector : $\alpha \in [0,1]^k$

Weight vector belongs to Δ_k (standard k-simplex)

Notation for linear interpolation: $(ab)_\alpha := (1-\alpha)a + \alpha b$

Bregman divergence:

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



Rewriting the vector skewed Jensen–Bregman divergences

Notation: $(ab)_\alpha := (1-\alpha)a + \alpha b$

We have: $(\theta_1\theta_2)_{\alpha_i} - (\theta_1\theta_2)_\gamma = (\gamma - \alpha_i)(\theta_1 - \theta_2),$

Therefore $\mathrm{JB}_F^{\alpha,\gamma,w}(\theta_1 : \theta_2) := \sum_{i=1}^k w_i B_F\left((\theta_1\theta_2)_{\alpha_i} : (\theta_1\theta_2)_\gamma\right) \geq 0,$ Rewrites as

$$\mathrm{JB}_F^{\alpha,\gamma,w}(\theta_1 : \theta_2) = \left(\sum_{i=1}^k w_i F\left((\theta_1\theta_2)_{\alpha_i}\right)\right) - F\left((\theta_1\theta_2)_\gamma\right) - \left\langle \sum_{i=1}^k w_i(\gamma - \alpha_i)(\theta_1 - \theta_2), \nabla F((\theta_1\theta_2)_\gamma) \right\rangle.$$

The *inner product vanishes* when we choose $\boxed{\gamma = \sum_{i=1}^k w_i\alpha_i := \bar{\alpha}.}$

And we get the **vector-skew α-JBD**:

$$\boxed{\mathrm{JB}_F^{\alpha,w}(\theta_1 : \theta_2) = \left(\sum_{i=1}^k w_i F\left((\theta_1\theta_2)_{\alpha_i}\right)\right) - F\left((\theta_1\theta_2)_{\bar{\alpha}}\right)}$$

Vector-skew Jensen–Shannon divergences

Definition 1 (Weighted vector-skew (α, w)-Jensen–Shannon divergence). *For a vector $\alpha \in [0,1]^k$ and a unit positive weight vector $w \in \Delta_k$, the (α, w)-Jensen–Shannon divergence between two densities $p, q \in \bar{\mathcal{P}}_1$ is defined by:*

$$\boxed{\mathrm{JS}^{\alpha,w}(p:q) := \sum_{i=1}^{k} w_i \mathrm{KL}((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}) = h\left((pq)_{\bar{\alpha}}\right) - \sum_{i=1}^{k} w_i h\left((pq)_{\alpha_i}\right),}$$

with $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$, where $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$ denotes the Shannon entropy [4] (i.e., $-h$ is strictly convex).

Theorem 1. *The vector-skew Jensen–Shannon divergences $\mathrm{JS}^{\alpha,w}(p:q)$ are f-divergences for the generator $f_{\alpha,w}(u) = \sum_{i=1}^{k} w_i(\alpha_i u + (1-\alpha_i)) \log \frac{(1-\alpha_i)+\alpha_i u}{(1-\bar{\alpha})+\bar{\alpha} u}$ with $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$.*

→ Invariant information-monotone divergences

Theorem 2 (Separable convexity). *The divergence $\mathrm{KL}_{\alpha,\beta}(p:q)$ is strictly separable convex for $\alpha \neq \beta$ and $x \in \mathcal{X}_p \cap \mathcal{X}_q$.*

→ Nice for optimization

Properties of the vector-skew JS divergences

Lemma 1 (KLD between two w-mixtures). *For $\alpha \in [0,1]$ and $\beta \in (0,1)$, we have:*

$$\boxed{\mathrm{KL}_{\alpha,\beta}(p:q) = \mathrm{KL}\left((pq)_\alpha : (pq)_\beta\right) \leq \log \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}.}$$

Lemma 2 (Bounded (w,α)-Jensen–Shannon divergence). $\mathrm{JS}^{\alpha,w}$ *is bounded by* $\log \frac{1}{\bar{\alpha}(1-\bar{\alpha})}$ *where* $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i \in (0,1)$.

Jensen–Shannon centroids on mixture families

Mixture family in information geometry (w-mixtures)

$$\mathcal{M} := \left\{ m(x;\theta) := \sum_{i=1}^{D} \theta^i p_i(x) + \left(1 - \sum_{i=1}^{D} \theta^i\right) p_0(x) \ : \ \theta^i > 0, \sum_{i=1}^{D} \theta^i < 1 \right\}.$$

Example: The *family of categorical distributions* is a mixture family:

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

The Kullback-Leibler divergence between two mixture distributions amount to a Bregman divergence for the negentropy generator:

$$\boxed{\mathrm{KL}(m_{\theta_1} : m_{\theta_2}) = B_F(\theta_1 : \theta_2) = B_{-h(m_\theta)}(\theta_1 : \theta_2).} \qquad F(\theta) = -h(m_\theta)$$

Jensen–Shannon centroids

Like the **Fréchet mean**, we define the **Jensen-Shannon centroid** as the minimizer(s) of

$$L(\theta) := \sum_{j=1}^{n} \omega_j \mathrm{JS}^{\alpha,w}(m_{\theta_k} : m_\theta),$$

$$L(\theta) = \sum_{j=1}^{n} \omega_j \left(\sum_{i=1}^{k} w_i F((\theta_j \theta)_{\alpha_i}) - F\left((\theta_j \theta)_{\bar{\alpha}}\right) \right)$$

This defines a **Difference of Convex (DC) program:** $\boxed{\min_\theta A(\theta) - B(\theta)}$

With convex functions:

$$\begin{aligned} A(\theta) &= \sum_{j=1}^{n} \sum_{i=1}^{\kappa} \omega_j w_i F((\theta_j \theta)_{\alpha_i}), \\ B(\theta) &= \sum_{j=1}^{n} \omega_j F\left((\theta_j \theta)_{\bar{\alpha}}\right). \end{aligned}$$

