ASYMPTOTIC EQUIPARTITION PROPERTIES

- Link probabilities to entropy and Kullback-Leibler divergence
- Convergence in probability →: $\forall \epsilon > 0, \Pr[|X_n - X| > \epsilon] \to 0$
- I.i.d. random vector: $X_{1:n} = (X_1, \ldots, X_n), X_i \sim p_1(x)$
- AEP with the entropy: $$\boxed{-\frac{1}{n}\log p(X_{1:n}) \to H[p_1]}$$
- AEP with the Kullback-Leibler divergence:

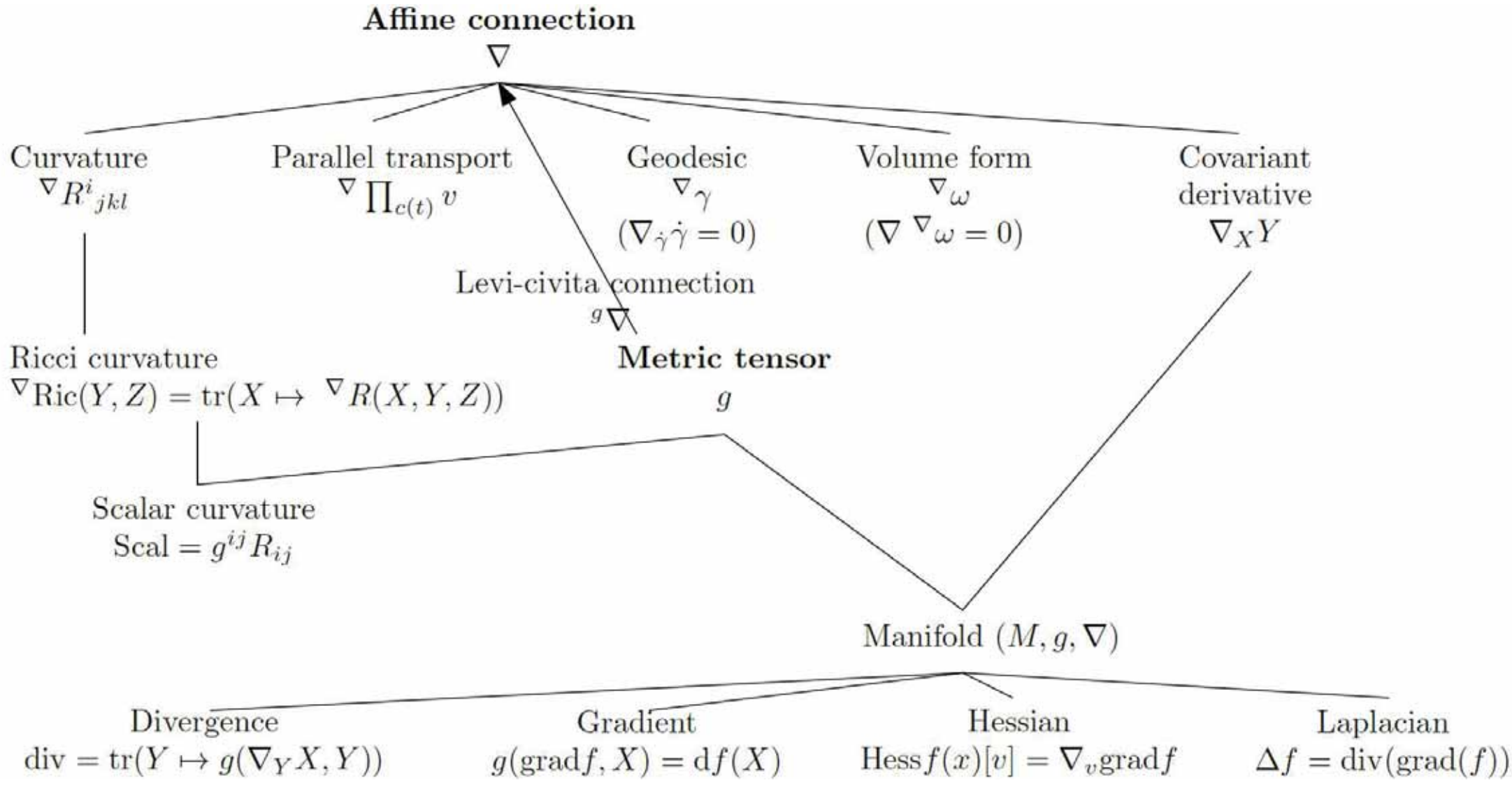
$$\boxed{\frac{1}{n}\log\frac{p_1(X_{1:n})}{p_2(X_{1:n})} \to D_{\mathrm{KL}}[p_1 : p_2], \quad (\forall p_2(x))}$$

Affine connection: Curvature and torsion

- Termed by Hermann K. Weyl[1] (1922) and fully investigated by Elie J. Cartan[2] (1923) for general relativity of spacetime with torsion = Einstein-Cartan theory
  $\rightarrow$ Palatini approach decoupling metric and connection in GR
- Affine connection: differential operator ∇ on tangent bundle (stitches infinitesimally close tangent spaces handled as affine spaces) allowing to define (1) covariant derivatives, (2) parallel transport, and (3) ∇-geodesics
- Parallel transport on an infinitesimally small loop yields an **affine displacement**: torsion is expressed by a **translation** and *curvature* by a **rotation**, also *homothetic curvature*
- Connections differing only by torsion have (pre)geodesics coinciding
- Riemann geometry (Einstein GR) is a special case using the torsion-free Levi-Civita metric connection.

[1] An elementary introduction to information geometry, `https://arxiv.org/abs/1808.08271`

[2] Sur les variétés à connexion affine et la théorie de la relativité généralisée, Annales scientifiques de l'ENS. Vol. 40. 1923.


Affine connection
∇
Curvature
${}^{\nabla}R^{i}{}_{jkl}$
Parallel transport
${}^{\nabla}\prod_{c(t)} v$
Geodesic
${}^{\nabla}\gamma$
$(\nabla_{\dot\gamma}\dot\gamma = 0)$
Volume form
${}^{\nabla}\omega$
$(\nabla\, {}^{\nabla}\omega = 0)$
Covariant derivative
$\nabla_X Y$
Levi-civita connection
${}^{g}\nabla$
Ricci curvature
${}^{\nabla}\mathrm{Ric}(Y,Z) = \mathrm{tr}(X \mapsto {}^{\nabla}R(X,Y,Z))$
Metric tensor
g
Scalar curvature
$\mathrm{Scal} = g^{ij}R_{ij}$
Manifold (M, g, ∇)
Divergence
$\mathrm{div} = \mathrm{tr}(Y \mapsto g(\nabla_Y X, Y))$
Gradient
$g(\mathrm{grad} f, X) = \mathrm{d}f(X)$
Hessian
$\mathrm{Hess} f(x)[v] = \nabla_v \mathrm{grad} f$
Laplacian
$\Delta f = \mathrm{div}(\mathrm{grad}(f))$

Affine immersion, conormal map & geometric divergence

- Dual vector space of $\mathbb{R}^D$ covector space of linear maps $\mathbb{R}_D := (\mathbb{R}^D)^*$
- Natural pairing product: $(l, v\rangle := l(v), \quad \forall l \in \mathbb{R}_D, \forall v \in \mathbb{R}^D$
- D-dimensional manifold M
- $\{i, \xi\}$: Immersion $i : M \to \mathbb{R}^{D+1}$ with transversal vector field xi ξ :

$$T_{i(p)}\mathbb{R}^{D+1} = i_*(T_pM) \oplus \mathbb{R}\{\xi_p\}$$

- **Conormal map** $c : M \to \mathbb{R}_{D+1}$ for non-degenerate equiaffine immersion:

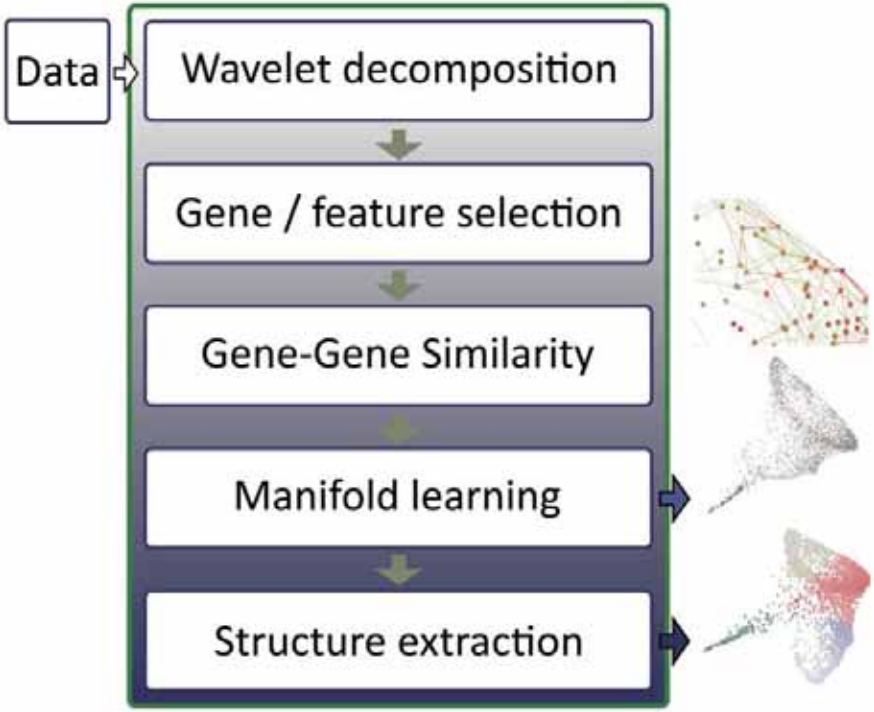
$$(c(p), \xi_p\rangle = 1, (c(p), i_*X_p\rangle = 0$$

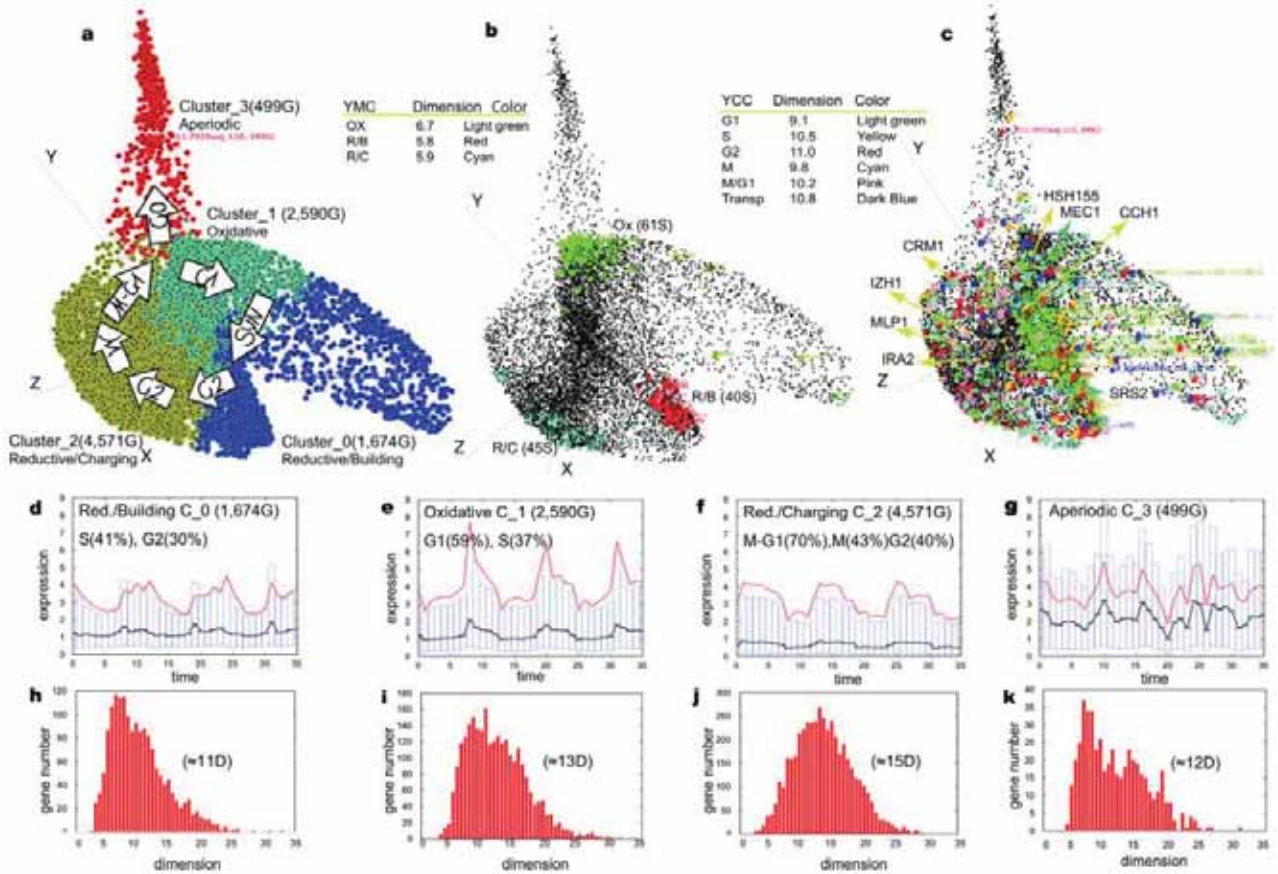
- **Geometric divergence** $\rho : M \times M \to \mathbb{R}_+$ $\boxed{\rho(p : q) := (c(p), i(p) - i(r)\rangle}$

- Dually flat space : i(θ)=(θ,F(θ)) and ξ=(0,…,0,1) → ρ is the **Bregman divergence**

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10)

AGCT: A Geometric Clustering Tool





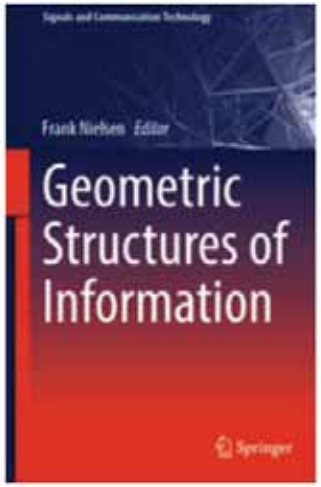
Nock R, Polouliakh N, Nielsen F, Oka K, Connell CR, Heimhofer C, et al. (2020) A Geometric Clustering Tool (AGCT) to robustly unravel the inner cluster structures of time-series gene expressions. PLoS ONE 15(7): e0233755. https://doi.org/10.1371/journal.pone.0233755

All norms are equivalent in finite dimensions

- In a *finite-dimensional vector space V*, all norms are **equivalent**: consider two norms $\|v\|_A$ and $\|v\|_B$. Then we have

for all v in V,

$$\exists 0 < c, C < \infty \qquad c\|v\|_A \leq \|v\|_B \leq C\|v\|_A$$



- Many applications: for example, **k-means++** in a finite-dimensional vector space yields O(log k) approximation factor with high probability.

Theorem 4 (k-means++ performance in a normed space) *In any normed space* $(\mathscr{X}, \|\cdot\|)$, *the k-means++ with* $D(x:y) = \|x-y\|^2$ *is* $16(2+\log k)$*-competitive.*

FN and Ke Sun, Clustering in Hilbert simplex geometry, arXiv:1704.00454
(Geometric Structures of Information 2019, Springer https://www.springer.com/gp/book/9783030025199)
FN and Richard Nock, Total Jensen divergences: Definition, properties and k-means++ clustering, arXiv:1309.7109
(ICASSP 2015)

Fat objects for slimming complexity: α-fat objects and (β,δ)-covered objects

- Goal: Design efficient algorithms and data-structures for real-world input data sets. Avoid pathological synthetic worst-cases.
- Object O is **α-fat** if the ratio of smallest enclosing ball on the largest inscribed ball is greater or equal to α.
- A triangle Δ is a (β, δ)-**triangle** of a planar object O (with $0 < \beta \leq \frac{\pi}{3}$ and $0 < \delta < 1$) if
  - $\Delta \subseteq O$,
  - each of the angles of Δ is at least β, and
  - the length of each of its edges is at least δ diameter(O).

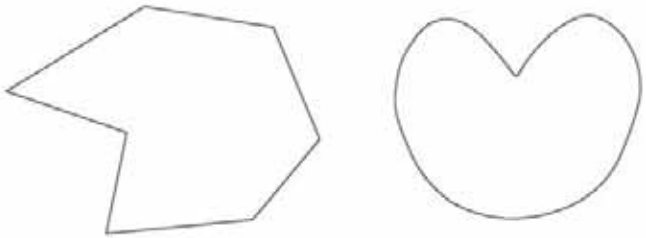


Fig. Two (β, δ)-covered objects.

Object O is (β, δ)-**covered** if for each point $p \in O$ there exists a (β, δ)-triangle that contains p.

Dynamic data structures for fat objects and their applications, Computational Geometry, 2000

α-representations of the Fisher Information Matrix

Usually, the Fisher Information Matrix (FIM) is introduced in two ways:

$$I(\theta) := (I_{ij}(\theta)), \quad I_{ij}(\theta) := E_{p(x;\theta)}[\partial_i l(x;\theta)\partial_j l(x;\theta)]$$

$$I'_{ij}(\theta) := 4\int \partial_i \sqrt{p(x;\theta)}\partial_j \sqrt{p(x;\theta)}\mathrm{d}\nu(x)$$

α-likelihood function $l^{(\alpha)}(x;\theta) := k_\alpha(p(x;\theta))$

α-Embedding

$$k_\alpha(u) = \begin{cases} \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1 \\ \log u, & \text{if } \alpha = 1. \end{cases}$$

α-representation of the FIM:

$$\boxed{I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)}(x;\theta)\partial_j l^{(-\alpha)}(x;\theta)d\nu(x)}$$

Corresponds to a basis choice in the tangent space (α-base)

@FrnkNlsn

https://tinyurl.com/yyukx86o

The many meanings of α in information geometry

- A scalar to parameterize **a family of α-divergences** including the Kullback-Leibler divergence and the reverse Kullback-Leibler divergence:

$$D_\alpha[p:q] := \begin{cases} \frac{4}{1-\alpha^2}\int\left(\frac{1-\alpha}{2}p+\frac{1+\alpha}{2}q-p^{\frac{1-\alpha}{2}}q^{\frac{1+\alpha}{2}}\right)\mathrm{d}\mu, & \alpha\in\mathbb{R}\backslash\{-1,1\} \\ D_1[p:q]=D_{\mathrm{KL}}[p:q] & \alpha=1 \\ D_{-1}[p:q]=D_{\mathrm{KL}}[q:p] & \alpha=-1 \end{cases}$$

- A scalar to parameterize **a family of torsion-free affine α-connections** from a pair of dual connections ∇ and ∇*: $\nabla^\alpha := \frac{\nabla+\nabla^*}{2}$

- A scalar to parameterize an **α-representation** defining vector basis of tangent spaces with Fisher inner product equivalent to an **α-expectation** on the **α-representation of tangent vectors** (gauge freedom): $a = a^i\partial_i \Rightarrow a_\alpha(x) = a^i\partial_i l_\alpha(x,\theta)$

$$h_\alpha(u) = \begin{cases} \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}} & \alpha\neq 1 \\ \log u & \alpha = 1 \end{cases} \Rightarrow l_\alpha(x,\theta) = h_\alpha(p_\theta(x))$$

$$b = b^i\partial_i \Rightarrow b_\alpha(x) = b^i\partial_i l_\alpha(x,\theta)$$

$$\langle a,b\rangle_{g_{\mathrm{Fisher}}(\theta)} := E^\alpha_{p_\theta}\left[a_\alpha(x)b_\alpha(x)\right] \qquad E^\alpha_{X\sim p}[f(x)] := \int p^\alpha(x)f(x)\mathrm{d}\mu(x)$$

- **Duality** α ↔ -α : $D_\alpha[q:p] = D_{-\alpha}[p:q]$ $\quad \nabla^0 = \nabla^{\mathrm{LC}}$ $\quad \partial_i l_\alpha \partial_j l_{-\alpha} = p_\theta \partial_i l \partial_j l$

An Elementary Introduction to Information Geometry, Entropy 2020

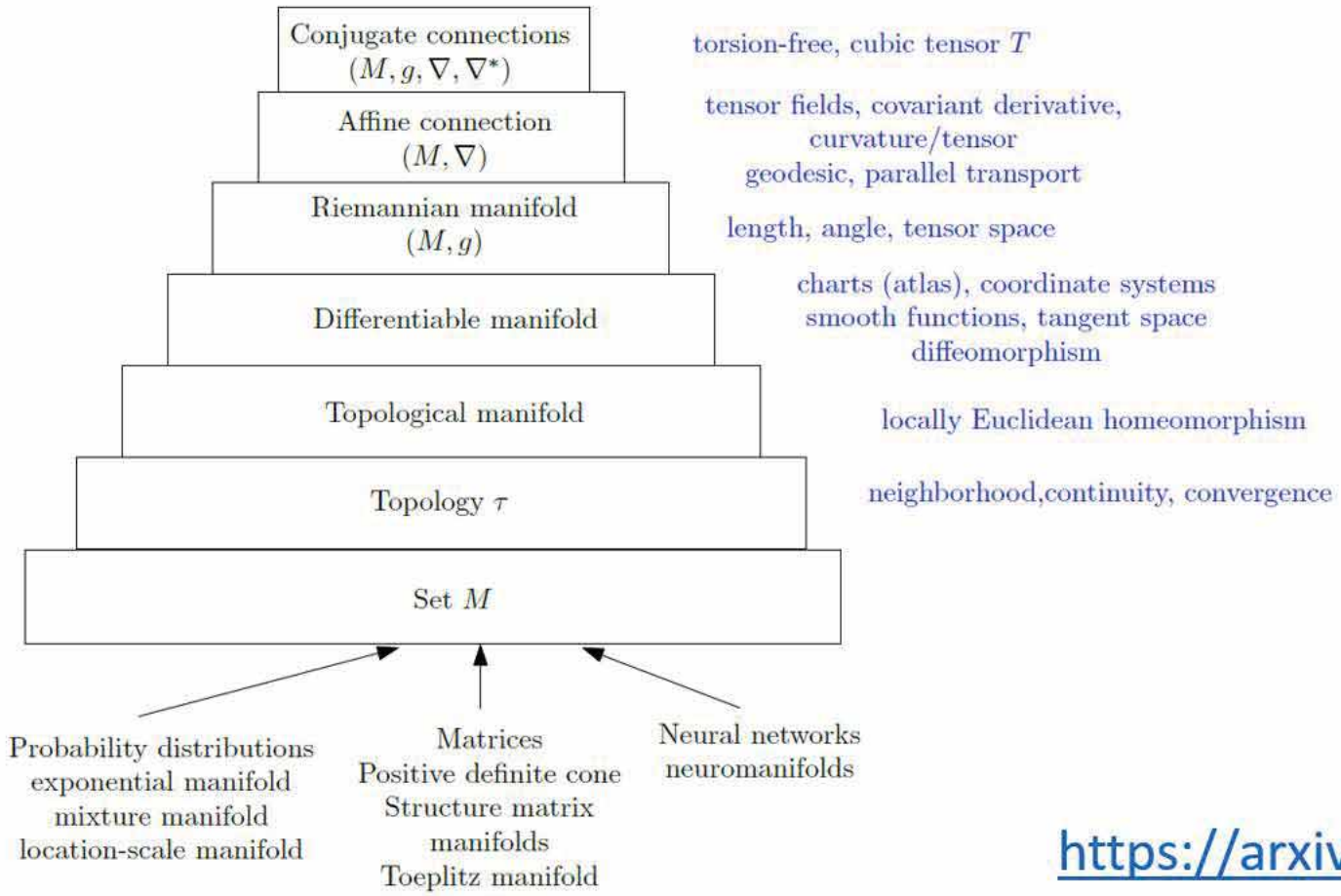
The α-divergences associated with a pair of strictly comparable quasi-arithmetic means, arxiv:2001.09660

α-topology in information geometry

- **Topology** T on set X: a collection of subsets of X called **open subsets** (defining neighborhoods) such that any **arbitrary union** and **finite intersection** of open subsets belong to T. Topology provides notions of **continuity** and **convergence**
- **Topology T(S) generated by S:** coarsest topology on X for which every element of S is open. Metric spaces are topological spaces generated by the collection of open balls (eg., total variation topology).
- **f-topology**: topology generated by open f-balls, open balls wrt ***f-divergences***
- A topology T is **stronger** than a topology T' if T contains all the open sets of T'
- **Csiszar's theorem**: When $|\alpha|<1$, the α -topology is equivalent to the total variation metric topology. Otherwise, the α-topology is stronger than the TV topology.

- Csiszár, I. (1967) Information-Type Measures of Difference of Probability Distributions and Indirect Observations Studia Scientiarum Mathematicarum Hungarica, 2, 299-318.
- An elementary introduction to information geometry, arXiv:1808.08271

Anatomy of a statistical manifold



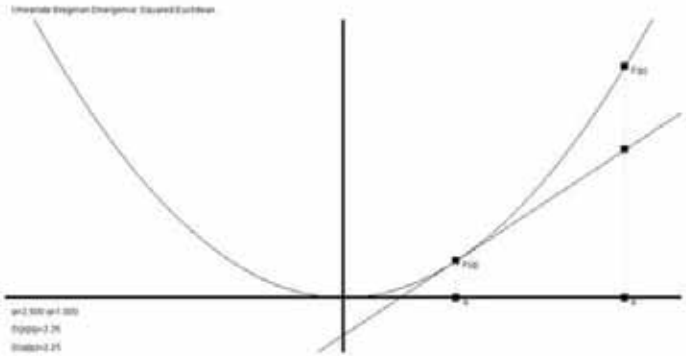
https://arxiv.org/abs/1808.08271

Two approximations of the Bregman divergences

- **Bregman divergences** for smooth strictly convex generators F:

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

→ Require the gradients ∇F



- **Scaled skew Jensen divergences**:

$$\begin{aligned} J'_\alpha(\theta_1 : \theta_2) &= \alpha F(\theta_1) + (1-\alpha) F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2), \quad \alpha \notin \{0,1\} \\ &= (F(\theta_1)F(\theta_2))_\alpha - F\left((\theta_1\theta_2)_\alpha\right) \end{aligned}$$

$$J_\alpha(\theta_1 : \theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J'_\alpha(\theta_1 : \theta_2) & \alpha \neq \{0,1\} \\ B(\theta_1 : \theta_2) & \alpha = 0 \\ B(\theta_2 : \theta_1) & \alpha = 1 \end{cases}$$

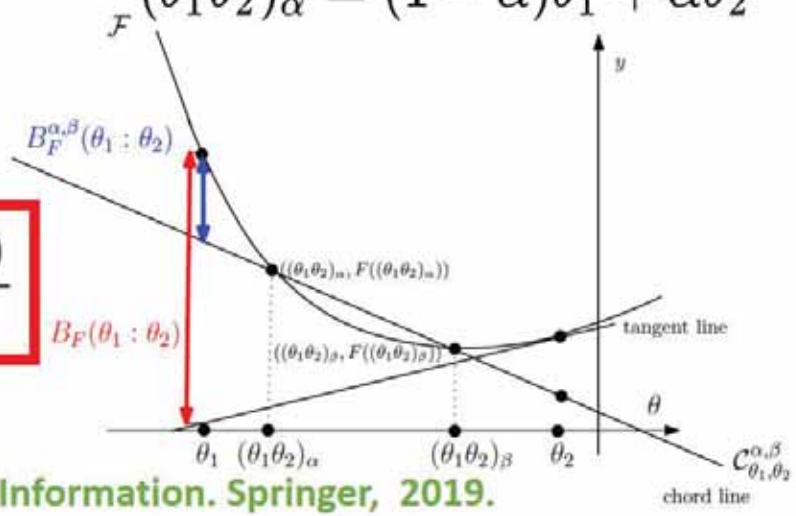
Linear interpolation notation:

$$(\theta_1\theta_2)_\alpha = (1-\alpha)\theta_1 + \alpha\theta_2$$

- ***Bregman chord divergences***:

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

$$B_F(\theta_1 : \theta_2) \simeq_{\epsilon \to 0} B_F^{1-\epsilon,1}(\theta_1 : \theta_2)$$

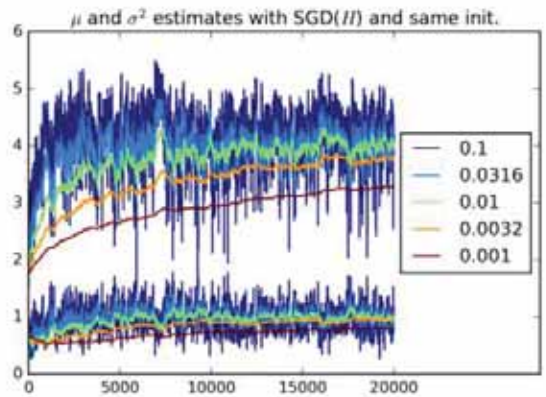


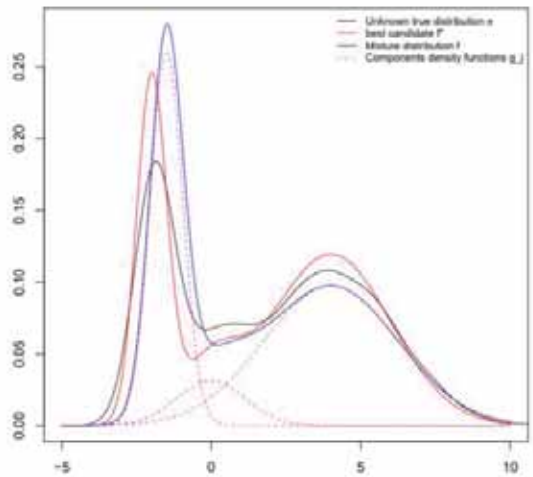
The Bregman chord divergence, International Conference on Geometric Science of Information. Springer, 2019.

Batch and online statistical mixture learning

$$\pi(x) \approx f(x;\theta) = \sum_{j=1}^{K} w_j g_j(x;\theta_j), \text{ with } w_j > 0 \text{ and } \sum_{j=1}^{K} w_j = 1.$$

- Gradient-based optimization
- Stochastic gradient descent methods:
  - Minibatch SGD
  - Momentum SGD
  - Average SGD
  - Adam (adaptive moment estimation)





(Adam) This method (named after adaptive moment estimation) is a first-order method which use estimates of first and second moments of the gradient w.r.t. each parameter to estimate.
At iteration t:

$$\boxed{\begin{aligned} &\text{for } j = 1,2 \qquad m_j^{(t+1)} = \beta_j\, m_j^{(t)} + (1-\beta_j)\left(\nabla_\theta C(x_t, \hat\theta^{(t)})\right)^{\circ j} \\ &\qquad\qquad\quad \hat m_j^{(t+1)} = m_j^{(t+1)}/(1-\beta_j^t) \\ &\hat\theta^{(t+1)} = \qquad \hat\theta^{(t)} - \alpha\, \hat m_1^{(t+1)} / \left(\sqrt{\hat m_2^{(t+1)}} + \epsilon\right) \end{aligned}} \qquad (11)$$

Algorithm 1: Exact Online MLE for regular exponential families

Input: a sequence $\mathcal{S} = x_1, x_2, \ldots$ of observations
Input: Functions s and $(\nabla F)^{-1}$ for some regular exponential family
Output: a sequence $\hat\eta^{(1)}, \hat\eta^{(2)}, \ldots$ of MLE where $\hat\eta^{(N)}$ is the exact MLE for the first N observations

$\hat\eta^{(0)} = 0;\quad N = 0;$
for $x_{N+1} \in \mathcal{S}$ **do**
  $\hat\eta^{(N+1)} = \hat\eta^{(N)} + \{N+1\}^{-1}(s(x_{N+1}) - \hat\eta^{(N)});$
  yield $\hat\eta^{(N+1)}$ or **yield** $\hat\theta^{(N+1)} = (\nabla F)^{-1}(\hat\eta^{(N+1)});$
  $N = N + 1;$

Batch and Online Mixture Learning: A Review with Extensions, Computational Information Geometry,Springer, 2017

Information Theory/Statistics viewpoint: [Azoury & Warmuth 2001]

KLD between exponential family densities = **reverse Bregman divergence**

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1)$$

Information geometry viewpoint: [Amari 2016, Nielsen 2020]

Dually flat space from convex cumulant function = **reverse KL divergence**

$$F(\theta) = \log\left(\int \exp\left(\sum_{i=1}^{D} \theta_i t_i(x)\right) \mathrm{d}\mu(x)\right)$$

Strictly convex cumulant function

Dually flat space

Canonical Bregman divergence

Reconstruct reverse KL

$$\boxed{B_F(\theta_1 : \theta_2) = D^*_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = D_{\mathrm{KL}}[p_{\theta_2} : p_{\theta_1}]}$$

Bregman divergences & generalized cross/relative entropies

- $(\mathcal{X}, 2^{\mathcal{X}}, \mu_{\#})$: probability space with the **counting measure** $\mu_{\#}$ on a finite alphabet X
- **Bregman divergence** for a strictly convex and smooth generator F(θ):

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

- Bregman divergence induces
  - A generalized strictly concave discrete entropy: $h_F(\theta) := -F(\theta)$
  - A discrete cross-entropy: $h_F^\times(\theta : \theta') := -F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$
  - A generalized discrete relative entropy: $\boxed{B_F(\theta : \theta') = h_F^\times(\theta : \theta') - h_F(\theta)}$
- **Not to be confused** with the Shannon entropy of an exponential family density:

$$p_\theta(x) = \exp(\theta^\top t(x) - F(\theta) + k(x)) \Rightarrow \boxed{h_{\text{Shannon}}(p_\theta) = -F^*(\eta) - E_{p_\theta}[k(x)]}$$

For example: $k_{\text{Poisson}}(x) = -\log(x!)$

Entropies and cross-entropies of exponential families, ICIP 2010

Binary entropy of Bernoulli and the Choose function

- Bernoulli distribution $X \sim \mathrm{Bernoulli}(p)$, p in (0,1) [Probability of "head"]
- Binary entropy: $H(X) = -p \log p - (1-p)\log(1-p)$
- Choose function: $\binom{n}{k} := \frac{n!}{k!(n-k)!}$
- Interpreting the Bernoulli entropy by means of the choose function:

$$\boxed{H(X) = \lim_{n\to\infty} \frac{1}{n} \log \binom{n}{np}}$$

- Proof using Stirling's approximation of the factorial

Cumulant-free formula for common (dis)similarities

$$D_{\mathrm{KL}}[p:q] := \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x), \qquad \text{(Kullback-Leibler divergence)}$$

$$D_J[p,q] := D_{\mathrm{KL}}[p:q] + D_{\mathrm{KL}}[q:p], \qquad \text{(Jeffreys' divergence)}$$

$$\rho[p,q] := \int_{\mathcal{X}} \sqrt{p(x)\ q(x)}\mathrm{d}\mu(x), \qquad \text{(Bhattacharyya coefficient)}$$

$$D_B[p,q] := -\log(\rho[p,q]), \qquad \text{(Bhattacharyya distance)}$$

$$D_H[p,q] := \sqrt{1-\rho[p,q]}. \qquad \text{(Hellinger distance)}$$

Exponential family:

$$p(x;\lambda) = \exp\left(\theta(\lambda)^\top t(x) - F(\theta(\lambda)) + k(x)\right)$$

characterized by its cumulant function F:

$$F(\theta) := \log\left(\int_{\mathcal{X}} \exp(\theta^\top t(x))\mathrm{d}\mu(x)\right)$$

Usual formula:

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_{F^*}(\theta_1 : \theta_2),$$

$$D_J[p_{\theta_1} : p_{\theta_2}] = (\theta_2 - \theta_1)^\top(\eta_2 - \eta_1),$$

$$\rho[p_{\theta_1} : p_{\theta_2}] = \exp(-J_F(\theta_1 : \theta_2)),$$

$$D_B[p_{\theta_1} : p_{\theta_2}] = J_F(\theta_1 : \theta_2)$$

$$D_H[p_{\theta_1} : p_{\theta_2}] = \sqrt{1-\exp(-J_F(\theta_1 : \theta_2))},$$

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

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

Cumulant-free formula:

$$\forall \omega \in \mathcal{X}, \quad \rho[p_{\lambda_1}, p_{\lambda_2}] = \frac{\sqrt{p_{\lambda_1}(\omega)p_{\lambda_2}(\omega)}}{p_{\bar\lambda}(\omega)} = \sqrt{\frac{p_{\lambda_1}(\omega)}{p_{\bar\lambda}(\omega)}}\sqrt{\frac{p_{\lambda_2}(\omega)}{p_{\bar\lambda}(\omega)}}$$

Quasi-arithmetic means:

$$\bar\lambda = \theta^{-1}\left(\frac{\theta(\lambda_1) + \theta(\lambda_2)}{2}\right) =: M_\theta(\lambda_1, \lambda_2)$$

$$\forall \omega \in \mathcal{X}, \quad D_B[p_{\lambda_1}, p_{\lambda_2}] = \log\left(\frac{p(\omega; M_\theta(\lambda_1,\lambda_2))}{\sqrt{p(\omega,\lambda_1)p(\omega,\lambda_2)}}\right)$$

$$\omega \in \mathcal{X}_\perp^{\mathcal{E}}(\lambda_1 : \lambda_2) \Rightarrow D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \log\left(\frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)}\right)$$

$$\mathcal{X}_\perp^{\mathcal{E}}(\lambda_1 : \lambda_2) := \left\{\omega \in \mathcal{X} \ : \ (\theta(\lambda_2) - \theta(\lambda_1))^\top(t(\omega) - \nabla F(\theta(\lambda_1))) = 0\right\}.$$

Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family
https://arxiv.org/abs/2003.02469

Lie groups admitting bi-invariant metrics

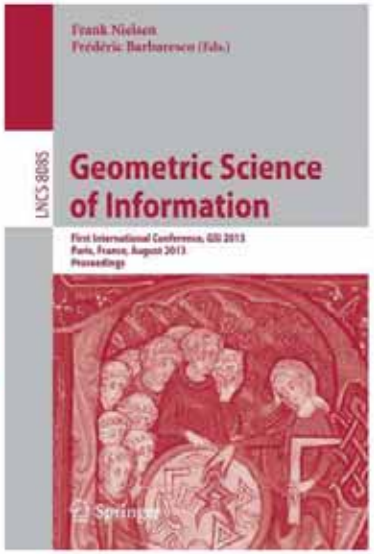
Sophus Lie
1842-1899

- Left translations $L_p(x)=p.x$, right translations $R_p(x)=x.p$ and conjugations $C_p(x)=p.x.p^{-1}$ on a Lie group G are **diffeomorphisms**
- Riemannian metric g on Lie group G is **left invariant** if and only if left translation is an isometry for p in G
- Riemannian metric g on Lie group G is **right invariant** if and only if right translation is an isometry for p in G
- Metric g is **bi-invariant** if and only if it is both left and right invariant
- **Lie groups always admit left-invariant metrics** (any inner product on the Lie algebra pushed forward using left translation)
- **Theorem: Compact Lie groups, commutative Lie groups and their products are the only Lie groups admitting bi-invariant metrics**
- When a Lie group admits a bi-invariant metric then the **Lie exponential coincides with the Riemannian exponential** at the identity element e in G

Lie groups admitting bi-invariant metrics

- Left translations $L_p(x)=p.x$, right translations $R_p(x)=x.p$ and conjugations $C_p(x)=p.x.p^{-1}$ on a Lie group G are **diffeomorphisms**
- Riemannian metric g on Lie group G is **left invariant** if and only if left translation is an isometry for p in G
- Riemannian metric g on Lie group G is **right invariant** if and only if right translation is an isometry for p in G
- Metric g is **bi-invariant** if and only if it is both left and right invariant
- **Lie groups always admit left-invariant metrics** (any inner product on the Lie algebra pushed forward using left translation)
- **Theorem: Compact Lie groups, commutative Lie groups and their products are the only Lie groups admitting bi-invariant metrics**
- When a Lie group admits a bi-invariant metric then the **Lie exponential coincides with the Riemannian exponential** at the identity element e in G

Sophus Lie
1842-1899



What is information geometry (IG)?

- Geometry of families of distributions: Term **geometrostatistics** coined by Kolmogorov for referring to the work of Chentsov, appeared in the preface to N. N. Chentsov's Russian book (1972) but lost in the English translation.

Предмет настоя[illegible]е всего было бы окрестить термино[illegible] «геометростатистика» В ней систематически изучается ал[illegible] и порождаемая ею

[illegible] name for the subject of this monograph might b[illegible] "geometrical statistics" The algebra of decision rules and the natural geometry that it

Precursors: Hotelling, Rao, Chentsov, Efron, Dawid, Barndorff-Nielsen, etc.

- Geometry of models + dualistic differential-geometric structure: Term **information geometry** appeared in the preface of S-.i Amari's book (1985)

Precursors: Norden, Sen, Chentsov, Amari, Nagaoka, etc.

Caianiello (1983) for applications to physics. They together will open a new field, which I would like to call information geometry.

An elementary introduction to information geometry, 2018

On geodesic triangles with right angles in a dually flat space, 2019

Biskewed Kullback-Leibler divergences and 1D mixture families of information geometry

- **Biskewed Kullback-Leibler divergence**[2] with $\alpha, \beta \in (0,1)$, $\alpha \neq \beta$

  $$\mathrm{KL}_{\alpha,\beta}(p:q) := \mathrm{KL}((1-\alpha)p+\alpha q : (1-\beta)p+\beta q) = \mathrm{KL}(m_\alpha : m_\beta),$$

  where $m_\theta(x) = (1-\alpha)p(x) + \alpha q(x)$ is the mixture and $\{m_\theta \ : \ \theta \in (0,1)\}$ a 1D **mixture family** ($q \neq p$ a.e.)
- Biskewed Kullback-Leibler divergence amounts to a **Bregman divergence**[3]:

  $$\mathrm{KL}_{\alpha,\beta}(p:q) = B_F(\alpha:\beta) = F(\alpha) - F(\beta) - (\alpha-\beta)F'(\beta)$$

  for $F(\theta) = \int m_\theta(x) \log m_\theta(x) \mathrm{d}\mu(x)$ (Shannon negentropy)
- **centroid** minimizing $\sum_i w_i \mathrm{KL}_{\alpha_i,\beta}(p:q) = \sum_i w_i B_F(\alpha_i:\beta)$ is a **right-sided Bregman centroid**[4] $m_{\bar\theta}$ with $\bar\theta = \sum_i w_i \alpha_i$
  $\Rightarrow$ **DON'T NEED F, which is usually not in closed form!!!**

[2] https://www.mdpi.com/1099-4300/22/2/221
[3] On the geometry of mixtures of prescribed distributions, ICASSP 2018
[4] Sided and Symmetrized Bregman Centroids, IEEE TIT 2009

Biskewed Kullback-Leibler divergences[1]

- **Biskewed Kullback-Leibler divergence**: $\alpha, \beta \in [0,1]$, $\alpha \neq \beta$

$$\boxed{\mathrm{KL}_{\alpha,\beta}(p:q) := \mathrm{KL}((1-\alpha)p + \alpha q : (1-\beta)p + \beta q)}$$

- Extend the ordinary $\mathrm{KL}(p:q) = \mathrm{KL}_{0,1}(p:q)$ with reference duality $\mathrm{KL}_{\alpha,\beta}(q:p) = \mathrm{KL}_{1-\alpha,1-\beta}(p:q)$. Always upper bounded when $\alpha, \beta \in (0,1)$ (but not KL).
- $\mathrm{KL}_{\alpha,\beta}$ is a f-**divergence**: $\mathrm{KL}_{\alpha,\beta}(p:q) = I_{f_{\alpha,\beta}}(p:q)$ with

$$\boxed{f_{\alpha,\beta}(u) = (\alpha u + 1 - \alpha) \log \frac{(1-\alpha) + \alpha u}{(1-\beta) + \beta u}}$$

- Write the **Jensen-Shannon divergence** as

$$\mathrm{JS}(p:q) = \frac{1}{2}\mathrm{KL}_{0,\frac{1}{2}}(p:q) + \frac{1}{2}\mathrm{KL}_{1,\frac{1}{2}}(p:q)$$

$\Rightarrow$ JS f-generator: $f_{\mathrm{JS}}(u) = \log \frac{2}{1+u} + u \log \frac{2u}{1+u}$

[1] On a Generalization of the Jensen-Shannon Divergence and the Jensen-Shannon Centroid, Entropy 2020
https://www.mdpi.com/1099-4300/22/2/221

Bonnet theorem:

Determination of 3D surfaces

- **Bonnet theorem**:

First fundamental form (metric g) and second fundamental form (II) determine uniquely a surface of R^3 up to a rigid motion

Pierre Ossian Bonnet
1819-1892

- Former student of Ecole Polytechnique (X1838)

- Bonnet theorem not to be confused with Gauss-Bonnet theorem

Boosting and additive modeling in machine learning

- Boosting is rooted in Valiant's PAC model
- Breakthrough in ML with AdaBoost (demonstrate a strong classifier from weak classifiers)
- Unify greedy boosting for decision trees with additive models
- Dual form of convex optimization

Input: $\mathcal{S}, \boldsymbol{w}_1$; **for** $t = 1, 2, ..., T$ $h_t \leftarrow \text{Tree}(\mathcal{S}, \boldsymbol{w}_t)$; $\alpha_t \leftarrow \arg\min_{\alpha \in \mathbb{R}} E_{\boldsymbol{w}_t}(\exp(-y\alpha h_t(\boldsymbol{x})))$; **for** $i = 1, 2, ..., m$ $w_{t+1,i} \leftarrow w_{t,i} \exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))/Z_t$;	**Input**: $\mathcal{S}, \boldsymbol{w}_1$; **for** $t = 1, 2, ..., T$ **let** $\ell \in \text{leaves}(H_{t-1})$ containing examples of both classes; $h_t \leftarrow \text{Stump}(\mathcal{S}_\ell, \boldsymbol{w}_1)$; $\ell \leftarrow h_t$;	**Input**: $\mathcal{S}, \boldsymbol{w}_1$; **for** $t = 1, 2, ..., T$ **let** $\ell \in \text{leaves}(H_{t-1})$ containing examples of both classes; $h_t \leftarrow \text{StumpLS}(\mathcal{S}_\ell, \boldsymbol{w}_1)$; $\ell \leftarrow h_t$;
AdaBoost + Trees	TopDown_DT (CART, C4.5)	TopDown_ODT (OC1, SADT)

The phylogenetic tree of boosting has a bushy carriage but a single trunk, PNAS letter, 2020
Bregman divergences and surrogates for learning, TPAMI 2009
Real boosting a la carte with an application to boosting oblique decision trees, IJCAI 2017

Infinitely many upper bounds on the entropy of a distribution: Applications to bounding the entropy of statistical mixtures

- **Simple trick**: Consider any **maximum entropy distribution** p (satisfying a moment constraint) for which we know a closed-form solution for the entropy.
- We use **exponential families with absolute monomials of order l** for which we have a closed-form of its entropy:

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

$$H(X) = \int_{\mathcal{X}} p(x) \log \frac{1}{p(x)} \mathrm{d}x \quad \Rightarrow \quad H_l(\theta) = \log 2 + \log \Gamma\left(\frac{1}{l}\right) - \log l + \frac{1}{l}(1 - \log(-\theta))$$

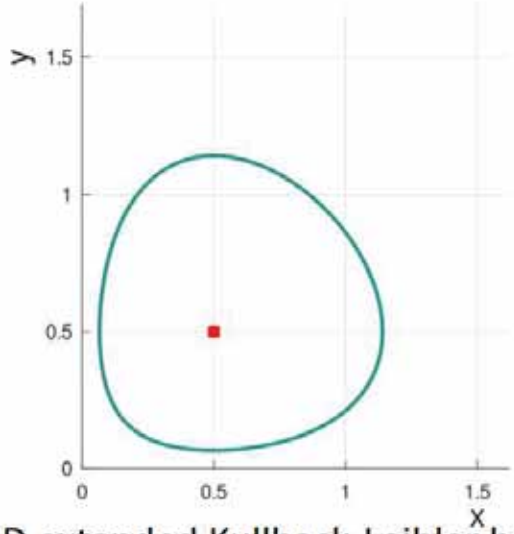
- Then any other distribution X (say, a mixture) has necessarily entropy less or equal to p for the same moment constraint: $H(X) \leq H_l^{\eta}\left(E_X\left[|X|^l\right]\right)$
- For bounding the entropy of a mixture, we need to calculate the absolute moment of order l of the mixture components. The upper bound is the minimum of all upper bounds.

MaxEnt upper bounds for the differential entropy of univariate continuous distributions, IEEE SPL 2017

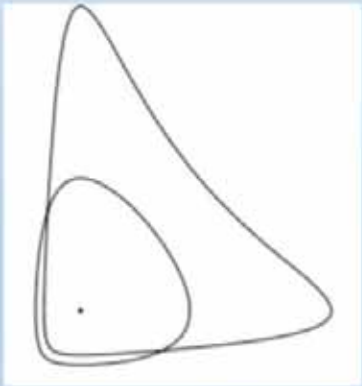
Drawing and printing Bregman balls...

```
# Draw using an implicit function an extended Kullback-Leibler ball in Octave
clear, clf, cla
figure('Position',[0,0,512,512]);
xm = 0:0.01:3;
ym = 0:0.01:3;
[x, y] = meshgrid(xm, ym);
xc = 0.5;
yc = 0.5;
r = 0.3;
f= (x.*log(x./xc).+xc.-x).+(y.*log(y./yc).+yc.-y)-r;
contour(x,y,f,[0,0],'linewidth',2)
grid on
xlabel('x', 'fontsize',16);
ylabel('y', 'fontsize',16)
hold on
plot(xc,yc,'+1','linewidth',5);
print ("eKL-ball.pdf", "-dpdf");
print('eKL-ball.png','-dpng','-r300');
hold off
```

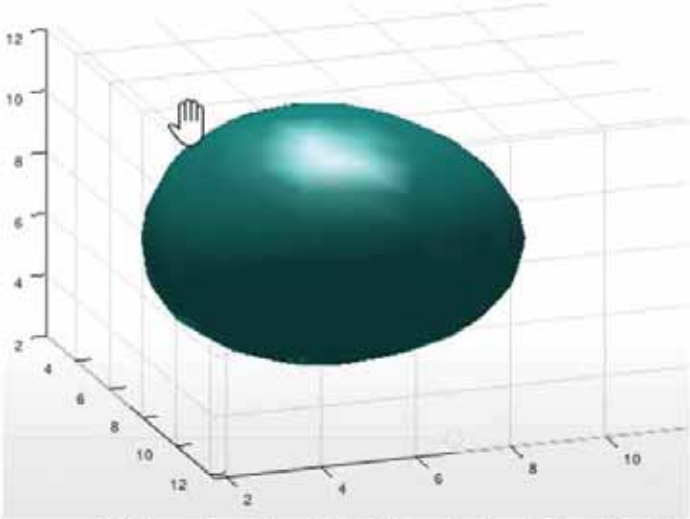
3D printing balls: KL, Itakura-Saito, Logistic



2D extended Kullback-Leibler ball



Itakura-Saito dual balls



3D extended Kullback-Leibler ball

Bregman Voronoi diagrams, Discrete & Computational Geometry (2010)

Dual Bregman sided centroids C_l and C_r[1] (1/2)

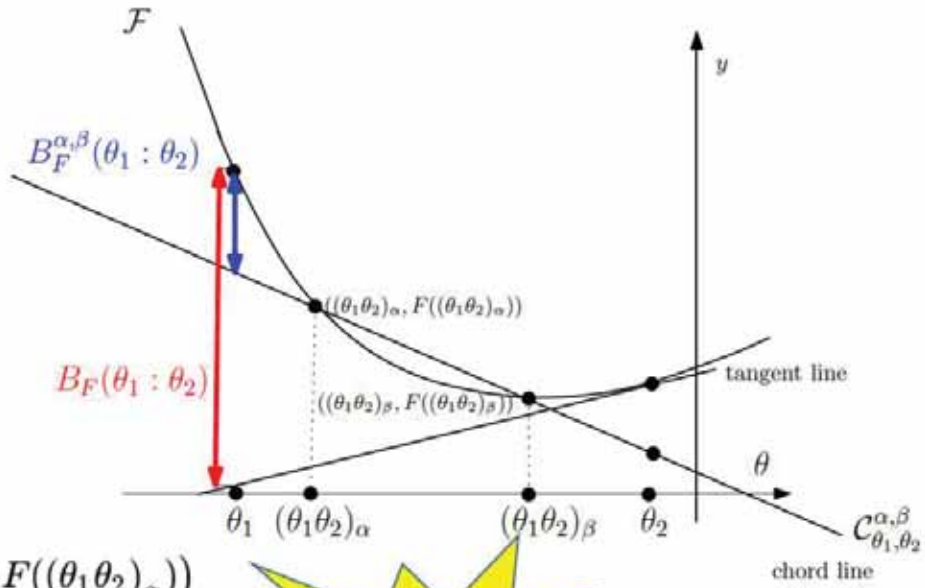
- **Bregman divergence**:
  $B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$ for strictly convex and differentiable generator F
- **Reverse Bregman divergence**:
  $B_F^r(\theta : \theta') := B_F(\theta' : \theta :) = B_{F^*}(\eta : \eta')$ where $\eta = \nabla F(\theta)$ and F^* is Legendre-Fenchel convex conjugate. $\nabla F^* = (\nabla F)^{-1}$.
- **Canonical dually flat divergence**:
  $A_F(\theta : \eta') = F(\theta) + F^*(\eta') - \theta^\top \eta = A_{F^*}(\eta' : \theta)$. We have $B_F(\theta : \theta') := A_F(\theta : \eta') = A_{F^*}(\eta' : \theta) = B_{F^*}(\eta' : \eta)$.
- Point P represented in **dual affine coordinate systems**:
  $\eta(P) = \nabla F(\theta(P))$ and $\theta(P) = \nabla F^*(\eta(P))$. Distance $D(P : P') = B_F(\theta(P) : \theta(P')) = A_F(\theta(P) : \eta(P'))$.
- **Left/right centroids** C_l / C_r for a distance D:
  $C_r = \arg\min_C \frac{1}{n} \sum_{i=1}^n D(P_i : C)$. $\theta(C_r) = \frac{1}{n} \sum_{i=1}^n \theta(P_i)$
  Similarly, $C_l = \arg\min_C \frac{1}{n} \sum_{i=1}^n D(C : C_i)$ with $\eta(C_l) = \frac{1}{n} \sum_{i=1}^n \eta(P_i)$
  Information radius $I_r(P_1, \ldots, P_n) = \frac{1}{n} \sum_{i=1}^n D(P_i : C_r)$.

[1] Sided and Symmetrized Bregman Centroids, IEEE TIT, 2009.

Bregman chord divergence: Free of gradient!

Ordinary Bregman divergence requires gradient calculation:

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



Bregman chord divergence uses two extra scalars α and β:

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

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

No gradient!

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

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

@FrnkNlsn

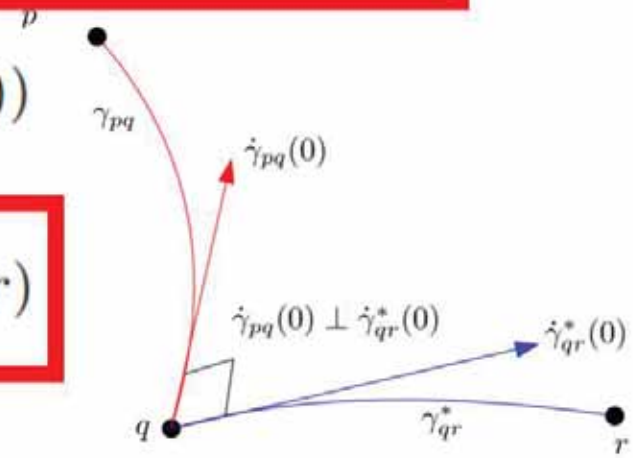
The Bregman chord divergence, arXiv:1810.09113

Bregman 3-parameter property: Generalized Law of cosines and Pythagoras' theorem

$$\boxed{B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3))}$$

$$D_F(p : q) = B_F(\theta(p) : \theta(q)) = D^*(q : p) = B_{F^*}(\eta(q) : \eta(p))$$

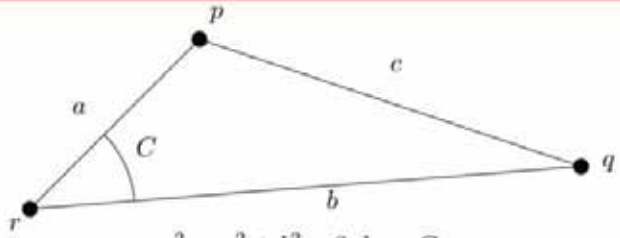
$$\boxed{\gamma_{pq} \perp_q \gamma^*_{qr} \quad \Longleftrightarrow \quad D_F(p : q) + D_F(q : r) = D_F(p : r)}$$



$D_F(p : r)) = D_F(p : q) + D_F(q : r)$

$B_F(\theta(p) : \theta(r)) = B_F(\theta(p) : \theta(q)) + B_F(\theta(q) : \theta(r))$

$(\theta(p) - \theta(q))^\top (\eta(r) - \eta(q)) = 0 \Leftrightarrow \dot\gamma_{pq}(0) \perp_q \dot\gamma^*_{qr}(0)$



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

Notice that when $F(\theta) = F_{\text{Euc}}(\theta) = \frac{1}{2}\theta^\top\theta$ (Euclidean geometry) and $\theta(r) = 0$, we get

$$\frac{1}{2}\|p - q\|^2 = \frac{1}{2}\|p\|^2 + \frac{1}{2}\|q\|^2 - \|p\|\|q\|\cos\alpha_r(p, q).$$ https://arxiv.org/abs/1910.03935

Dual parallel transport in a Bregman manifold

- Bregman manifold (=dually flat space): Two convex potential functions linked by Legendre transformations defining two dual global **Hessian structures**
- Primal/dual geodesics are straight in the primal/dual global affine coordinate charts

$$\begin{aligned}\gamma_{pq} &= \{x_\lambda \in M \ : \ \theta(x_\lambda) = (1-\lambda)\theta(p) + \lambda\theta(q), \quad \lambda \in [0,1]\},\\ \gamma^*_{pq} &= \{x_\lambda \in M \ : \ \eta(x_\lambda) = (1-\lambda)\eta(p) + \lambda\eta(q), \quad \lambda \in [0,1]\}.\end{aligned}$$

- Primal parallel transport of a vector does not change the contravariant vector components, and dual parallel transport does not change the covariant vector components. Because the dual connections are flat, path-independent dual parallel transports
- Property: Dual parallel transport preserves the metric:

$$\boxed{g_q\left(\prod_{p,q}(u), \overset{*}{\prod_{p,q}}(v)\right) = g_p(u,v).}$$

https://arxiv.org/abs/1910.03935

Bregman divergence: 4-parameter identity

In a Bregman manifold, divergences between points amount to **Bregman divergences** between corresponding parameters:

$$D_F(p:q) = B_F(\theta(p):\theta(q)) = F(\theta(p)) - F(\theta(q)) - (\theta(p)-\theta(q))^\top \nabla F(\theta(q)).$$

<u>4-parameter/4-point identity</u>: (Dual coordinate system $\eta(\theta) = \nabla F(\theta)$ **)**

$$\boxed{\begin{aligned} B_F(\theta(p_1):\theta(q_1)) + B_F(\theta(p_2):\theta(q_2)) \quad & -B_F(\theta(p_1):\theta(q_2)) - B_F(\theta(p_2):\theta(q_1)) \\ & -(\theta(p_2)-\theta(p_1))^\top(\eta(q_1)-\eta(q_2)) = 0. \end{aligned}}$$

Geometric interpretation:

$$D_F(p_1:p_3) - D_F(p_4:p_3) = D_F(p_1:p_2) - D_F(p_4:p_2) - g_{p_2}(\dot\gamma_{p_2p_3}(0), \dot\gamma^*_{p_2p_1}(0)) + g_{p_2}(\dot\gamma_{p_2p_3}(0),, \dot\gamma^*_{p_2p_4}(0)).$$

Recover the **Euclidean parallelogram identity** for $F_{\mathrm{Euc}}(\theta) = \frac{1}{2}\theta^\top\theta = \frac{1}{2}\|\theta\|_2^2$

$$\begin{aligned} \frac{1}{2}\|\theta_1\|^2 + \frac{1}{2}\|\theta_2\|^2 &= \left\|\frac{\theta_1-\theta_2}{2}\right\|^2 + \left\|\frac{\theta_1+\theta_2}{2}\right\|^2 \\ 2\|\theta_1\|^2 + 2\|\theta_2\|^2 &= \|\theta_1-\theta_2\|^2 + \|\theta_1+\theta_2\|^2 \end{aligned}$$

$$B_{F_{\mathrm{Euc}}}(\theta:\theta') = \tfrac{1}{2}\|\theta-\theta'\|^2$$

https://arxiv.org/abs/1910.03935

k-MLE: Learning mixtures with Classification EM

Bijection between exponential families and dual Bregman divergences:

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

MLE

Maximum likelihood estimate (MLE) for an exponential family = dual Bregman centroid for the sufficient statistics

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

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

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

Bregman centroid

Complete likelihood of the mixture: $\bar{l}(x_1, ..., x_n|w,\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\kappa}\delta_j(z_i)(-B_{F^*}(t(x_i):\eta_j) + F^*(t(x_i)) + k(x_i) + \log w_j)$

$$\begin{aligned}\bar{l}' &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\kappa}\delta_j(z_i)(B_{F^*}(t(x_i):\eta_j) - \log w_j),\\ &= \frac{1}{n}\sum_{i=1}^{n}\min_{j=1}^{k}(B_{F^*}(y_i:\eta_j) - \log w_j),\\ &= \mathrm{kmeans}_{F^*,\log w}(\mathcal{Y}:H),\end{aligned}$$

Classification Expectation-Maximization for a mixture of an exponential family is equivalent to dual Bregman k-means.

k-MLE

k-MLE: A fast algorithm for learning statistical mixture models, IEEE ICASSP 2012 https://arxiv.org/abs/1203.5181

Flat manifolds (M, ∇) : Two equivalent definitions

- A **connection** ∇ is **flat** if both its curvature tensor R^{∇} and its torsion T^{∇} vanish
- A **manifold** M with a flat connection ∇ is said **flat**

If M admits a flat connection ∇ Then there exist *local coordinate systems* such that $\nabla_{\partial_i}\partial_i = 0$ and the changes between local coordinate systems are affine transformations

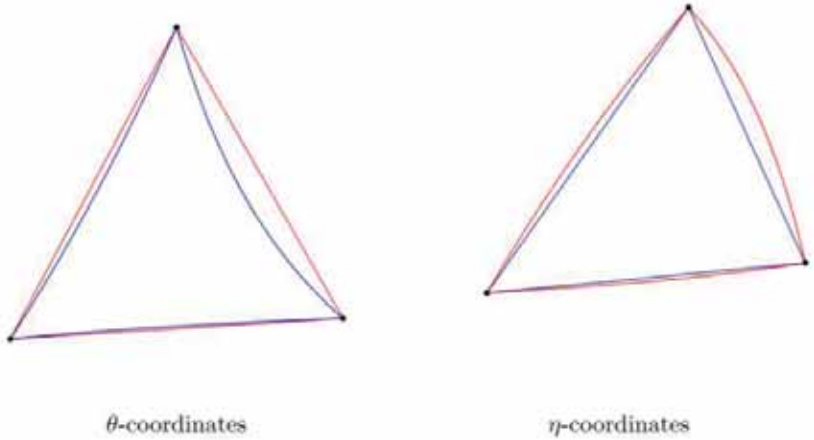
$\Longleftrightarrow$

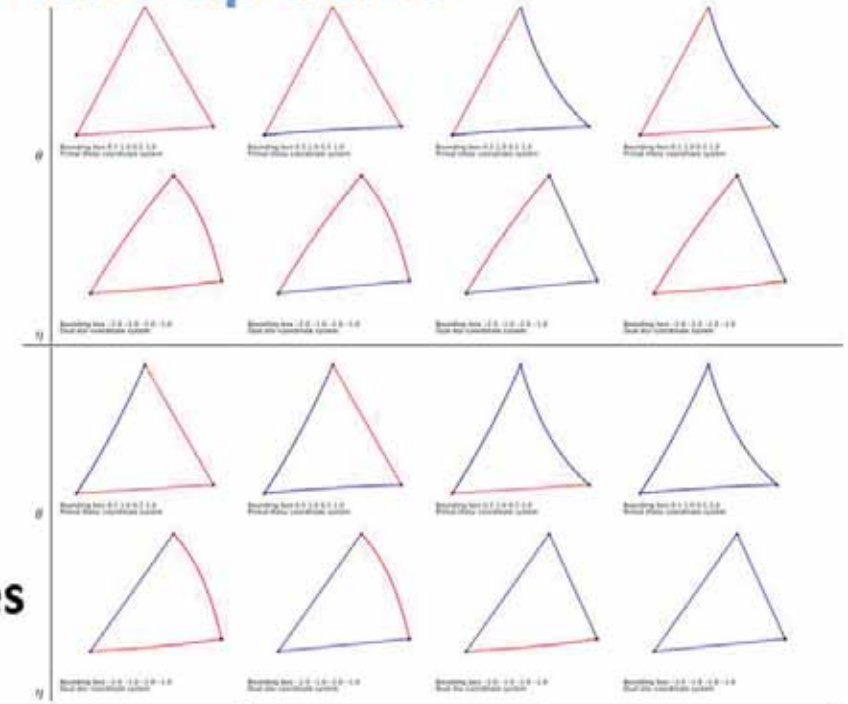
If M admits *local coordinate systems* such that changes of coordinate systems are affine transformations then there exists a flat connection ∇ such that $\nabla_{\partial_i}\partial_i = 0$

- Coordinate system satisfying $\nabla_{\partial_i}\partial_i = 0$ is called **∇-affine coordinate system**
- **Bregman manifolds M** are *dually flat manifolds*: (M, ∇) and (M, ∇^*) both flat

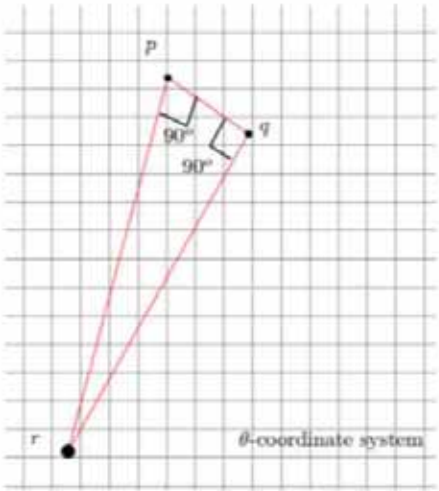
On geodesic triangles with right angles in a dually flat space

Bregman manifolds: Dually flat spaces





3 vertices define 6 geodesic edges from which
8 geodesic triangles can be built, defining 18 interior angles

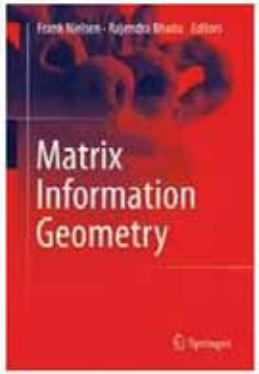


Geodesic triangle
with two right angles

Triple of points where the
dual Pythagorean theorems hold

https://arxiv.org/abs/1910.03935

Matrix Bregman divergences



For real symmetric matrices:

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

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

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

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

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

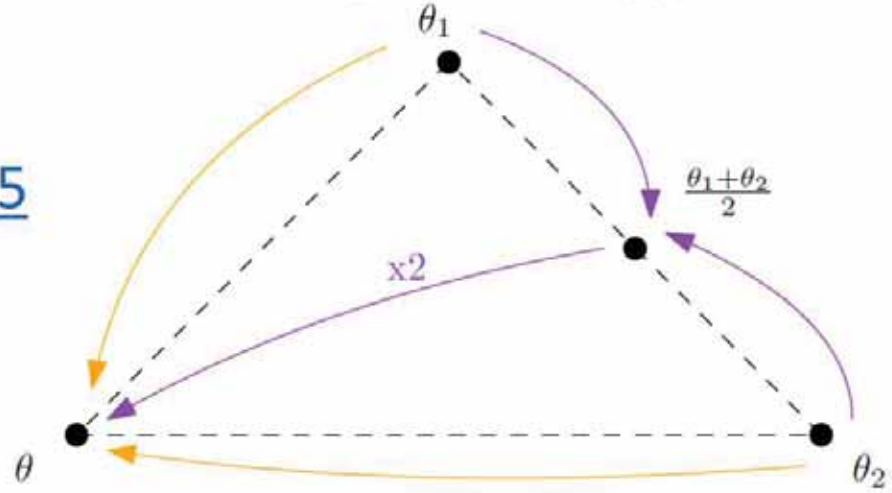
$$D_{\mathrm{ld}}(X:Y) = \mathrm{tr}\left(XY^{-1}\right) - \log\det\left(XY^{-1}\right) - n$$

Bregman–Schatten p-divergences…

@FrnkNIsn **Mining Matrix Data with Bregman Matrix Divergences for Portfolio Selection, 2013**

Bregman divergence: Parallelogram-type identity

https://arxiv.org/abs/1910.03935



$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta)=B_F\left(\theta_1:\tfrac{\theta_1+\theta_2}{2}\right)+B_F\left(\theta_2:\tfrac{\theta_1+\theta_2}{2}\right)+2B_F\left(\tfrac{\theta_1+\theta_2}{2}:\theta\right)$$

Jensen-Bregman Divergence JB

$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = B_F\left(\theta_1:\frac{\theta_1+\theta_2}{2}\right)+B_F\left(\theta_2:\frac{\theta_1+\theta_2}{2}\right)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right)$$

$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = 2\mathrm{JB}_F(\theta_1,\theta_2)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right),$$

Jensen divergence J

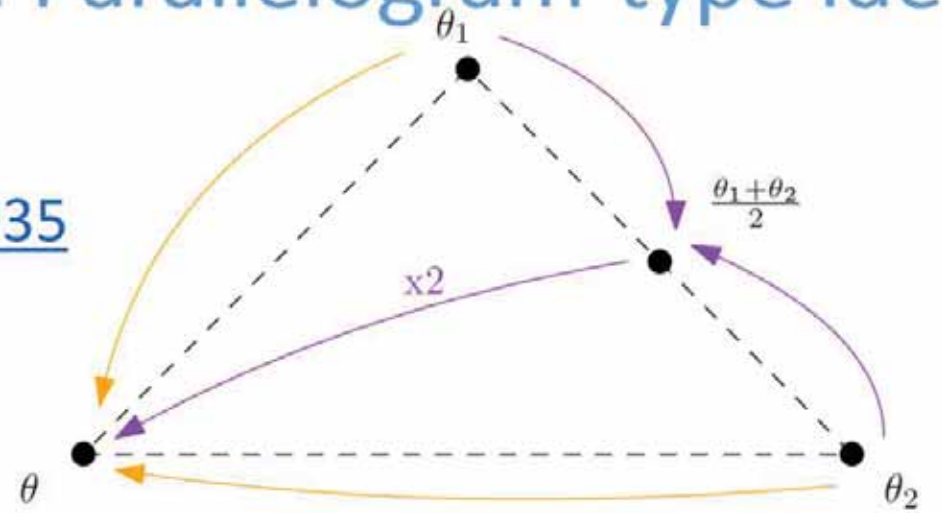
$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = 2J_F(\theta_1,\theta_2)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right),$$

$F_{\mathrm{Euc}}(\theta)=\frac{1}{2}\theta^\top\theta=\frac{1}{2}\|\theta\|_2^2$, then we have $B_{F_{\mathrm{Euc}}}(\theta:\theta')=\frac{1}{2}\|\theta-\theta'\|^2$ and $\theta=0$

recover the parallelogram identity: $2\|\theta_1\|^2+2\|\theta_2\|^2 = \|\theta_1-\theta_2\|^2+\|\theta_1+\theta_2\|^2,$

Bregman divergence: Parallelogram-type identity

https://arxiv.org/abs/1910.03935



$$\boxed{B_F(\theta_1:\theta)+B_F(\theta_2:\theta)=B_F\left(\theta_1:\tfrac{\theta_1+\theta_2}{2}\right)+B_F\left(\theta_2:\tfrac{\theta_1+\theta_2}{2}\right)+2B_F\left(\tfrac{\theta_1+\theta_2}{2}:\theta\right)}$$

Jensen-Bregman Divergence (JB)

$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = B_F\left(\theta_1:\frac{\theta_1+\theta_2}{2}\right)+B_F\left(\theta_2:\frac{\theta_1+\theta_2}{2}\right)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right)$$

$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = 2\mathrm{JB}_F(\theta_1,\theta_2)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right),$$

Jensen divergence (J)

$$B_F(\theta_1:\theta)+B_F(\theta_2:\theta) = 2J_F(\theta_1,\theta_2)+2B_F\left(\frac{\theta_1+\theta_2}{2}:\theta\right),$$

$F_{\mathrm{Euc}}(\theta)=\frac{1}{2}\theta^\top\theta=\frac{1}{2}\|\theta\|_2^2$, then we have $B_{F_{\mathrm{Euc}}}(\theta:\theta')=\frac{1}{2}\|\theta-\theta'\|^2$ and $\theta=0$

Recover the Euclidean parallelogram identity: $2\|\theta_1\|^2+2\|\theta_2\|^2 = \|\theta_1-\theta_2\|^2+\|\theta_1+\theta_2\|^2,$

Radical hyperplane of Bregman spheres

- Consider two left-sided Bregman balls: $\sigma_1 = \{\theta \in \Theta \ : \ B_F(\theta_1 : \theta) = r_1\}$

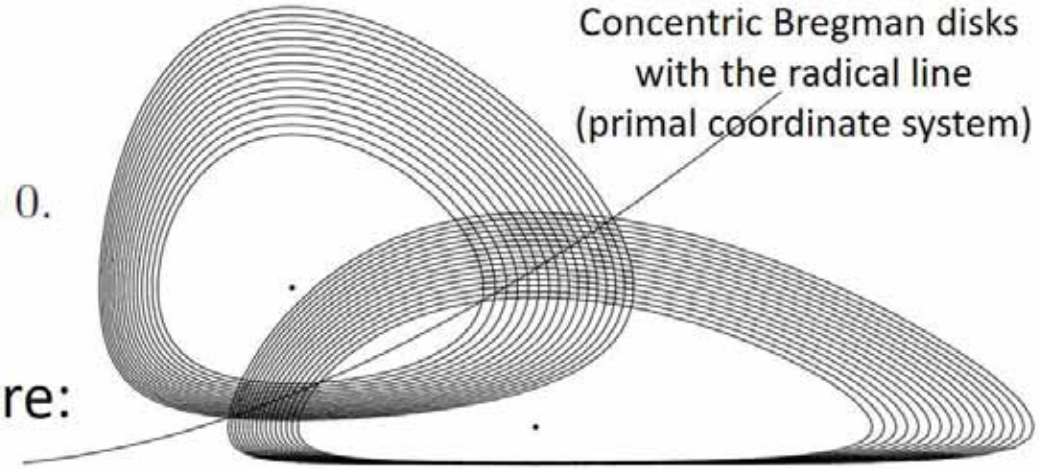
  $\sigma_2 = \{\theta \in \Theta \ : \ B_F(\theta_2 : \theta) = r_2\}$

- Define the **power to a Bregman ball** as: $\mathrm{Pow}_{\sigma_i}(\theta) := B_F(\theta_i : \theta) - r_i$
- The **radical hyperplane** is defined by:

$$H_{12} \ : \ \mathrm{Pow}_{\sigma_1}(\theta) = \mathrm{Pow}_{\sigma_2}(\theta),$$
$$H_{12} \ : \ \langle \eta, \theta_2 - \theta_1 \rangle + F(\theta_1) - F(\theta_2) - r_1 + r_2 = 0.$$

- It is a **dual (d-1)-flat** which supports the intersection of the (d-2)-dimensional sphere:

$$\boxed{\sigma_{12} = \sigma_1 \cap \sigma_2 \in H_{12}}$$



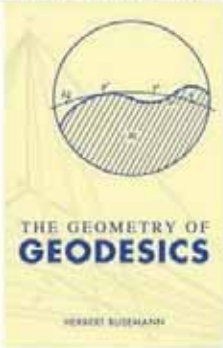
Tailored Bregman ball trees for effective nearest neighbors, EWCG 2009.
Bregman vantage point trees for efficient nearest neighbor queries, IEEE ICME 2009.
Bregman Voronoi diagrams, Discrete & Computational Geometry 44.2 (2010): 281-307.

Busemann's G-spaces

A **metric homogeneous space** (X,d) which **further satisfies**:

Herbert Busemann
(1905-1994)

- $\forall x \neq z, \exists y \neq x, z$ such that $d(x,y) + d(y,z) = d(x,z)$ (**Menger convexity**)
- $\forall p, \exists$ radius $\rho_p > 0$ such that $\forall x \neq y \in \mathrm{Ball}^\circ(p, \rho_p)$
  $\exists z \neq x, y$ such that $d(x,y) + d(y,z) = d(x,z)$ (**local extendability**)
- For all $x \neq y$, if $d(x,y) + d(y,z_1) = d(x,z_1)$ and $d(x,y) + d(y,z_2) = d(x,z_2)$ and $d(y,z_1) = d(y,z_2)$ then $z_1 = z_2$ (**uniqueness of prolongation**)
- A bounded infinite set has at least one accumulation point (**finite compactness**)



<u>Busemann Conjecture</u>: Every G-space is a topological manifold

Thurston, 4-dimensional Busemann G-space are 4-manifolds, Differential Geometry and its Applications 6.3 (1996).

Some paintings by H. Busemann:

Calculating statistical distances, (cross)-entropies, etc.

- **Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family**. arxiv:2003.02469 → Use the fact that Bregman generator is defined up to a modulo term to choose $F(\theta)=-\log p(x;\theta)$
- **A note on Onicescu's informational energy and correlation coefficient in exponential families**. arXiv:2003.13199
→ Get closed-form expressions for exponential families with a natural parameter cone space
- **On power chi expansions of f-divergences**. arXiv:1903.05818
  On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 2013
→ f-divergence using higher order chi square divergences which have closed-form expressions for affine natural parameter space
- **On the Kullback-Leibler divergence between location-scale densities**. arXiv:1904.10428
→ use the group structure of location-scale family paramaters. Highlight some geometric properties
- **A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions**, arXiv:1905.10965
→ The KLD between Cauchy distributions is remarkably symmetric!
- **MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions**. IEEE SPL 2017
→ Simple trick to bound the entropy of mixtures by a series of bounds using the entropy of maxent distributions
- **Patch Matching with Polynomial Exponential Families and Projective Divergences**. SISAP 2016: 109-116
→ γ-divergence are projective divergences (=invariant to scale) which can be applied to unnormalized densities, and tend to KLD!
- **Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series**. IEEE SSP 2016: 1-5
→ Which distance between dependent random-variables: Optimal transport compared with information geometry
- **An Information-Geometric Characterization of Chernoff Information**. IEEE SPL, 2013
→ Chernoff distance between exponential families are Bregman divergences. Can be calculated exactly for uni-order families
- **On Rényi and Tsallis entropies and divergences for exponential families**. arXiv: 1105.3259
- **A closed-form expression for the Sharma-Mittal entropy of exponential families**. Journal of Physics A: Math. and Theo., 2011
- **Entropies and cross-entropies of exponential families**. IEEE ICIP 2010 → with examples (Poisson, Rayleigh, etc.)
- **Earth Mover Distance on superpixels**. IEEE ICIP 2010 → Optimal transport with topological constraints

Calvo-Oller distance between multivariate normals

- Embed d-dimensional normal N(μ,Σ) into a (d+1)-dimensional SPD matrix S:

$$S : (\mu, \Sigma) \to \mathbb{P}_{d+1}$$
$$(\mu, \Sigma) \to \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$$

- Calculate geodesic distance between two a **(d+1)-dimensional** SPD matrices:

$$D_{\mathrm{SPD}}(S_1, S_2) = \|\mathrm{Log}(S_1^{-1} S_2)\|_F = \sqrt{\left(\sum_{i=1}^{d} (\ln \lambda_i (S_1^{-1} S_2))^2\right)}$$

- Since S yields a **non-geodesic submanifold** embedding, we have

$$D_{\mathrm{Fisher-Rao}}(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) \geq D_{\mathrm{SPD}}(S(\mu_1, \Sigma_1), S(\mu_2, \Sigma_2))$$

$$D_{\mathrm{Calvo-Oller}}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = D_{\mathrm{SPD}}(S(\mu_1, \Sigma_1), S(\mu_2, \Sigma_2))$$

$$\boxed{D_{Calvo-Oller}(\lambda_1, \lambda_2) = \frac{1}{\sqrt{2}} \left\| \mathrm{Log} \begin{bmatrix} \Sigma_1^{-1/2} (\Sigma_2 + \Delta\mu\Delta\mu') \Sigma_1^{-1/2} & \Sigma_1^{-1/2} \Delta\mu \\ \Delta\mu' \Sigma_1^{-1/2} & 1 \end{bmatrix} \right\| \qquad \Delta\mu = \mu_2 - \mu_1}$$

Calvo and Oller, A distance between multivariate normal distributions based in an embedding into the Siegel group, JMA 1990
On approximating the Riemannian 1-center, CG 2013

Riemannian distance on a Cartan-Hadamard manifold

- **Definition**: complete and simply connected Riemannian manifold with non-positive sectional curvatures
- **Examples**: Hyperbolic space, Positive-definite matrix cone, Siegel upper space
- **Properties**:
  - diffeomorphic to the Euclidean space (Cartan–Hadamard theorem)
  - unique geodesic connecting a pair of points (Hopf–Rinow theorem)
- **Riemannian distance** expressed using the Riemannian logarithmic map:

$$\rho(p, q) = \| \log_p q \|_p$$

- **Geodesics** expressed using the Riemannian exponential & logarithmic maps:

$$\gamma_{pq}(t) = \exp_p(t \log_p(q)), \quad t \in [0, 1]$$

"On approximating the Riemannian 1-center." *Computational Geometry* 46.1 (2013): 93-104.

Cauchy-Binet formula generalizes Pythagoras' theorem

- Consider A and B two rectangular matrices of dimension (m,n) with $n \geq m$
- **Cauchy-Binet formula** for the determinant of (n,n) square matrix $A^T B$:

$$\boxed{|A^\top B| = \sum_{|I|=n} |A_I^\top B_I|}$$

(more columns n than rows m)

A_I is the matrix with rows of A indexed by the subset I of $2^{[m]}$ $[m] = \{1, \ldots, m\}$

- When A=B with $A = [v_1, \ldots, v_m]^\top$ with rows $v_i \in \mathbb{R}^n$, we have

$$|A^\top A| = \mathrm{vol}^2(\mathrm{parallelotope}(v_1, \ldots, v_m))$$

$$|A^\top A| = \sum_{|I|=n} |A_I^\top A_I|$$

Squared volume of m-dimensional parallelotope

- In particular when **n=1**, we recover **Pythagoras' theorem** for $A = [v]$

$$v = (v_{(1)}, \ldots, v_{(n)}) \in \mathbb{R}^n \qquad \mathrm{vol}^2(v) = |v^\top v| = \|v\|^2 = \sum_{i=1}^{d} v_{(i)}^2$$

f-divergences between Cauchy densities

- The Cauchy family is a location-scale family with standard density $p(x) := \frac{1}{\pi(1+x^2)}$

$$\mathcal{C} = \left\{ p_{l,s}(x) := \frac{1}{s} p\left(\frac{x-l}{s}\right) \ : \ (l,s) \in \mathbb{R} \times \mathbb{R}_{++} \right\}$$

- Csiszar' s f-divergences (for a convex function f with f(1)=0): $I_f(p:q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x$

Theorem: All f-divergences between Cauchy densities are <u>symmetric</u> and can be expressed as a <u>function of the chi-squared divergence</u>

f-divergence name	f	$h_f(u)$ for $I_f[p_{\lambda_1} : p_{\lambda_2}] = h_f(\chi[p_{\lambda_1} : p_{\lambda_2}])$
Kullback-Leibler divergence	$-\log u$	$\log(1+\frac{1}{2}u)$
LeCam divergence	$\frac{(u-1)^2}{1+u}$	$2-4\sqrt{\frac{1}{2(u+2)}}$
squared Hellinger	$\frac{1}{2}(\sqrt{u}-1)^2$	$1-\frac{2K\left(1-\left(1+u+\sqrt{u(2+u)}\right)^{-2}\right)}{\pi\sqrt{1+u+\sqrt{u(2+u)}}}$

$$\chi[p_\lambda : p_{\lambda'}] = \frac{\|\lambda-\lambda'\|^2}{2\lambda_2\lambda_2'}$$

On f-divergences between Cauchy distributions, 2021

Jensen-Shannon divergence between Cauchy distributions

- Cauchy densities form a location-scale family with **even standard density** p(x)=p(-x):

$$p(x) := \frac{1}{\pi(1+x^2)} \qquad \mathcal{C} = \left\{ p_{l,s}(x) := \frac{1}{s} p\left(\frac{x-l}{s}\right) \ : \ (l,s) \in \mathbb{R} \times \mathbb{R}_{++} \right\}$$

- **Jensen-Shannon divergence** is a symmetrization of the Kullback-Leibler divergence bounded by log 2:

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

- The JSD between two Cauchy densities can be expressed using the **chi square divergence** :

$$\boxed{\begin{aligned} D_{\mathrm{JS}}(p_{l_1,s_1} : p_{l_2,s_2}) &= \log\left(\frac{2\sqrt{(l_1-l_2)^2+(s_1+s_2)^2}}{\sqrt{(l_1-l_2)^2+(s_1+s_2)^2}+2\sqrt{s_1 s_2}}\right) \\ &=: h_{\mathrm{JS}}(\chi(p_{l_1,s_1}, p_{l_2,s_2})), \end{aligned}}$$

$$\begin{aligned} \chi(p_{l_1,s_1}, p_{l_2,s_2}) &= \int \frac{(p_{l_1,s_1}(x) - p_{l_2,s_2}(x))^2}{p_{l_2,s_2}(x)} \mathrm{d}x \\ &= \frac{(l_2-l_1)^2+(s_2-s_1)^2)}{2 s_1 s_2} \end{aligned}$$

$$\Rightarrow \quad h_{\mathrm{JS}}(u) = \log\left(\frac{2\sqrt{2+u}}{\sqrt{2+u}+\sqrt{2}}\right)$$

On f-divergences between Cauchy distributions, arXiv 2101.12459

Cauchy-Schwarz inequality & divergence

Cauchy-Schwarz inequality: $|\langle u, v\rangle| \leq \|u\|\,\|v\|$ where $\|v\| := \sqrt{\langle v, v\rangle}$

A simple proof mnemonic:

$$\langle u, v\rangle = \|u\|\,\|v\|\,\underbrace{\cos(\alpha_{uv})}_{\in[-1,1]}$$

$$|\langle u, v\rangle| = \|u\|\,\|v\|\,|\cos(\alpha_{uv})| \leq \|u\|\,\|v\|$$

The **Cauchy-Schwarz divergence**:

- closed-form for Gaussian mixtures
  Wishart mixtures, etc.
- Do not need normalized densities
- Extends to **Hölder divergences**

$$\boxed{D_{\mathrm{CS}}[p:q] := -\log \frac{\langle p, q\rangle}{\|p\|\,\|q\|} \geq 0}$$

On Hölder projective divergences, Entropy 19.3 (2017)
Closed-form information-theoretic divergences for statistical mixtures, IEEE ICPR 2012

Cauchy-Schwarz divergence in exponential families

Cauchy-Schwarz divergence

$$D_{\mathrm{CS}}(p,q) := -\log\left(\frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}\mu(x)}{\sqrt{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}\mu(x)\right)\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}\mu(x)\right)}}\right) \geq 0.$$

Exponential family $\mathcal{E} = \left\{ p_\theta(x) = \exp\left(\theta^\top t(x) - F(\theta) + k(x)\right) \ : \ \theta \in \Theta \right\}$

Closed-form expression

$$\boxed{D_{\mathrm{CS}}(p_{\theta_1}, p_{\theta_2}) = J_F(2\theta_1 : 2\theta_2) + \log\left(\frac{\sqrt{E_{p_{2\theta_1}}\left[\exp(k(x))\right] E_{p_{2\theta_2}}\left[\exp(k(x))\right]}}{E_{p_{\theta_1+\theta_2}}\left[\exp(k(x))\right]}\right)}$$

where J is a Jensen divergence: $J_F(\theta_1,\theta_2) := \frac{F(\theta_1)+F(\theta_2)}{2} - F\left(\frac{\theta_1+\theta_2}{2}\right) \geq 0$

For multivariate Gaussians (conic natural parameter space)

$$D_{\mathrm{CS}}(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) = \frac{1}{2}\log\left(\frac{1}{2^d}\frac{\sqrt{|\Sigma_1|\ |\Sigma_2||}}{|(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}|}\right)$$
$$+\frac{1}{2}\mu_1^\top\Sigma_1^{-1}\mu_1 + \frac{1}{2}\mu_2^\top\Sigma_2^{-1}\mu_2$$
$$-\frac{1}{2}(\Sigma_1^{-1}\mu_1+\Sigma_2^{-1}\mu_2)^\top(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1+\Sigma_2^{-1}\mu_2).$$

A note on Onicescu's informational energy and correlation coefficient in exponential families

Cauchy-Schwarz divergence in exponential families

Cauchy-Schwarz divergence $$D_{\mathrm{CS}}(p,q) := -\log\left(\frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}\mu(x)}{\sqrt{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}\mu(x)\right)\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}\mu(x)\right)}}\right) \geq 0.$$

Exponential family $$\mathcal{E} = \left\{p_\theta(x) = \exp\left(\theta^\top t(x) - F(\theta) + k(x)\right) \ : \ \theta \in \Theta\right\}$$

When the natural parameter space is a cone (eg., Gaussian, Wishart, etc.):

$$\boxed{\begin{aligned} D_{\mathrm{CS}}(p_{\theta_1}, p_{\theta_2}) &= -\log\left(\frac{\sqrt{p_{2\theta_1}(\omega)}\sqrt{p_{2\theta_2}(\omega)}}{p_{\theta_1+\theta_2}(\omega)}\right), \\ &= l_{\theta_1+\theta_2}(\omega) - \frac{l_{2\theta_1}(\omega) + l_{2\theta_2}(\omega)}{2}, \quad \forall \omega \in \mathcal{X}. \end{aligned}}$$ (Log-likelihood)

A note on Onicescu's informational energy and correlation coefficient in exponential families, arXiv 2003.13199

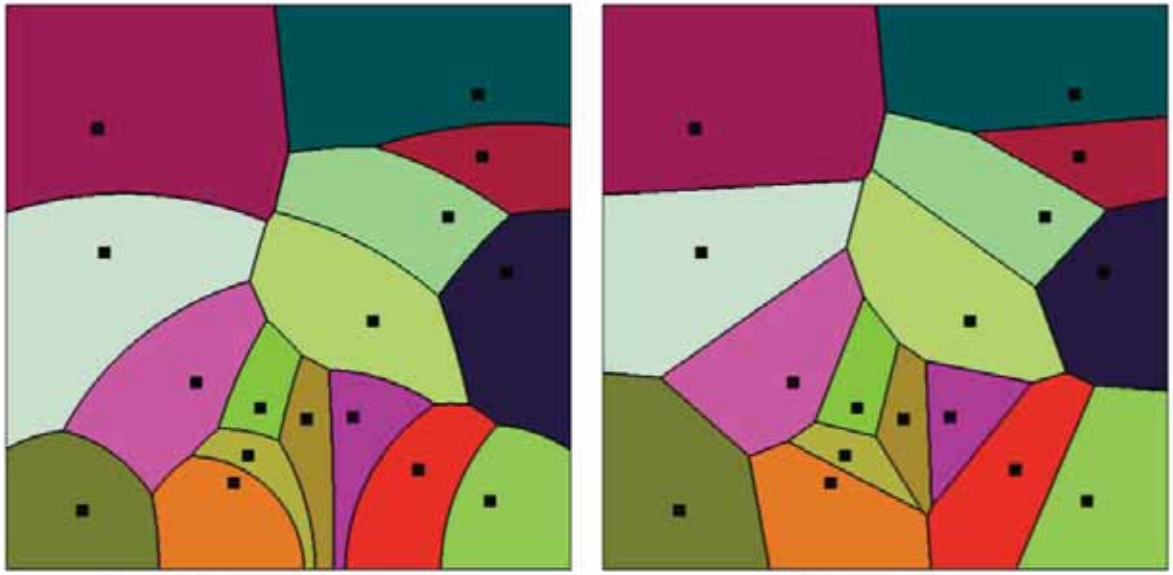
Voronoi diagrams and Delaunay complex on the Cauchy manifolds

Dual Voronoi cells wrt dissimilarity D(.:.):

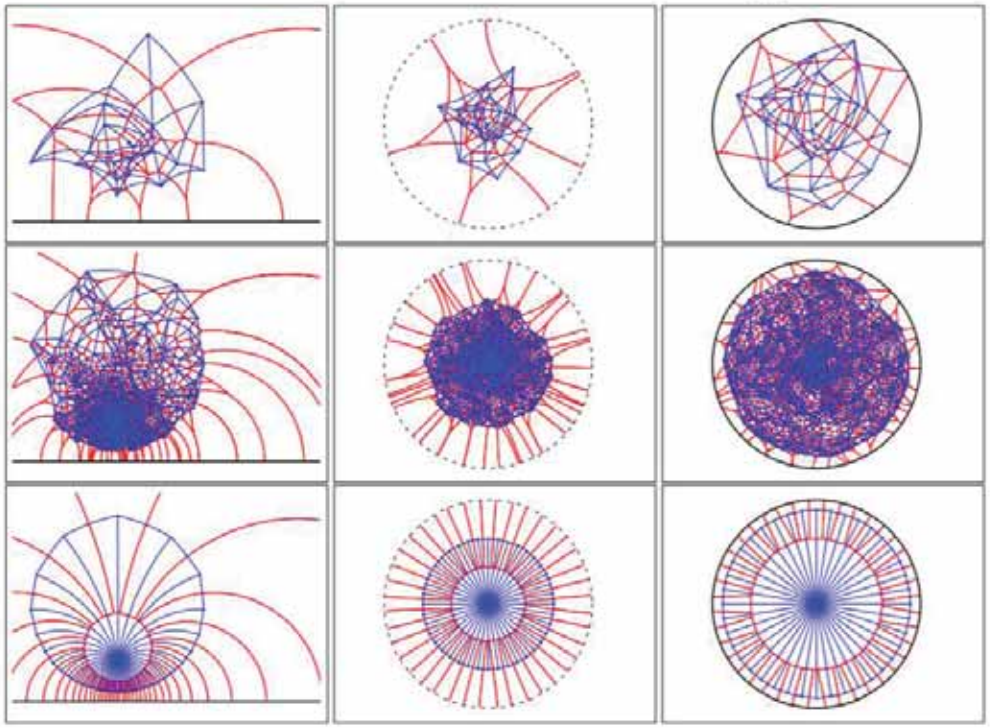
$$\mathrm{Vor}_D(P_i) \;:=\; \{X \in \mathbb{X},\quad D(P_i : X) \le D(P_j : X),\quad \forall j \in \{1,\ldots,n\}\}$$

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

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.

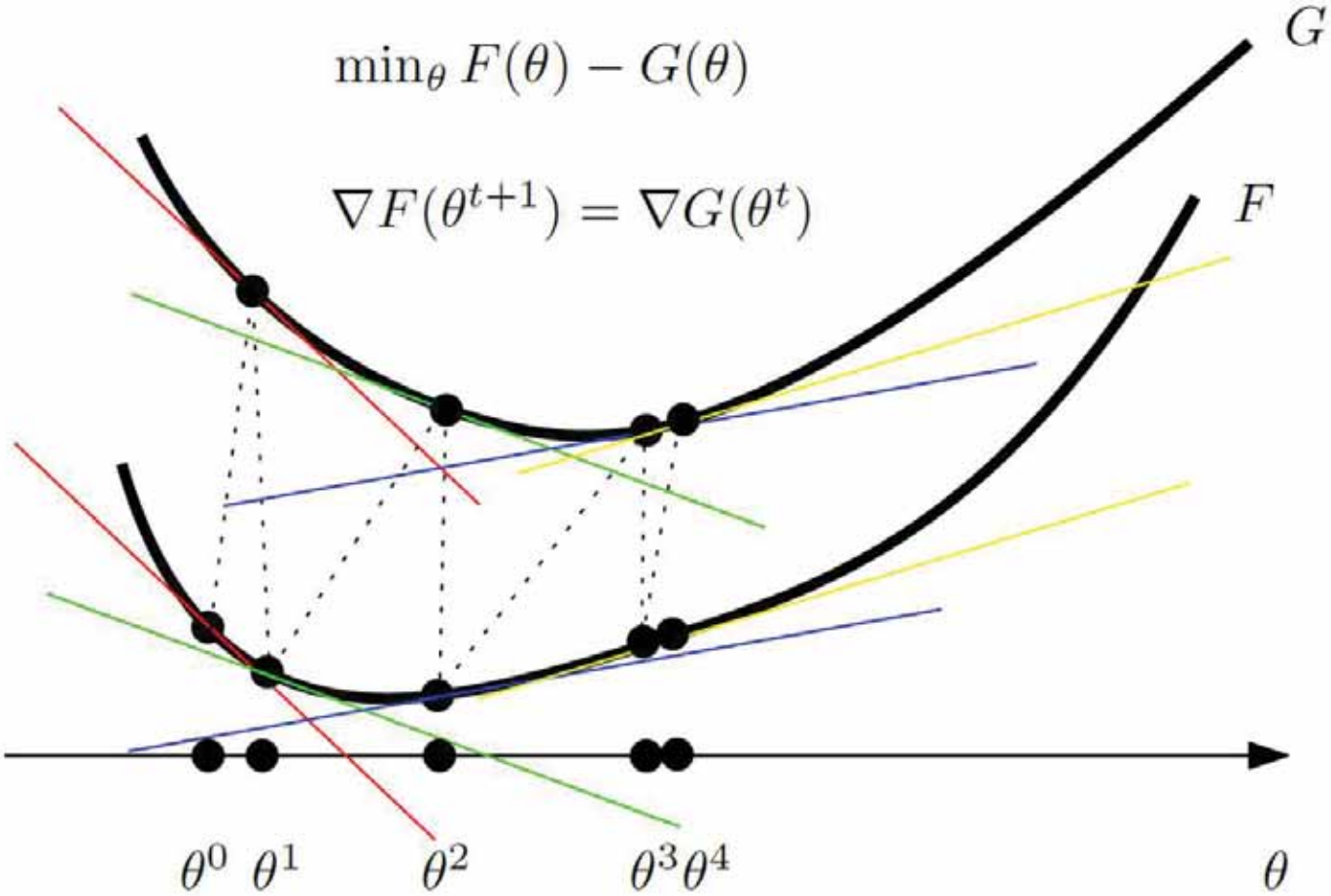


$\mathrm{Vor}_{\rho_{\mathrm{FR}}} = \mathrm{Vor}_{\rho_{\mathrm{KL}}} = \mathrm{Vor}_{\rho^2_{\chi}} = \mathrm{Vor}_{D_{\mathrm{flat}}}$ $\qquad$ $\mathrm{Vor}_{D^*_{\mathrm{flat}}} = \mathrm{Vor}_{\rho_E}$.



On Voronoi Diagrams on the Information-Geometric Cauchy Manifolds, *Entropy* 22, no. 7: 713. arxiv 2006.07020

ConCave-Convex Procedure (CCCP)



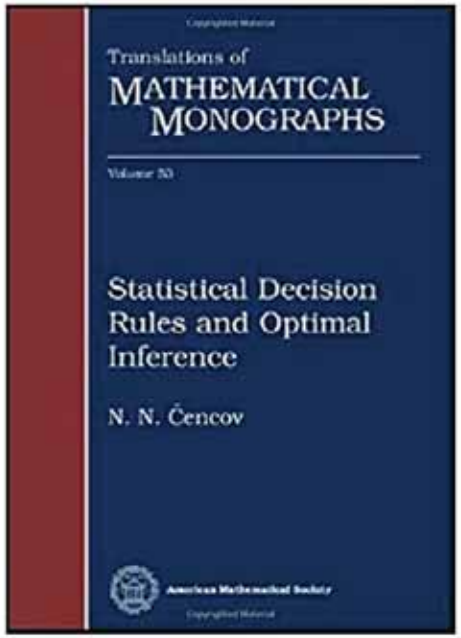
- Write any energy/loss with lower bounded Hessian as the sum of a **convex function F** plus a **concave function -G**
- Optimization to a local minimum by matching points of the graph plots which have the same tangent hyperplane

  (no learning rate!)

$$\nabla F(\theta^{t+1}) = \nabla G(\theta^t)$$

Yuille, Alan L., and Anand Rangarajan. "The concave-convex procedure (CCCP)." , NeurIPS 2002.
The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory* 57.8 (2011)

Geometrostatistics (Information Geometry)



Nikolai N. Chentsov (1930-1992)

INTRODUCTION 1
1. Statistical problems and statistical decisions 1
2. Probability measures and conditional distributions 13
3. Smooth manifolds and their mappings 37
4. Category theory and geometry 50

CHAPTER I. THE FORMAL DECISION PROBLEM 65
5. The category of statistical decisions 65
6. Markov geometry of families of probability distributions 76
7. Geometry of dominated families of probability distributions 92
8. Invariant information characteristics 113

CHAPTER II. EQUIVARIANT DIFFERENTIAL GEOMETRY OF A COLLECTION OF PROBABILITY DISTRIBUTIONS 127
9. Geometry of the simplex of probability distributions on a finite algebra 127
10. Projective geometry of a collection of probability distributions 140
11. Invariant Riemannian metric on a manifold of probability distributions 156
12. Geodesic mean of probability distributions 165

CHAPTER III. SMOOTH FAMILIES OF PROBABILITY DISTRIBUTIONS AND THE INFORMATION INEQUALITY 185
13. Finite-dimensional approximation of infinite-dimensional manifolds of probability distributions 185
14. Differentiable families of probability distributions 199
15. The information inequality 213
16. Continuously differentiable families of probability distributions 228

CHAPTER IV. GEOMETRY OF EXPONENT FAMILIES OF PROBABILITY DISTRIBUTIONS 245
17. Convex functions and the Legendre transformation 245
18. Exponent families of probability distributions 262
19. Natural parametrization of exponent families 279
20. Conjugate parametrizations of an exponent family of probability distributions 290
21. Distributions of values of the directional statistic of an exponent family and related families 303
22. Nonsymmetric Pythagorean geometry of the information deviation 319
23. Charts of an exponent family of probability distributions 334

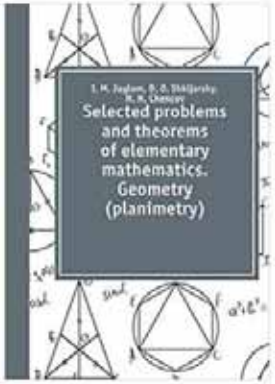
CHAPTER V. OPTIMAL DECISION RULES IN THE EQUIVARIANT PROBLEM OF POINT ESTIMATION 355
24. Estimation of the unknown mean of a multivariate normal distribution 355
25. Estimation of the unknown density of a probability distribution 370
26. Invariant loss functions in problems of mathematical statistics 391
27. Optimal estimators for smooth families of probability distributions 403
28. Quasi-homogeneous families of probability distributions 437

APPENDIX 451
29. Random probability measures 451

NOTES AND COMMENTS 465

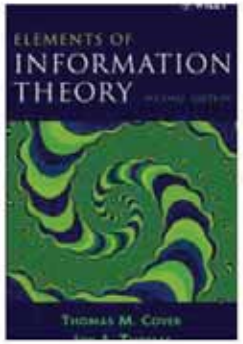
BIBLIOGRAPHY 477

INDEX 495

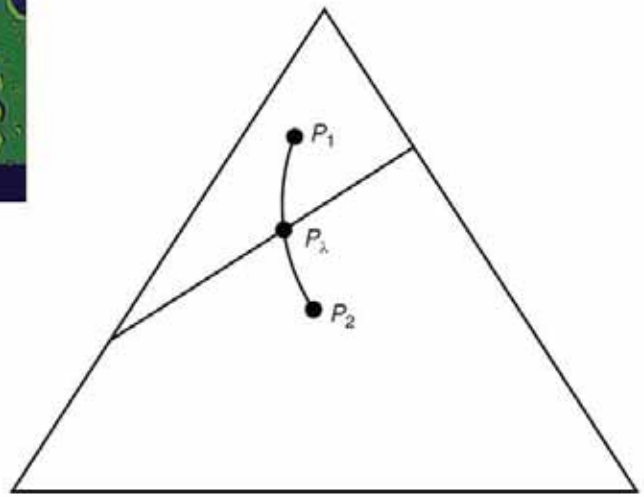


Chernoff information: A geometric characterization

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



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



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

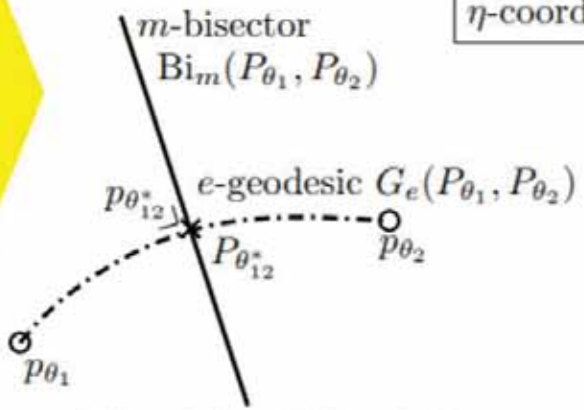
Probability simplex

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

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

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

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



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

Exponential family manifold

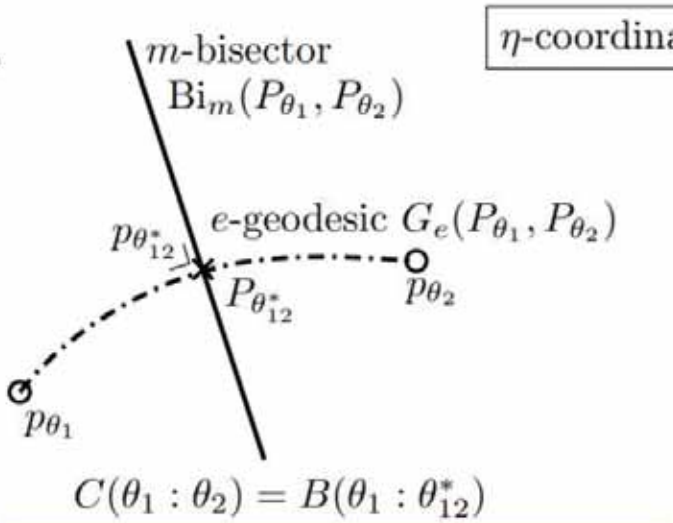
Bayesian hypothesis testing:
A geometric characterization of the best error exponent

Dually flat Exponential Family Manifold (EFM):
Chernoff information amounts to a Bregman divergence

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

$$C(P_1, P_2) = -\log \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}\nu(x) \geq 0,$$

Chernoff Information



This geometric characterization yields to an exact closed-form solution in 1D EFs, and a simple geodesic bisection search for arbitrary dimension

@FrnkNlsn

An Information-Geometric Characterization of Chernoff Information, IEEE SPL, 2013 (arXiv:1102.2684)

Chernoff information between two Cauchy distributions

Cauchy location-scale family: $p_{l,s}(x) := \frac{1}{\pi s\left(1+\left(\frac{x-l}{s}\right)^2\right)} = \frac{s}{\pi(s^2+(x-l)^2)}$

Chernoff information: $D_{\text{Chernoff}}((p_1 : p_2) := -\log \min_{a\in(0,1)} \int p_1(x)^a p_2(x)^{1-a}\, \mathrm{d}x$

$$D_{\text{Chernoff}}(p_{\theta_1}, p_{\theta_2}) = D_{\text{Bhattacharyya}}(p_{\theta_1}, p_{\theta_2}) = -\log\left(\frac{2K\left(1-\left(1+\chi(\theta_1,\theta_2)+\sqrt{\chi(\theta_1,\theta_2)(2+\chi(\theta_1,\theta_2))}\right)^{-2}\right)}{\pi\sqrt{1+\chi(\theta_1,\theta_2)+\sqrt{\chi(\theta_1,\theta_2)(2+\chi(\theta_1,\theta_2))}}}\right)$$

K: elliptic integral

$$\theta_i = (l_i, s_i) \qquad \chi(\theta_1,\theta_2) = \frac{(l_1-l_2)^2+(s_1-s_2)^2}{2s_1s_2} \qquad K(t) := \int_0^{\pi/2} \frac{1}{\sqrt{1-t\sin^2\theta}}d\theta,\ 0 \le t < 1.$$

$\sqrt{D_{\text{Chernoff}}(p_{\theta_1}, p_{\theta_2})}$ is a metric distance

Nielsen and Okamura, On f-divergences between Cauchy distributions, arXiv:2101.12459

Chernoff minimax discrimination as a generalized Sibson's information radius

- Generalized **information radius** wrt a mean M(x,y) and a divergence D:

(M,D)-information radius: $D^{M,D}_{\mathrm{InformationRadius}}[p:q] = \min_c M(D[p:c], D[q:c])$

- **Jensen-Shannon divergence** is a (A, D_{KL}) -information radius:

$$D_{\mathrm{JS}}[p:q] = D^{A,D_{\mathrm{KL}}}_{\mathrm{InformationRadius}}[p:q] \qquad A(x,y) := \frac{x+y}{2}$$

- **Chernoff minimax discrimination** is a $(\max, D^*_{\mathrm{KL}})$ -information radius (wrt to the reverse KL divergence) :

$$D_{\mathrm{ChernoffMinimax}}[p:q] := \min_c \max\{D_{\mathrm{KL}}[c:p], D_{\mathrm{KL}}[c:q]\}$$

$$D_{\mathrm{ChernoffMinimax}}[p:q] = D^{\max,D^*_{\mathrm{KL}}}_{\mathrm{InformationRadius}}[p:q] \qquad \max(x,y) := \max\{x,y\}$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, Entropy 2021

Transformation rule of Christoffel symbols

- (M,g, ∇): D-dimensional manifold with a metric tensor g and affine connection ∇
- Christoffel symbols are D^3 functions that defines the **∇ -geodesic equation**:

$$\forall i \in \{1, \ldots, D\}, \qquad \ddot{\theta}^i + \sum_{j=1}^{D} \sum_{k=1}^{D} \Gamma^i_{jk}(\theta) \dot{\theta}^j \dot{\theta}^k = 0$$

- Under a smooth invertible transformation: $\eta(\theta) \Leftrightarrow \theta(\eta)$
- The Christoffel symbols' transformation rule show they are **not tensors**:

$$\boxed{\eta = \eta(\theta) \longrightarrow \bar{\Gamma}^a_{bc}(\eta) = \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} \frac{\partial \eta^a}{\partial \theta^i} \frac{\partial \theta^j}{\partial \eta^b} \frac{\partial \theta^k}{\partial \eta^c} \Gamma^i_{jk}(\theta(\eta)) + \sum_{i=1}^{D} \frac{\partial^2 \theta^i}{\partial \eta^b \partial \eta^c} \frac{\partial \eta^a}{\partial \theta^i}}$$

$$\forall a \in \{1, \ldots, D\}, \qquad \ddot{\eta}^a + \sum_{b=1}^{D} \sum_{c=1}^{D} \bar{\Gamma}^a_{bc}(\eta) \dot{\eta}^b \dot{\eta}^c = 0$$

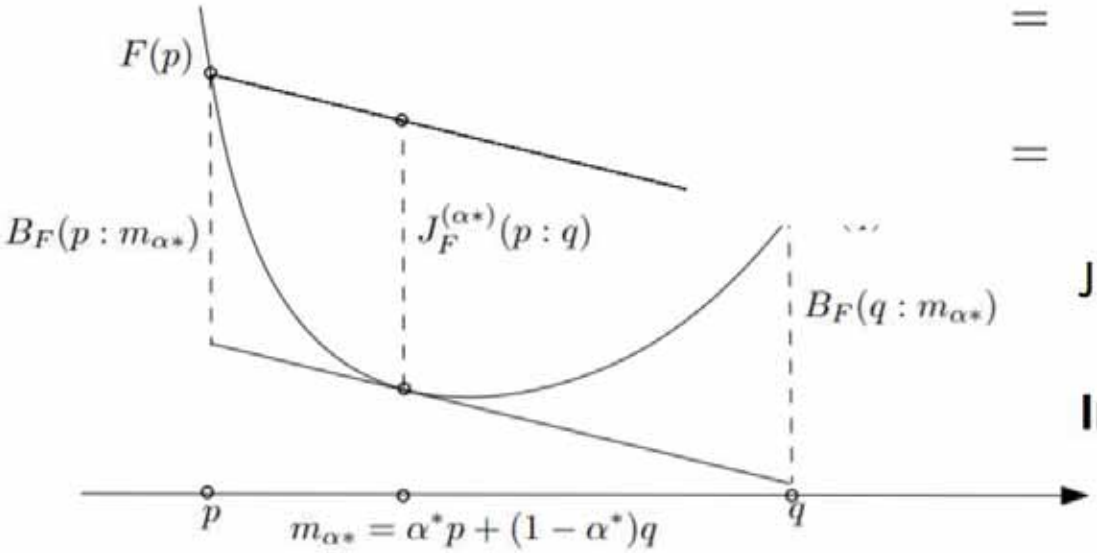
Chernoff information is a Bregman divergence

(when densities belong to the same exponential families)

Herman Chernoff

Chernoff information:

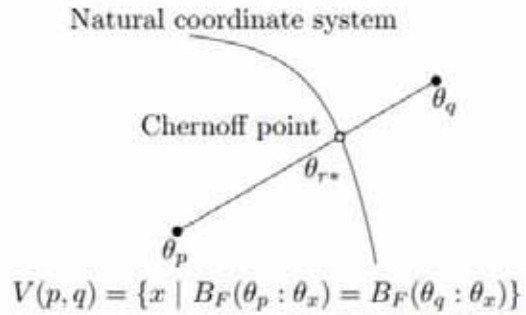
$$\begin{aligned} C^*(p_1 : p_2) &= C_{\alpha^*}(p_1 : p_2) \\ &= -\log \min_{\alpha \in (0,1)} \int p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}x \geq 0. \\ &= \max_{\alpha \in (0,1)} -\log \int p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}x \end{aligned}$$



Jensen divergence: $J_F^{(\alpha)}(p : q) = \int (\alpha F(p(x)) + (1-\alpha) F(q(x)) - F(\alpha p(x) + (1-\alpha) q(x))) \mathrm{d}x \geq 0, \forall \alpha \in (0,1)$

Interpretation on a statistical manifold:

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



Best exponent α* (in Bayesian hypothesis testing):

$$\boxed{J_F^{(\alpha^*)}(p : q) = B_F(p : m_{\alpha^*}) = B_F(q : m_{\alpha^*})}$$

An Information-Geometric Characterization of Chernoff Information, IEEE SPL 2013

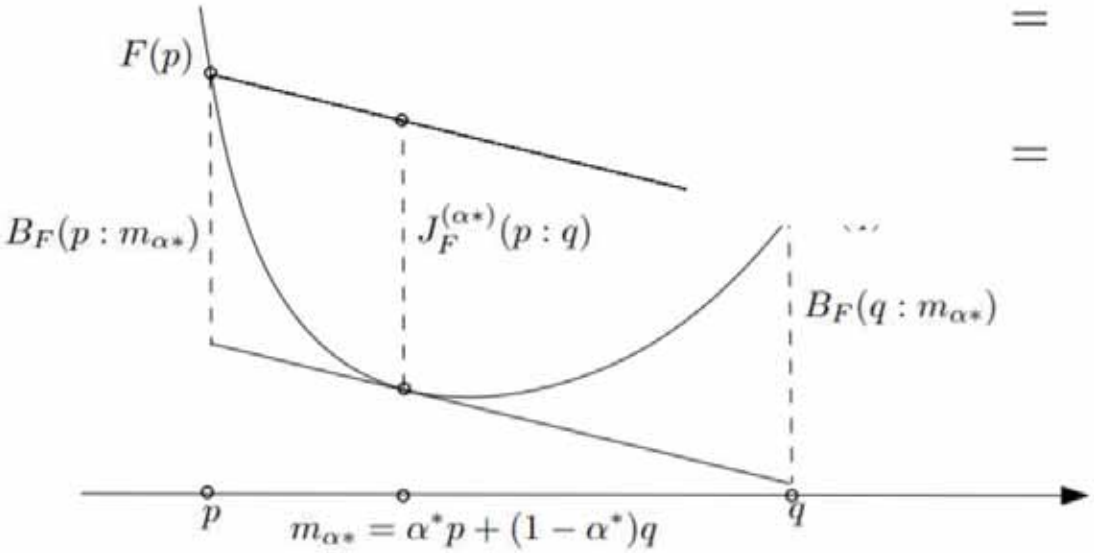
Chernoff information is a Bregman divergence

(when densities belong to the same exponential families)

Herman Chernoff

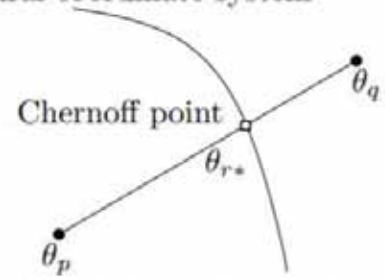
Chernoff information:

$$\begin{aligned} C^*(p_1 : p_2) &= C_{\alpha^*}(p_1 : p_2) \\ &= -\log \min_{\alpha \in (0,1)} \int p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}x \geq 0. \\ &= \max_{\alpha \in (0,1)} -\log \int p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}x \end{aligned}$$



$$\begin{aligned} J_F^{(\alpha)}(p : q) &= \int (\alpha F(p(x)) + (1-\alpha) F(q(x)) - \\ &\quad F(\alpha p(x) + (1-\alpha) q(x))) \, \mathrm{d}x \geq 0, \\ &\forall \alpha \in (0,1) \end{aligned}$$

Natural coordinate system



$V(p,q) = \{x \mid B_F(\theta_p : \theta_x) = B_F(\theta_q : \theta_x)\}$

Best exponent α* (in Bayesian hypothesis testing):

$$\boxed{J_F^{(\alpha^*)}(p : q) = B_F(p : m_{\alpha^*}) = B_F(q : m_{\alpha^*})}$$

An Information-Geometric Characterization of Chernoff Information, IEEE SPL 2013

KL & f-divergences between wrapped Cauchy distributions

Circular Cauchy distributions: $p_w^{\mathrm{cc}}(\phi) := \frac{1}{2\pi}\frac{1-|w|^2}{|e^{i\phi}-w|^2}\,\mathrm{d}z, \quad \phi \in [-\pi,\pi) \qquad z := e^{i\phi} \in \mathbb{C}$

Using the polar transform: $p_{\rho,\phi_0}^{\mathrm{cc}}(\phi) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\phi-\phi_0)}\,\mathrm{d}\phi, \quad \phi \in [-\pi,\pi).$

Wrapped Cauchy density: $p_w^{\mathrm{cc}}(\phi) = p^{\mathrm{wc}}\left(\phi,\eta(w)\right), \quad \eta(w) = \frac{w-i}{w+i}$

$$p_{\mu,\gamma}^{\mathrm{wc}}(\phi) = \sum_{n=-\infty}^{\infty} \frac{\gamma}{\pi\left(\gamma^2+(\phi-\mu+2\pi n)^2\right)}, \quad -\pi \le \phi < \pi$$

$X \sim \mathrm{Cauchy}(\theta)$ $\Rightarrow$ $Y = \frac{1+iX}{1-iX}$ follows a circular Cauchy distributions $\mathrm{CCauchy}\left(\frac{1+i\theta}{1-i\theta}\right)$

Theorem (f-divergences between wrapped Cauchy distributions). *The f-divergence between two wrapped Cauchy distributions amounts to the f-divergence between two corresponding Cauchy distributions: $I_f(p_{\eta_1}^{\mathrm{wc}} : p_{\eta_2}^{\mathrm{wc}}) = I_f(p_{\theta(\eta_1)} : p_{\theta(\eta_2)})$. It follows that the f-divergence between wrapped Cauchy distributions is symmetric and can be expressed as a scalar function of the chi square divergence.*

ClickRemoval (2005): Remove elements in one click

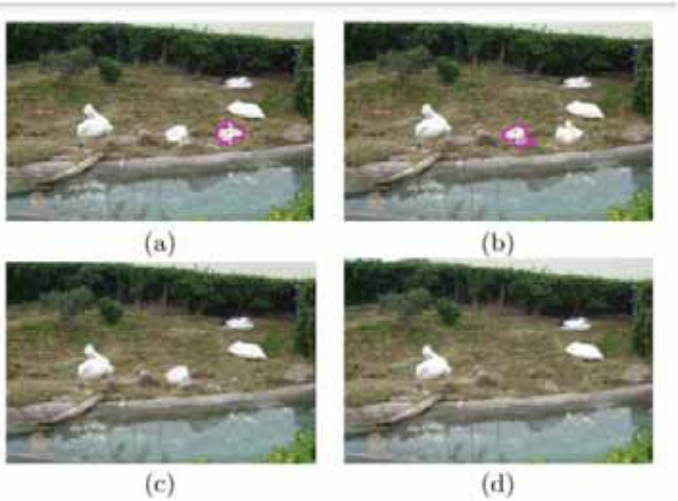


Figure 1: ClickRemoval is an interactive image object removal system. As the user's mouse cursor rolls over the image, the current implicitly selected object is highlighted, (a) and (b). By simple mouse clickings objects are removed and background portions are instantaneously synthesized, (c) and (d).

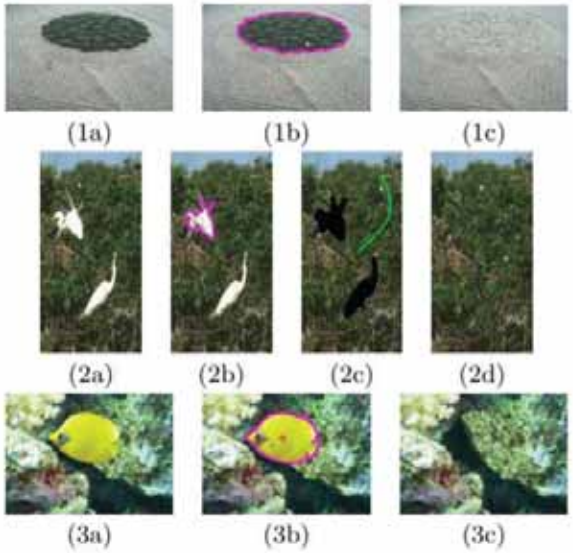


Figure 3: The three different modes of the ClickRemoval interface: fully automatic, manual texture selection, and both user-steered segmentation/texture selection.

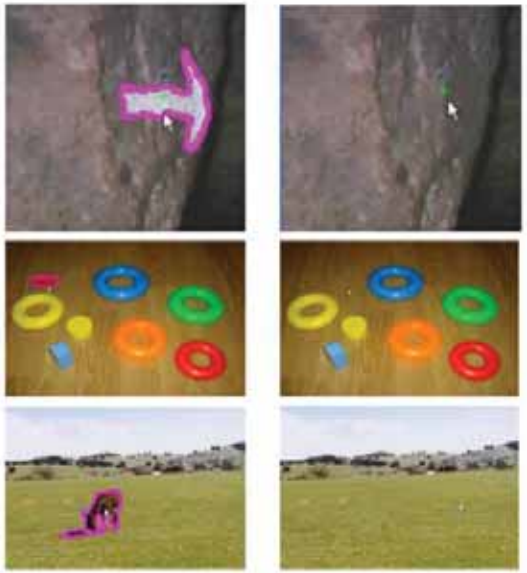


Figure 4: 1-Click removal results (fully automatic segmentation).

ClickRemoval: interactive pinpoint image object removal. ACM Multimedia 2005: 315-318

On Region Merging: The Statistical Soundness of Fast Sorting, with Applications. CVPR (2) 2003: 19-26

Clustering with mixed α-Divergences

$$M_{\lambda,\alpha}(p:x:q) \;=\; \lambda D_\alpha(p:x) + (1-\lambda) D_\alpha(x:q) \quad \text{with} \quad D_\alpha(p:q) \doteq \sum_{i=1}^{d} \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}p^i + \frac{1+\alpha}{2}q^i - (p^i)^{\frac{1-\alpha}{2}}(q^i)^{\frac{1+\alpha}{2}}\right)$$

K-means (hard/flat clustering)

Algorithm 1: Mixed α-seeding; $\mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$

Input: Weighted histogram set $\mathcal{H}$, integer $k \geq 1$, real $\lambda \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} \leftarrow h_j$ with uniform probability ;

for $i = 2, 3, \ldots, k$ **do**

  Pick at random histogram $h \in \mathcal{H}$ with probability:

$$\pi_{\mathcal{H}}(h) \;\doteq\; \frac{w_h M_{\lambda,\alpha}(c_h : h : c_h)}{\sum_{y\in\mathcal{H}} w_y M_{\lambda,\alpha}(c_y : y : c_y)},$$

  //where $(c_h, c_h) \doteq \arg\min_{(z,z)\in\mathcal{C}} M_{\lambda,\alpha}(z : h : z)$;

  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(h,h)\}$;

Output: Set of initial cluster centers $\mathcal{C}$;

Input: Weighted histogram set $\mathcal{H}$, integer $k > 0$, real $\lambda \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^{k} \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

  //Assignment

  for $i = 1, 2, \ldots, k$ **do**

    $\mathcal{A}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

  // Centroid relocation

  for $i = 1, 2, \ldots, k$ **do**

    $r_i \leftarrow \left(\sum_{h\in\mathcal{A}_i} w_i h^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

    $l_i \leftarrow \left(\sum_{h\in\mathcal{A}_i} w_i h^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

until *convergence*;

Output: Partition of $\mathcal{H}$ in k clusters following $\mathcal{C}$;

$$J_\alpha(\tilde{p} : \tilde{q}) = \frac{8}{1-\alpha^2}\left(1 + \sum_{i=1}^{d} H_{\frac{1-\alpha}{2}}(\tilde{p}^i, \tilde{q}^i)\right)$$

$$H_\beta(a,b) = \frac{a^\beta b^{1-\beta} + a^{1-\beta} b^\beta}{2}$$

Heinz means interpolate the arithmetic and the geometric means

$$\sqrt{ab} = H_{\frac{1}{2}}(a,b) \leq H_\alpha(a,b) \leq H_0(a,b) = \frac{a+b}{2}$$

EM (soft/generative clustering)

Input: Histogram set $\mathcal{H}$ with $|\mathcal{H}| = m$, integer $k > 0$, real $\lambda \leftarrow \lambda_{\text{init}} \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^{k} \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

  //Expectation

  for $i = 1, 2, \ldots, m$ **do**

    for $j = 1, 2, \ldots, k$ **do**

      $p(j|h_i) = \frac{\pi_j \exp(-M_{\lambda,\alpha}(l_j : h_i : r_j))}{\sum_{j'} \pi_{j'} \exp(-M_{\lambda,\alpha}(l_{j'} : h_i : r_{j'}))}$;

  //Maximization

  for $j = 1, 2, \ldots, k$ **do**

    $\pi_j \leftarrow \frac{1}{m}\sum_i p(j|h_i)$;

    $l_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

    $r_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

  //Alpha - Lambda

  $\alpha \leftarrow \alpha - \eta_1 \sum_{j=1}^{k}\sum_{i=1}^{m} p(j|h_i) \frac{\partial}{\partial\alpha} M_{\lambda,\alpha}(l_j : h_i : r_j)$;

  if $\lambda_{\text{init}} \neq 0, 1$ **then**

    $\lambda \leftarrow \lambda - \eta_2 \Big(\sum_{j=1}^{k}\sum_{i=1}^{m} p(j|h_i) D_\alpha(l_j : h_i) -$

    $\sum_{j=1}^{k}\sum_{i=1}^{m} p(j|h_i) D_\alpha(h_i : r_j)\Big)$;

    //for some small η_1, η_2; ensure that $\lambda \in [0,1]$.

until *convergence*;

Output: Soft clustering of $\mathcal{H}$ according to k densities $p(j|.)$ following $\mathcal{C}$;

@FrnkNlsn

On Clustering Histograms with k-Means by Using Mixed α-Divergences. Entropy 16(6): 3273-3301 (2014)

zbMATH Documents Authors Serials Classification Software Formulæ

Collaboration Distance

Compute co-authors path between

nielsen.frank and erdos.paul

Co-authors path:

The co-authors path between Nielsen, Frank and Erdős, Pál has **3** edges.

0. Nielsen, Frank

Efrat, Alon; Katz, Matthew J.; Nielsen, Frank; Sharir, Micha
Dynamic data structures for fat objects and their applications. (English)
Zbl 0952.68147
Comput. Geom. 15, No. 4, 215-227 (2000).

1. Sharir, Micha

Agarwal, Pankaj K.; Apfelbaum, Roel; Purdy, George; Sharir, Micha
Similar simplices in a d-dimensional point set. (English) Zbl 1218.52021
Proceedings of the 23rd annual symposium on computational geometry 2007, Gyeongiu, South Korea, June 6–8, 2007. New York, NY: Association for Computing Machinery (ISBN 978-1-59593-705-6). 232-238 (2007).

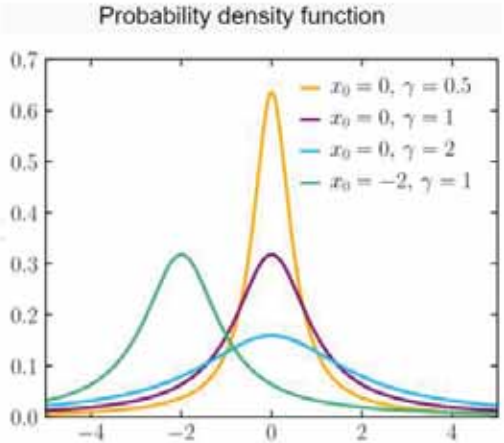
2. Purdy, George B.

Erdős, Paul; Purdy, George
Extremal problems in combinatorial geometry. (English) Zbl 0852.52009
Graham, R. L. (ed.) et al., Handbook of combinatorics. Vol. 1-2. Amsterdam: Elsevier (North-Holland), 809-874 (1995).

3. Erdős, Pál

Complex parameterization of the Cauchy distributions

- McCullagh (1992) proposed to parameterize the Cauchy distributions with a **complex parameter** θ instead of a 2D parameter λ:



$$p_\lambda(x) = \frac{s}{\pi(s^2 + (x-l)^2)} \quad \Leftrightarrow \quad p_\theta(x) = \frac{\mathrm{Im}(\theta)}{\pi|x-\theta|^2}$$

$$\lambda(\theta) = (l, s) = (\mathrm{Re}(\theta), |\mathrm{Im}(\theta)|) \in \mathbb{R}^2 \quad \Leftrightarrow \quad \theta(\lambda) = l + is \in \mathbb{C}$$

- Action of SL(2,R) = **real linear fractional transformation**:

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{R}), \quad M.\theta := \frac{a\theta + b}{c\theta + d}$$

- **<u>Benefit 1</u>**: $$X \sim \mathrm{Cauchy}\,(\theta) \Rightarrow \frac{aX+b}{cX+d} = M.X \sim \mathrm{Cauchy}\left(\frac{a\theta+b}{c\theta+d}\right)$$

- **<u>Benefit 2</u>**: $I_f(p:q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x$ $$\boxed{\forall M \in \mathrm{SL}(2,\mathbb{R}), \quad I_f(p_{M.\theta_1} : p_{M.\theta_2}) = I_f(p_{\theta_1} : p_{\theta_2})}$$

On f-divergences between Cauchy distributions, arXiv 2101.12459 2021

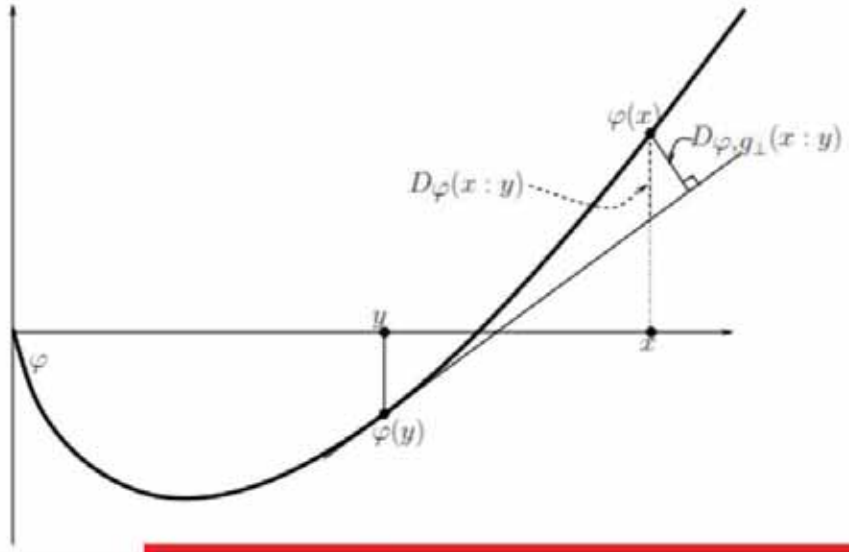
Conformal Bregman divergences and their centroids

A (right-sided) *conformal divergence*, $D_{\varphi,g}$, is parameterized by two real-valued functions φ and g with $\mathrm{img} \subseteq (0, +\infty)$, whose domains are a compact convex of $\mathbb{R}^d$. The expression of $D_{\varphi,g}$ is:

$$\boxed{D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y}) \doteq g(\boldsymbol{y}) D_\varphi(\boldsymbol{x} : \boldsymbol{y})}$$

φ is real-valued strictly convex twice differentiable, and $D_\varphi(\boldsymbol{x} : \boldsymbol{y})$ is the ordinary Bregman divergence with generator φ:

$$D_\varphi(\boldsymbol{x} : \boldsymbol{y}) = \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^\top \nabla\varphi(\boldsymbol{y})$$



	ordinary Bregman divergence	total Bregman divergence
Right population minimizer	centroid	this paper (Corollary 1)
Left population minimizer	$\nabla\varphi$-mean, weights $w_i = 1/n$	$\nabla\varphi$-mean, weights $w_i = 1/\sqrt{1 + \lVert\nabla\varphi(\boldsymbol{x}_i)\rVert^2}$
Robustness of pop. minimizers	no	yes
Information	Bregman information	this paper (Lemma 3)
Exhaustiveness	right pop. minimizer = sample average	this paper (Theorem 2)
Bijection	exponential families	lifted exponential families
Inference	MLE = observed point ($\hat{\boldsymbol{\eta}}$=right population minimizer)	Bayesian MAP = total observed point (left population minimizer)

On conformal divergences and their population minimizers,IEEE Transactions on Information Theory 62.1 (2015)

Conformal divergences

Rescale divergence D by a conformal factor ρ:
(induced Riemannian tensor is scaled, conformal geometry
conformal = preserves angles)

$$\boxed{D'(p:q) = \rho(p,q)D(p:q)}$$

Examples:

- Total Bregman divergences: $\mathrm{tB}(p:q) = \rho_B(q)B(p:q), \quad \rho_B(q) = \sqrt{\frac{1}{1+\langle \nabla F(q), \nabla F(q)\rangle}}$
- Total Jensen divergences: $\mathrm{tJ}_\alpha(p:q) = \rho_J(p,q)J_\alpha(p:q), \quad \rho_J(p,q) = \sqrt{\frac{1}{1+\frac{(F(p)-F(q))^2}{\langle p-q,p-q\rangle}}}$
- (M,N)-Bregman divergences: $$B_F^{\rho,\tau}(p:q) = \frac{\tau(F(p))-\tau(F(q))}{\tau'(F(q))} - \frac{\rho(p)-\rho(q)}{\rho'(q)}F'(q),$$
$$= \frac{1}{\tau'(F(q))}B_{\tau\circ F\circ\rho^{-1}}(\rho(p):\rho(q)) \geq 0,$$

On conformal divergences and their population minimizers, IEEE Trans. IT 2016
Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity, IEEE SPL 2017
Total Jensen divergences: definition, properties and clustering, IEEE ICASSP 2015
Shape retrieval using hierarchical total Bregman soft clustering, IEEE TPAMI, 2012

Conformally-projectively equivalent statistical manifolds

Conformal divergence: $\boxed{D_\rho(p:q) = e^{\rho(p:q)} D(p:q)}$

- Case $D_{\rho_l}(p:q) = e^{\rho_l(p)} D(p:q)$

  $(M, {}^D g, {}^D\nabla)$ and $(M, {}^{D_{\rho_l}} g, {}^{D_{\rho_l}}\nabla)$ are **(-1)-conformally equivalent**

- Case $D_{\rho_r}(p:q) = e^{\rho_r(q)} D(p:q)$ **(eg., total Bregman divergence)**

  $(M, {}^D g, {}^D\nabla)$ and $(M, {}^{D_{\rho_r}} g, {}^{D_{\rho_r}}\nabla)$ are **1-conformally equivalent**

- Case $D_\rho(p:q) = e^{\rho(p:q)} D(p:q)$ **(eg., total Jensen divergence)**

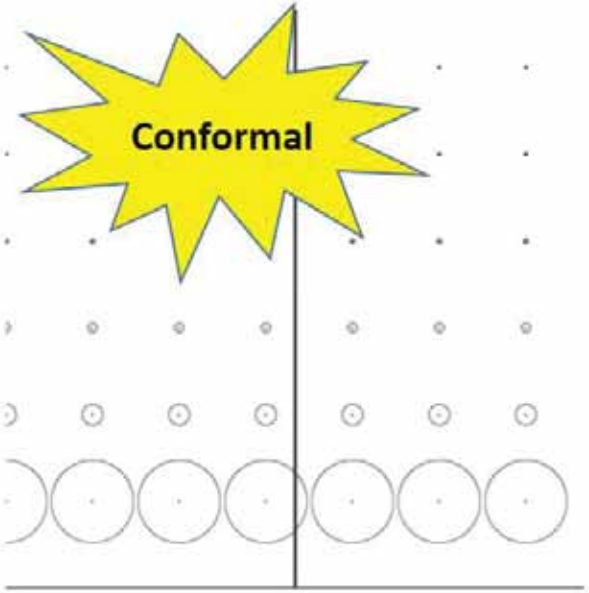
  $(M, {}^D g, {}^D\nabla)$ and $(M, {}^{D_\rho} g, {}^{D_\rho}\nabla)$ are **conformally projectively equivalent**

@FrnkNlsn

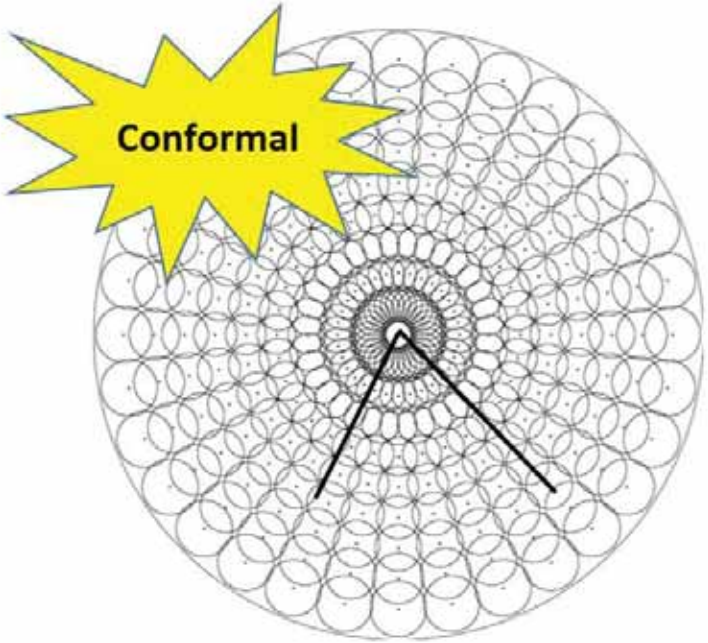
Total Jensen divergences: definition, properties and clustering, IEEE ICASSP 2015
Shape retrieval using hierarchical total Bregman soft clustering, IEEE TPAMI 2012

Conformal versus non-conformal metric tensor field:

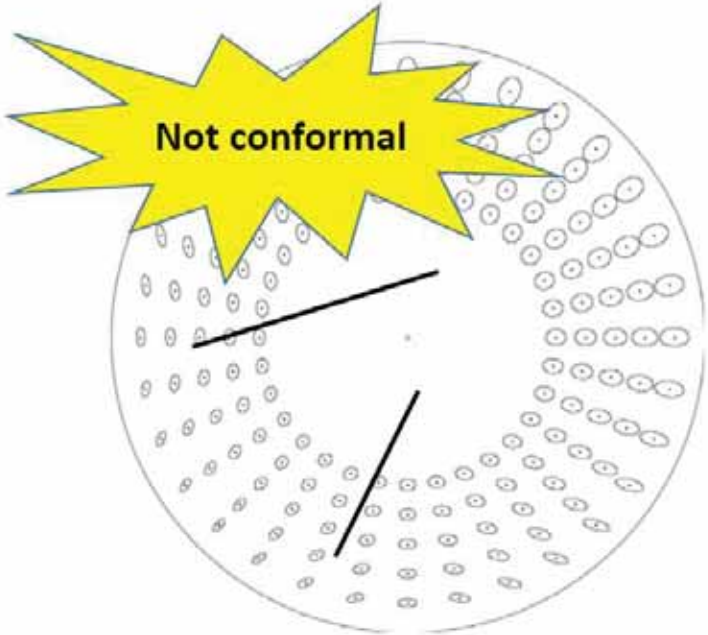
Hyperbolic geometry



Upper Poincare plane
(conformal)



Poincare disk
(conformal)



Klein disk
(non-conformal)

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

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

In conformal geometry, we can measure angles without distortions

∇ connection: Affine coordinate system, flatness & geodesics

- (Local) **affine coordinate system** θ wrt ∇: $\nabla_{\partial_i}\partial_j = 0$ where $\partial_i := \frac{\partial}{\partial\theta^i}$
- Since $\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}^k \partial_k = 0 \Rightarrow \Gamma_{ik}^k(\theta) = 0$

  →**Christoffel symbols vanish** in the θ-coordinate system
- ∇ connection is **flat** iff $\Gamma_{ik}^k(\theta) = 0$

  Parallel transport does not depend on the smooth curve but only its endpoints
- ∇-**geodesic**: autoparallel curve $\nabla_{\dot\gamma}\dot\gamma = 0$

Geodesic equation: $\ddot\theta^i + \sum_{j,k} \Gamma_{jk}^i \dot\theta^j \dot\theta^k = 0$

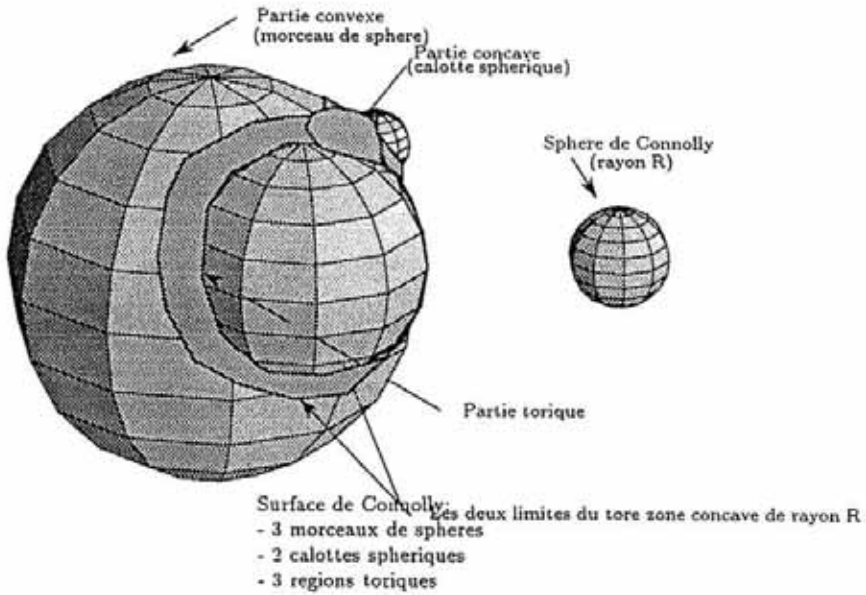
When ∇ is flat,

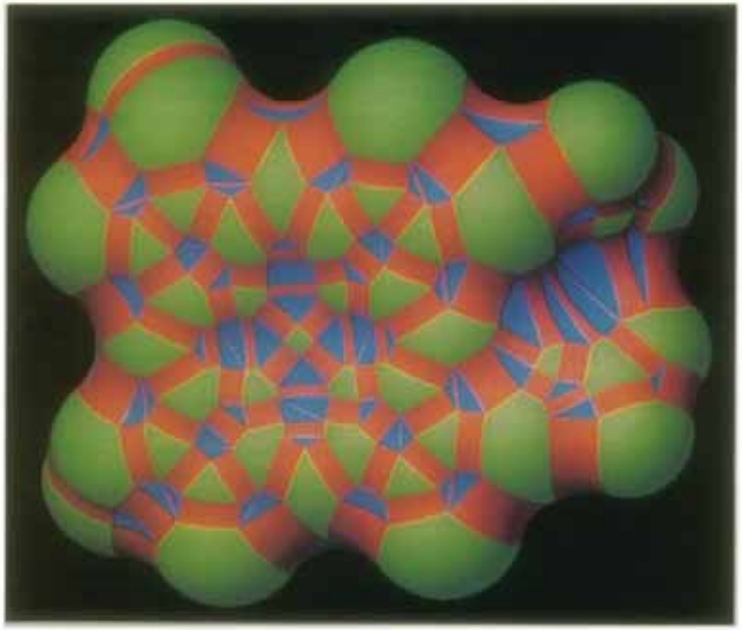
$$\ddot\theta^i = 0 \Rightarrow \theta_1 + t(\theta_2 - \theta_1)$$

$\gamma(t)$ and $\gamma(at+b)$: same geodesics: t is called an **affine parameter**

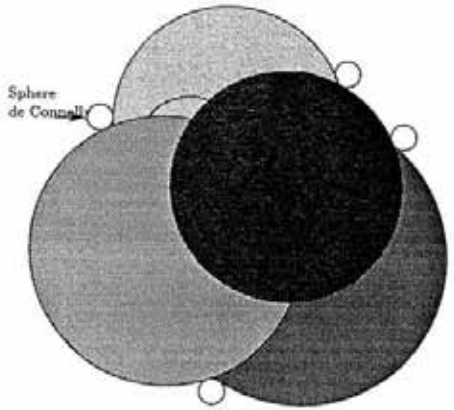
An Elementary Introduction to Information Geometry, Entropy 2020, 22(10)

Connolly surfaces and convex hulls of spheres

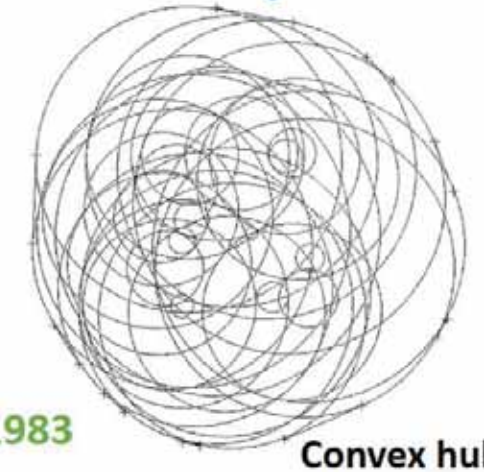




Connolly surface (© Connolly, 1983)



Probe radius → ∞



Convex hull

- Rolling a probe sphere (solvent) on a set of spheres (molecules)….
- Smooth molecular water-tight surface (analytic surface/volume calculation)
- Concave spherical triangles, saddle-shaped rectangles, convex spherical polygons)

Connolly, Solvent-accessible surfaces of proteins and nucleic acids. Science 1983
Connolly, Analytical molecular surface calculation, Journal of applied crystallography 1983
An Output-Sensitive Convex Hull Algorithm for Planar Objects, IJCGA 1995

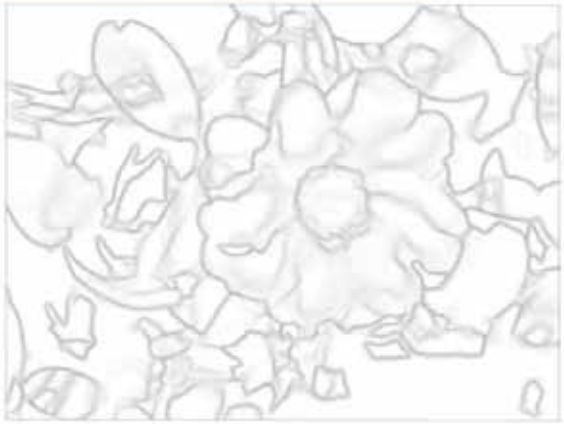
Consensus-based image segmentation

A never-ending image segmentation algorithm which self-improves:

- Design a **randomized image segmentation** algorithm
  e.g., Statistical Region Merging (SRM)
- Every run yields a different image segmentation (for a different seed)
  → kind of "statistical algorithm"
- Make a consensus of all segmentation results
  → kind of calculate the "expectation" of the statistical algorithm

Fig. 2. A population of segmentations: Several random segmentation variates ($l = 12$) using non-deterministic RRM.

Consensus



Soft contour map

Consensus region merging for image segmentation, IEEE ACPR 2013
Statistical Region Merging, IEEE TPAMI 2004 (CVPR 2003)

Contextual Bregman divergences via Bregman projections

- Context*ual* dissimilarity:

$$\boxed{\mathrm{cs}_f(q,p|u) = \arg\min_{0\le\omega\le 1} \Phi_g(\omega;q,p,u)}$$

where $\Phi_f(\omega;q,p,u) = f(q,\omega p+(1-\omega)u)$

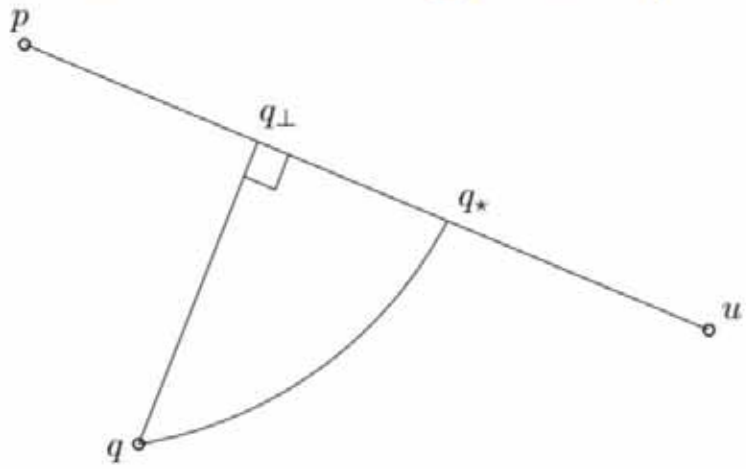
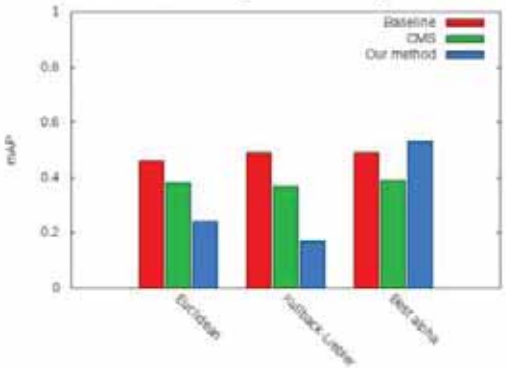


Figure : Contextual similarity and Bregman projection.

- Reranking with multiple contexts (CBIR):

$$\mathrm{cs}_f(q,p) = \sum_{u\in\mathcal{U}} \mathrm{cs}_f(q,p|u)$$



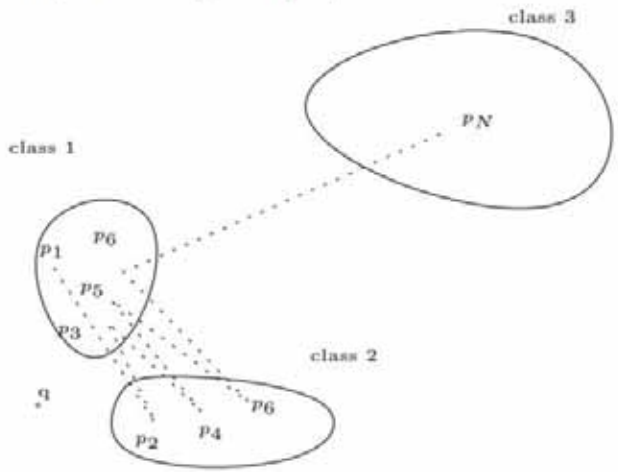


Figure : Different contexts and potential rank lists.

Reranking with Contextual Dissimilarity Measures from Representational Bregman k-Means, VISAPP 2010

Contravariant Cramér–Rao lower bound

- Fisher information matrix is **covariant** under reparameterization:

$$I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial \theta_i}{\partial \eta_j}\right]$$

covariant

- **Loewner partial ordering** on the SPD matrix cone:

$$A \succeq B \Leftrightarrow \forall x, x^\top (A-B)x \geq 0$$

- **Cramér–Rao-Darmois- Fréchet lower bound** (CRLB):

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I_\theta^{-1}(\theta)$$

Variance of an unbiased estimator:

- **Jacobian of an inverse function f is the inverse of the Jacobian of f**:

$$\mathrm{Jac}_{\eta(\theta)} := \left[\frac{\partial \eta_i}{\partial \theta_j}\right] = (\mathrm{Jac}_{\eta^{-1}(\theta)})^{-1} = (\mathrm{Jac}_{\theta(\eta)})^{-1} := \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^{-1} \Rightarrow \left[\frac{\partial \theta_i}{\partial \eta_j}\right] \times \left[\frac{\partial \eta_i}{\partial \theta_j}\right] = I_{D\times D}$$

- CRLB is **contravariant** under reparameterization:

$$\boxed{\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I_\theta^{-1}(\theta) \xrightarrow{\eta=\eta(\theta)} \mathrm{Var}[\hat{\eta}_n] \succeq \frac{1}{n} \left[\frac{\partial \eta_i}{\partial \theta_j}\right] I_\theta(\theta(\eta))^{-1} \left[\frac{\partial \eta_i}{\partial \theta_j}\right]^\top}$$

contravariant

Converting similarities S $\leftrightarrow$ distances D

D: Distance measure S: Similarity measure

$$\boxed{\begin{aligned} S(p,q) &= \exp(-D(p,q)) \in (0,1] \\ D(p,q) &= -\log S(p,q) \in [0,\infty) \end{aligned}}$$

Additive triangular inequality of metric distances:

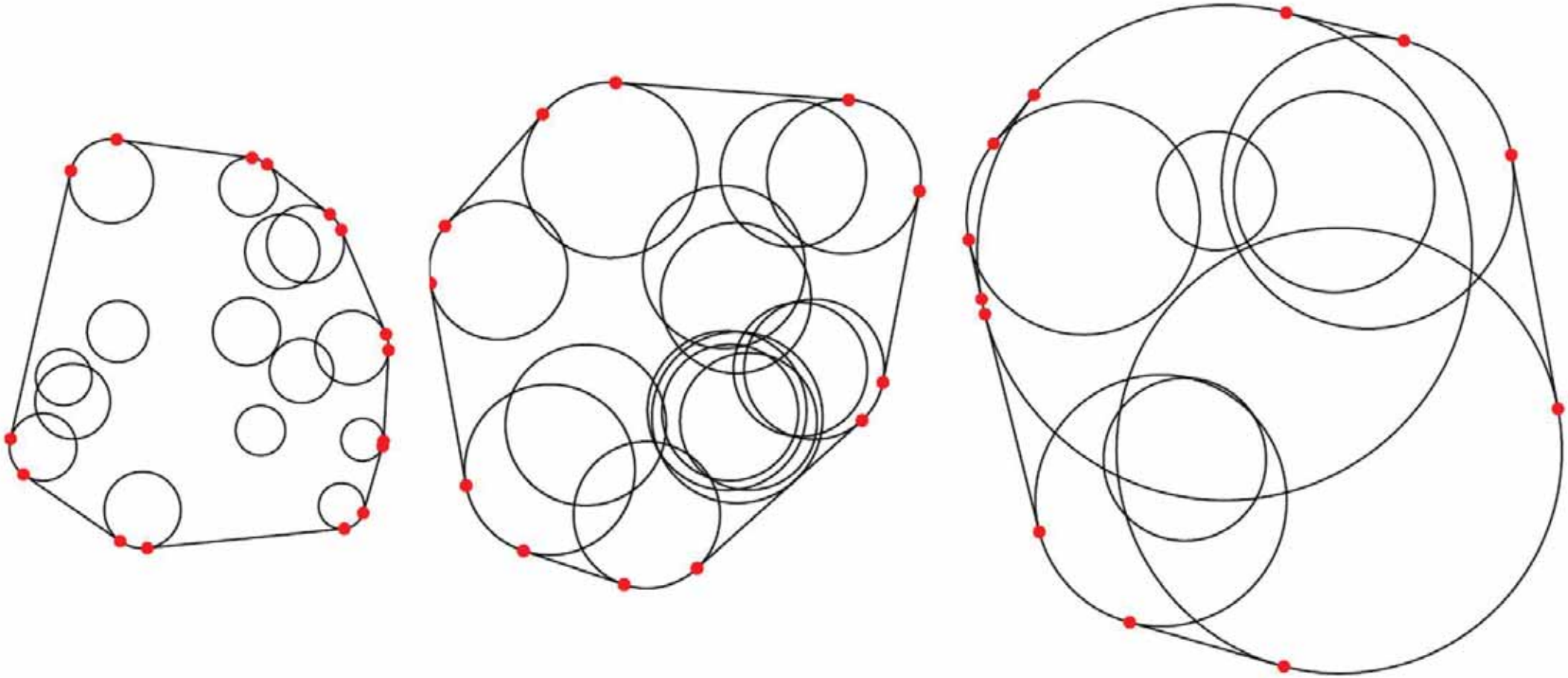
$$D(p,q) + D(q,r) \geq D(p,r)$$

$$\Updownarrow$$

Multiplicative triangular inequality of similarities:

$$S(p,q) \times S(q,r) \leq S(p,r)$$

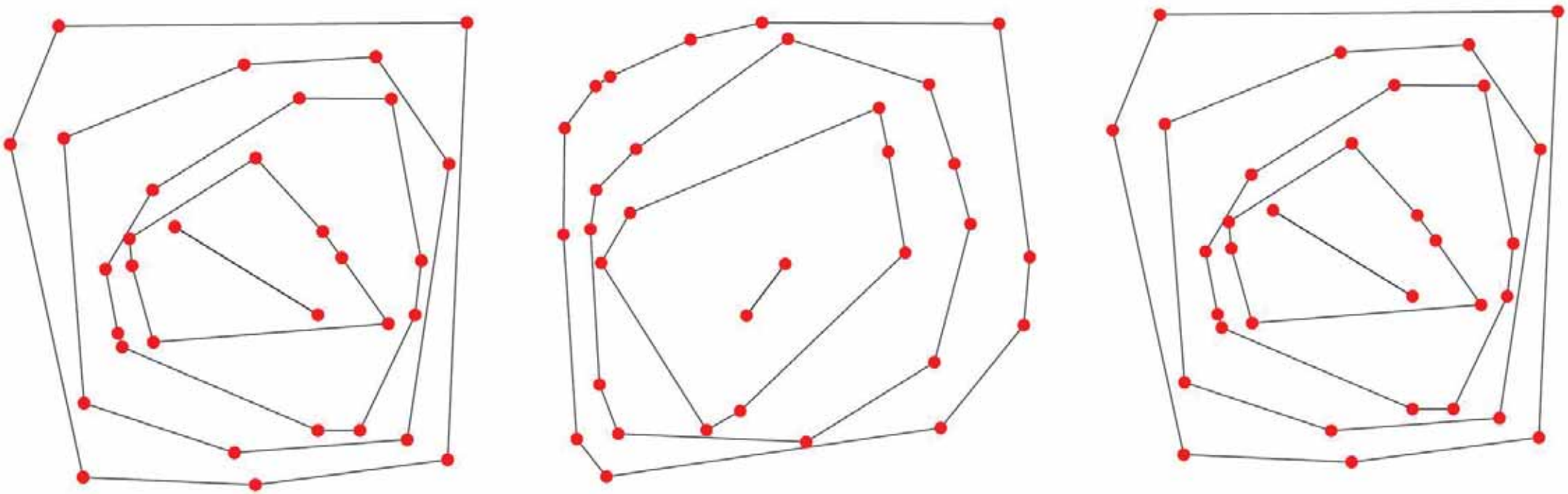
Optimal output-sensitive convex hull algorithm of 2D disks



Output-Sensitive Convex Hull Algorithms of Planar Convex Objects, IJCGA (1998)

https://franknielsen.github.io/ConvexHullDisk/

Convex layers: convex peeling of point sets



Optimal O(nlog h)-time algorithm for peeling the first l layers where h is the number of points on first first l layers

Output-sensitive peeling of convex and maximal layers, IPL, 1996

COSH distance: A symmetrized Bregman divergence

$$
\begin{aligned}
D_{\mathrm{IS}}(\theta_1 : \theta_2) &= B_{F_{\mathrm{Burg}}}(\theta_1 : \theta_2) \\
&= \frac{1}{d}\sum_{i=1}^{d}\left(\frac{\theta_1^i}{\theta_2^i} - \log\frac{\theta_1^i}{\theta_2^i} - 1\right)
\end{aligned}
$$

$$\theta = (\theta^1, \ldots, \theta^d)$$

$$F_{\mathrm{Burg}}(\theta) = -\frac{1}{d}\sum_{i=1}^{d}\log\theta^i$$

$$
\boxed{
\begin{aligned}
D_{\mathrm{COSH}}(\theta_1 : \theta_2) &= \frac{1}{2}(B_{F_{\mathrm{Burg}}}(\theta_1 : \theta_2) + B_{F_{\mathrm{Burg}}}(\theta_2 : \theta_1)) \\
&= \underline{\frac{1}{2d}\sum_{i=1}^{d}\left(\frac{\theta_1^i}{\theta_2^i} + \frac{\theta_2^i}{\theta_1^i} - 2\right)} \\
&= \frac{1}{2}(\theta_2 - \theta_1)^\top(\eta_2 - \eta_1) \\
&= \frac{1}{2}\sum_{i=1}^{d}(\theta_2^i - \theta_1^i)\left(\frac{1}{\theta_1^i} - \frac{1}{\theta_2^i}\right)
\end{aligned}
}
$$

$$\eta^i = \frac{\partial}{\partial\theta^i}F_{\mathrm{Burg}}(\theta) = -\frac{1}{d\theta^i}$$

$$\eta = \nabla F_{\mathrm{Burg}}(\theta) = -\frac{1}{d}\left(\frac{1}{\theta^1}, \ldots, \frac{1}{\theta^d}\right)$$

Gray and. Markel, "COSH Measure for Speech Processing," 90th Mtg., Acoustical Society Of America, Vol. 58 (1), p. S97, 1975.
Sided and Symmetrized Bregman Centroids, IEEE Transactions on Information Theory, 2009

Covariance matrices & centering matrices

$$E[AX+b] = A\,E[X] + b, \quad \mathrm{Cov}[AX+b] = A\,\mathrm{Cov}[X]\,A^\top$$

Empirical covariance matrix: $C = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top. \quad \bar{x} = \frac{1}{n}\sum_i x_i$

$$\boxed{\begin{aligned} C &= X\,J J^\top X^T \\ J &= \frac{1}{\sqrt{n}}(I_n - \frac{1}{n}1_{n,n}) \end{aligned}}$$

$X = [x_1 | \vdots | x_n]$

Centering matrix:

$$1_{n,n} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

Covariance and invariance on a Bregman manifold

- **Bregman manifold** induced by a strictly convex *Legendre-type function* F(θ)
- **Convex conjugate** F*(η) = *Legendre-type function:* η=∇F(θ) and θ =∇F*(η)
- In information geometry, **exponential family** and **mixture family** yield Bregman manifolds, also called dually flat spaces
- **Crouzeix identity** of Hessians: **$\nabla^2 F(\theta) \times \nabla^2 F^*(\eta) = I$**, the identity matrix
- **Fisher Information Matrix** (FIM) is **covariant**: $\left[\frac{\partial \theta}{\partial \eta}\right] = \nabla\nabla F^*(\eta) =: \nabla^2 F^*(\eta)$ and $(\nabla^2 F^*(\eta))^\top = \nabla^2 F^*(\eta)$

$$I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top I_\theta(\theta(\eta)) \left[\frac{\partial \theta_i}{\partial \eta_j}\right] \Rightarrow I_\eta(\eta) = \underbrace{\nabla^2 F^*(\eta)\nabla^2 F(\theta)}_{=I} \nabla^2 F^*(\eta) = \nabla^2 F^*(\eta)$$

- **Fisher-Rao geodesic distance** is **invariant**: $\mathrm{d}\eta = \nabla^2 F(\theta)\mathrm{d}\theta$ and $\mathrm{d}\theta = \nabla^2 F^*(\eta)\mathrm{d}\eta$

$$\begin{aligned} \mathrm{d}s_\theta^2 &= \mathrm{d}\theta^\top \nabla^2 F(\theta)\mathrm{d}\theta \\ &= \mathrm{d}\eta^\top \underbrace{\nabla^2 F^*(\eta)\nabla^2 F(\theta)}_{=I} \nabla^2 F^*(\eta)\mathrm{d}\eta \\ &= \mathrm{d}s_\eta^2 \end{aligned} \Rightarrow \rho_{\mathrm{FR}}(p_{\theta_1}, p_{\theta_2}) = \rho_{\mathrm{FR}}(p_{\eta_1}, p_{\eta_2})$$

On Geodesic Triangles with Right Angles in a Dually Flat Space, Progress in Information Geometry, Springer 2021

Contravariant and covariant transformation rules

- On a D-dimensional manifold M, consider **local coordinate systems**: λ and $\hat{\lambda}$
- A vector in a tangent plane can be expressed using either the **natural basis** $\left\{e_i = \frac{\partial}{\partial \lambda^i}\right\}_i$ with respect to λ or the natural basis with respect to $\hat{\lambda}$: $\left\{\hat{e}_i = \frac{\partial}{\partial \hat{\lambda}^i}\right\}_i$
- Vector expressed in the two basis: $v = \sum_{i=1}^{D} v^i e_i = \sum_{i=1}^{D} \hat{v}^i \hat{e}_i$
- The **contravariant transformation rule** is $\hat{v}^i = \sum_{j=1} v^j \frac{\partial \hat{\lambda}^i}{\partial \lambda^j}$
- The **covariant transformation rule (e.g., gradient vector)** is $\hat{v}_i = \sum_{j=1}^{D} v_j \frac{\partial \lambda^j}{\partial \hat{\lambda}^i}$

An elementary introduction to information geometry, arXiv:1808.08271

Cramér l_p distances

- For X and Y two independent **univariate** random variables with **cumulative distributions functions** F_X and F_Y:

$$\boxed{D_{\text{Cramér}}[X,Y] := \int_{-\infty}^{+\infty} (F_X(x) - F_Y(x))^2 \, \mathrm{d}x}$$

- $\sqrt{D_{\text{Cramér}}[X,Y]}$ Square root of Cramér distance is a metric
- Can be extended to **Cramér l_p distances** for $p \geq 1$:

$$D_{\text{Cramér},p}[X,Y] := \left(\int_{-\infty}^{\infty} (F_X(x) - F_Y(x))^p \, \mathrm{d}x \right)^{\frac{1}{p}}$$

- When p=1, Cramér distance coincides with the **Wasserstein l_1-distance**
- Can be generalized to multivariate cases with the **Székely energy distance**

Harald Cramér
1893-1985

Cramér , On the composition of elementary errors, Scandinavian Actuarial Journal (1928)

Fisher information: Chain rule and Cramér-Rao bound

- Fisher information quantifies the information a random variable X carries about a parameter θ: $I_X(\theta) := E_{X\sim p_\theta(x)}\left[(\nabla_\theta \log p_\theta(X))^\top \nabla_\theta \log p_\theta(X)\right]$

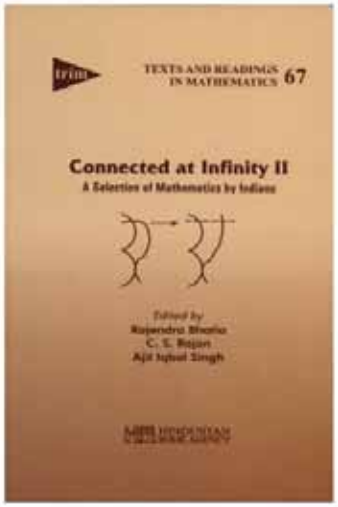
- Cramér-Rao lower bound (CRLB) gives the minimum variance of any unbiased estimator (with respect to Lowner ordering): $\mathrm{Var}_{X\sim p_\theta(x)}[\hat{\theta}(X)] \succeq I_X(\theta)^{-1}$

- Chain rule of the Fisher information for joint distributions (X,Y):

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_{Y|X}(\theta)$$

- For a iid random vector $X_{1:n} = (X_1, \ldots, X_n)$ have the CRLB:

$$\boxed{I_{X_{1:n}} = nI_X(\theta) \quad \Rightarrow \quad \mathrm{Var}_{X_{1:n}\sim\prod p_\theta(x_i)}[\hat{\theta}(X_{1:n})] \succeq \frac{1}{n} I_X(\theta)^{-1}}$$



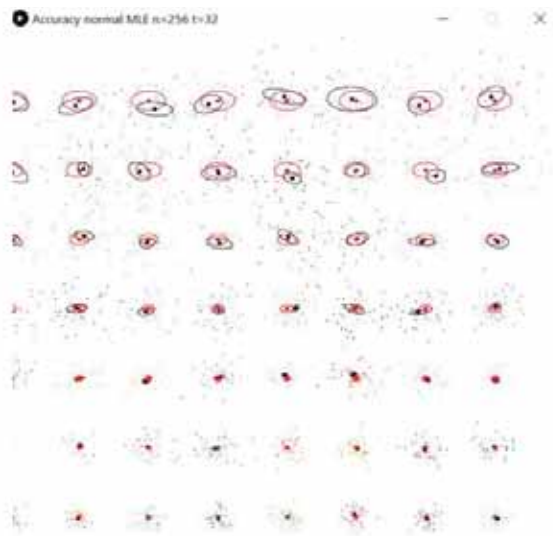
Cramér-Rao lower bound and information geometry, Connected at Infinity II, p. 18-37 (2013)

Sensitivity and accuracy of estimators

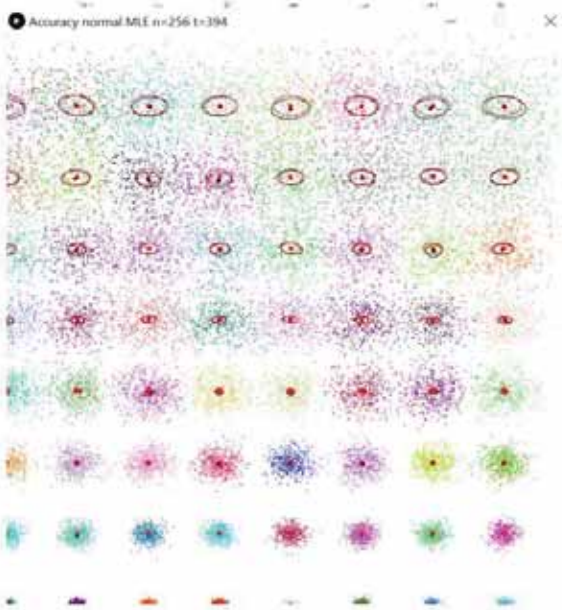
The accuracy of an estimator is not homogeneous and depends on the underlying true parameter: Consider the family of normal distributions N(μ,σ)

For each grid position (μ, σ),

- For $t \in \{1, \ldots, T\}$ (number of trials):
  - Draw i.i.d. s normal variates, estimate the MLE X with $(\hat{\mu}_t, \hat{\sigma}_t)$
- Calculate sample mean $(\tilde{\mu}, \tilde{\sigma})$ and the sample covariance matrix $\tilde{\Sigma}$ for $\{(\hat{\mu}_t, \hat{\sigma}_t)\}_{i \in [T]}$.



T=32



T=394

Cramer-Rao lower bound:

The variance of an unbiased estimator is lower bounded by the inverse Fisher information matrix (IFIM):

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

$$\boxed{\mathrm{Var}[X - (\mu, \sigma)] \underset{\text{Lowner ordering}}{\succeq} \frac{1}{n} I^{-1}(\mu, \sigma)}$$

Asymptotic normality of the MLE: $X \sim N\left((\mu, \sigma), \frac{1}{n} I^{-1}(\mu, \sigma)\right)$

Cramér-Rao lower bound and information geometry, Connected at Infinity II, 2013. 18-37.

Cramér-Rao lower bound & Fisher additive information

Random vector: $X_{1:n} = (X_1, \ldots X_n) \sim_{\text{ind}} p_\theta$ $\qquad X \sim p_\theta$

Cramér-Rao lower bound on the variance of any unbiased estimator:

$$\boxed{\mathrm{Var}[\hat{\theta}_n] \succeq I^{-1}_{X_{1:n}}(\theta)} \Longleftrightarrow \mathrm{Var}[\hat{\theta}_n] - I^{-1}_{X_{1:n}}(\theta) \text{ is positive-semidefinite}$$

Usually **written as**

$$\boxed{\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I^{-1}_X(\theta)}$$

because Fisher information of iid random vector is **additive**:

$$I_{X_{1:n}}(\theta) = n I_X(\theta)$$

Cramér-Rao lower bound and information geometry, *Connected at Infinity II*. Hindustan Book Agency, Gurgaon, 2013. 18-37.