Jensen–Shannon centroids: CCCP

Convex-ConCave Procedure (CCCP) is *step-size free* optimization for *smooth* DC programs:

- Initialize $\theta^{(0)}$ arbitrarily (eg, centroid)
- Iteratively update: $\boxed{\theta^{(t+1)} = (\nabla B)^{-1}(\nabla A(\theta^{(t)}))}$

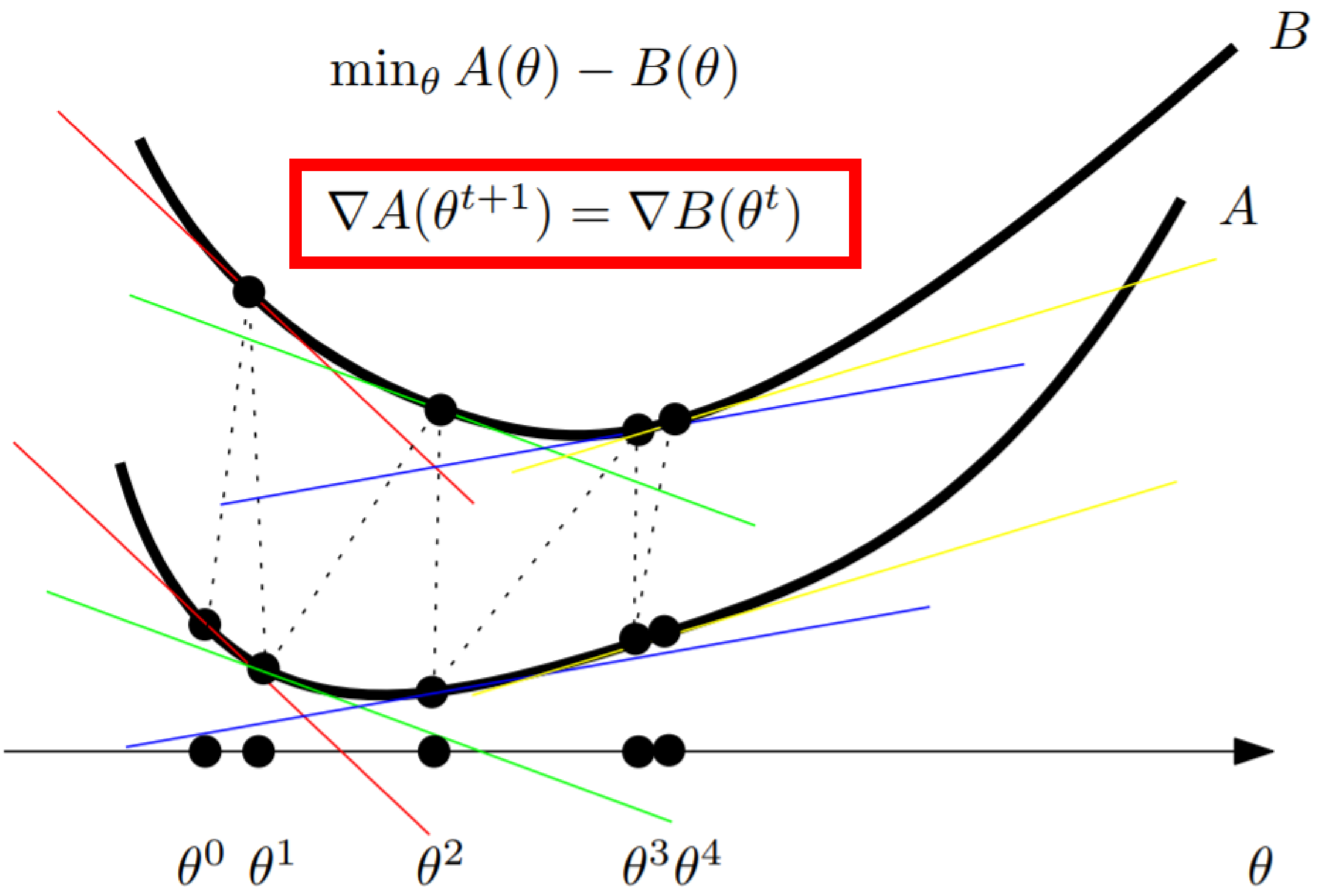
$$A(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{\kappa} \omega_j w_i F((\theta_j \theta)_{\alpha_i}),$$

$$B(\theta) = \sum_{j=1}^{n} \omega_j F\left((\theta_j \theta)_{\bar{\alpha}}\right).$$

$$\nabla A(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{k} \omega_j w_i \alpha_i \nabla F((\theta_j \theta)_{\alpha_i})$$

$$\nabla B(\theta) = \sum_{j=1}^{n} \omega_j \bar{\alpha} \nabla F\left((\theta_j \theta)_{\bar{\alpha}}\right)$$

Visualization of the CCCP



Interpretation: Support hyperplanes to *A graph* shall be parallel to *B graph*

Jensen-Shannon centroid for categorical distributions

Mixture family (mixture of mixtures is a mixture):

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

Shannon neg-entropy is a strictly convex and differentiable **Bregman generator**:

$$\boxed{\begin{aligned} F(\theta) &= -h(m_\theta) = \sum_{i=1}^{D} \theta_i \log \theta_i + \left(1 - \sum_{i=1}^{D} \theta_i\right) \log \left(1 - \sum_{i=1}^{D} \theta_i\right). \\ \mathrm{KL}(m_{\theta_1} : m_{\theta_2}) &= B_F(\theta_1 : \theta_2) = B_{-h(m_\theta)}(\theta_1 : \theta_2). \end{aligned}}$$

$$\nabla F(\theta) = \left[\frac{\partial}{\partial \theta_i}\right]_i, \quad \frac{\partial}{\partial \theta_i} F(\theta) = \log \frac{\theta_i}{1 - \sum_{j=1}^{D} \theta_j}.$$