Upper bounding distances between mixtures: Joint convexity

Definition (Jointly convex distance). *A distance $D(\cdot:\cdot)$ on a statistical manifold $\mathcal{M}$ is jointly convex if and only if*

$$D((1-\alpha)p_1+\alpha p_2 : (1-\alpha)q_1+\alpha q_2) \leq (1-\alpha)D(p_1:p_2)+\alpha D(p_2:q_2), \quad \forall \alpha \in [0,1],\ p_1,p_2 \in \mathcal{M}.$$

Examples of joint convex distances: **Kullback-Leibler divergences** and **f-divergences**

Two mixtures: $m_1(x) = \sum_{i=1}^{k_1} \alpha_i p_i(x)$ and $m_2(x) = \sum_{i=1}^{k_2} \beta_i q_i(x)$

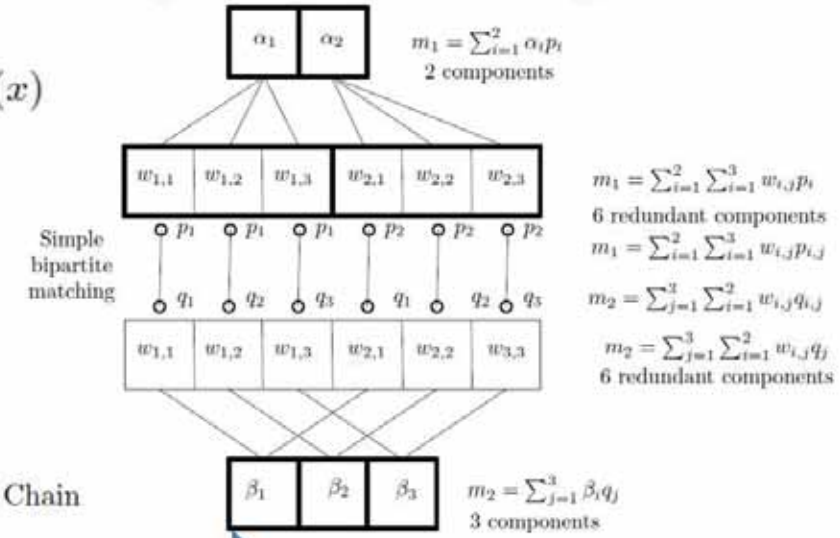
$$\boxed{H_D(m_1,m_2) = \min_{W \in U(\alpha,\beta)} \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{ij} D(p_i,q_j).}$$

Transport polytope

when k1=k2=d:

$$U(\alpha,\beta) = \{P \in \mathbb{R}_+^{d\times d} : P1_d = \alpha, P^\top 1_d = \beta\}$$

Sinkhorn entropic relaxation



Definition (CROT distance). *Given two multivariate distributions $p(x,y)$ and $q(x,y)$, we define the* Chain Rule Optimal Transport *as follows:*

$$\boxed{H_D(p,q) := \inf_{r \in \Gamma(p(y),q(z))} E_{r(y,z)}\left[D\Big(p(x|y), q(x|z)\Big)\right]}$$

where $D(\cdot,\cdot)$ is a ground distance defined on conditional density manifold $\mathcal{C} = \{p(x|y)\}$ (e.g., the Total Variation), $\Gamma(p(y),q(z))$ is the set of all probability measures on $\mathcal{Y}^2$ satisfying the constraints $\int r(y,z)\mathrm{d}z = p(y)$ and $\int r(y,z)\mathrm{d}y = q(z)$, and $E_{r(y,z)}$ denotes the expectation with respect to $r(y,z)$.

CROT is a metric when D is a metric

Chain Rule Optimal Transport (CROT) arXiv:1812.08113

Some statistical distances with closed-form expressions for statistical mixtures…

- **Cauchy-Schwarz divergence**: $\mathrm{CS}(P:Q) = -\log \frac{\int p(x)q(x)\mathrm{d}x}{\sqrt{\int p(x)^2\mathrm{d}x \int q(x)^2\mathrm{d}x}},$

$$p_\theta(x) = \exp\left(t(x)^\top \theta - F(\theta)\right)$$

- For mixtures of *exponential families* like the Gaussian family with **conic natural parameter space**:

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

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

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

Only holds when natural parameter space is a cone

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

Sided f-divergences and duality: Historical note

- Csiszar defined the **f-divergences** in 1964 (defended his thesis in 1966):

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 329—339.

ON TOPOLOGICAL PROPERTIES OF *f*-DIVERGENCES

by

$$\mathscr{I}_f(\mu_1, \mu_2) = \int p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right) \lambda(dx) \tag{1}$$

- Nowadays, the f-divergences are conventionally defined by

$$I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x).$$

- The **sided f-divergences** are **reference dual divergences** $I_f{}^*[p:q] := I_f[q:p]$ related by **dual f-divergence generators** $f^\diamond(u) := uf\left(\frac{1}{u}\right)$:

Reference duality
(forward/backward divergences)

$$\boxed{I_f{}^*[p:q] = I_{f^\diamond}[p:q]}$$

Dual generators
(perspective function)

Csiszár, I. "On topological properties of f-divergences." Studia Math. Hungar. 2 (1967): 329-339

Curved Mahalanobis distances (Cayley-Klein geometry)

Usual squared **Mahalanobis distance** (Bregman divergence with dually flat geometry)

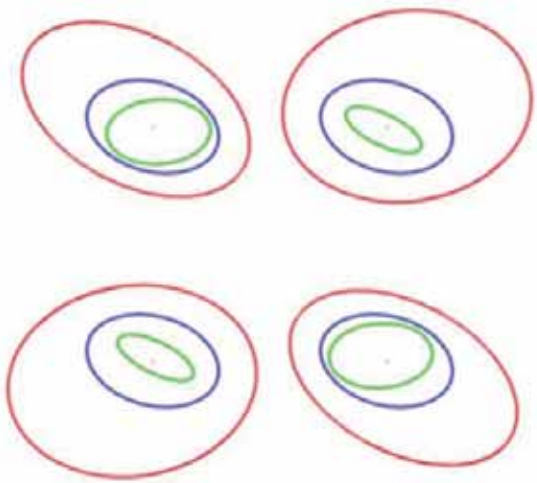
$$D_Q(p,q) = \sqrt{(p-q)^\top Q(p-q)}$$ where Q is positive-definite matrix

Curved Mahalanobis distance (centered at μ and of curvature κ):

$$\boxed{D_S(p,q) = \frac{1}{2|\kappa|}\arccos\left(\frac{|S(p,q)|}{\sqrt{S(p,p)S(q,q)}}\right)}$$

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

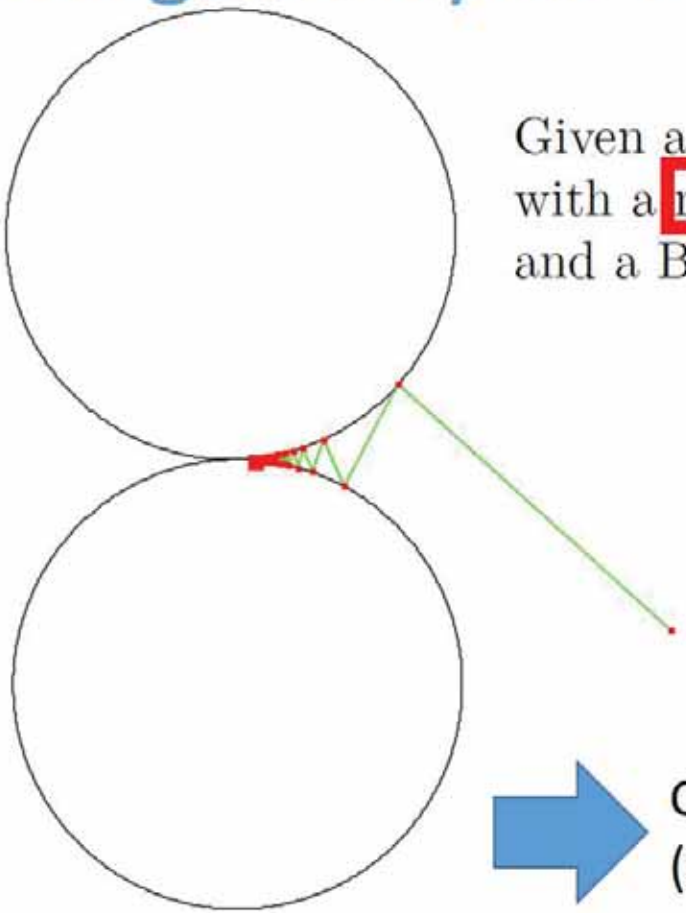
$$S(p,q) = S_{\Sigma,\mu,\kappa}(p,q) = (p-\mu)^\top \Sigma (q-\mu) + \operatorname{sign}(\kappa)\frac{1}{\kappa^2}$$



Some curved Mahalanobis balls (Mahalanobis in blue)

@FrnkNlsn **Classification with mixtures of curved Mahalanobis metrics, ICIP 2016.**

Bregman cyclic D-projections



Given a set of n convex objects $\{O_i\}$ of X with a non-empty intersection $\cap O_i \neq \emptyset$ and a Bregman divergence D.

Cyclic projections:

1. Let $p_0 \in X$, $t \leftarrow 0$
2. D-projection: $p_{t+1} = \arg\min_{o \in O_{(1+t) \bmod (n+1)}} D(p_t : o)$.
3. $t \leftarrow t + 1$. Goto 2.

Converge to a common point in the intersection
(Cyclic projections diverge when no common point of intersection)

THE RELAXATION METHOD OF FINDING THE COMMON POINT OF CONVEX SETS AND ITS APPLICATION TO THE SOLUTION OF PROBLEMS IN CONVEX PROGRAMMING*

L.M. BREGMAN
Leningrad

(Received 20 May 1966)

s paper we consider an iterative method of finding the common of convex sets. This method can be regarded as a generalization of the methods discussed in [1 - 4]. Apart from problems which can be reduced to finding some point of the intersection of convex sets, the method considered can be applied to the approximate solution of problems in linear and convex programming.

1. The problem of finding the common point of convex sets

L. M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." USSR Computational Mathematics and Physics, 7:200-217, 1967

De Bruijn identity: differential entropy & Fisher information

Nicolaas Govert de Bruijn
1918–2012

- Let X be a random variable with **finite variance**: $|E[X^2]| < \infty$
- Define $Z_\theta := X + \sqrt{\theta} Y$ $\qquad Y \sim \mathcal{N}(0,1)$
- Let h(.) denotes the **differential entropy** and I(.) the **Fisher information**
- Then we have the **De Bruijn identity**: $$\boxed{\frac{d}{d\theta} h(Z_\theta) = \frac{1}{2} I_{Z_\theta}(\theta)}$$

Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon (1959)

Deep transposition-invariant distances on sequences

Kendall's tau distance

$$\tau(x,y) = \frac{\sharp\{\text{concordant pairs}\} - \sharp\{\text{discordant pairs}\}}{N(N-1)/2}$$

Concordant pair $\Pi(x)_i < \Pi(x)_j$ and $\Pi(y)_i < \Pi(y)_j$

Disconcordant pair $\Pi(x)_i < \Pi(x)_j$ and $\Pi(y)_i > \Pi(y)_j$

Spearman's rho distance

(= l2-norm between their rank vectors)

$$\rho(x,y) = \sqrt{\sum_{i=1}^{N} (\Pi(x)_i - \Pi(y)_i)^2}.$$

Truncated Spearman's rho distance

(consider l most important coordinates)

$$\rho_l(x,y) = \sqrt{\sum_{i=1}^{l} (\Pi(x)_i - \Pi(y)_i)^2}.$$

Distance between two sequences

(encθ = encoder RNN)

Corpus-dependent distance

$$\boxed{D_{\mathcal{D}}(s, s'; l) = \rho_l(\mathrm{enc}_\theta(s), \mathrm{enc}_\theta(s'))}$$

$$\mathcal{L}_{\text{invariant}}(\theta; \mathcal{D}, \mathcal{T}, \lambda) := \sum_{t,t' \in \mathcal{T}_k} \sum_{k=1}^{K} \sum_{i=1}^{N} \left(\mathrm{H}\left[(t'.s^k)_i \,;\, \mathrm{dec}_\theta\left(\overline{\mathrm{enc}}_\theta\left(s^k, t.s^k\right), t' \right)_i \right] + \lambda \ell_1 \left[\mathrm{enc}_\theta(s^k) - \mathrm{enc}_\theta(t.s^k) \right] \right)$$

Deep rank-based transposition-invariant distances on musical sequences

https://arxiv.org/abs/1709.00740

Special issue of “Information geometry” (Springer)

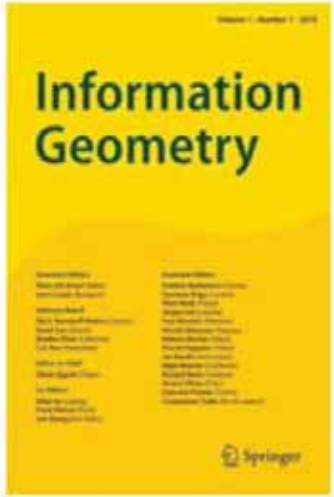
Call for papers for

"Information Geometry for Deep Learning"

- Submission Deadline: **1st April 2020**
- Submit manuscript to https://www.editorialmanager.com/inge/default.aspx

  select 'S.I.: Information Geometry for Deep Learning' in the submission

Deep neural networks (DNNs) are artificial neural network architectures with parameter spaces geometrically interpreted as neuromanifolds (with singularities) and learning algorithms visualized as trajectories/flows on neuromanifolds.



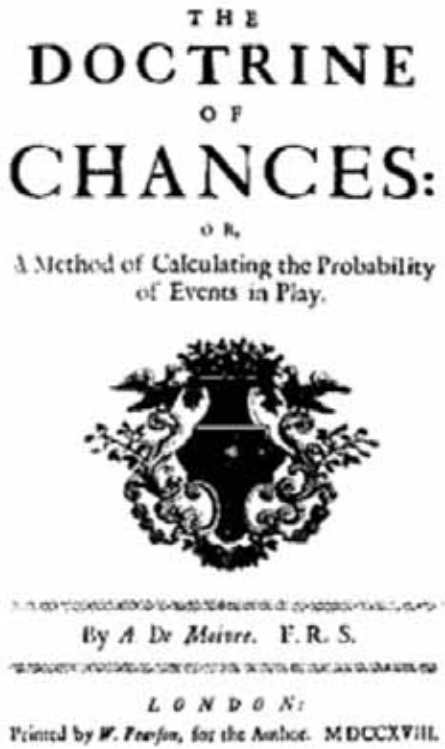
The aim of this special issue is to comprise original theoretical/experimental research articles which address the recent developments and research efforts on information-geometric methods in deep learning.

https://www.springer.com/journal/41884

Did Bayes learnt probability from de Moivre's book?

de Moivre linked complex numbers with trigonometry:

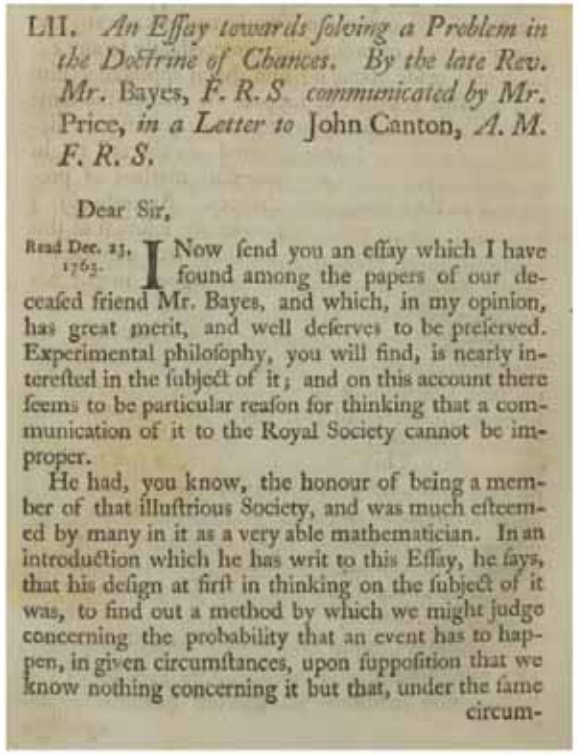
$$(\cos x + i \sin x)^n = \cos(nx) + i \sin(nx)$$



THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability of Events in Play.

By *A De Moivre*. F. R. S.

LONDON:
Printed by *W. Pearson*, for the Author. MDCCXVIII.



LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many in it as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circum-

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Abraham de Moivre (1667-1754)

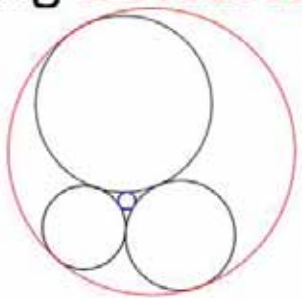
Thomas Bayes (1701-1761)

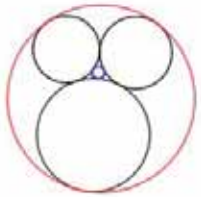
Descartes's theorem (1643) as a poem in Nature (1936)

- Consider three *mutually touching circles* C1, C2 and C3 (=kissing circles)
- ***Inner*** *and* ***outer*** *tangential circles* to (C1,C2,C3) using **Descartes's theorem**:

$\kappa_1 = \frac{1}{r_1}$

$$\begin{aligned} \kappa_4 &= \kappa_1 + \kappa_2 + \kappa_3 + 2\sqrt{\kappa_1\kappa_2 + \kappa_2\kappa_3 + \kappa_3\kappa_1}, \\ \kappa_5 &= \kappa_1 + \kappa_2 + \kappa_3 - 2\sqrt{\kappa_1\kappa_2 + \kappa_2\kappa_3 + \kappa_3\kappa_1}. \end{aligned}$$





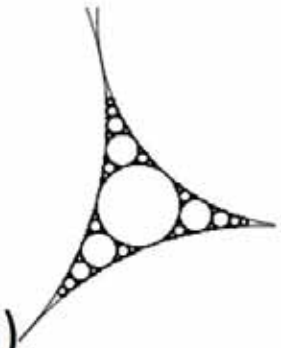
- Build the circle centers using the **complex Descartes' theorem**

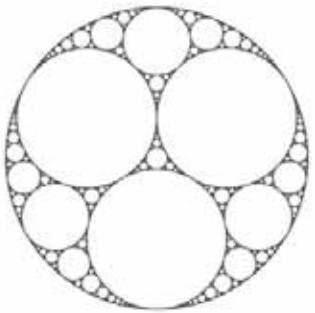
$z_i = x_i + iy_i$

$$a = \frac{z_1\kappa_1 + z_2\kappa_2 + z_3\kappa_3 + 2\sqrt{\kappa_1\kappa_2 z_1 z_2 + \kappa_2\kappa_3 z_2 z_3 + \kappa_1\kappa_3 z_1 z_3}}{\kappa_4},$$

$$b = \frac{z_1\kappa_1 + z_2\kappa_2 + z_3\kappa_3 - 2\sqrt{\kappa_1\kappa_2 z_1 z_2 + \kappa_2\kappa_3 z_2 z_3 + \kappa_1\kappa_3 z_1 z_3}}{\kappa_4}$$

- Apply recursively to build the **Apollonius gasket** (fractal)
- Descartes' theorem published as a **poem** in Nature (1936)



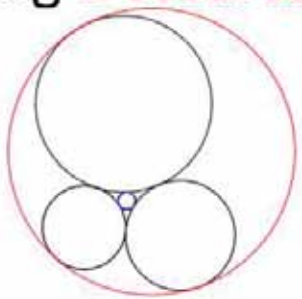


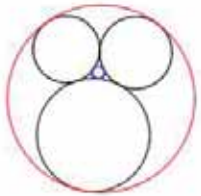
Descartes's theorem (1643) as a poem in Nature (1936)

- Consider three *mutually touching circles* C1, C2 and C3 (=kissing circles)
- ***Inner*** *and* ***outer*** *tangential circles* to (C1,C2,C3) using **Descartes's theorem**:

$\kappa_1 = \frac{1}{r_1}$

$$\begin{aligned} \kappa_4 &= \kappa_1 + \kappa_2 + \kappa_3 + 2\sqrt{\kappa_1\kappa_2 + \kappa_2\kappa_3 + \kappa_3\kappa_1}, \\ \kappa_5 &= \kappa_1 + \kappa_2 + \kappa_3 - 2\sqrt{\kappa_1\kappa_2 + \kappa_2\kappa_3 + \kappa_3\kappa_1}. \end{aligned}$$





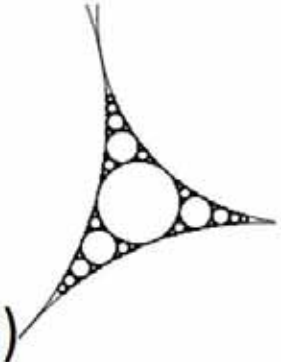
- Build the circle centers using the **complex Descartes' theorem**

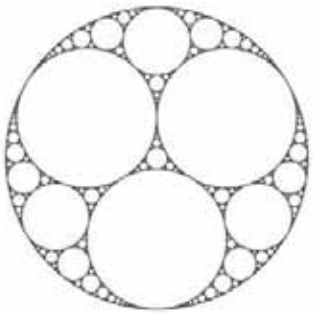
$z_i = x_i + iy_i$

$$a = \frac{z_1\kappa_1 + z_2\kappa_2 + z_3\kappa_3 + 2\sqrt{\kappa_1\kappa_2 z_1 z_2 + \kappa_2\kappa_3 z_2 z_3 + \kappa_1\kappa_3 z_1 z_3}}{\kappa_4},$$

$$b = \frac{z_1\kappa_1 + z_2\kappa_2 + z_3\kappa_3 - 2\sqrt{\kappa_1\kappa_2 z_1 z_2 + \kappa_2\kappa_3 z_2 z_3 + \kappa_1\kappa_3 z_1 z_3}}{\kappa_4}$$

- Apply recursively to build the **Apollonius gasket** (fractal)
- Descartes' theorem published as a **poem** in Nature (1936)





Natural gradient on a Hessian manifold = ordinary gradient on dually parameterized function

- **Hessian manifold** = Riemannian manifold (M,g) with a Hessian metric:

$$g_\theta(\theta) = \nabla^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$$

- Two coordinate systems related by Legendre-Fenchel transformation:

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

- **Natural gradient = ordinary gradient on dually parameterized function**:

$$L_\theta(\theta) = L_\theta(\theta(\eta)) \qquad L_\eta(\eta) = L_\eta(\eta(\theta))$$

$$\boxed{{}^{\mathrm{NG}}\nabla_\theta L_\theta = \nabla_\eta L_\eta}$$

Chain rule of differentiation:

$$\nabla_\theta L_\theta(\theta) = \nabla_\theta(L_\eta(\eta(\theta))) = (\nabla_\theta \eta)(\nabla_\eta L_\eta(\eta)).$$

Proof:

$$\begin{aligned} {}^{\mathrm{NG}}\nabla L_\theta(\theta) &:= g_\theta^{-1}(\theta)\nabla_\theta L_\theta(\theta), \\ &= (\nabla_\theta \eta)^{-1}(\nabla_\theta \eta)\nabla_\eta L_\eta(\eta), \\ &= \nabla_\eta L_\eta(\eta). \end{aligned}$$

[BM 2019] On geodesic triangles with right angles in a dually flat space, arxiv:1910.03935
A note on the natural gradient and its connections with the Riemannian gradient, the mirror descent, and the ordinary gradient

Basics of differential forms ω

- **k-differential forms** are smooth sections of the bundle of alternating k-tensors:

$\forall v_1, \ldots, v_k \in T_pM, \quad \omega(v_{\sigma(1)}, \ldots, v_{\sigma(k)}) = (\sigma)\,\omega(v_1, \ldots, v_k)$ $\qquad \Lambda^k T^*M$ of dimension $\binom{n}{k}$

for which there exists a differentiation operator d which is *different from the connection-based covariant derivative*. Space $\Omega^k(M)$ of k-forms

Examples: 1-differential forms are covector fields, 0-dimensional forms are smooth functions

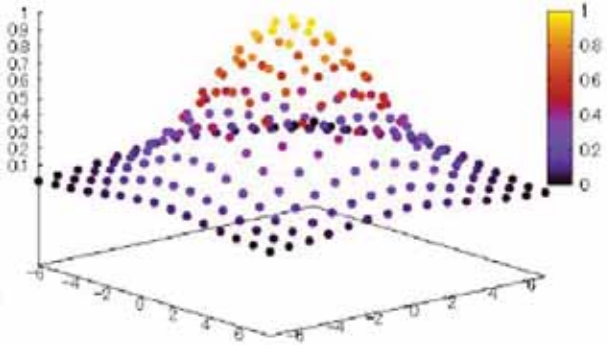
- Natural projection $P : \otimes^k T_p^* M \to \Lambda^k T_p^* M$: $\quad PT(v_1, \ldots, v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} T(v_{\sigma(1)}, \ldots, v_{\sigma(k)})$

(space of alternating tensors interpreted as signed determinant)

- **Exterior derivative** d (linear operator): $\forall \omega_1, \omega_2 \in \Omega^k(M), d(\omega_1 \wedge \omega_2) = (d\omega_1) \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2, \quad d^2 = 0$
- In R3, the exterior derivative of a 0-form is a 1-form equivalent to the *gradient vector field*, the exterior derivative of a 1-form is a 2-form equivalent to the *curl*, the exterior derivative of a 2-form is a 3-form equivalent to the *divergence*
- **Theorem: A smooth D-dimensional manifold M is orientable iff there exists a D-form which is nowhere vanishing on M.**

Discrete normal distributions & Riemann theta function

Maximum entropy distribution on the integer lattice support with prescribed mean and covariance:



$$\boxed{p_\xi(l) = \frac{\exp\left(2\pi\left(-\frac{1}{2}l^\top \xi_2 l + l^\top \xi_1\right)\right)}{\theta(-i\xi_1, i\xi_2)} =: \frac{\tilde{p}_\xi(l)}{\theta(-i\xi_1, i\xi_2)}, \quad l \in \mathbb{Z}^d}$$

Normalizer related to Riemann holomorphic theta function:

$\theta \ : \ \mathbb{C}^d \times \mathcal{H}_d \to \mathbb{C}$ where $\mathcal{H}_d = \{R \in M(d, \mathbb{C}) \ : \ R = R^\top, \mathrm{Im}(R) \in \mathcal{P}_d\}$ is the Siegel upper space

$$\theta(z, \Omega) := \sum_{l \in \mathbb{Z}^d} \exp\left(2\pi i \left(\frac{1}{2} l^\top \Omega l + l^\top z\right)\right)$$

Approximately calculate the **Kullback-Leibler divergence** for small value of γ>0

$$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] \approx \frac{1}{\gamma(\gamma-1)} \log\left(\frac{(I_{\gamma,E_\xi}(\xi:\xi)\, I_{\gamma,E'_\xi}(\xi':\xi')^{\gamma-1}}{I_{\gamma,E_\xi \cup E_{\xi'}}(\xi:\xi')^\gamma}\right) \qquad I_{\gamma,R}(\xi:\xi') := \sum_{x \in R} \tilde{p}_\xi\, \tilde{p}_{\xi'}(x)^\gamma$$

On the Kullback−Leibler divergence between discrete normal distribution http://arxiv.org/abs/2109.14920

2-point distances and n-point diversity indices

- A **dissimilarity** D(p,q) measures the *separation* between two points p and q. It is a 2-point function.
- A **diversity index** D(p1,…,pn) measures the *variation* of a set of n points. A diversity index is a n-point function. Diversity indices generalize dissimilarities.
- Usually, the diversity index is calculated using a notion of centrality (i.e., centroid).

$$\boxed{\mathrm{Diversity}_D(p_1,\ldots,p_n) := \min_c \frac{1}{n}\sum_{i=1}^{n} D(p_i : c)}$$

- For example, the **Bregman information** is a diversity index calculated from the **Bregman centroid** (center of mass, *independent* of the Bregman generator) which generalizes the variance of a point set. It yields a Jensen-Bregman diversity index:

$$\mathrm{Diversity}_{B_F}(\theta_1,\ldots,\theta_n) = \frac{1}{n}\sum_{i=1}^{n} F(\theta_i) - F\left(\frac{1}{n}\sum_{i=1}^{n}\theta_i\right) = J_F(\theta_1,\ldots,\theta_n)$$

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory 55.6 (2009)

2-point distances and n-point diversity indices

- A **dissimilarity** D(p,q) measures the *separation* between two points p and q. It is a 2-point function.
- A **diversity index** D(p1,…,pn) measures the *variation* of a set of n points. A diversity index is a n-point function. Diversity indices generalize dissimilarities.
- Usually, the diversity index is calculated using a notion of centrality (i.e., centroid).

$$\boxed{\mathrm{Diversity}_D(p_1,\ldots,p_n) := \min_c \frac{1}{n}\sum_{i=1}^{n} D(p_i : c)}$$

- For example, the **Bregman information** is a diversity index calculated from the **Bregman centroid** (center of mass, *independent* of the Bregman generator) which generalizes the variance of a point set. It yields a Jensen-Bregman diversity index:

$$\mathrm{Diversity}_{B_F}(\theta_1,\ldots,\theta_n) = \frac{1}{n}\sum_{i=1}^{n} F(\theta_i) - F\left(\frac{1}{n}\sum_{i=1}^{n}\theta_i\right) = J_F(\theta_1,\ldots,\theta_n)$$

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory 55.6 (2009)

Divergence geometry (Eguchi's method, 1983)

- In information theory, a **divergence** (eg., Kullback-Leiber divergence) means a **dissimilarity measure between probability distributions**, potentially oriented which may fail the triangle inequality of metric distances
- In information geometry, a **divergence** (formerly called contrast function) means a **dissimilarity measure between parameters** such that

$$\begin{aligned} D(\theta:\theta') &\geq 0 && \forall \theta,\theta' \text{ with equality iff } \theta=\theta' \\ \partial_i\, D(\theta:\theta')|_{\theta'=\theta} = \partial'_j\, D(\theta:\theta')|_{\theta'=\theta} &= 0 && \forall i,j \\ -\left[\partial_i \partial'_j\, D(\theta:\theta')|_{\theta'=\theta}\right]_{ij} &\succeq 0 && \text{positive-definite matrix} \end{aligned}$$

where $\partial_l := \frac{\partial}{\partial \theta_l}$ and $\partial'_l := \frac{\partial}{\partial \theta'_l}$

- Eguchi proposed a method to build a **dual statistical structure** (M, g, ∇, ∇^*) from any (parameter) divergence as follows:

$$\boxed{\begin{aligned} g_{ij}(\theta) &= -\partial_i \partial'_j\, D(\theta:\theta')|_{\theta'=\theta}\,, \\ \Gamma_{ij,k}(\theta) &= -\partial_i \partial_j \partial'_k\, D(\theta:\theta')|_{\theta'=\theta}\,, \\ \Gamma^*_{ij,k}(\theta) &= -\partial_k \partial'_i \partial'_j\, D(\theta:\theta')|_{\theta'=\theta}\,. \end{aligned}}$$

- The **Bregman divergence geometry** is **dually flat**
- The **f-divergence geometry** is the **α-geometry** for $\alpha = 3 + 2\frac{f'''(1)}{f''(1)}$
- The **half-squared Rao divergence geometry** is the **0-geometry** for $D_{\mathrm{Rao}}(\theta_1:\theta_2) := \frac{1}{2}\rho^2_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})$

Scale-invariant, projective and sided-projective divergences

- A smooth **statistical dissimilarity** is called a **divergence**:

  For example, the Kullback-Leibler divergence: $\mathrm{KL}[p:q] = \int p \log \frac{p}{q} \mathrm{d}\mu$

- A **scale-invariant divergence** is such that $D[\lambda p : \lambda q] = D[p:q], \forall \lambda > 0$

  For example, the Itakura-Saito divergence: $D_{\mathrm{IS}}[p:q] = \int (\frac{p}{q} - \log \frac{p}{q} - 1) \mathrm{d}\mu$

  (a Bregman divergence)

- A **projective divergence** is such that $D[\lambda p : \lambda' q] = D[p:q], \forall \lambda > 0, \lambda' > 0$

  The γ –divergence: $D_\gamma[p:q] = \frac{1}{\gamma(\gamma+1)} \log \int p^{1+\gamma} \mathrm{d}\mu - \frac{1}{\gamma} \log \int p q^\gamma \mathrm{d}\mu + \frac{1}{\gamma+1} \log \int q^{1+\gamma} \mathrm{d}\mu$

  $\lim_{\gamma \to 0} D_\gamma[p:q] = \mathrm{KL}[\frac{p}{\int p \mathrm{d}\mu} : \frac{q}{\int q \mathrm{d}\mu}]$

- A **sided-projective divergence** is such that: $D[p : \lambda q] = D[p:q], \forall \lambda > 0$

  For example, the Hyvärinen divergence: $D_H[p:q] = \frac{1}{2} \int p(x) \left\| \nabla_x \log p(x) - \nabla_x \log q(x) \right\|^2 \mathrm{d}\mu(x)$

Sided and symmetrized Bregman centroids, 2009

Patch matching with polynomial exponential families and projective divergences, 2016

The Mahalanobis distance (1936)

ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

2. It will be convenient at this stage to introduce the idea of a statistical field, such that at each point in this field there is a specified set of parameters (α_i, α_{ij}) which define a particular population.

It has been shown elsewhere* that the distance between the two statistical populations can be conveniently measured by a Δ^2-statistics defined by

$$P \cdot \Delta^2 = \mathop{S}_{i=1}^{i=P} \frac{(\alpha_i - \alpha'_i)^2}{\alpha_{ii}} \quad \ldots \quad \ldots \quad \ldots \quad (2{\cdot}1)$$

The generalized Δ^2=statistics for P correlated variates can be now written in the form

$$P \cdot \Delta^2 = \mathop{S}_{i,j=1}^{i,j=P} [\alpha^{ij} \cdot (\alpha_i - \alpha'_i) \cdot (\alpha_j - \alpha'_j)] \quad \ldots \quad \ldots \quad (2{\cdot}4)$$

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

Comparing with the formula for ds^2

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

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

Statistical field
(=Population space)

Prasanta C. Mahalanobis
(1893-1972)

Diagonal covariance matrix
(uncorrelated case)

$$\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$$

$$D_\Sigma^{\mathrm{Mah}}(\mu_1, \mu_2) = \sum_{i=1}^{D} \frac{(\mu_1^i - \mu_2^i)^2}{\sigma_i^2}$$

Generic covariance/precision matrix

$$P = \Sigma^{-1}$$

$$D_\Sigma^{\mathrm{Mah}}(\mu_1, \mu_2) = \sqrt{(\mu_2 - \mu_1)^\top \Sigma^{-1} (\mu_2 - \mu_1)}.$$

$$D_\Sigma^{\mathrm{Mah}}(\mu_1, \mu_2) = \sqrt{\sum_{i,j} P_{ij} (\mu_1^i - \mu_2^i)(\mu_1^j - \mu_2^j)}$$

Similar expression of a metric g
in Riemannian geometry
(in general relativity)

Mahalanobis, On the generalized distance in statistics, National Institute of Science of India, 1936.

Mirror descent on Bregman manifold = natural gradient on dual Hessian manifold

- **Mirror descent** extends the ordinary descent by minimizing with respect to a proximity function: $\theta_{t+1} \leftarrow \arg\min_{\theta}\{\theta^\top \nabla L(\theta_t) + \frac{1}{\alpha_t}\Phi(\theta, \theta_t)\}$

Recover ordinary gradient when $\Phi(\theta, \theta_t) = \frac{1}{2}\|\theta - \theta_t\|^2$

- Consider a **Hessian structure** on a Riemannian manifold (M,g): g is the hessian metric of a strictly convex potential function (=*Bregman manifold* **[BM 2019]**)
- Mirror descent with respect to a Bregman divergence:

$$\theta(\eta) = \nabla F^*(\eta) \quad \eta(\theta) = \nabla F(\theta) \qquad \theta_{t+1} \leftarrow \arg\min_{\theta}\{\theta^\top \nabla L(\theta_t) + \frac{1}{\alpha_t} B_F(\theta : \theta_t)\}$$

- Mirror descent on Bregman manifold (M,g) amounts to **natural gradient** on the **dual Hessian manifold** (M,g*): $\eta_{t+1} \leftarrow \eta_t - \alpha_t (g^*)^{-1}(\theta_t)\nabla_\eta L(\theta(\eta_t))$

[BM 2019] On geodesic triangles with right angles in a dually flat space, arxiv:1910.03935

[MDIG] Raskutti and Mukherjee, The information geometry of mirror descent, IEEE Information Theory (2015)

Dual numbers

- A **dual number** is expressed as **a+εb** where ε is **nilpotent** (ε ε=0)
- Introduced independently by Clifford and Study
- The basic addition/multiplication operations are:

$$(a_1 + \varepsilon b_1) + (a_2 + \varepsilon b_2) = (a_1 + a_2) + \varepsilon(b_1 + b_2)$$
$$(a_1 + \varepsilon b_1) \cdot (a_2 + \varepsilon b_2) = (a_1 a_2) + \varepsilon(a_1 b_2 + a_2 b_1)$$

- Binomial expansion of $(a + \varepsilon b)^k$ is $a^k + \varepsilon k a^{k-1} b$
- Taylor expansions of an analytic function with dual number argument:

$$\boxed{f(a + b\varepsilon) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a) b^n \varepsilon^n}{n!} = f(a) + b f'(a)\varepsilon}$$

- Thus we have for example: $\sin(x + \varepsilon y) = \sin(x) + \varepsilon y \cos(x)$
- Generalized to higher dimensions by Grassmann numbers
- Recently used in automatic differentiation and machine learning

William Clifford
(1845–1879)

C. H. Eduard Study
(1862-1930)

Automatic Differentiation (AD) with Dual numbers: $\mathbb{D}$-AD

- A **dual number** is **a+εb** where ε is **nilpotent** ($\epsilon^2 = 0$) : $\mathbb{D} = \{a + b\epsilon \ : \ (a, b) \in \mathbb{R}^2\}$
- The basic algebra addition/multiplication operations of dual numbers are:

$$(a_1 + \varepsilon b_1) + (a_2 + \varepsilon b_2) = (a_1 + a_2) + \varepsilon(b_1 + b_2)$$
$$(a_1 + \varepsilon b_1) \cdot (a_2 + \varepsilon b_2) = (a_1 a_2) + \varepsilon(a_1 b_2 + a_2 b_1)$$

- Taylor expansion of an analytic function f yields: $f(a + b\epsilon) = f(a) + bf'(a)\epsilon$
- **Recover the Leibniz's rules of differentiation**:

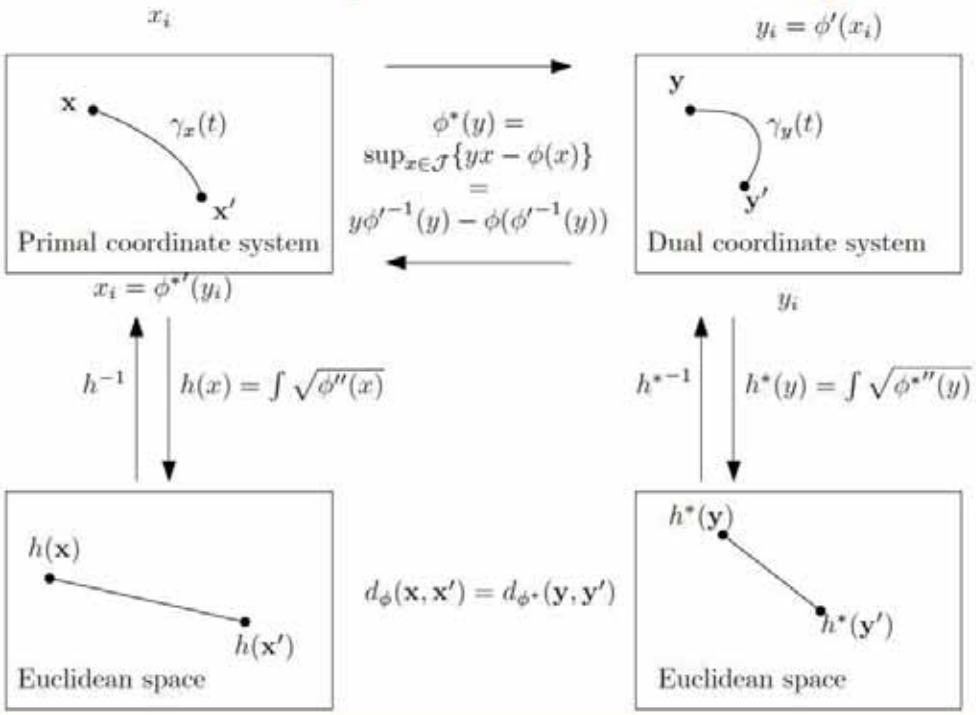
$$(f + g)(a + \epsilon) = (f(a) + g(a)) + \underbrace{(f'(a) + g'(a))}_{(f+g)'(a)}\epsilon$$

$$(fg)(a + \epsilon) = (f(a)g(a)) + \underbrace{(f'(a)g(a) + f(a)g'(a))}_{(fg)'(a)}\epsilon$$

- AD with *dual numbers (exact)* better than AD with *complex numbers* ($\mathbb{C}$-AD) :

$$f(a + ib) = f(a) + if'(a)b + O\left(b^2\right)$$ ← yields truncation errors

Dual Riemann geodesic distances induced by a separable Bregman divergence



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

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

Bregman divergence:

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

Separable Bregman generator:

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

Riemannian metric tensor:

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

Geodesics:

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

Riemannian distance (metric):

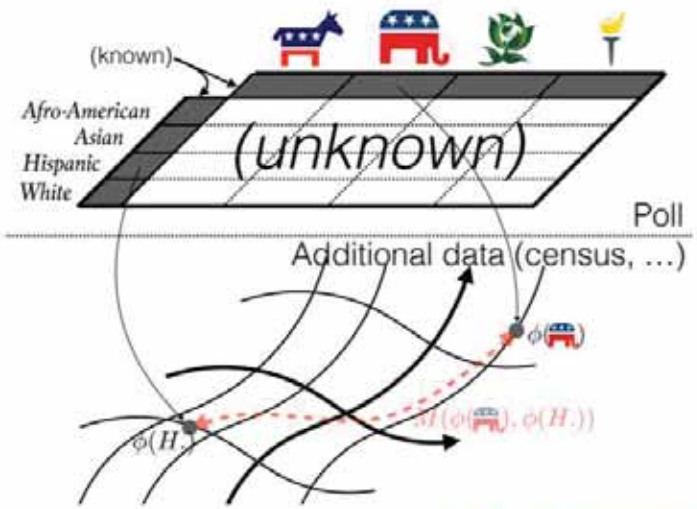
$$\rho_\phi(x,x') = \sqrt{\sum_{j=1}^K \left(h(x_j) - h(x_j')\right)^2}$$

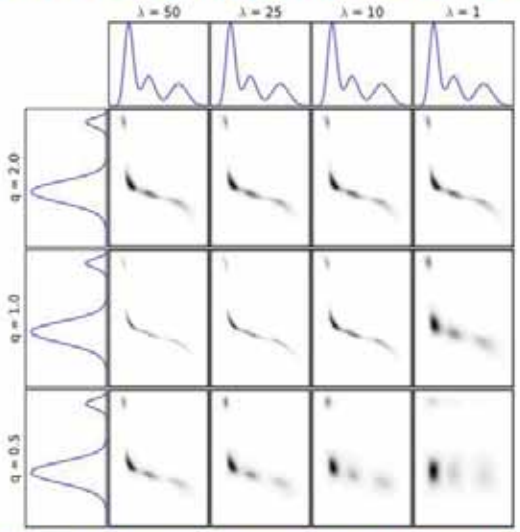
where $h(x) := \int \sqrt{\phi''(x)}$

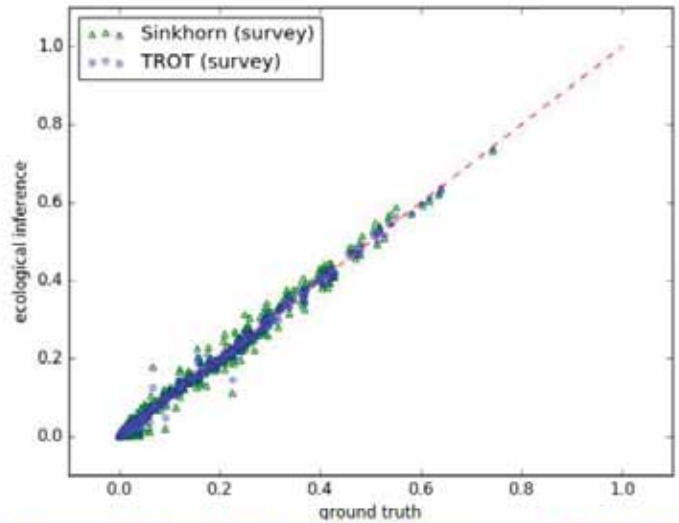
Geometry and clustering with metrics derived from separable Bregman divergences, arXiv:1810.10770

Ecological inference: Reconstructing joint distributions

- Use **macroscopic aggregates** (=``ecological'') to infer **microscopic individual** information. Roots in socials sciences (elections, polls), epidemiology, etc.
- Standard techniques: combine deterministic bounds (e.g., knowing a ratio has to be in [0,1]) with statistical approaches (i.e., ecological regression)
- **New technique**: Introduce **Tsallis regularized optimal transport** (TROT) for **reconstructing joint distributions from their marginal** (transportation plan)







Muzellec et al, Tsallis Regularized Optimal Transport and Ecological Inference, AAAI 2017
On Rényi and Tsallis entropies and divergences for exponential families, arXiv:1105.3259

Exponential families: Bregman divergence and Jensen divergence

$$
\begin{aligned}
D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] &= B_F{}^*(\theta_1 : \theta_2), && \text{Kullback-Leibler divergence} \\
D_J[p_{\theta_1} : p_{\theta_2}] &= (\theta_2 - \theta_1)^\top (\eta_2 - \eta_1), && \text{Jeffreys' divergence} \\
\rho[p_{\theta_1} : p_{\theta_2}] &= \exp(-J_F(\theta_1 : \theta_2)), && \text{Bhattacharyya coefficient} \\
D_B[p_{\theta_1} : p_{\theta_2}] &= J_F(\theta_1 : \theta_2) && \text{Bhattacharyya distance} \\
D_H[p_{\theta_1} : p_{\theta_2}] &= \sqrt{1 - \exp(-J_F(\theta_1 : \theta_2))}, && \text{Hellinger distance}
\end{aligned}
$$

Bregman divergence B_F and **Jensen divergence** J_F for a **strictly convex and smooth generator** F:

$$
\begin{aligned}
B_F(\theta_1 : \theta_2) &:= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \\
J_F(\theta_1 : \theta_2) &:= \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right).
\end{aligned}
$$

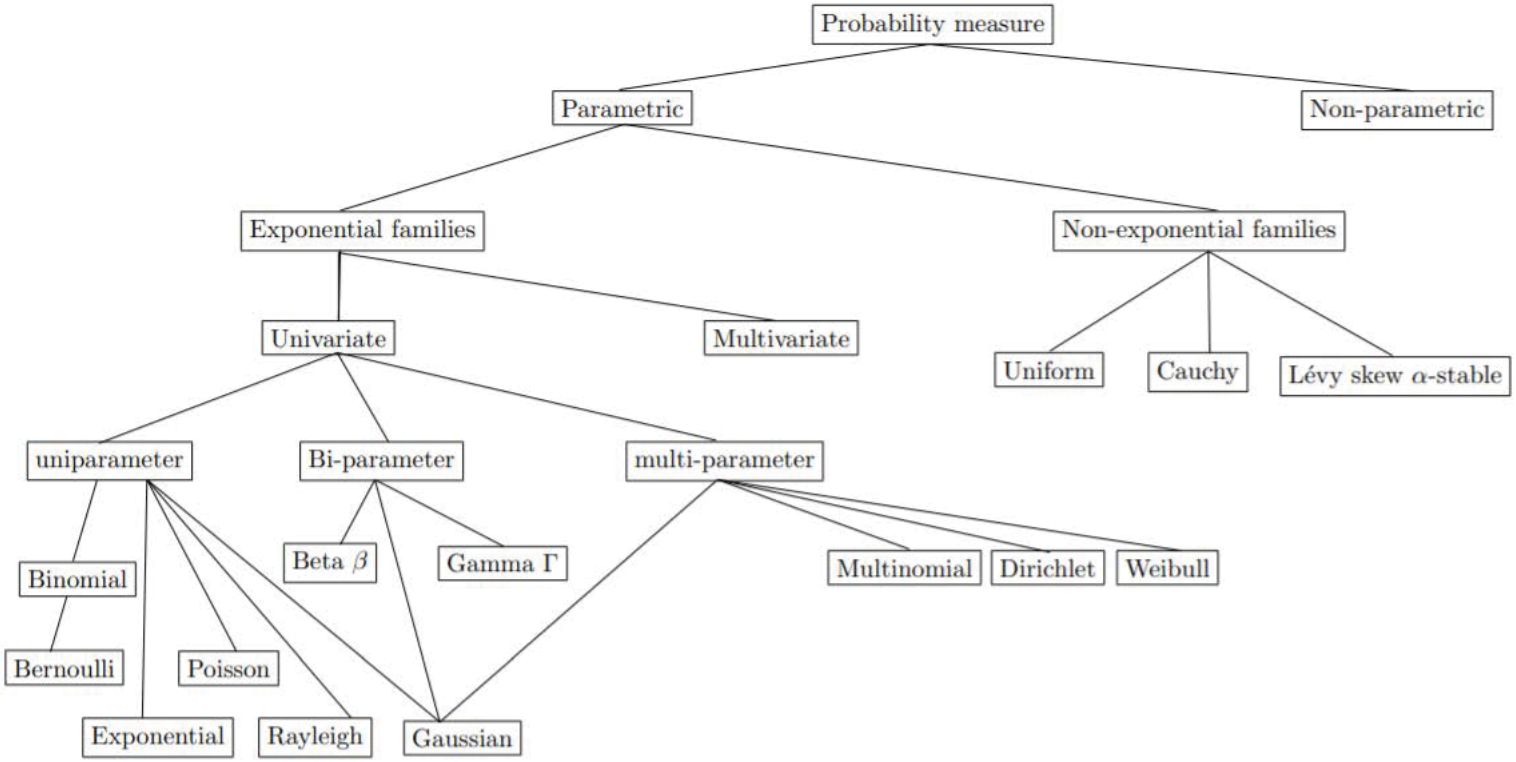


Table 1: Partial taxonomy of statistical distributions.

Open Access Article

An Elementary Introduction to Information Geometry

by Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan

Entropy **2020**, *22*(10), 1100; https://doi.org/10.3390/e22101100 (registering DOI)

Received: 6 September 2020 / Revised: 25 September 2020 / Accepted: 27 September 2020 / Published: 29 September 2020

(This article belongs to the Special Issue Review Papers for Entropy)

Download PDF Cite This Paper

Abstract

In this survey, we describe the fundamental differential-geometric structures of information

Deflation method: Eigenpairs of a Hermitian matrix

Hermitian matrix M: matrix which equals its conjugate transpose: M=M* (complex generalization of symmetric matrices). Hermitian matrices have real diagonal elements, are diagonalizable, and have all real eigenvalues.

Deflation method : numerically calculate the eigenvalues and *normalized eigenvectors* of Hermitian matrix: $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_d|$

1. Let $l = 1$ and $M_1 = M$.
2. Initialize at random a normalized vector $x_0 \in \mathbb{C}^d$ (i.e., x_0 on the unit sphere with $x_0^* x_0 = 1$)
3. For j in $(0, \ldots, L_l - 1)$:

$$\begin{aligned} x_{j+1} &\leftarrow M_l x_j \\ x_{j+1} &\leftarrow \frac{x_{j+1}}{x_{j+1}^* x_{j+1}} \end{aligned}$$

4. Let $v_l = x_{L_l}$ and $\lambda_l = x_{L_l}^* M_l x_{L_l}$
5. Let $l \leftarrow l + 1$. If $l \leq d$ then let $M_l = M_{l-1} - \lambda_{l-1} v_{l-1} v_{l-1}^*$ and goto 2.

Hilbert geometry of the Siegel disk: The Siegel-Klein disk model, arXiv 2004.08160

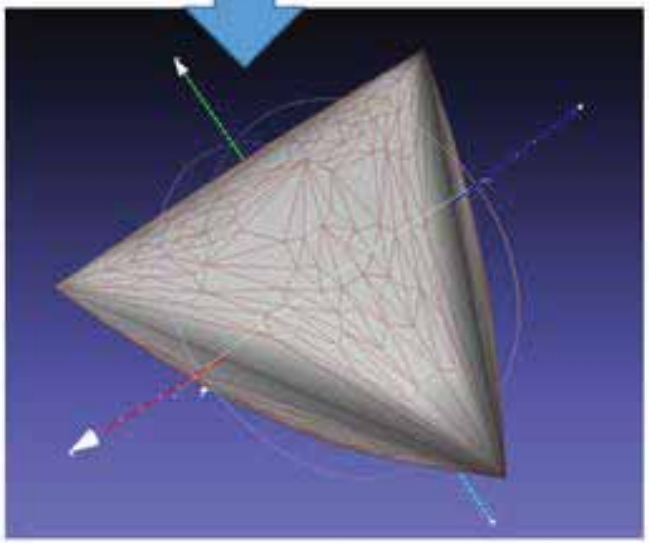
相関行列の空間 ＝ Elliptope　エリプトープ

Space of correlation matrices

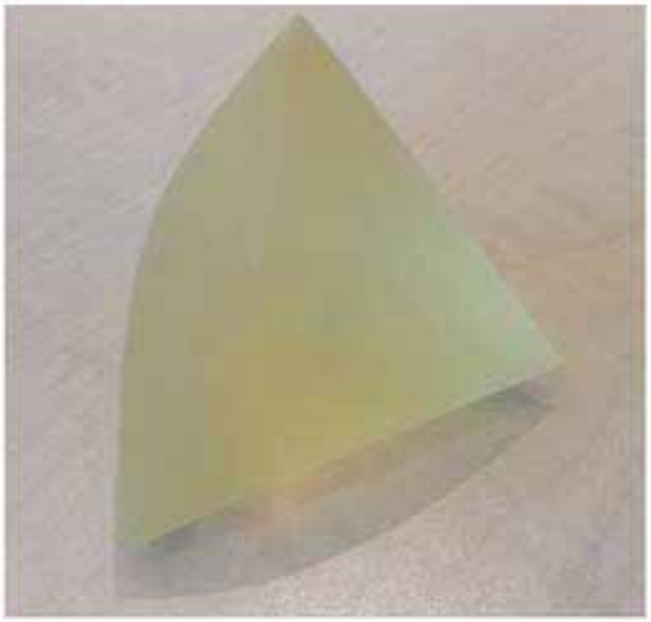
$$\mathbf{C} = \begin{pmatrix} 1 & r_{XY} & r_{XZ} \\ r_{XY} & 1 & r_{YZ} \\ r_{XZ} & r_{YZ} & 1 \end{pmatrix}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

A point represents a matrix



Inside: Full rank 3
Facet: Rank 2



$$\mathcal{E}_3 = \left\{ (x, y, z) \in \mathbb{R}^3 : \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \right\}$$

$$x^\top A x \geq 0$$

Optimization with odd/even loss decomposition

ML Trick

- Any function f(x) on the real can be decomposed (or factorized) uniquely into the sum of an **odd function** g(-x)=-g(x) plus an **even function** h(-x)=h(x) :

f(x)=g(x)+h(x) with g(x)=(f(x)-f(-x))/2=-g(x) and h(x)=(f(x)+f(-x))/2=h(-x)

- Odd/even loss decomposition useful in **supervised learning** with labels in {-1,+1} for **empirical risk optimization with linear odd part**:

Theorem (Factorization) *Let $\mathcal{H}$ be the space of linear hypotheses. Assume that a loss l is such that $l_o(x) \doteq (l(x) - l(-x))/2$ is linear. Then, for any sample $\mathcal{S}$ and hypothesis $h \in \mathcal{H}$ the empirical l-risk can be written as*

$$R_{\mathcal{S},l}(h) = \frac{1}{2} R_{\mathcal{S}_{2x},l}(h) + l_o(h(\boldsymbol{\mu}_{\mathcal{S}})) \ ,$$

where $\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \forall \sigma \in \mathcal{Y}\}$.

	loss l	odd term l_o
LOL	$l(x)$	$-ax$
ρ-loss	$\rho\lvert x\rvert - \rho x + 1$	$-\rho x \ (\rho \geq 0)$
unhinged	$1 - x$	$-x$
perceptron	$\max(0, -x)$	$-x$
2-hinge	$\max(-x, 1/2 \max(0, 1 - x))$	$-x$
SPL	$a_l + l^*(-x)/b_l$	$-x/(2b_l)$
logistic	$\log(1 + e^{-x})$	$-x/2$
square	$(1 - x)^2$	$-2x$
Matsushita	$\sqrt{1 + x^2} - x$	$-x$

Patrini et al. "Loss factorization, weakly supervised learning and label noise robustness." *International conference on machine learning* (ICML), 2016.

Entropy for discrete/continuous exponential families

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

Using natural parameter θ:

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

Using expectation parameter:

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

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

Consider yet another univariate exponential family: the Poisson distribution with probability mass function $p(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$. The entropy is $\lambda(1 - \log\lambda) - E[k(x)]$ Since $k(x) = -\log x!$ (see [4]), we have:
$-E[k(x)] = \sum_{k=0}^{\infty} p_F(x;\lambda)\log k! = e^{-\lambda}\sum\frac{\lambda^k \log k!}{k!}$.

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

When is the entropy of a mixture in closed-form?

- Shannon's (discrete/differential) **entropy**: $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$
- Density in a mixture family: $m(\theta) = \sum_{i=1}^{D} \theta_i p_i(x) + \left(1 - \sum_{i=1}^{D} \theta_i\right) p_0(x)$ $\quad \theta \in \Delta_D$ $\quad \theta_0 := 1 - \sum_{i=1}^{D} \theta_i$
- The differential entropy of a mixture is in closed-form when the component distributions have **pairwise disjoint supports**: $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_D \quad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, \quad i \neq j$

$$m(\theta) = \sum_{i=1}^{D} \theta_i p_i(x) + \left(1 - \sum_{i=1}^{D} \theta_i\right) p_0(x) \Rightarrow m_\theta(x) = \theta_i p_i(x), \forall x \in \mathcal{X}_i$$

$$\mathcal{X} = \cup_i \mathcal{X}_i$$

$$\boxed{\begin{aligned} h(m_\theta) &= -\int_{\mathcal{X}} m_\theta(x) \log m_\theta(x) \\ &= -\sum_{i=0}^{D} \theta_i \int_{\mathcal{X}_i} p_i(x) \log(\theta_i p_i(x)) \mathrm{d}\mu(x) \\ &= -\sum_{i=0}^{D} \theta_i \log \theta_i + \sum_{i=0}^{D} \theta_i h(p_i) \end{aligned}}$$

Simple expression

- The differential entropy of a Gaussian mixture model is **NOT analytic**

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid, Entropy 22.2 (2020)

Open Access Article

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius

by Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan

Academic Editor: Yong Deng

Entropy **2021**, *23*(4), 464; https://doi.org/10.3390/e23040464 (registering DOI)

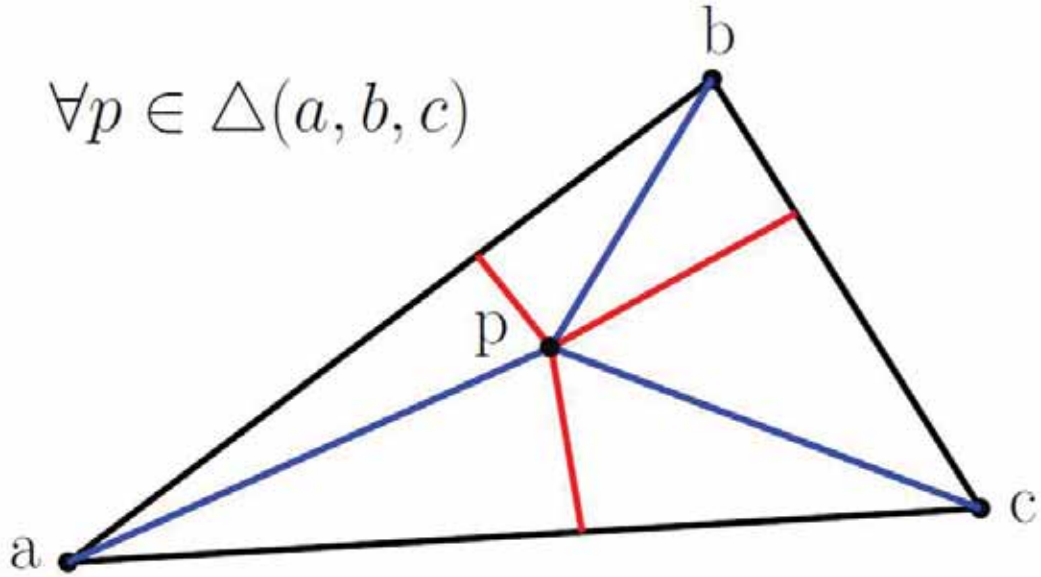
Received: 12 March 2021 / Revised: 9 April 2021 / Accepted: 9 April 2021 / Published: 14 April 2021

(This article belongs to the Special Issue **Selected Papers from the 5th conference on Geometric Science of Information**)

Download PDF Citation Export

An example of a geometric inequality: The Erdős–Mordell inequality

$\forall p \in \triangle(a,b,c)$



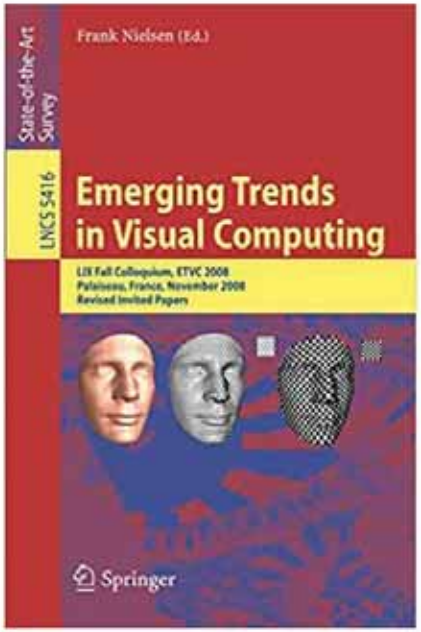
$$d(p,a) + d(p,b) + d(p,c) \geq 2\left(d(p,(ab)) + d(p,(bc)) + d(p,(ca))\right)$$

Equivality iff T is equilateral and P is its centroid

Paul Erdős
1913-1996

Louis J. Mordell
1888-1972

Alsina, Claudi, and Roger B. Nelsen. "A visual proof of the Erdös-Mordell inequality." Forum Geom. Vol. 7. 2007.



Geometric Computing

Abstracts of the LIX Fall Colloquium 2008: Emerging Trends in Visual Computing 1
Frank Nielsen

From Segmented Images to Good Quality Meshes Using Delaunay Ren ement 13
Jean-Daniel Boissonnat, Jean-Philippe Pons, and Mariette Yvinec

Information Geometry and Applications

Discrete Curvature Flows for Surfaces and 3-Manifolds 38
Xiaotian Yin, Miao Jin, Feng Luo, and Xianfeng David Gu

Information Geometry and Its Applications: Convex Function and Dually Flat Manifold 75
Shun-ichi Amari

Computational Geometry from the Viewpoint of Affine Differential Geometry 103
Hiroshi Matsuzoe

Interactions between Symmetric Cone and Information Geometries: Bruhat-Tits and Siegel Spaces Models for High Resolution Autoregressive Doppler Imagery 124
Frederic Barbaresco

Clustering Multivariate Normal Distributions 164
Frank Nielsen and Richard Nock

Computer Graphics and Vision

Intrinsic Geometries in Learning 175
Richard Nock and Frank Nielsen

Shape from Depth Discontinuities 216
Gabriel Taubin, Daniel Crispell, Douglas Lanman, Peter Sibley, and Yong Zhao

Computational Photography: Epsilon to Coded Photography 238
Ramesh Raskar

Unifying Subspace and Distance Metric Learning with Bhattacharyya Coefficient for Image Classification 254
Qingshan Liu and Dimitris N. Metaxas

Information Retrieval

Constant-Working-Space Algorithms for Image Processing 268
Tetsuo Asano

Sparse Multiscale Patches for Image Processing 284
Paolo Piro, Sandrine Anthoine, Eric Debreuve, and Michel Barlaud

Medical Imaging and Computational Anatomy

Recent Advances in Large Scale Image Search 305
Herve Jegou, Matthijs Douze, and Cordelia Schmid

Information Theoretic Methods for Diffusion-Weighted MRI Analysis ... 327
Angelos Barmpoutis and Baba C. Vemuri

Statistical Computing on Manifolds: From Riemannian Geometry to Computational Anatomy 347
Xavier Pennec

Author Index 387

Etymology of the Jensen-Shannon divergence

- **Jensen's midpoint inequality** defining the convexity of a function:

$$\varphi(x) + \varphi(y) \geq 2\varphi\left(\frac{x+y}{2}\right),$$

on dit que $\varphi(x)$ est une fonction convexe dans cet intervalle.

SUR LES FONCTIONS CONVEXES ET LES INÉGALITÉS ENTRE LES VALEURS MOYENNES[1]

PAR

J. L. W. V. JENSEN

À COPENHAGUE.

Jensen, 1906

Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

because *midpoint+continuous → convex:*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \qquad \text{for all} \quad t \in \langle 0, 1 \rangle$$

A Mathematical Theory of Communication

By C. E. SHANNON

Shannon, 1948

- *Shannon entropy h is* ***concave*** *(→ -h is convex):*

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

- *Take the* ***inequality gap*** *of midpoint convexity:* $J_\phi[p:q] = \int \left(\frac{1}{2}(\phi(p(x)) + \phi(q(x))) - \phi\left(\frac{p(x)+q(x)}{2}\right)\right) \mathrm{d}\mu$

$$\boxed{D_{\mathrm{JS}}[p:q] = J_{-h}[p:q] = h\left(\frac{p+q}{2}\right) - \frac{1}{2}\left(h(p) + h(q)\right)}$$

Usual equivalent definition using KLD:

$$\boxed{\mathrm{JS}[p:q] = \frac{1}{2}\mathrm{KL}\left[p : \frac{p+q}{2}\right] + \frac{1}{2}\mathrm{KL}\left[q : \frac{p+q}{2}\right]}$$

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid, 2020, Entropy 22 (2), 221

On the Jensen–Shannon symmetrization of distances relying on abstract means, 2019, Entropy 21 (5), 485

The Euclidean Riemannian metric Jacobian decomposition

- In a Cartesian coordinate system x, the Euclidean metric is encoded by the identity matrix I:

$$^{E}g(e_i, e_j) = \delta_{ij}, \quad ^{E}g(x) = I$$

- In any new coordinate system λ' (eg., spherical, polar, etc.), a metric expressed in the coordinate system λ is rewritten as:

$$^{\lambda'}g(\lambda') = J^{\top}_{\lambda,\lambda'} \times {}^{\lambda}g(\lambda(\lambda')) \times J_{\lambda,\lambda'}$$

- Thus a Euclidean metric can be expressed as:

$$\boxed{^{E}g(\lambda') = J^{\top}_{x,\lambda'} \times J_{x,\lambda'}}$$

- When the metric can be decomposed as above, it is necessarily the Euclidean metric.

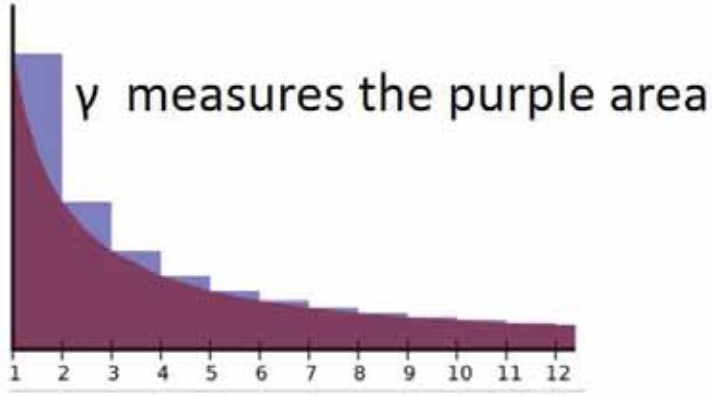
An elementary introduction to information geometry, 2018

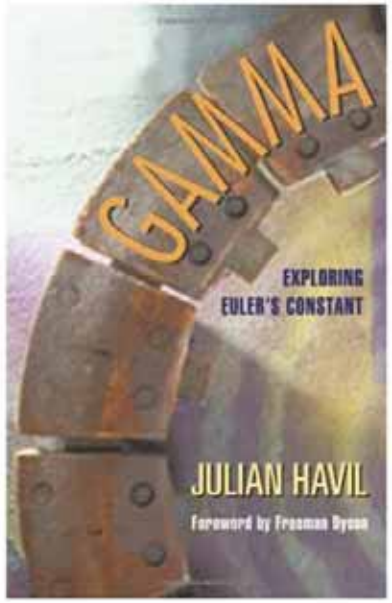
Is Euler–Mascheroni γ constant irrational or not?

n-th Harmonic number: $H_n := \sum_{i=1}^{n} \frac{1}{i}$

Euler–Mascheroni γ constant:

$$\boxed{\gamma := \lim H_n - \log n \simeq 0.577}$$





Link with Gamma function:

$$\gamma = -\Gamma'(1)$$

Modern definition: $\gamma = -\int_0^\infty e^{-x} \log x \, dx.$

Rayleigh distributions $p(x;\sigma^2) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ have entropy $1 + \ln \frac{\sigma}{\sqrt{2}} + \frac{\gamma}{2}$. [ICIP 2010]

Conjecture: Euler–Mascheroni γ constant is irrational

Lagarias, Jeffrey. "Euler's constant: Euler's work and modern developments." *Bulletin of AMS* 50.4 (2013

Some inequalities for expectation of means

- X,Y: positive random variables

$$
\begin{aligned}
A(x,y) &= \tfrac{x+y}{2} &\Rightarrow\ & E[A(X,Y)] = A(E[X],E[Y]) \\
G(x,y) &= \sqrt{xy} &\Rightarrow\ & E[G(X,Y)] \le G(E[X],E[Y]) \\
H(x,y) &= \tfrac{2}{x^{-1}+y^{-1}} &\Rightarrow\ & E[H(X,Y)] \le H(E[X],E[Y])
\end{aligned}
$$

Student (1996), Vol.1, No.3, 149-158

Seven Inequalities in Statistical Estimation Theory

C. Radhakrishna Rao

Department of Statistics, The Pennsylvania State University, United States

If $X > 0$ and $Y > 0$ are random variables, then

$$E\left(\frac{XY}{X+Y}\right) \le \frac{E(X)E(Y)}{E(X+Y)}.$$

Rao, Calyampudi Radhakrishna. *Seven inequalities in statistical estimation theory*. 1996.

Exponential mean, quasi-arithmetic mean and log-sum-exp

- **Exponential mean** is a quasi-arithmetic mean for generator f(u)=exp(u):

$$\begin{aligned}\overline{\mathrm{Exp}}(x_1,\ldots,x_n;w_1,\ldots,w_n) &:= \log\left(\sum_{i=1}^n w_i \exp(w_i)\right) \qquad w\in\Delta_n\\ &= M_{\exp}(x_1,\ldots,x_n;w_1,\ldots,w_n)\end{aligned}$$

- **Quasi-arithmetic mean** for strictly monotone function f(u) (hence inversible):

$$M_f(x_1,\ldots,x_n;w_1,\ldots,w_n) := f^{-1}\left(\sum_{i=1}^n w_i f(x_i)\right)$$

- **Log-sum-exp function** (LSE):

$$\mathrm{LSE}(a_1,\ldots,a_n) := \log\left(\sum_{i=1}^n \exp(a_i)\right)$$

- LSE is convex but not strictly convex:

$$\boxed{\overline{\mathrm{Exp}}(x_1,\ldots,x_n;w_1,\ldots,w_n) = \mathrm{LSE}(x_1+\log w_1,\ldots,x_n+\log w_n)}$$

- Quasi-arithmetic means (incl. LSE) yields associative operator: MapReduce

Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities, Frank Nielsen and Ke Sun, Entropy 2016

f-divergences as weighted integrals of scalar Bregman divergences

f-divergence: $I_f(p:q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$

Bregman divergence: $B_f(a:b) = f(a) - f(b) - (a-b) f'(b)$

Extended f-divergence to positive densities:

$$\boxed{I_f^e(p:q) = \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) \mathrm{d}\mu(x) \geq 0}$$

$$I_f^e(p:q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right) \mathrm{d}\mu(x)$$

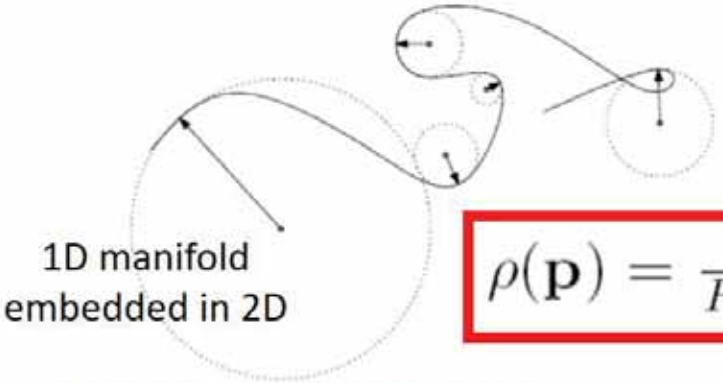
Standard f-divergence: f convex, strictly convex at 1, with f(1)=f'(1)=0 and f''(1)=1

Non-negative Monte Carlo estimation of f-divergences, 2020

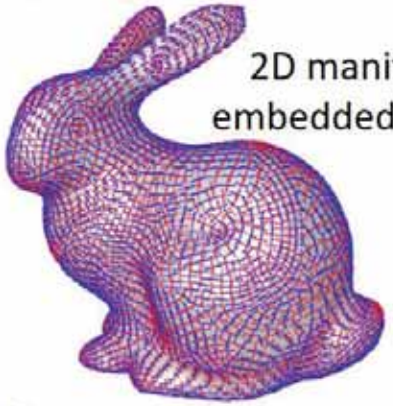
On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 2013

Extrinsic curvatures versus intrinsic curvatures

- **Extrinsic curvature** measured on the *embedded manifold* in higher Euclidean space: For 1D curves, curvature is the inverse of the radius of the *osculating circle*. **Min**/**Max** sectional curvatures are perpendicular to each other



$$\rho(\mathbf{p}) = \frac{1}{R(\mathbf{p})}$$



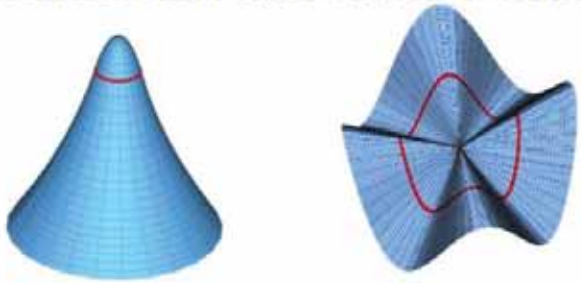
Gaussian curvature:

$$\rho_G = \rho_{\max} \times \rho_{\min}$$

Mean curvature:

$$\frac{\rho_{\max} + \rho_{\min}}{2}$$

- **Intrinsic curvature** measured from the **Riemann curvature (1,3)-tensor** (Ricci curvature, Ricci curvature tensor). Do not need ambient space (no embedding). Gaussian curvature measured from the circumference C(r) of a circle:



$$\rho_G(p) = \lim_{r \to 0} \frac{6}{r^2} \left(1 - \frac{C_p(r)}{2\pi r} \right)$$

https://images.math.cnrs.fr/Visualiser-la-courbure.html?lang=fr

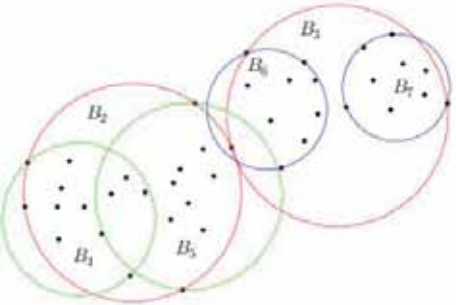
An elementary introduction to information geometry, *arXiv:1808.08271*

Fast Proximity queries for Bregman divergences (incl. KL)

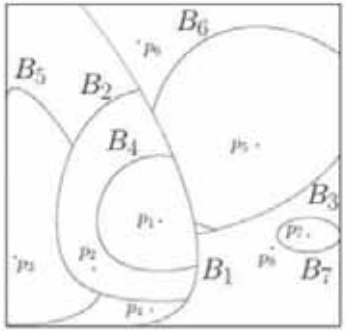
Fast <u>Nearest Neighbour Queries</u> for Bregman divergences

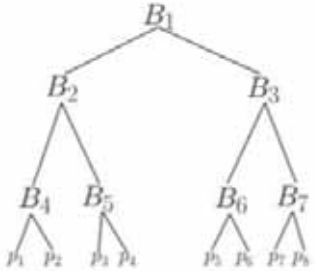
Space partition induced by

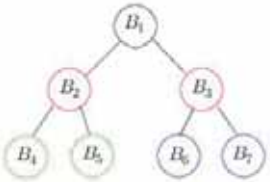
Bregman vantage point trees



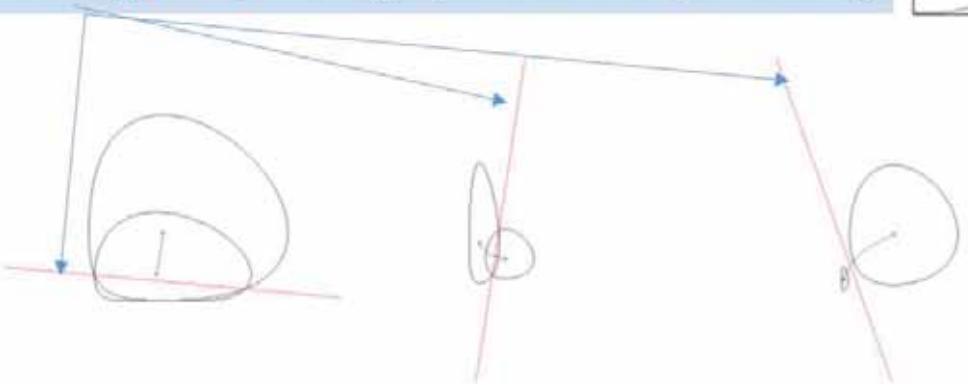
<u>Key property</u>:
Check whether two Bregman spheres
Intersect or not easily
(radical hyperplane, space of spheres)

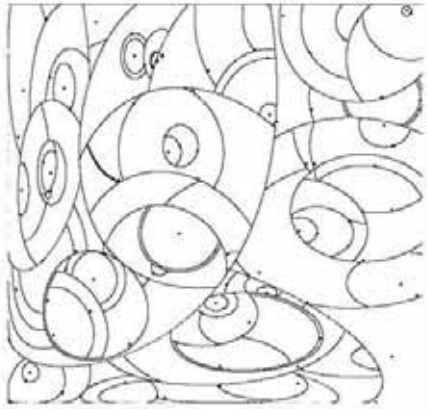






Bregman ball trees





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

@FrnkNlsn

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

E.g., Extended Kullback-Leibler

Fast Hessian-vector multiplication: Pearlmutter's R-operator

- When training deep neural networks, need to perform Hessian-vector matrix multiplications. The **R-differential operator** (= Gateau derivative) allows to perform the multiplication Hessian-vector in linear time.
- Goal is to calculate $H(w) \times v$ where $H(w) = \nabla_w^2 f(w)$
- First-order Taylor expansion: $\nabla_w f(w + \Delta w) = \nabla_w f(w) + \nabla_w^2 f(w) \Delta w + O(\|\Delta w\|^2)$

Approximation: $Hv \approx \dfrac{\nabla_w f(w + \lambda v) - \nabla_w f(w)}{\lambda}$ **Exact:** $Hv = \lim_{\lambda \to 0} \dfrac{\nabla_w f(w + \lambda v) - \nabla_w f(w)}{\lambda}$

Operator: $\mathcal{R}_v(f(w)) := \left. \dfrac{\partial f(w + \lambda v)}{\partial \lambda} \right|_{\lambda=0}$ $\Rightarrow$ $Hv = \mathcal{R}_v(\nabla_w)$, $\mathcal{R}_v(w) = v$

f is a computational graph (neural network) -> automatic differentiation

Pearlmutter, Fast exact multiplication by the Hessian, Neural computation 6.1 (1994)
Schraudolph, Fast curvature matrix-vector products for second-order gradient descent, Neural computation 14.7 (2002)

Arimoto α-entropies, f_β-divergences and their metrizations

- Arimoto's α-entropies: $$h_\alpha(t) = \begin{cases} \frac{1}{1-\alpha}\left[1-\left(t^{1/\alpha}+(1-t)^{1/\alpha}\right)^\alpha\right] & \text{if} \quad \alpha \in (0,\infty)\backslash\{1\} \\ -t\ln t-(1-t)\ln(1-t) & \text{if} \quad \alpha = 1 \\ \min\{t, 1-t\} & \text{if} \quad \alpha = 0 \end{cases}$$
- f-divergences (f convex, strictly convex at 1): $I_f[p:q] := \int p f\left(\frac{q}{p}\right) \mathrm{d}\mu,$
- f_β-divergences: $$f_\beta(u) = \begin{cases} \frac{1}{1-1/\beta}\left\{\left(1+u^\beta\right)^{1/\beta} - 2^{1/\beta-1}(1+u)\right\} & \text{if} \quad \beta \in (0,\infty)\backslash\{1\} \\ (1+u)\ln(2)+u\ln(u)-(1+u)\ln(1+u) & \text{if} \quad \beta = 1 \\ |1-u|/2 & \text{if} \quad \beta = \infty \end{cases}$$
- Include squared Hellinger, capacitory discrimination (2x Jensen-Shannon divergence), and total variation

$$I_{f_{\frac{1}{2}}}[p:q] = D^2_{\text{Hellinger}}[p:q], \quad I_{f_1}[p:q] = 2D_{\text{JS}}[p:q], \quad I_{f_\infty}[p:q] = D_{\text{TV}}[p:q] = \frac{1}{2}\int |p-q| \mathrm{d}\mu$$

- **Theorem [Metrization of f_β-divergences]:**

$$\rho_\beta[p,q] := \left(I_{f_\beta}(p,q)\right)^{\min\{\beta,\frac{1}{2}\}}$$ is a metric distance (triangle inequality)

Osterreicher & Vajda, A new class of metric divergences on probability spaces and its applicability in statistics, ISM 55.3 (2003)
On the chi square and higher-order chi distances for approximating f-divergences, IEEE Signal Processing Letters 21.1 (2013)

f-divergences as integrals of weighted Bregman divergences

Scalar Bregman divergence induced by a strictly convex and differentiable generator f called a **Bregman generator** :

$$B_f(a:b) = f(a) - f(b) - (a-b)f'(b) \geq 0.$$

Statistical f-divergence for a Bregman generator f such that **f(1)=0**

$$\boxed{I_f(p:q) = \int_{x\in\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x \Rightarrow I_f(p:q) = \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) \mathrm{d}x = E_p\left[B_f\left(\frac{q(x)}{p(x)} : 1\right)\right]}$$

Proof:

$$\begin{aligned}\int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) \mathrm{d}x &= \int p(x)\left(f\left(\frac{q(x)}{p(x)}\right) - \underbrace{f(1)}_{=0} - \left(\frac{q(x)}{p(x)} - 1\right) f'(1)\right) \mathrm{d}x, \\ &= \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x - f'(1) \underbrace{\int (q(x) - p(x)) \mathrm{d}x}_{0} =: I_f(p:q).\end{aligned}$$

(this does not include the total variation f-divergence)

On information projections between multivariate elliptical and location-scale families, arXiv 2101.03839

Approximating f-divergences with chi2 distances

- For a generator f convex, strictly convex at 1 with f(1)=0, f-divergence is:

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

Include Kullback-Leibler divergence, alpha-divergences, Jeffreys divergence Jensen-Shannon divergences, etc.

- Chi2 distance (non-metric, Pearson and Neyman version):

$$\chi^2_P(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^2}{x_1(x)} \mathrm{d}\nu(x) \qquad \chi^2_N(X_1 : X_2) = \int \frac{(x_1(x) - x_2(x))^2}{x_2(x)} \mathrm{d}\nu(x)$$

- Approximating f-divergence with chi2:

$$\boxed{I_f(X_1 : X_2) \sim \frac{f''(1)}{2} \chi^2_N(X_1 : X_2)}$$

(since f is strictly convex at 1, f''>0)

$$\Rightarrow \quad \chi^2_P(X_1 : X_2) \sim 2\,\mathrm{KL}(X_1 : X_2). \quad \Rightarrow \quad I_{\mathrm{JS}}(X_1 : X_2) \sim \tfrac{1}{4}\chi^2_N(X_1 : X_2)$$

On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 2013
On power chi expansions of f -divergences, arXiv:1903.05818

Taylor expansions of f-divergences with order-k chi divergences/power chi divergences

- Generalization of the Pearson/Neyman chi_2 divergences for integers k>1
- Proper f-divergences when **k is even** but may be negative or violate law of the indiscernibles when k is odd:

$$D^P_{\chi,k}(p:q) = I_{f^P_{\chi,k}}(p:q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x), \quad f^P_{\chi,k}(u) = (u-1)^k$$

$$D^N_{\chi,k}(p:q) = I_{f^N_{\chi,k}}(p:q) = D_{\chi,k}(q:p) = \int \frac{(p(x)-q(x))^k}{q(x)^{k-1}} \mathrm{d}\mu(x), f^N_{\chi,k}(u) = \frac{(1-u)^k}{u^{k-1}}$$

- **Power chi expansions of f-divergences** under converging conditions:

$$f(x) = \sum_{n=1}^{\infty} a_n (x-1)^n, f \in C^{\omega} \quad \Rightarrow \quad I_f(p_{\theta_1} : p_{\theta_2}) = \sum_{i=2}^{\infty} a_i D_{\chi,i}(p_{\theta_1} : p_{\theta_2})$$

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

f-divergences as power chi divergence series

- **Power chi divergence** (non-negative for even order, can be negative for odd order):

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

Generalize chi-square divergence for k=2

Theorem (**f-divergences as power chi divergence series**) *Let X be a topological space and μ be a Borel measure on X with full support. Assume that $p(x)$ and $q(x)$ are positive and continuous probability density functions with respect to x with full support, and that there exists a constant $C := C(p,q)$ such that $p(x) \leq Cq(x)$ for every $x \in X$. Let $f(u) = \sum_{n=1}^{\infty} a_n (u-1)^n$ be an analytic function with convergence radius $r_f \geq 1$.*

- *If $\frac{q(x)}{p(x)} < 1 + r_f$ for every x then*

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

- *If $\frac{q(x)}{p(x)} > 1 + r_f$ for some x, then, the infinite sum $\sum_{n=2}^{\infty} a_n D_{\chi,n}(p:q)$ diverges*

arXiv:2101.12459

Examples:

r_f=1

$$D_{\mathrm{KL}}(p:q) = \sum_{n=2}^{\infty} \frac{(-1)^n}{n} D_{\chi,n}(p:q)$$

$$D_{\mathrm{JS}}(p:q) = \sum_{n=2}^{\infty} \frac{(-1)^n (2^{n-1}-1)}{n(n-1)2^{n-1}} D_{\chi,n}(p:q)$$

f-divergence between location-scale densities[1]

- $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$: open half-space of 2D (l, s) location-scale parameters
- Location-scale families $\mathcal{P} = \{\frac{1}{s_1} p\left(\frac{x-l_1}{s_1}\right) \ : \ (l_1, s_1) \in \mathbb{H}\}$ and $\mathcal{Q} = \{\frac{1}{s_2} q\left(\frac{x-l_2}{s_2}\right) \ : \ (l_2, s_2) \in \mathbb{H}\}$, where p and q are *arbitrary* densities
- f-**divergence** for a positive convex function on $\mathbb{R}_{++}$, strictly convex at 1, with $f(1) = 0$:

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

  Kullback-Leibler divergence: f-divergence for $f(u) = -\log u$
- Action of the **location-scale group** on f-divergences:

$$\boxed{I_f(p_{l_1,s_1} : q_{l_2,s_2}) = I_f\left(p : q_{\frac{l_2-l_1}{s_1}, \frac{s_2}{s_1}}\right) = I_f\left(p_{\frac{l_1-l_2}{s_2}, \frac{s_1}{s_2}} : q\right)}$$

- For submanifolds $\mathcal{P}$ and $\mathcal{Q}$ of the density manifold, $I_f(p_{l_1,s_1} : \mathcal{Q}) := \min_{(l_2,s_2)\in\mathbb{H}} I_f(p_{l_1,s_1} : q_{l_2,s_2}) = I_f(p : \mathcal{Q})$ is **independent** of (l_1, s_1). Similarly, $I_f(\mathcal{P} : q_{l_2,s_2}) = I_f(\mathcal{P} : q)$, independent of (l_2, s_2).

[1] https://arxiv.org/abs/1904.10428

Fisher information matrix (FIM) of multivariate normals

- Covariance matrix: $\Sigma = [\sigma_{ij}]$
- Precision matrix (inverse covariance matrix): $\Sigma^{-1} = [\sigma^{ij}]$
- Fisher information matrix (mnemonic block expression):

$$\boxed{I(\mu,\Sigma) = \begin{array}{c} \begin{array}{cc} \mu & \Sigma = [\sigma_{ij}] \end{array} \\ \left[\begin{array}{c|c} \sigma^{ij} & 0 \\ \hline 0 & \sigma^{il}\sigma^{jk} + \sigma^{ik}\sigma^{jl} \end{array}\right] \end{array} \begin{array}{c} \mu \\ \Sigma = [\sigma_{kl}] \end{array}}$$

Orthogonal parameters

$$\mu \perp \Sigma$$

- For univariate normals, $\sigma^{11} = \frac{1}{\sigma^2}$ $\Rightarrow$ $I(\mu,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$

Skovgaard, "A Riemannian geometry of the multivariate normal model", Scandinavian journal of statistics (1984)
An elementary introduction to information geometry, *arXiv:1808.08271*

Fisher information of a natural exponential family

- **Fisher information identity** under regularity conditions:

$$I(\theta) = [g_{ij}(\theta)] := E_{X\sim p_\theta}\left[(\nabla_\theta \log p_\theta(x))\ (\nabla_\theta \log p_\theta(x))^\top\right]$$
$$= -E_{X\sim p_\theta}\left[\nabla_\theta \nabla_\theta \log p_\theta(x)\right] = -E_{X\sim p_\theta}\left[\nabla^2_\theta \log p_\theta(x)\right]$$

- Natural exponential family (NEF) of order D: $\mathcal{E} := p_\theta(x) = \exp\left(\sum_{i=1}^{D} \theta_i t_i(x) - F(\theta) + k(x)\right)$
- Fisher information of a NEF with sufficient statistic D-dimensional vector t(X):

$$\boxed{\begin{aligned} I(\theta) &= -E_{X\sim p_\theta}\left[\nabla^2_\theta \log p_\theta(x)\right] = \nabla^2 F(\theta) \qquad I(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta(\theta) \\ &= \mathrm{Cov}[t(X)] := E_{X\sim p_\theta}\left[t(X)t(X)^\top\right] - E_{X\sim p_\theta}\left[t(X)\right] E_{X\sim p_\theta}\left[t(X)\right]^\top \end{aligned}}$$

- Fisher information of a NEF from moments of sufficient statistics:

$$\eta_i(\theta) = E_{X\sim p_\theta}\left[t_i(X)\right] = \frac{\partial}{\partial \theta_i} F(\theta) =: \partial_i F(\theta)$$
$$\eta_{ij}(\theta) = E_{X\sim p_\theta}\left[t_i(X) t_j(X)\right]$$

$$\Rightarrow \boxed{g_{ij}(\theta) = \eta_{ij}(\theta) - \eta_i(\theta)\eta_j(\theta)}$$

An Elementary Introduction to Information Geometry, Entropy 2020
Statistical exponential families: A digest with flash cards, arXiv:0911.4863

Fisher information matrix with (μ,log(σ)) parameterization

Fisher information

Usual parameterization: $\theta = (\theta_1 = \mu, \theta_2 = \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$

$$I_\theta(\theta) = \begin{bmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{bmatrix}$$

New parameterization: $\eta = \left(\eta_1 = \theta_1, \eta_2 = \frac{1}{2}\log\theta_2\right) = (\mu, \log\sigma) \in \mathbb{R}^2$

Fisher information matrix is covariant under reparameterization:

$$\begin{aligned} I_\eta(\eta) &= \left[\frac{\partial\theta_i}{\partial\eta_j}\right]_{ij}^\top I(\theta(\eta)) \left[\frac{\partial\theta_i}{\partial\eta_j}\right]_{ij} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 2\exp(2\eta_2) \end{bmatrix} \begin{bmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2\exp(2\eta_2) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 4\exp(4\eta_2) \end{bmatrix} \begin{bmatrix} \frac{1}{\exp(2\eta_2)} & 0 \\ 0 & \frac{1}{2\exp(4\eta_2)} \end{bmatrix} \end{aligned}$$

with Jacobian matrix: $\left[\frac{\partial\theta_i}{\partial\eta_j}\right]_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 2\exp(2\eta_2) \end{bmatrix}$

$$\boxed{I_\eta(\eta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2 \end{bmatrix}}$$

Poisson family: Fisher information reparameterizations

- **Fisher information** is *invariant by diffeomorphism of the sample space* and *covariant by reparameterization of the parameter space*:

$$I_\theta(\theta) = E_{p_\theta}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\right)^2\right] \Rightarrow \boxed{I_\lambda(\lambda) = \frac{\mathrm{d}\theta}{\mathrm{d}\lambda} I_\theta(\theta(\lambda)) \frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = \frac{1}{\lambda}\exp(\log(\lambda))\frac{1}{\lambda} = \frac{1}{\lambda}}$$

- Consider the Poisson family (univariate and uni-order family):

$$\mathcal{P} = \left\{p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda \in (0,\infty)\right\}$$
$$= \{p_\theta(x) = \exp(\theta t(x) - F(\theta))\mathrm{d}\nu(x) \ : \ \theta \in \Theta\}$$

Exponential family decomposition

$$\nu(x) = \frac{\mu(x)}{x!} = e^{k(x)}\mu(x)$$
$$t(x) = x$$
$$\theta = \theta(\lambda) = \log(\lambda) \in \Theta$$
$$F(\theta) = \exp(\theta)$$
$$\eta = F'(\theta) = \lambda$$

$$\boxed{I_\theta(\theta) = \mathrm{Var}_{p_\theta}[x] = F''(\theta) = \exp(\theta)}$$

$$F^*(\eta) = \eta\log\eta - \eta$$

$$E_{p_\theta}[x] = F'(\theta) = \exp(\theta)$$

$$\boxed{I_\eta(\eta) = (F^*)''(\eta) = \frac{1}{\eta}}$$

Crouzeix identity $F''(\theta)(F^*)''(\eta) = \lambda \times \frac{1}{\lambda} = 1$

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = B_F(\theta(\lambda_2) : \theta(\lambda_1)) = \lambda_1 \log\frac{\lambda_1}{\lambda_2} + \lambda_2 - \lambda_1$$

Rao distance: $D_\rho(\lambda_1, \lambda_2) = 2\left|\sqrt{\lambda_1} - \sqrt{\lambda_2}\right|$

https://arxiv.org/abs/1808.08271

Reparameterization of the Fisher information matrix

For two parameterizations λ and λ′ of a parametric family of densities, the Fisher information matrix relates to each other by:

$$\boxed{I_{\lambda'}(\lambda') = J_{\lambda,\lambda'}^{\top} I_{\lambda}(\lambda(\lambda')) J_{\lambda,\lambda'}}$$

Jacobian matrix: $J_{\lambda,\lambda'} = \left[\frac{\partial \lambda_i}{\partial \lambda'_j}\right]$

Example: the Gaussian family

Mean-standard deviation param.: $p(x;\lambda=(\mu,\sigma)) = \frac{1}{\sqrt{2\pi}\lambda_2^2}\exp\left(-\frac{1}{2}\frac{(x-\lambda_1)^2}{\lambda_2^2}\right)$ $\quad I_{\lambda}(\lambda) = \begin{bmatrix} \frac{1}{\lambda_2^2} & 0 \\ 0 & \frac{2}{\lambda_2^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$

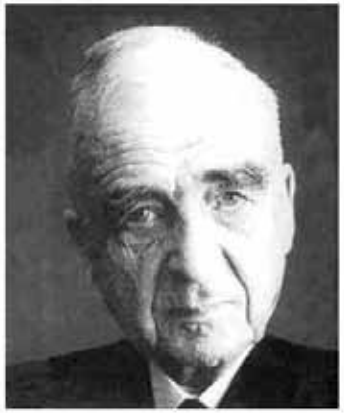
Mean-variance parameterization: $p(x;\lambda'=(\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi}\lambda'_2}\exp\left(-\frac{1}{2}\frac{(x-\lambda'_1)^2}{\lambda'_2}\right)$ $\quad I_{\lambda'}(\lambda') = \begin{bmatrix} \frac{1}{\lambda_2} & 0 \\ 0 & \frac{1}{2\lambda_2^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$

$$J_{\lambda',\lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} \Rightarrow I_{\lambda}(\lambda) = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

An elementary introduction to information geometry, 2018

https://arxiv.org/abs/1808.08271

Medians and means in Finsler geometry



Paul Finsler
(1894-1970)
German/Swiss

Several generalizations of the centroids in Riemannian geometry :
Karcher means or exponential means or Fréchet means (set)

Finsler geometry generalizes Riemannian geometry:
In Finsler geometry, tangent space is equipped with a **Minkowski norm**

(forward) p-mean minimizes (p>1): $$\mathscr{E}_{\mu,p}(x) = \int_M \rho^p(x,z)\mu(dz)$$

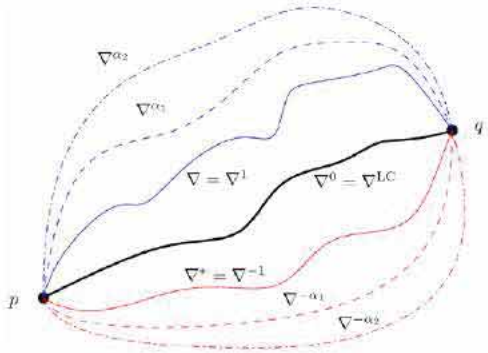
(forward) median minimizes: $$\mathscr{F}_\mu(x) = \int_M \rho(x,z)\mu(dz)$$

Existence and uniqueness conditions + algorithms reported in:

LMS Journal of Computation and Mathematics 15 (2012): 23-37.
https://arxiv.org/abs/1011.6076

@FrnkNlsn

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

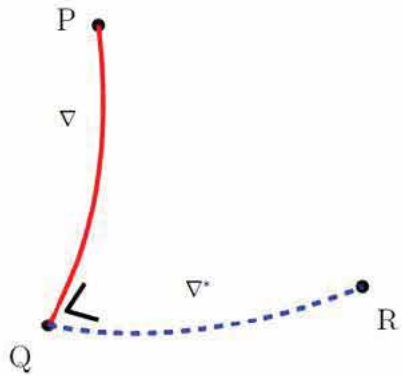


Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/



An elementary introduction to information geometry

https://arxiv.org/abs/1808.08271

Fisher's fundamental theorem of natural selection

"The rate of increase in the average fitness of a population is equal to the genetic variance of fitness of that population" (1930)

Sir Ronald Fisher (1890-1962)

At any time t

$$\frac{\partial_{\text{NaturalSelection}}}{\partial t}\text{MeanFitness}_t = \text{GeneticFitnessVariance}_t$$

For a population P_i with fitness functior f_i , the Lotka–Volterra equation is:

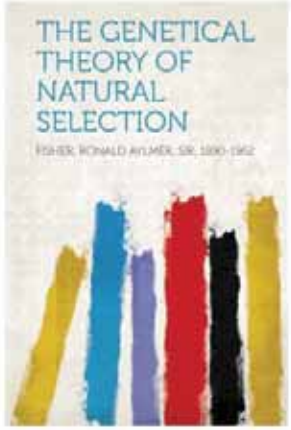
$$\dot{P}_i = f_i(P)P_i$$

Variance of the fitness: $\text{Var}(f(P)) = \sum_j \left(f_j(P) - \bar{f}(P)\right)^2 p_j$ where $p_i = \frac{P_i}{\sum_i P_i}$

$$g(\dot{p}, \dot{p}) = \text{Var}[f(P)]$$

where g is the Fisher metric of the probability simplex:

$$g_p(v, v') = \sum_{i=1} \frac{v_i v_i'}{p_i}$$



Kagan's Fisher information theorem (1985): Exponential family vs non-exponential family

- (Curved) Exponential families have **finite dimensional sufficient statistics** and are geometrically **autoparallel submanifolds** wrt to the e-connection
- Rationale: 1D CEFs E should be "simpler" compared to other "competing" parametric families F indexed by the same scalar parameter θ.

$$\mathcal{E} = \left\{ p_\theta(x) = \exp\left(\sum_{i=1}^{D} c_i(\theta) t_i(x) + k(x) \right) \ : \ \theta \in I \subset \mathbb{R} \right\} \quad \text{versus} \quad \mathcal{F} = \{ q_\theta(x) : \ \theta \in I \subset \mathbb{R} \}$$

- **Kagan's theorem:** conditions to compare **Fisher information**: $\boxed{I_{\mathcal{E}}(\theta) \leq I_{\mathcal{F}}(\theta)}$
- Example: location Gaussian family vs location family with standard density q

$$\mathcal{E} = \left\{ p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} \right) \ : \ \theta \in \mathbb{R} \right\} \quad \text{versus} \quad \mathcal{F} = \{ q_\theta(x) = q(x-\theta) : \ \theta \in I \subset \mathbb{R} \}$$

Conditions: $\int x q(x) \mathrm{d}x = 0$ and $\int x^2 q(x) \mathrm{d}x = \sigma^2$ $\Longrightarrow$ $I_{\mathcal{E}}(\theta) = \frac{1}{\sigma^2} \leq I_{\mathcal{F}}(\theta)$

Kagan, "Information property of exponential families." *Theory of Probability & Its Applications* 30.4 (1986)
English translation of 1985 Russian paper

Fisher-Rao distance on location-scale manifolds

Proposition 4 (Fisher-Rao distance on a location-scale manifold). *The Fisher-Rao distance between two densities p_{l_1,s_1} and p_{l_2,s_2} of a location-scale family $\mathcal{F}_p$ with even standard density $p(x) = p(-x)$ on the support $\mathbb{R}$ is*

$$\boxed{\rho_p((l_1,s_1),(l_2,s_2)) = b\ \rho_U\left(\left(\frac{a}{b}l_1, s_1\right), \left(\frac{a}{b}l_2, s_2\right)\right)},$$

where $a = \sqrt{E_p\left[\left(\frac{p'(x)}{p(x)}\right)^2\right]}$ *and* $b = \sqrt{E_p\left[\left(x\frac{p'(x)}{p(x)}+1\right)^2\right]} > 0$, *and*

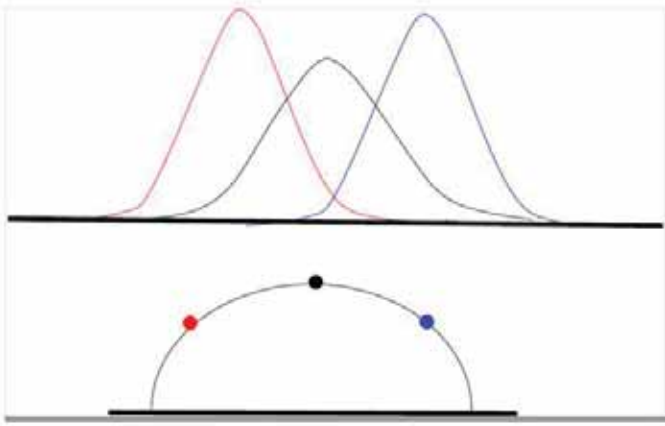
$$\rho_U((l_1,s_1),(l_2,s_2)) = \operatorname{arccosh}\left(\frac{1-(l_1l_2+s_1s_2)}{\sqrt{(1-(l_1^2+s_1^2))\,(1-(l_2^2+s_2^2))}}\right).$$

Proposition 3 (Fisher information of a location-scale family). *The Fisher information matrix $I(\lambda)$ of a location-scale family with continuously differentiable standard density $p(x)$ with full support $\mathbb{R}$ is* $I(\lambda) = \frac{1}{s^2}\begin{bmatrix} a^2 & c \\ c & b^2 \end{bmatrix}$, *where* $a^2 = E_p\left[\left(\frac{p'(x)}{p(x)}\right)^2\right]$, $b^2 = E_p\left[\left(1+x\frac{p'(x)}{p(x)}\right)^2\right]$ *and* $c = E_p\left[\frac{p'(x)}{p(x)}\left(1+x\frac{p'(x)}{p(x)}\right)\right]$.

Example 9. *The Fisher-Rao distance between two normal densities $p^{\mathcal{N}}_{\mu_1,\sigma_1}$ and $p^{\mathcal{N}}_{\mu_12,\sigma_2}$ is*

$$\boxed{\rho_{p^{\mathcal{N}}}((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{2}\ \rho_U\left(\left(\frac{1}{\sqrt{2}}\mu_1, \sigma_1\right), \left(\frac{1}{\sqrt{2}}\mu_2, \sigma_2\right)\right)}$$

since $a^2 = 1$, $b^2 = 2$, $\kappa = -\frac{1}{2}$.



Hyperbolic upper plane

$$\mathcal{N} := \left\{ p^{\mathcal{N}}_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \ :\ (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_{++} \right\}$$

On information projections between multivariate elliptical and location-scale families, arXiv 2101.03839

Beware that Fisher information may be infinite!

- **Fisher information matrix** quantifies the information that a random variable X carries about a vector parameter θ:

$$I_X(\theta) := E_{X\sim p_\theta(x)}\left[(\nabla_\theta \log p_\theta(X))^\top \nabla_\theta \log p_\theta(X)\right]$$

- Fisher information of a **location model** is *independent of the parameter*

$$\mathcal{L} := \{p_\theta(x) = p_0(x-\theta) \ : \ \theta \in \mathbb{R}\} \quad \Rightarrow \quad \boxed{I(\theta) = \int_{-\infty}^{\infty} \frac{(p_0'(x))^2}{p_0(x)} dx}$$

- Fisher information matrix may be **infinite**
- For example, consider the location model with standard distribution p0:

$$p_0(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if} \quad |x| < 1 \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow$$

Then the integral defining the Fisher information diverges! I(θ)=∞

Fisher-Rao Riemannian geometry (Hotelling precursor)

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

Metric tensor = **Fisher information metric**

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

Infinitesimal squared length element:

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

Fisher-Rao distance satisfying the metric axioms:

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

Geodesic length distance
(shortest path)

- Statistical data analysis and inference, Yadolah Dodge (Ed), 1989
- An elementary introduction to information geometry, arXiv:1808.08271
- Cramér-Rao Lower Bound and Information Geometry,
  Connected at Infinity II, 2013 - Springer

Fisher scoring: Natural gradient ascent for the MLE

- Parametric family of distributions {p(x;θ)} with twice differentiable densities
- Score s(θ) = gradient of the log-likelihood function with respect to parameter θ:

Likelihood $L(\theta; x) = p(x; \theta)$ → $l(\theta; x) = \log L(\theta; x)$ → $s(\theta) = \nabla_\theta l(\theta; x)$

- Maximum Likelihood Estimator (MLE):

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} l(\theta; x_i)$$

- When the MLE is not in closed-form, use **Fisher scoring** = an iterative procedure

Fisher scoring: $\theta_{t+1} = \theta_t + I^{-1}(\theta_t) s(\theta_t)$ with $I(\theta) := -E_{X \sim p(x;\theta)}[\nabla^2 l(\theta; x)]$

(Fisher information matrix)

Natural gradient ascent: (NGA)

$$\boxed{\theta_{t+1} = \theta_t + \underbrace{I^{-1}(\theta_t)\nabla_\theta}_{:=\nabla^{\mathrm{NG}}} \sum_{i=1}^{n} l(\theta; x_i)}$$

- Fischer scoring = **natural gradient ascent on the log-likelihood loss function**

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10)

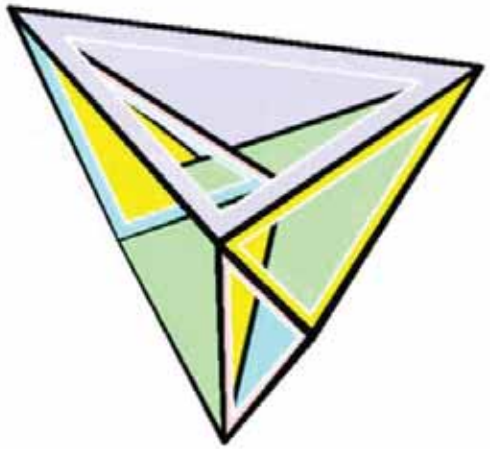
Rigid and flexible polyhedra

- A polyhedron with **hinged edges** is **flexible** if its shape can be smoothly deformed (dihedral angles change continuously).

  If not, it is said **rigid**.
- **Cauchy's theorem** (1813): **3D convex polyhedra are rigid**. That is, polyhedra with same face lattice and congruent faces are congruent.
- **Alexandrov's theorem** (1950): **dD convex polyhedra are rigid** for d>2.
- Connelly (1977) constructed **the first flexible non-convex 3D polyhedron**. Steffen reported another simpler flexible polyhedron

  (Bricard's octahedra have intersecting faces)
- Flexible 3D polyhedra have **invariant volume** while flexing (1997)
- Surface of a Euclidean convex polyhedron is a **geodesic metric space**.

Hinged flexible polyhedron (Connelly, Courtesy of IHES)



Steffen's flexible polyhedron (Courtesy of Wikipedia)

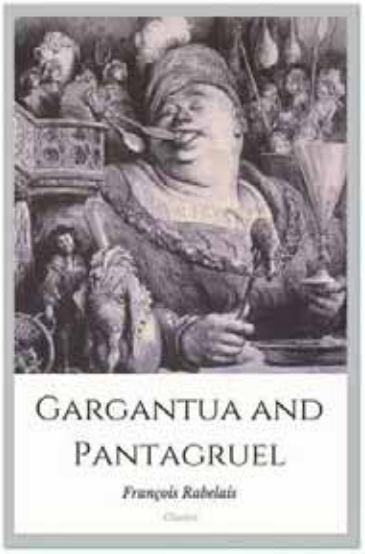
Shaping Space: Exploring Polyhedra in Nature, Art, and the Geometrical Imagination, Senechal, Marjorie (Ed.), Springer, 2013

Quote for thoughts…

“Science without conscience is the soul's perdition.”

François Rabelais

(>=1483,<=1553)



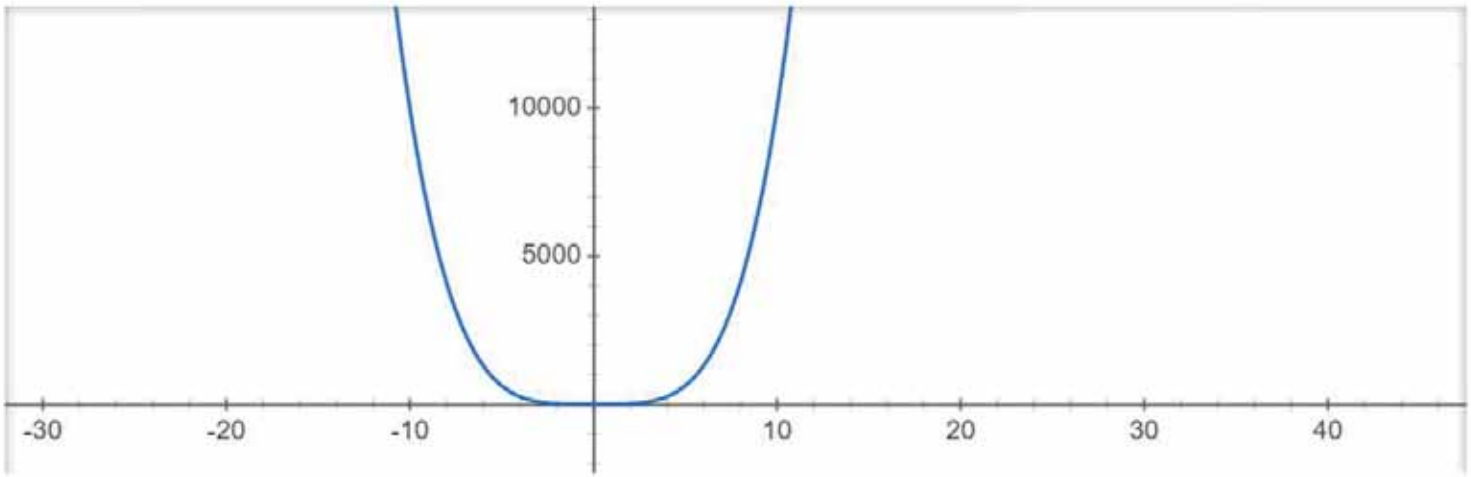
Convexity vs strict convexity wrt 2nd order derivatives:

- A twice differentiable function f is **convex if and only if** $f''(x) \geq 0$ for x in I

- **If** $f''(x) > 0$ for x in I then f(x) is **strictly convex** on I (⚠ but not iff !!!)

  For example: $f(x)=x^4$ with $f''(x)=12x^2$ is strictly convex although $f''(0)=0$

Graph for x^4



Riemann-Christoffel curvature tensors and the fundamental theorem of information geometry

- **Curvature** is a fundamental notion in geometry: from scalar curvature, sectional curvature, Gaussian curvature of surfaces to **Riemann-Christoffel (RC) 4-tensor**, Ricci symmetric 2-tensor, synthetic Ricci, etc.
- In **information geometry**, a manifold M is equipped with a pair of dual torsion-free affine connections (∇, ∇*) coupled to the metric g: (M,g,∇, ∇*)
- **Definition**: A statistical structure (M,g,∇) is said of **constant curvature κ** when

$$R^{\nabla}(X,Y)Z = \kappa\{g(Y,Z)X - g(X,Z)Y\}, \quad \forall X,Y,Z \in \Gamma(TM)$$

- The fundamental theorem of information geometry relates the RC tensors of the dual connection:

$$g\left(R^{(\alpha)}(X,Y)Z,W\right) + g\left(Z,R^{(-\alpha)}(X,Y)W\right) = 0$$

$$g\left(R^{\nabla^*}(X,Y)Z,W\right) = -g\left(Z,R^{\nabla}(X,Y)W\right)$$

- **Fundamental Theorem**: **κ-constant (M,g,∇) iff κ-constant (M,g,∇*)**
- **Corollary**: A manifold (M,g,∇, ∇*) is ∇-flat iff it is ∇*-flat

Zhang, A note on curvature of α-connections of a statistical manifold, ISM 2007.
An elementary introduction to information geometry, *arXiv:1808.08271*

The fundamental theorem of information geometry

- A manifold M equipped with a pair of torsion-free affine connections (∇, ∇^*) coupled to the metric g (= dual connections): (M,g,∇, ∇^*)
- **Definition**: A statistical structure (M,g,∇) is said of **constant curvature κ** when

$$R^{\nabla}(X,Y)Z = \kappa\{g(Y,Z)X - g(X,Z)Y\}, \quad \forall X,Y,Z \in \Gamma(TM)$$

Fundamental theorem of information geometry relates the **Riemann-Christoffel** tensors of the dual torsion-free connections:

Fundamental Theorem of IG: **(M,g,∇) is κ-constant iff (M,g,∇^*) is κ-constant**

Corollary: An information manifold (M,g,∇, ∇^*) is ∇-flat (i.e.,κ=0) iff it is ∇^*-flat

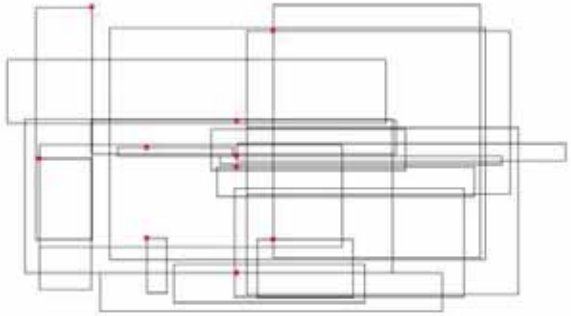
An elementary introduction to information geometry, *Entropy 2020*

How many points to pierce pairwise intersecting objects?

Gallai numbers with $\prod^q$ = Piercing/stabbing property by q points:

Definition 1. If every subset of $\mathscr{S}$ ($\mathscr{S} \in \mathscr{C}$) of size p is $\prod^q$ (let $\prod_p^q$ denote that property) then $\mathscr{S}$ is pierceable with $\mathfrak{G}(p,q,\mathscr{C})$ points, where $\mathfrak{G}(p,q,\mathscr{C})$ is the Gallai number of class $\mathscr{C}$ for the $\prod^q$ property of its p-subsets.

- Pairwise intersecting disks on the plane pierced by **4 points**: $\mathfrak{G}(2,1,\mathscr{D}^2)=4$ (with a linear-time algorithm)
- In dimension d, pairwise intersecting of balls pierced by **exponential number** of points: $\mathfrak{G}(2,1,\mathscr{D}^d)>1.003^d$
- Minimum example of pairwise intersecting disks which cannot be intersected by 3 points: 21 [**Grunbaum 1959**], 13 [**Har-Peled et al.2018**]



On piercing sets of objects, ACM SoCG 1996
On Point Covers of c-Oriented Polygons, TCS 2000
Carmi et al., Stabbing pairwise intersecting disks by four points, arXiv:1812.06907

Constant Gaussian curvatures: Hyperquadric isometries

- When the **Gaussian curvature** κ is **constant**, the D-dimensional manifold can be immersed in $\mathbb{R}^{D+1}$ using a **hyperquadric**

$$\sum_{i=1}^{D+1} a_i z_i^2 = \frac{1}{\kappa}, \quad \forall i,\ a_i \in \{-1, 1\}.$$

- For example, $\kappa = -\frac{1}{2}$ for Gaussian family , and hyperquadric is the pseudo-sphere $z_1^2 + z_2^2 - z_3^2 = -2$ (pseudosphere) with

$$\begin{aligned} z_1 &= \frac{\mu}{\sigma} \in \mathbb{R} \\ z_2 &= \frac{\mu^2 - 2}{2\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}} \in \mathbb{R} \\ z_3 &= \frac{\mu^2 + 2}{2\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}} \in \mathbb{R}_{++} \end{aligned}$$

- Location-scale families have constant negative Gaussian curvatures

Some recently generalization of distances:

- **α-divergences based on comparable *weighted means***. arXiv: 2001.09660

→ Obtain also generalized Kullback-Leibler divergences

- ***Vector-skewed* Jensen-Shannon divergence/Jensen-Shannon** Centroid. 10.3390/e22020221

→ Obtain also vector-skewed Jensen-Bregman divergences

- ***(M,N)*-Jensen-Shannon divergences.** 10.3390/e21050485

→ Examples: Geometric JSD for exponential families, Power JSDs for t-Student distributions

- ***Conformal* Divergences** and their centroids 10.1109/TIT.2015.2448072

→ Generalize the dually flat (u,v)-structures of information geometry

- **Bhattacharyya/Chernoff using *quasi-arithmetic means*.** 10.1016/j.patrec.2014.01.002
- ***Matrix* Bregman divergences.** ICML 2011

→ Spectral matrix distances

- ***Higher-order* chi divergences** and f-divergences. 10.1109/LSP.2013.2288355

→ Taylor series for f-divergences based on k-order chi divergences

A generalization of α-divergences

α-divergences

$$I_\alpha(p:q) := \begin{cases} \frac{1}{\alpha(1-\alpha)} \int \left(\alpha p + (1-\alpha)q - p^\alpha q^{1-\alpha}\right) \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\} \\ I_1(p:q) = \mathrm{KL}_e(p:q), & \alpha = 1 \\ I_0(p:q) = \mathrm{KL}_e(q:p), & \alpha = 0 \end{cases}$$

Extended Kullback-Leibler divergence: $\mathrm{KL}_e(p:q) := \int \left(p \log \frac{p}{q} + q - p\right) \mathrm{d}\mu$

Quasi-arithmetic weighted mean: $M_\alpha^f(p,q) = f^{-1}\left((1-\alpha)f(p) + \alpha f(q)\right)$

Theorem 4 (Quasi-arithmetic α-divergences). *Let f and g be two strictly continuously increasing and differentiable functions on $(0,\infty)$ such that $f \circ g^{-1}$ is strictly convex. Then the quasi-arithmetic α-divergences induced by (f,g) for $\alpha \in [0,1]$ is*

$$\boxed{I_\alpha^{f,g}(p:q) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \int \left(M_{1-\alpha}^f(p:q) - M_{1-\alpha}^g(p:q)\right) \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\}. \\ I_1^{f,g}(p:q) = \int \left(\frac{f(q)-f(p)}{f'(p)} - \frac{g(q)-g(p)}{g'(p)}\right) \mathrm{d}\mu & \alpha = 1, \\ I_0^{f,g}(p:q) = I_1^{f,g}(q:p) = \int \left(\frac{f(p)-f(q)}{f'(q)} - \frac{g(p)-g(q)}{g'(q)}\right) \mathrm{d}\mu, & \alpha = 0. \end{cases}} \quad (32)$$

Power (r,s) α-divergences

$$I_\alpha^{r,s}(p:q) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \int \left((\alpha p^r + (1-\alpha)q^r)^{\frac{1}{r}} - (\alpha p^s + (1-\alpha)q^s)^{\frac{1}{s}}\right) \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\}. \\ I_1^{r,s}(p:q) = \int \left(\frac{q^r - p^r}{rp^{r-1}} - \frac{q^s - p^s}{sp^{s-1}}\right) \mathrm{d}\mu & \alpha = 1, \\ I_0^{r,s}(p:q) = I_1^{r,s}(q:p) & \alpha = 0. \end{cases}$$

A generalization of the α-divergences based on comparable and distinct weighted means, arxiv 2001.09660

Generalized (f, g)-Kullback-Leibler divergences[1]

- f and g: two **strictly increasing and differentiable functions** such that $f \circ g^{-1}$ is **strictly convex**.
- (f, g)-cross-entropy, entropy $h^{f,g}$ and Kullback-Leibler divergence $\mathrm{KL}^{f,g}$ extended to **positive densities**:

$$
\begin{aligned}
h_{\times}^{f,g}(p:q) &= \int \left(\frac{f(q)}{f'(p)} - \frac{g(q)}{g'(p)} \right) \mathrm{d}\mu, h^{f,g}(p) = h_{\times}^{f,g}(p:p) \\
\mathrm{KL}^{f,g}(p:q) &= h_{\times}^{f,g}(p:q) - h^{f,g}(p) \geqslant 0, \\
&= \int \left(\frac{f(q)}{f'(p)} - \frac{g(q)}{g'(p)} - \frac{f(p)}{f'(p)} + \frac{g(p)}{g'(p)} \right) \mathrm{d}\mu
\end{aligned}
$$

- When $f(u) = u$ and $g(u) = \log u$, $f \circ g^{-1}(u) = \exp(u)$ is strictly convex, $\rightarrow$ Shannon's measures:

$$
\begin{aligned}
h_{\times}^{f,g}(p:q) &= \int (q - p \log q) \mathrm{d}\mu = h_{\times}(p:q) \\
h^{f,g}(p:q) &= h_{\times}^{f,g}(p:p) = \int (p - p \log p) \mathrm{d}\mu = h(p) \\
\mathrm{KL}_e(p:q) &= h_{\times}(p:q) - h(p) = \int (p \log \frac{p}{q} + q - p) \mathrm{d}\mu
\end{aligned}
$$

[1] https://arxiv.org/abs/2001.09660

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 expression, and these distances need to be **approximated** in applications.

  (machine learning, eg., deep learning in GANs)

- 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 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/

Statistical distances vs. mutual information

- A **statistical distance** is a measure of *distortion* (discrepancy) between **probability distributions** represented either by the probability density (eg., Kullback-Leibler divergence), cumulative distribution function (Kolmogorov-Smirnov distance), etc.
- A **mutual information (MI)** measures the *dependence* between **random variables**. It can be calculated as a statistical distance between the joint distribution and the product of marginal:

I(X,Y)=0
iff X is independent of Y

$$\boxed{I(X;Y) := D_{\mathrm{KL}}[p_{XY} : p_X p_Y] = \int p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \mathrm{d}x\mathrm{d}y}$$

Generalization of MI: $I(X;Y) := D[p_{XY} : p_X p_Y]$ or any statistical distance D

For example, **Rényi α-divergence -> α-mutual information**

- Sklar's theorem factorizes a multivariate joint distribution as a **copula** encoding all **dependence** times the product of **marginals**. Mutual information amounts to the negative copula entropy.

Optimal copula transport for clustering multivariate time series. ICASSP 2016
Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series, SSP 2016
On Rényi and Tsallis entropies and divergences for exponential families, arXiv:1105.3259

Three flavors of the Jensen-Shannon divergence

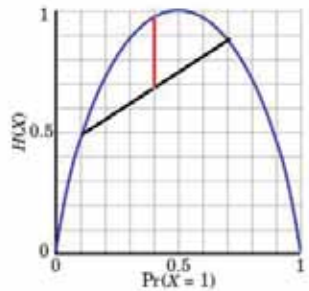
- **Symmetrization** of the relative entropy:

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

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means https://doi.org/10.3390/e22020221

- **Convexity gap** of Shannon negentropy:

$$D_{\mathrm{JS}}[p,q] \quad := \quad h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2}$$



On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid https://doi.org/10.3390/e22020221

- **Diversity index** of two probability measures:

$$D_{\mathrm{JS}}[p,q] \quad := \quad \min_{c \in \mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right)$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius https://doi.org/10.3390/e23040464

Generalization of the Jensen-Shannon divergence[1]

- Bregman divergence for F strictly convex and differentiable: $B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle$
- **Vector-skewed** α-Jensen-Bregman divergence $\alpha \in [0,1]^k$, $w \in \Delta_k$ with $\gamma \in [0,1]$, $\bar{\alpha} = \sum_{i=1}^k w_i \alpha_i$:

$$\begin{aligned} \mathrm{JB}_F^{\alpha,\gamma,w}(\boldsymbol{\theta_1} : \boldsymbol{\theta_2}) &:= \sum_{i=1}^k w_i B_F\left((\boldsymbol{\theta_1}\boldsymbol{\theta_2})_{\alpha_i} : (\boldsymbol{\theta_1}\boldsymbol{\theta_2})_{\gamma}\right) \geqslant \mathbf{0}, \\ \mathrm{JB}_F^{\alpha,w}(\boldsymbol{\theta_1} : \boldsymbol{\theta_2}) &= \left(\sum_{i=1}^k w_i F\left((\boldsymbol{\theta_1}\boldsymbol{\theta_2})_{\alpha_i}\right)\right) - F\left((\boldsymbol{\theta_1}\boldsymbol{\theta_2})_{\bar{\alpha}}\right) \end{aligned}$$

- **Vector-skewed** α-Jensen-Shannon divergence for $F = -h$ Shannon's entropy: (1) always bounded, (2) does not require matching supports, (3) extend to unnormalized densities.

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

A f-divergence for the generator $f_{\alpha,w}(u) = \sum_{i=1}^k w_i(\alpha_i u + (1-\alpha_i)) \log \frac{(1-\alpha_i)+\alpha_i u}{(1-\bar{\alpha})+\bar{\alpha} u}$

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

Bhattacharyya distance as a Jensen-Shannon symmetrization

- The **reverse Kullback-Leibler divergence**:

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

- Generic statistical **M-mixture** with respect to an abstract mean M(x,y):

$$(pq)_\alpha^M(x) := \frac{M_\alpha(p(x), q(x))}{Z_\alpha^M(p:q)} \qquad Z_\alpha^M(p:q) := \int_{t \in \mathcal{X}} M_\alpha(p(t), q(t)) \mathrm{d}\mu(t)$$

- Generic **Jensen-Shannon symmetrization** of a divergence D:

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

- **Bhattacharyya distance**:

$$B_\alpha(p:q) = -\log\left(\int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}\mu(x)\right)$$

- **<u>Bhattacharyya distance interpreted as a JS-symmetrization of the reverse KL</u>**:

Geometric mean:

$$G_\alpha(x,y) = x^{1-\alpha} y^\alpha$$

$$\mathrm{JS}_{\mathrm{KL}^*}^{G_\alpha}(p:q) := (1-\alpha)\mathrm{KL}^*(p : (pq)_\alpha^G) + \alpha \mathrm{KL}^*(q : (pq)_\alpha^G).$$

$$\boxed{\mathrm{JS}_{\mathrm{KL}^*}^{G_\alpha}(p:q) = B_{1-\alpha}(p:q)}$$

On a generalization of the Jensen-Shannon divergence and the JS-symmetrization of distances relying on abstract means

Gen. variances & determinants of moment matrices

- Variance expressed as the determinant of a moment matrix: $\mu_i = E[X^i]$

$$\mathrm{Var}[X] = E[(X-\mu)^2] = E[X^2] - E[X]^2 = \mu_2 - \mu_1^2 = \det\left(\begin{bmatrix} 1 & \mu \\ \mu & \mu_2 \end{bmatrix}\right) \geq 0$$

- Generalization of the variance to (d+1) RVs:

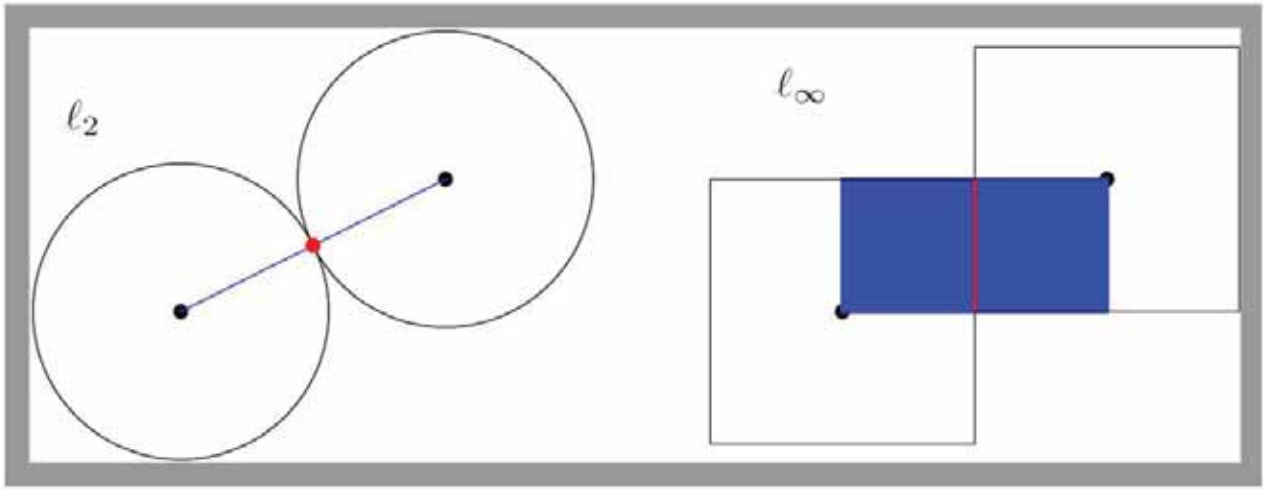
$$\boxed{X_1,,X_{d+1} \sim_{iid} F_X \Rightarrow E\left[\prod_{j>i}(X_i - X_j)^2\right] = (d+1)!\det(M_d) \geq 0}$$

- Moment matrix: $M_d := [\mu_{i+j-2}]_{ij} = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_d \\ \mu_1 & \mu_2 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ \mu_d & \cdots & \cdots & \mu_{2d} \end{bmatrix}$ $\det(M_d) \geq 0$

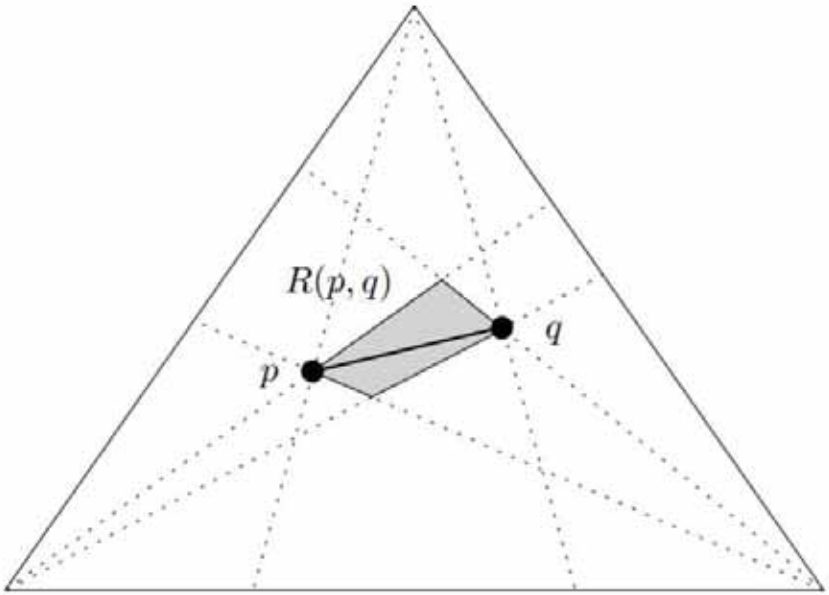
Lindsay, Bruce G. "On the determinants of moment matrices." *The Annals of Statistics* (1989)

(Non)-uniqueness of geodesics induced by convex norms

- Unique when the norm is smooth convex (eg., L2)
- Not-unique when the norm is polyhedral convex (eg., L1)



Smooth L2 norm **Polyhedral L∞ norm**



Hilbert log cross-ratio
(isometric to a polygonal normed vector space)

Clustering in Hilbert simplex geometry, https://arxiv.org/abs/1704.00454

∇-Geodesics: Boundary Value Problems (BVPs) versus Initial Value Problems (IVPs)

- On a manifold M equipped with an **affine connection ∇**, the geodesic is a straight line = **autoparallel smooth curve.** When the affine connection is the metric Levi-Civita connection, geodesics are locally minimizing length curves.
- Thus the geodesic equation is given by: $\boxed{\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0}$
- Geodesics with **initial value problems (IVPs)**: $\gamma(0) = p, \dot\gamma(0) = v (v \in T_p)$

  There exists a unique geodesic passing through p with tangent vector v at Tp.
- Geodesics with **boundary value problems (BVPs)**: geodesics passing through two points p and q: $\gamma(0) = p, \gamma(1) = q$
- Example: The **manifold of symmetric positive-definite matrices** (the SPD cone)

$$\boxed{\begin{aligned} P_{12}^{\mathrm{IV}}(u) &= P_1^{\frac{1}{2}}\mathrm{Exp}(u\ P_1^{-\frac{1}{2}} S P_1^{-\frac{1}{2}})P_1^{\frac{1}{2}} \\ P_{12}^{\mathrm{BV}}(t) &= P_1^{\frac{1}{2}}\mathrm{Exp}(t\ \mathrm{Log}(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}}))P_1^{\frac{1}{2}} \end{aligned}}$$

$$P(0) = P_1,\ P(1) = P_2,\ \dot P_1(0) = S \in \mathrm{Sym}(d, \mathbb{R})$$

$$\mathrm{Exp}(P) = \sum_{i=0}^{\infty} \frac{1}{i!} P^i$$

An elementary introduction to information geometry, arXiv:1808.08271

∇-Geodesics: Boundary Value Problems (BVPs) versus Initial Value Problems (IVPs)

- On a manifold M equipped with an **affine connection ∇**, the geodesic is a straight line = **autoparallel smooth curve.** When the affine connection is the metric Levi-Civita connection, geodesics are locally minimizing length curves.
- Thus the geodesic equation is given by: $\boxed{\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0}$
- Geodesics with **initial value problems (IVPs)**: $\gamma(0) = p, \dot\gamma(0) = v (v \in T_p)$

  There exists a unique geodesic passing through p with tangent vector v at Tp.
- Geodesics with **boundary value problems (BVPs)**: geodesics passing through two points p and q: $\gamma(0) = p, \gamma(1) = q$
- Example: The **manifold of symmetric positive-definite matrices** (the SPD cone)

$$\boxed{\begin{aligned} P_{12}^{\mathrm{IV}}(u) &= P_1^{\frac{1}{2}} \mathrm{Exp}(u\, P_1^{-\frac{1}{2}} S P_1^{-\frac{1}{2}}) P_1^{\frac{1}{2}} \\ P_{12}^{\mathrm{BV}}(t) &= P_1^{\frac{1}{2}} \mathrm{Exp}(t\, \mathrm{Log}(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})) P_1^{\frac{1}{2}} \end{aligned}}$$

$$P(0) = P_1,\ P(1) = P_2,\ \dot P_1(0) = S \in \mathrm{Sym}(d, \mathbb{R})$$

$$\mathrm{Exp}(P) = \sum_{i=0}^{\infty} \frac{1}{i!} P^i$$

An elementary introduction to information geometry, arXiv:1808.08271

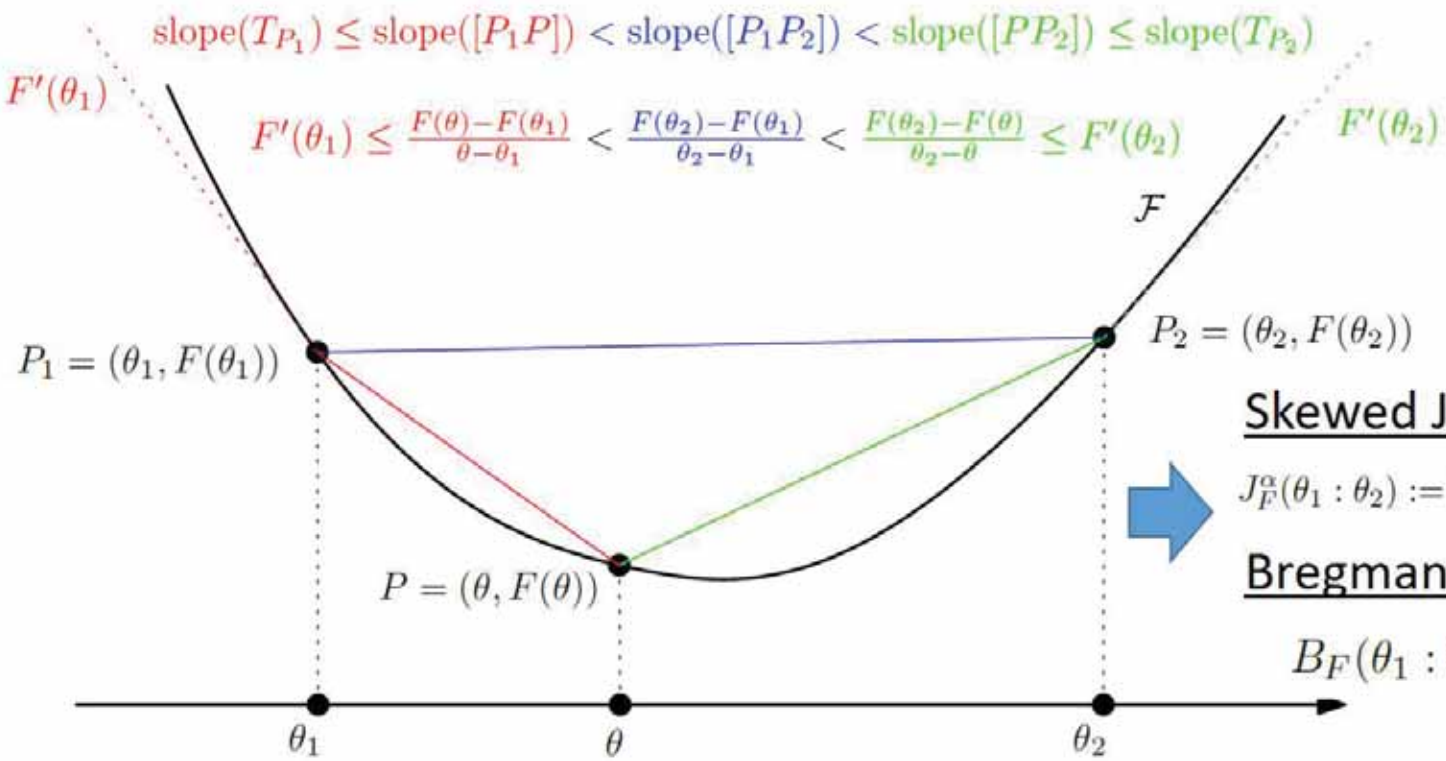
Geometric interpretations of the Jensen and Bregman divergences from the chordal slope lemma

For a strictly convex function F, the **chordal slope lemma** states that

$$\theta_1 < \theta < \theta_2 \quad \Rightarrow \quad \frac{F(\theta) - F(\theta_1)}{\theta - \theta_1} < \frac{F(\theta_2) - F(\theta_1)}{\theta_2 - \theta_1} < \frac{F(\theta_2) - F(\theta)}{\theta_2 - \theta}$$

Consequence for strictly convex and differentiable function F:

$$F'(\theta_1) \leq \frac{F(\theta_2) - F(\theta_1)}{\theta_2 - \theta_1} \leq F'(\theta_2)$$



<u>Skewed Jensen divergences:</u>

$$J_F^\alpha(\theta_1 : \theta_2) := (1-\alpha)F(\theta_1) + \alpha F(\theta_2) - F((1-\alpha)\theta_1 + \alpha\theta_2) > 0.$$

<u>Bregman divergences:</u>

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)F'(\theta_2).$$

Geometric Jensen-Shannon divergence[1]

- **Geometric Jensen-Shannon divergence**:

$$\mathrm{JS}^{G_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p:(pq)_\alpha^G\right) + \alpha\mathrm{KL}\left(q:(pq)_\alpha^G\right)$$

  for the G-**mixture**: $(pq)_\alpha^G(x) \propto G_\alpha(p(x), q(x))$, where geometric weighted mean $G_\alpha(a,b) = a^{1-\alpha}b^\alpha$ for $a, b > 0$ and $\alpha \in [0,1]$.

- When $p = p_{\theta_1}$ and $q = p_{\theta_2}$ belong to the same exponential family, $(p_{\theta_1}p_{\theta_2})_\alpha^G(x) = p_{(\theta_1\theta_2)_\alpha}(x)$ where $(\theta_1\theta_2)_\alpha := (1-\alpha)\theta_1 + \alpha\theta_2$.
- For multivariate normal distributions $p = N(m_1, \Sigma_1)$ and $q = N(m_2, \Sigma_2)$, the G-mixture is $r_\alpha = N(m_\alpha, \Sigma_\alpha)$ with $\Sigma_\alpha = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1}$ $m_\alpha = \Sigma_\alpha((1-\alpha)M_1^{-1}m_1 + \alpha M_2^{-1}m_2)$, and get the closed-form formula:

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p:q) &= (1-\alpha)\mathrm{KL}(p:r_\alpha) + \alpha\mathrm{KL}(q:r_\alpha), \\
\mathrm{KL}(p_1:p_2) &= \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta_m^\top\Sigma_2^{-1}\Delta_m + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\right\},
\end{aligned}$$

  with $\Delta_m = m_2 - m_1$. Ordinary JS divergence not in closed-form!

[1] `https://arxiv.org/abs/1904.04017`

(Information) Geometry of convex cones

- A cone in a vector space V yields a **dual cone** of positive linear functionals in the dual vector space V*:

$$\mathcal{C}^* := \{\omega \in V^* \ : \ \forall v \in C, \omega(v) \geq 0\}$$

- A cone is **homogeneous** if the automorphism group acts transitively on the cone
- On a homogeneous cone, define a **characteristic function :**

$$\chi_{\mathcal{C}}(\theta) := \int_{\mathcal{C}^*} \exp(-\omega(\theta)) \mathrm{d}\omega \geq 0.$$

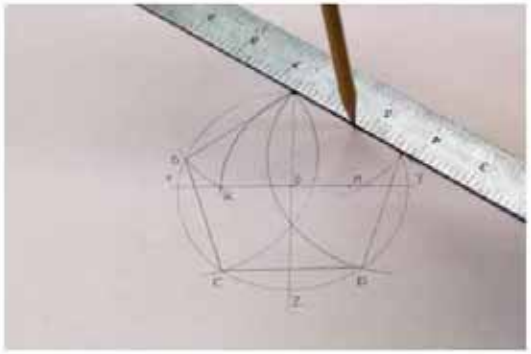
- The logarithm of the characteristic function is a Bregman generator which yields a dually flat space: Hessian geometry

Ernest Vinberg
(1937-2020)

Jean-Louis Koszul
(1921-2018)

- Vinberg, Theory of homogeneous convex cones, Trans. Moscow Math. Soc., 1967
- Koszul, Ouverts convexes homogenes des espaces affines, Mathematische Zeitschrift, 1962
- An elementary introduction to information geometry, arXiv:1808.08271
- On geodesic triangles with right angles in a dually flat space, arXiv:1910.03935

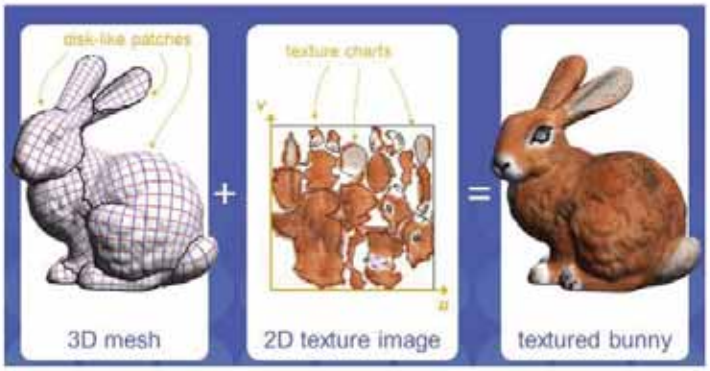
Figures in geometry



- Figures by construction with tools (eg., ruler and compass), synthetic geometry (no formula nor coordinates)

- Schematic figures by mental construction: A picture worth a thousand words, visualizing concepts, draw in your head!

- Visualization in charts/atlas: plot in local coordinate systems



Some geometries of matrix domains

- **Riemannian geometry** of the symmetric positive-definite (SPD) cone manifold

On approximating the Riemannian 1-center, Computational Geometry 46.1 (2013)

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

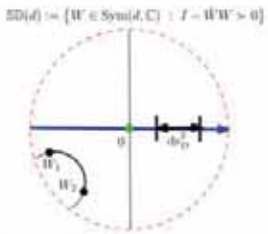
- **Dualistic information geometry** (-log det dually flat space)

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory 55.6 (2009)

$$F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log\pi$$

- **Hilbert geometry** of the space of correlation matrices (=elliptope)

Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices, Geometric Structures of Information, 2019



- **Symplectic geometry** of the Riemann matrices (Siegel domains)

The Siegel–Klein Disk: Hilbert Geometry of the Siegel Disk Domain, Entropy 22.9 (2020)

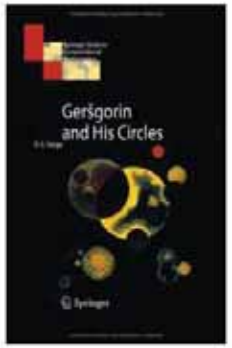
Complex matrix eigenvalues and Gershgorin circle theorem

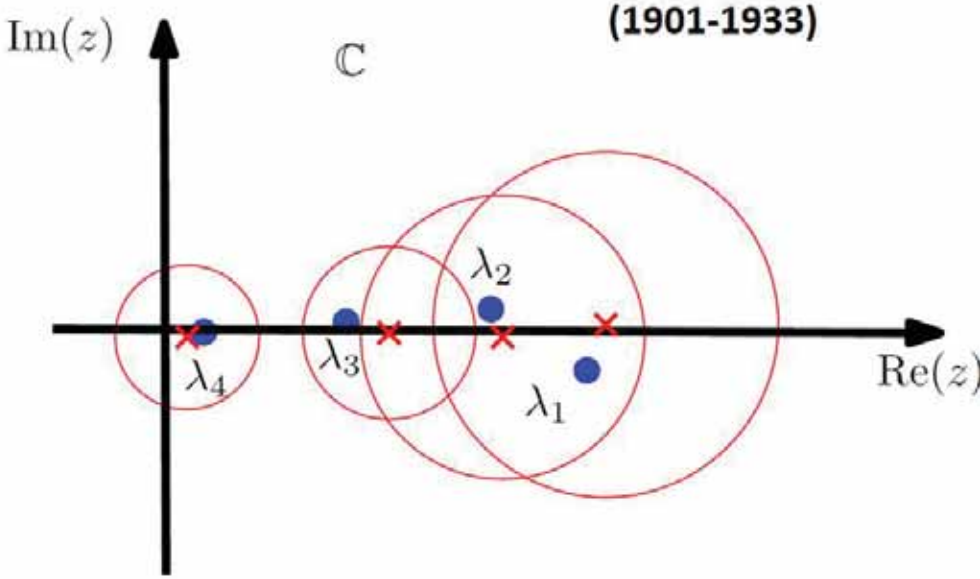
- Complex matrix: $M = [m_{ij}] \in \mathrm{Mat}_{\mathbb{C}}(d, d)$
- Eigenspectrum: $\Lambda(M) := \{\lambda_1(M), \ldots, \lambda_d(M)\}$
- Gershgorin circle: $\mathrm{Disk}(c, r) = \{z \in \mathbb{C} \ : \ |z - c| \leq r\}$

Semyon Aronovich Gershgorin
(1901-1933)

Theorem (Gershgorin, 1931):

$$\boxed{\Lambda(M) \subset \cup_{i=1}^{d} \mathrm{Disk}\left(m_{ii}, r_i = \sum_{j \neq i} |m_{ij}|\right)}$$

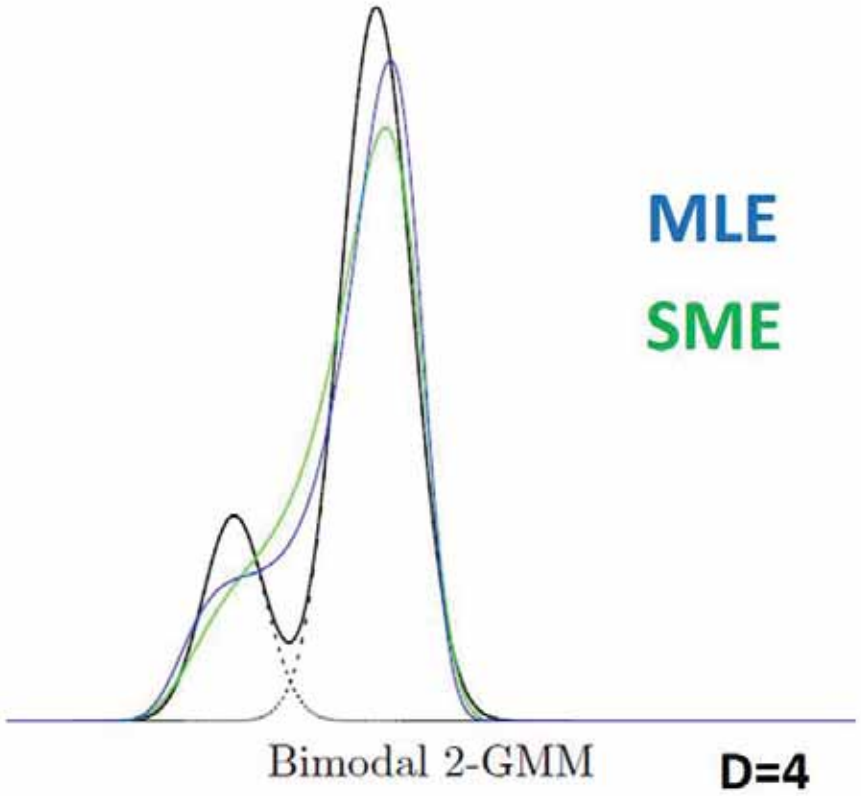


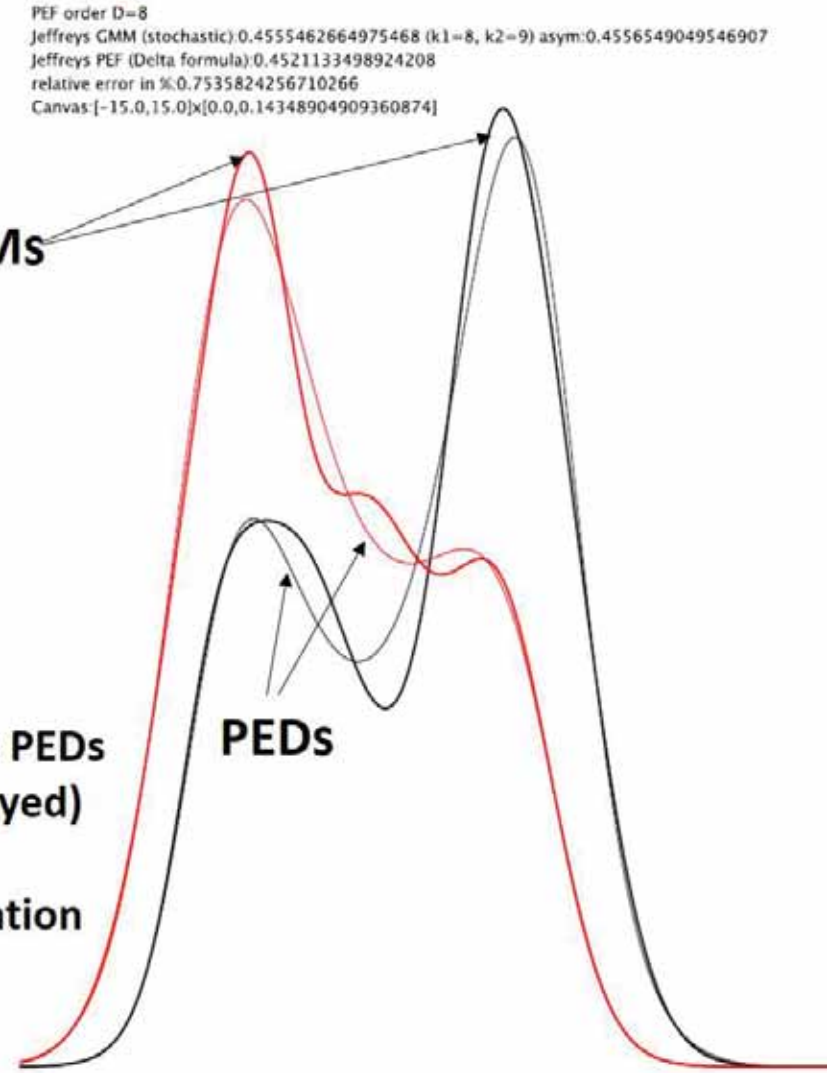


Converting GMMs to PEDs for Jeffreys approximation

A mixture is converted into a pair of quartic exponential family densities

$$m \rightarrow (\tilde{p}_{\bar{\theta}_{\mathrm{MLE}}}, \tilde{p}^{\bar{\eta}_{\mathrm{SME}}})$$





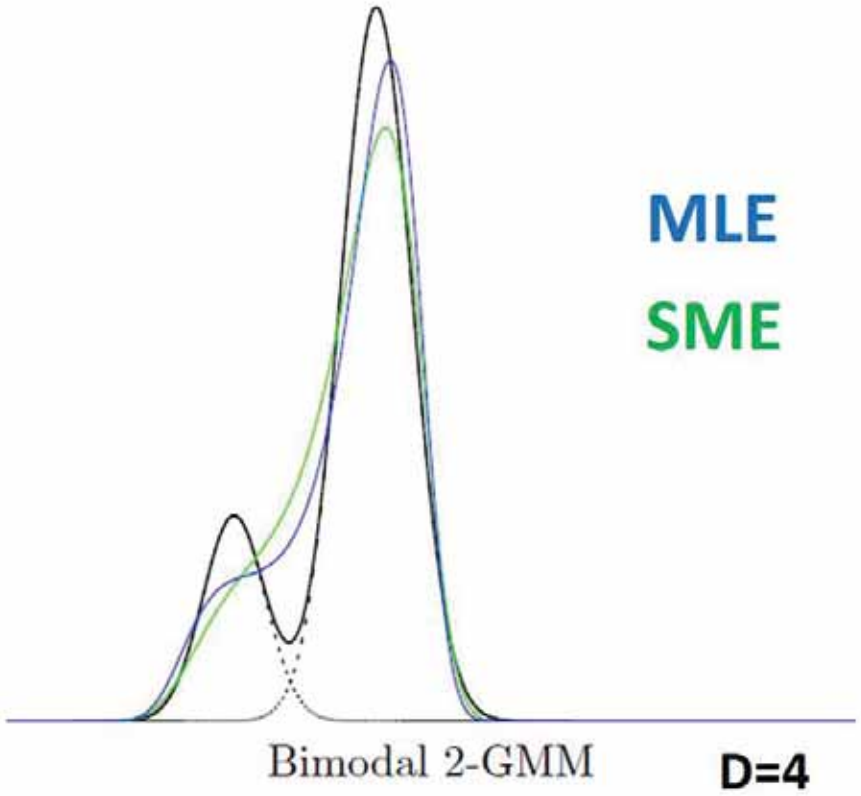
Convert two GMMs to PEDs (only SME PED displayed) 0.7% error In Jeffreys approximation

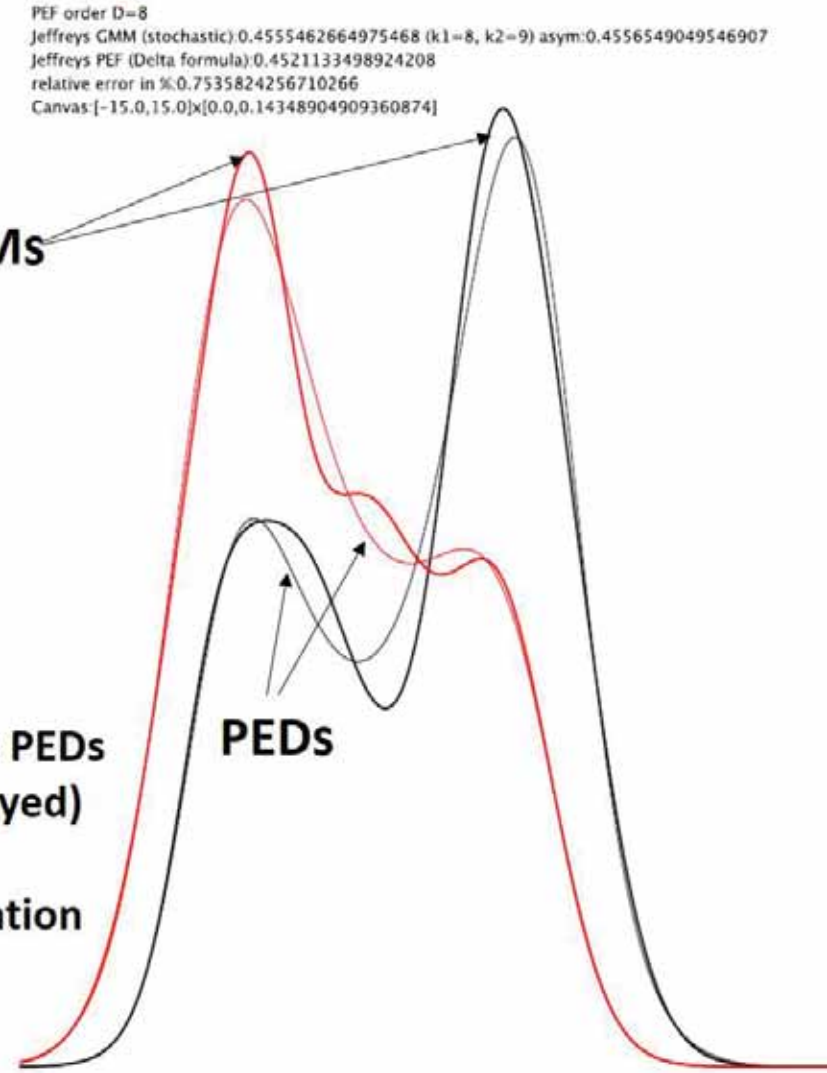
$$D = 8$$

Converting GMMs to PEDs for Jeffreys approximation

A mixture is converted into a pair of quartic exponential family densities

$$m \rightarrow (\tilde{p}_{\bar{\theta}_{\mathrm{MLE}}}, \tilde{p}^{\bar{\eta}_{\mathrm{SME}}})$$





Convert two GMMs to PEDs (only SME PED displayed) 0.7% error In Jeffreys approximation

$$D = 8$$

Gradient and Hessian on a manifold **(M,g,∇)**

Directional derivative of f at point x in direction of vector v:

$$Df(x)[v] = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$$

Gradient (requires metric tensor g): unique vector $\mathrm{grad}_p f(x)$ satisfying

$$\langle \mathrm{grad}_p f(x), v\rangle = Df(x)[v], \quad \forall v \in T_pM$$

Hessian (requires an affine connection, usually Levi-Civita metric conn.)

$$\mathrm{hess}_p f(x)[v] = \nabla_v^{\mathrm{LC}} \mathrm{grad}_p f(x), \quad \forall v \in T_pM$$

Property: $\langle \mathrm{hess}_p f(v)[v], w\rangle_p = \langle v, \mathrm{hess}_p f(x)[w]\rangle_p, \forall v, w \in T_pM$

@FrnkNlsn

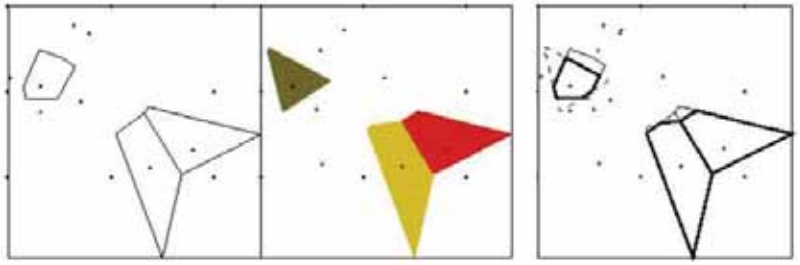
https://arxiv.org/abs/1808.08271

Group action on a set

- Group $G = \{g_1, g_2, \ldots\}$ with identity element e, composition operation '.' (group law), and inverse element g^{-1}
- Set of points $X = \{x_1, x_2, \ldots, \}$
- Action of G on X: a set of **one-to-one maps** $A_g : X \to X$ such that
  - $A_{g_2} \circ A_{g_1} = A_{g_2}(A_{g_1}(x)) = A_{g_2 . g_1}(x)$ for all $x \in X$
  - $A_{g^{-1}}(x) = (A_g)^{-1}(x)$ for all $x \in X$
  - $A_e(x) = \mathrm{Id}_X(x)$ for all $x \in X$
- Informally, group $(G, .)$ is realized by a set of one-to-one mappings with function composition following the group law

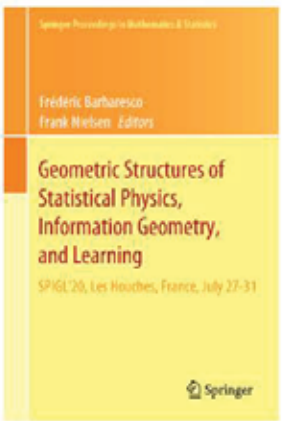
Paradigms to get output-sensitive algorithms

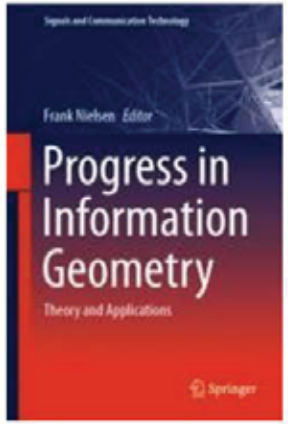
- An **output-sensitive algorithm** is an algorithm whose complexity depends on the combinatorial size of the output. The output size can widely range in computational geometry: e.g., 3D triangulations, union of rectangles, etc.
- The **marriage-before-conquer** is an output-sensitive paradigm which differs from the **divide-and-conquer** paradigm by merging the solutions of the subproblems **before** recursively solving the subproblems: The advantage is to reduce the subproblem sizes according to the merged solution (eg, convex hull)
- The **grouping-and-querying paradigm** consists in partitioning the input into groups, solving the problem on the groups in a non-output-sensitive way, and building the solution in output-sensitive way by **iteratively querying the solutions of the groups**.

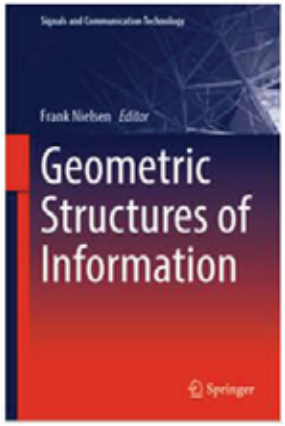


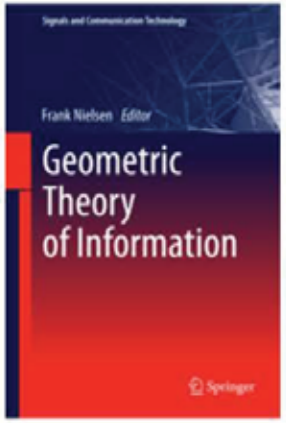
Computing a few Voronoi cells

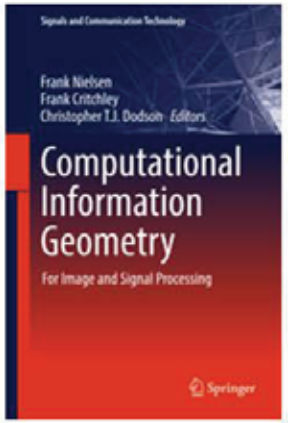
Grouping and querying: A paradigm to get output-sensitive algorithms, JCDCG 1998

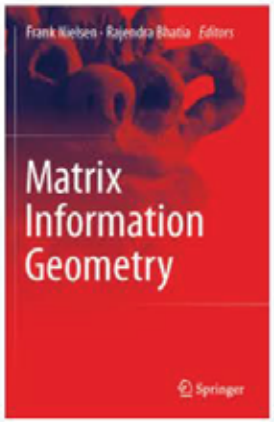

Frédéric Barbaresco
Frank Nielsen Editors
Geometric Structures of Statistical Physics, Information Geometry, and Learning
SPIGL'20, Les Houches, France, July 27-31
Springer

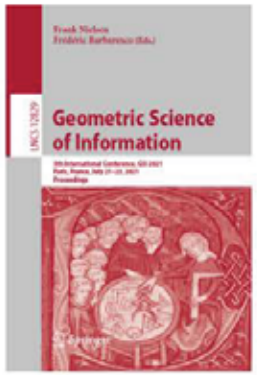

Signals and Communication Technology
Frank Nielsen Editor
Progress in Information Geometry
Theory and Applications
Springer

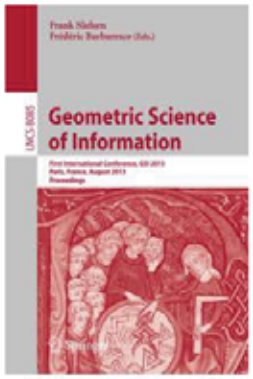

Signals and Communication Technology
Frank Nielsen Editor
Geometric Structures of Information
Springer


Signals and Communication Technology
Frank Nielsen Editor
Geometric Theory of Information
Springer

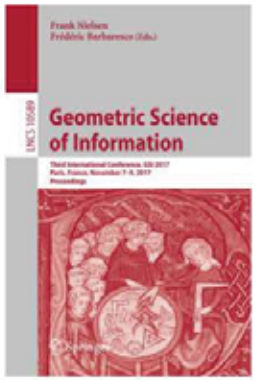

Signals and Communication Technology
Frank Nielsen
Frank Critchley
Christopher T.J. Dodson Editors
Computational Information Geometry
For Image and Signal Processing
Springer

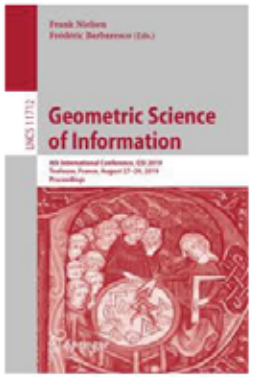

Frank Nielsen · Rajendra Bhatia Editors
Matrix Information Geometry
Springer

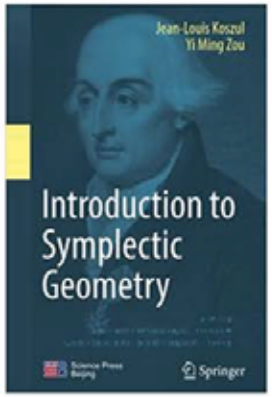

Geometric Science of Information

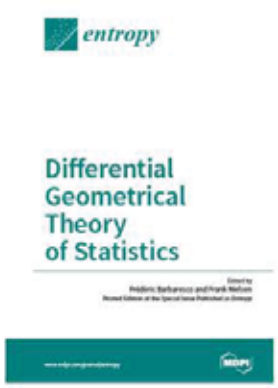

Geometric Science of Information

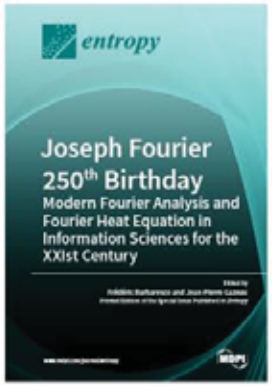
Geometric Science of Information

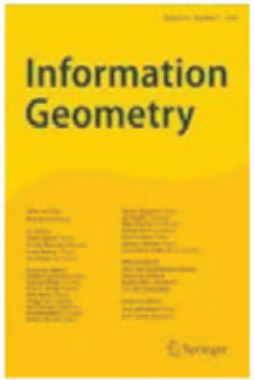

Geometric Science of Information


Geometric Science of Information


Jean-Louis Koszul
Yi Ming Zou
Introduction to Symplectic Geometry
Springer


entropy
Differential Geometrical Theory of Statistics


entropy
Joseph Fourier 250th Birthday
Modern Fourier Analysis and Fourier Heat Equation in Information Sciences for the XXIst Century


Information Geometry

Hadamard manifolds (complete CAT(0)-spaces)

Jacques Hadamard
1865-1963

Definition: Riemannian manifold (M,g) is a Hadamard manifold if:

- **Complete** (geodesics extend to infinity)
- **Simply connected**
- **Negative sectional curvature**

Properties of Hadamard manifolds:

- Riemannian exponential map = **diffeomorphism** [Cartan-Hadamard theorem]
- **Unique geodesic** between any two points
- **Geodesic balls** are **convex** wrt **geodesic convex combination $\#_t$**: $p\#_t q = \gamma_{pq}(t)$

$$\forall p, q \in \mathrm{Ball}(c, r),\ \forall t \in [0, 1],\ p\#_t q \in \mathrm{Ball}(c, r)$$

Some examples:

- Hyperbolic space/ Symmetric Positive-Definite matrix cone (+Hermitian PD cone)
- Siegel domains (Siegel ball, Siegel upper space)

On approximating the Riemannian 1-center, Computational Geometry 46.1 (2013)
The Siegel–Klein Disk: Hilbert Geometry of the Siegel Disk Domain, Entropy 22.9 (2020)

Hadamard spaces = Complete CAT(0) spaces

- Metric space (X,d) is **geodesic** iff exists a mapping γ (not necessarily unique):

$$\boxed{\gamma(0) = p, \gamma(1) = q, \quad \forall s, t \in [0,1], \quad d(\gamma(s), \gamma(t)) = |s-t|\, d(p,q)}$$

- **CAT(0)** = geodesic metric space (X,d) which satisfies (→ unique geodesics):

$$\forall x \in X, \quad d(x, \gamma(t))^2 \leq (1-t)d(x, \gamma(0))^2 + t d(x, \gamma(1))^2 - t(1-t)d(\gamma(0), \gamma(1))^2$$

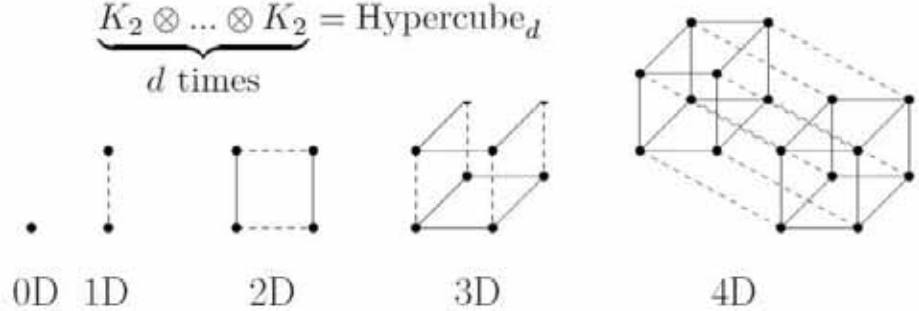
- **Hadamard space** = complete CAT(0) space
- In Hadamard spaces, we have the following **4-point inequality**:

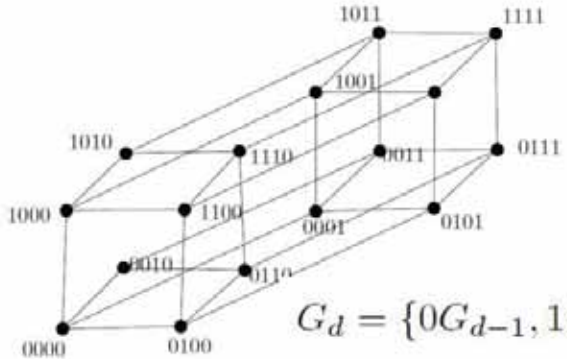
$$\forall x, y, u, v \in X, \quad d(x,y)^2 + d(u,v)^2 \leq d(x,v)^2 + d(y,u)^2 + 2d(x,u)d(y,v)$$

- Examples of Hadamard spaces: Euclidean spaces, hyperbolic spaces, complete simply connected Riemannian manifolds of nonpositive sectional curvature(SPD manifold), Euclidean buildings, locally compact R-trees, CAT(0) complexes, BHV tree space, etc.

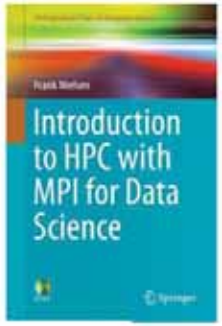
Hamming finite metric geometry and the hypercube

$$\underbrace{K_2 \otimes \ldots \otimes K_2}_{d \text{ times}} = \text{Hypercube}_d$$





$$G_d = \{0G_{d-1}, 1G_{d-1}^{\text{mirrored}}\}, \quad G_{-1} = \emptyset.$$



Hamming distance is a **metric** distance on binary strings

$$\boxed{\text{Hamming}(P, Q) = \sum_{i=0}^{d-1} 1_{P_i \neq Q_i}}$$

$G(1)$		$G(2)$		$G(3)$
	mirror	prefix	mirror	prefix
0	0	00	00	000
1	1	01	01	001
	1	11	11	011
	0	10	10	010
			10	110
			11	111
			01	101
			00	100

Gray code

- The *complement* of a Hamming ball is a Hamming ball
- Computing the *smallest enclosing Hamming ball* is **NP-hard** (= circumcenter of sets of strings)
- **Gray code** on the hypercube for network routing

Hierarchical clustering, in "Introduction to HPC with MPI for Data Science", Springer 2016

Hamming and Lee metric distances

- Consider a finite alphabet A of d letters {0,…,d-1} and words w and w' of n letters
- **Hamming distance**: $D_H(w,w') = \sum_{i=1}^{n} \delta(w_i, w'_i), \quad \delta(a,b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{when } a \neq b \end{cases}$
- For binary words Hamming distance amount to a XOR: $D_H(b,b') = w \text{ XOR } w'$
- **Lee distance**: $D_L(w,w') = \sum_{i=1}^{n} \min\left(|w_i - w'_i|, d - |w_i - w'_i|\right)$
- Both Hamming and Lee distances are **metric** distances.
- Hamming and Lee distances coincide when d=2 or d=3

Harmonic Jensen-Shannon divergence[1]

- **Harmonic Jensen-Shannon divergence**:

$$\mathrm{JS}^{H_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p:(pq)_\alpha^H\right) + \alpha\mathrm{KL}\left(q:(pq)_\alpha^H\right)$$

  for the H-**mixture**: $(pq)_\alpha^H(x) \propto H_\alpha(p(x), q(x))$, where harmonic weighted mean $H_\alpha(a,b) = \frac{ab}{\alpha a + (1-\alpha)b}$ for $a, b > 0$ and $\alpha \in [0,1]$.