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

$$\nabla F^*(\eta) = (\nabla F)^{-1}(\eta) = \frac{1}{1 + \sum_{j=1}^{D} \exp(\eta_j)} [\exp(\eta_i)]_i,$$

$$\theta_i = (\nabla F^{-1}(\eta))_i = \frac{\exp(\eta_i)}{1 + \sum_{j=1}^{D} \exp(\eta_j)},$$

Jensen-Shannon centroid: Implementing CCCP

Initialize: $\theta^{(0)} = \frac{1}{n}\sum_i \theta_i$

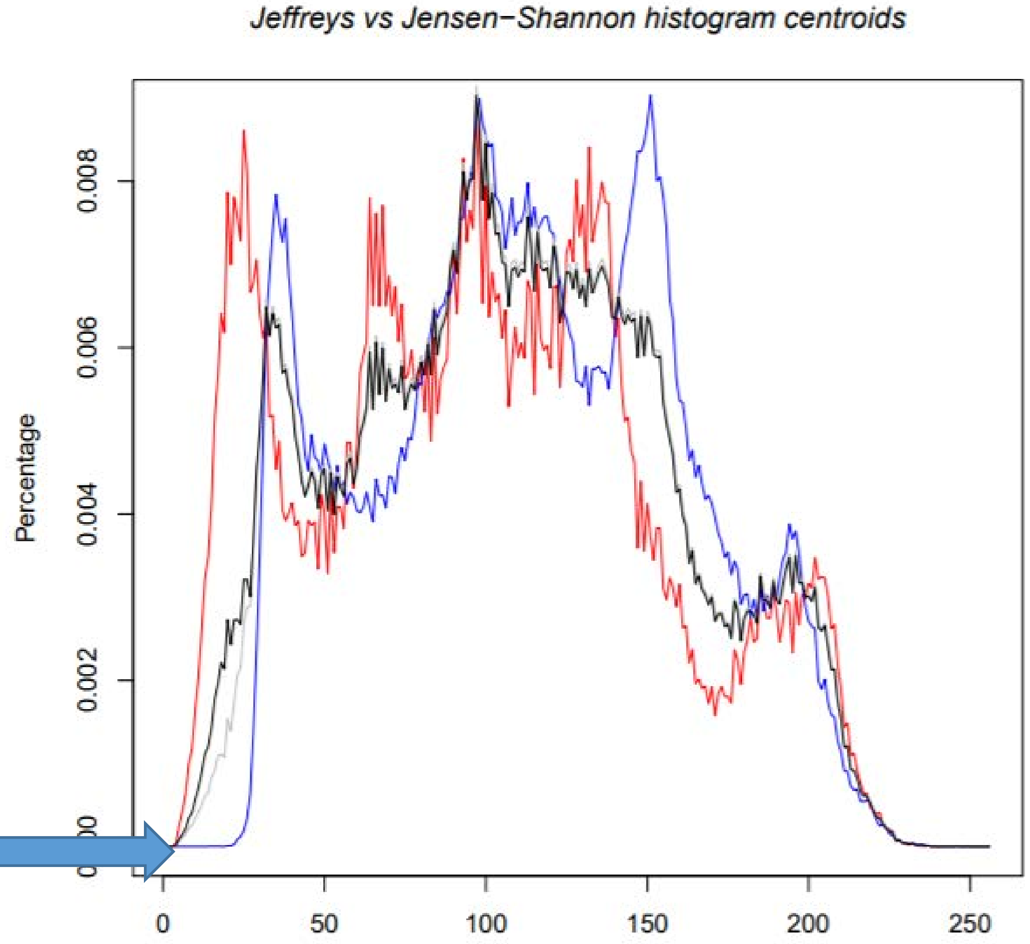
Iterate: $$\theta^{(t+1)} = (\nabla F)^{-1}\left(\frac{1}{n}\sum_i \nabla F\left(\frac{\theta_i + \theta^{(t)}}{2}\right)\right)$$

$$\nabla F(\theta) = \left[\frac{\partial}{\partial \theta_i}\right]_i, \quad \frac{\partial}{\partial \theta_i} F(\theta) = \log \frac{\theta_i}{1 - \sum_{j=1}^{D} \theta_j}.$$