- When $p = p_{\gamma_1}$ and $q = p_{\gamma_2}$ belong to the scale Cauchy family, $(p_{\gamma_1} p_{\gamma_1})_\alpha^H(x) = p_{(\gamma_1\gamma_2)_\alpha}(x)$ where $(\gamma_1\gamma_2)_\alpha := (1-\alpha)\gamma_1 + \alpha\gamma_2$.

$$\begin{aligned}
\mathrm{JS}^H(p:q) &= \frac{1}{2}\left(\mathrm{KL}\left(p:(pq)_{\frac{1}{2}}^H\right) + \mathrm{KL}\left(q:(pq)_{\frac{1}{2}}^H\right)\right), \\
\mathrm{JS}^H(p_{\gamma_1}:p_{\gamma_2}) &= \frac{1}{2}\left(\mathrm{KL}\left(p_{\gamma_1}:p_{\frac{\gamma_1+\gamma_2}{2}}\right) + \mathrm{KL}\left(p_{\gamma_2}:p_{\frac{\gamma_1+\gamma_2}{2}}\right)\right), \\
&= \log\frac{(3\gamma_1+\gamma_2)(3\gamma_2+\gamma_1)}{8\sqrt{\gamma_1\gamma_2}(\gamma_1+\gamma_2)}.
\end{aligned}$$

- Ordinary JS divergence for scale Cauchy distributions not in closed-form!

[1] https://arxiv.org/abs/1904.04017

A generalization of Hartigan k-means heuristic: The merge-and-split heuristic

- **K-Means** minimize the average squared length distance of points to their closest centers (cluster centroids):

$$e(\mathcal{P},\mathcal{K}) = \sum_{i=1}^{n} \min_{j=1}^{k} D(p_i, c_j) = \sum_{i=1}^{n} D(p_i, c_{l_i}) = \sum_{j=1}^{k} \sum_{p \in \mathcal{K}_j} D(p, c_j)$$

- K-means loss is **NP-hard** when k>1 and d>1, *polynomial* when d=1
- **Hartigan's swap heuristic**: move a point to another cluster if the loss decreases: always guarantee *same number of clusters.*
- **Lloyd's batched heuristic:** may end up with empty clusters
- **Merge-and-split heuristic**: merge two clusters Ci and Cj, and split them according to new centers c'_i and c'_j. (e.g. , use 2-means++ on Ci cup Cj)

Accept when difference of loss is **positive**: $\boxed{\Delta(\mathcal{C}_i,\mathcal{C}_j) = e_1(\mathcal{C}_i, c_i) + e_1(\mathcal{C}_j, c_j) - (e_1(\mathcal{C}'_i, c'_i) + e_1(\mathcal{C}'_j, c'_j))}$

Further heuristics for k-means: The merge-and-split heuristic and the (k,l)-means, arxiv 1406.6314
Optimal interval clustering: Application to Bregman clustering and statistical mixture learning, IEEE SPL 2014
Hartigan's method for k-MLE: Mixture modeling with Wishart distributions and its application to motion retrieval, GTI, 2014

A generalization of Hartigan k-means heuristic: The split-and-merge heuristic

- **K-Means** minimize the average squared length distance of points to their closest centers (cluster centroids):

$$e(\mathcal{P},\mathcal{K}) = \sum_{i=1}^{n} \min_{j=1}^{k} D(p_i, c_j) = \sum_{i=1}^{n} D(p_i, c_{l_i}) = \sum_{j=1}^{k} \sum_{p \in \mathcal{K}_j} D(p, c_j)$$

- K-means loss is **NP-hard** when k>1 and d>1, *polynomial* when d=1
- **Hartigan's swap heuristic**: move a point to another cluster if the loss decreases: always guarantee *same number of clusters.*
- **Lloyd's batched heuristic:** may end up with empty clusters
- **Merge-and-split heuristic**: merge two clusters Ci and Cj, and split them according to new centers c'_i and c'_j. (e.g. , use 2-means++ on Ci cup Cj)

Accept when difference of loss is **positive**: $\boxed{\Delta(\mathcal{C}_i,\mathcal{C}_j) = e_1(\mathcal{C}_i, c_i) + e_1(\mathcal{C}_j, c_j) - (e_1(\mathcal{C}'_i, c'_i) + e_1(\mathcal{C}'_j, c'_j))}$

Further heuristics for k-means: The merge-and-split heuristic and the (k,l)-means, arxiv 1406.6314

Optimal interval clustering: Application to Bregman clustering and statistical mixture learning, IEEE SPL 2014

Hartigan's method for k-MLE: Mixture modeling with Wishart distributions and its application to motion retrieval, GTI, 2014

A generalization of Hartigan k-means heuristic: The split-and-merge heuristic

- **K-Means** minimize the average squared length distance of points to their closest centers (cluster centroids):

$$e(\mathcal{P},\mathcal{K}) = \sum_{i=1}^{n} \min_{j=1}^{k} D(p_i, c_j) = \sum_{i=1}^{n} D(p_i, c_{l_i}) = \sum_{j=1}^{k} \sum_{p \in \mathcal{K}_j} D(p, c_j)$$

- K-means loss is **NP-hard** when k>1 and d>1, *polynomial* when d=1
- **Hartigan's swap heuristic**: move a point to another cluster if the loss decreases: always guarantee *same number of clusters.*
- **Lloyd's batched heuristic:** may end up with empty clusters
- **Split-and-merge heuristic**: merge two clusters Ci and Cj, and split them according to new centers c'_i and c'_j. (e.g. , use 2-means++ on Ci cup Cj)

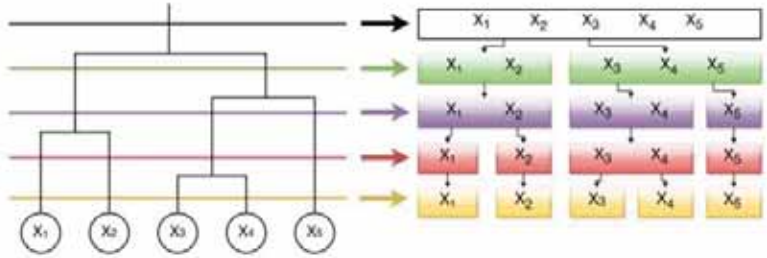
Accept when difference of loss is **positive**: $\Delta(\mathcal{C}_i, \mathcal{C}_j) = e_1(\mathcal{C}_i, c_i) + e_1(\mathcal{C}_j, c_j) - (e_1(\mathcal{C}'_i, c'_i) + e_1(\mathcal{C}'_j, c'_j))$

Further heuristics for k-means: The merge-and-split heuristic and the (k,l)-means, arxiv 1406.6314

Optimal interval clustering: Application to Bregman clustering and statistical mixture learning, IEEE SPL 2014

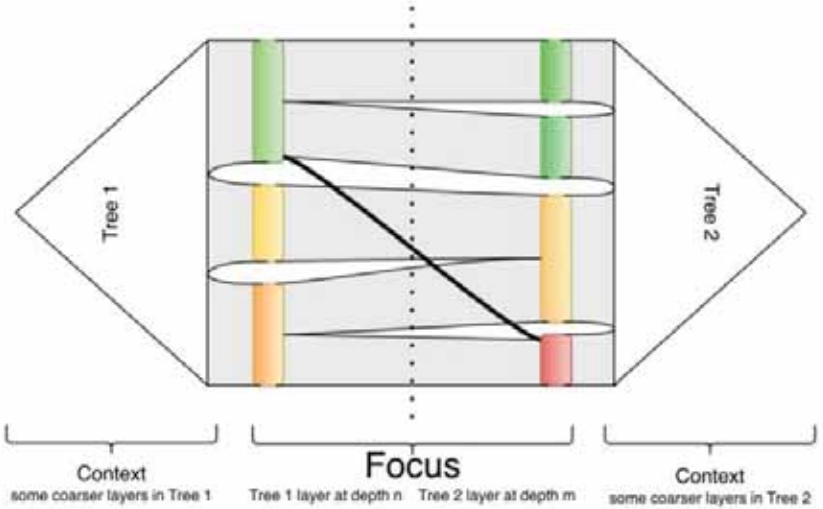
Hartigan's method for k-MLE: Mixture modeling with Wishart distributions and its application to motion retrieval, GTI, 2014

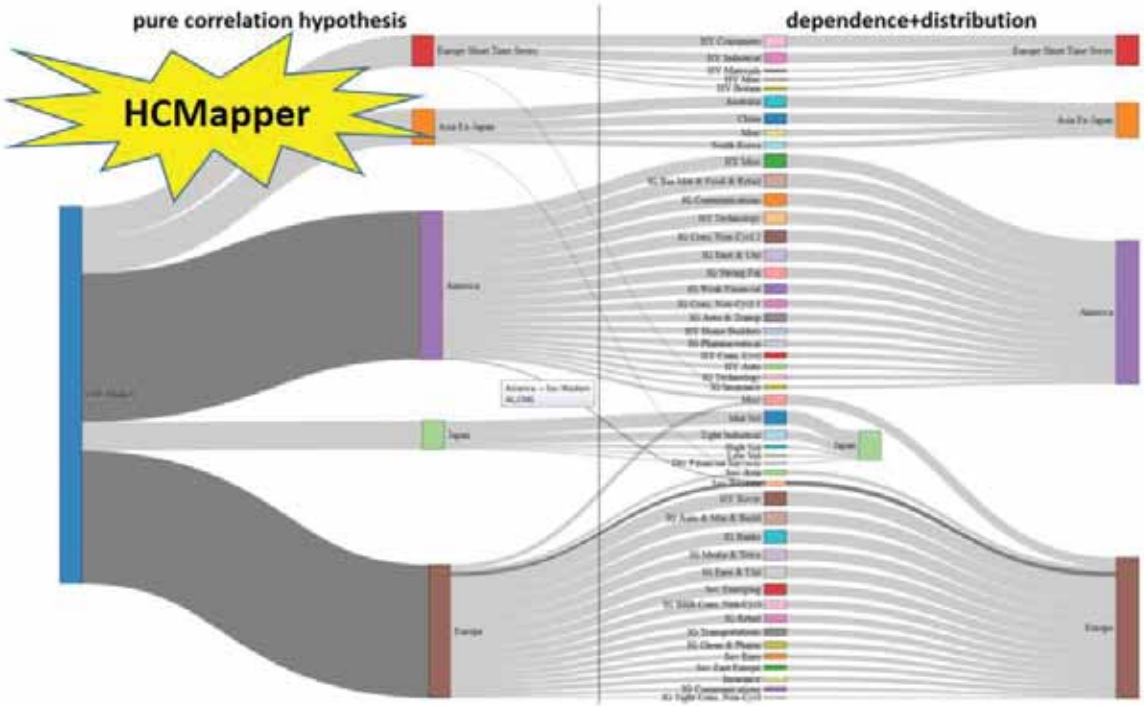
HCMapper: Visualization tool for comparing dendrograms



Hierarchical clustering (dendrograms) → **flat clustering (partitions)**

Hypothesis 1 — Hypothesis 2





Sankey diagram

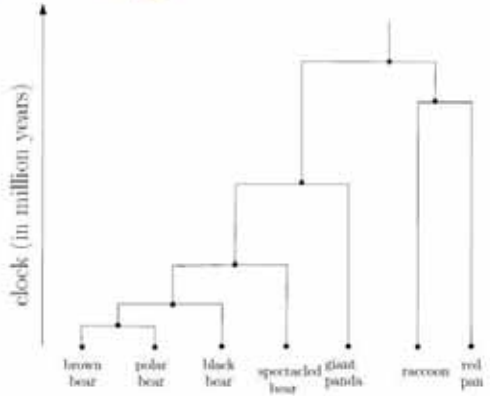
Compare two dendrograms on the same set by displaying multiscale partition-based layered structures

- **HCMapper: An interactive visualization tool to compare partition-based flat clustering extracted from pairs of dendrograms**

Gautier Marti, Philippe Donnat, Frank Nielsen, Philippe Very. https://arxiv.org/abs/1507.08137

- **Hierarchial custering, Introduction to HPC with MPI for Data Science, Springer 2016**

Phylogenetic trees from hierarchical clustering

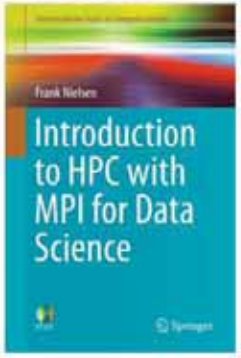


- **Ultrametric distance**: strengthens triangle inequality of metrics with $\boxed{D(x,y) \leq \max_z(D(x,z), D(z,y))}$
- **Additive tree**: tree with *real-valued edges* so that the distance between two leaves is the sum of edge values linking them.
- **Ultrametric tree**: the common ancestor k of two leaves i and j satisfies $D_{i,k} = D_{j,k}$

Visualize the dendrogram of an ultrametric tree by letting the height function $\frac{1}{2}D_{i,j}$

- **UPGMA**= Unweighted Pair Group Method using arithmetic Averages algorithm (= Hierarchical clustering with *group average linkage*)

> Theorem
>
> When the matrix distance $M = [D_{i,j}]_{i,j}$ with $D_{i,j} = D(x_i, x_j)$ of a data-set X satisfies the ultra-metric property, then there exists a unique ultra-metric tree that can be built with the UPGMA algorithm.



Hierarchical clustering, in "Introduction to HPC with MPI for Data Science", Springer 2016

Upper bounding the differential entropy (of mixtures)

Idea: compute the differential entropy of a **MaxEnt exponential family** with **given sufficient statistics** in closed form. *Any other* distribution has less entropy for the same moment expectations. Applies to statistical mixtures.

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

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

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

Legendre-Fenchel conjugate

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

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

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

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

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

@FrnkNlsn

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

Symmetrized α-divergences and Heinz means

α-divergences on **positive arrays** (not necessarily normalized probabilities):

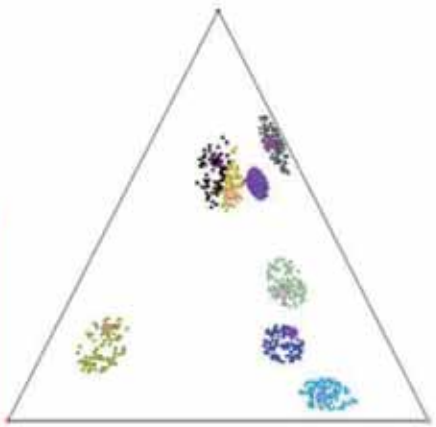
$$D_\alpha(p:q) := \sum_{i=1}^d \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\tilde{p}^i + \frac{1+\alpha}{2}\tilde{q}^i - \left(\tilde{p}^i\right)^{\frac{1-\alpha}{2}}\left(\tilde{q}^i\right)^{\frac{1+\alpha}{2}}\right)$$

$$D_\alpha(\tilde{p}:\tilde{q}) = D_{-\alpha}(\tilde{q}:\tilde{p})$$

Symmetrized α-divergences: $J_\alpha(\tilde{p},\tilde{q}) = \frac{1}{2}(D_\alpha(\tilde{p}:\tilde{q}) + D_\alpha(\tilde{q}:\tilde{p})) = J_{-\alpha}(\tilde{p},\tilde{q})$

also called **Jeffreys α-divergences** between categorical distributions:

$$J_\alpha(p:q) = \frac{4}{1-\alpha^2}\left(1 + \sum_{i=1}^d H_{\frac{1-\alpha}{2}}\left(p^i : q^i\right)\right)$$



Heinz mean:

Interpolate between arithmetic and geometric means

$$H_\beta(a,b) = \frac{a^\beta b^{1-\beta} + a^{1-\beta} b^\beta}{2}$$

On Clustering Histograms with *k*-Means by Using Mixed *α*-Divergences, Entropy 2014

Generalized Hellinger distances and statistical Minkowski distances

Jeffreys (1946) considered the **generalized Hellinger distances**:

An invariant form for the prior probability in estimation problems

By Harold Jeffreys, F.R.S.

(*Received* 23 *November* 1945)

$$\int |(dP')^m - (dP)^m|^{1/m} \qquad (2)$$

Hellinger distance, generalized Hellinger distance:

$$D_{H,2}[p:q] = \sqrt{\frac{1}{2}}\sqrt{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)} \Rightarrow D_{H,\alpha}[p:q] = \left(\frac{1}{2}\int \left(p(x)^{\frac{1}{\alpha}} - q(x)^{\frac{1}{\alpha}}\right)^{\alpha} \mathrm{d}\mu(x)\right)^{\frac{1}{\alpha}}$$

Minkowski statistical distances from Minkowski inequality:

Minkowski inequality:

$$\|p+q\|_\alpha \leq \|p\|_\alpha + \|q\|_\alpha$$

$$\left(\int |p(x)+q(x)|^\alpha \mathrm{d}\mu(x)\right)^{\frac{1}{\alpha}} \leq \left(\int |p(x)|^\alpha \mathrm{d}\mu(x)\right)^{\frac{1}{\alpha}} + \left(\int |q(x)|^\alpha \mathrm{d}\mu(x)\right)^{\frac{1}{\alpha}}$$

$$\Rightarrow$$

$$D_\alpha(p,q) := \|p\|_\alpha + \|q\|_\alpha - \|p+q\|_\alpha \geq 0.$$

$$L_\alpha(p,q) := -\log \frac{\|p+q\|_\alpha}{\|p\|_\alpha + \|q\|_\alpha} = \log \frac{\|p\|_\alpha + \|q\|_\alpha}{\|p+q\|_\alpha} \geq 0.$$

The Statistical Minkowski Distances: Closed-Form Formula for Gaussian Mixture Models, GSI 2019, arxiv 1901.03732

Heron α-means

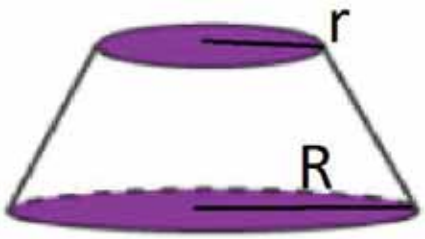
- **Heron α-means** interpolate between the geometric mean $G(a,b) := \sqrt{ab}$ and the arithmetic means $A(a,b) := \dfrac{a+b}{2}$

$$H_\alpha(a,b) := (1-\alpha)G(a,b) + \alpha A(a,b)$$

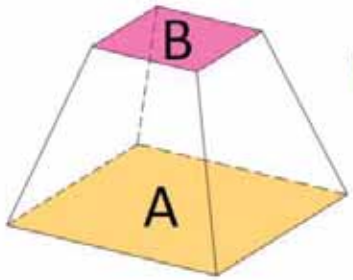
$$\boxed{H_\alpha(a,b) := (1-\alpha)\sqrt{ab} + \alpha\frac{a+b}{2}}$$

Heron of Alexandria
circa 10-70 AD

- **Heron 2/3-means** $H_{\frac{2}{3}}(a,b) = \dfrac{1}{3}(\sqrt{ab} + a + b)$ is used to calculate the **volume of the frustrum** of a cone C or a pyramid P:



$$\begin{aligned} V_{\mathcal{C}} &= hH_{\frac{2}{3}}(\pi r^2, \pi R^2) \\ &= \frac{\pi h}{3}(r^2 + R^2 + rR) \end{aligned}$$



$$\begin{aligned} V_{\mathcal{P}} &= hH_{\frac{2}{3}}(A,B) \\ &= \frac{h}{3}(A + B + \sqrt{AB}) \end{aligned}$$

Bregman manifolds are (special) Hessian manifolds

- Hessian metric tensor field induced from convex potential F(θ):

$$\boxed{g = \partial_i \partial_j F(\theta) \mathrm{d}\theta^i \otimes \mathrm{d}\theta^j} \quad \text{where} \quad \partial_i := \frac{\partial}{\partial \theta^i}$$

- Legendre transformation yields dual convex potential F*(η):

$$F(\theta) + F^*(\eta) = \theta^i \eta_j$$

Hessian metric induced from the dual convex potential F*(η):

$$\boxed{g = \partial^i \partial^j F^*(\eta) \mathrm{d}\eta_i \otimes \mathrm{d}\eta_j} \quad \text{where} \quad \partial^i := \frac{\partial}{\partial \eta_i}$$

- **Mixed (θ,η) representation** of the metric tensor: $\boxed{g = \mathrm{d}\theta^i \otimes \mathrm{d}\eta_j}$
- Inner product wrt coordinates: $g(v, w) = g_{ij} v^i w^j = v_i w^j = g^{ij} v_i w_j = v^i w_j$

On Geodesic Triangles with Right Angles in a Dually Flat Space, Progress in Information Geometry, Springer 2021

Hessian Fisher Information: Mixed parameterization

- **Fisher Information Matrix** (FIM) of a parametric model is said **Hessian** when it can be expressed as a Hessian of a strictly convex potential function:

$$I_\theta(\theta) = E_{p_\theta}\left[\frac{\partial}{\partial\theta_i}\log p_\theta(x)\frac{\partial}{\partial\theta_j}\log p_\theta(x)\right] = \nabla^2_\theta F(\theta)$$

examples of **HFIM**: exponential families, mixtures families, etc. [**1803.07225**]

- Dual convex conjugate potential function F* yields dual coordinate system η :

$$I_\eta(\eta) = \nabla^2 F^*(\eta), \quad \eta = \nabla F(\theta), \quad \theta = \nabla F^*(\eta)$$

- **Crouzeix identity** (meaning θ-basis is the reciprocal basis of η-basis [arXiv:1808.08271]):

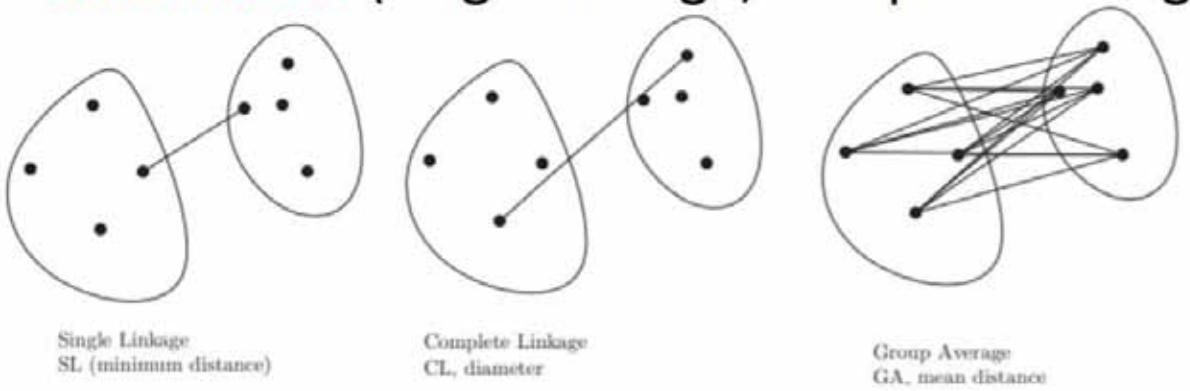
$$\nabla^2_\theta F(\theta)\nabla^2_\eta F^*(\eta) = I$$

- **Mixed parameterization** by choosing l coordinates from θ and D-l coordinates from η : HFIM becomes **block diagonal**. **In 2D, we can always diagonalize HFIM:** HFIM wrt (θ_1, η_2) or wrt (η_1, θ_2) is always diagonal!

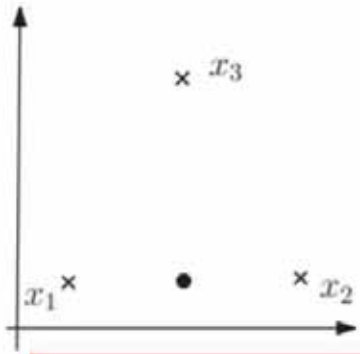
Ke Sun, FN, Relative Fisher information and natural gradient for learning large modular models, ICML 2017
An elementary introduction to information geometry, arXiv:1808.08271
Monte Carlo information-geometric structures, Geometric Structures of Information, 2019 (arXiv:1803.07225)

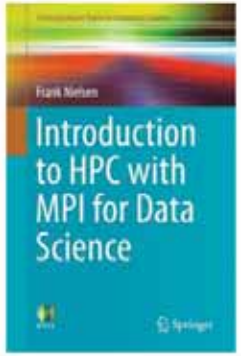
Inversion in hierarchical clustering with Ward criterion

- Agglomerative hierarchical clustering with **base distance D** and **linkage distance Δ** (single linkage, complete linkage, group average linkage, etc.)



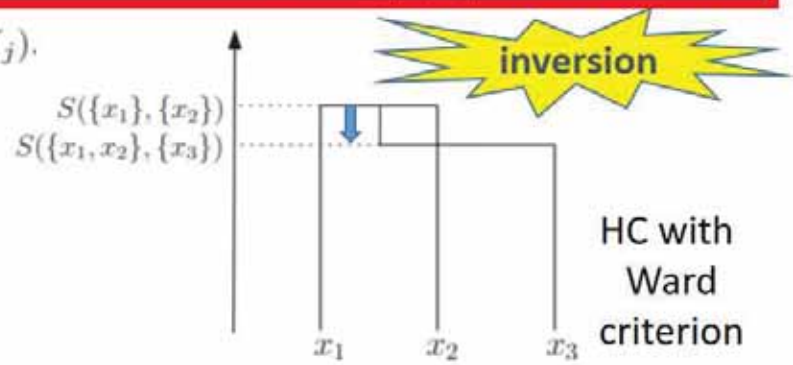
$$\Delta(X_i, X_j) = \min_{x_i \in X_i, x_j \in X_j} D(x_i, x_j) \qquad \Delta(X_i, X_j) = \max_{x_i \in X_i, x_j \in X_j} D(x_i, x_j) \qquad \Delta(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{x_i \in X_i} \sum_{x_j \in X_j} D(x_i, x_j)$$





Ward: $\Delta(X_i, X_j) = \frac{n_i n_j}{n_i + n_j} \|c(X_i) - c(X_j)\|^2$

- Tree structure called a **dendrogram** $S(X_i, X_j) = -\Delta(X_i, X_j)$.
- Ward's criterion based on subset **centroids** c(.):
- **Inversion** when there exists a path from a leaf to the root with **non monotonous height function S**



HC with Ward criterion

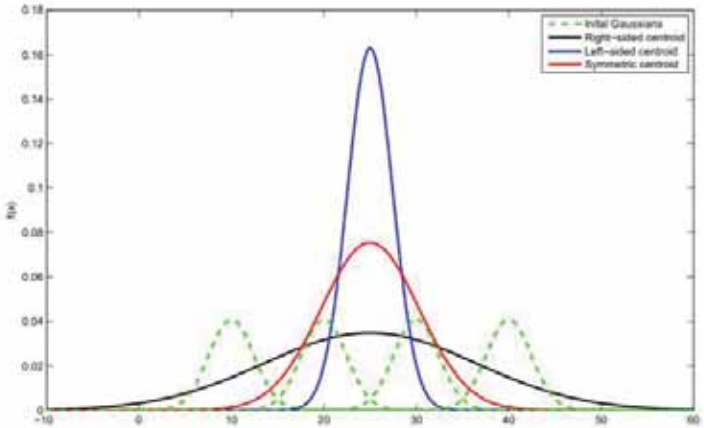
Hierarchical clustering, in "Introduction to HPC with MPI for Data Science", Springer 2016

Hierarchical mixtures of exponential families

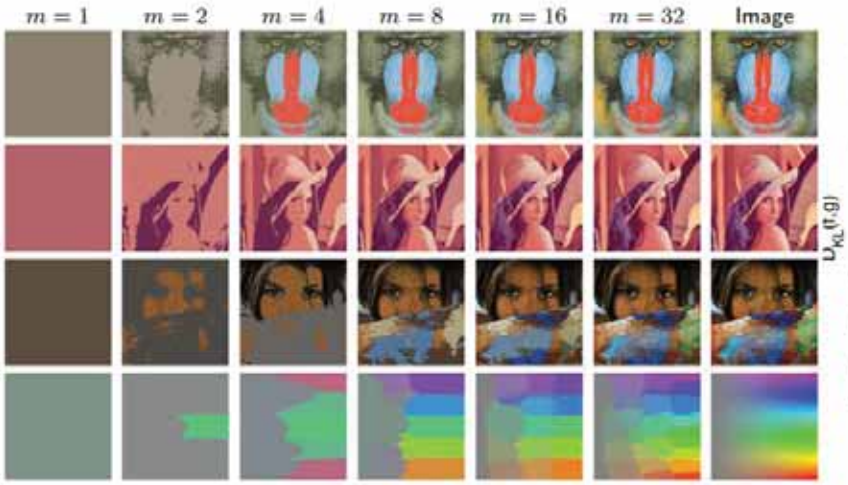
Hierarchical clustering with Bregman sided and symmetrized divergences

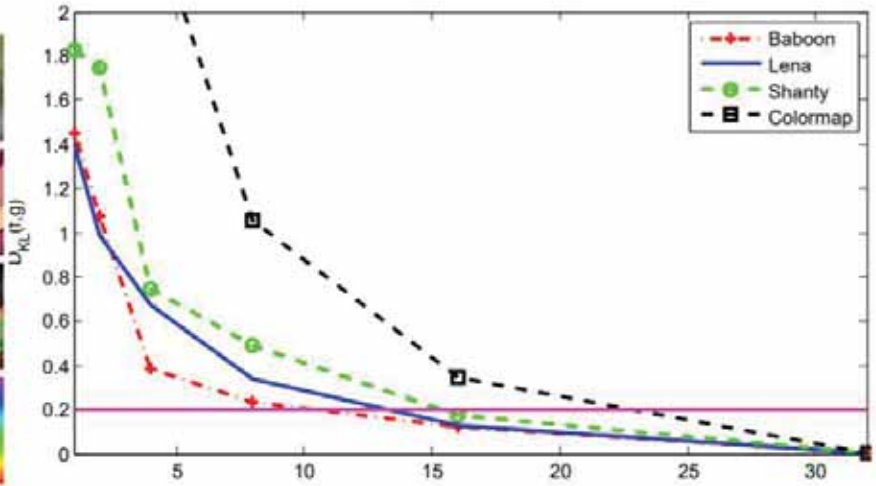
Learning & simplifying
Gaussian mixture models (GMMs)

- Agglomerative method:
  1. Find the two closest subsets $\mathcal{S}_i$ and $\mathcal{S}_j$
  2. Merge the subsets $\mathcal{S}_i$ and $\mathcal{S}_j$
  3. Go back to 1. until one single set remains



Criterion	Formula
Minimum distance	$\mathrm{D}_{\min}(A,B) = \min\{d(a,b) \mid a \in A,\ b \in B\}$
Maximum distance	$\mathrm{D}_{\max}(A,B) = \max\{d(a,b) \mid a \in A,\ b \in B\}$
Average distance	$\mathrm{D}_{\mathrm{av}}(A,B) = \frac{1}{\lvert A\rvert\lvert B\rvert} \sum_{a\in A} \sum_{b\in B} d(a,b)$



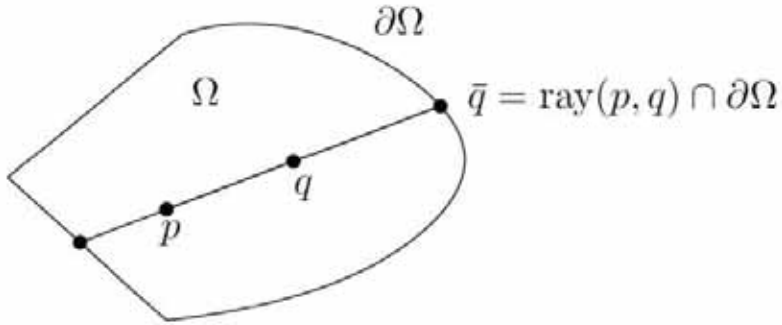


@FrnkNlsn

Simplification and hierarchical representations of mixtures of exponential families. Signal Processing 90(12): (2010)

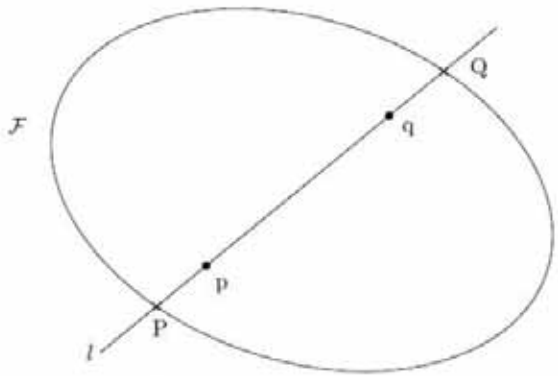
Hilbert geometry: Finsler, Riemann, Cayley-Klein geometries

- Hilbert geometry (log cross-ratio metric) has straight geodesics (not necessarily unique)
- Hilbert geometry can be studied from Finslerian viewpoint (smooth domain, Minkowski norm)
- When the domain is a simplex, Hilbert geometry is isometric to normed (Hilbert) space
- Hilbert Finslerian geometry is Riemannian **iff domain=ellipsoid**: **Cayley-Klein geometry**



$\bar{p} = \mathrm{ray}(q,p) \cap \partial\Omega$

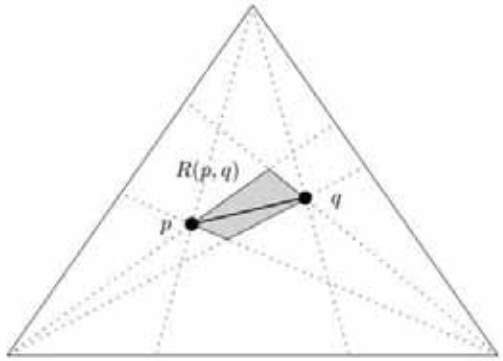
$$H_\Omega(p,q) = \log \frac{\|\bar{q}-p\|\|\bar{p}-q\|}{\|\bar{q}-q\|\|\bar{p}-p\|}$$



(a) distance measurement

$\mathrm{dist}(p,q) = c_{\mathrm{dist}}\ \mathrm{Log}((p,q;P,Q))$

Cayley-Klein geometry



Hilbert log cross-ratio
(isometric normed vector space)

On approximating the Riemannian 1-center, CGTA, 2013, **arXiv:1101.4718**

Clustering in Hilbert simplex geometry, 2017, https://arxiv.org/abs/1704.00454

Classification with mixtures of curved Mahalanobis metrics, 2016, **arXiv**:1609.07082, 2016

Hilbert's projective distance is information monotone on the probability simplex

f-divergences are separable information monotone:

$$I_f[p:q] := \sum_{i=1}^{d} p[i] f(q[i]/p[i]) \qquad I_f[p_{|\mathcal{X}} : q_{|\mathcal{X}}] \leq I_f[p:q], \quad \forall \mathcal{X}$$

Hilbert's projective metric on the positive orthant
(metric on the probability simplex)

$$\boxed{D_{\mathrm{Hilbert}}[p,q] = \log \frac{\max_i \frac{p_i}{q_i}}{\min_i \frac{p_i}{q_i}}}$$

$$\Delta_d = \{(x_1, \ldots, x_d) \in \mathbb{R}_+^d \ : \ \textstyle\sum_{i=1}^d x_i = 1\}$$

$$D_{\mathrm{Hilbert}}[\lambda p, \lambda' q] = D_{\mathrm{Hilbert}}[p,q] \text{ for any } \lambda, \lambda' > 0.$$

Hilbert's distance is non-separable information monotone:

$$D_{\mathrm{Hilbert}}[p_{|\mathcal{X}}, q_{|\mathcal{X}}] \leq D_{\mathrm{Hilbert}}[p,q]$$

Hilbert metric distance on the probability simplex

- The d-dimensional **positive orthant cone** cut by the hyperplane $\sum_{i=1}^{d} x_i = 1$ yields the **(d-1)-dimensional probability simplex** (standard simplex)
- **Hilbert distances** are **projective distances** defined on cones:

$$\rho(\lambda p, \lambda' q) = \rho(p, q) \text{ for any } \lambda, \lambda' > 0$$

- For the positive orthant, we get

$$\boxed{\rho(p, q) = \log \frac{\max_i \frac{p_i}{q_i}}{\min_j \frac{p_j}{q_j}}}$$

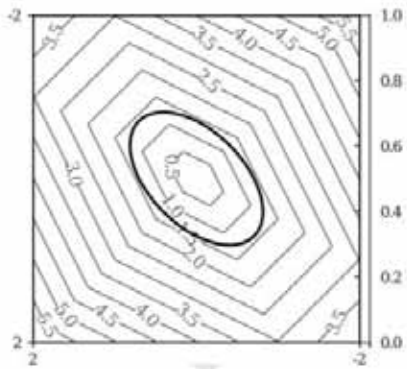
Metric distance

- **Metric** on the probability simplex
- Can be calculated in linear time
- Hilbert geometry for the probability simplex is

isometric to a (polytope) **normed vector space**

$$\rho(p, q) = \|v(p) - v(q)\|_{\mathrm{NH}}$$

$$v_i(p) = \frac{1}{d}\left((d-1)\log p_i - \sum_{j \neq i} \log p_j\right)$$



Polytope norm

(isocontours)

Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices, Geometric Structures of Information, 2019. 297-331

Hilbert cone projective metric and Hilbert convex metric

- A **cone** C in vector space V induces a **partial order** $\leq_C$:

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

- **Projective metric Hilbert distance**: $D_C(p,q) = 0 \Leftrightarrow p = \lambda q$

$$\boxed{D_C(p,q) := \log\left(M(p:q)\ M(q:p)\right)}$$

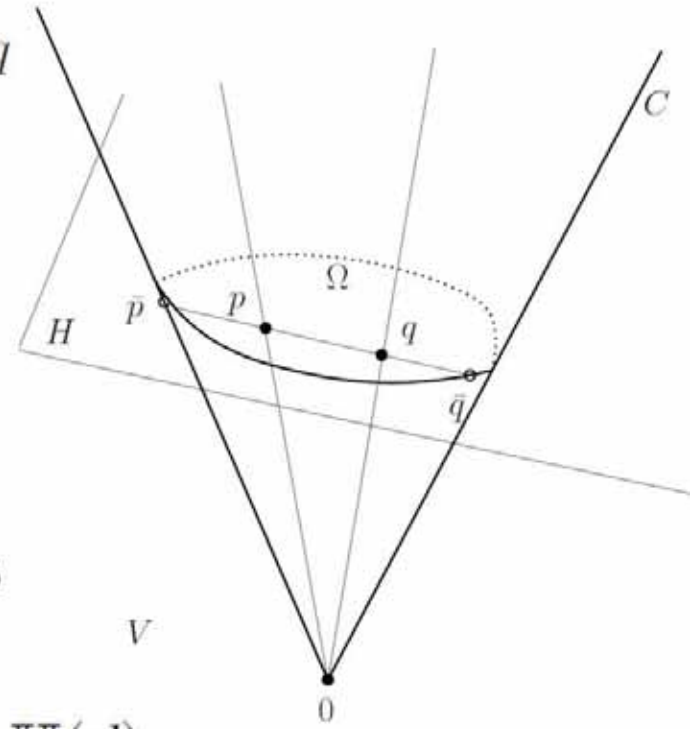
$$M(q:p) := \inf\{\lambda \in \mathbb{R}\ :\ q \leq_C \lambda p\}$$

- **Metric Hilbert distance** for a bounded domain Ω :

$$\boxed{D_\Omega(p,q) := \log\left(\frac{\|\bar{q}-p\|\|\bar{p}-q\|}{\|\bar{q}-q\|\|\bar{p}-p\|}\right)}$$

$$D_C(p,q) = D_\Omega(p,q)$$

$$\Omega = H \cap C$$



- The Hilbert geometry of the *Lorentz (icecream) cone* is is isometric to the d-dimensional *hyperbolic geometry*:

$$L := \{(y, x_1, \ldots, x_d) \in \mathbb{R}^{d+1}\ :\ y^2 > x_1^2 + \ldots x_d^2\} \qquad (L, D_L) \cong \mathbb{H}(d)$$

- FN and Ke Sun, Clustering in Hilbert simplex geometry, Geometric Structures of Information 2019, Springer, arXiv:1704.00454
- Hilbert geometry of the Siegel disk: The Siegel-Klein disk model, Entropy 2020, 22(9), 1019 arXiv 2004.08160

Hilbert distance on the symmetric positive-definite matrix cone vs on the correlation elliptope

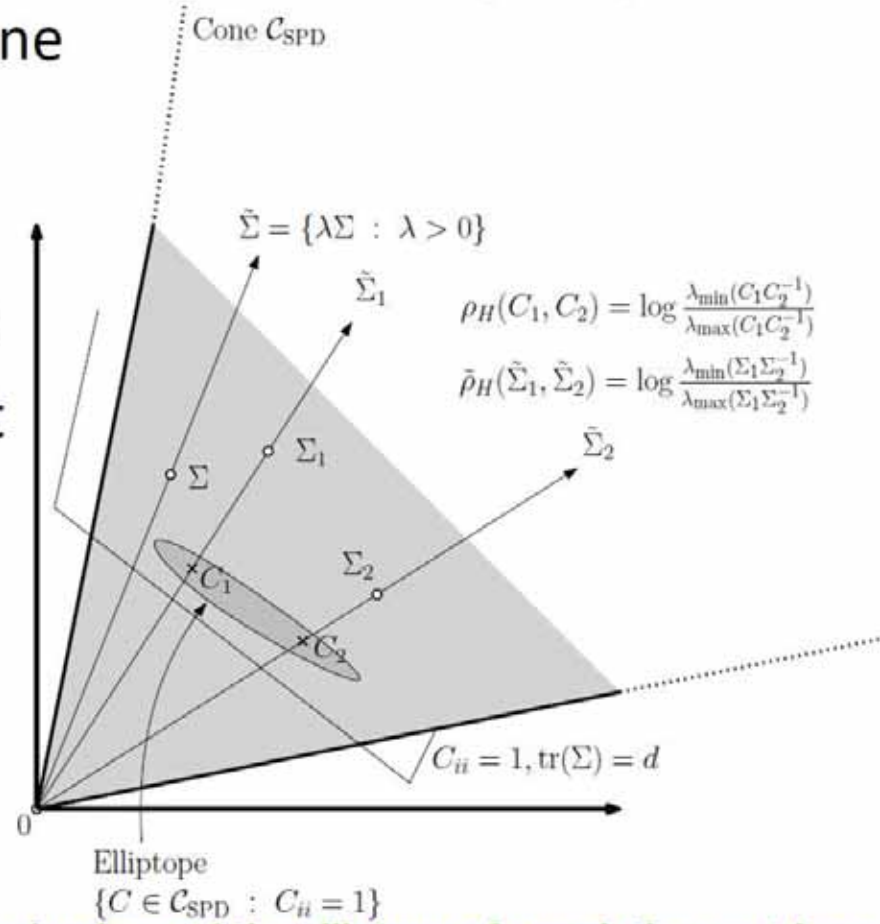
- **Hilbert projective distance** is defined for any cone

SPD cone ➡ $\tilde{\rho}_H(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = \log \frac{\lambda_{\min}(\Sigma_1 \Sigma_2^{-1})}{\lambda_{\max}(\Sigma_1 \Sigma_2^{-1})}$

- Projective means that $\rho_H(\lambda_1 p_1, \lambda_2 p_2) = \rho_H(p_1, p_2), \quad \forall \lambda_1, \lambda_2 > 0$
- On the subspace of correlation matrices, Hilbert distance becomes a **proper metric distance**:

$$\boxed{\rho_H(C_1, C_2) = \log \frac{\lambda_{\min}(C_1 C_2^{-1})}{\lambda_{\max}(C_1 C_2^{-1})}}$$

- Fast to calculate: needs only the extremal eigenvalues (power iteration methods)



Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices, Geometric Structures of Information, 2019. 297-331

Hölder projective divergences (incl. Cauchy-Schwarz div.)

A divergence D is **projective** when $D(\lambda p : \lambda' q) = D(p : q), \quad \forall \lambda, \lambda' > 0$

For α>0, define **conjugate exponents**: $\frac{1}{\alpha} + \frac{1}{\beta} = 1$

For α ,γ>0, define the family of **Hölder projective divergences**:

$$\boxed{D^H_{\alpha,\gamma}(p : q) = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta} \mathrm{d}\mu(x)}{\left(\int_{\mathcal{X}} p(x)^{\gamma} \mathrm{d}\mu(x)\right)^{1/\alpha} \left(\int_{\mathcal{X}} q(x)^{\gamma} \mathrm{d}\mu(x)\right)^{1/\beta}}\right)}$$

When α=β=γ=2, we get the Cauchy-Schwarz divergence:

$$\mathrm{CS}(p : q) = -\log \frac{\int_{\mathcal{X}} p(x) q(x) \mathrm{d}\mu(x)}{\sqrt{\left(\int_{\mathcal{X}} p(x)^2 \mathrm{d}\mu(x)\right)\left(\int_{\mathcal{X}} q(x)^2 \mathrm{d}\mu(x)\right)}}$$

@FrnkNlsn **On Hölder projective divergences, Entropy, 2017 (arXiv:1701.03916)**

Every homogeneous cone is isomorphic to a slice of the SPD matrix cone

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 500–506

© 2003 Society for Industrial and Applied Mathemati

RELATING HOMOGENEOUS CONES AND POSITIVE DEFINITE CONES VIA T-ALGEBRAS*

CHEK BENG CHUA†

COROLLARY 4.3. *If K is a homogeneous cone in $\mathbb{R}^n$, then there exist an $m \leq n$ and an injective linear map $M : \mathbb{R}^n \to \mathbb{S}^{m\times m}$ such that $M(K) = \mathbb{S}^{m\times m}_{++} \cap M(\mathbb{R}^n)$, where $\mathbb{S}^{m\times m}$ is the space of m-by-m symmetric matrices and $\mathbb{S}^{m\times m}_{++}$ is the cone of positive definite symmetric m-by-m matrices.*

Computing homographies by mapping to a square

$$\mathbf{r}_i = \begin{bmatrix} x'_i \\ y'_i \\ w'_i \end{bmatrix} = \mathbf{H}\mathbf{l}_i = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ w_i \end{bmatrix}$$

$$x'_i = \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + 1}, \qquad y'_i = \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + 1}$$

$$\begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1x'_1 & -y_1x'_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1y'_1 & -y_1y'_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2x'_2 & -y_2x'_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x_2y'_2 & -y_2y'_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3x'_3 & -y_3x'_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -x_3y'_3 & -y_3y'_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x_4x'_4 & -y_4x'_4 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -x_4y'_4 & -y_4y'_4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & 1 & 0 & 0 & 0 & -x_nx'_n & -y_nx'_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -x_ny'_n & -y_ny'_n \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ y'_2 \\ x'_3 \\ y'_3 \\ x'_4 \\ y'_4 \\ \vdots \\ x'_n \\ y'_n \end{bmatrix}$$

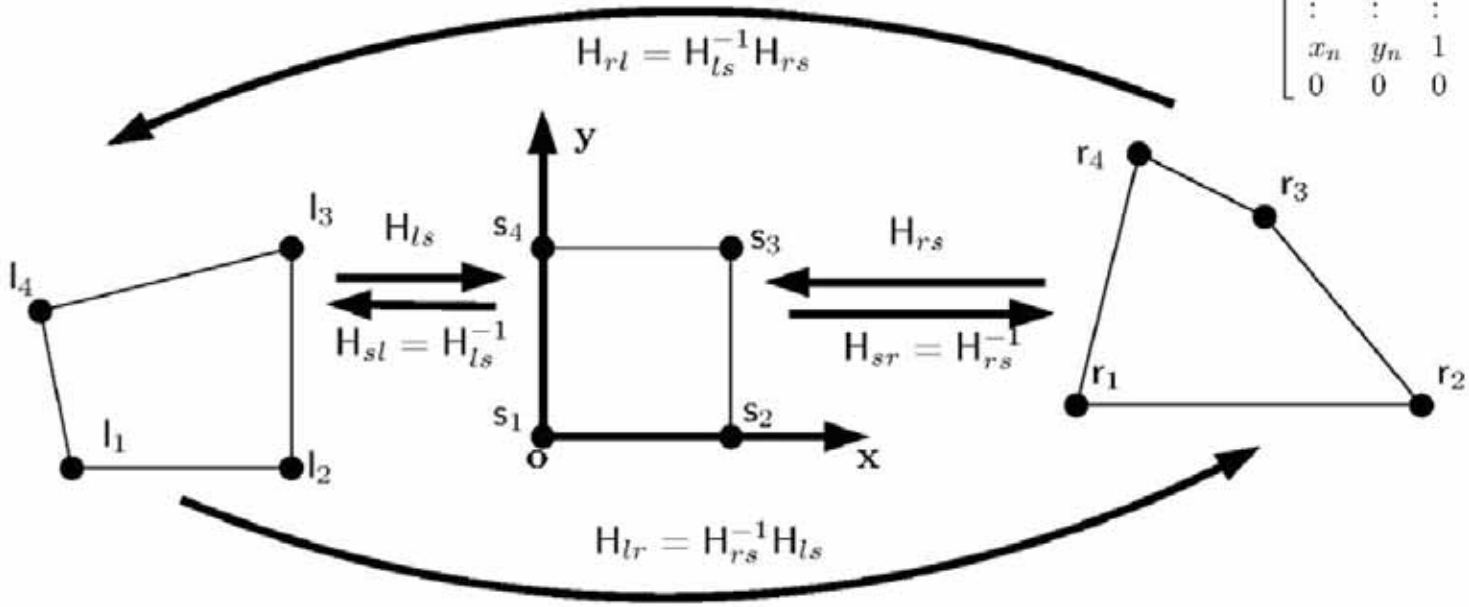
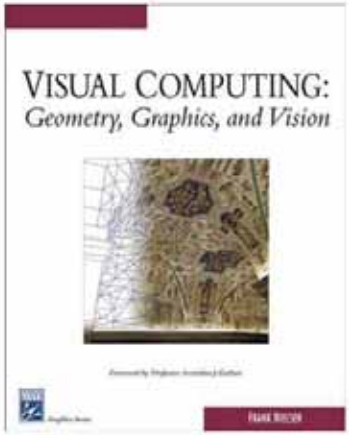
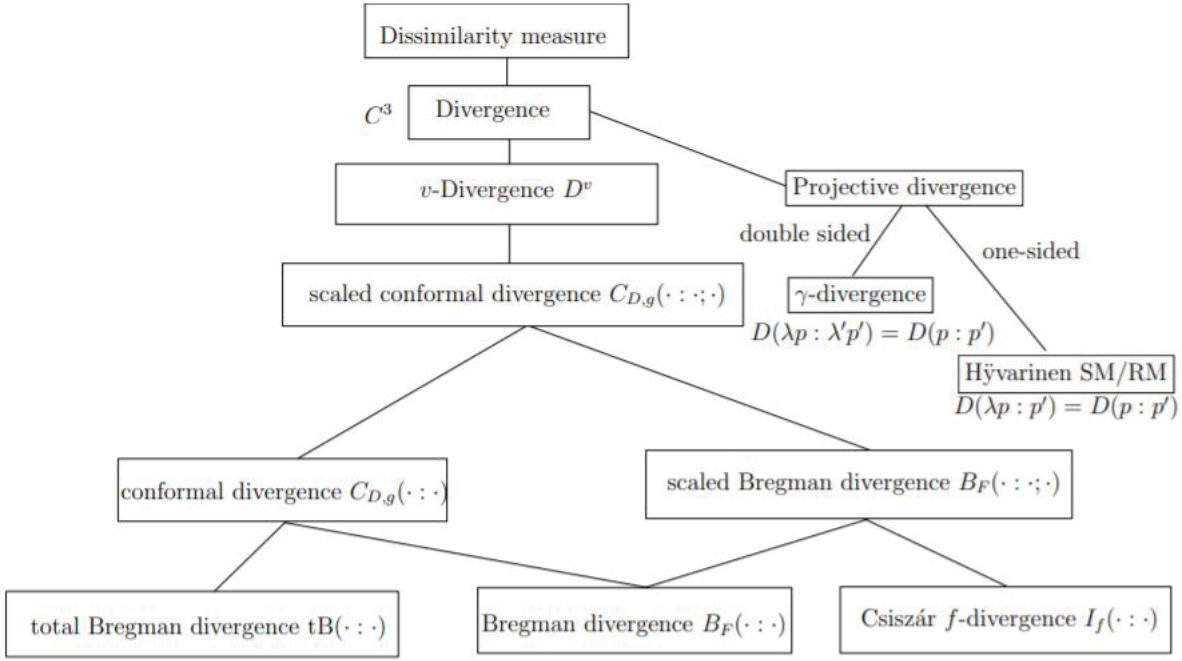


FIGURE 3.31 *Computing a homography as the composition of two quadrilateral-to-square homographies.*

simpler inhomogenous linear system with

$\mathbf{s}_1 = [x'_1\ y'_1]^T$	$[0\ 0]^T$
$\mathbf{s}_2 = [x'_2\ y'_2]^T$	$[1\ 0]^T$
$\mathbf{s}_3 = [x'_3\ y'_3]^T$	$[1\ 1]^T$
$\mathbf{s}_4 = [x'_4\ y'_4]^T$	$[0\ 1]^T$





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

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

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

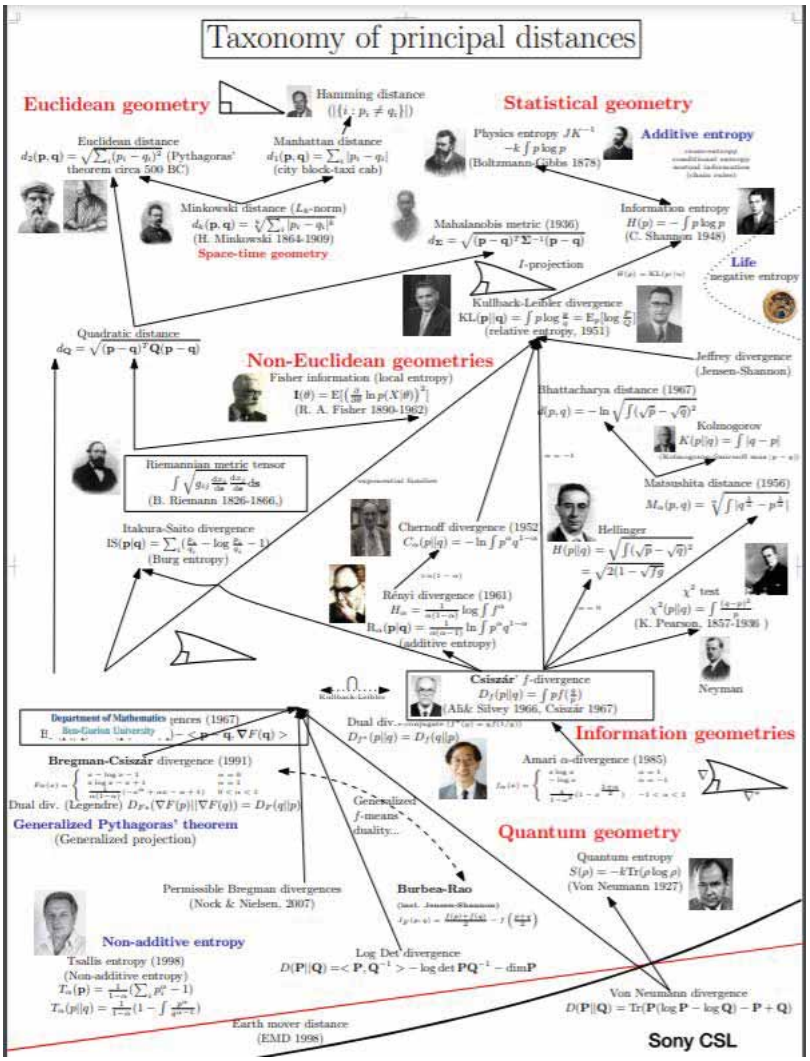
$\mathrm{tB}_F(P:Q) = \frac{B_F(P:Q)}{\sqrt{1+\|\nabla F(Q)\|^2}}$

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

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

Figure 17. Principled classes of distances/divergences.

Taxonomy of principal distances



Part I

High Performance Computing (HPC) with the Message Passing Interface (MPI)

Front Matter
Pages 1-1

A Glance at High Performance Computing (HPC)
Frank Nielsen
Pages 3-20

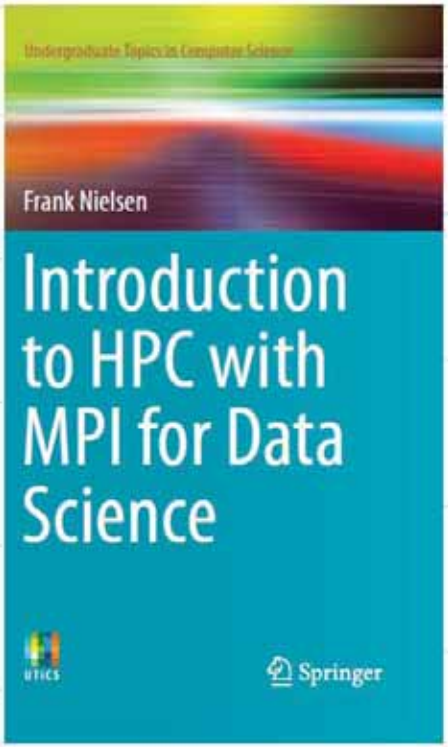
Introduction to MPI: The Message Passing Interface
Frank Nielsen
Pages 21-62

Topology of Interconnection Networks
Frank Nielsen
Pages 63-97

Parallel Sorting
Frank Nielsen
Pages 99-120

Parallel Linear Algebra
Frank Nielsen
Pages 121-145

The MapReduce Paradigm
Frank Nielsen
Pages 147-160



Part II

High Performance Computing (HPC) for Data Science (DS)

Front Matter
Pages 161-161

Partition-Based Clustering with k-Means
Frank Nielsen
Pages 163-193

Hierarchical Clustering
Frank Nielsen
Pages 195-211

Supervised Learning: Practice and Theory of Classification with the k-NN Rule
Frank Nielsen
Pages 213-229

Fast Approximate Optimization in High Dimensions with Core-Sets and Fast Dimension Reduction
Frank Nielsen
Pages 231-244

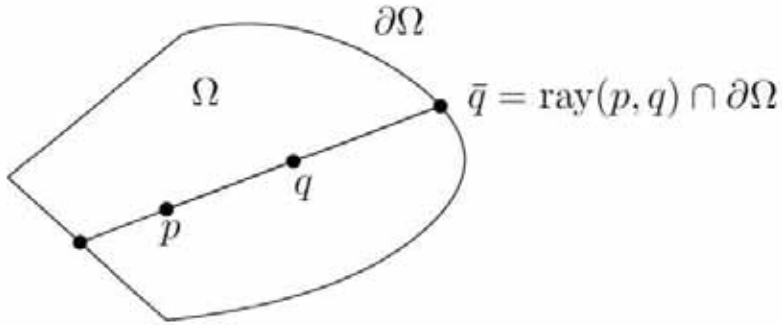
Parallel Algorithms for Graphs
Frank Nielsen
Pages 245-259

Back Matter
Pages 261-282

https://franknielsen.github.io/HPC4DS/index.html

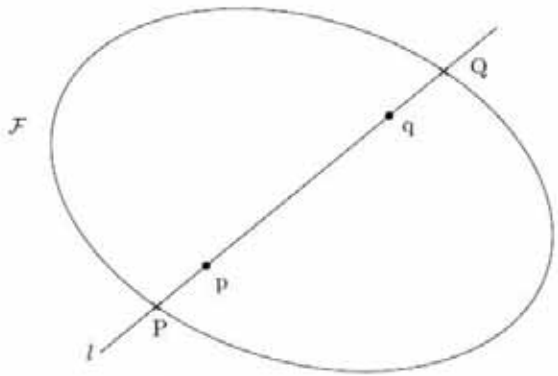
Hilbert geometry: Finsler, Riemann, Cayley-Klein geometries

- Hilbert geometry (log cross-ratio metric) has straight geodesics (not necessarily unique)
- Hilbert geometry can be studied from Finslerian viewpoint (smooth domain, Minkowski norm)
- When the domain is a simplex, Hilbert geometry is isometric to normed (Hilbert) space
- Hilbert Finslerian geometry is Riemannian **iff domain=ellipsoid**: **Cayley-Klein geometry**



$\bar{p} = \mathrm{ray}(q,p) \cap \partial\Omega$

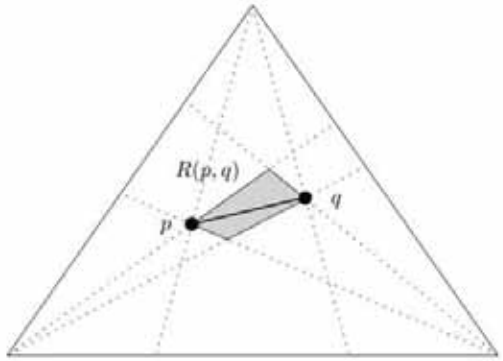
$$H_\Omega(p,q) = \log \frac{\|\bar{q}-p\|\|\bar{p}-q\|}{\|\bar{q}-q\|\|\bar{p}-p\|}$$



(a) distance measurement

$\mathrm{dist}(p,q) = c_{\mathrm{dist}}\ \mathrm{Log}((p,q;P,Q))$

Cayley-Klein geometry



Hilbert log cross-ratio
(isometric normed vector space)

On approximating the Riemannian 1-center, CGTA, 2013, **arXiv:1101.4718**

Clustering in Hilbert simplex geometry, 2017, https://arxiv.org/abs/1704.00454

Classification with mixtures of curved Mahalanobis metrics, 2016, **arXiv**:1609.07082, 2016

Hilbert cone projective metric and Hilbert convex metric

- A **cone** C in vector space V induces a **partial order** $\leq_C$:

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

- **Projective metric Hilbert distance**: $D_C(p,q) = 0 \Leftrightarrow p = \lambda q$

$$\boxed{D_C(p,q) := \log\left(M(p:q)\ M(q:p)\right)}$$

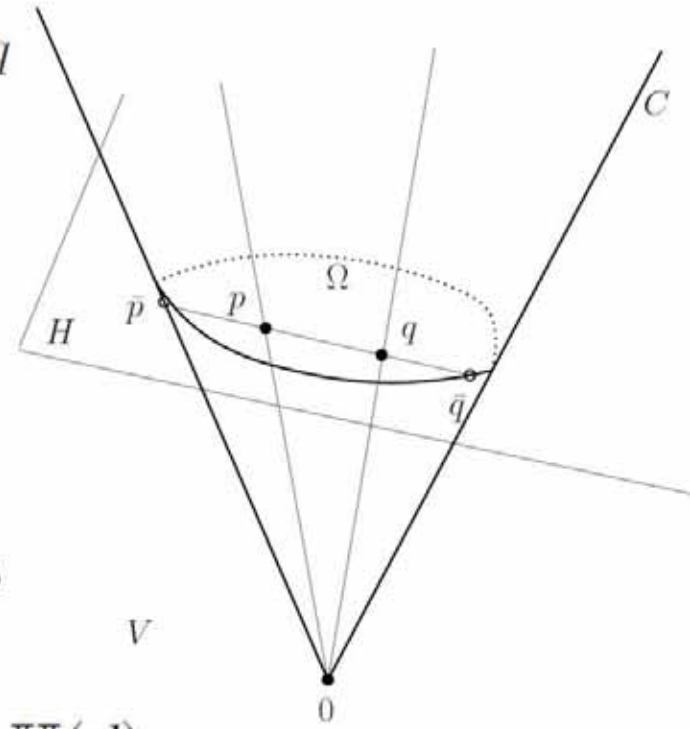
$$M(q:p) := \inf\{\lambda \in \mathbb{R}\ :\ q \leq_C \lambda p\}$$

- **Metric Hilbert distance** for a bounded domain Ω :

$$\boxed{D_\Omega(p,q) := \log\left(\frac{\|\bar{q}-p\|\|\bar{p}-q\|}{\|\bar{q}-q\|\|\bar{p}-p\|}\right)}$$

$$D_C(p,q) = D_\Omega(p,q)$$

$$\Omega = H \cap C$$



- The Hilbert geometry of the *Lorentz (icecream) cone* is is isometric to the d-dimensional *hyperbolic geometry*:

$$L := \{(y, x_1, \ldots, x_d) \in \mathbb{R}^{d+1}\ :\ y^2 > x_1^2 + \ldots x_d^2\} \qquad (L, D_L) \cong \mathbb{H}(d)$$

- FN and Ke Sun, Clustering in Hilbert simplex geometry, Geometric Structures of Information 2019, Springer, arXiv:1704.00454
- Hilbert geometry of the Siegel disk: The Siegel-Klein disk model, Entropy 2020, 22(9), 1019 arXiv 2004.08160

Hilbert metric distance on the probability simplex

- The d-dimensional **positive orthant cone** cut by the hyperplane $\sum_{i=1}^{d} x_i = 1$ yields the **(d-1)-dimensional probability simplex** (standard simplex)
- **Hilbert distances** are **projective distances** defined on cones:

$$\rho(\lambda p, \lambda' q) = \rho(p, q) \text{ for any } \lambda, \lambda' > 0$$

- For the positive orthant, we get

$$\boxed{\rho(p, q) = \log \frac{\max_i \frac{p_i}{q_i}}{\min_j \frac{p_j}{q_j}}}$$

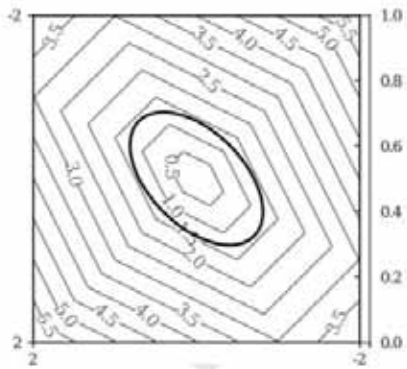
Metric distance

- **Metric** on the probability simplex
- Can be calculated in linear time
- Hilbert geometry for the probability simplex is

isometric to a (polytope) **normed vector space**

$$\rho(p, q) = \|v(p) - v(q)\|_{\mathrm{NH}}$$

$$v_i(p) = \frac{1}{d}\left((d-1)\log p_i - \sum_{j \neq i} \log p_j\right)$$



Polytope norm

(isocontours)

Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices, Geometric Structures of Information, 2019. 297-331

Hilbert distance on the symmetric positive-definite matrix cone vs on the correlation elliptope

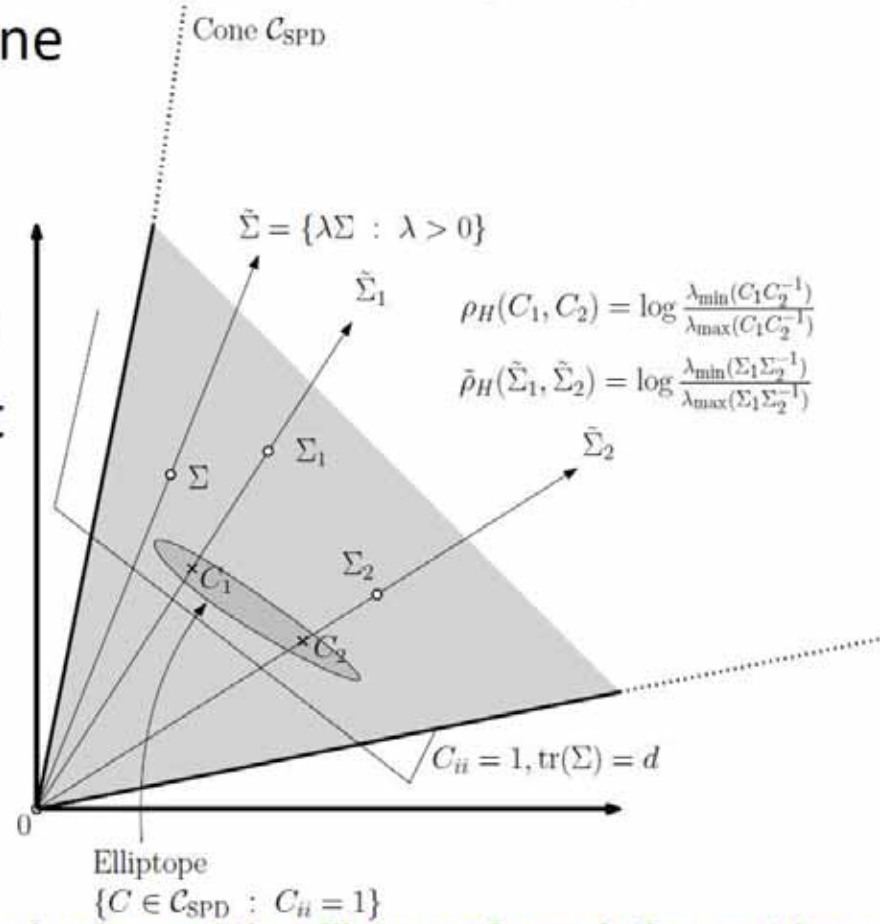
- **Hilbert projective distance** is defined for any cone

SPD cone ➡ $\tilde{\rho}_H(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = \log \frac{\lambda_{\min}(\Sigma_1 \Sigma_2^{-1})}{\lambda_{\max}(\Sigma_1 \Sigma_2^{-1})}$

- Projective means that $\rho_H(\lambda_1 p_1, \lambda_2 p_2) = \rho_H(p_1, p_2), \quad \forall \lambda_1, \lambda_2 > 0$
- On the subspace of correlation matrices, Hilbert distance becomes a **proper metric distance**:

$$\boxed{\rho_H(C_1, C_2) = \log \frac{\lambda_{\min}(C_1 C_2^{-1})}{\lambda_{\max}(C_1 C_2^{-1})}}$$

- Fast to calculate: needs only the extremal eigenvalues (power iteration methods)



Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices, Geometric Structures of Information, 2019. 297-331

Hilbert's projective distance is information monotone on the probability simplex

f-divergences are separable information monotone:

$$I_f[p:q] := \sum_{i=1}^{d} p[i] f(q[i]/p[i]) \qquad I_f[p_{|\mathcal{X}} : q_{|\mathcal{X}}] \leq I_f[p:q], \quad \forall \mathcal{X}$$

Hilbert's projective metric on the positive orthant
(metric on the probability simplex)

$$\boxed{D_{\mathrm{Hilbert}}[p,q] = \log \frac{\max_i \frac{p_i}{q_i}}{\min_i \frac{p_i}{q_i}}}$$

$$\Delta_d = \{(x_1, \ldots, x_d) \in \mathbb{R}_+^d \ : \ \textstyle\sum_{i=1}^d x_i = 1\}$$

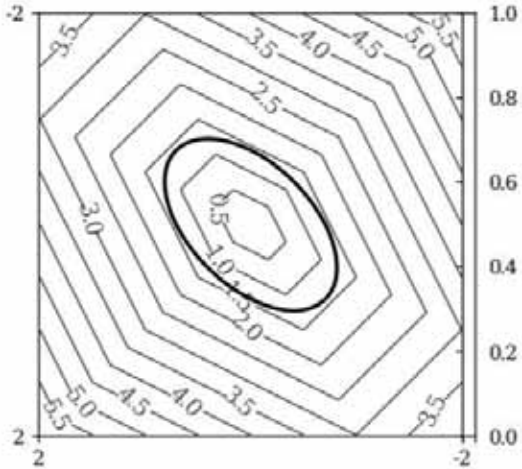
$$D_{\mathrm{Hilbert}}[\lambda p, \lambda' q] = D_{\mathrm{Hilbert}}[p,q] \text{ for any } \lambda, \lambda' > 0.$$

Hilbert's distance is non-separable information monotone:

$$D_{\mathrm{Hilbert}}[p_{|\mathcal{X}}, q_{|\mathcal{X}}] \leq D_{\mathrm{Hilbert}}[p,q]$$

Variation norm and Hilbert projective distances

- **Variation norm**: $\boxed{\|x\|_{\mathrm{var}} := \max_{i\in[d]}\{x_i\} - \min_{i\in[d]}\{x_i\}}$



- Hilbert projective distance on the **probability simplex**:

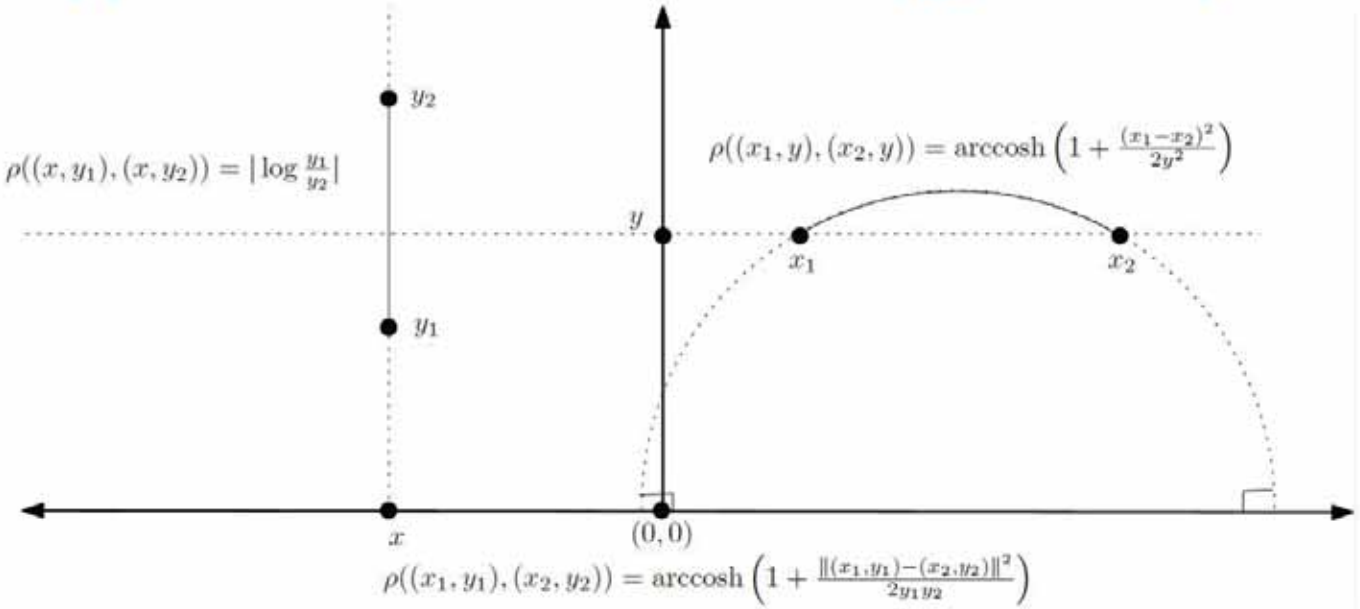
$$D_{\mathrm{Hilbert}}(p,q) = \|\log\frac{p}{q}\|_{\mathrm{var}} = \log\frac{\max_i \frac{p_i}{q_i}}{\min_j \frac{p_j}{q_j}}$$

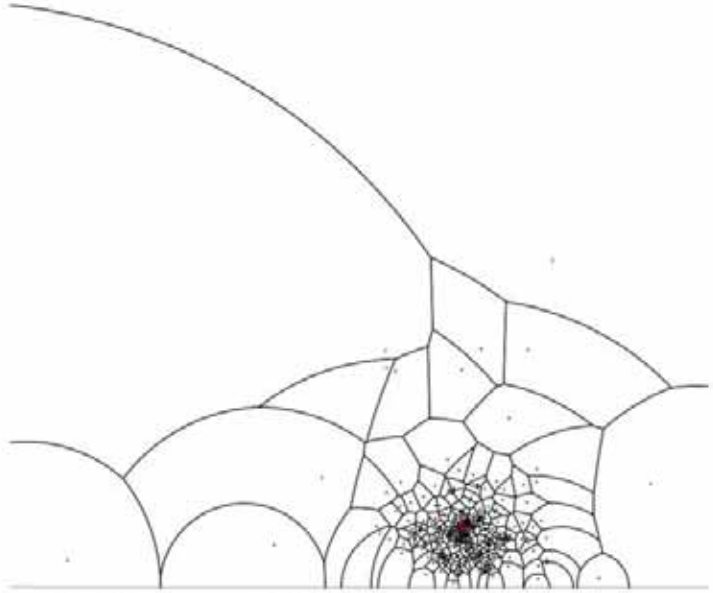
- Hilbert projective distance between two **correlation matrices**:

$$D_{\mathrm{Correlation}}(C_1,C_2) = \|\log\lambda(C_1^{-1}C_2)\|_{\mathrm{var}} := \log\frac{\lambda_{\max}(C_1^{-1}C_2)}{\lambda_{\min}(C_1^{-1}C_2)}$$

- Nielsen and Sun, Clustering in Hilbert simplex geometry, arXiv:1704.00454
- Nielsen and Sun, Clustering in Hilbert's projective geometry: The case studies of the probability simplex, Geom. Struct. Inf (2018): 297-331.

Hyperbolic Voronoi diagrams (Poincaré disk model)

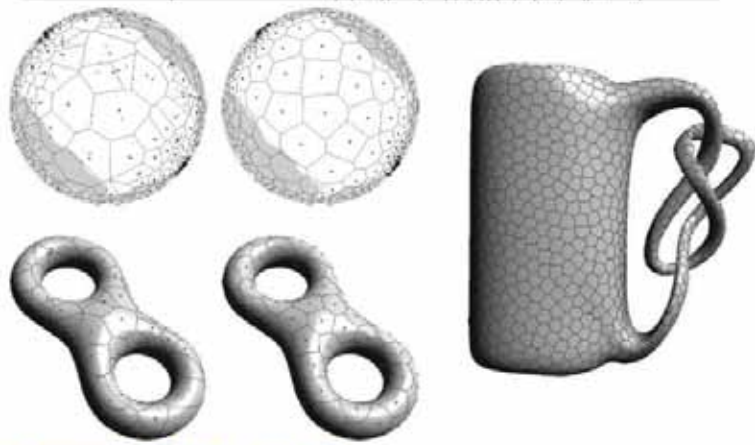




Easy to calculate

- **Hyperbolic Voronoi diagrams (HVD) and**
- **Hyperbolic centroidal Voronoi tessellations**

using Klein non-conformal model

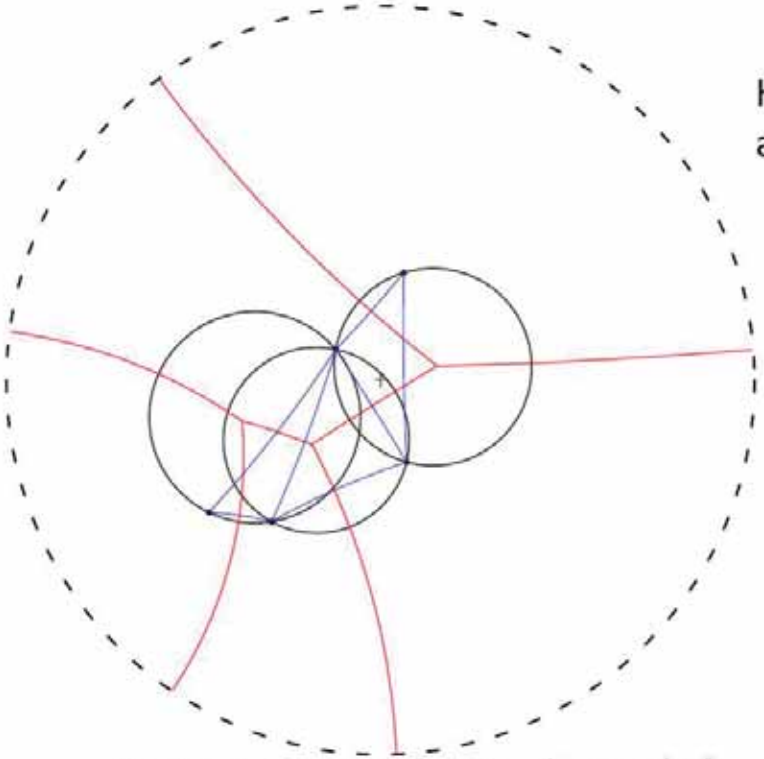


Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010

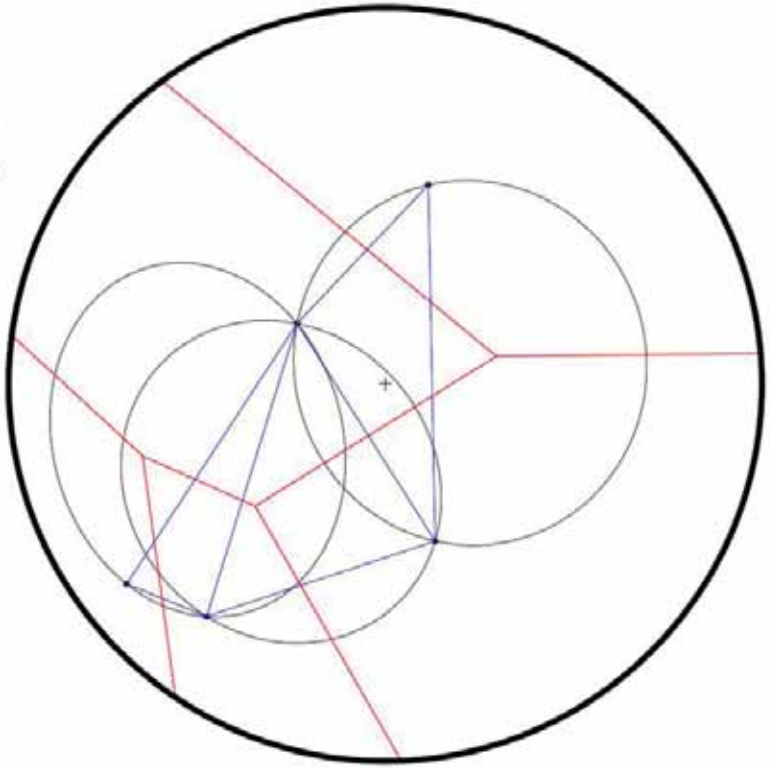
Rong et al., Centroidal Voronoi tessellation in universal covering space of manifold surfaces, CAGD, 2011

Hyperbolic Delaunay complex: Empty-sphere property

The hyperbolic sphere passing through the vertices of a Delaunay triangle is empty of other sites



Hyperbolic sphere centers
are on Voronoi T-junctions

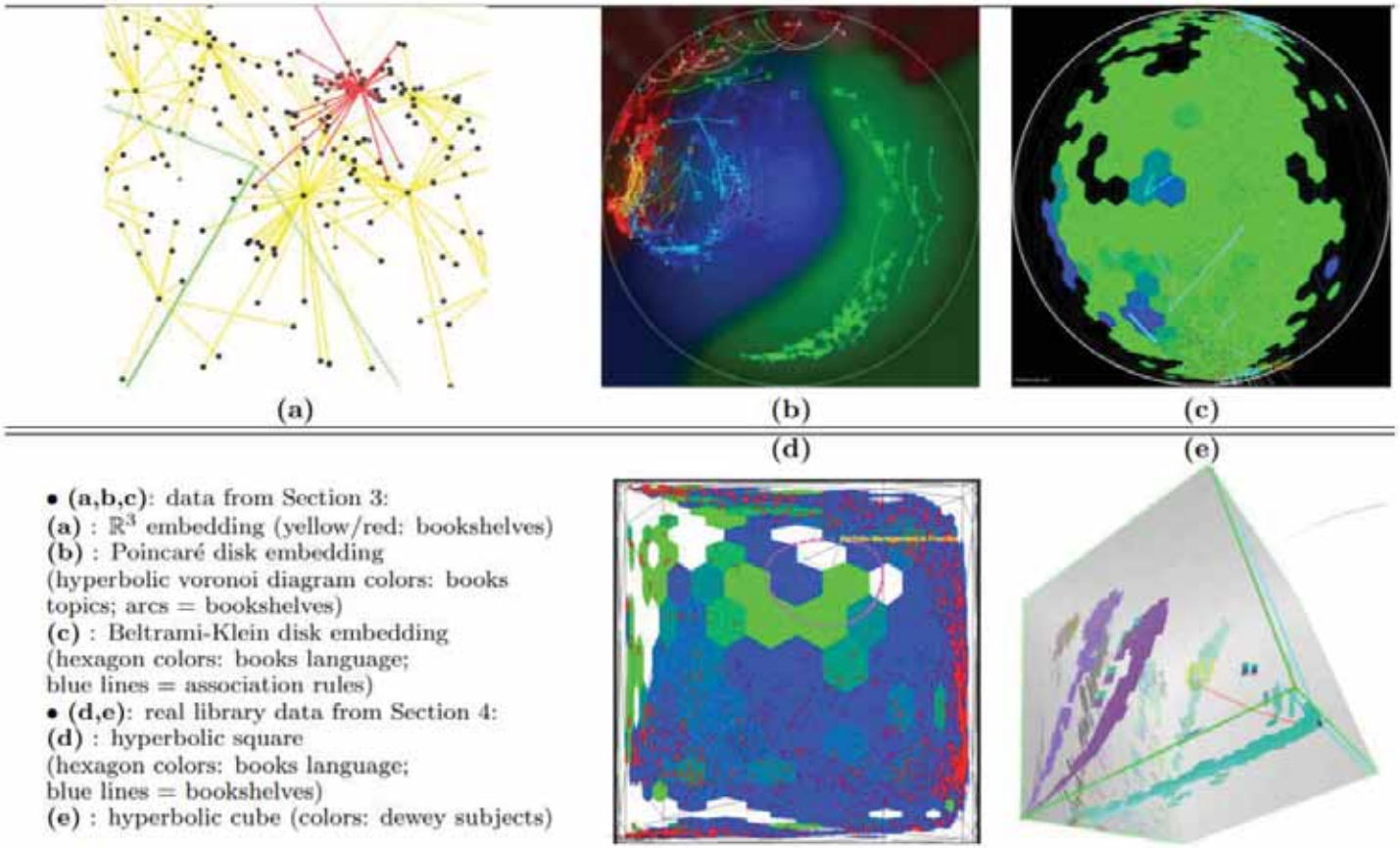


Poincare conformal model
Hyperbolic balls have Euclidean ball shape
(with displaced center)

Klein non-conformal model
Hyperbolic balls have Euclidean ellipsoid shape
(with displaced center)

Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010.

3D focus+context visualizations of book library: hyperbolic geometric and mappings in cubes



- **(a,b,c)**: data from Section 3:
(a) : $\mathbb{R}^3$ embedding (yellow/red: bookshelves)
(b) : Poincaré disk embedding
(hyperbolic voronoi diagram colors: books topics; arcs = bookshelves)
(c) : Beltrami-Klein disk embedding
(hexagon colors: books language;
blue lines = association rules)
- **(d,e)**: real library data from Section 4:
(d) : hyperbolic square
(hexagon colors: books language;
blue lines = bookshelves)
(e) : hyperbolic cube (colors: dewey subjects)

Non-linear book manifolds: learning from associations the dynamic geometry of digital libraries ACM/IEEE DL 2013

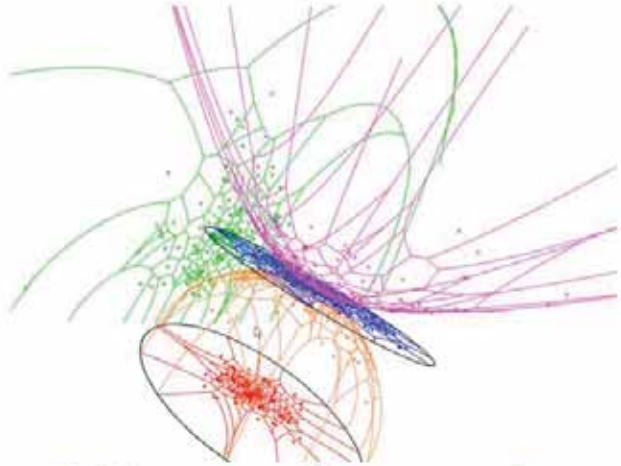
Riemannian metrics of hyperbolic manifold models

Five main models:

- the **Poincaré Upper half-space** (U)
- the Poincaré ball (P)
- the Klein ball (K)
- the Lorentz hyperboloid (L)
- the Beltrami hemisphere models (B)

$$
\begin{aligned}
\mathcal{U}^+ &= \{(x_0, r, x_2, \ldots x_d) \mid x_0 > 0\} \text{ (upper space)},\\
\mathcal{P} &= \left\{(0, x_1, \ldots, x_d) \,\middle|\, \sum_{i=1}^{d} x_i^2 < r^2\right\} \text{ (Poincaré ball)},\\
\mathcal{S}^+ &= \left\{(x_0, x_1, \ldots, x_d) \,\middle|\, \sum_{i=0}^{d} x_i^2 < r^2 \text{ and } x_0 > 0\right\} \text{ (Beltrami hemisphere)},\\
\mathcal{K} &= \left\{(r, x_1, \ldots, x_d) \,\middle|\, \sum_{i=1}^{d} x_i^2 < r^2\right\} \text{ (Klein ball)},\\
\mathcal{L}^+ &= \left\{(x_0, x_1, \ldots, x_d) \,\middle|\, \sum_{i=1}^{d} x_i^2 - x_0^2 = -r^2 \text{ and } x_0 > 0\right\} \text{ (Lorentz hyperboloid)}.
\end{aligned}
$$

Model	Riemannian metric $\mathrm{d}s^2$	$\cosh\frac{d(p,q)}{\sqrt{-\kappa}}$, $d(\cdot,\cdot)$: Riemannian distance
Klein (K)	$-\kappa\left(\frac{\mathrm{d}s_E^2}{1-\lVert x\rVert_E^2} + \frac{\langle x,\mathrm{d}x\rangle_E}{(1-\lVert x\rVert_E^2)^2}\right)$	$\frac{1-\langle p,q\rangle_E}{\sqrt{(1-\lVert p\rVert_E^2)(1-\lVert q\rVert_E^2)}}$
Poincaré (P)	$-\kappa\frac{4\mathrm{d}s_E^2}{(1-\lVert x\rVert_E^2)^2}$	$1+\frac{2\lVert p-q\rVert_E^2}{(1-\lVert p\rVert_E^2)(1-\lVert q\rVert_E^2)}$
Upper (U)	$-\kappa\frac{\mathrm{d}s_E^2}{x_d^2}$	$1+\frac{\lVert p-q\rVert_E^2}{2p_dq_d}$
Hyperboloid (L)	$\kappa\mathrm{d}x_0^2+\mathrm{d}s_E^2$	$\frac{\langle p,q\rangle_L}{\kappa}$
Hemisphere (B)	$\frac{\sum_{i=0}^{d}\mathrm{d}x_i^2}{x_0^2}$	$1+\frac{1-\langle p,q\rangle_E}{\kappa p_0q_0}$



The hyperbolic Voronoi diagram in arbitrary dimension

https://arxiv.org/abs/1210.8234

Hyvärinen one-sided projective divergence

Hyvärinen divergence between two densities p and q wrt to positive measure μ

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

Projective divergence on one side because: $D_{\mathrm{Hyv}}[p:q] = D_{\mathrm{Hyv}}[p:\lambda q], \forall \lambda > 0$

The trick: $\nabla_x \log q(x) = \nabla_x \log \tilde{q}(x)$

Useful to handle unnormalized densities $\tilde{q}$: $\boxed{D_{\mathrm{Hyv}}[p:\tilde{q}] = D_{\mathrm{Hyv}}[p:q]}$

with $q(x) = \dfrac{\tilde{q}(x)}{\int \tilde{q}(x)\mathrm{d}\mu(x)}$

Hyvärinen score matching: allows one to estimate exponential families with computationally intractable normalizing constants

Hyvärinen,"Estimation of non-normalized statistical models by score matching", JMLR 2005
"Patch matching with polynomial exponential families and projective divergences", SISAP 2016

Hyvärinen divergence for polynomial exponential families

- **Hyvärinen divergence** is a right-sided projective divergence:

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

$$\forall \lambda > 0, \quad D_{\mathrm{Hyv}}[p:\lambda q] = D_{\mathrm{Hyv}}[p:q]$$

Rhs can be an unormalized density:

$$D_{\mathrm{Hyv}}[p:q] = D_{\mathrm{Hyv}}[p:\tilde{q}]$$

- For densities belonging to an exponential family: $p_\theta(x) = \exp\left(\langle \theta, t(x)\rangle - F(\theta) + k(x)\right)$

$$D_{\mathrm{Hyv}}[p_{\theta_1}:p_{\theta_2}] = \frac{1}{2}\int \|\langle \theta, \nabla_x t(x)\rangle\|^2 p_{\theta_1}(x)\mathrm{d}\mu(x).$$

- In particular, for **polynomial exponential families** (PEFs) including the Gaussian family, we have:

$$D_{\mathrm{Hyv}}[p^{\mathrm{PEF}}_{\theta_1}:p^{\mathrm{PEF}}_{\theta_2}] = \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}\Delta\theta_i\Delta\theta_j ij E_{p_{\theta_1}}\left[x^{i+j-2}\right]$$

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

with

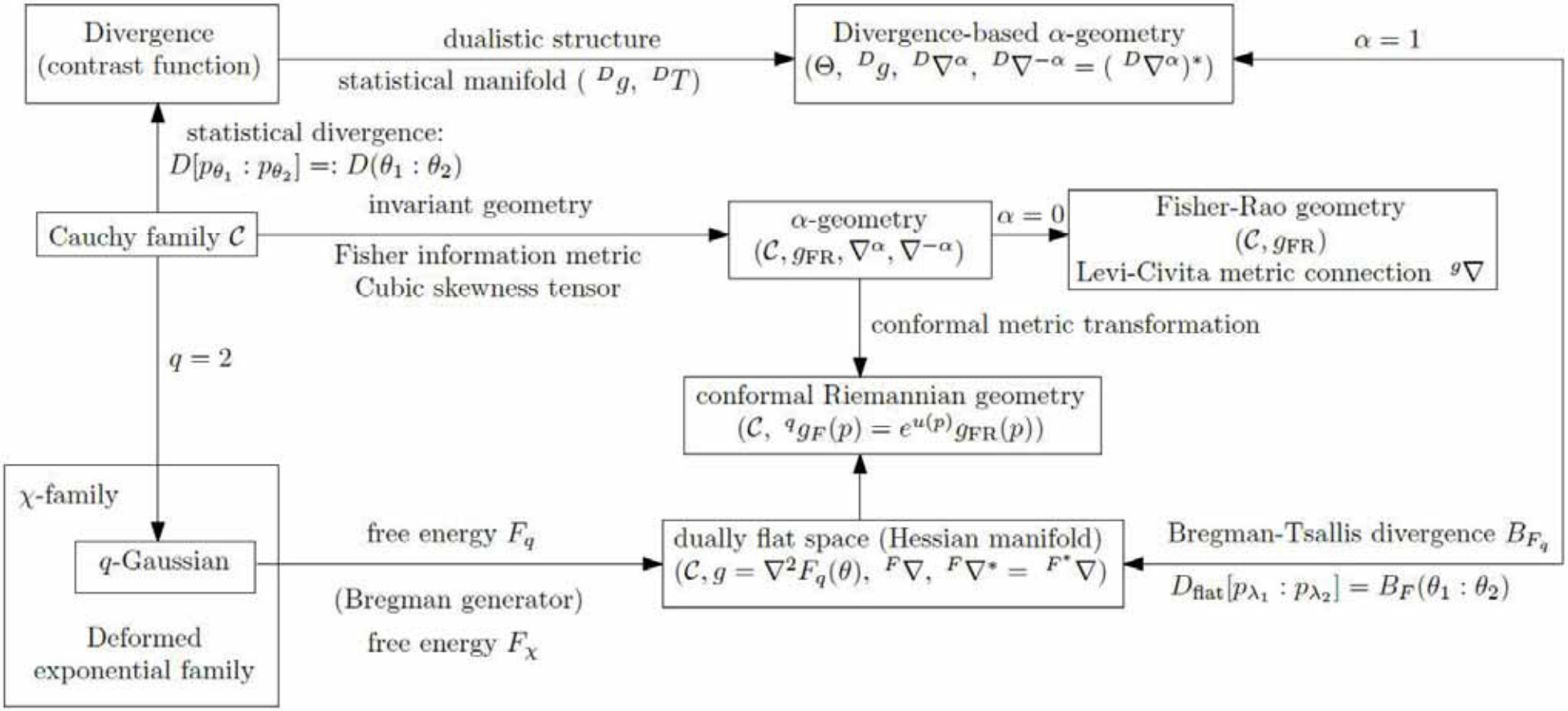
$$t(x) = (x, \ldots, x^D)$$

$$\boxed{D_{\mathrm{Hyv}}[p^{\mathrm{Nor}}_{\mu_1,\sigma_1}:p^{\mathrm{Nor}}_{\mu_2,\sigma_2}] = \frac{(\sigma_1^2-\sigma_2^2)^2}{2\sigma_1^2\sigma_2^4} + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^4}}$$

Shun-ichi Amari
Information Geometry and Its Applications
Springer
Information Geometry and Its Applications
Springer
Translations of MATHEMATICAL MONOGRAPHS
Statistical Decision Rules and Optimal Inference
N. N. Čencov
Frank Nielsen
Frank Critchley
Christopher T.J. Dodson Editors
Computational Information Geometry
For Image and Signal Processing
Springer
Ovidiu Calin · Constantin Udrişte
Geometric Modeling in Probability and Statistics
Springer
111
AFFINE DIFFERENTIAL GEOMETRY
Julian Hofrichter
Jürgen Jost
Tat Dat Tran
Information Geometry and Population Genetics
The Mathematical Structure of the Wright-Fisher Model
Springer
Translations of MATHEMATICAL MONOGRAPHS
Methods of Information Geometry
AMS
OXFORD
Masahito Hayashi
Quantum Information Theory
Mathematical Foundation
Second Edition
Springer
HIROHIKO SHIMA
THE GEOMETRY OF HESSIAN STRUCTURES
Frank Nielsen · Rajendra Bhatia Editors
Matrix Information Geometry
Springer
Algebraic Geometry and Statistical Learning Theory
Sumio Watanabe
Mathematical Foundations of Infinite-Dimensional Statistical Models
Evarist Giné
Richard Nickl
Algebraic and Geometric Methods in Statistics
Frank Nielsen Editor
Progress in Information Geometry
Theory and Applications
Springer
Geometrical Foundations of Asymptotic Inference
Robert E. Kass
Paul W. Vos
Monographs on Statistics and Applied Probability 48
Differential Geometry and Statistics
Michael K. Murray and John W. Rice
CHAPMAN & HALL/CRC
Frank Nielsen Editor
Geometric Theory of Information
Springer
Memoirs
APPLICATIONS OF DIFFERENTIAL GEOMETRY TO ECONOMETRICS
Paul Marriott and Mark Salmon
Nihat Ay
Jürgen Jost
Hông Vân Lê
Lorenz Schwachhöfer
Information Geometry
Springer
Frank Nielsen Editor
Geometric Structures of Information
Springer
Information and Exponential Families in Statistical Theory
Ole E. Barndorff-Nielsen
WILEY

Information-geometric structures of the Cauchy manifolds

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



On Voronoi diagrams and dual Delaunay complexes on the information-geometric Cauchy manifolds, arxiv 2006.07020

Equivariance, covariance and invariance in information geometry

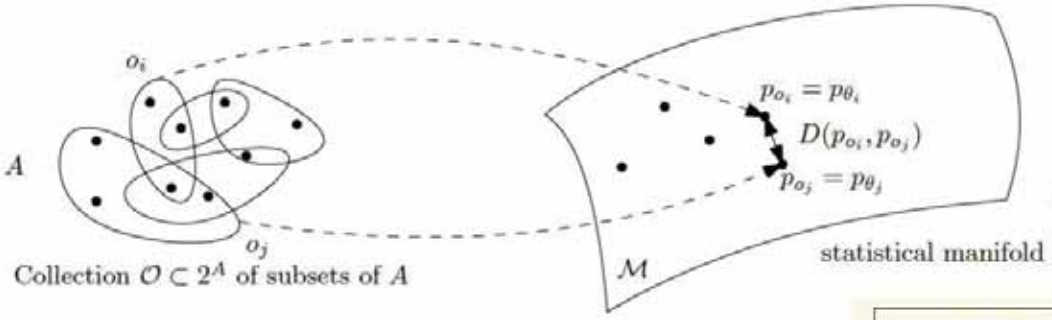
Let $\{p_\theta\}$ be a parametric family of densities and $\eta = \eta(\theta)$ a **new parameterization** (differentiable and invertible: $\theta = \theta(\eta)$):

- **Maximum Likelihood Estimator (MLE) is equivariant**: $\hat{\eta}_{\mathrm{MLE}} = \widehat{\eta(\theta)}_{\mathrm{MLE}} = \eta\left(\hat{\theta}_{\mathrm{MLE}}\right)$
- **Fisher Information Matrix (FIM) is covariant**: $I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top I_\theta(\theta(\eta)) \left[\frac{\partial \theta_i}{\partial \eta_j}\right]$
- **Fisher-Rao geodesic distance is invariant**: $\rho_{\mathrm{FR}}(p_{\theta_1}, p_{\theta_2}) = \rho_{\mathrm{FR}}(p_{\eta_1}, p_{\eta_2})$

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10)

IGSE: Information-Geometric Set Embedding

Embed subsets onto a statistical manifold



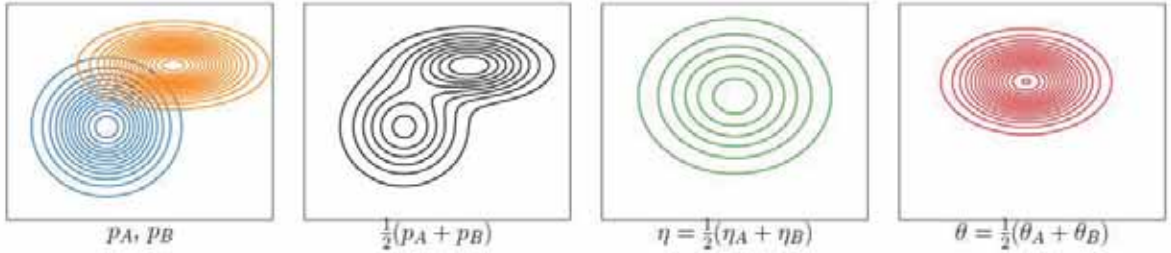
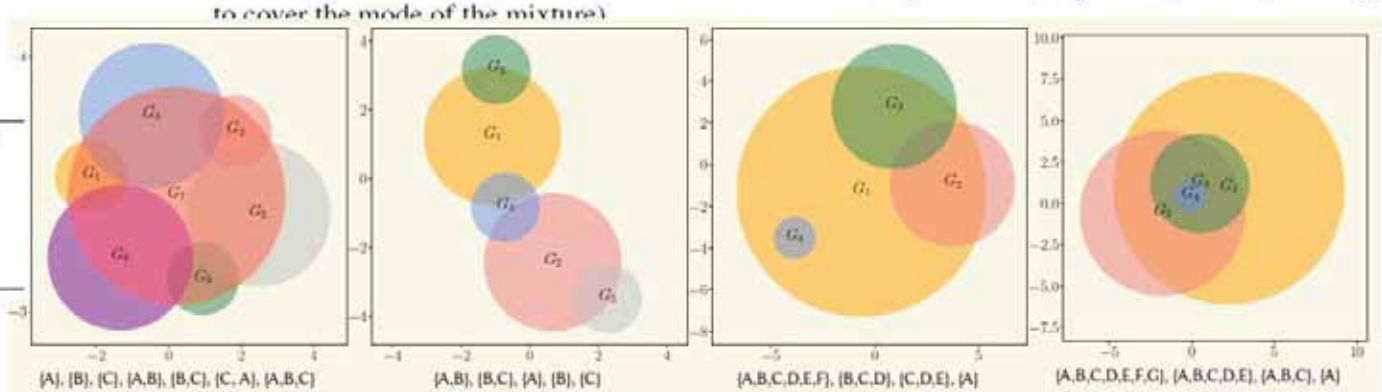


Figure : From left to right: (1) Two Gaussian distributions; (2) their mixture; (3) Their m-centroid (center of mass in the expectation parameters; mixture centroid tending to cover the support of the mixture); (4) Their e-centroid (center of mass in the natural parameters; exponential centroid tending to cover the mode of the mixture)

Table 1: Three different families of subsets

$\mathcal{O}_1$	{A}, {B}, {C}, {A,B}, {B,C}, {C, A}, {A,B,C}
$\mathcal{O}_2$	{A,B}, {B,C}, {A}, {B}, {C}
$\mathcal{O}_3$	{A,B,C,D,E,F}, {B,C,D}, {C,D,E}, {A}
$\mathcal{O}_4$	{A,B,C,D,E,F,G}, {A,B,C,D,E}, {A,B,C}, {A}



Embedding subsets onto a isotropic Gaussian manifold wrt. Jensen-Shannon divergence

https://arxiv.org/abs/1911.12463 https://informationgeometry.xyz/IGSE/

Bhattacharyya α-distances: Gaussian distributions

- **Log-density**: $l(x;\mu,\Sigma) := \log p(x;\mu,\Sigma)$ with $p(x;\mu,\Sigma) := \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$.
- **Partial factorization** (exponential family) to get **parameter barycenter** $\lambda_\alpha = (\mu_\alpha, \Sigma_\alpha)$ (with $\lambda = (\mu,\Sigma)$): $\theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, \frac{1}{2}\lambda_M^{-1}\right)$ and $\lambda(\theta) = (\frac{1}{2}\theta_M^{-1}\theta_v, \frac{1}{2}\theta_M^{-1})$ yields

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

- Apply **generic Bhattacharyya distance formula for exponential families** (choose any $\omega \in \mathbb{R}^d$, eg, $\omega = 0$):

$$\begin{aligned} &D_{B,\alpha}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] \\ &\quad = \alpha l(\omega;\mu_1,\Sigma_1) + (1-\alpha) l(\omega;\mu_2,\Sigma_2) - l(\omega;\mu_\alpha,\Sigma_\alpha) \end{aligned}$$

- Recover the classical formula:

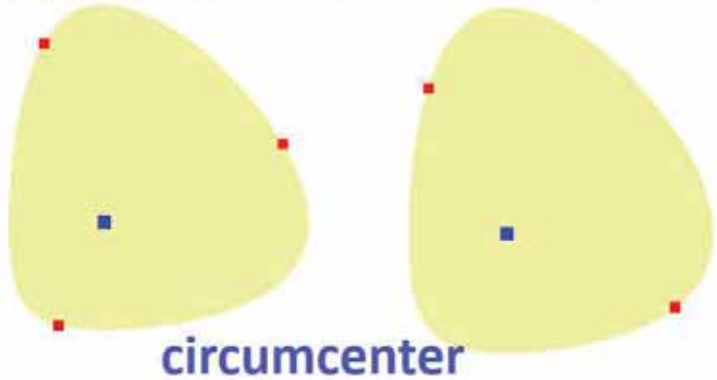
$$\boxed{D_{B,\alpha}[p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}] = \frac{1}{2}\left(\alpha\mu_1^\top\Sigma_1^{-1}\mu_1 + (1-\alpha)\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^\alpha|\Sigma_2|^{1-\alpha}}{|\Sigma_\alpha|}\right)}$$

Inside/outside a Bregman ball? or on a Bregman sphere?

Unique **Bregman ball** passing through 2 to (d+1) points

$$\mathrm{Ball}_F(\theta_c, r) := \{\theta_x \in \Theta \ : \ B_F(\theta_x : \theta_c) \leq r\}$$

Examples of extended Kulback-Leibler balls:



To determine if a point falls *inside/outside a Bregman ball*, calculate the **sign of a (d+2)x(d+2) determinant**:

Negative = inside, Zero= co-spherical, positive = outside

$$\boxed{\mathrm{InBregmanSphere}_F(p_1, \ldots, p_{d+1}; p) := \mathrm{sign}\left(\begin{vmatrix} 1 & \ldots & 1 & 1 \\ p_1 & \ldots & p_{d+1} & x \\ F(p_1) & \ldots & F(p_{d+1}) & F(x) \end{vmatrix}\right)}$$

$$\mathrm{InBregmanSphere}_F(p_1, \ldots, p_{d+1}; p) : \begin{cases} = -1 & \Leftrightarrow \quad p \in \mathrm{InBregmanSphere}_F^{\circ}(p_1, \ldots, p_{d+1}; p) \\ = 0 & \Leftrightarrow \quad p \in \partial\mathrm{InBregmanSphere}_F(p_1, \ldots, p_{d+1}; p) \\ = +1 & \Leftrightarrow \quad p \notin \mathrm{InBregmanSphere}_F(p_1, \ldots, p_{d+1}; p) \end{cases}$$

Do not need to explicitly calculate the Bregman circumcenter!

Bregman voronoi diagrams, Discrete & Computational Geometry 44.2 (2010): 281-307.

Inductive arithmetic-geometric mean (AGM)

- An **inductive mean** is defined as the limit of a sequence
- The **arithmetic-geometric mean** (AGM)

$$\begin{aligned} a_0 &= a \\ g_0 &= b \end{aligned} \quad \Rightarrow \quad \begin{aligned} a_{n+1} &= \frac{a_n + g_n}{2} \\ g_{n+1} &= \sqrt{a_n g_n} \end{aligned}$$

- AGM is a mean **in-between** the geometric mean (G) and the arithmetic mean (A, G ≤A), hence its name: $g_0 \leq \ldots \leq g_n \leq \ldots \leq \mathrm{AGM}(a,b) \leq \ldots \leq a_n \leq \ldots \leq a_0$
- AGM is expressed in closed-form using **the elliptic integral of the first kind** K(t):

$$\begin{aligned} \mathrm{AGM}(a,b) &= \lim_{n\to\infty} a_n = \lim_{n\to\infty} g_n \\ &= \frac{\pi}{4}\frac{a+b}{K\left(\frac{a-b}{a+b}\right)} \end{aligned}$$

$$K(t) := \int_0^{\pi/2} \frac{1}{\sqrt{1 - t\sin^2\theta}} d\theta, \ 0 \leq t < 1$$

$$K(t) = \frac{\pi}{2\mathrm{AGM}(1, \sqrt{1-t^2})}$$

- The **Hellinger divergence** between two Cauchy densities *has closed-form* relying on K(t) [arXiv 2101.12459 2021]

Infinite powers of GEOMETRIZATION!

- **Geometrizing** problems in engineering/theory brings **invariance principles** and yields **geometric structures**.
- **Coordinates** are necessary for **calculus**, and geometry yields invariance of calculus with respect to coordinate transformations: **intrinsic calculus**.
- Geometry can be carved using tools in space (compass & ruler), plotted in coordinate charts, or imagined with (abstract) pictures in human heads!
- A geometric structure emanates from some invariance of a given problem

(eg., geometry of family of distributions invariant by sufficient Markov kernel mappings = information geometry) but the geometric structure can be used to **any other problem domains**: **Geometrization yields abduction.**

- Geometry of models: **Information geometry** (Amari), **Geometrostatistics** (Chentsov/Kolmogorov), **Geometrothermodynamics** (Ruppeiner), etc.

An elementary introduction to information geometry, arXiv:1808.08271

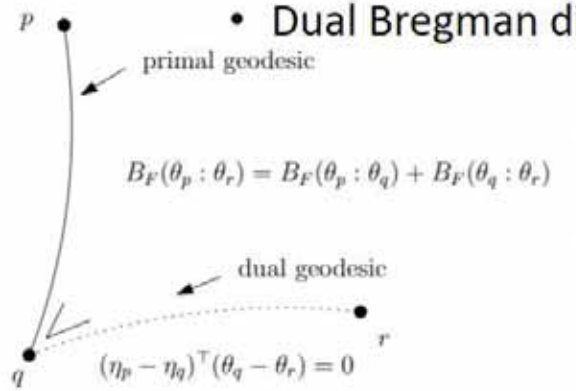
Information projections and Bregman projections

- (M,g, ∇) manifold with *metric tensor* g and *torsion-free affine connection* ∇
- Affine connection defines **geodesics** (= autoparallel curves): γ_{pq}^{∇}
- Orthogonal (wrt g) **information projection** of point p to set S:

$$\boxed{\mathrm{Proj}_S^{\nabla}(p) := \left\{ q \in S \ : \ \gamma_{pq}^{\nabla} \perp_g S \right\}}$$

- **Bregman manifold** induced by a smooth strictly convex function F(θ):
  - F(θ) → $(M, {}^F\nabla, {}^Fg)$, Convex conjugate F*(η) induces **dual structure**: $(M, {}^{F^*}\nabla, {}^{F^*}g)$
  - Dually flat space (DFS, dual connections coupled to the metric): ${}^{F^*}\nabla = {}^F\nabla^*$ ${}^{F^*}g = {}^Fg^*$
    - Dual affine coordinate systems: $\eta = \nabla F(\theta)$ $\theta = \nabla F^*(\eta)$
    - Dual Bregman divergences: B_F B_{F^*}

Dual metric



Pythagorean theorem in DFS: **Projection = Divergence minimization**

$$\boxed{\mathrm{Proj}_S^{\nabla}(p) := \left\{ q \in S \ : \ \gamma_{pq}^{\nabla} \perp_g S \right\} \Longleftrightarrow \mathrm{Proj}_S^{\nabla}(p) = \arg\min_{q \in S} B_F(q : p)}$$

Orthogonality in DFS: $(\theta_p - \theta_q)^\top(\eta_q - \eta_r) = 0$

${}^F\nabla$-projection is **unique** if S is ${}^F\nabla^*$-autoparallel=η-flat

What is... an information projection?, Notices of the AMS 65.3 (2018): 321-324

em-algorithm: Minimization by alternating projections

- **Information divergence view**:

P and Q convex sets

$$
\begin{aligned}
p_0 &\in P \\
q_i &:= \min_{q \in Q} D_{\mathrm{KL}}[p_i : q] \\
p_{i+1} &:= \min_{p \in P} D_{\mathrm{KL}}[p : q_i]
\end{aligned}
$$

Theorem

$$\boxed{\lim_{n \to \infty} D_{\mathrm{KL}}[p_n : q_n] = \inf_{p \in P, q \in Q} D_{\mathrm{KL}}[p : q]}$$

INFORMATION GEOMETRY AND ALTERNATING MINIMIZATION PROCEDURES

I. Csiszár and G. Tusnády*)

Received: Revised version: July 4, 1983

*) This paper was presented at the Colloquium on Information Theory, Budapest, August 1981.

- **Geometric geodesic projection view** (wrt Fisher information metric, FIM):

$$
\begin{aligned}
p_0 &\in P \qquad {}^{m}\nabla \\
q_i &:= \text{FIM orthogonal } m\text{-projection}(p_i) \\
p_{i+1} &:= \text{FIM orthogonal } e\text{-projection}(q_i) \qquad {}^{e}\nabla
\end{aligned}
$$

The EM Algorithm and Information Geometry in Neural Network Learning

Shun-ichi Amari
Department of Mathematical Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan

Neural Computation **7, 13–18** (1995) © 1994 Massachusetts Institute of Technology

Csiszár and Tusnády, Information Geometry and Alternating Minimization Problems, Statistics & Decision (1984).
Amari, Information geometry of the EM and em algorithms for neural networks, Neural networks (1995)
What is... an information projection?, Notices of the AMS 65.3 (2018): 321-324

Information radius as variational divergences

(M, D)-information	Definition $I_{M,D}[p,q] := \min_c\{M(D[p:c], D[q:c])\}$
Information radius of order 1 aka Jensen-Shannon divergence	$I_{A,\mathrm{KL}}[p,q] = \frac{1}{2}D_{\mathrm{KL}}[p : \frac{p+q}{2}] + \frac{1}{2}D_{\mathrm{KL}}[q : \frac{p+q}{2}]$
Information radius of order α aka Sibson's information radius	$I_{M_\alpha^R, D_\alpha^R}[p,q] = \frac{\alpha}{\alpha-1}\log_2 \int_{\mathcal{X}} \left(\frac{p(x)^\alpha + q(x)^\alpha}{2}\right)^{\frac{1}{\alpha}} \mathrm{d}\mu(x)$
Bregman information skewed Jensen–Bregman divergence	$I_{A,B_F}(\theta_1,\theta_2) = \frac{1}{2}B_F(\theta_1 : \frac{\theta_1+\theta_2}{2}) + \frac{1}{2}B_F(\theta_2 : \frac{\theta_1+\theta_2}{2})$ $I_{A_\beta,B_F}(\theta_1:\theta_2) = (1-\beta)B_F(\theta_1 : (1-\beta)\theta_1 + \beta\theta_2) + \beta B_F(\theta_2 : (1-\beta)\theta_1 + \beta\theta_2)$
Chernoff minimax discrimination	$I_{\max, D_{\mathrm{KL}}^*}[p,q] = \min_c \max\{D_{\mathrm{KL}}[c:p], D_{\mathrm{KL}}[c:q]\}$
Amari's α-risk	$I_{A,D_\alpha}[p,q] = \frac{1}{2}D_\alpha[p : (pq)_\alpha^A] + \frac{1}{2}D_\alpha[q : (pq)_\alpha^A]$
Annealing geometric paths $(pq)_{G_\beta}$	minimizer of $I_{G_\beta, D_{\mathrm{KL}}^*}[p:q]$

arithmetic mean	$A(a,b) = \frac{a+b}{2}$
weighted arithmetic mean	$A_\beta(a,b) = (1-\beta)a + \beta b$
geometric mean	$G_\beta(a,b) = a^{1-\beta}b^\beta$
maximum (mean)	$\mathrm{MAX}(a,b) = \max\{a,b\}$
Rényi's α-mean	$M_\alpha^R(a,b) = \frac{1}{\alpha-1}\log_2\left(\frac{2^{(\alpha-1)a}+2^{(\alpha-1)b}}{2}\right)$
Amari's α-integration	$(pq)_\alpha^A(x) \propto f_\alpha^{-1}(\frac{1}{2}f_\alpha(p(x)) + \frac{1}{2}f_\alpha(q(x)))$, α-representation $f_\alpha(u) = \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}}$
M-mixture	$(pq)_{M_\beta}(x) \propto M_\beta(p(x), q(x))$

Kullback-Leibler divergence	$D_{\mathrm{KL}}[p:q] = \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$
reverse Kullback-Leibler divergence	$D_{\mathrm{KL}}^*[p:q] = \int_{\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\mu(x) = D_{\mathrm{KL}}[q:p]$
Rényi's α-divergence	$D_\alpha^R[p:q] = \frac{1}{\alpha-1}\log_2\left(\int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}\mu(x)\right)$
Amari's α-divergence	$D_\alpha^A[p:q] = \frac{4}{1-\alpha^2}(1 - \int_{\mathcal{X}} p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\mathrm{d}\mu(x))$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius, Entropy 23.4 (2021)

Statistical divergence versus information

- **Information** I_D is a **variational divergence** with respect to divergence D:

$$I_D[p,q] := \min_c \frac{1}{2}\left(D[p:c] + D[q:c]\right)$$

- When more than two distributions, information is also called **generalized variational divergence** or **diversity index** (eg, Jensen diversity index)

Examples:

- **Information radius** (Jensen-Shannon divergence) = variational Kullback-Leibler divergence
- **Bregman information** = variational Bregman divergence
- **Chernoff information** = variational minmax reverse KL divergence

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy* (2021)

Integrating stochastic models, mixtures and clustering

- For a statistical dissimilarity D, define the **D-optimal integration** of n weighted probability distributions as the minimizer of

$$R_D[q(x)] := \sum_{i=1}^{n} w_i D[p_i(x) : q(x)]$$

- **Theorem**: Optimal integration for the α-divergences are the **α-mixtures**:

$$\min_{q(x)} R_{D_\alpha}[q(x)] = m_\alpha(p_1, \ldots p_n; w_1, \ldots, w_n)$$

$$m_\alpha(p_1, \ldots p_n; w_1, \ldots, w_n) := \frac{f_\alpha^{-1}\left(\sum_{i=1}^{n} w_i f_\alpha(p_i(x))\right)}{\int f_\alpha^{-1}\left(\sum_{i=1}^{n} w_i f_\alpha(p_i(x))\right) \mathrm{d}\mu(x)}$$

$$f_\alpha[p(s)] = \begin{cases} \frac{2}{1-\alpha} p(s)^{(1-\alpha)/2}, & \alpha \neq 1 \\ \log p(s), & \alpha = 1 \end{cases}$$

Amari, Integration of Stochastic Models by Minimizing α-Divergence, Neural Computation 2007
On clustering histograms with k-means by using mixed \alpha-divergences, Entropy 2014

Intersection of Bregman balls and spheres

Unique **Bregman balls (2 kinds)** passing through **2 to (d+1) points** (in general position)

$$\mathrm{Ball}_F(\theta_c, r) := \{\theta_x \in \Theta \ : \ B_F(\theta_x : \theta_c) \leq r\}$$

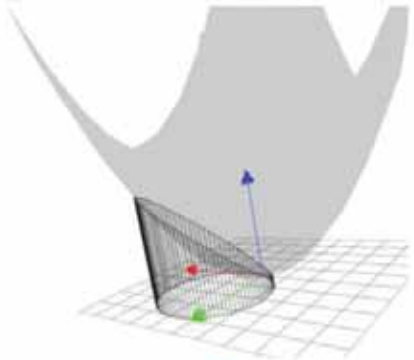
$$\mathrm{Ball}^*_F(\theta_c, r) := \{\theta_x \in \Theta \ : \ B_F(\theta_c : \theta_x :) \leq r\}$$

Bregman divergence:

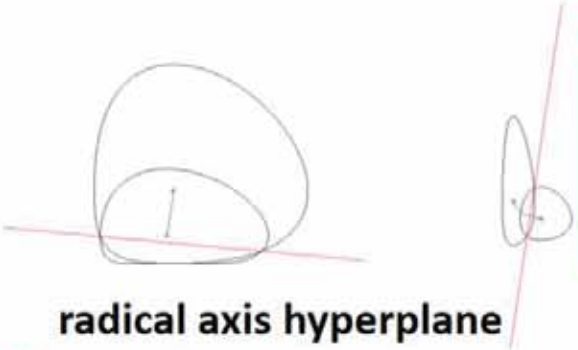
$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$$

Intersection of two (d-1)-dim. Bregman balls = (d-2)-dim. Bregman ball

(proof using the potential lifting transform that generalizes the Euclidean paraboloid transform)



Lifting transform



Intersection of two (d-1)-dim Bregman sphere is a (d-2)-dim Bregman sphere which lies in the radical axis hyperplane

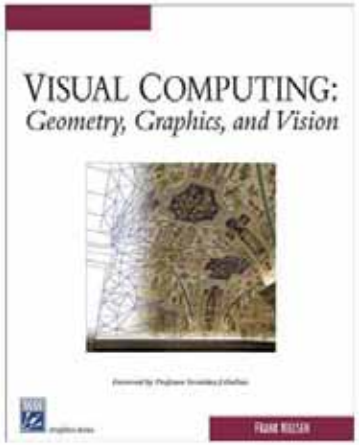
Distinct **Bregman disks** (d=2) intersects in at most 2 points (=**pseudo-circles**):

Calculate the convex hull of n Bregman disks in O(nlog h), where h is output size

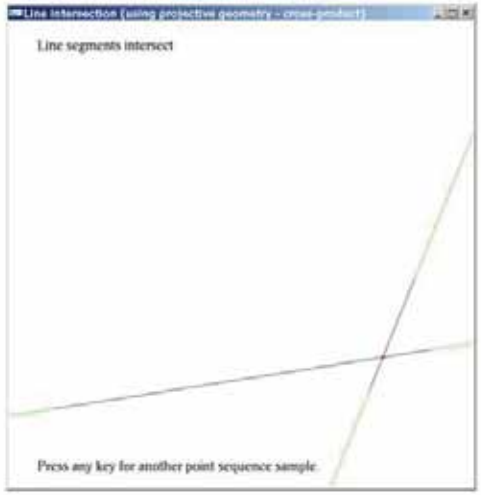
Bregman Voronoi diagrams, Discrete & Computational Geometry 44.2 (2010): 281-307.

Bregman vantage point trees for efficient nearest neighbor queries, IEEE ICME 2009,

An output-sensitive convex hull algorithm for planar objects, IJCGA 1998



$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} \simeq \begin{bmatrix} \frac{x}{w} \\ \frac{y}{w} \end{bmatrix}, \text{for } w \neq 0.$$



Thus, to check whether two line segments $[\mathbf{pq}]$ and $[\mathbf{rs}]$ intersect or not, we compute the intersection point $\mathbf{v} = \mathbf{l}_1 \times \mathbf{l}_2$ of the two lines $\mathbf{l}_1 = \mathbf{p} \times \mathbf{q}$ and $\mathbf{l}_2 = \mathbf{r} \times \mathbf{s}$ supporting the respective line segments. That is,

$$\boxed{\mathbf{v} = (\mathbf{p} \times \mathbf{q}) \times (\mathbf{r} \times \mathbf{s}).} \tag{3.29}$$

Then, we check that inhomogeneous point $\mathbf{v}$ (obtained from homogeneous point $\mathbf{v}$ by perspective division) belongs to both line segments.

Invariance of f-divergences:

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

f convex, strictly convex at 1 with f(1)=0

- Invariance of f-divergences to diffeomorphisms m of the sample

$X \sim p, Y \sim q$ $\quad Y = m(X)$ $\Rightarrow$ $\boxed{I_f[q_1 : q_2] = I_f[p_1 : p_2]}$

- In particular, for parametric densities: $Y_\sigma = m(X_\lambda)$

$$I_f[p_\lambda : p_{\lambda+\mathrm{d}\lambda}] = I_f[q_\sigma : q_{\sigma+\mathrm{d}\sigma}] = \frac{1}{2} f''(1) \mathrm{d}\lambda^\top I_\lambda(\lambda) \mathrm{d}\lambda = \frac{1}{2} f''(1) \mathrm{d}\sigma^\top I_\sigma(\sigma) \mathrm{d}\sigma$$

(= invariance of **Fisher length element** by reparameterization)

Example: $X \sim \mathrm{Exponential}(\lambda) \Leftrightarrow Y = m(X) = \sqrt{X} \sim \mathrm{Rayleigh}\left(\sigma = \frac{1}{\sqrt{2\lambda}}\right)$.

$p_\lambda(x) = \lambda \exp(-\lambda x)$ with support $\mathcal{X} = [0, \infty)$.

$q_\sigma(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ with support $\mathcal{X} = [0, \infty)$.

$$D_{\mathrm{KL}}[q_{\sigma_1} : q_{\sigma_2}] = \log\left(\frac{\lambda_2^2}{\lambda_1^2}\right) + \frac{\sigma_1^2 - \sigma_2^2}{\sigma_2^2}$$

$\Rightarrow$

$$\begin{aligned} D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] &= D_{\mathrm{KL}}[q_{\frac{1}{\sqrt{2\lambda_1}}} : q_{\frac{1}{\sqrt{2\lambda_1}}}], \\ &= \log \frac{2\lambda_1}{2\lambda_2} + 2\lambda_2 \left(\frac{1}{2\lambda_1} - \frac{1}{\lambda_2}\right) \\ &= \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_2}{\lambda_1} - 1. \end{aligned}$$

An elementary introduction to information geometry, 2018 https://arxiv.org/abs/1808.08271

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

Invariance principles in information geometry

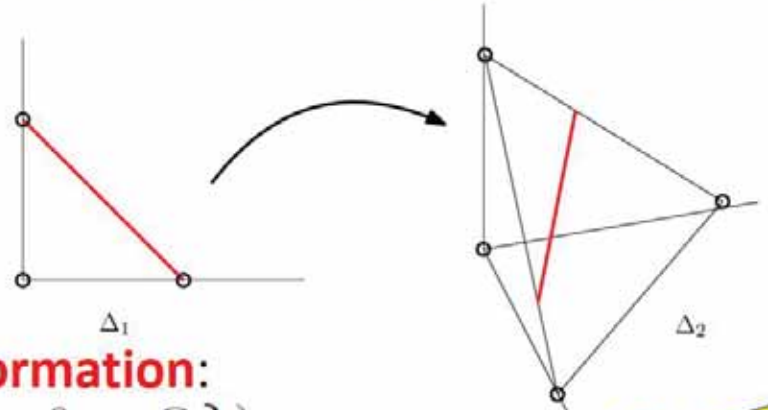
- **<u>Finite discrete case (= family of D-dimensional categorical distributions):</u>**
  - Geometry invariant by **permutations** of atoms (≠optimal transport):

$$\forall \sigma \in P, \quad \text{Geometry}\left(\{\Pr(X = x_i) \ : \ x \in \Delta_D\}\right) = \text{Geometry}\left(\{\Pr(X = x_{\sigma(i)}) \ : \ x \in \Delta_D\}\right)$$

  - Geometry invariant by **imbedding** into higher-dimensional simplex (Markov morphism):

$$i : \Delta_D \to \Delta_{D'}$$

$$\text{Geometry}(\Delta_D) = \text{Geometry}(i(\Delta_D))$$



- **<u>Continuous case:</u>**
  - Geometry invariant by invertible **parameter transformation**:

$$\text{Geometry}\left(\{p_\lambda : \lambda \in \Lambda\}\right) = \text{Geometry}\left(\{p_\theta : \theta \in \Theta\}\right)$$

$$D[p_{\lambda_1} : p_{\lambda_2}] = D[p_{\theta(\lambda_1)} : p_{\theta(\lambda_2)}], \quad \forall \theta = \theta(\lambda) \leftrightarrow \lambda = \lambda(\theta)$$

f-divergences

  - Geometry invariant by **monotonically smooth function** y=t(x):

$$\text{Geometry}\left(\{p_\lambda(x) : \lambda \in \Lambda\}\right) = \text{Geometry}\left(\{q_\lambda(y) : \lambda \in \Lambda\}\right)$$

$$p_\lambda(x) = q_\lambda(t(x)) \det(\nabla_x t(x))$$

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10), 1100

Complex parameter for Cauchy distributions

$$p_{l,s}(x) := \frac{1}{\pi s\left(1+\left(\frac{x-l}{s}\right)^2\right)} = \frac{s}{\pi(s^2+(x-l)^2)}. \quad \xrightarrow{\theta = l + is} \quad p_\theta(x) = \frac{|\mathrm{Im}(\theta)|}{\pi|x-\theta|^2}.$$

Lemma **(Invariance of Cauchy f-divergence under $\mathrm{SL}(2,\mathbb{R})$)** *For any $A \in \mathrm{SL}(2,\mathbb{R})$ and $\theta_1, \theta_2 \in \mathbb{H}$, we have*

$$I_f(p_{A.\theta_1} : p_{A.\theta_2}) = I_f(p_{\theta_1} : p_{\theta_2}).$$

$$\mathrm{SL}(2,\mathbb{R}) := \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \; : \; a,b,c,d \in \mathbb{R}, ad - bc = 1 \right\}$$

Theorem *The f-divergence between two univariate Cauchy densities is symmetric and expressed as a function of the chi-squared divergence:*

$$I_f(p_{\theta_1} : p_{\theta_2}) = I_f(p_{\theta_2} : p_{\theta_1}) = h_f(\chi(\theta_1, \theta_2)), \quad \theta_1, \theta_2 \in \mathbb{H}.$$

arXiv:2101.12459

Inner product versus linear symplectic form

- **Inner product space** (V,G) = vector space V equipped with **inner product** G, $G : V \times V \to \mathbb{R}$: **positive definite symmetric bilinear form**. G can measure vector magnitudes and angles (orthogonality):
  - Symmetric $\forall v_1, v_2 \in V, \quad G(v_1, v_2) = G(v_2, v_1)$
  - Positive definite $\forall v \in V, \quad G(v, v) \geq 0$ with equality iff v=0

- **Symplectic vector space** (V,ω) = vector space V equipped with a **linear symplectic form** $\omega : V \times V \to \mathbb{R}$: **non-degenerate skew-symmetric bilinear form:**
  - Skew-symmetric: $\omega(v_1, v_2) = -\omega(v_2, v_1)$
  - Non-degenerate: $\forall v_2 \in V \ \omega(v_1, v_2) = 0 \Rightarrow v_1 = 0$

Symplectic vector spaces have even dimensions, and ω can measure the skew-orthogonality: $v_1 \perp_\omega v_2 \Leftrightarrow \omega(v_1, v_2) = 0$. In particular, $\forall v \in V, \quad \omega(v, v) = 0$

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10), 1100

Two isometry theorems in Riemannian geometry

- A smooth connected Riemannian manifold (M,g) yields a metric space (M,ρ) with an isometry (= distance preserving transformation).

- **Theorem (MSP, 1957): The differentiable manifold can be canonically reconstructed from the metric space**

  (= reconstruct g from ρ)

- **Theorem (MS, 1939): The isometry group of a Riemannian manifold is a Lie group**

Sumner Byron Myers (1910-1955)

Norman Earl Steenrod (1910-1971)

Richard Sheldon Palais (1931-)

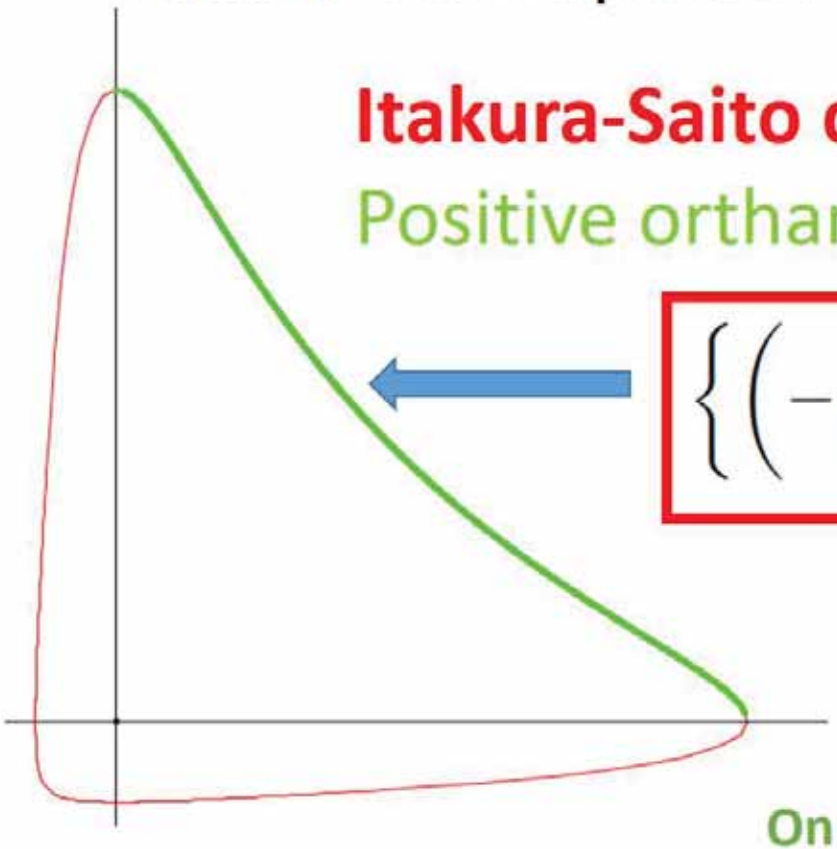
Myers, Steenrod (1939), "The group of isometries of a Riemannian manifold"
Palais, (1957), "On the differentiability of isometries"

Itakura-Saito divergence: $D_{\mathrm{IS}}[p,q] = D_{\mathrm{IS}}[p_x,q_x] + D_{\mathrm{IS}}[p_y,q_y],$

(separable divergence)

$$= \frac{p_x}{q_x} - \log\frac{p_x}{q_x} - 1 + \frac{p_y}{q_y} - \log\frac{p_y}{q_y} - 1$$

Itakura-Saito sphere: $\mathrm{sphere}_{\mathrm{IS}}(\theta_c, r) := \{\theta \ : \ D_{\mathrm{IS}}[\theta_c,\theta] = r\}.$



Itakura-Saito circle of center (0.5,0.5) and radius r=1

Positive orthant parametric equation:

$$\boxed{\left\{\left(-\frac{c_x}{W_0(-e^{-u-1})}, -\frac{c_y}{W_0(-e^{-(r-u)-1})}\right), \quad u \in [0,r]\right\}}$$

Branches of the Lambert W function:

$$\exists k \in \mathbb{Z}, \quad xe^x = y \Rightarrow x = W_k(y)$$

Squared Mahalanobis divergences are the only Jeffreys-Bregman divergences which are Bregman divergences

- D univariate convex functions $F_i : \theta_i \in \Theta_i \to \mathbb{R}$ with convex domain Θ_i. Build **separable multivariate function** $F : \theta = (\theta_1, \ldots, \theta_D) \in \Theta = \prod_{i=1}^D \theta_i \to \mathbb{R}$ defined by $F(\theta) = \sum_{i=1}^D F_i(\theta_i)$. F **is convex iff** Θ **is convex**. Thus when all univariate functions F_i's are **identical**, F is convex
- **Jeffreys-Bregman divergence**[1] (symmetrization of Bregman divergences) $\mathrm{JB}(\theta, \theta') := B_F(\theta : \theta') + B_F(\theta' : \theta) = (\theta - \theta')^\top (\nabla F(\theta) - \nabla F(\theta'))$.
- $\mathrm{JB}(\theta, \theta') = B_F(\theta : \theta') + B_{F*}(\eta : \eta') =: B_{\hat{F}}(\hat{\theta} : \hat{\theta}')$ with $\hat{F}(\hat{\theta}) := F(\theta) + F^*(\eta)$ and $\hat{\theta} = (\theta, \eta = \nabla F(\theta))$ and F^* the Legendre-Fenchel convex conjugate. Symmetric $B_{\hat{F}}$ is a Bregman divergence iff $\hat{\Theta} = \Theta \times H$ is convex. Occurs only for **squared Mahalanobis divergences**[2] with $\hat{\Theta} = \mathbb{R}^{2D}$.

[1] Sided and symmetrized Bregman centroids, IEEE TIT 2009. On the Jensen–Shannon Symmetrization of Distances, Entropy 2019.

[2] Bregman voronoi diagrams, DCG 2010.

Retrospective: Jeffreys' invariant prior (1945/46)

An invariant form for the prior probability in estimation problems, H. Jeffreys (submitted 1945, published 1946) https://doi.org/10.1098/rspa.1946.0056

Sir Harold Jeffreys FRS
1891-1989

$$I_1 = \int (\sqrt{dP'} - \sqrt{dP})^2, \quad I_2 = \int \log \frac{dP'}{dP}\, d(P' - P), \qquad (1)$$

- I1: Twice squared Hellinger divergence I2: Jeffreys' divergence

that they are invariant for all transformations of the parameters in the laws and also for all non-singular transformations of x the range of integration being over the

$$I_2 = \lim \Sigma \frac{1}{p_r}\left(\frac{\partial p'_r}{\partial \alpha_i}\Delta\alpha_i\right)\left(\frac{\partial p'_r}{\partial \alpha_k}\Delta\alpha_k\right)$$

$$= g_{ik}\Delta\alpha_i\Delta\alpha_k, \qquad (6)$$

$$g_{ik} = \lim_{\delta x_r \to 0} \Sigma \frac{1}{p_r}\frac{\partial p'_r}{\partial \alpha_i}\frac{\partial p'_r}{\partial \alpha_k}. \qquad (7)$$

(7) Fisher information metric tensor (FIM)

$\| g_{ik} \|^{\frac{1}{2}} d\alpha_1 d\alpha_2 \ldots d\alpha_m$ is invariant.

For Gaussians, $p_r \doteqdot \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{(x_r-\alpha)^2}{2\sigma^2}\right\}\delta x_r$ $\doteqdot 2\left(\frac{\delta\sigma}{\sigma}\right)^2 + \frac{(\delta\alpha)^2}{\sigma^2}$

Jeffreys' invariant prior: and if the prior probability density of the parameters is taken as proportional to $\| g_{ik} \|^{\frac{1}{2}}$ the prior probability over any region will be invariant for all ways of choosing the parameters.

Jeffreys Centroids: A Closed-Form Expression for Positive Histograms and a Guaranteed Tight Approximation for Frequency Histograms. IEEE SPL 2013

Symmetrized KL divergences between GMMs

- **Jeffreys J-divergence** symmetrizes the Kullback-Leibler divergence (KLD):

$$D_J[m, m'] \;:=\; D_{\mathrm{KL}}[m : m'] + D_{\mathrm{KL}}[m' : m] \;=\; \int_{\mathcal{X}} (m(x) - m'(x)) \log \left(\frac{m(x)}{m'(x)} \right) \mathrm{d}x.$$

- KLD between **Gaussian mixture models** (GMMs) is provably **not analytic**
- J-divergence between densities of an exponential family *does not use explicitly* the cumulant function but difference of natural and expectation parameters:

$$D_J(p_\theta, p_{\theta'}) = (\theta' - \theta)^\top (\eta' - \eta).$$

- Approximate J-divergence: convert GMMs m into pairs of exponential family densities parameterized in natural and expectation parameters using extensions of Maximum Likelihood Estimator (MLE) and Score Matching Estimator (SME):

$$\begin{array}{l} m \to (\tilde{p}_{\bar{\theta}_{\mathrm{MLE}}}, \tilde{p}^{\bar{\eta}_{\mathrm{SME}}}) \\ m' \to (\tilde{p}_{\bar{\theta}'_{\mathrm{MLE}}}, \tilde{p}^{\bar{\eta}'_{\mathrm{SME}}}) \end{array} \Longrightarrow \quad D_J(m, m') \approx (\bar{\theta}'_{\mathrm{MLE}} - \bar{\theta}_{\mathrm{MLE}})^\top (\bar{\eta}'_{\mathrm{MLE}} - \bar{\eta}_{\mathrm{MLE}}).$$

Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities, arXiv:2107.05901

SKL = Jeffreys divergence & exponential families

Jeffreys' divergence D_J: arithmetic sum of the forward & reverse Kullback-Leibler divergence

$$\begin{aligned} D_J[m, m'] &:= D_{\mathrm{KL}}[m : m'] + D_{\mathrm{KL}}[m' : m], \\ &= \int_{\mathcal{X}} (m(x) - m'(x)) \log \left(\frac{m(x)}{m'(x)} \right) \mathrm{d}x \end{aligned}$$

For densities belonging to an exponential family, **J-divergence** does not use explicitly the cumulant function (= **symmetrized Bregman divergence**)

$$\boxed{\begin{aligned} D_J(p_\theta, p_{\theta'}) &= D_{\mathrm{KL}}[p_\theta : p_{\theta'}] + D_{\mathrm{KL}}[p_{\theta'} : p_\theta], \\ &= B_F(\theta' : \theta) + B_F(\theta : \theta'), \\ &= (\theta' - \theta)^\top (\eta' - \eta), \\ &= (\theta' - \theta)^\top (\nabla F(\theta') - \nabla F(\theta)), \\ &= (\nabla F^*(\eta') - \nabla F^*(\eta))^\top (\eta' - \eta). \end{aligned}}$$

Symmetrized BD = Bregman divergence only for the **generalized quadratic divergence wrt** $Q \succ 0$ $\quad D_Q(\theta_1 : \theta_2) := (\theta_2 - \theta_1)^\top Q (\theta_2 - \theta_1)$

= squared Mahalanobis distance

$$D^2_{\mathrm{Mahalanobis}}[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] := D_{\Sigma^{-1}}(\mu_1, \mu_2)$$

$$D_{\mathrm{Mahalanobis}}[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] := \sqrt{D_{\Sigma^{-1}}(\mu_1, \mu_2)}$$

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory 55.6 (2009): 2882-2904

Power *metrizable* or *non-metrizable* f-divergences

- f-divergence for a convex generator f with f(1)=0: $I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$.
  (strictly convex at 1)
- <u>Jeffreys divergence</u>:

$$\boxed{D_J[p,q] = D_{\mathrm{KL}}[p:q] + D_{\mathrm{KL}}[q:p]}$$

$$D_J[p,q] = I_{f_J}[p:q] \qquad f_J(u) := (u-1)\log u$$

- <u>Jensen-Shannon divergence</u>: $D_{\mathrm{JS}}(p,q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(p : \frac{p+q}{2}\right) + D_{\mathrm{KL}}\left(q : \frac{p+q}{2}\right)\right)$

$$D_{\mathrm{JS}}[p,q] = I_{f_{\mathrm{JS}}}[p:q] \quad f_{\mathrm{JS}}(u) := \frac{u}{2}\log\frac{2u}{1+u} - \frac{1}{2}\log\frac{1+u}{2}$$

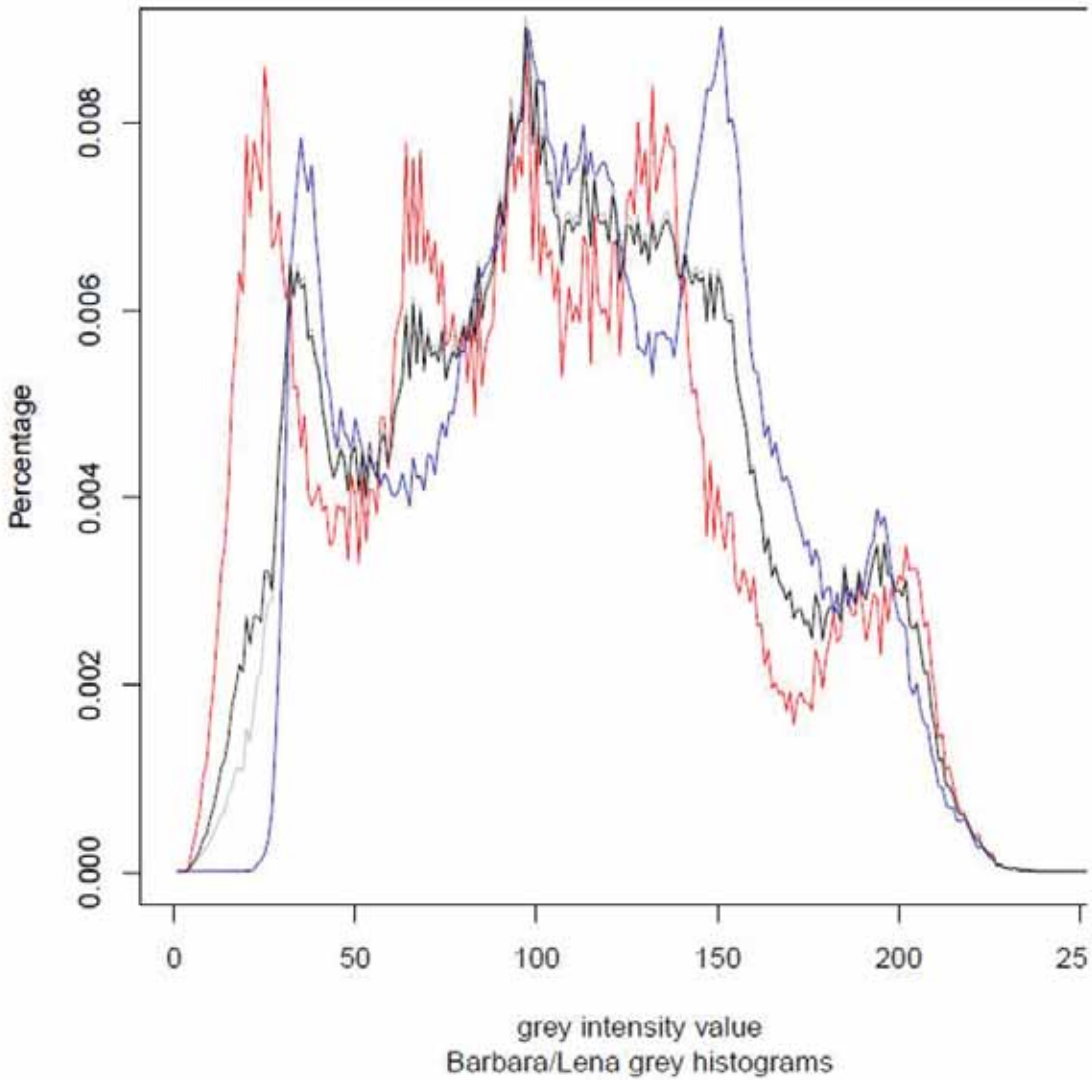
- Symmetric f-divergence is **<u>power metrizable</u>** is there exists a positive power α such that I_f^{α} satisfies the triangle inequality
- **JSD is power metrizable** $\sqrt{D_{\mathrm{JS}}[p,q]}$ is a metric
- **Jeffreys divergence is not power metrizable**:

$$\boxed{\forall \alpha > 0, \exists p,q,r, \quad D_J[p,q]^{\alpha} + D_J[q,r]^{\alpha} \not\geq D_J[p,r]^{\alpha}}$$

Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms IEEE SPL

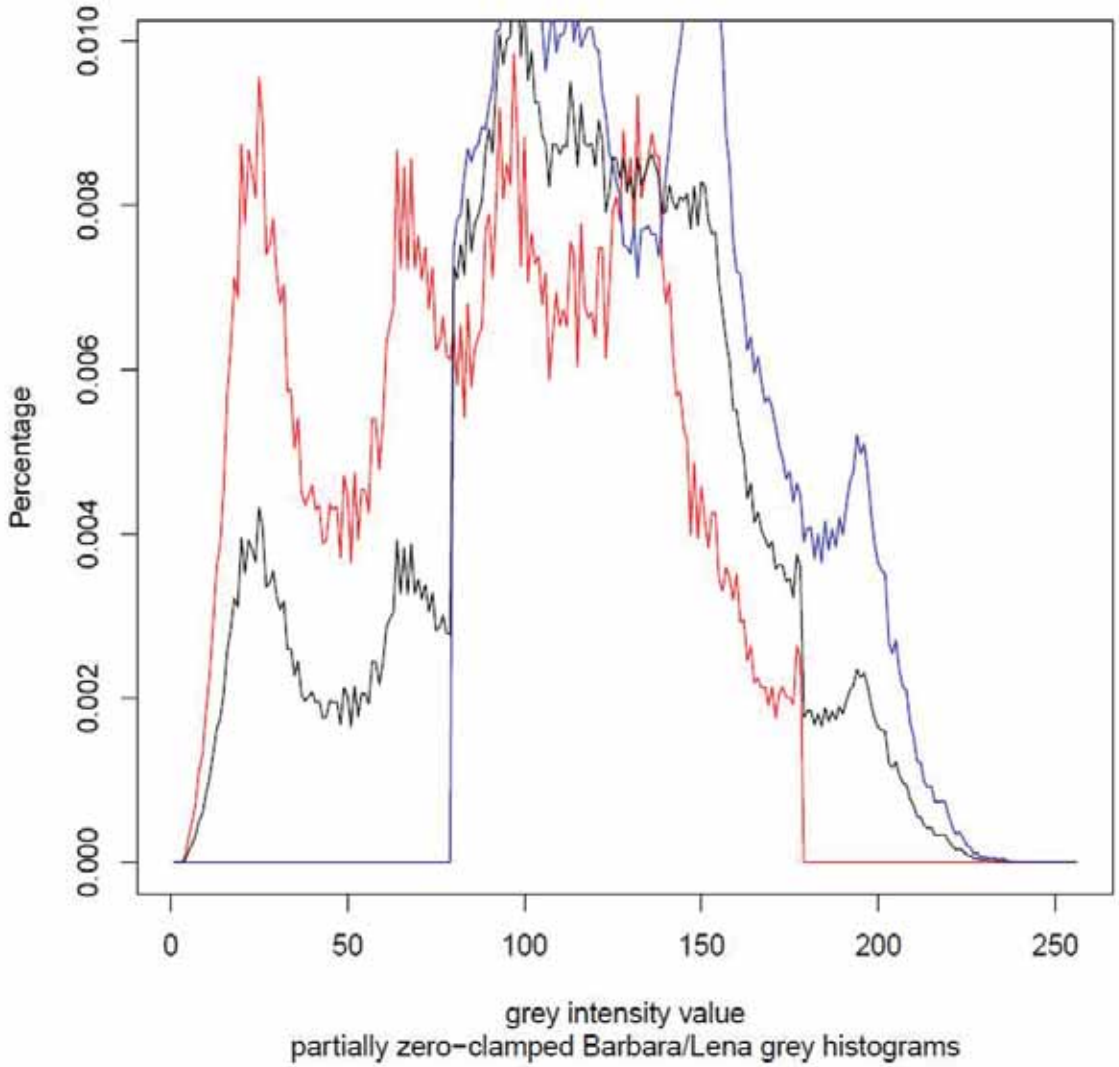
Coinciding supports

Jeffreys vs Jensen−Shannon histogram centroids



Non-matching supports

Jensen−Shannon histogram centroid (non−matching support)



Jensen-Chernoff divergence

- α-skewed Jensen divergence for a strictly convex F:
  $J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2)$
- The **Jensen-Chernoff** divergence is defined by:

$$\mathrm{JC}_F(\theta_1 : \theta_2) = \max_{\alpha \in (0,1)} J_F^{(\alpha)}(\theta_1 : \theta_2)$$

- Jensen-Chernoff amount to a **Bregman divergence**:

$$\mathrm{JC}_F(\boldsymbol{\theta_1} : \boldsymbol{\theta_2}) = B_F(\boldsymbol{\theta_1} : \alpha^* \boldsymbol{\theta_1} + (\mathbf{1} - \alpha^*)\boldsymbol{\theta_2}) = B_F(\boldsymbol{\theta_2} : \alpha^* \boldsymbol{\theta_1} + (\mathbf{1} - \alpha^*)\boldsymbol{\theta_2}),$$

  where $\alpha^* = \arg\max_{\alpha \in (0,1)} J_F^{(\alpha)}(\theta_1 : \theta_2)$.
- Binary hypothesis testing between p_{θ_1} and p_{θ_2} of an exponential family. **Chernoff information**[1]:

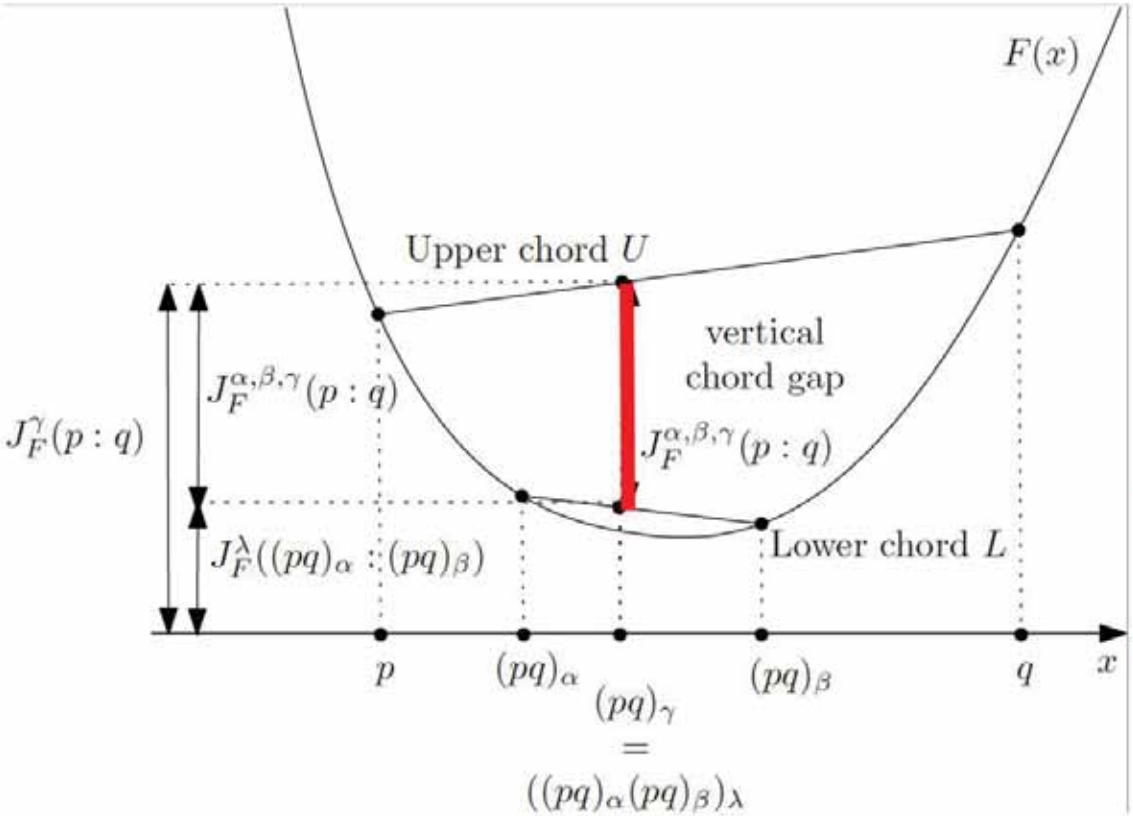
$$C(p_{\theta_1} : p_{\theta_2}) = \max_{\alpha \in (0,1)} -\log \rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \mathrm{JC}_F(\theta_1 : \theta_2),$$

  where ρ_α is the α-**skewed Bhattacharrya coefficient**:

$$\rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \int p_{\theta_1}^{\alpha}(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x)$$

[1] An information-geometric characterization of Chernoff information, IEEE SPL 2013, `https://arxiv.org/abs/1102.2684`

The Jensen chord divergence: Truncated skew Jensen divergences



Linear interpolation (LERP):

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

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

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

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

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

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

(truncated skew Jensen divergence)

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

The two faces of the Jensen-Shannon divergence

- The Jensen-Shannon divergence (Lin 1991) is a **Jensen-Bregman divergence** for the Shannon negentropy generator F=-h:

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

$$\mathrm{JB}_F(\theta:\theta') := \frac{1}{2}\left(B_F\left(\theta:\frac{\theta+\theta'}{2}\right)+B_F\left(\theta':\frac{\theta+\theta'}{2}\right)\right)$$

- The Jensen-Shannon divergence is a **Jensen divergence** for the Shannon negentropy generator F=-h:

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

$$J_F(p,q) = \frac{F(p)+F(q)}{2}-F\left(\frac{p+q}{2}\right).$$

Jensen-Shannon divergence is a **f-divergence** always upper bounded by log 2

$$I_f(p:q) = \int pf\left(\frac{q}{p}\right)\mathrm{d}\mu. \qquad f_{\mathrm{JS}}(u) := -(u+1)\log\frac{1+u}{2}+u\log u.$$

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid, 2020, Entropy 22 (2), 221

On the Jensen–Shannon symmetrization of distances relying on abstract means, 2019, Entropy 21 (5), 485

Jensen-Shannon divergences wrt. abstract means

f-divergence for a convex generator f with f(1)=0: $I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x).$

Ordinary skewed Jensen-Shannon divergence is a f-divergence:

$$\boxed{\begin{aligned} D_{\mathrm{JS}}^{\alpha,\beta}[p:q] &:= (1-\beta) D_{\mathrm{KL}}[p:m_\alpha] + \beta D_{\mathrm{KL}}[q:m_\alpha] \\ &= I_{f_{\mathrm{JS}}^{\alpha,\beta}}[p:q] \\ & f_{\mathrm{JS}}^{\alpha,\beta}(u) = -\left((1-\beta)\log(\alpha u + (1-\alpha)) + \beta u \log\left(\frac{1-\alpha}{u} + \alpha\right)\right) \end{aligned}}$$

Generalization of skewed Jensen-Shannon divergence wrt weighted means M and N

$$\boxed{D_{\mathrm{JS}}^{M_\alpha, N_\beta}[p:q] := N_\beta\left(D_{\mathrm{KL}}(p:(pq)_{M_\alpha}), D_{\mathrm{KL}}(q:(pq)_{M_\alpha})\right)}$$

Statistical M-mixtures: $(pq)_{M_\alpha}(x) = \frac{M_\alpha(p(x), q(x))}{\int M_\alpha(p(x), q(x)) \mathrm{d}\mu(x)}$

On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. Entropy (2019)

Three flavors of the Jensen-Shannon divergence

- **Symmetrization** of the relative entropy:

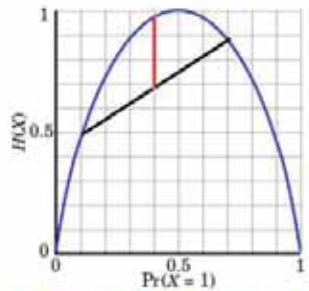
$$D_{\mathrm{JS}}[p,q] \quad := \quad \frac{1}{2}\left(D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q : \frac{p+q}{2}\right]\right)$$

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

https://doi.org/10.3390/e22020221

- **Convexity gap** of Shannon negentropy:

$$D_{\mathrm{JS}}[p,q] \quad := \quad h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2}$$



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

https://doi.org/10.3390/e22020221

- **Diversity index** of two probability measures:

$$D_{\mathrm{JS}}[p,q] \quad := \quad \min_{c \in \mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right)$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius

https://doi.org/10.3390/e23040464

Jensen-Shannon divergence expanded as higher-order triangular discriminations

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

$$\boxed{D_{\mathrm{JS}}[p,q] = \frac{1}{2}\sum_{i=1}^{\infty} \frac{1}{2i(2i-1)} \Delta_{2i}(p,q)}$$

higher-order triangular discriminations: $\Delta_{2i}(p,q) := \int \frac{(p(x)-q(x))^{2i}}{(p(x)+q(x))^{2i-1}} \mathrm{d}\mu(x)$

Harremoës and Vignat, An entropy power inequality for the binomial family, JIPAM. J. Inequal. Pure Appl. Math 4.5 (2003)
Nielsen and Nock, On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 21.1 (2013)

The Jensen-Shannon divergence as a f-divergence

Consider the convex generator:

$$\boxed{f(u) := -\frac{1+u}{2}\log\left(\frac{1+u}{2}\right) + \frac{u}{2}\log(u)}$$

Mid-distribution: $m = \frac{p+q}{2}$

$$\boxed{\mathrm{JS}[p:q] = \frac{1}{2}\mathrm{KL}\left[p : \frac{p+q}{2}\right] + \frac{1}{2}\mathrm{KL}\left[q : \frac{p+q}{2}\right]}$$

f-divergence yields the **Jensen-Shannon divergence**

$$\begin{aligned}
I_f[p:q] &:= \int p f\left(\frac{q}{p}\right) \mathrm{d}\mu, \\
&= \frac{1}{2}\int \left((p+q)\log\frac{2p}{p+q} + q\log\frac{q}{p}\right)\mathrm{d}\mu, \\
&= \frac{1}{2}\int p\log\frac{p}{m}\mathrm{d}\mu + \frac{1}{2}\int q\log\frac{p}{m}\frac{q}{p}\mathrm{d}\mu, \\
&= \frac{1}{2}\int p\log\frac{p}{m}\mathrm{d}\mu + \frac{1}{2}\int q\log\frac{q}{m}\mathrm{d}\mu, \\
&= \frac{1}{2}\mathrm{KL}\left[p : \frac{p+q}{2}\right] + \frac{1}{2}\mathrm{KL}\left[q : \frac{p+q}{2}\right], \\
&=: \mathrm{JS}[p:q].
\end{aligned}$$

Symbolic computation using Maxima:

```
(%i3) f(u):=-((1+u)/2)*log((1+u)/2)+(u/2)*log(u);
      p*f(q/p);
      ratsimp(%);
```

(%o1) $f(u) := \left(-\frac{1+u}{2}\right)\log\left(\frac{1+u}{2}\right) + \frac{u}{2}\log(u)$

(%o2) $p\left(\frac{q\log\left(\frac{q}{p}\right)}{2p} - \frac{\left(\frac{q}{p}+1\right)\log\left(\frac{\frac{q}{p}+1}{2}\right)}{2}\right)$

(%o3) $-\frac{(q+p)\log\left(\frac{q+p}{2p}\right) - q\log\left(\frac{q}{p}\right)}{2}$

On the chi square and higher-order chi distances for approximating f-divergences, IEEE Signal Processing Letters, 2013.
On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid, Entropy 2020, 22(2), 221

2020 Paris Kanellakis Award

The 2020 Paris Kanellakis Theory and Practice Award is presented to Yossi Azar, Andrei Broder, Anna Karlin, Michael Mitzenmacher and Eli Upfal for the discovery and analysis of balanced allocations, known as the power of two choices, and their extensive applications to practice.

Balanced Allocations

Yossi Azar* Andrei Z. Broder† Anna R. Karlin‡ Eli Upfal§

Abstract

Suppose that we sequentially place n balls into n boxes by putting each ball into a randomly chosen box. It is well known that when we are done, the fullest box has with high probability $(1 + o(1)) \ln n / \ln \ln n$ balls in it. Suppose instead, that for each ball we choose two boxes at random and place the ball into the one which is less full at the time of placement. We show that with high probability, the fullest box contains only $\ln \ln n / \ln 2 + O(1)$ balls – exponentially less than before. Furthermore, we show that a similar gap exists in the infinite process, where at each step one ball, chosen uniformly at random, is deleted, and one ball is added in the manner above. We discuss consequences of this and related theorems for dynamic resource allocation, hashing, and on-line load balancing.

The Power of Two Random Choices: A Survey of Techniques and Results

Michael Mitzenmacher * Andréa W. Richa †
Ramesh Sitaraman ‡

1 Introduction

To motivate this survey, we begin with a simple problem that demonstrates a powerful fundamental idea. Suppose that n balls are thrown into n bins, with each ball choosing a bin independently and uniformly at random. Then the *maximum load*, or the largest number of balls in any bin, is approximately $\log n / \log \log n$ with high probability.[1] Now suppose instead that the balls are placed sequentially, and each ball is placed in the least loaded of $d \geq 2$ bins chosen independently and uniformly at random. Azar, Broder, Karlin, and Upfal showed that in this case, the maximum load is $\log \log n / \log d + \Theta(1)$ with high probability [ABKU99].

The important implication of this result is that even a small amount of choice can lead to drastically different results in load balancing. Indeed, having just two random choices (i.e., $d = 2$) yields a large reduction in the maximum load over having one choice, while each additional choice beyond two decreases the maximum load by just a constant factor. Over

Kullback-Leibler divergence: Scale Cauchy family[1]

- Cauchy location-scale family $\{p_{l,s}\}$ from Cauchy standard density p:

$$p(x) = \frac{1}{\pi(1+x^2)}, \quad p_{l,s}(x) = \frac{s}{\pi(s^2+(x-l)^2)}.$$

- Cross-entropy between two scale Cauchy densities p_{s_1} and p_{s_2}:

$$h^{\times}(p_{s_1} : p_{s_2}) = \log \pi \frac{(s_1+s_2)^2}{s_2}$$

- Differential entropy of a location-scale Cauchy density $p_{l,s}$:

$$h(p_{l,s}) = \log 4\pi s$$

- Kullback-Leibler divergence between two scale Cauchy densities p_{s_1} and p_{s_2}:

$$\mathrm{KL}(p_{s_1} : p_{s_2}) = 2\log\left(\frac{s_1+s_2}{2\sqrt{s_1 s_2}}\right) = 2\log\frac{A(s_1,s_2)}{G(s_1,s_2)},$$

where $A(s_1,s_2) = \frac{s_1+s_2}{2}$ = (arithmetic mean) and $G(s_1,s_2) = \sqrt{s_1 s_2}$ (geometric mean), and $A \geq G \Rightarrow \mathrm{KL} \geq 0$

[1] https://tinyurl.com/y5wcd5gu

Kullback-Leibler divergence related to Fisher divergence

Fisher divergence used for score-matching (estimation of unnormalized densities):

$$D_{\mathrm{Fisher}}[p:q] = \frac{1}{2}\int p(x)\,\|\nabla_x \log p(x) - \nabla_x \log q(x)\|_2^2\,\mathrm{d}\mu(x)$$

Kullback-Leibler divergence expressed as a definite integral of Fisher divergences:

$$\boxed{D_{\mathrm{KL}}[p:q] = \int_0^\infty D_{\mathrm{Fisher}}[p * \mathcal{N}(0,\lambda I) : q * \mathcal{N}(0,\lambda I)]\,\mathrm{d}\lambda}$$

where * denotes the convolution: $(f * g)(x) = \int f(y)g(x-y)\mathrm{d}y$

Convergence wrt Fisher divergence is stronger than convergence wrt KLD

Lyu, Siwei. "Interpretation and generalization of score matching." *arXiv preprint arXiv:1205.2629* (2012).