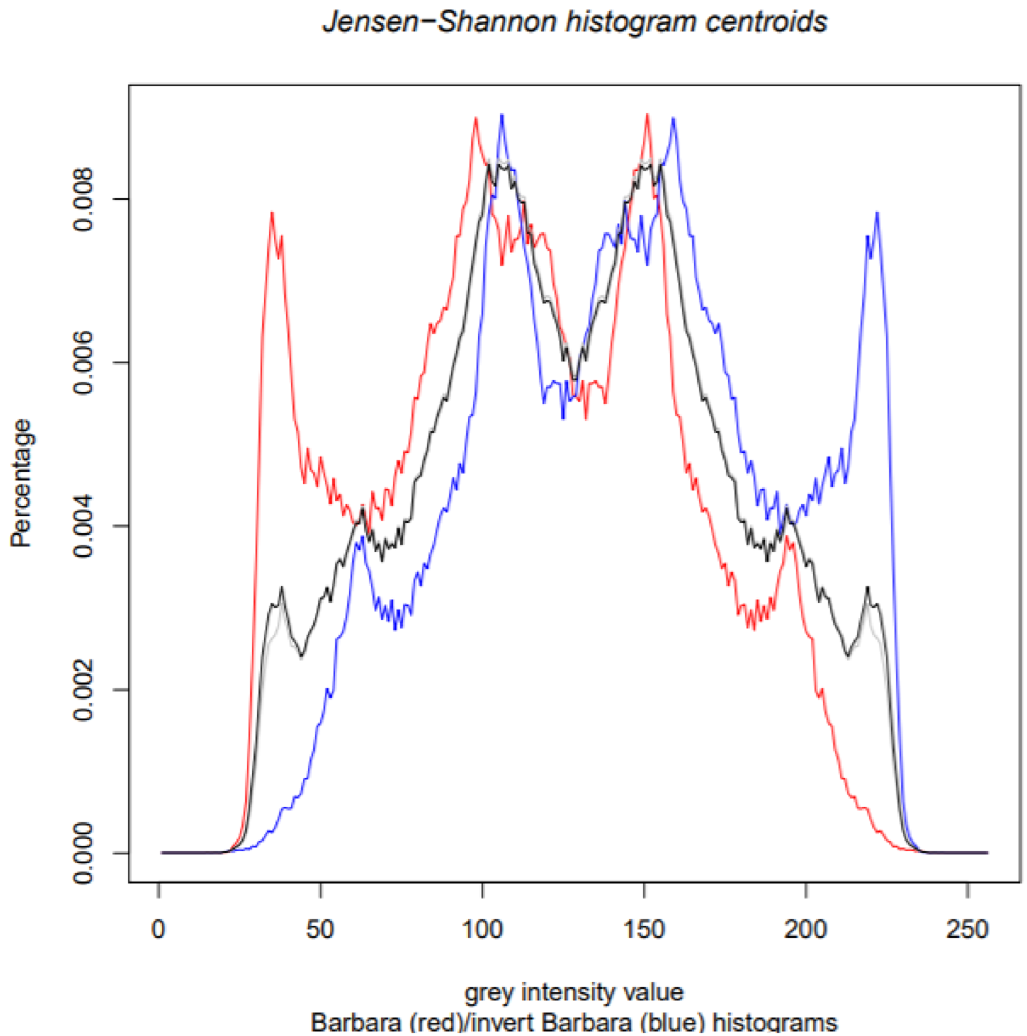
$$\nabla F^*(\eta) = (\nabla F)^{-1}(\eta) = \frac{1}{1 + \sum_{j=1}^{D} \exp(\eta_j)} [\exp(\eta_i)]_i,$$

Experiments:

Jeffreys centroid (grey histogram)
Jensen–Shannon centroid (black histogram)
Lena image (red histogram)
Barbara image (blue histogram)



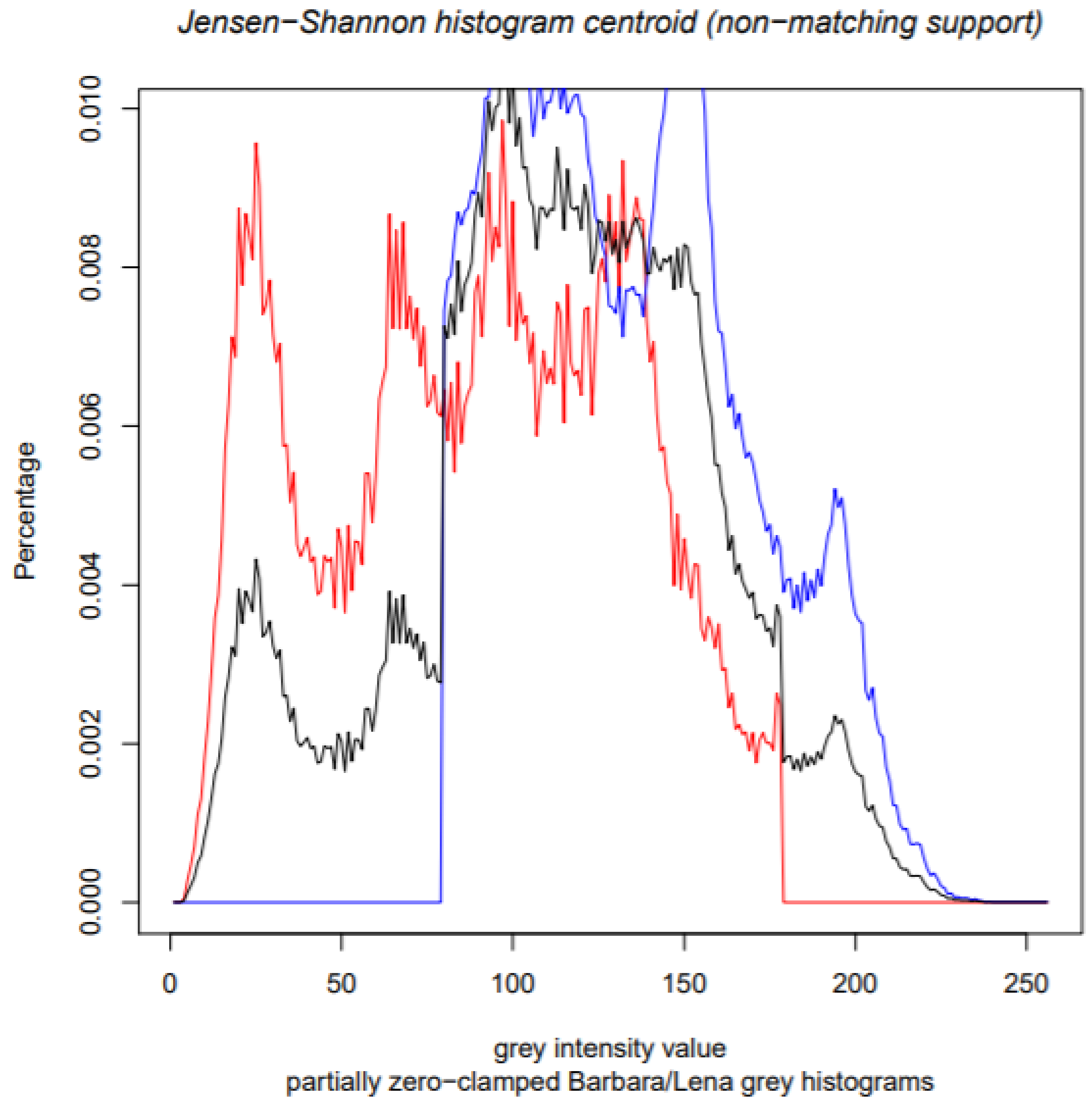
Close to zero in [0,20]