Kullback-Leibler divergence between two densities of different exponential families

First exponential family: $\mathcal{P} = \{p_{\theta_1}(x) = \exp\left({\theta_1}^\top t_{\mathcal{P}}(x) - F_{\mathcal{P}}(\theta_1) + k_{\mathcal{P}}(x)\right) \ : \ \theta_1 \in \Theta_1\}$

Second exponential family: $\mathcal{Q} = \{q_{\theta_2}(x) = \exp\left({\theta_2}^\top t_{\mathcal{Q}}(x) - F_{\mathcal{Q}}(\theta_2) + k_{\mathcal{Q}}(x)\right) \ : \ \theta_2 \in \Theta_2\}$

Kullback-Leibler divergence: $D_{\mathrm{KL}}[p:q] = \begin{cases} \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, & \mathrm{supp}(p) \subset \mathrm{supp}(q) \\ +\infty, & \mathrm{supp}(p) \not\subset \mathrm{supp}(q) \end{cases}$

Example: KLD between $p_{\theta_1} \sim \mathrm{Exponential}(\lambda_1)$ and $q_{\theta_2} \sim \mathrm{Laplacian}(\sigma_2)$

Generic formula

$$\boxed{\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= F_{\mathcal{Q}}(\theta_2) + F^*_{\mathcal{P}}(\eta_1) - E_{p_{\theta_1}}[t_{\mathcal{Q}}(x)]^\top \theta_2 + E_{p_{\theta_1}}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)] \\ &= \log(2\sigma_2) + \log(\lambda_1) - 1 - \left(-\frac{1}{\lambda_1 \sigma_2}\right), \\ &= \frac{1}{\lambda_1 \sigma_2} - \log \frac{1}{\lambda_1 \sigma_2} + \log \frac{2}{e} \end{aligned}}$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, Entropy 2021

Summary: KLD between exp. fam. densities

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] \quad = \quad B_F(\theta(\lambda_2) : \theta(\lambda_1)) = \sum_{i=1}^{s} w_i B_{F_{\omega_i}}(\theta(\lambda_2) : \theta(\lambda_1))$$

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] \quad = \quad \sum_{i=1}^{s} w_i \log\left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)}\right) - (\theta(\lambda_2) - \theta(\lambda_1))^\top \underbrace{\left(\sum_{i=1}^{s} w_i \nabla_\theta F_{\omega_i}(\theta(\lambda_1))\right)}_{\omega_i \text{ and } w_i \text{ such that } = 0}$$

Condition for KLD = **weighted average of log density ratios**:

$$\begin{aligned}\sum_{i=1}^{s} w_i \nabla_\theta F_{\omega_i}(\theta(\lambda_1)) &= -\sum_{i=1}^{s} (w_i(t(\omega_i) - \nabla F(\theta(\lambda_1)))) \\ &= \nabla F(\theta(\lambda_1)) - \sum_{i=1}^{s} w_i t(\omega_i)\end{aligned}$$

$$E_{p_{\lambda_1}}[t(x)] = \sum_{i=1}^{s} w_i t(\omega_i) \quad \Rightarrow \quad \sum_{i=1}^{s} w_i \nabla_\theta F_{\omega_i}(\theta(\lambda_1)) = 0$$

Recovering the KLD formula for multivariate normals

Corollary 3 (KLD between location-scale densities). *The KLD between two regular location-scale densities is equivalent to the f-divergence between one standard location-scale density and another affinely shifted location-scale density:*

$$D_{\mathrm{KL}}(p_{l_1,P_1} : q_{l_2,P_2}) = D_{\mathrm{KL}}\left(p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2}\right) = D_{\mathrm{KL}}\left(p_{P_2^{-1}(l_1-l_2),P_2^{-1}P_1} : q\right). \tag{42}$$

$$D_{\mathrm{KL}}\left(p^{\mathcal{N}} : p^{\mathcal{N}}_{\mu,\Sigma}\right) = \frac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}) + \mu^\top \Sigma^{-1}\mu + \log|\Sigma| - d\right).$$

$$\begin{aligned} D_{\mathrm{KL}}\left(p^{\mathcal{N}}_{\mu_1,\Sigma_1} : p^{\mathcal{N}}_{\mu_2,\Sigma_2}\right) &= D_{\mathrm{KL}}\left(p : p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1),\Sigma_1^{-1}\Sigma_2}\right), \\ &= \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^\top \Sigma_2^{-1}(\mu_2-\mu_1) + \log|\Sigma_1^{-1}\Sigma_2| - d\right). \end{aligned}$$

$$D_{\mathrm{KL}}\left(p^{\mathcal{N}}_{\mu_1,\Sigma_1} : p^{\mathcal{N}}_{\mu_2,\Sigma_2}\right) = D^{\Sigma_2^{-1}}_{\mathrm{Mah}}(\mu_1,\mu_2) + D_{\mathrm{IS}}(\Sigma_1,\Sigma_2).$$

$$D^{Q}_{\mathrm{Mah}}(\mu_1,\mu_2) := \frac{1}{2}(\mu_2-\mu_1)^\top Q(\mu_2-\mu_1), \qquad D_{\mathrm{IS}}(\Sigma_1,\Sigma_2) := \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1 - I) - \log|\Sigma_2^{-1}\Sigma_1|\right),$$

On information projections between multivariate elliptical and location-scale families, arXiv 2101.03839

Monomial exponential families (MEFs)

- MEFs are exponential families with sufficient statistics a monomial x^D:

D is an even integer

$$\boxed{p_{D,\theta}(x) := \exp(\theta x^D - F_D(\theta))}$$

- The cumulant function and convex conjugates are obtained in **closed-form**:

$$\begin{aligned} F_D(\theta) &= \log \int_{\mathbb{R}} \exp(\theta x^D) \mathrm{d}x, \\ &= \log\left(\frac{2\Gamma(1/D)}{D}\right) - \frac{1}{D}\log(-\theta) \end{aligned}$$

$$F_D^*(\eta) = -\log\left(\frac{2\Gamma(1/D)}{D}\right) - \frac{1}{D}(1 + \log(D\eta)$$

Fenchel-Young inequality: $F_D(\theta) + F_D^*(\eta) - \theta\eta = 0.$

- Dual parameterization: $\eta = \nabla F_D(\theta) = -\frac{1}{D\theta} > 0. \Longleftrightarrow \theta = \nabla F_D^*(\eta_D) = -\frac{1}{D\eta} < 0$

- Kullback-Leibler divergence (KL) = scaled Itakura-Saito divergence (IS):

$$\boxed{D_{\mathrm{KL}}[p_{D,\theta_1} : p_{D,\theta_2}] = B_F(\theta_2 : \theta_1) = \frac{1}{D}\left(\frac{\theta_2}{\theta_1} - \log\frac{\theta_2}{\theta_1} - 1\right) = \frac{1}{D} D_{\mathrm{IS}}[\theta_2 : \theta_1]}$$

$$D_{\mathrm{IS}}[p : q] := \frac{p}{q} - \log\left(\frac{p}{q}\right) - 1$$

Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities, arXiv:2107.05901

Kullback-Leibler divergence between two exponential family densities: An example

$$D_{\mathrm{KL}}[P_\lambda : Q_p] \quad = \quad F_{\mathcal{Q}}(\theta') + F^*_{\mathcal{P}}(\eta) - E_{P_\theta}[t_{\mathcal{Q}}(x)] \cdot \theta' + E_{P_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]$$

	Poisson family $\mathcal{P}$	Geometric family $\mathcal{Q}$
support	$\mathbb{N} \cup \{0\}$	$\mathbb{N} \cup \{0\}$
base measure	counting measure	counting measure
ordinary parameter	rate $\lambda > 0$	success probability $p \in (0,1)$
pmf	$\frac{\lambda^x}{x!} \exp(-\lambda)$	$(1-p)^x p$
sufficient statistic	$t_{\mathcal{P}}(x) = x$	$t_{\mathcal{Q}}(x) = x$
natural parameter	$\theta(\lambda) = \log \lambda$	$\theta(p) = \log(1-p)$
cumulant function	$F_{\mathcal{P}}(\theta) = \exp(\theta)$	$F_{\mathcal{Q}}(\theta) = -\log(1-\exp(\theta))$
	$F_{\mathcal{P}}(\lambda) = \lambda$	$F_{\mathcal{Q}}(p) = -\log(p)$
auxiliary measure term	$k_{\mathcal{P}}(x) = -\log x!$	$k_{\mathcal{Q}}(x) = 0$
moment parameter $\eta = E[t(x)]$	$\eta = \lambda$	$\eta = \frac{e^\theta}{1-e^\theta} = \frac{1}{p} - 1$
negentropy (convex conjugate) $(F^*(\eta) = \theta \cdot \eta - F(\theta))$	$F^*_{\mathcal{P}}(\theta(\lambda)) = \lambda \log \lambda - \lambda$	$F^*_{\mathcal{Q}}(\theta(p)) = (1 - \frac{1}{p}) \log(1-p) + \log p$

$$D_{\mathrm{KL}}[P_\lambda : Q_p] = -\log p + \lambda \log \lambda - \lambda - \lambda \log(1-p) - \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k \log(k!)}{k!}$$

Kullback-Leibler divergence: scale densities[1]

- KLD between scale densities $\{\frac{1}{s}p\left(\frac{x}{s}\right),\ s>0\}$ depends only on the $\boxed{\text{scale ratio}}$:

$$\mathrm{KL}(p_{\lambda s_1} : p_{\lambda s_2}) = \mathrm{KL}(p : p_{\frac{\lambda s_2}{\lambda s_1}}) = \boxed{\mathrm{KL}(p : p_{\frac{s_2}{s_1}})} = \mathrm{KL}(p_{s_1} : p_{s_2})$$

- KLD between scale densities is symmetric ($\mathrm{KL}(p_{s_1} : p_{s_2}) = \mathrm{KL}(p_{s_2} : p_{s_1})$) iff

$$\boxed{\int_{\mathcal{X}} p(x) \log \frac{p(\frac{x}{s})}{p(sx)} \mathrm{d}x = 2 \log s, \quad \forall s \in \mathbb{R}_{++}}$$

- For example, KLD between scale Cauchy densities ($p(x) = \frac{1}{\pi(1+x^2)}$):

$$\begin{aligned}
\mathrm{KL}(p_{s_1} : p_{s_2}) &= 2 \log \left(\frac{s_1 + s_2}{2\sqrt{s_1 s_2}} \right), \\
&= 2 \log \left(\frac{1 + \frac{s_2}{s_1}}{2\sqrt{\frac{s_2}{s_1}}} \right) = 2 \log \left(\frac{1 + \frac{s_1}{s_2}}{2\sqrt{\frac{s_1}{s_2}}} \right) = \mathrm{KL}(p_{s_2} : p_{s_1})
\end{aligned}$$

[1] https://arxiv.org/abs/1904.10428

Kullback-Leibler divergence between two densities of different exponential families

First exponential family: $\mathcal{P} = \{p_\theta(x) = \exp\left(\theta^\top t_{\mathcal{P}}(x) - F_{\mathcal{P}}(\theta) + k_{\mathcal{P}}(x)\right) \ : \ \theta \in \Theta\}$

Second exponential family: $\mathcal{Q} = \{q_{\theta'}(x) = \exp\left(\theta'^\top t_{\mathcal{Q}}(x) - F_{\mathcal{Q}}(\theta') + k_{\mathcal{Q}}(x)\right) \ : \ \theta' \in \Theta'\}$

Kullback-Leibler divergence (Can be used to calculate the KLD between two Weibull densities):

$$\boxed{D_{\mathrm{KL}}[p_\theta : q_{\theta'}] = F_{\mathcal{Q}}(\theta') + F^*_{\mathcal{P}}(\eta) - E_{p_\theta}[t_{\mathcal{Q}}(x)]^\top \theta' + E_{p_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]}$$

When $\mathcal{Q} = \mathcal{P} \Rightarrow k_{\mathcal{Q}}(x) = k_{\mathcal{P}}(x), F(\theta) = F_{\mathcal{Q}}(\theta) = F_{\mathcal{P}}(\theta)$ we recover the **Fenchel-Young divergence**:

$$\Rightarrow \boxed{D_{\mathrm{KL}}[p_\theta : q_{\theta'}] = F(\theta') + F^*(\eta) - \theta'^\top \eta.}$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, arXiv:2102.09728

KLD between different exponential family densities

$$\boxed{D_{\mathrm{KL}}[p_\theta : q_{\theta'}] = F_{\mathcal{Q}}(\theta') + F^*_{\mathcal{P}}(\eta) - E_{p_\theta}[t_{\mathcal{Q}}(x)]^\top \theta' + E_{p_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)].}$$

Zero-centered Laplacian family:

$$\mathcal{P} = \mathcal{L} = \left\{ p_\sigma(x) := \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) \ : \ \sigma > 0 \right\}$$

$$k_{\mathcal{L}}(x) = 0 \qquad F^*_{\mathcal{L}}(\eta(\sigma)) = -1 - \log(2\sigma)$$

$$t_{\mathcal{L}}(x) = |x|$$

Zero-mean Gaussian family:

$$\mathcal{Q} = \mathcal{N}_0 = \left\{ q_{\sigma'}(x) = \frac{1}{\sqrt{2\pi(\sigma')^2}} \exp\left(-\frac{x^2}{2(\sigma')^2}\right) \right\}$$

$$t_{\mathcal{N}_0}(x) = x^2 \qquad k_{\mathcal{N}_0}(x) = 0 \qquad \theta'(\sigma') = -\frac{1}{2(\sigma')^2}$$

$$F_{\mathcal{N}_0}(\theta'(\sigma')) = \frac{1}{2}\log(2\pi(\sigma')^2)$$

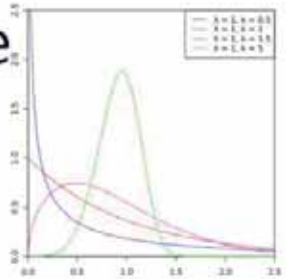
Cross-expectation of sufficient statistics : $E_{p_\sigma}[t_{\mathcal{N}_0}(x)] = E_{p_\sigma}[x^2] = 2\sigma^2$

$$\begin{aligned} D_{\mathrm{KL}}[p_\sigma : q_{\sigma'}] &= F_{\mathcal{Q}}(\theta'(\sigma')) + F^*_{\mathcal{P}}(\eta(\sigma)) - E_{p_\sigma}[t_{\mathcal{Q}}(x)]^\top \theta'(\sigma') + E_{p_\sigma}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)] \\ &= \frac{1}{2}\log(2\pi(\sigma')^2) - 1 - \log(2\sigma) - 2\sigma^2\left(-\frac{1}{2(\sigma')^2}\right), \\ &= \log\left(\frac{\sigma'}{\sigma}\right) + \left(\frac{\sigma}{\sigma'}\right)^2 + \frac{1}{2}\log\left(\frac{\pi}{2}\right) - 1. \end{aligned}$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, arXiv:2102.09728

Kullback-Leibler divergence between two Weibull densities

- **Weibull distribution** with **shape** κ>0 and **scale** σ>0 has density on the positive reals defined by:



κ=1→exponential distributions

κ=2→Rayleigh distributions

$$p^{\mathrm{Wei}}_{\kappa,\sigma}(x) := \frac{\kappa}{\sigma}\left(\frac{x}{\sigma}\right)^{\kappa-1}\exp\left(-\left(\frac{x}{\sigma}\right)^{\kappa}\right)$$

- For a fixed shape κ, the Weibull distributions form an exponential family:

Natural parameter $\theta = -\frac{1}{\sigma^{\kappa}}$ Sufficient statistics $t_{\kappa}(x) = x^{\kappa}$ $F_{\kappa}(\theta) = -\log(-\theta)$

Auxiliary carrier measure term $k_{\kappa}(x) = (\kappa-1)\log x + \log\kappa$ Cumulant function: $F_{\kappa}(\theta(\sigma)) = F_{\kappa}(\sigma) = \kappa\log\sigma$

- Kullback-Leibler divergence between two Weibull densities calculated from the KLD formula between two exponential family densities:

$$D_{\mathrm{KL}}\left[p^{\mathrm{Wei}}_{\kappa_1,\sigma_1} : p^{\mathrm{Wei}}_{\kappa_2,\sigma_2}\right] = F_{\kappa_2}(\theta') + F^*_{\kappa_1}(\eta) - E_{p_{\kappa_1,\sigma_1}}[x^{\kappa_2}]^{\top}\theta' + E_{p_{\kappa_1,\sigma_1}}[k_{\kappa_1}(x) - k_{\kappa_2}(x)]$$

$$\boxed{D_{\mathrm{KL}}\left[p^{\mathrm{Wei}}_{\kappa_1,\sigma_1} : p^{\mathrm{Wei}}_{\kappa_2,\sigma_2}\right] = \log\frac{\kappa_1}{\sigma_1^{\kappa_1}} - \log\frac{\kappa_2}{\sigma_2^{\kappa_2}} + (\kappa_1-\kappa_2)\left[\log\sigma_1 - \frac{\gamma}{\kappa_1}\right] + \left(\frac{\sigma_1}{\sigma_2}\right)^{\kappa_2}\Gamma\left(\frac{\kappa_2}{\kappa_1}+1\right) - 1}$$

$\gamma \approx 0.5772156649$ *Euler–Mascheroni constant*

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, arXiv:2102.09728

Kullback-Leibler divergence & exponential families

Theorem *The Kullback-Leibler divergence between two densities p_{λ_1} and p_{λ_2} belonging to a full regular exponential family $\mathcal{E}$ of order D can be expressed as the averaged sum of logarithms of density ratios:*

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s} \sum_{i=1}^{s} \log \left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)} \right),$$

where $\omega_1, \ldots, \omega_s$ are $s \leq D+1$ distinct samples of $\mathcal{X}$ chosen such that $\frac{1}{s}\sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]$.

The bound $s \leq D+1$ follows from Carathéodory's theorem

Example:

$$D_{\mathrm{KL}}[p_{\mu_1,\sigma_1} : p_{\mu_2,\sigma_2}] = \frac{1}{2} \left(\log \left(\frac{p_{\mu_1,\sigma_1}(\mu_1 - \sigma_1)}{p_{\mu_2,\sigma_2}(\mu_1 - \sigma_1)} \right) + \log \left(\frac{p_{\mu_1,\sigma_1}(\mu_1 + \sigma_1)}{p_{\mu_2,\sigma_2}(\mu_1 + \sigma_1)} \right) \right)$$

Example:

$$D_{\mathrm{KL}}[p_{\Sigma_1} : p_{\Sigma_2}] = \frac{1}{d} \sum_{i=1}^{d} \log \left(\frac{p_{\Sigma_1}(\sqrt{d\lambda_i} e_i)}{p_{\Sigma_2}(\sqrt{d\lambda_i} e_i)} \right),$$

where the λ_i's are the eigenvalues of Σ_1 with the e_i's the corresponding eigenvectors.

Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family
https://arxiv.org/abs/2003.02469

Kullback-Leibler divergence & exponential families

Theorem *The Kullback-Leibler divergence between two densities p_{λ_1} and p_{λ_2} belonging to a full regular exponential family $\mathcal{E}$ of order D can be expressed as the averaged sum of logarithms of density ratios:*

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s}\sum_{i=1}^{s} \log\left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)}\right),$$

where $\omega_1, \ldots, \omega_s$ are $s \leq D+1$ distinct samples of $\mathcal{X}$ chosen such that $\frac{1}{s}\sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]$.

The bound $s \leq D+1$ follows from Carathéodory's theorem

$$\begin{aligned} D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] &= \frac{1}{2d}\sum_{i=1}^{d}\left(\log\left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1 - [\sqrt{d\Sigma_1}]_{\cdot,i}\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1 - [\sqrt{d\Sigma_1}]_{\cdot,i}\right)}\right) + \log\left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1 + [\sqrt{d\Sigma_1}]_{\cdot,i}\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1 + [\sqrt{d\Sigma_1}]_{\cdot,i}\right)}\right)\right) \\ &= = \frac{1}{2}\left((\mu_2-\mu_1)^\top \Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}\left(\Sigma_2\Sigma_1^{-1}\right) + \log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - d\right). \end{aligned}$$

where $[\sqrt{d\Sigma_1}]_{\cdot,i} = \sqrt{\lambda_i}e_i$ denotes the vector extracted from the i-th column of the square root matrix of $d\Sigma_1$, and e_i the i-th eigenvector.

Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family
https://arxiv.org/abs/2003.02469

Summary and conclusion

- Since Bregman divergences are defined for generators up to an affine term, considered $\boxed{\mathrm{F}(\theta) = -\log \mathrm{p}_\theta(\omega)}$ for **any** ω in the convex hull of the support $\mathcal{X}$ of the exponential family
- Can express the KLD (as a reverse BD) between two densities of an exponential family as an **average sum of log density-ratio**:

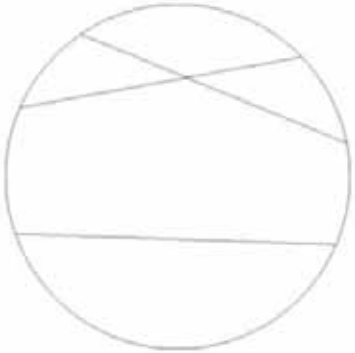
$$\boxed{D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{s}} \log \left(\frac{\mathrm{p}_{\lambda_1}(\omega_\mathrm{i})}{\mathrm{p}_{\lambda_2}(\omega_\mathrm{i})} \right) \text{ s.t. } \frac{1}{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{s}} \mathrm{t}(\omega_\mathrm{i}) = \mathrm{E}_{\mathrm{p}_{\lambda_1}}[\mathrm{t}(\mathrm{x})] = \eta_1}$$

  where $\omega_1, \ldots, \omega_s$ are $s \leqslant D+1$ distinct points of $\mathcal{X}$ chosen such that $\frac{1}{s}\sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)] = \eta_1$.
- **assess the error in the Monte Carlo estimators of EF KLD**
- Approach works whenever a statistical divergence between two parametric densities amount to a Bregman divergence:
  exponential families $\rightarrow$ q-exponential families $\rightarrow$ χ-exponential families)
- Other statistical divergences like **Bhattacharyya**, **Hellinger**, **α-divergences** between EF densities and report formula without the explicit use of the log-normalizer $F(\theta)$, arXiv:2003.02469 [13]

Closed-form hyperbolic geodesics in the Klein model

- Distance in the d-dimensional **Klein unit disk model** of hyperbolic geometry:

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



- **Geodesics** in Klein geodesic metric space : Straight line segments

$$\gamma_K(p,q;\alpha) = (1-u(\alpha))p + u(\alpha)q$$

$$u(\alpha) = \frac{ad(\alpha)\sqrt{(ac+b^2)(d(\alpha)^2-1)} + ab(1-d(\alpha)^2)}{acd(\alpha)^2 + b^2}$$

where

$$\begin{aligned} a &= 1 - p^\top p, \\ b &= p^\top (q-p), \\ c &= (q-p)^\top (q-p), \\ d(\alpha) &= \cosh(\alpha d_K(p,q)) \end{aligned}$$

satisfies $d_K(\gamma(p,q;s), \gamma(p,q;t)) = |s-t|\, d_K(p,q)$

A closed-form expression of geodesics in the Klein model of hyperbolic geometry

Hyperbolic Voronoi diagram is orthogonal to dual hyperbolic Delaunay complex

Distance in Klein Riemannian unit ball $\mathbb{B}$ model of hyperbolic geometry (non-conformal except at the origin) :

$$D^H_{\text{Klein}}(p,q) = \operatorname{arccosh}\left(\frac{1-\langle p,q\rangle}{\sqrt{(1-\langle p,p\rangle)(1-\langle q,q\rangle)}}\right)$$

with $\operatorname{arccosh}(x) := \log(x+\sqrt{x^2-1})$

Klein pregeodesic (line segment):

$$\gamma_{\text{Klein}}(p,q) = \{(pq)\cap\mathbb{B}\}$$

Klein Voronoi bisector:

$$\text{Bi}_{\text{Klein}}(p,q) := \{x\in\mathbb{B} \;:\; D^H_{\text{Klein}}(p,x) = D^H_{\text{Klein}}(q,x)\}$$

Orthogonality property:

$$\boxed{\gamma_{\text{Klein}}(p,q) \perp \text{Bi}_{\text{Klein}}(p,q)}$$

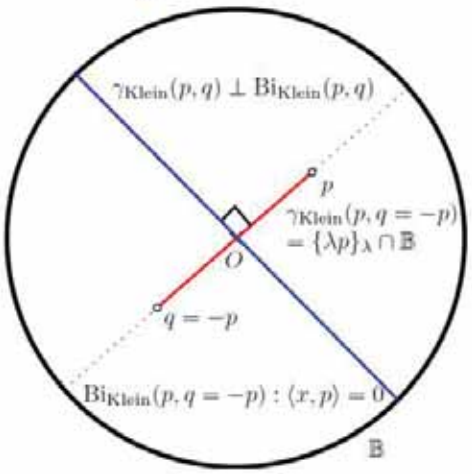
Proof: Wlog, assume q=-p

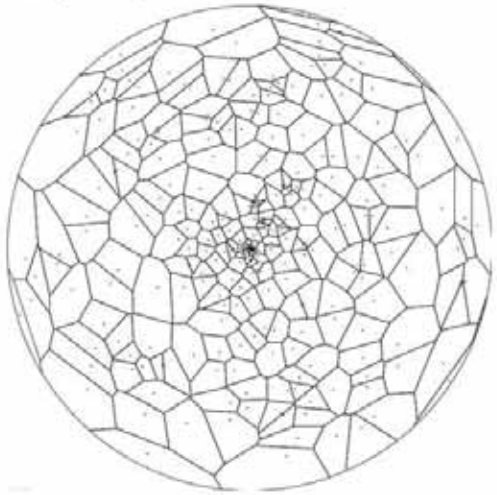
(hyperbolic motion via Moebius transformation)

Bisector: $\langle x,p\rangle = 0$

Pregeodesic: $\gamma_{pq} = \{\lambda p\}_\lambda \cap \mathbb{B}$

$\gamma_{\text{Klein}}(p,q) \perp \text{Bi}_{\text{Klein}}(p,q)$





Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010

Kullback-Leibler divergence and Fisher-Rao distance

- Kullback-Leibler (KL) oriented divergence (non-metric relative entropy):

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

- Fisher-Rao (FR) distance (metric):

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

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

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

- For densities of a statistical model:

$$\boxed{D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = \frac{1}{2} \rho^2_{\mathrm{FR}}(\theta_1, \theta_2) + o(\rho^2_{\mathrm{FR}}(\theta_1, \theta_2))}$$

An elementary introduction to information geometry, 2018

https://arxiv.org/abs/1808.08271

Kullback-Leibler divergence between Gaussians[1]

- MultiVariate Gaussian/Normal density with $\lambda{:=}(\lambda_v, \lambda_M) = (\mu, \Sigma)$:

$$\begin{aligned} p_\lambda(x;\lambda) &:= \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2}(x-\lambda_v)^\top \lambda_M^{-1}(x-\lambda_v)\right) \\ p_\theta(x;\theta) &:= \exp\left(\langle t(x), \theta\rangle - F_\theta(\theta)\right) = p_\lambda(x;\lambda(\theta)) \end{aligned}$$

  with $\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right) = \theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, -\frac{1}{2}\lambda_M^{-1}\right)$, $t(x) = (x, -xx^\top)$, and $\langle\theta,\theta'\rangle{:=}\theta_v^\top\theta'_v + \mathrm{tr}\left(\theta_M'^{\top}\theta_M\right)$

- Cumulant $F_\theta(\theta) = \frac{1}{2}\left(d\log\pi - \log|\theta_M| + \frac{1}{2}\theta_v^\top\theta_M^{-1}\theta_v\right)$
- Moment parameters $\eta = (\eta_v, \eta_M) = E[t(x)] = \nabla F(\theta)$:

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

- Legendre convex conjugate $F_\eta^*(\eta) = -\frac{1}{2}\left(\log(1+\eta_v^\top\eta_M^{-1}\eta_v) + \log|-\eta_M| + d(1+\log 2\pi)\right)$
- The Kullback-Leibler between $p_{(\mu_1,\Sigma_1)}$ and $p_{(\mu_2,\Sigma_2)}$ (with $\Delta_\mu = \mu_2 - \mu_1$) is

$$\begin{aligned} \mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= \boxed{\frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\right\}} = \mathrm{KL}(p_{\lambda_1} : p_{\lambda_2}), \\ &= B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2) = A_F(\theta_2 : \eta_1) = A_{F^*}(\eta_1 : \theta_2) \end{aligned}$$

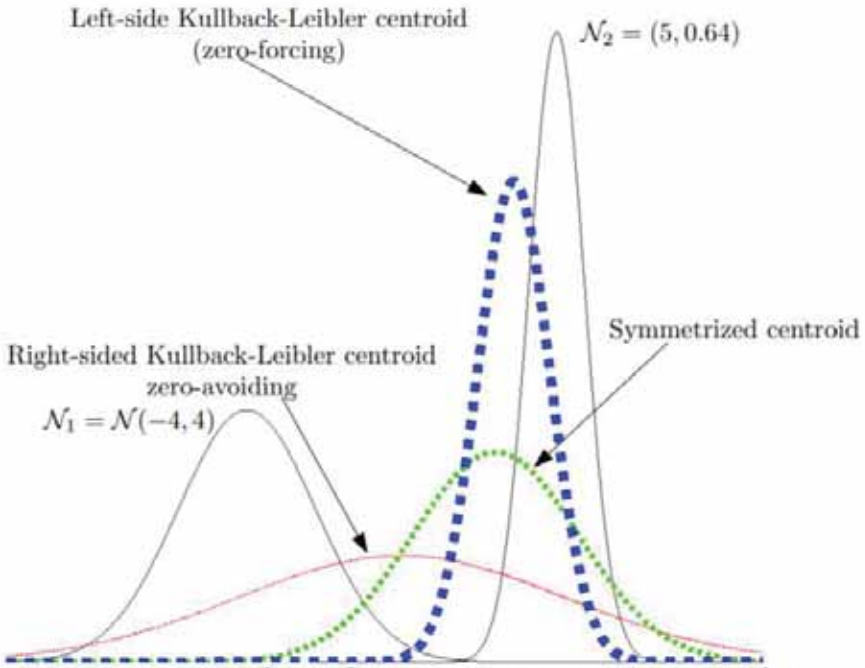
- Bregman divergence B_F and canonical divergence A_F:

$$\begin{aligned} B_F(\theta : \theta') &:= F(\theta) - F(\theta') - \langle\theta - \theta', \nabla F(\theta')\rangle \\ A_F(\theta_1 : \eta_2) &:= F(\theta_1) + F^*(\eta_2) - \langle\theta_1, \eta_2\rangle = A_{F^*}(\eta_2 : \theta_1) \end{aligned}$$

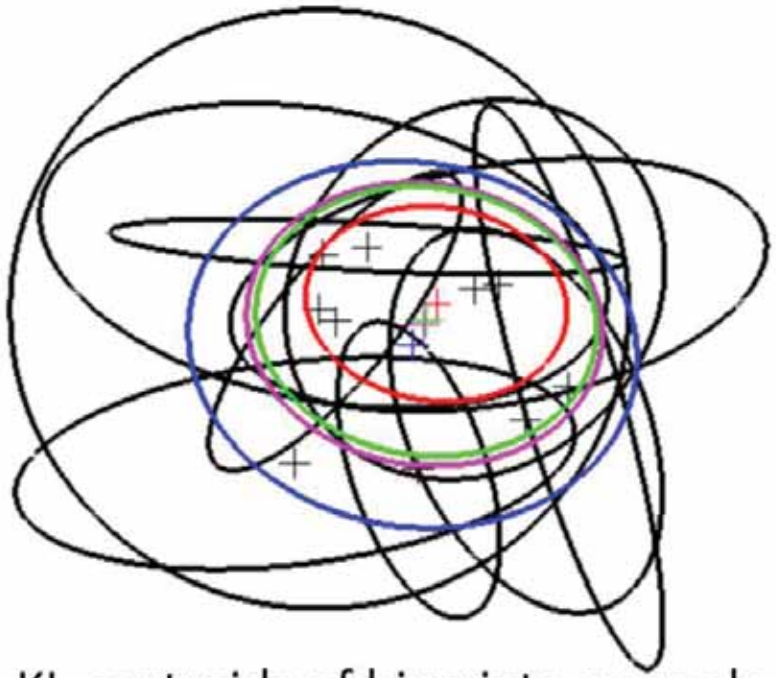
[1] https://arxiv.org/abs/1904.10428

Minimizing Kullback-Leibler divergence: Which side?

- Kullback-Leibler divergence (KL, relative entropy): $\mathrm{KL}(p(x)||q(x)) = \int_x p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x.$
- KL right-sided centroid is **zero-avoiding** or **mass covering** $c_R = \arg_c \min \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(p_i||c).$
- KL left-sided centroid is **zero-forcing** or **mode attracting** $c_L = \arg_c \min \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(c||p_i)$



KL centroids of univariate normals



KL centroids of bivariate normals

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory (2009)

Kullback-Leibler divergence: Cauchy densities

- Location-scale Cauchy densities:

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

- Kullback-Leibler divergence

$$\begin{aligned} \mathrm{KL}(p_{l_1,s_1} : p_{l_2,s_2}) &= \int p_{l_1,s_1}(x) \log \frac{p_{l_1,s_1}}{p_{l_2,s_2}} \mathrm{d}x, \\ &= \log \frac{(s_1+s_2)^2 + (l_1-l_2)^2}{4 s_1 s_2}. \end{aligned}$$

- Finite and symmetric (not true in general for location-scale families)
- Relies on calculating this novel definite integral $4df - e^2, 4ac - b^2 \geq 0$

$$A(a,b,c;d,e,f) = \int_{-\infty}^{\infty} \frac{\log(dx^2+ex+f)}{ax^2+bx+c} \mathrm{d}x$$

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

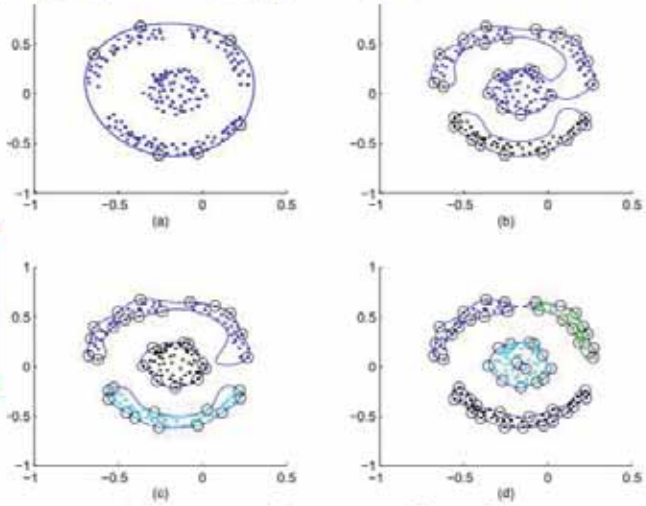
Approximating the kernelized minimum enclosing ball

Kernel $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$ Feature map $\Phi(x) : \mathbb{R}^d \to \mathbb{R}^D$

(D may be infinite)

Trick: **Encode implicitly** the circumcenter φ of the enclosing ball as a convex combination of the data points: $\varphi = \sum_i \alpha_i \phi(p_i) = \sum_i \alpha_i \phi_i$

$$\begin{aligned}\|\varphi - \phi(p)\|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \langle \phi_i, \phi_j \rangle - 2\sum_{i=1}^{n} \alpha_i \langle \phi_i, \phi(p) \rangle + \langle \phi(p), \phi(p) \rangle, \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(p_i, p_j) - 2\sum_{i} \alpha_i k(p_i, p) + k(p,p).\end{aligned}$$



Update weights iteratively: $\boxed{\alpha^{(i+1)} = \frac{i}{i+1}\alpha^{(i)} + \frac{1}{i+1} e_{f_i}}$

Index of the current farthest point e_{f_i}

Applications: Support Vector Data Description, Support Vector Data Description

A note on kernelizing the smallest enclosing ball for machine learning, 2017

Fitting the smallest enclosing Bregman ball, ECML 2005

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

arxiv:1203.5181

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

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

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

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

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

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

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

Online k-MLE for Mixture Modeling with Exponential Families, GSI 2015
Hartigan's Method for k-MLE: Mixture Modeling with Wishart Distributions and Its Application to Motion Retrieval, GTI 2014
A New Implementation of k-MLE for Mixture Modeling of Wishart Distributions, GSI 2013
Fast Learning of Gamma Mixture Models with k-MLE, SIMBAD 2013
k-MLE: A fast algorithm for learning statistical mixture models, ICASSP 2012
k-MLE for mixtures of generalized Gaussians, ICPR 2012

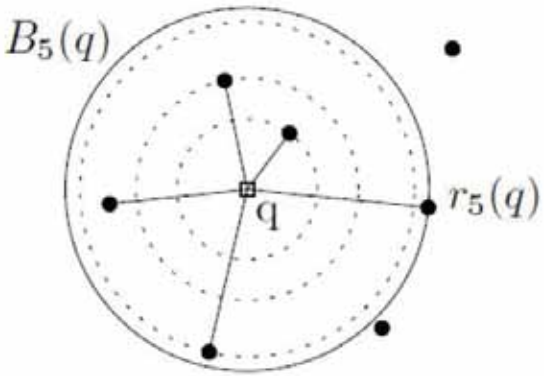
@FrnkNlsn

k-NN: Balloon estimator, Bayes' error and HPC

K-NN rule: Classify x by taking the majority label of the k-nearest neighbors of x

$$p(x) \approx \frac{k}{nV(B_k(x))} \propto \frac{k}{nr_k(x)^d}$$

Balloon estimator



Theorem 10

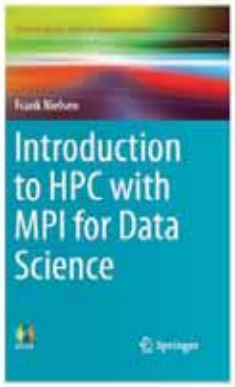
The balloon estimator allows to approximate a smooth density $p(x)$ with support in $\mathbb{R}^d$ by $p(x) \approx \frac{k}{nV(B)}$, where k is the number of samples of X that is contained in the ball B, and $V(B)$ its volume.

For the the multi-class case ($m \geq 2$ classes) and the NN-rule, one can further prove that we have the following guaranteed upper-bound:

$$P_e \leq \tau_{\text{error}}(\text{NN}) \leq P_e \left(2 - \frac{m}{m-1} P_e\right).$$

Theorem 11

The optimal Bayesian MAP rule can be approximated by the k-NN voting rule within a multiplicative error factor of 2 when we estimate non-parametrically the class probabilities using the k-NN balloon estimator.



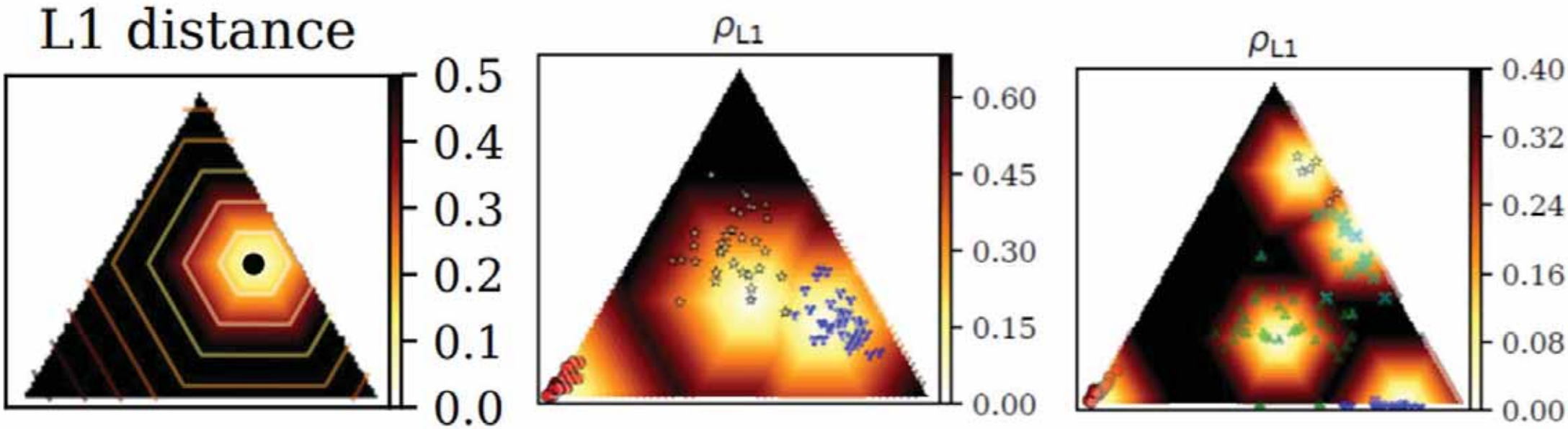
Implementation on a HPC cluster with decomposable k-NN queries:

$$\text{NN}_k(x, X) = \text{NN}_k(x, \cup_{l=1}^{p} \text{NN}_k(x, X_l)).$$
$$X = \uplus_{l=1}^{p} X_l$$

Introduction to HPC with MPI for Data Science, 2016 https://franknielsen.github.io/HPC4DS/index.html

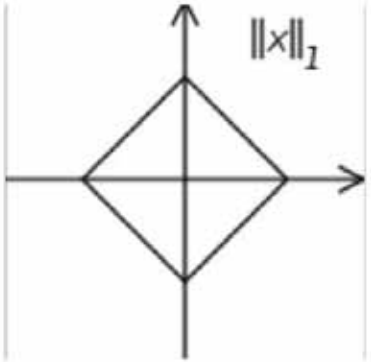
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means, PRL 2014

L1 norm restricted to the standard simplex: Hexagonal norm



Clustering in Hilbert simplex geometry, arxiv 1704.00454

L1 norm restricted to the standard simplex: Hexagonal norm



L1 norm

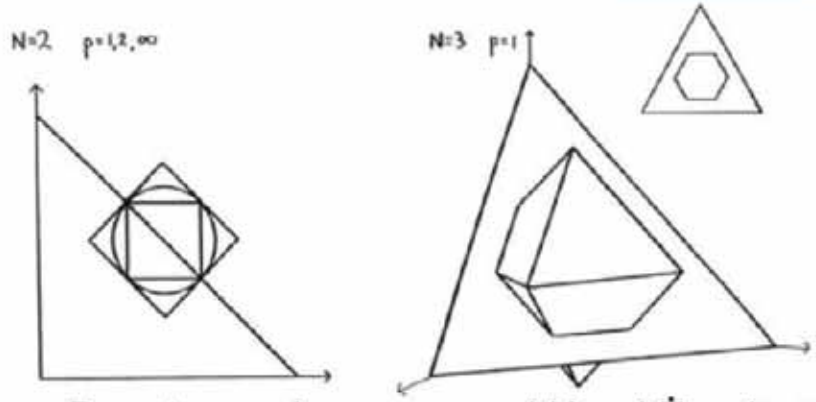
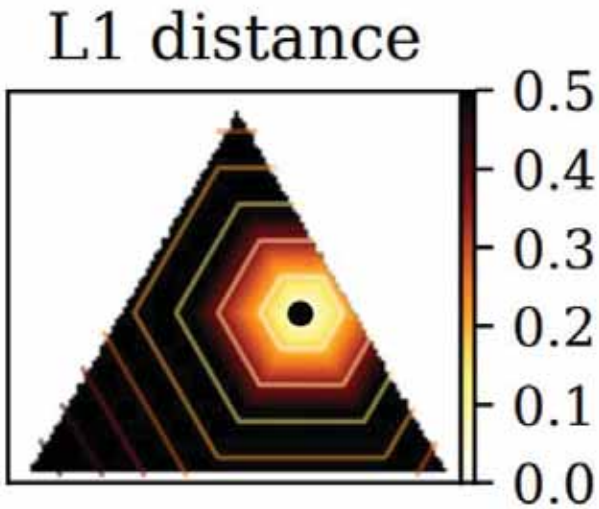
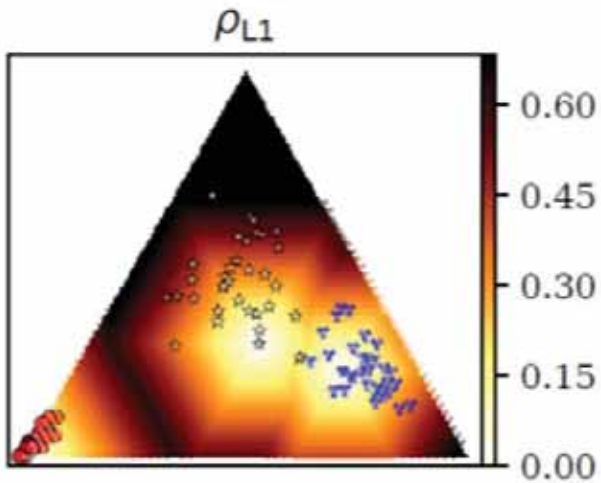
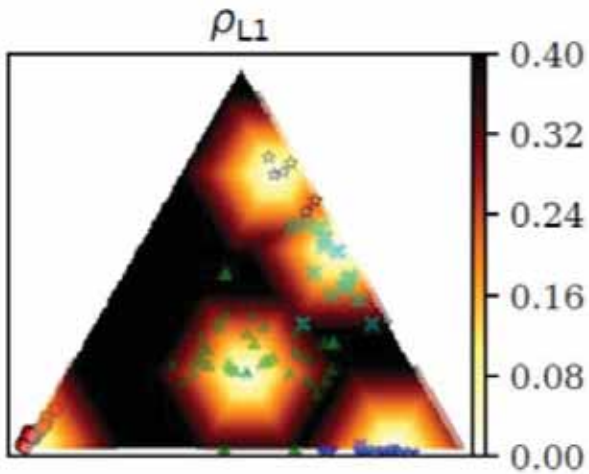


Figure from Bengtsson, Ingemar, and Karol Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017.

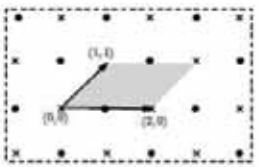


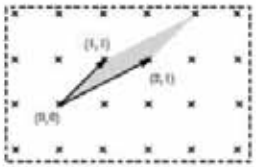




Clustering in Hilbert simplex geometry, arxiv 1704.00454

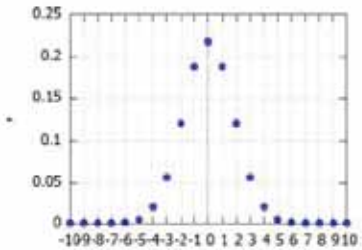
Lattice Gaussian distributions

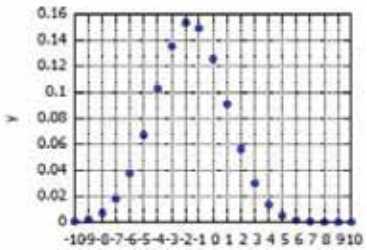




$$p_\xi(l) = \frac{1}{\theta_\Lambda(\xi)} \exp\left(2\pi\left(-\frac{1}{2} l^\top \xi_2 l + l^\top \xi_1\right)\right), \quad l \in \Lambda,\ \Lambda = \Lambda(L) = L\mathbb{Z}^d = \{Lz \ : \ z \in \mathbb{Z}^d\}$$

Partition function $\theta_\Lambda(\xi) = \theta(iL^\top \xi_1, -iL^\top \xi_2 L)$, expressed with the **Riemann theta function**:





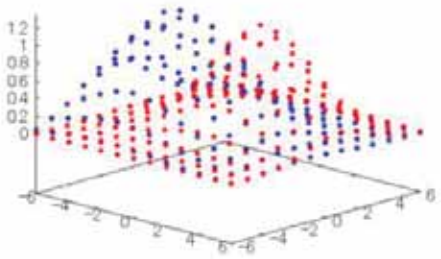
$$\theta : \mathbb{C}^d \times \mathcal{H}_d \to \mathbb{C}$$

$$\theta(\omega, \Omega) := \sum_{z \in \mathbb{Z}^d} \exp\left(2\pi i \left(\frac{1}{2} z^\top \Omega z + z^\top \omega\right)\right).$$

Kullback-Leibler divergence between two lattice Gaussian distributions:

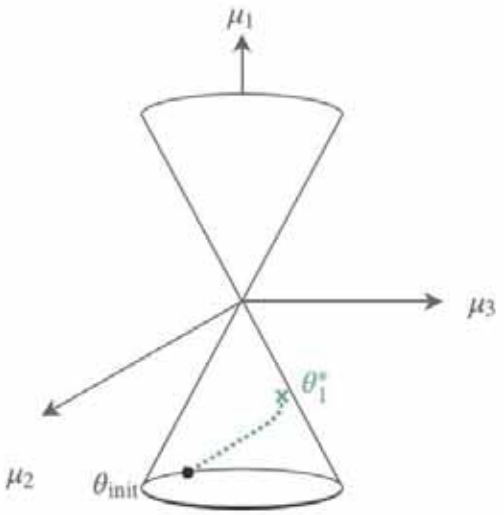
Proposition *The Kullback-Leibler divergence between two lattice Gaussian distributions $p_\xi \sim N_\Lambda(\mu, \Sigma)$ and $p'_\xi \sim N_\Lambda(\mu', \Sigma')$ is:*

$$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right) - 2\pi\mu^\top(\xi'_1 - \xi_1) + \pi\,\mathrm{tr}\left((\xi'_2 - \xi_2)(\Sigma + \mu\mu^\top)\right).$$



On the Kullback-Leibler divergence between discrete normal distributions, arxiv:2109.14920

Dynamics of learning on singular models & stratifolds



Log-likelihood:

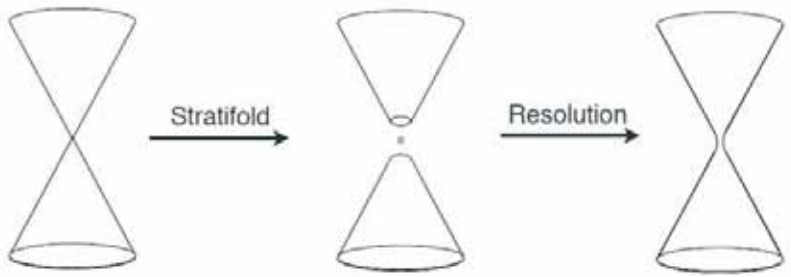
$$\mathcal{L}(x;\xi,\theta) = \frac{1}{2}\left((x_1-\xi)^2 + (x_2-\xi\cos\theta)^2 + (x_3-\xi\sin\theta)^2\right)$$

Stochastic Gradient Descent (SGD)

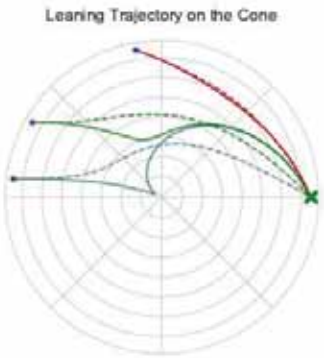
$$\begin{bmatrix}\xi^{(t+1)}\\ \theta^{(t+1)}\end{bmatrix} = -\vartheta\begin{bmatrix}\bar{x}_1 + \bar{x}_2\cos\theta^{(t)} + \bar{x}_3\sin\theta^{(t)} - 2\xi^{(t)}\\ -\xi^{(t)}(\bar{x}_2\sin\theta^{(t)} + \bar{x}_3\cos\theta^{(t)})\end{bmatrix}$$

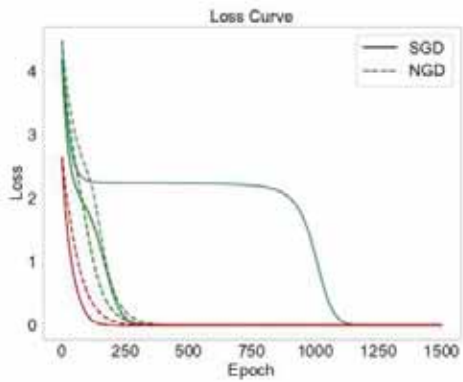
Natural Gradient Descent (NGD)

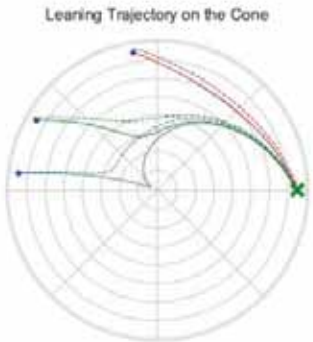
$$\begin{bmatrix}\xi^{(t+1)}\\ \theta^{(t+1)}\end{bmatrix} = -\vartheta\begin{bmatrix}2 & 0\\ 0 & (\xi^{(t)})^2\end{bmatrix}^{-1}\nabla\mathcal{L}(x;\xi,\theta)$$

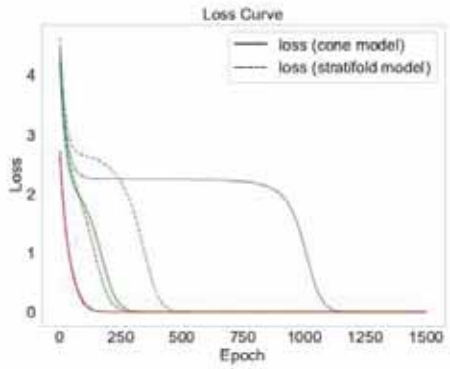


$$\mu_1 = \xi$$
$$\mu_2 = \xi\cos\theta$$
$$\mu_3 = \xi\sin\theta$$





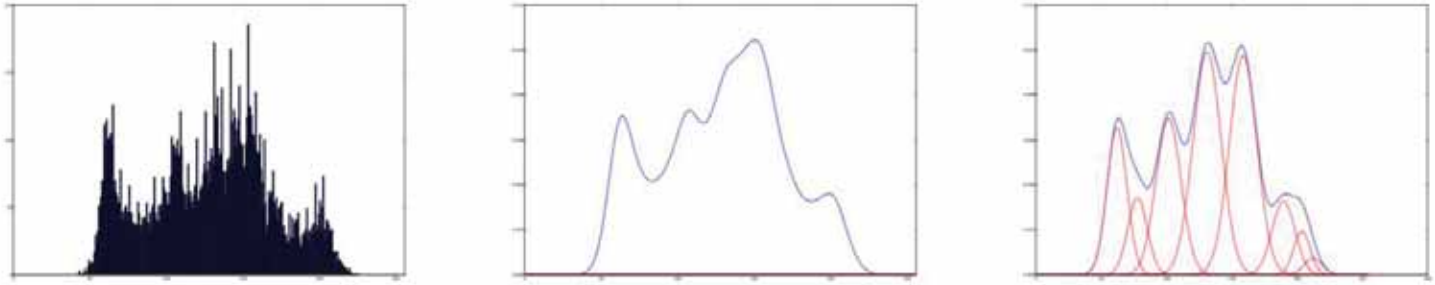




$$\Theta := \{\xi,\theta\} \in \mathbb{R}^2$$

https://www.youtube.com/watch?v=W3HhjQoOxk8

Learning a mixture by simplifying a kernel density estimator



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

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

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

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

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

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

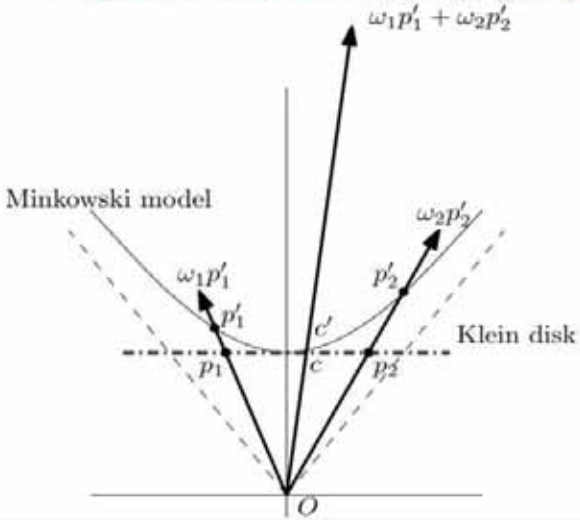
$$\mathrm{KLD}(f_p, f_q) = \frac{1}{2}\log\left(\frac{\det \Sigma_p}{\det \Sigma_q}\right) + \frac{1}{2}\mathrm{tr}(\Sigma_q^{-1}\Sigma_p) + \frac{1}{2}(\mu_q - \mu_p)^T \Sigma_q^{-1}(\mu_q - \mu_p) - \frac{d}{2}$$

or Fisher-Rao distance (hyperbolic distance)

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

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

Galperin's model centroid (HG)



Simple model centroid algorithm:

Embed Klein points to points of the Minkowski hyperboloid
Centroid = center of mass c, scaled back to c' of the hyperboloid
Map back c' to Klein disk

@FrnkNlsn

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

LeCam divergence between Cauchy densities

- **LeCam divergence** is akind of symmetrization of the **chi-squared divergence**:

$$D_{\mathrm{LC}}[p,q] := \int \frac{(p(x)-q(x))^2}{p(x)+q(x)}\mathrm{d}x \qquad \chi[p:q] := \int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}x$$

- The square root of LeCam divergence yields a **metric distance:** $\sqrt{D_{\mathrm{LC}}[p,q]}$
- LeCam divergence is a **f-divergence** :

$$I_f(p:q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \qquad f_{\mathrm{LC}}(u) = \frac{(u-1)^2}{1+u}$$

- Closed-form formula for the **LeCam divergence between Cauchy densities**:

$$p_{l,s}(x) = \frac{s}{\pi(s^2+(x-l)^2)}$$

$$\boxed{\begin{aligned} D_{\mathrm{LC}}[p_{l_1,s_1}, p_{l_2,s_2}] &= 2 - 4\sqrt{\frac{1}{2(\chi[p_{l_1,s_1}:p_{l_2,s_2}]+2)}} \\ \chi[p_{l_1,s_1}, p_{l_2,s_2}] &= \frac{(l_1-l_2)^2+(s_1-s_2)^2}{2s_1s_2} \end{aligned}}$$

On f-divergences between Cauchy distributions, arXiv 2101.12459 2021

Legendre transform, convexity and parameterization

- F(θ) a proper lower semi-continuous (lsc) convex function
- Legendre transform:

$$\begin{aligned} F^*(\eta) &:= \sup_{\theta\in\Theta}\langle\theta,\eta\rangle - F(\theta) \\ &= \langle(\nabla F)^{-1}(\eta),\eta\rangle - F((\nabla F)^{-1}(\eta)) \end{aligned}$$

- Parameter transformation: $\eta = \nabla F(\theta) = (\nabla F^*)^{-1}(\theta) \Longleftrightarrow \theta = \nabla F^*(\eta) = (\nabla F)^{-1}(\eta)$

- **Moreau biconjugation theorem**: $(F^*)^* = F$
- Convexity of functions is preserved by affine transformations: $F(A\theta + b)$
- F*(η) is proper lsc convex function, but $F^*(\nabla F(\theta))$ may not be convex!

<u>Example:</u>

$$\begin{aligned} F(\theta) &= -\log\theta \\ F'(\theta) &= -\frac{1}{\theta} =: \eta \\ \theta &= -\frac{1}{\eta} \\ F^*(\eta) &= \theta\eta - F(\theta) = -1 + \log\theta = -1 - \log(-\eta) \end{aligned}$$

$\Longrightarrow$ **F*(η) convex but F*(η(θ)) not convex!**

Legendre transformation and information geometry (2010)

Legendre type functions and Legendre transform

- A set S is convex iff the segments linking any two points of S fall inside S
- F is convex if its **epigraph** epi(F) is convex: $\mathrm{epi}(F) := \{(\theta, y) \quad : \quad \theta \in \Theta, y \geq F(\theta)\}$
- **Gradient mapping** of a convex function : $\nabla F : \theta \to \eta = \nabla F(\theta)$
- If ∇F is one-to-one, the **Legendre transform** $(\Theta, F)^* = (H, F^*)$:

$$F^*(\eta) := \langle \eta, (\nabla F)^{-1}(\eta) \rangle - F((\nabla F)^{-1}(\eta)) \qquad H = \{\nabla F(\theta) \ : \ \theta \in \Theta\}$$

- (Θ, F) **Legendre type** if is a non-empty open convex set satisfying:

$$\lim_{\alpha \to 0^+} \frac{d}{\mathrm{d}\alpha} F(\alpha\theta + (1-\alpha)\bar{\theta}) = -\infty, \forall \theta \in \Theta, \forall \bar{\theta} \in \partial\bar{\Theta}$$

= function become infinitely steep near boundary points of its effective domain

<u>Theorem</u>: If (Θ, F) Legendre type then $(\Theta, F)^* = (H, F^*)$ is Legendre type and $\nabla F^*(\eta) = \nabla F^{-1}(\eta), \forall \eta \in H$. Moreover, $(\Theta, F)^{*\uparrow} = (\Theta, F)$ (involution)

Rockafellar, "Conjugates and Legendre transforms of convex functions" *Canadian Journal of Math.* (1967)

The seven Liberal arts: Trivium + Quadrivium

- Trivium (lower part of curriculum): grammar, rhetoric, logic
- Quadrivium (upper part of curriculum): geometry, arithmetic, music, astronomy

Quadrivium

Trivium

Seven Liberal Arts (circa 1450, painting by Domenico di Michelino)

Fast approximation of the Löwner extremal matrix

Problem 1 (Löwner maximal matrices)

$$\bar{S} = \inf\{X \in \mathrm{Sym}(\mathbb{R}) \ : \ \forall i \in [n], X \succeq S_i\}$$

Finding the extremal matrix of positive-definite matrices amount to compute the smallest enclosing ball of cone basis balls

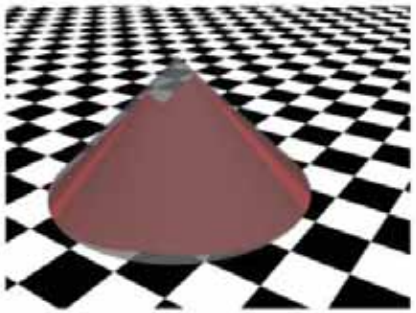
<u>Visualizations of a positive-definite matrix:</u>

a/Covariance ellipsoids

b/Translated positive-define cone

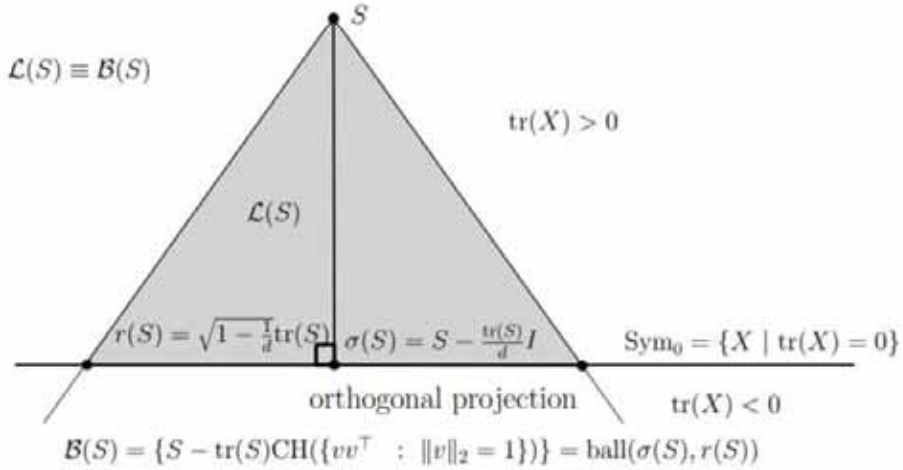
c/Basis balls of (b)

(a)



(b)

(c)



https://arxiv.org/abs/1604.01592

Log-likelihood trick in machine learning

- Since $\nabla_\theta \log p_\theta(x) = \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}$, we have

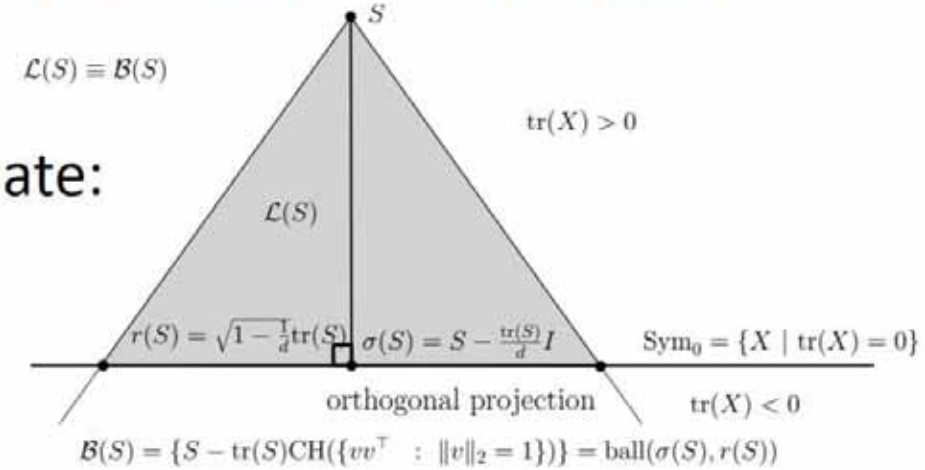
$$\boxed{\nabla_\theta p_\theta(x) = p_\theta(x) \nabla_\theta \log p_\theta(x)}$$

$$\boxed{\nabla_\theta E_{p_\theta}[f(X)] = E_{p_\theta}\left[f(X) \nabla_\theta \log p_\theta(X)\right]}$$

- How to use this trick?
  - when $\nabla_\theta \log p_\theta(x)$ is easier to calculate than $\nabla_\theta p_\theta(x)$: For example, when p is a product: $p = \prod_j p_j$
  - to *estimate with Monte Carlo* the gradient of $E_{p_\theta}[f(X)]$ using s independent samples $x_1, \ldots, x_s$ from propositional distribution $p_\theta(x)$:

$$\nabla_\theta E_{p_\theta}[f(X)] = E_{p_\theta}\left[f(X) \nabla_\theta \log p_\theta(X)\right] \simeq \frac{1}{s} \sum_{i=1}^{s} f(x_i) \nabla_\theta \log p_\theta(x_i)$$

The cone of symmetric positive-definite matrices: The SPD cone



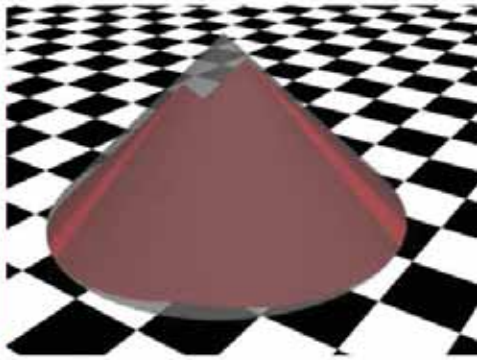
To a SPD matrix, $S \succ 0 \Leftrightarrow \forall x \neq 0, x^\top S x > 0$ associate:

- Dominance matrix cone L(S)
- Matrix ball ball(S)
  (intersection of L(S) with zero-trace hyperplane)

Löwner partial ordering: $S_1 \succ S_2 \Leftrightarrow S_1 - S_2 \succ 0 \Leftrightarrow \forall x \neq 0, x^\top (S_1 - S_2) x > 0$
of SPD matrices

Dominance of SPD matrices
as geometric containments
- Enclosing dominance cones
- Enclosing balls

$$\boxed{\begin{aligned} S_1 \succ S_2 \quad &\Leftrightarrow \quad \mathcal{L}(S_1) \supset \mathcal{L}(S_2) \\ &\Leftrightarrow \quad \mathrm{ball}(S_1) \supset \mathrm{ball}(S_2) \end{aligned}}$$



2x2 SPD matrices in R3

Fast (1+ε)-Approximation of the Löwner Extremal Matrices of High-Dimensional Symmetric Matrices, Computational Information Geometry: For Image and Signal Processing, 2017. arxiv 1604.01592

Mahalanobis distance and Cholesky decomposition

Mahalanobis metric distance for a *symmetric positive-definite* matrix Q:

$$\boxed{M_Q(\theta,\theta') = \sqrt{(\theta'-\theta)^\top Q(\theta'-\theta)} = \sqrt{\Delta\theta^\top Q \Delta\theta}.} \qquad \Delta\theta = \theta' - \theta.$$

The Mahalonobis distance is induced by a norm and amount to the Euclidean distance (i.e., Q=I) on affinely transformed parameters:

$$\begin{aligned} M_Q^2(\theta,\theta') &= \Delta\theta^\top Q \Delta\theta, \\ &= \Delta\theta^\top L L^\top \Delta\theta, \\ &= M_I^2(L^\top\theta, L^\top\theta') = \|L^\top\theta - L^\top\theta'\|^2 \end{aligned}$$

We can calculate a Mahalanobis distance wrt Q1 as another Mahalanobis distance wrt Q2:

$$M_Q(\theta_1,\theta_2) = M_I(L^\top\theta_1, L^\top\theta_2) \Leftrightarrow M_I(\theta_1,\theta_2) = M_Q((L^\top)^{-1}\theta_1, ((L^\top)^{-1}\theta_2).$$

$$M_{Q_1}(\theta_1,\theta_2) = M_I(L_1^\top\theta_1, L_1^\top\theta_2) = M_{Q_2}((L_2^\top)^{-1}L_1^\top\theta_1, (L_2^\top)^{-1}L_1^\top\theta_2).$$

$$\boxed{M_{Q_1}(\theta_1,\theta_2) = M_{Q_2}((L_2^\top)^{-1}L_1^\top\theta_1, (L_2^\top)^{-1}L_1^\top\theta_2).}$$

- Mahalanobis geometry is the *extrinsic geometry of Riemannian tangent planes*
- In *deep learning*, Cholesky decomposition often used to ensure matrix Q is SPD

The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory* 57.8 (2011)

Dual parametrization of the Mahalanobis distance

- **Mahalanobis distance** is a metric distance (induced by a norm):

$$\mathrm{Mah}_Q(\theta,\theta') := \sqrt{(\theta-\theta')^\top Q(\theta-\theta')} = \|\theta-\theta'\|_Q$$

Usually positive-definite matrix Q is the inverse of the covariance matrix

- Mahalanobis is the **only symmetric Bregman divergence** [BDV'07]:

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

$$F(\theta) = \tfrac{1}{2}\theta^\top Q\theta \quad \Longrightarrow \quad B_F(\theta:\theta') = \frac{1}{2}\mathrm{Mah}_Q^2(\theta,\theta')$$

- The convex conjugate F* yields a **dual parameterization**: $F^*(\eta) = \frac{1}{2}\eta^\top Q^{-1}\eta$
- Since we have $B_F(\theta:\theta') = B_{F^*}(\eta':\eta)$, it follows that: $\eta = Q\theta$

$$\boxed{\mathrm{Mah}_Q(\theta,\theta') = \mathrm{Mah}_{Q^{-1}}(\eta,\eta') = \sqrt{(Q\theta-Q\theta')^\top Q^{-1}Q(\theta-\theta')} = \sqrt{(\theta-\theta')^\top Q(\theta-\theta')}}$$

[BVD'07] Bregman Voronoi Diagrams, arXiv:0709.2196

Sided and symmetrized Bregman centroids, IEEE transactions on Information Theory, 2009

Katalin Marton's f-divergence

1941-2019

DEFINITION: If p and r are probability measures on $\mathcal{X}$ then

$$\boxed{d_2(p,r) = \left[\int \left| 1 - \frac{dr}{dp}(y) \right|_+^2 dp(y) \right]^{1/2}.}$$

$$d_2(p,r) = \sqrt{I_{f_{\text{Marton}}}[p:r]} \qquad \text{with} \qquad f_{\text{Marton}}(u) := (u-1)^2 1\{u<1\}$$

Marton's square root f-divergence satisfies triangle inequality but not symmetric

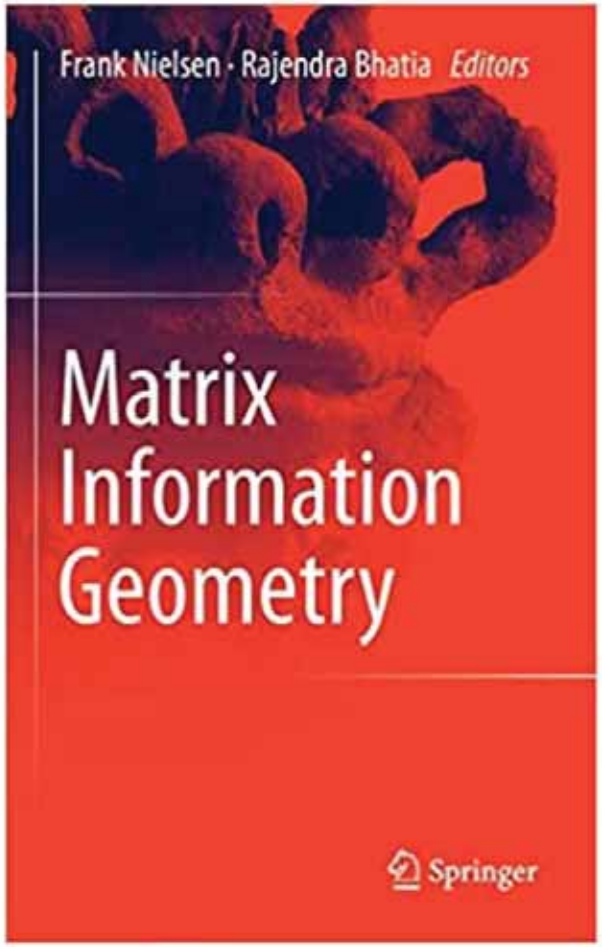
Geometric And Functional Analysis
Vol. 6, No. 3 (1996)

1016-443X/96/0300556-16$1.50+0.20/0
© 1996 Birkhäuser Verlag, Basel

A MEASURE CONCENTRATION INEQUALITY FOR CONTRACTING MARKOV CHAINS

K. MARTON

K. Marton
A measure concentration inequality for contracting Markov chains
Geometric and Functional Analysis, vol. 6, pp. 556–571, 1996.



Part I State-of-the-Art Surveys and Original Matrix Theory Work

1 **Supremum/Infimum and Nonlinear Averaging of Positive Definite Symmetric Matrices** 3
Jesús Angulo

2 **The Riemannian Mean of Positive Matrices** 35
Rajendra Bhatia

3 **The Geometry of Low-Rank Kalman Filters** 53
Silvère Bonnabel and Rodolphe Sepulchre

4 **KV Cohomology in Information Geometry** 69
Michel Nguiffo Boyom and Paul Mirabeau Byande

5 **Derivatives of Multilinear Functions of Matrices** 93
Priyanka Grover

6 **Jensen Divergence-Based Means of SPD Matrices** 111
Frank Nielsen, Meizhu Liu and Baba C. Vemuri

7 **Exponential Barycenters of the Canonical Cartan Connection and Invariant Means on Lie Groups** 123
Xavier Pennec and Vincent Arsigny

Part II Advanced Matrix Theory for Radar Processing

8 **Medians and Means in Riemannian Geometry: Existence, Uniqueness and Computation** 169
Marc Arnaudon, Frédéric Barbaresco and Le Yang

9 **Information Geometry of Covariance Matrix: Cartan-Siegel Homogeneous Bounded Domains, Mostow/Berger Fibration and Fréchet Median** 199
Frédéric Barbaresco

10 **On the Use of Matrix Information Geometry for Polarimetric SAR Image Classification** 257
Pierre Formont, Jean-Philippe Ovarlez and Frédéric Pascal

11 **Doppler Information Geometry for Wake Turbulence Monitoring** 277
Zhongxun Liu and Frédéric Barbaresco

Part III Matrix-Based Signal Processing Applications

12 **Review of the Application of Matrix Information Theory in Video Surveillance** 293
M. K. Bhuyan and T. Malathi

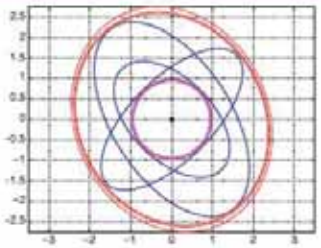
13 **Comparative Evaluation of Symmetric SVD Algorithms for Real-Time Face and Eye Tracking** 323
Tapan Pradhan, Aurobinda Routray and Bibek Kabi

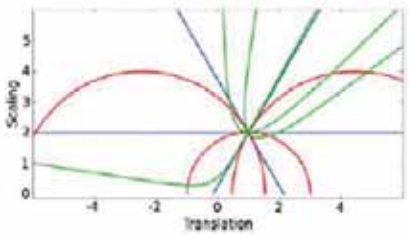
14 **Real-Time Detection of Overlapping Sound Events with Non-Negative Matrix Factorization** 341
Arnaud Dessein, Arshia Cont and Guillaume Lemaitre

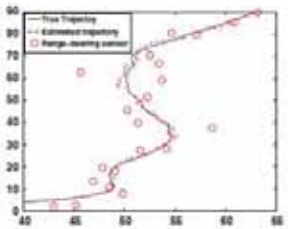
15 **Mining Matrix Data with Bregman Matrix Divergences for Portfolio Selection** 373
Richard Nock, Brice Magdalou, Eric Briys and Frank Nielsen

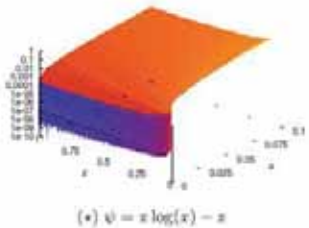
16 **Learning Mixtures by Simplifying Kernel Density Estimators** . . . 403
Olivier Schwander and Frank Nielsen

17 **Particle Filtering on Riemannian Manifolds. Application to Covariance Matrices Tracking** 427
Hichem Snoussi









Matrix spectral distances

- A d-variate function f is **symmetric** if it is invariant by any permutation σ of its arguments: $\forall \sigma, \quad f(x_1, \ldots, x_d) = f(\sigma(x_1), \ldots, \sigma(x_d))$
- The **eigenvalue map** Λ (M) of a matrix M gives its (unsorted) eigenvalues
- **Matrix spectral distance** with matrix combinator C:

$$\boxed{D(M_1, M_2) = f(\Lambda(C(M_1, M_2)))}$$

- Example of spectral matrix distances: Kullback-Leibler divergence between same mean Gaussians (see also the Siegel distance):

$$\begin{aligned} D_{\mathrm{KL}}(p_{\Sigma_1}, p_{\Sigma_2}) &= \int p_{\Sigma_1}(x) \log \frac{p_{\Sigma_1}(x)}{p_{\Sigma_2}(x)} \mathrm{d}x, \\ &= \frac{1}{2}\left(\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - d\right) \\ &= \frac{1}{2}\sum_{i=1}^{d}\left(\lambda_i(\Sigma_2^{-1}\Sigma_1) - \log \lambda_i(\Sigma_2^{-1}\Sigma_1) - 1\right) \\ &= (f_{\mathrm{KL}} \circ \Lambda)(\Sigma_2^{-1}\Sigma_1) \end{aligned}$$

$$C(\Sigma_1, \Sigma_2) = \Sigma_2^{-1}\Sigma_1$$

$$f_{\mathrm{KL}}(u_1, \ldots, u_d) = \frac{1}{2}\left(\sum_{i=1}^{d}(u_i - 1 - \log u_i)\right)$$

Hilbert geometry of the Siegel disk: The Siegel-Klein disk model, arxiv 2004.08160

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

MaxEnt distributions are exponential families: Application: MaxEnt upper bounds on entropy

Maximum entropy problem:

$\max_p h(p)$ such that $E_p[t(x)] = \eta$

Moment constraints

Exponential family:

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

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

$$h(p) = -F^*(\eta) = F(\theta) - \theta^\top \nabla F(\theta)$$

MaxEnt upper bound:

By definition, $\forall q, \quad h(q) \leq h(p)$

Example: Gaussian distribution is MaxEnt distribution for $t(x)=(x,x^2)$

$$\max_p h(p) \text{ such that } E_p[x] = \eta_1 = \mu \text{ and } E_p[x^2] = \eta_2 = \mu^2 + \sigma^2$$

MaxEnt upper bound

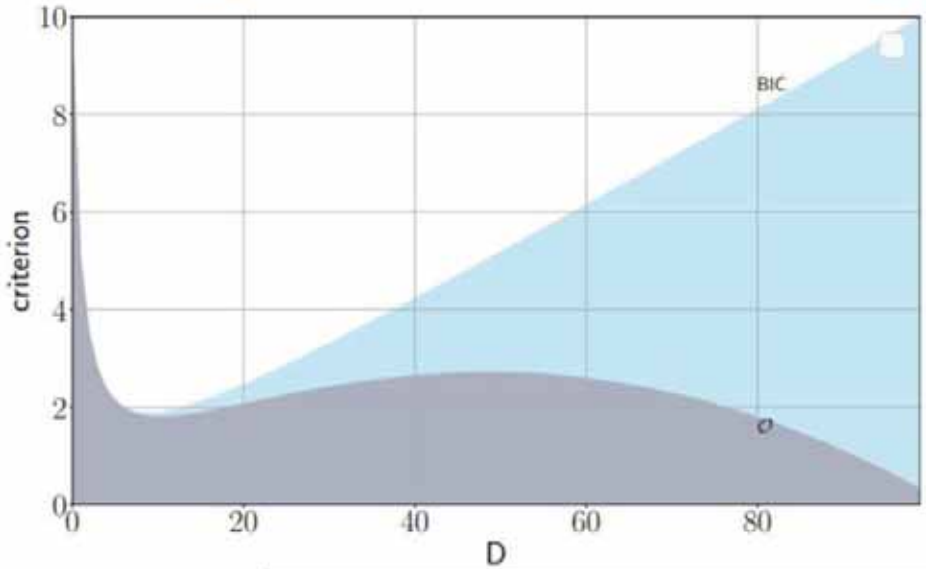
$$\forall q, \quad h(q) \leq h(p) = \frac{1}{2} \log (2\pi e \mathrm{Var}[q])$$

$$\mathrm{Var}[q] = \int x^2 q(x) \mathrm{d}\mu(x) - \left(\int x q(x) \mathrm{d}\mu(x) \right)^2$$

MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions

IEEE SPL 2017, arxiv:1612.02954

Minimum Description Length for Deep nets:
A singular differential geometric approach



- Varying local dimensionality of lightlike manifolds
- Prior interpolating Jeffreys's prior with Gaussian prior
- MDL which explains the "negative complexity" term in DNNs (similar to double descent risk curve)
- Intrinsic complexity of DNNs related to Fisher information spectrum

$$\mathcal{O} := -\log p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}) + \frac{D}{2}\log\frac{N}{2\pi} + \log V + \frac{1}{2}\log\frac{\left|\mathcal{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N\varepsilon_2^2}\boldsymbol{I}\right|}{\left|\mathcal{I}(\hat{\boldsymbol{\theta}}) + \varepsilon_1\boldsymbol{I}\right|}.$$

K. Sun, F. Nielsen. Lightlike Neuromanifolds, Occam's Razor and Deep Learning

https://arxiv.org/abs/1905.11027 web: InformationGeometry.xyz

Mediant inequalities: Application to bounding the Kullback-Leibler divergence between w-mixtures

Proposition **(Weighted mediant inequality)** *Let $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, and $w_1, \ldots, w_k$ be positive reals. We have:*

$$\boxed{\min_{i\in[k]} \frac{a_i}{b_i} \leq \frac{\sum_{i=1}^k w_i a_i}{\sum_{i=1}^k w_i b_i} \leq \max_{i\in[k]} \frac{a_i}{b_i}}$$

Mixture family:
(information geometry)

$$\mathcal{M} = \{m(x;w) = \sum_{i=1}^k w_i p_i(x) \ : \ w \in \Delta_k\}$$

$$m(x) = m(x;w)$$

$$m'(x) = m(x;w')$$

Bounding the Kullback-Leibler divergence:

$$\boxed{\log \min_{i\in[k]} \frac{w_i}{w'_i} \leq \mathrm{KL}[m:m'] \leq \mathrm{KL}[w:w'] \leq \log \max_{i\in[k]} \frac{w_i}{w'_i}}$$

Metric transformations

- Let $d(p,q)$ be a **metric distance** and $T(u)$ a real continue function defined for $u \geqslant 0$ with $T(u) > 0$ for $u > 0$ and $T(0) = 0$.
- T is a **metric transformation** if $D(p,q) := T(d(p,q))$ is a metric distance.
- First example: $T_\alpha(u) = u^\alpha$ for $\alpha \in (0,1)$ is a metric transformation.
- Second example, $T(u) = \frac{u}{1+u}$ is a metric transformation:

$$D(p,q) := \frac{d(p,q)}{1 + d(p,q)}$$

yields always a **bounded** metric distance. In that case, (X,d) and $(X, T(d))$ are topologically equivalent.

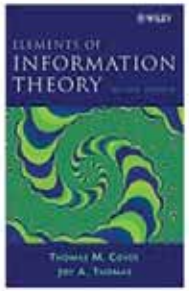
Metric transforms and subadditive functions

- Let (X, ρ) be a **metric space**, and $t : \mathbb{R}_+ \to \mathbb{R}_+$ be a function such that
  1. $t(0) = 0$
  2. t is strictly increasing
  3. t is subadditive: $t(a + b) \leqslant t(a) + t(b)$

  Then $(X, t \circ \rho)$ **is a metric space**
- Concave functions are subadditive
- Examples of metric transforms:
  - $t(u) = u^\alpha$ for $\alpha \in (0, 1]$
  - $t(u) = \frac{u}{u+c}$ for any $c > 0$
  - $t(u) = \log(1 + u)$
- Distances and (dis)similarities:

  `https://tinyurl.com/yamn6sbs`

Minimax redundancy code as a smallest enclosing Bregman ball



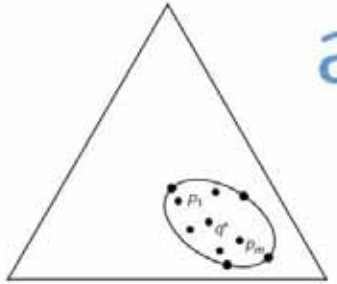


FIGURE 13.1. Minimum radius information ball containing all the p_λ's

Discrete alphabet $\mathcal{X} = \{A_1, \ldots, A_d\}$

Huffman codeword length for x: $l(x) = -\log p(x)$

Expected codeword length: Shannon entropy $H(X) = -\sum_x p(x) \log p(x)$

Redundancy of coding with q instead of true p: Kullback-Leiber KL(q:p)

Assume p belongs to $\mathcal{P} = \{p_{\lambda_1}, \ldots, p_{\lambda_n}\}$ then minimax redundancy:

$$R^* = \min_{p_\lambda} \max_{i \in \{1,\ldots,n\}} \mathrm{KL}(p_{\lambda_i} : p_\lambda)$$

Use natural coordinates of categorical distributions (exponential family with cumulant F)

$$\boxed{R^* = \min_{\theta \in \in \mathbb{R}^D} \max_{i \in [n]} B_{F_{\text{categorical}}}(p_\theta : p_{\theta_i})}$$

Smallest enclosing Bregman ball

On the smallest enclosing information disk, IPL 2008

Fitting the smallest enclosing Bregman ball, ECML 2005

Minkowski distances: Banach spaces versus F-spaces

- **Minkowski distances** are metric distances induced by **norms** for p ≥1:

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \quad p \geq 1 \quad \Rightarrow \quad M_p(x,y) = \|x-y\|_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

→ include **Euclidean distance** (p=2), **Manhattan distance** (p=1) and **Chebyshev distance** (p= ∞)

- For p in (0,1), $M_p^p(x,y)$ is a metric distance: $\|x\|_p := M_p^p(0,x)$ is a **F-norm**

- **Lebesgue space L^p** = measurable functions f s.t. $|f|^p$ is Lebesgue integrable
  (Banach space which is a Hilbert space with inner product only when p=2)
- Define the **Minkowski log ratio distance** between densities of L^p :

$$\boxed{M_p(f_1,f_2) := -\log \frac{\|f_1+f_2\|_p}{\|f_1\|_p + \|f_2\|_p} \geq 0} \quad \Rightarrow \quad \forall \lambda > 0, \quad M_p(\lambda f_1, \lambda f_2) := M_p(f_1,f_2)$$

Homogeneous distance

The Statistical Minkowski Distances: Closed-Form Formula for Gaussian Mixture Models, GSI 2019

Minkowski distances: Old and new ☺

- Sir Harold Jeffreys ruled out Minkowski **statistical distances** because they did not yield the Fisher information (but all f-divergences incl. Kullback-Leibler do)

$$I_1 = \int (\sqrt{dP'} - \sqrt{dP})^2, \quad I_2 = \int \log \frac{dP'}{dP} d(P' - P), \tag{1}$$

$$\int | (dP')^m - (dP)^m |^{1/m} \tag{2}$$

would behave similarly; but I_1 and I_2 are apparently the only ones that are ordinarily of the second order in the differences of the parameters in the laws when these differences are small.