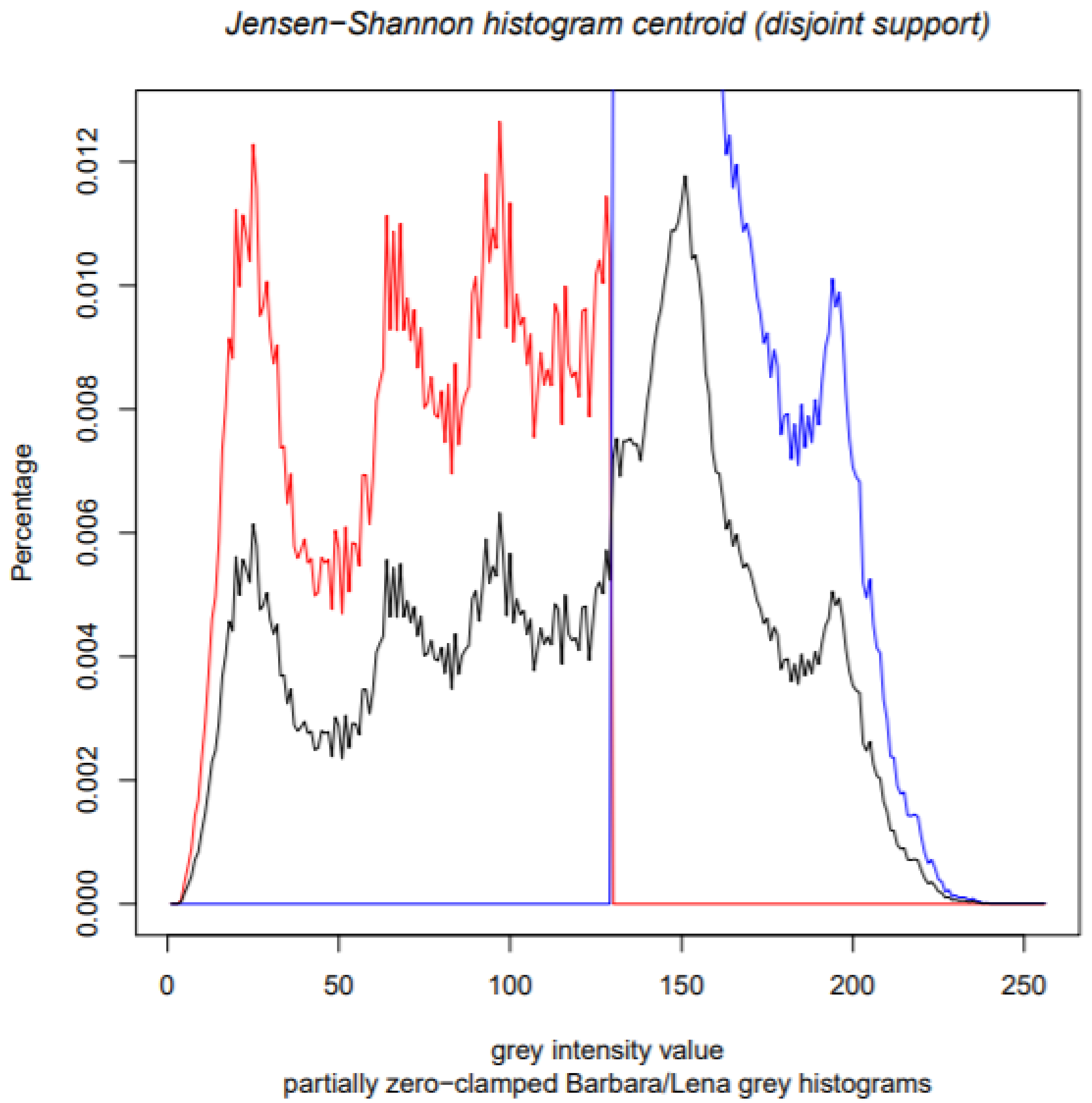


Barbara histogram

negative image histogram

JSD always bounded even on different supports





Summary: Vector-skewed Jensen-Shannon divergence

- Jensen-Shannon divergence is a **bounded symmetrization** of the Kullback-Leibler divergence (KLD) which allows to measure the distance between distributions with **potentially different supports** (useful in ML like GANs)
- Jensen-Shannon divergence is a **f-divergence** which satisfies the **data processing inequality**
- Generalize the weighted skewed Jensen-Shannon divergence by using a **skew vector parameter** $\alpha \in [0,1]^k$: $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$ $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$

$$\boxed{\mathrm{JS}^{\alpha,w}(p:q) := \sum_{i=1}^{k} w_i \mathrm{KL}((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}) = h((pq)_{\bar{\alpha}}) - \sum_{i=1}^{k} w_i h((pq)_{\alpha_i})}$$

- The vector-skewed Jensen-Shannon divergence is an information monotone f-divergence
- The (vector-skewed) Jensen-Shannon centroids can be modeled using a smooth **Difference of Convex (DC) program** and solved using
- the **Convex-ConCave Procedure** (CCCP)

https://www.mdpi.com/1099-4300/22/2/221

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

Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means
Entropy 2019, 21(5), 485; https://doi.org/10.3390/e21050485
https://www.mdpi.com/1099-4300/21/5/485
Code: https://franknielsen.github.io/M-JS/

Unbounded Kullback-Leibler divergence (KLD)

$\mathrm{KL} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$

$$\boxed{\mathrm{KL}(P : Q) := \int p \log \frac{p}{q} \mathrm{d}\mu}$$

$P, Q \ll \mu$

Also called **relative entropy**:

$$\boxed{\mathrm{KL}(p : q) = h_\times(p : q) - h(p)}$$

Cross-entropy: $h_\times(p : q) := \int p \log \frac{1}{q} \mathrm{d}\mu,$

Shannon's entropy: $h(p) := \int p \log \frac{1}{p} \mathrm{d}\mu = h_\times(p : p),$

(self cross-entropy)

Reverse KLD:

(KLD=forward KLD)

$$\boxed{\mathrm{KL}^*(P : Q) := \mathrm{KL}(Q : P) = \int q \log \frac{q}{p} \mathrm{d}\mu.}$$

Symmetrizations of the Kullback-Leibler divergence

Jeffreys' divergence (twice the arithmetic mean of oriented KLDs):

$$\boxed{J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p-q)\log\frac{p}{q}\mathrm{d}\mu = J(q;p)}$$

Resistor average divergence (harmonic mean of forward+reverse KLD)

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

Question: Role and extensions of the mean in symmetrization ?

Bounded Jensen-Shannon divergence (JSD)

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

$$= \quad \frac{1}{2}\int\left(p\log\frac{2p}{p+q}+q\log\frac{2q}{p+q}\right)\mathrm{d}\mu.$$

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

(Shannon entropy h is strictly concave, JSD>=0)

Do not require same support

JSD is **bounded**: $0 \leq \mathrm{JS}(p:q) \leq \log 2$

Proof: $\mathrm{KL}\left(p:\frac{p+q}{2}\right)=\int p\log\frac{2p}{p+q}\mathrm{d}\mu \leq \int p\log\frac{2p}{p}\mathrm{d}\mu = \log 2.$

$\sqrt{\mathrm{JS}}$: Square root of the JSD is a **metric distance** (moreover Hilbertian)

Invariant f-divergences, symmetrized f-divergences

Convex generator f, strictly convex at 1
with f(1)=0 (standard when f'(1)=0, f''(1)=1)

$$\boxed{I_f(p:q) = \int pf\left(\frac{q}{p}\right) \mathrm{d}\mu}$$

f-divergences are said **invariant** in *information geometry* because they satisfy **coarse-graining** (data processing inequality)

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p

coarse graining

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

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

f-divergences can always be symmetrized: **Reverse f-divergence** for $f^*(x) = xf(\frac{1}{x})$

$$\text{Jeffreys f-generator:} \quad f_J(u) := (u-1)\log u,$$

$$\text{Jensen-Shannon f-generator:} \quad f_{JS}(u) := -(u+1)\log\frac{1+u}{2} + u\log u.$$

Statistical distances vs parameter vector distances

A **statistical distance D** between two parametric distributions of a same family (eg., Gaussian family) amount to a **parameter distance P**:

$$\boxed{P(\theta:\theta'):=D(p_\theta:p_{\theta'})}$$

For example, the KLD between two densities of a same exponential family amounts to a **reverse Bregman divergence** for the *Bregman cumulant generator*:

$$\boxed{\mathrm{KL}(p_\theta:p_{\theta'})=B_F^*(\theta:\theta')}=B_F(\theta':\theta).$$

$$B_F(\theta:\theta'):=F(\theta)-F(\theta')-\langle\theta-\theta',\nabla F(\theta')\rangle$$

From a smooth C3 parameter distance (= *contrast function*), we can build a dualistic information-geometric structure

Skewed Jensen-Bregman divergences

JS-kind symmetrization of the *parameter Bregman divergence*:

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

Notation for the **linear interpolation**: $(\theta_p\theta_q)_\alpha := (1-\alpha)\theta_p+\alpha\theta_q$

$$
\boxed{
\begin{aligned}
\mathrm{JB}_F^\alpha(\theta:\theta') &:= (1-\alpha)B_F\left(\theta:(\theta\theta')_\alpha\right)+\alpha B_F\left(\theta':(\theta\theta')_\alpha\right)\\
&= (F(\theta)F(\theta'))_\alpha - F\left((\theta\theta')_\alpha\right) =: J_F^\alpha(\theta:\theta'),
\end{aligned}
}
$$

J-Symmetrization and JS-Symmetrization

J-symmetrization of a statistical/parameter distance D:

$$\boxed{J_D^\alpha(p:q) := (1-\alpha) D(p:q) + \alpha D(q:p)} \qquad \alpha \in [0,1]$$

JS-symmetrization of a statistical/parameter distance D:

$$\boxed{\begin{aligned} \mathrm{JS}_D^\alpha(p:q) &:= (1-\alpha) D(p:(1-\alpha)p+\alpha q) + \alpha D(q:(1-\alpha)p+\alpha q) \\ &= (1-\alpha) D(p:(pq)_\alpha) + \alpha D(q:(pq)_\alpha). \end{aligned}}$$

Example: J-symmetrization and JS-symmetrization of f-divergences:

$$f_\alpha^J(u) = (1-\alpha) f(u) + \alpha f^\diamond(u),$$

$$I_{f^\diamond}(p:q) = I_f^*(p:q) = I_f(q:p)$$

$$I_f^\alpha(p:q) := (1-\alpha) I_f(p:(pq)_\alpha) + \alpha I_f(q:(pq)_\alpha)$$

$$f_\alpha^{\mathrm{JS}}(u) := (1-\alpha) f(\alpha u + 1 - \alpha) + \alpha f\left(\alpha + \frac{1-\alpha}{u}\right).$$

Conjugate f-generator:

$$f^\diamond(u) = u f\left(\tfrac{1}{u}\right)$$

Generalized Jensen-Shannon divergences: Role of abstract weighted means, generalized mixtures

Quasi-arithmetic weighted means for a strictly increasing function h:

$$\boxed{M_\alpha^h(x,y) := h^{-1}\left((1-\alpha)h(x) + \alpha h(y)\right)}$$

Definition 1 (*M*-mixture). *The M_α-interpolation $(pq)_\alpha^M$ (with $\alpha \in [0,1]$) of densities p and q with respect to a mean M is a α-weighted M-mixture defined by:*

$$\boxed{(pq)_\alpha^M(x) := \frac{M_\alpha(p(x), q(x))}{Z_\alpha^M(p:q)}}$$

where

$$Z_\alpha^M(p:q) = \int_{t\in\mathcal{X}} M_\alpha(p(t), q(t))\mathrm{d}\mu(t) =: \langle M_\alpha(p,q)\rangle. \quad (45)$$

is the normalizer function (or scaling factor) ensuring that $(pq)_\alpha^M \in \mathcal{P}$. (The bracket notation $\langle f\rangle$ denotes the integral of f over $\mathcal{X}$.)

When M=A arithmetic mean, normalizer Z is 1

Definitions: M-JSD and M-JS symmetrizations

Definition 2 (*M*-Jensen–Shannon divergence)**.** *For a mean M, the skew M-Jensen–Shannon divergence (for $\alpha \in [0,1]$) is defined by*

$$\boxed{\mathrm{JS}^{M_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p:(pq)^M_\alpha\right) + \alpha\mathrm{KL}\left(q:(pq)^M_\alpha\right)} \tag{48}$$

When $M_\alpha = A_\alpha$, we recover the ordinary Jensen–Shannon divergence since $A_\alpha(p:q) = (pq)_\alpha$ (and $Z^A_\alpha(p:q) = 1$).

We can extend the definition to the JS-symmetrization of any distance:

Definition extended for generic distance D (not necessarily KLD):

Definition 3 (*M*-JS symmetrization)**.** *For a mean M and a distance D, the skew M-JS symmetrization of D (for $\alpha \in [0,1]$) is defined by*

$$\boxed{\mathrm{JS}^{M_\alpha}_D(p:q) := (1-\alpha)D\left(p:(pq)^M_\alpha\right) + \alpha D\left(q:(pq)^M_\alpha\right)} \tag{49}$$

Generic definition: (M,N)-JS symmetrization

Consider two **abstract means** M and N

(eg, N harmonic as in resistor average distortion):

Definition 5 (Skew (M,N)-D divergence). *The skew (M,N)-divergence with respect to weighted means M_α and N_β as follows:*

$$\boxed{\mathrm{JS}_D^{M_\alpha,N_\beta}(p:q) := N_\beta\left(D\left(p:(pq)_\alpha^M\right), D\left(q:(pq)_\alpha^M\right)\right).} \tag{61}$$

The main advantage of (M,N)-JSD is to get **closed-form formula** for distributions belonging to given parametric families by carefully choosing the M-mean.

For example, *geometric mean for exponential families*, or the *harmonic mean for Cauchy or t-Student families*, etc.

(A,G)-Jensen-Shannon divergence for exponential families

Exponential family: $\mathcal{E}_F = \left\{ p_\theta(x)\mathrm{d}\mu = \exp(\theta^\top x - F(\theta))\mathrm{d}\mu : \theta \in \Theta \right\}$

Natural parameter space: $\Theta = \left\{ \theta : \int_{\mathcal{X}} \exp(\theta^\top x)\mathrm{d}\mu < \infty \right\}$

Geometric statistical mixture:

$$\boxed{\forall x \in \mathcal{X}, \quad (p_{\theta_1} p_{\theta_2})_\alpha^G(x) := \frac{G_\alpha(p_{\theta_1}(x), p_{\theta_2}(x))}{\int G_\alpha(p_{\theta_1}(t), p_{\theta_2}(t))\mathrm{d}\mu(t)} = \frac{p_{\theta_1}^{1-\alpha}(x) p_{\theta_2}^\alpha(x)}{Z_\alpha^G(p:q)},}$$

Normalization coefficient: $\boxed{Z_\alpha^G(p:q) = \exp(-J_F^\alpha(\theta_1 : \theta_2)),}$

Jensen parameter divergence: $J_F^\alpha(\theta_1 : \theta_2) := (F(\theta_1)F(\theta_2))_\alpha - F((\theta_1\theta_2)_\alpha).$

(A,G)-Jensen-Shannon divergence for exponential families

Closed-form formula the KLD between two geometric mixtures in term of a Bregman divergence between interpolated parameters:

$$\begin{aligned}\mathrm{KL}\left(p_\theta : (p_{\theta_1} p_{\theta_2})^G_\alpha\right) &= \mathrm{KL}\left(p_\theta : p_{(\theta_1\theta_2)_\alpha}\right),\\ &= B_F((\theta_1\theta_2)_\alpha : \theta).\end{aligned}$$

$$\boxed{\begin{aligned}\mathrm{JS}^G_\alpha(p_{\theta_1} : p_{\theta_2}) &:= (1-\alpha)\mathrm{KL}(p_{\theta_1} : (p_{\theta_1} p_{\theta_2})^G_\alpha) + \alpha \mathrm{KL}(p_{\theta_2} : (p_{\theta_1} p_{\theta_2})^G_\alpha),\\ &= (1-\alpha) B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2).\end{aligned}}$$

Theorem 2 (G-JSD and its dual JS-symmetrization in exponential families). *The α-skew G-Jensen–Shannon divergence JS^{G_α} between two distributions p_{θ_1} and p_{θ_2} of the same exponential family $\mathcal{E}_F$ is expressed in closed form for $\alpha \in (0,1)$ as:*

$$\boxed{\begin{aligned}\mathrm{JS}^{G_\alpha}(p_{\theta_1} : p_{\theta_2}) &= (1-\alpha) B_F\left((\theta_1\theta_2)_\alpha : \theta_1\right) + \alpha B_F\left((\theta_1\theta_2)_\alpha : \theta_2\right), && (80)\\ \mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}(p_{\theta_1} : p_{\theta_2}) &= \mathrm{JB}^\alpha_F(\theta_1 : \theta_2) = J^\alpha_F(\theta_1 : \theta_2). && (81)\end{aligned}}$$

Example: Multivariate Gaussian exponential family

Family of Normal distributions: $\{N(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \succ 0\}$. $\lambda := (\lambda_v, \lambda_M) = (\mu, \Sigma)$

$$p_\lambda(x;\lambda) := \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2}(x-\lambda_v)^\top \lambda_M^{-1}(x-\lambda_v)\right),$$

Canonical factorization: $p_\theta(x;\theta) := \exp\left(\langle t(x),\theta\rangle - F_\theta(\theta)\right) = p_\lambda(x;\lambda(\theta)),$

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

Sufficient statistics: $t(x) = (x, -xx^\top)$

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

$$F_\lambda(\lambda) = \frac{1}{2}\left(\lambda_v^\top\lambda_M^{-1}\lambda_v + \log|\lambda_M| + d\log 2\pi\right) = \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right).$$

Example: Multivariate Gaussian exponential family

Dual moment parameterization: $\eta = (\eta_v, \eta_M) = E[t(x)] = \nabla F(\theta)$

<u>Conversions between ordinary/natural/expectation parameters:</u>

$$\begin{cases} \theta_v(\lambda) = \lambda_M^{-1}\lambda_v = \Sigma^{-1}\mu \\ \theta_M(\lambda) = \frac{1}{2}\lambda_M^{-1} = \frac{1}{2}\Sigma^{-1} \end{cases} \Leftrightarrow \begin{cases} \lambda_v(\theta) = \frac{1}{2}\theta_M^{-1}\theta_v = \mu \\ \lambda_M(\theta) = \frac{1}{2}\theta_M^{-1} = \Sigma \end{cases}$$

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

$$\begin{cases} \lambda_v(\eta) = \eta_v = \mu \\ \lambda_M(\eta) = -\eta_M - \eta_v\eta_v^\top = \Sigma \end{cases} \Leftrightarrow \begin{cases} \eta_v(\lambda) = \lambda_v = \mu \\ \eta_M(\lambda) = -\lambda_M - \lambda_v\lambda_v^\top = -\Sigma - \mu\mu^\top \end{cases}$$

Dual potential function (=negative differential Shannon entropy):

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

Corollary 1 (G-JSD between Gaussians). *The skew G-Jensen–Shannon divergence JS_α^G and the dual skew G-Jensen–Shannon divergence JS_α^{*G} between two multivariate Gaussians $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ is*

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}) + \alpha\mathrm{KL}(p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}), & (106)\\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2), & (107)\\
&= \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \right.\\
&\quad \left.(1-\alpha)(\mu_\alpha-\mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_1) + \alpha(\mu_\alpha-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_2) - d\right) & (108)\\
\mathrm{JS}_*^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}), & (109)\\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha), & (110)\\
&= J_F(\theta_1 : \theta_2), & (111)\\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right), & (112)
\end{aligned}$$

where

$$\Sigma_\alpha = (\Sigma_1\Sigma_2)_\alpha^\Sigma = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1}, \tag{113}$$

(matrix harmonic barycenter) and

$$\mu_\alpha = (\mu_1\mu_2)_\alpha^\mu = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right). \tag{114}$$

More examples: Abstract means and M-mixtures

Weighted mean	M_α, $\alpha \in (0,1)$
Arithmetic mean	$A_\alpha(x,y) = (1-\alpha)x + \alpha y$
Geometric mean	$G_\alpha(x,y) = x^{1-\alpha} y^\alpha$
Harmonic mean	$H_\alpha(x,y) = \frac{xy}{(1-\alpha)y+\alpha x}$
Power mean	$P_\alpha^p(x,y) = ((1-\alpha)x^p + \alpha y^p)^{\frac{1}{p}}, \quad p \in \mathbb{R}\backslash\{0\}, \lim_{p\to 0} P_\alpha^p = G$
Quasi-arithmetic mean	$M_\alpha^f(x,y) = f^{-1}((1-\alpha)f(x) + \alpha f(y))$, f strictly monotonous
M-mixture	$Z_\alpha^M(p,q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t)$ with $Z_\alpha^M(p,q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t)$

JS^{M_α}	**Mean M**	**Parametric Family**	$Z_\alpha^M(p:q)$
JS^{A_α}	arithmetic A	mixture family	$Z_\alpha^M(\theta_1:\theta_2) = 1$
JS^{G_α}	geometric G	exponential family	$Z_\alpha^G(\theta_1:\theta_2) = \exp(-J_F^\alpha(\theta_1:\theta_2))$
JS^{H_α}	harmonic H	Cauchy scale family	$Z_\alpha^H(\theta_1:\theta_2) = \sqrt{\frac{\theta_1\theta_2}{(\theta_1\theta_2)_\alpha(\theta_1\theta_2)_{1-\alpha}}}$

https://www.mdpi.com/1099-4300/21/5/485

Summary: Generalized Jensen-Shannon divergences

- Jensen-Shannon divergence (JSD) is a **bounded symmetrization** of the Kullback-Leibler divergence (KLD). Jeffreys divergence (JD) is an **unbounded symmetrization** of KLD. Both JSD and JD are invariant f-divergences.
- Although KLD and JD between Gaussians (or densities of a same exponential family) admits closed-form formulas, the JSD between Gaussians does not have a closed-form expression, and these distances need to be **approximated** in applications. (machine learning, eg., GANs in deep learning)
- The skewed Jensen-Shannon divergence is based on statistical arithmetic mixtures. We define generic **statistical M-mixtures** based on an abstract mean, and define accordingly the **M-Jensen-Shannon divergence**, and further the **(M,N)-JSD**.
- When M=G is the **geometric weighted mean**, we obtain closed-form formula for the **G-Jensen-Shannon divergence** between **Gaussian distributions**. Applications to machine learning (eg, deep learning GANs) https://arxiv.org/abs/2006.10599

Code: https://franknielsen.github.io/M-JS/