I1= **Squared Hellinger divergence**

I2= **Jeffreys divergence**

- Studied the Minkowski statistical distances and give closed-form for mixtures:

For $\alpha \in (1, \infty)$ $\quad D_\alpha(p,q) := \|p\|_\alpha + \|q\|_\alpha - \|p+q\|_\alpha \geq 0.$

$L_\alpha(\lambda p, \lambda q) = L_\alpha(p,q)$ $\quad L_\alpha(p,q) := -\log \frac{\|p+q\|_\alpha}{\|p\|_\alpha + \|q\|_\alpha} = \log \frac{\|p\|_\alpha + \|q\|_\alpha}{\|p+q\|_\alpha} \geq 0.$

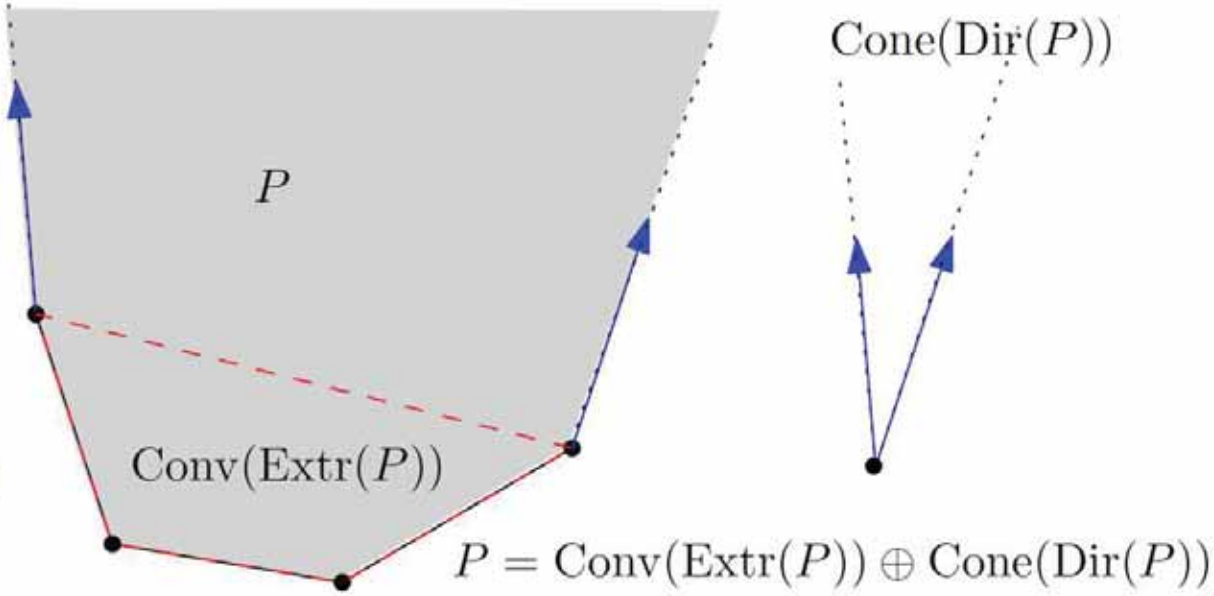
- **An invariant form for the prior probability in estimation problems, H. Jeffreys, 1946**
- **The statistical Minkowski distances: Closed-form formula for Gaussian Mixture Models, Springer LNCS GSI 2019, https://arxiv.org/abs/1901.03732**

Minkowski-Weyl theorem: duality H-polytope / V-polytope

Polytope:

$$P := \left\{ x \in \mathbb{R}^d \ : \ \underbrace{A}_{\mathbb{R}^{m \times d}} x \geq \underbrace{b}_{\mathbb{R}^m} \right\}$$

(e.g., feasible set of Linear Program)

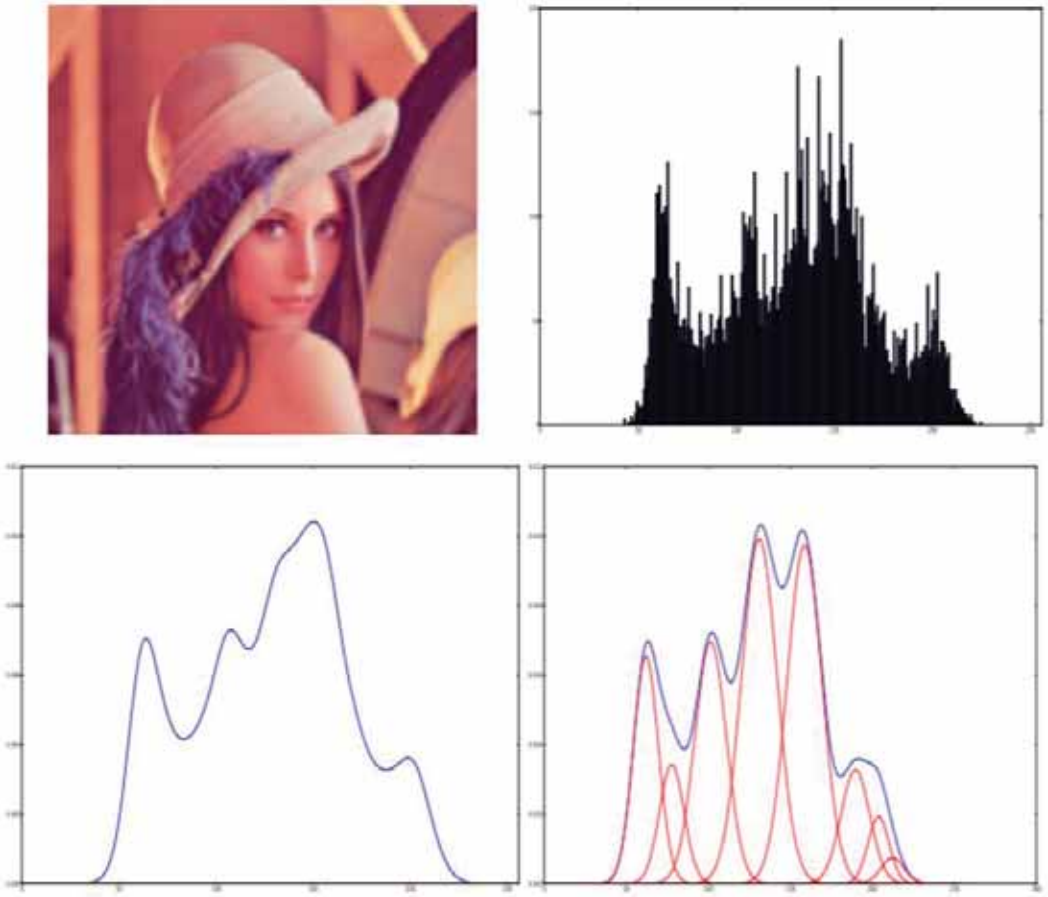


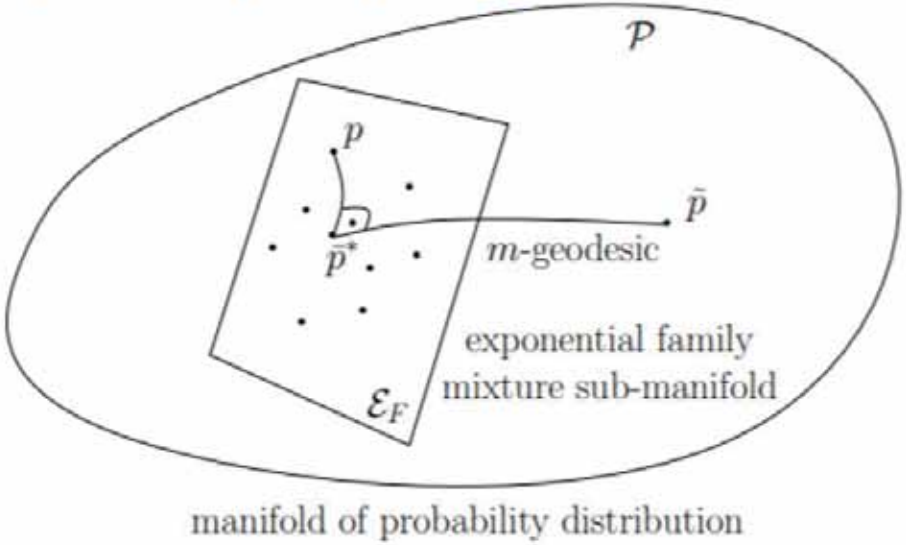
<u>Minkowski-Weyl decomposition theorem:</u>

$$\boxed{P = \mathrm{Conv}(\mathrm{Extr}(P)) \oplus \mathrm{Conv}(\mathrm{Dir}(P))}$$

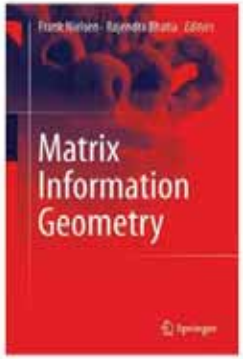
Bounded polytope (null cone): Equivalence between halfspace representation **H-polytope** (= intersection of supporting halfspaces) and vertex representation **V-polytope** (= convex hull of extreme points)

Learning statistical mixtures by simplifying KDEs





Kullback-Leibler divergence minimization as an information projection



"Learning mixtures by simplifying kernel density estimators."
***Matrix Information Geometry*. Springer, Berlin, Heidelberg, 2013. 403-426.**

Estimating the Kullback-Leibler divergence between densities with computationally intractable normalization factors

- Estimate the γ –divergence for small value of γ>0,

$$D_\gamma[p:q] := \frac{1}{\gamma(1+\gamma)} \log \int p^{1+\gamma} \mathrm{d}\mu - \frac{1}{\gamma} \log \int p q^\gamma \mathrm{d}\mu + \frac{1}{1+\gamma} \log \int q^{1+\gamma} \mathrm{d}\mu.$$

- γ –divergence is a projective divergence: $D_\gamma[\lambda p : \lambda' q] = D_\gamma[p:q]$
- γ –divergence tends to the Kullback-Leibler divergence:

$$\lim_{\gamma \to 0} D_\gamma[p:q] = D_{\mathrm{KL}}\left[\frac{p}{\int p \mathrm{d}\mu} : \frac{q}{\int q \mathrm{d}\mu}\right]$$

- Estimate with Monte Carlo stochastic sampling: $I_{\alpha,\beta}[p:q] = \int p^\alpha q^\beta \mathrm{d}\mu$

$$\hat{I}^{r,m}_{\alpha,\beta}[p:q] := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{r(x_i)} p^\alpha(x_i) q^\beta(x_i)$$

$x_1, \ldots, x_m$ are iid from a proposal distribution r

$$\lim_{m \to \infty} \hat{I}^{r,m}_{\alpha,\beta}[p:q] = I_{\alpha,\beta}[p:q]$$

Patch matching with polynomial exponential families and projective divergences, 2016

On estimating the Kullback-Leibler divergence between two densities with computationally intractable normalization factors, 2020

Some ML tricks

- **kernel trick**
  (Euclidean to Hilbert space ML, RKHS)
- **probability integral transform trick**
  (continuous distribution->uniform)
- **Maximum Mean Discrepancy trick**
  (avoid density estimation)
- **log-sum-exp (softmax) trick**
  (smoothen maximum function)
- **replica trick**
  (thermostatistics/log partition function)
- **Gumbel-max trick**
  (approximate expectation)
- **Russian Roulette trick**
  (unbiased estimator for sums/series)
- **Pearlmutter's fast Hessian-vector product**
- **contrastive divergence** (graphical models)
- **majorization minimization trick** (EM etc.)
- **Bayesian conjugacy trick**
  (conjugate priors of exponential families)
- **reparameterization trick**
  (backprop in DL through stochastic nodes)
- **duality tricks**
  (dual problem easier, Fenchel, OT, MI, etc.)
- **Feynman integral trick (Thermodynamic VO)**
  (complex integral solves easier)
- **doubly reparameterized Gradient Estimators**
- **backpropagation Through Time (BPTT)**
  (LSTM, etc.)
- **Woodbury matrix inversion Lemma**
  (natural gradient)
- **Hutchinson's stochastic trace trick**
  (randomized linear algebra for large scale ML)
- **concentration inequalities (theo. bounds)**
- **Kronecker factorization** (natural gradient K-FAC)
- **Bootstrapping trick**

Some mathematical tricks for machine learning

- **kernel trick**
  (Euclidean to Hilbert space ML, RKHS)
- **probability integral transform trick**
  (continuous distribution->uniform)
- **Maximum Mean Discrepancy trick**
  (avoid density estimation)
- **Log-sum-exp (softmax) trick**
  (smoothen maximum function)
- **Replica trick**
  (thermostatistics/log partition function)
- **Gumbel-max trick**
  (approximate expectation)
- **Russian Roulette trick**
  (unbiased estimator for sums/series)
- **Bayesian conjugacy trick**
  (conjugate priors of exponential families)
- **Reparameterization trick**
  (backprop in DL through stochastic nodes)
- **Duality trick**
  (dual problem easier Fenchel)
- **Feynman integral trick**
  (complex integral solves easier)
- **Doubly Reparameterized Gradient Estimators**
  (ICLR 2019)

Some ML tricks

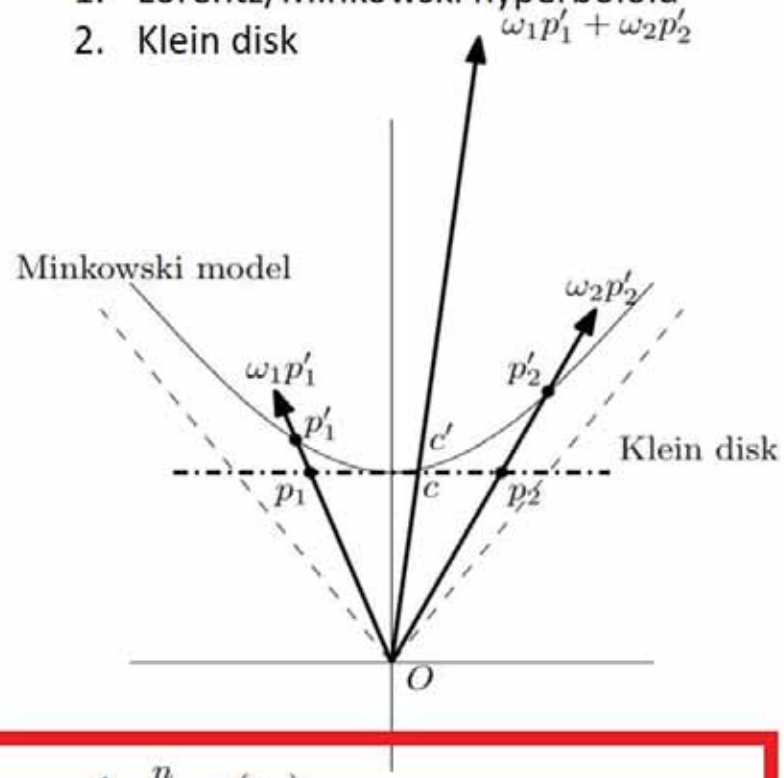
- **kernel trick**
  (Euclidean to Hilbert space ML, RKHS)
- **probability integral transform trick**
  (continuous distribution->uniform)
- **Maximum Mean Discrepancy trick**
  (avoid density estimation)
- **log-sum-exp (softmax) trick**
  (smoothen maximum function)
- **replica trick**
  (thermostatistics/log partition function)
- **Gumbel-max trick**
  (approximate expectation)
- **Russian Roulette trick**
  (unbiased estimator for sums/series)
- **Pearlmutter's fast Hessian-vector product**
- **contrastive divergence** (graphical models)
- **majorization minimization trick** (EM etc.)
- **Bayesian conjugacy trick**
  (conjugate priors of exponential families)
- **reparameterization trick**
  (backprop in DL through stochastic nodes)
- **duality tricks**
  (dual problem easier, Fenchel, OT, MI, etc.)
- **Feynman integral trick (Thermodynamic VO)**
  (complex integral solves easier)
- **doubly reparameterized Gradient Estimators**
- **backpropagation Through Time (BPTT)**
  (LSTM, etc.)
- **Woodbury matrix inversion Lemma**
  (natural gradient)
- **Hutchinson's stochastic trace trick**
  (randomized linear algebra for large scale ML)
- **concentration inequalities (theo. bounds)**
- **Kronecker factorization** (natural gradient K-FAC)
- **Bootstrapping trick**

Hyperbolic centroids/midpoints

- Karcher-Fréchet Riemannnian centroid not in closed form in hyperbolic manifolds

- Use Galperin's **model centroid** (defined for any constant curvature space, preserve invariance translation/rotation):
    1. Lift Klein hyperbolic point p_i to the upper hyperboloid sheet (Lorentz/Minkowski factor)
    2. Perform vector additions
    3. Renormalize so that the mean falls on the hyperboloid sheet (c')
    4. Convert back from the hyperboloid sheet to the Klein disk (c)
- Also called **Einstein midpoint** (Einstein gyrovector space) (e.g., Hyperbolic Attention Networks. ICLR 2019)

Two models of hyperbolic geometry:
1. Lorentz/Minkowski hyperboloid
2. Klein disk



$$c = \frac{1}{n}\sum_{i=1}^{n} \frac{\gamma(p_i)}{\Gamma} p_i \qquad \Gamma = \sum_{i=1}^{n} \gamma(p_i)$$

$$\gamma(p_i) = \frac{1}{\sqrt{1 - \|p_i\|^2}}$$

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

Möbius distance on the complex unit disk

Metric Möbius distance: $\boxed{\boldsymbol{m}(\lambda', \lambda'') := \left| \frac{\lambda' - \lambda''}{1 - \lambda' \bar{\lambda}''} \right|}$ $\lambda', \lambda'' \in \mathbb{D}$

Schwarz–Pick lemma: $\boldsymbol{m}(f(\lambda'), f(\lambda'')) \leq \boldsymbol{m}(\lambda', \lambda''), \lambda', \lambda'' \in \mathbb{D}$

Do not contract for automorphisms of the disk:

$f \in \mathrm{Aut}(\mathbb{D}) \Leftrightarrow \boldsymbol{m}(f(\lambda'), f(\lambda'')) = \boldsymbol{m}(\lambda', \lambda''), \lambda', \lambda'' \in \mathbb{D}$

August F. Möbius (1790-1868)

Poincaré hyperbolic Riemannian metric distance:

$$\boxed{\tanh^{-1}(\boldsymbol{m}(a, b)) \quad \text{with} \quad \tanh^{-1}(t) = \frac{1}{2} \log \frac{1+t}{1-t}}$$

Poincaré distance is lower bounded by m

Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010

Raw moments of multiparameter exponential families

- Exponential family with sufficient statistics $t_i(x)$: $\mathcal{E} := \left\{ p_\theta(x) := \exp\left(\sum_{i=1}^{D} \theta_i t_i(x) - F(\theta) \right) \; : \; \theta \in \Theta \right\}$
- Moment generating function of t(X): $m_\theta(u) := E\left[\exp\left(\sum_{i=1}^{D} u_i t_i(X) \right) \right] = \exp(F(\theta + u) - F(\theta))$
- Raw moments of t(X): $$\boxed{E_{p_\theta}[t_1(x)^{n_1} \dots t_D(x)^{n_D}] = \left. \frac{\partial^{\sum_{i=1}^{D} n_i}}{\partial u_1^{n_1} \dots \partial u_D^{n_D}} m_{\theta(u)} \right|_{u=0}}$$
- Example for the raw moments of the normal distributions: $t(x) = [t_1(x) \; t_2(x)]^\top = [x \; x^2]^\top$

$$E_{p_{\mu,\sigma}}[X^l] = \left(\frac{\partial^l}{\partial u_1^l} m_\theta(u) \right) \Bigg|_{u=0} \qquad F_N(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\theta_2} \right) \qquad \theta = (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$$

Maxima Code for $E[X^8]$

```
(%i4) F(theta1,theta2):=-(theta1·theta1/(4·theta2))+(1/2)·log(-%pi/theta2)$
MGF(theta1,theta2,u1,u2):=exp(F(theta1+u1,theta2+u2)-F(theta1,theta2))$
derivative(MGF(theta1,theta2,u1,u2),u1,8)$
ratsimp(ev(%, u1=0,u2=0,theta1=(mu/(sigma··2)),theta2=-1/(2·(sigma··2))));
```

(%o4) $105\ sigma^8 + 420\ mu^2\ sigma^6 + 210\ mu^4\ sigma^4 + 28\ mu^6\ sigma^2 + mu^8$

Muirhead's means and inequalities

Robert F. Muirhead
1860-1941

- For a n-dimensional positive normalized weight vector $w \in \Delta_n$ and a positive vector $x \in \mathbb{R}^n_{++}$, the **Muirhead's mean** is

$$\boxed{M_{[w]}(x_1, \ldots, x_n) = \frac{1}{n!} \sum_{\sigma} \prod_{i=1}^{n} x_{\sigma(i)}^{w_i}}$$

- Muirhead means include the *arithmetic*, *geometric* and *Heinz means*:

$$M_{[(1,0,\ldots,0)]}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} x_i}{n} := A(x_1, \ldots, x_n) \qquad M_{[(\frac{1}{n},\ldots,\frac{1}{n})]}(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i^{\frac{1}{n}} := G(x_1, \ldots, x_n)$$

$$M_{[\alpha,1-\alpha]}(x_1, x_2) = \mathrm{Heinz}(x_1, x_2; \alpha, 1-\alpha) := \frac{x_1^{\alpha} x_2^{1-\alpha} + x_1^{1-\alpha} x_2^{\alpha}}{2}$$

- Muirhead are distinct for different weight vectors:

$$M_{[w_1]}(x_1, \ldots, x_n) = M_{[w_2]}(x_1, \ldots, x_n) \Leftrightarrow w_1 = w_2 \text{ or } x_1 = \ldots = x_n$$

- Muirhead's inequality generalizes the Arithmetic-Geometric inequality:

$$\boxed{M_{[w_1]}(x_1, \ldots, x_n) \leq M_{[w_2]}(x_1, \ldots, x_n) \Leftrightarrow \exists \text{ doubly stochastic matrix } P \text{ such that } w_1 = P w_2}$$

Muti-continued fractions

$$\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}}$$

n-th convergent	Continued Fraction Expansion	Rational	Decimal Rep.
0	[1]	$\frac{1}{1}$	1
1	[1/2]	$\frac{3}{2}$	1.5
2	[1/2/3]	$\frac{10}{7}$	1.428571429...
3	[1/2/3/4]	$\frac{43}{30}$	1.433333333...
4	[1/2/3/4/5]	$\frac{225}{157}$	1.433121019...

Convergents of $\frac{225}{157} = [1/2/3/4/5]$.

1.000000000	$\frac{1}{1}$	$1/\infty$
1.500000000	$\frac{3}{2}$	$1/2/\infty$
1.570000000	$\frac{157}{100}$	$1/1/1/3/14/\infty$
1.570000000	$\frac{157}{100}$	$1/1/1/3/14/\infty$
1.570700000	$\frac{15707}{10000}$	$1/1/1/3/27/1/2/1/1/1/4/\infty$
1.570790000	$\frac{157079}{100000}$	$1/1/1/3/31/1/2/16/4/2/\infty$
1.570796000	$\frac{392699}{250000}$	$1/1/1/3/31/1/41/1/1/2/1/3/\infty$
1.570796300	$\frac{15707963}{10000000}$	$1/1/1/3/31/1/121/3/1/4/3/1/1/2/\infty$
1.570796320	$\frac{9817477}{6250000}$	$1/1/1/3/31/1/138/1/2/2/1/4/4/\infty$
1.570796326	$\frac{785398163}{500000000}$	$1/1/1/3/31/1/144/1/18/8/2/1/7/4/\infty$

Approximations of $\frac{\pi}{2}$.

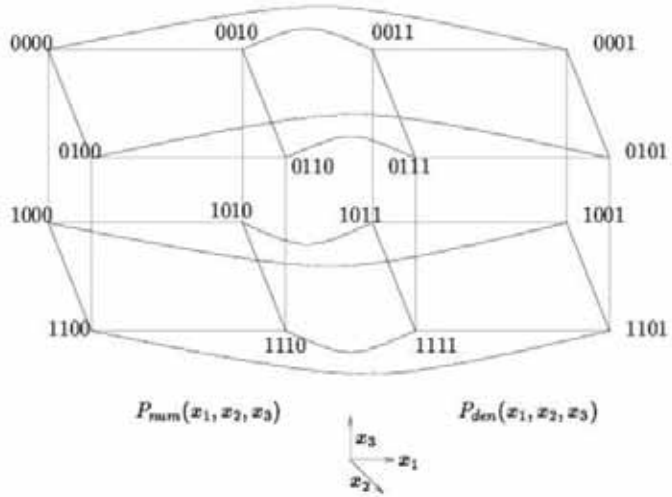
Matrix representation of continued fractions $\frac{p}{q} = [a_0/\ldots/a_n]$

If n is odd:

$$\frac{p}{q} \equiv (p\ q) = (x_i\ 1) = (1\ 0) \times \begin{pmatrix} 1 & a_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ a_{n-1} & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 \\ a_{2j} & 0 \end{pmatrix} \begin{pmatrix} 1 & a_{2j-1} \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 \\ a_0 & 1 \end{pmatrix}$$

If n is even:

$$\frac{p}{q} \equiv (p\ q) = (x_i\ 1) = (0\ 1) \times \begin{pmatrix} 1 & 0 \\ a_n & 1 \end{pmatrix} \begin{pmatrix} 1 & a_{n-1} \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & a_{2j+1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ a_{2j} & 0 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 \\ a_0 & 1 \end{pmatrix}$$



@FrnkNlsn

Algorithms on Continued and Multi-continued fractions, 1993

Multinoulli Fisher-Rao manifold = Einstein manifold

Multinoulli distribution

= multinomial for one trial

(sample space = alphabet of size d)

Multinoulli family

= **exponential family** of order D=d-1

Natural parameter space

= open D-dim. standard simplex

Fisher information matrix:

$$\boxed{g_{ij} = \frac{\delta_{ij}}{\theta_i} + \frac{1}{1 - \sum_{i=1}^{D} \theta_i}}$$

Fisher metric = Shahshahani metric

Constant sectional curvature $\kappa_{\text{sec}} = \frac{1}{4}$

Ricci curvature $\boxed{R_{ij} = \frac{D-1}{4} g_{ij}}$

Einstein manifold

Scalar curvature: $\kappa = \frac{D(D-1)}{4}$

	Multinomial ($n = 1$ trial)
density	$\frac{1}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
support	$\{E_1, ..., E_k\}$
base measure	Counting measure
auxiliary carrier $k(x)$	$-\sum_{i=1}^{k} \log x_i!$
sufficient statistics $t(x)$	$(x_1, \cdots, x_{k-1})$
$\theta(\lambda)$	$\left(\log\left(\frac{p_i}{p_k}\right)\right)_i$
Order D	$d-1$
log-normalizer $F(\theta)$	$\log\left(1 + \sum_{i=1}^{k-1} \exp\theta_i\right)$
$\eta = \nabla F(\theta) = E[t(X)]$	$\left(\frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j}\right)_i$
$\theta = \nabla F^*(\eta)$	$\left(\log\left(\frac{\eta_i}{1-\sum_{j=1}^{k-1}\eta_j}\right)\right)_i$
$F^*(\eta) = \langle\theta, \eta\rangle - F(\theta)$	$\left(\sum_{i=1}^{k-1}\eta_i \log\eta_i\right) + \left(1 - \sum_{i=1}^{k-1}\eta_i\right)\log\left(1 - \sum_{i=1}^{k-1}\eta_i\right)$
Kullback-Leibler divergence	$p_{1,k}\log\frac{p_{1,k}}{p_{2,k}} - \sum_{i=1}^{k-1} p_{1,i}\log\frac{p_{2,i}}{p_{1,i}}$
Fisher information $I(\theta) = \nabla^2 F(\theta)$	$I_{ii} = \frac{1}{p_i} + \frac{1}{p_d}, \quad I_{ij} = \frac{1}{p_d} (i \neq j)$
MLE $\hat{\eta} = \bar{t}$	$\hat{p}_i = \frac{n_i}{n}$
$\lambda(\theta)$	$\begin{cases} p_i = \frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j} & \text{if } i < k \\ p_k = \frac{1}{1+\sum_{j=1}^{k-1}\exp\theta_j} \end{cases}$
$\lambda(\eta)$	$\begin{cases} p_i = \eta_i & \text{if } i < k \\ p_k = 1 - \sum_{i=1}^{k-1}\eta_j \end{cases}$

Pattern learning and recognition on statistical manifolds, SIMBAD 2013

Invariance of mutual information (MI)

- Let X and Y be two random variables of R^d
- The **mutual information** between X and Y is the **symmetric information dissimilarity**:

$$\begin{aligned} I(X;Y) &:= D_{\mathrm{KL}}[p_{(X,Y)}(x,y) : p_X(x)p_Y(y)] \geq 0 \\ &= \int\int p_{(X,Y)}(x,y) \log \frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} \mathrm{d}x\mathrm{d}y \\ &= I(Y;X) \end{aligned}$$

- MI measures the **mutual dependence** between X and Y (0 if independent) :

$$X_1 \perp X_2 : p_{(X,Y)}(x,y) = p_X(x)p_Y(y) \Leftrightarrow I(X_1;X_2) = 0$$

- For any invertible and differentiable transformations t_1 and t_2, $t_1, t_2 : \mathbb{R}^d \to \mathbb{R}^d$

we have the following invariance:

$$\boxed{I(t_1(X_1);t_2(X_2)) = I(X_1;X_2)}$$

Mutual information = negentropy of copula density

- For any random variable X with cumulative distribution function F_X, we can get the uniform distribution U on [0,1] by taking **$U=F_X(X)$**
- **Copula** $C(u_1,u_2)$ of (X_1,X_2) = distribution function of (U_1,U_2):

$$C(u_1, u_2) := \Pr(U_1 \leq u_1, U_2 \leq u_2)$$

- **Copula density** is

$$c(u_1, u_2) := \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}$$

- Copula captures all the **dependence information between X and Y:**

$$I(X;Y) = 0 \Leftrightarrow c(u_1, u_2) = 0 \quad \forall (u_1, u_2) \in [0,1]^2$$

- Mutual information I(X1,X2) is the negative entropy of the copula density:

$$\boxed{I(X;Y) = \int_0^1 \int_0^1 c(u_1, u_2) \log c(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2}$$

(with 0log 0=0)

Optimal Transport vs. Fisher-Rao distance between Copulas for Clustering Multivariate Time Series, IEEE SSP 2016
Optimal Copula Transport for Clustering Multivariate Time Series, IEEE ICASSP 2016

Multivariate Bregman Divergences as univariate BDs

- **MultiVariate Bregman divergence** (MV BD) for F strictly convex and smooth:

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

- MV BDs can be interpreted as a *family of UniVariate BDs*:

$$\boxed{B_F(\theta_1 : \theta_2) = B_{F_{\theta_1,\theta_2}}(0:1), \quad \forall \theta_1 \neq \theta_2,}$$

with

$$\boxed{F_{\theta_1,\theta_2}(u) := F(\theta_1 + u(\theta_2 - \theta_1))}$$

Strictly convex

- Proof:

Directional Derivative:

$$\begin{aligned} \nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(u) &:= \lim_{\epsilon \to 0} \frac{F(\theta_1 + (\epsilon + u)(\theta_2 - \theta_1)) - F(\theta_1 + u(\theta_2 - \theta_1))}{\epsilon} \\ &= (\theta_2 - \theta_1)^\top \nabla F(\theta_1 + u(\theta_2 - \theta_1)), \end{aligned}$$

Therefore we have:

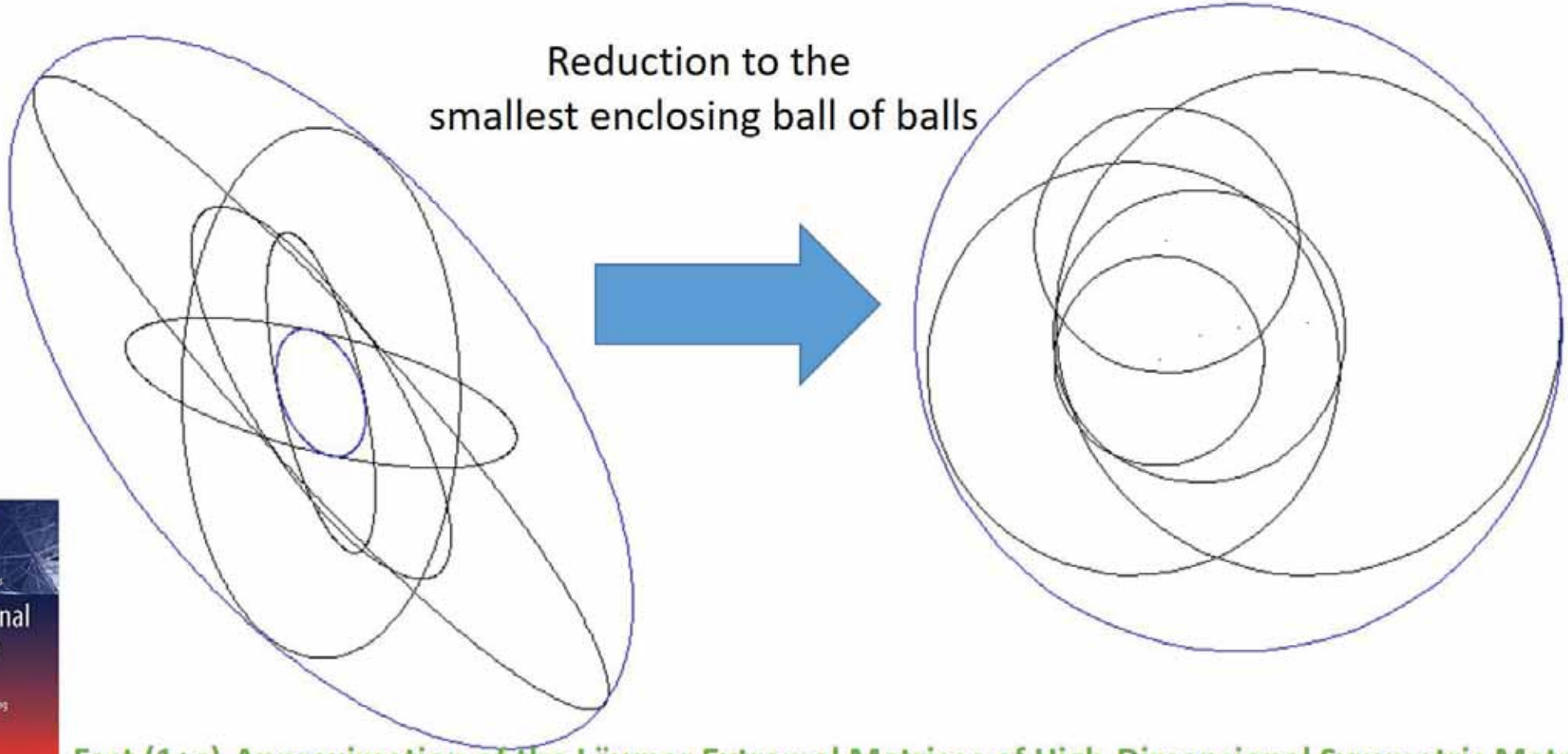
$$\begin{aligned} B_{F_{\theta_1,\theta_2}}(0:1) &= F_{\theta_1,\theta_2}(0) - F_{\theta_1,\theta_2}(1) - (0-1)\nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(1), \\ &= F(\theta_1) - F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2), \\ &= B_F(\theta_1 : \theta_2). \end{aligned}$$

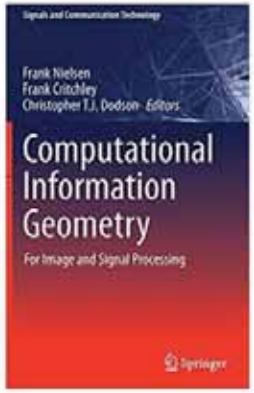
Bregman chord divergence, GSI 2019

An Elementary Introduction to Information Geometry, Entropy 2020, 22(10), 1100

Minimum volume enclosing ellipsoid of zero-centered ellipsoids
Maximum volume enclosed ellipsoid of zero-centered ellipsoids

Reduction to the
smallest enclosing ball of balls





Fast (1+ε)-Approximation of the Löwner Extremal Matrices of High-Dimensional Symmetric Matrices, Computational Information Geometry: For Image and Signal Processing, 2017

Multivariate normals: A multivariate location-scale family

- **MultiVariate (MV) location-scale families** generalize location-scale families:

$$\mathcal{P} = \left\{ p_{\mu,P}(x) := |P|^{-1} p_{\mathrm{std}}(P^{-1}(x-\mu)) \ : \ P \succ 0, \mu \in \mathbb{R}^d \right\}$$

$$X \sim p_{\mathrm{std}}(x) \quad \Rightarrow \quad X_{\mu,P} \sim p_{\mu,P}(x) \quad \text{with} \quad X_{\mu,P} := PX + \mu$$

- **MV normals** and **MV Cauchy** are examples of MV location-scale families:

$$p^{\mathcal{N}}_{\mathrm{std}}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2} x^\top x\right) \quad \text{with} \quad P(\Sigma) := \Sigma^{\frac{1}{2}}$$ (Symmetric square-root matrix)

- MV location-scale families include **MV elliptical distributions**
- **MV location-scale group** and **action** of the group onto densities:

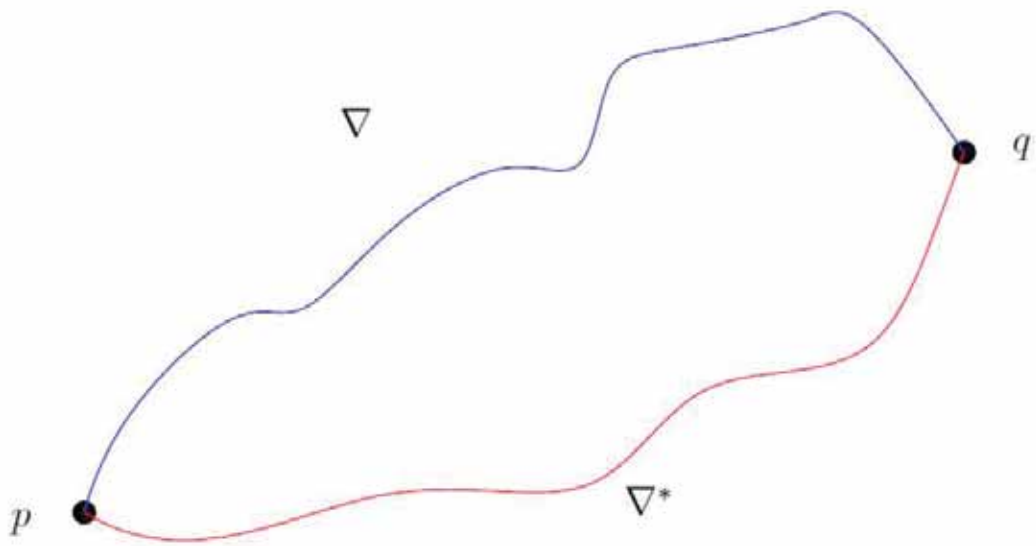
$$g_{l,P} \odot p := |P|^{-1} p\left(|P|^{-1}(x-l)\right) \qquad g_{l_2,P_2} \cdot g_{l_1,P_1} \odot p(x) \ = \ (P_2P_1)^{-1} p\left((P_2P_1)^{-1}(x - l_2 - P_2 l_1)\right)$$

- Invariance of f-divergences wrt group action: **arXiv:2101.03839**

$$I_f(p_{l_1,P_1} : q_{l_2,P_2}) = I_f\left(p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2}\right) = I_f\left(p_{P_2^{-1}(l_1-l_2),P_2^{-1}P_1} : q\right)$$

The Gaussian dually flat manifold: (M, g, ∇, ∇^*)

A Bregman manifold



$$M = \{p_{\mu,\Sigma} \ : \ \mu \in \mathbb{R}^D, \Sigma \in \mathbb{P}_D\}$$

g: Fisher information metric

∇: e-connection

∇*: m-connection

∇*-geodesic
(=mixture geodesic)

$$(p_1 p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ \Sigma^m_\alpha = \bar{\Sigma}_\alpha + (1-\alpha)\mu_1\mu_1^\top - \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha{}^\top \end{cases}$$

∇-geodesic
(=exponential geodesic)

$$(p_1 p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \Sigma^e_\alpha((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2) \\ \Sigma^e_\alpha = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1} \end{cases}$$

An elementary introduction to information geometry,arXiv:1808.08271 (2018)

Invariance is the only constant in change.

Naming the Kullback-Leibler distance!

Kullback mentioned already several different aliases in 1983:

The information measure and its properties defined and discussed herein, may be found in the technical literature under various names: Kullback–Leibler information, directed divergence, discrimination information, Rényi's information gain, expected weight of evidence, entropy distance, entropy*, cross-entropy. The term "discrimination information" is used herein. Con-

Kullback–Leibler information,
directed divergence,
discrimination information,
Renyi's information gain,
expected weight of evidence
entropy distance,
entropy*,
cross-entropy.

Kullback preferred to call it the **discrimination information**
minimum discrimination information (MDI)

(1987) Letters to the Editor, The American Statistician, 41:4, 338-341, DOI: 10.1080/00031305.1987.10475510

THE KULLBACK–LEIBLER DISTANCE

As the Kullback of the Kullback–Leibler (KL) distance discussed by Hastie (1987), I feel constrained to comment on that nomenclature. As mentioned in Kullback (1983), this measure may be found in the technical literature under nine different names. Although Hastie introduces the KL distance with a reference to Kullback (1959), a reader looking for the KL distance in Kullback (1959) would not find it there. The KL distance was introduced by some writers based on Kullback and Leibler (1951).

My own preference has been the term *discrimination information* and its basis for the principle of minimum discrimination information (MDI) (Gokhale and Kullback 1978; Kullback 1959, 1983; Kullback and Keegel 1984; Kullback, Keegel, and Kullback 1987), as more indicative of its statistical origins. The relevance of the principle of MDI to Hastie's consideration of the family of one-parameter exponential densities is that the principle of MDI leads quite naturally to estimates that are members of multiparameter exponential families. Kullback et al. (1987) includes illustrations and applications of MDI to stochastic processes and random functions.

Solomon Kullback
Professor Emeritus of Statistics
George Washington University
Washington, D.C. 20052

References

Gokhale, D. V., and Kullback, S. (1978), *The Information in Contingency Tables*, New York: Marcel Dekker.

Hastie, T. (1987), "A Closer Look at the Deviance," *The American Statistician*, 41, 16–20.

Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley. [Reprinted in 1968 by Dover Publications (New York) and in 1978 by Peter Smith Publisher (Magnolia, MA).]

© *1987 American Statistical Association*

——— (1983), "Kullback Information," in *Encyclopedia of Statistical Sciences* (Vol. 4), eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp. 421–425.

Kullback, S., and Keegel, J. C. (1984), "Categorical Data Problems Using Information Theoretic Approach," in *Handbook of Statistics* (Vol. 4), eds. P. R. Krishnaiah and P. K. Sen, New York: Elsevier, pp. 831–871.

Kullback, S., Keegel, J. C., and Kullback, J. H. (1987), *Topics in Statistical Information Theory* (*Lecture Notes in Statistics*, Vol. 42), New York: Springer-Verlag.

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Natural gradient application: Natural Evolution Strategies

- Minimize a d-dimensional **fitness function** f: R^d→R

$$\min_{x \in \mathbb{R}^d} f(x)$$

- **Stochastic relaxation**: Minimize the ***expected fitness*** wrt a **mutation** distribution (eg, multivariate normal):

mutation distribution space: $\mathcal{P} = \{p_\theta \ : \ \theta \in \Theta\}$

$$\min_{\theta \in \Theta} L(\theta) = E_{p_\theta}[f(x)]$$

- **Monte Carlo approximation** of the gradient:

$$x_1, \ldots, x_n \simeq_{\text{iid}} p_\theta$$

$$\nabla_\theta L(\theta) \simeq \frac{1}{n} f(x_i) \nabla_\theta \log p_\theta(x_i)$$

- **Natural gradient**:

$$\theta_{t+1} = \theta_t - \alpha_t I_{\mathcal{P}}^{-1}(\theta_t) \nabla_\theta L(\theta_t)$$

Fisher information matrix (FIM): $I_{\mathcal{P}}(\theta) = E_{p_\theta}\left[(\nabla_\theta \log p_\theta(x))(\nabla_\theta \log p_\theta(x))^\top\right]$

- Efficient incremental update of the FIM inverse: $I_{\mathcal{P}}(\theta)^{-1}$

Yi Sun et al., Efficient Natural Evolution Strategies, Journal of Machine Learning Research 15 (2014) 949-980

FN, An elementary introduction to information geometry, arXiv:1808.08271

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

In a dually flat space (Hessian manifold), we have

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

Natural gradient

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

Ordinary gradient

→ Used in variational inference (VI)

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

© Frank Nielsen

Natural gradient as Riemannian gradient with rectraction exponential map approximation

- **Natural gradient** descent on a Riemannian manifold M with metric tensor g:
- L: loss function to minimize: $\theta_{t+1} \leftarrow \theta_t - \alpha_t g^{-1}(\theta_t)\nabla_\theta L(\theta_t)$
- Natural gradient may leave the manifold! **Riemannian gradient** relies on the **Riemannian exponential map** and ensure it stays on the manifold:

$$\boxed{\theta_{t+1} \leftarrow \exp_{\theta_t}(-\alpha_t \nabla_g L(\theta_t))}$$

- Riemannian exponential difficult to calculate, use a computable **retraction** R.

When $R_\theta(v) = \theta + v$, we get a 1st order Taylor approximation of the exponential map, and we **recover the natural gradient** since $\nabla_g L(\theta) = g^{-1}(\theta)\nabla L(\theta)$

[EIG] An elementary introduction to information geometry, arxiv.org:1808.08271

[BM] On geodesic triangles with right angles in a dually flat space, arxiv:1910.03935

[B] Stochastic gradient descent on Riemannian manifolds, IEEE TAC, 2013

Variance function of natural exponential families

- Let $(\mathcal{X}, \Sigma, \nu)$ be a measurable space with positive measure ν
- A **Natural Exponential Family** (NEF) has density wrt to ν

$$p_\theta(x) = \exp\left(\sum_i x_i\theta_i - F(\theta)\right) \Rightarrow \boxed{F(\theta) = \log \int_{\mathcal{X}} \exp\left(\sum_i x_i\theta_i\right) \mathrm{d}\nu(x)}$$

- Function F is strictly convex and real-analytic (= Bregman generator)
- **Variance function** is parameterized by the dual parameterization: $\eta(\theta) = \nabla F(\theta)$

$$\boxed{\eta \mapsto V_F(\eta) = \nabla^2 F(\nabla F^*(\eta)) = \left(\nabla^2 F^*(\nabla F^*(\eta))\right)^{-1}}$$

Convex Conjugate F* $\Updownarrow$

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

- **<u>Theorem</u>: Variance function fully characterizes the NEF**
- Only six 1D NEFs with quadratic variance functions (Morris, 1982), twelve with cubic variance function (Letac and Mora, 1990), etc.

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

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

In a dually flat space (Hessian manifold), we have

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

Natural gradient

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

Ordinary gradient

Yet another proof for the non-negativeness of f-divergences

- **f-divergence** for a convex generator f, strictly convex at 1, with f(1)=0:

$$I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$$

- Kullback-Leibler divergence recovered for the generator f(u)=-log(u)
- Classic proof of **non-negativeness** (=Gibbs inequality) of f-divergences uses the Jensen's inequality:

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

- Another proof uses **Bregman scalar divergences** since the generator f is defined up to a term: $I_f[p:q] = I_g[p:q] \leftrightarrow f(u) = g(u) + \lambda(u-1), \lambda \in \mathbb{R}$

Choose $g(u) = f(u) - (u-1)f'(1) = f(u) - f(1) - (u-1)f'(1) =: B_f(u:1)$

$$\begin{aligned} I_f[p:q] &= \int p(x)\left(f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right)\mathrm{d}\mu(x) \\ &= \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right)\mathrm{d}\mu(x) \geq 0 \end{aligned}$$

Non-negative Monte Carlo estimation of f-divergences:

$$\boxed{\hat{I}_{f,n}[p:q] := \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\frac{q(x_i)}{p(x_i)}\right) - f'(1)\left(\frac{q(x_i)}{p(x_i)} - 1\right)\right) \geq 0 \quad x_1, \ldots, x_n \sim_{\text{iid}} p}$$

Non-negative estimations of the Jensen-Shannon divergence

- Traditionally estimate the Kullback-Leibler divergence:

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{r(x_i)} \log \frac{p(x_i)}{q(x_i)}, \quad x_1, \ldots, x_n \sim_{\mathrm{iid}} r.$$

$\longrightarrow$ but may be negative because $\sum_i p(x_i) \neq 1$ and $\sum_i q(x_i) \neq 1$

- Non-negative KLD estimation:

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n}\sum_{i=1}^{n} \left(\log \frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1 \right) \geq 0.$$

(with proposal distribution r=p)

- Jensen-Shannon divergence is a f-divergence $I_f[p:q] := \int p f\left(\frac{q}{p}\right) \mathrm{d}\mu.$ for

$$\mathrm{JS}[p:q] = \frac{1}{2}\mathrm{KL}\left[p : \frac{p+q}{2}\right] + \frac{1}{2}\mathrm{KL}\left[q : \frac{p+q}{2}\right]$$

$$f(u) := -\frac{1+u}{2}\log\left(\frac{1+u}{2}\right) + \frac{u}{2}\log(u)$$

- Non-negative estimation of the f-divergence:

$$\hat{I}_{f,n}(p:q) = \frac{1}{n}\sum_{i=1}^{n} f\left(\frac{q(x_i)}{p(x_i)}\right) - f'(1)\left(\frac{q(x_i)}{p(x_i)} - 1\right) \geq 0.$$

fJSD'(1)=-1/2

$$\boxed{\frac{1}{2s}\sum_{i=1}^{s}\left\{\frac{q(x_i)}{p(x_i)}\log\left(\frac{2q(x_i)}{p(x_i)+q(x_i)}\right) + \log\left(\frac{2p(x_i)}{p(x_i)+q(x_i)}\right) + \frac{q(x_i)}{p(x_i)}\right\} - \frac{1}{2}, \quad x_1, \ldots, x_s \sim p}$$

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid, Entropy 2020

Non-negative Monte Carlo estimation of f-divergences, 2020

Non-negative Monte Carlo estimator of f-divergences

f-divergence: $I_f(p:q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$

Monte Carlo estimator (r is proposal distribution):

$x_1, \ldots, x_n \sim_{\text{iid}} r$ $\qquad \widehat{\mathrm{KL}}_{r,n}(p:q) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{r(x_i)} \log \frac{p(x_i)}{q(x_i)}$

Problem: can be potentially negative!!!

Solution: MC estimation on **extended f-divergences**:

$$\boxed{I_f(p:q) = \int p(x)\left(f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right) \mathrm{d}\mu(x)}$$

Non-negative MC estimation:

$$\hat{I}_f^{r,n}(p:q) = \frac{1}{n}\sum_{i=1}^{n} f\left(\frac{q(x_i)}{p(x_i)}\right) - f'(1)\left(\frac{q(x_i)}{p(x_i)} - 1\right) \geq 0$$

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n}\sum_{i=1}^{n} \left(\log \frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1\right) \geq 0$$

Non-negative Monte Carlo Estimation of f -divergences

Two non-parametric expressions of the Fisher information

- **Parametric** Fisher information matrix expressions:

$$I(\theta) = [I_{ij}(\theta)]_{ij}, \quad I_{ij}(\theta) := \mathrm{Cov}\left(\frac{\partial}{\partial\theta_i}\log p_\theta, \frac{\partial}{\partial\theta_j}\log p_\theta\right)$$

Regularity conditions I → $I_1(\theta) = E_{p_\theta}\left[\nabla_\theta \log p_\theta(x)(\nabla_\theta \log p_\theta(x))^\top\right]$

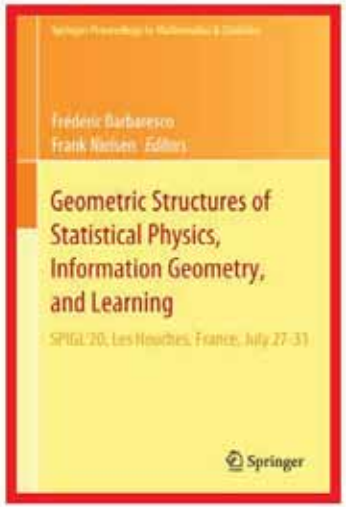
Regularity conditions II → $I_2(\theta) = -E_{p_\theta}\left[\nabla_\theta^2 \log p_\theta(x)\right]$

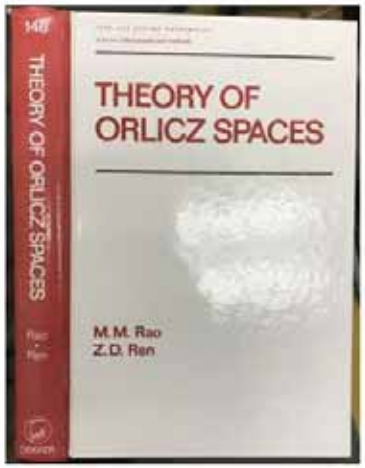
- **Nonparametric** Fisher information expressions (univariate densities):

→ $$I_1(p) = \int \frac{\nabla_x p(x) \times \nabla_x p(x)}{p(x)} \mathrm{d}\mu(x) = \int \nabla_x p(x) \times \nabla_x \log p(x) \mathrm{d}\mu(x)$$

→ $$I_2(p) = -\int \nabla_x^2 (p(x)\log p(x)) \mathrm{d}\mu(x)$$

Orlicz spaces for non-parametric information geometry





Władysław R. Orlicz
(1903-1990)

Consider a σ-finite measure μ on space X:

Young function: convex $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ and

$$\lim_{x\to+\infty} \frac{\Phi(x)}{x} = +\infty \quad \text{and} \quad \lim_{x\to 0} \frac{\Phi(x)}{x} = 0$$

L_Φ is **vector space** spanned by $\tilde{L}_\Phi := \left\{ f \ : \ \int_X \Phi(|f(x)|)\mathrm{d}\mu(x) < \infty \right\}$

Pistone, A Lecture About the Use of Orlicz Spaces in Information Geometry
Geometric Structures of Statistical Physics, Information Geometry, and Learning, Springer 2021

Neuromanifolds, Occam's Razor and Deep Learning

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

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

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

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

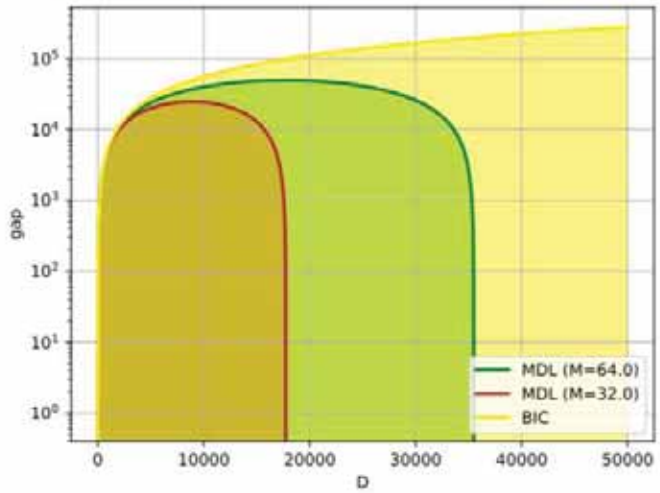
Θ: parameters of the DNN, $\hat{\Theta}$: estimated parameters

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

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

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

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



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

https://arxiv.org/abs/1905.11027

Quadratic entropies and Onicescu informational energy

Onicescu's informational energy: $I(p) := \int p^2(x)\mathrm{d}\mu(x) = E_p[p] > 0.$

Strictly convex function

Renyi quadratic entropy: $R_2(p) := -\log \int p^2(x)\mathrm{d}\mu(x) = -\log I(p)$

Vajda quadratic entropy: $V_2(p) := 1 - \int p^2(x)\mathrm{d}\mu(x) = 1 - I(p).$

Onicescu's **correlation coefficient** : $\rho(p,q) := \frac{I(p,q)}{\sqrt{I(p)I(q)}},\ I(p,q) := \int p(x)q(x)\mathrm{d}\mu(x)$

Closed-form formula for exponential families:

$$I(p_\theta) = \exp\left(F(2\theta) - 2F(\theta)\right) E_{p_{2\theta}}\left[\exp(k(x))\right]$$

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

$$\boxed{\rho(p_{\theta_1}, p_{\theta_2}) = \exp(-J_F(2\theta_1 : 2\theta_2)) \times \frac{E_{p_{\theta_1+\theta_2}}\left[\exp(k(x))\right]}{\sqrt{E_{p_{2\theta_1}}\left[\exp(k(x))\right] E_{p_{2\theta_2}}\left[\exp(k(x))\right]}}}$$

A note on Onicescu's informational energy and correlation coefficient in exponential families, arxiv 2003.13199

Onicescu's informational correlation & Cauchy-Schwarz divergence

Cauchy-Schwarz divergence (**projective divergence** $D(\lambda p : \lambda' q) = D(p : q), \quad \forall \lambda, \lambda' > 0.$):

$$D_{\mathrm{CS}}(p, q) := -\log \left(\frac{\int_{\mathcal{X}} p(x) q(x) \mathrm{d}\mu(x)}{\sqrt{\left(\int_{\mathcal{X}} p(x)^2 \mathrm{d}\mu(x)\right) \left(\int_{\mathcal{X}} q(x)^2 \mathrm{d}\mu(x)\right)}} \right) \geq 0.$$

Onicescu's informational correlation (affinity coefficient):

$$\rho(p, q) := \frac{I(p, q)}{\sqrt{I(p) I(q)}}, \quad I(p, q) := \int p(x) q(x) \mathrm{d}\mu(x)$$

informational energy:

$$I(p) = I(p, p)$$

$$I(p) := \int p^2(x) \mathrm{d}\mu(x)$$

Onicescu-Cauchy-Schwarz identity: $\boxed{D_{\mathrm{CS}}(p, q) = -\log\left(\rho(p, q)\right)}$

A note on Onicescu's informational energy and correlation coefficient in exponential families, 2003.13199

Origin of Fisher information (1922, 1925)

If a frequency curve is defined by

$$df = y dx$$

where y is a function of a parameter θ, then the intrinsic accuracy of the curve, as a means of estimating θ, is

$$\boxed{-\int y \frac{\partial}{\partial \theta^2} (\log y) . dx} \Rightarrow -\int p_\theta(x) \frac{d^2}{d\theta^2} \log p_\theta(x) d\mu(x) = \boxed{-E_{p_\theta}\left[\frac{d^2}{d\theta^2} \log p_\theta(x)\right]}$$

over the whole range of possible values. Since

$$y \frac{\partial^2}{\partial \theta^2} (\log y) = \frac{\partial^2 y}{\partial \theta^2} - \frac{1}{y}\left(\frac{\partial y}{\partial \theta}\right)^2,$$

while the integral of the first term over all possible values must vanish, the intrinsic accuracy may equally be written

$$\boxed{\int \frac{1}{y}\left(\frac{\partial y}{\partial \theta}\right)^2 dx,} \Rightarrow \int \frac{1}{p_\theta(x)}\left(\frac{d}{d\theta} p_\theta(x)\right)^2 d\mu(x) = \boxed{E_{p_\theta}\left[\left(\frac{d}{d\theta} \log p_\theta(x)\right)^2\right]}$$

over all possible values; in this form it is clearly seen to be necessarily positive.

the amount of information provided by a combination of two or more independent observations is thus merely the sum of the amounts of information in each piece separately.

Ronald Aylmer Fisher
(1890-1962)

Fisher, Ronald Aylmer. "Theory of statistical estimation." *Mathematical proceedings of the Cambridge philosophical society*. Vol. 22. No. 5. Cambridge University Press, 1925.

The origin of the Jensen-Shannon divergence
Why choosing the mid-mixture?

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 37, NO. 1, JANUARY 1991

Divergence Measures Based on the Shannon Entropy

[1991] Jianhua Lin, *Member, IEEE*

$$K(p_1, p_2) = I\left(p_1, \frac{1}{2}p_1 + \frac{1}{2}p_2\right). \qquad (3.2)$$

$$L(p_1, p_2) = K(p_1, p_2) + K(p_2, p_1). \qquad (3.4)$$

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

Theorem 2.2.

$$K_\alpha\begin{pmatrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{pmatrix} = K_\alpha\left(\begin{matrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{matrix}\middle| v\right)$$

for precisely one value $v = v_0$ of v, and v_0 is given by

$$q_0 = \sum w_i p_i / \sum w_i \qquad \text{if } \alpha = 1,$$

Mid-mixture

Z. Wahrscheinlichkeitstheorie verw. Geb. 14, 149–160 (1969)

[1969] Information Radius*

ROBIN SIBSON

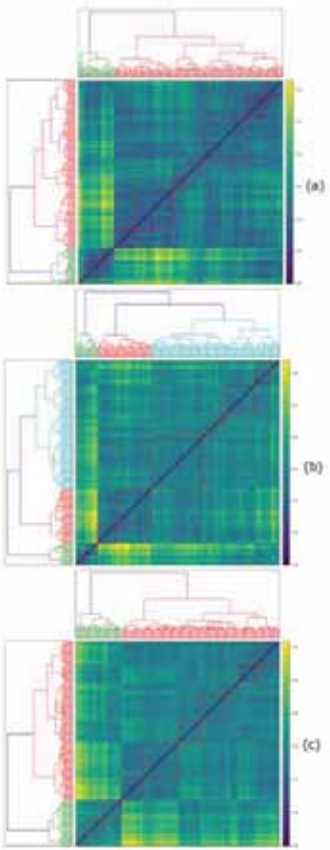
Corollary 2.3. *K_α takes the following forms:*

$$K_1\begin{pmatrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{pmatrix} = \int_X \sum_i \left\{\frac{w_i p_i}{\sum_j w_j} \log_2 \frac{p_i \sum_j w_j}{\sum_j w_j p_j}\right\} d\lambda;$$

$$D_{\mathrm{JS}}[p, q] := \min_{c \in \mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p : c] + D_{\mathrm{KL}}[q : c]\right)$$

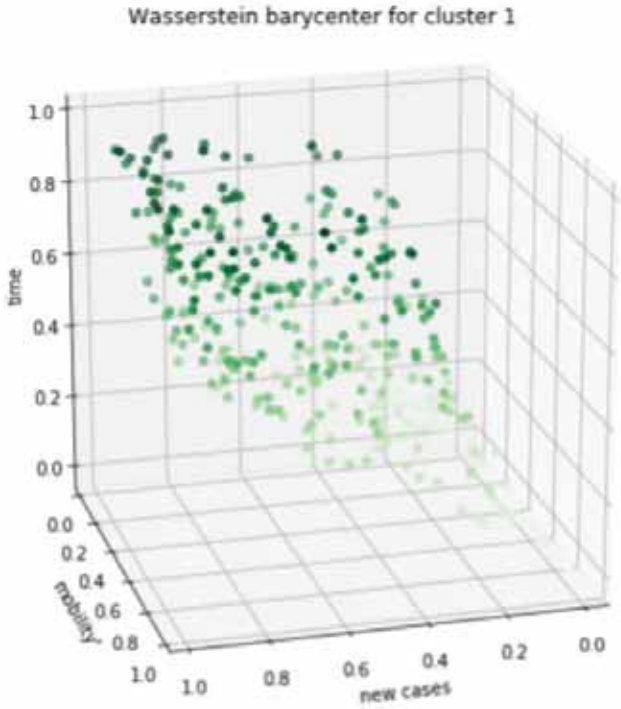
On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius, Entropy 2021

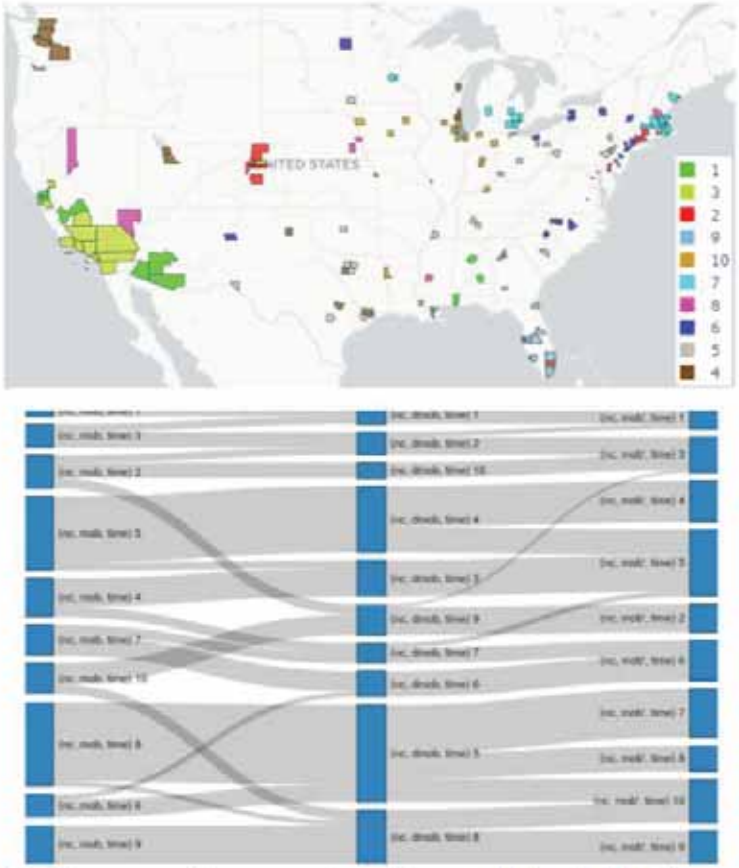
Optimal transport clustering: COVID-19 dynamics/Human mobility



$$W_p(\mu,\nu) = \left(\inf_{\lambda \in \tau(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^p d\lambda(x,y) \right)^{1/p},$$

Wasserstein Lp metric distance





HCMapper: Comparing hierarchical clusterings

Clustering patterns connecting COVID-19 dynamics and Human mobility using optimal transport
FN, Gautier Marti, Sumanta Ray, Saumyadipta Pyne **https://arxiv.org/abs/2007.10677**

Optimal Copula Transport: Clustering Time Series

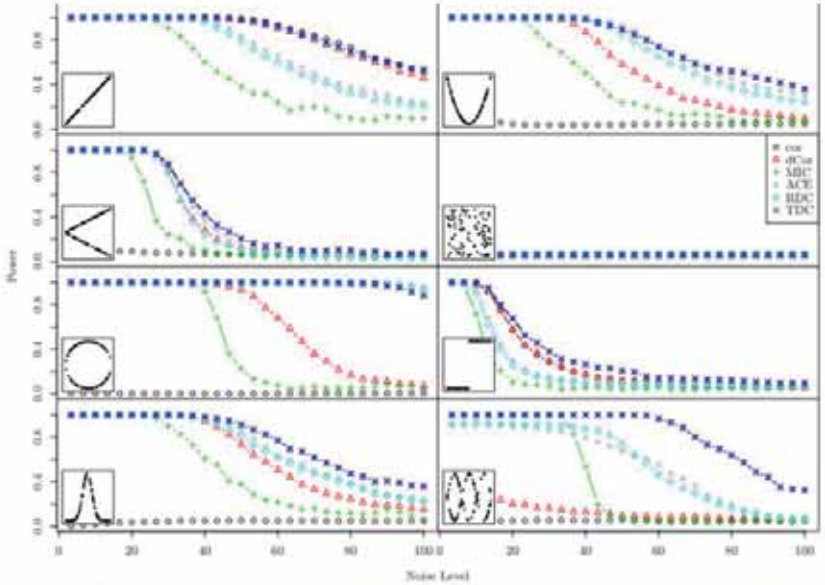
Distance between random variables (Mutual Information, similarity: correlation coefficient)

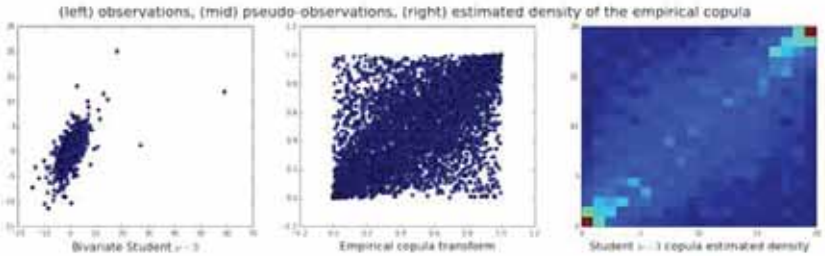
Spearman correlation ρ_S more resilient to outliers than Pearson correlation ρ

(a) $\rho = 0.66$ (b) $\rho = 0.23$ (c) $\rho_S = 0.65$ (d) $\rho_S = 0.64$

Sklar's theorem: $F(x_i, x_j) = C_{ij}(F_i(x_i), F_j(x_j))$

Copulas C = encode dependence between marginals F





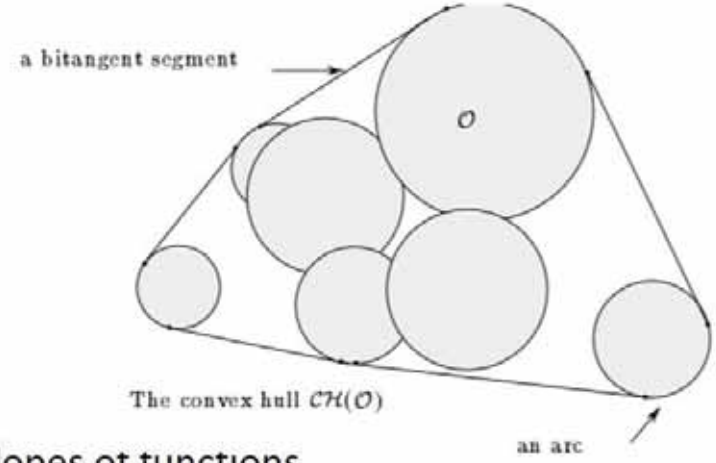
- Build the forget-dependence copulas $\{C_l^F\}_l$
- Build the target-dependence copulas $\{C_k^T\}_k$
- Compute the empirical copula C_{ij} from x_i, x_j

$$\mathrm{TDC}(C_{ij}) = \frac{\min_l \mathrm{EMD}(C_l^F, C_{ij})}{\min_l \mathrm{EMD}(C_l^F, C_{ij}) + \min_k \mathrm{EMD}(C_{ij}, C_k^T)}$$

@FrnkNlsn Optimal Copula Transport for Clustering Multivariate Time Series, ICASSP 2016 Arxiv 1509.08144

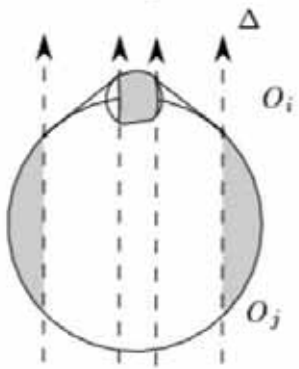
Output-sensitive convex hull construction of 2D objects

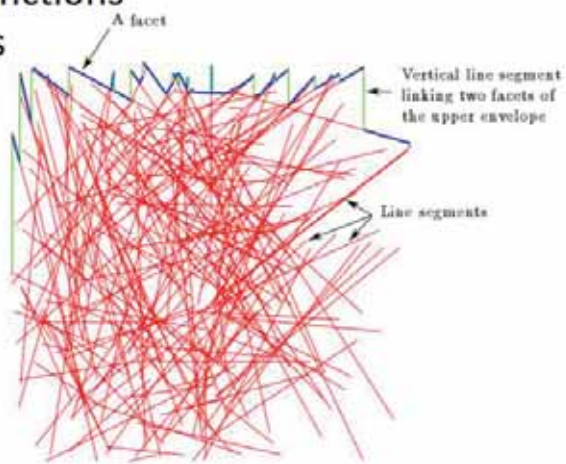
N objects, boundaries intersect pairwise in at most m points
Convex hull of disks (m=2), of ellipses (m=4), etc.
Complexity bounded using Ackermann's inverse function α



$m \backslash \lambda(n,m)$	Lower Bound Ω	Upper Bound O
1	n	n
2	$2n-1$	$2n-1$
3	$\Omega(n \times \alpha(n))$	$O(n \times \alpha(n))$
4	$\Omega(n \times 2^{\alpha(n)})$	$O(n \times 2^{\alpha(n)})$
$2s+1$	$\Omega(n \times 2^{O(\alpha(n)^{s-1})})$	$O(n \times \alpha(n)^{O(\alpha(n)^{s-1})})$
$2s+2$	$\Omega(n \times 2^{O(\alpha(n)^{s})})$	$O(n \times 2^{O(\alpha(n)^{s})})$

Extend to upper envelopes of functions pairwise intersecting in m points

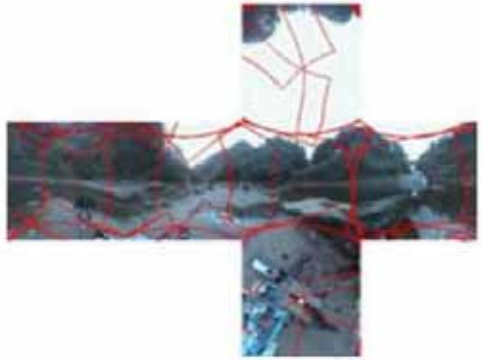




Output-Sensitive Convex Hull Algorithms of Planar Convex Objects, IJCGA (1998)

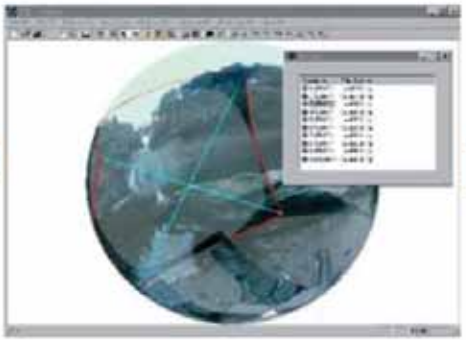
Representing spherical panoramic images

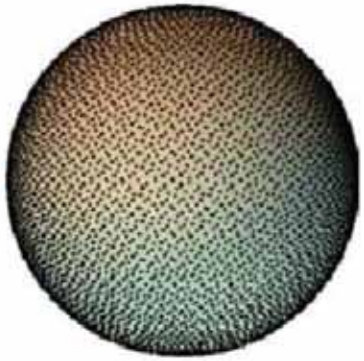
Called **environment maps** in Computer Graphics

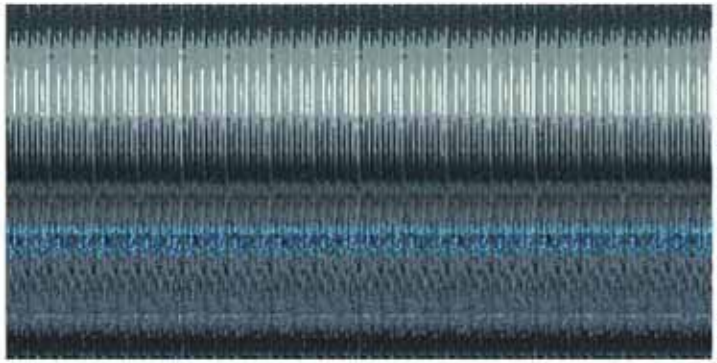


Usual maps: latitude-longitude, cube, front/back double paraboloids, etc.

Best quality map related to incremental sampling on the sphere: The **Hammersley map**





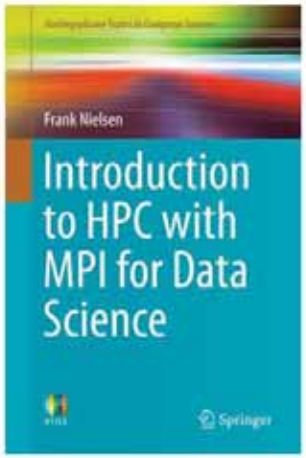


1D map stored in 2D image

On Representing Spherical Videos, IEEE CVPR worshop, 2001.

Surround video: a multihead camera approach, The visual computer 2005.

Parallel Algorithms for Graphs $G = (V, E)$



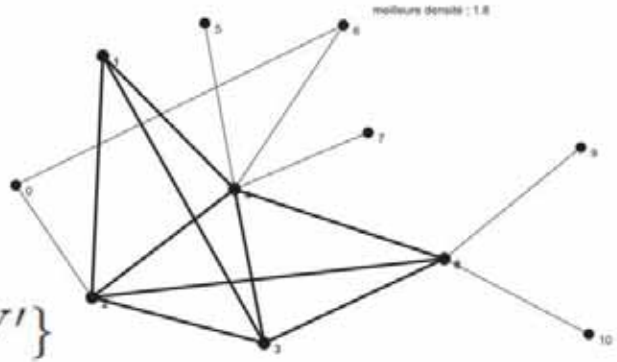
Finding densest subgraphs in graphs: maximize the graph density

$$\boxed{\rho^* = \max_{V' \subseteq V} \rho(V')}$$

where

$$\rho(V') = \frac{|E(V')|}{|V'|}$$

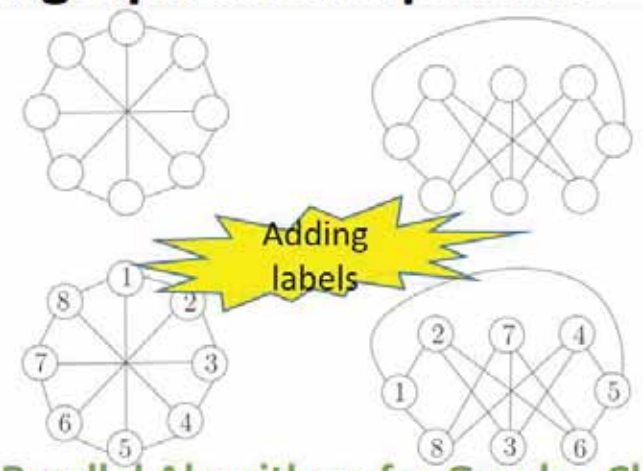
$$E(V') = \{(u, v) \in E \mid (u, v) \in V' \times V'\}$$



$$\rho(G_{|V^*}) = 1.8$$

Greedy parallel heuristic

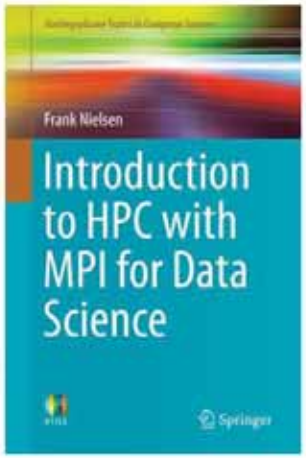
Subgraph isomorphisms:



Parallelizing enumerating algorithms (load balancing)

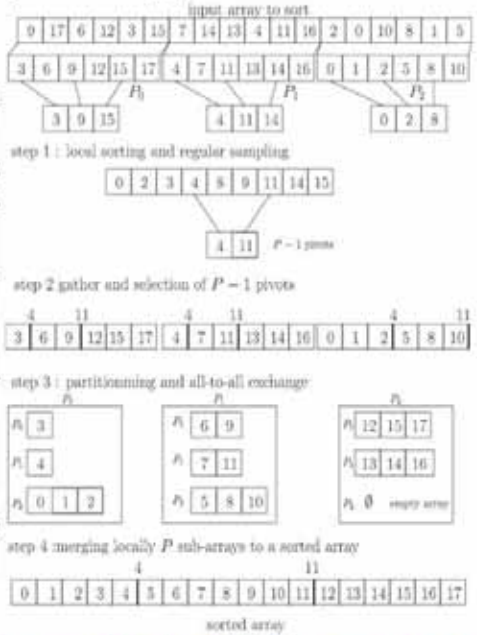
Parallel Algorithms for Graphs, Chapter 11 of "Introduction to HPC with MPI for Data Science", Springer 2016

Parallel sorting on a cluster of machines

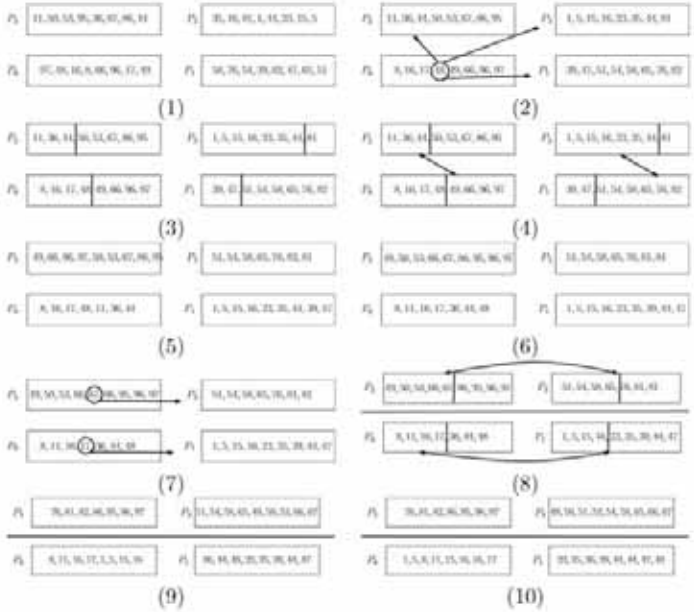


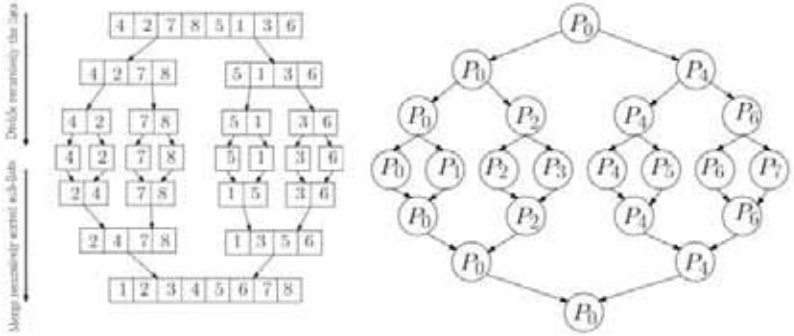
- **Logical topology** (eg., 2d torus) versus **physical topology** (eg., hypercube) of interconnection network of machines
- Send/receive data packets using **the Message Passing Interface** (MPI)
- Parallel merge sort, HyperQuicksort, Parallel Sort Regular Sampling (PSRS), bitonic merge sort, shearsort, odd-even sort: **Load balancing**

Parallel sort regular sampling



Parallel quicksort





Parallel merge sort,

Parallel sorting, Chapter 4 of "Introduction to HPC with MPI for Data Science", Springer 2016

Perspective functions: f-divergences and Fisher information

- **Perspective function** of a convex function $f : H \to (-\infty, +\infty]$ defined on a real Hilbert space: $P_f : \mathbb{R} \times H \to (-\infty, +\infty]$

$$(p, q) \mapsto P_f(p, q) = \begin{cases} pf\left(\frac{q}{p}\right) & \text{if } p > 0, \\ +\infty & \text{otherwise} \end{cases}$$

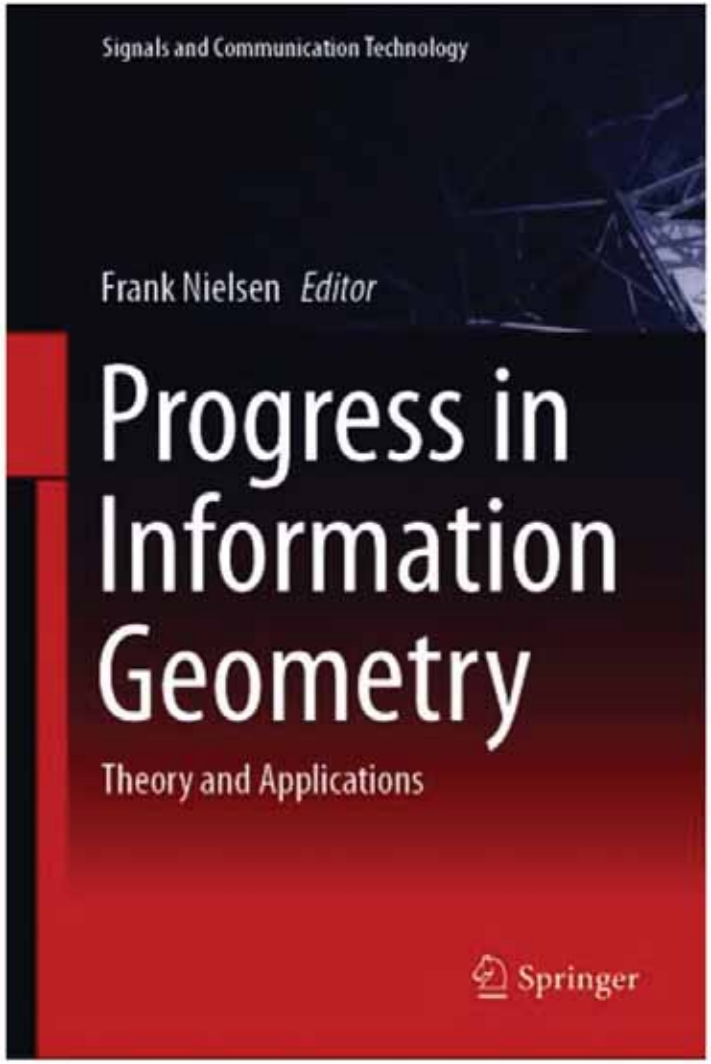
- P_f is convex if and only if *f* is convex
- Perspective function used to define **f-divergences**:

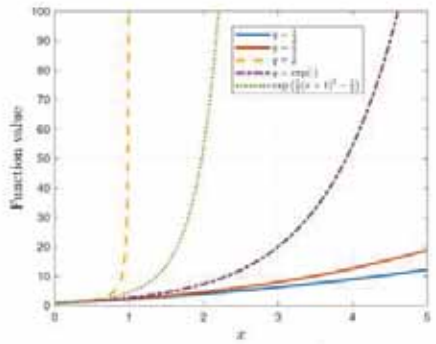
$$I_f[p : q] = \int_{\mathcal{X}} P_f(p(x), q(x)) \mathrm{d}\mu(x) = \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$$

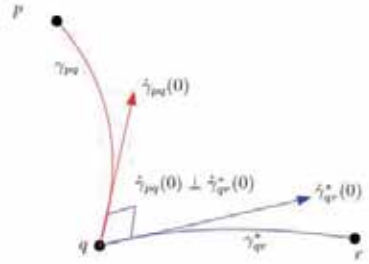
- Perspective function used to define **non-parametric Fisher information**:

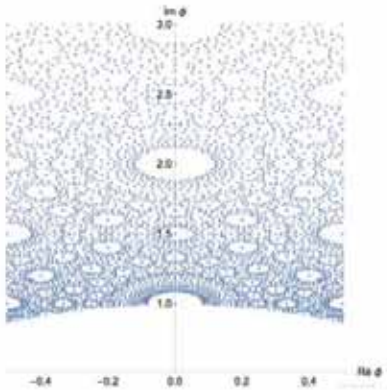
$$I(p) = \int_{\mathbb{R}^d} P_{\|\cdot\|^2}(p(x), \nabla p(x)) \mathrm{d}x \mathrm{d}x = \int_{\mathbb{R}^d} \frac{\|\nabla p(x)\|^2}{p(x)} \mathrm{d}x = 4 \int_{\mathbb{R}^d} \|\nabla \sqrt{p(x)}\|^2 \mathrm{d}x$$

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









Hessian Information Geometry | Shape Space | Monotone Embedding

Stochastic Geometric Mechanics | Riemannian Manifolds | Geometric Deep Learning

Lie Group Thermodynamics | Probability Density Estimation | Geometry of quantum states

1 **Information Geometry of Smooth Densities on the Gaussian Space: Poincaré Inequalities** 1
Giovanni Pistone

2 **On Normalization Functions and φ-Families of Probability Distributions** 19
Luiza H. F. de Andrade, Francisca L. J. Vieira, and Charles C. Cavalcante

3 **Affine Connections with Torsion in (Para-)complexified Structures** 37
Jun Zhang and Gabriel Khan

4 **Contact Hamiltonian Systems for Probability Distribution Functions and Expectation Variables: A Study Based on a Class of Master Equations** 57
Shin-itiro Goto and Hideitsu Hino

5 **Invariant Koszul Form of Homogeneous Bounded Domains and Information Geometry Structures** 89
Frédéric Barbaresco

6 **Gauge Freedom of Entropies on q-Gaussian Measures** 127
Hiroshi Matsuzoe and Asuka Takatsu

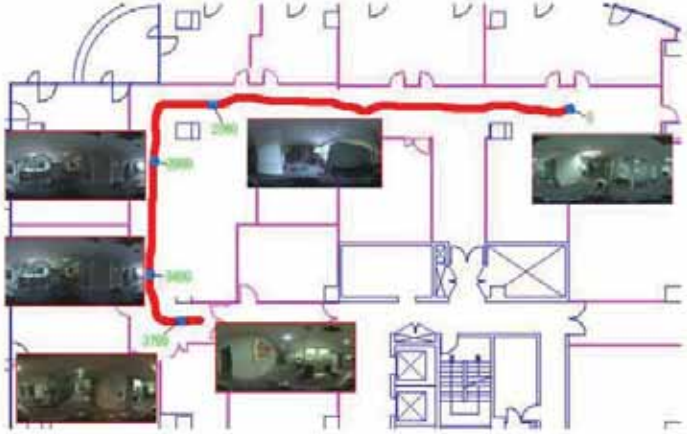
7 **On Geodesic Triangles with Right Angles in a Dually Flat Space** ... 153
Frank Nielsen

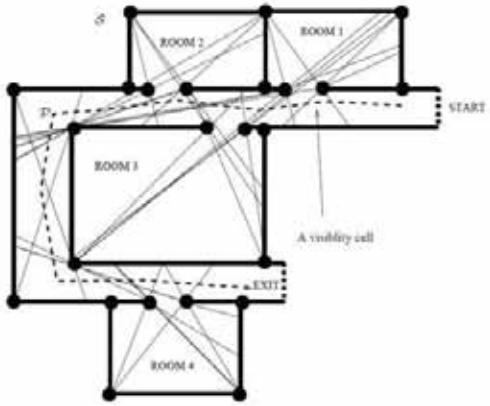
8 **Chain Rule Optimal Transport** 191
Frank Nielsen and Ke Sun

9 **Towards the "Shape" of Cosmological Observables and the String Theory Landscape with Topological Data Analysis** 219
Alex Cole and Gary Shiu

10 **A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets** 245
Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, and Philippe Donnat

Plenoptic camera for VR image-based navigation





Plenoptic path and its applications, ICIP 2003

EF KLD as log density-ratio: Poisson example

$$\boxed{D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \log\left(\frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)}\right), \quad t(\omega) = F'(\theta(\lambda_1)) \in \overline{\mathrm{CH}(\mathcal{X})}}$$

- Poisson PMFs $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $\mathcal{X} = \{0, 1, \ldots\}$.
- $\theta = \log \lambda$, $F(\theta) = e^\theta$ and $\eta = F'(\theta) = e^\theta = \lambda$.
- rewrite density $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{\Gamma(x+1)}$ (with $\Gamma(x+1) = x!$): $p_\lambda(x)$ is defined for $x \in \mathbb{R}_+$
- Apply theorem for $\omega = \lambda_1 \notin \mathcal{X}$ but in $\overline{\mathrm{CH}(\mathcal{X})} = [0, \infty]$:

$$\begin{aligned}
D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] &= \log \frac{p_{\lambda_1}(\lambda_1)}{p_{\lambda_2}(\lambda_1)}, \\
&= \log \frac{\frac{\lambda_1^{\lambda_1} \exp(-\lambda_1)}{\Gamma(\lambda_1+1)}}{\frac{\lambda_2^{\lambda_1} \exp(-\lambda_2)}{\Gamma(\lambda_1+1)}}, \\
&= \lambda_1 \log \frac{\lambda_1}{\lambda_2} + \lambda_2 - \lambda_1.
\end{aligned}$$

Polarization identity: Reconstructing the inner product

ANNALS OF MATHEMATICS
Vol. 36, No. 3, July, 1935

ON INNER PRODUCTS IN LINEAR, METRIC SPACES

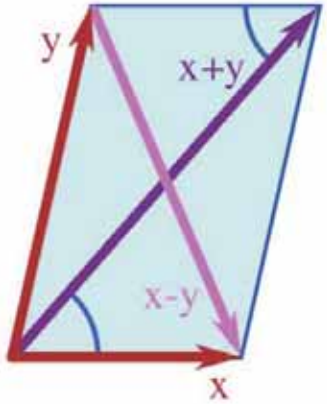
BY P. JORDAN AND J. V. NEUMANN

Ernst Pascual Jordan
(1902-1980)

John von Neumann
(1903-1957)

THEOREM I. *Let L be a generalized linear, metric space, with the operations $\alpha \cdot f$, $f + g$, $\| f \|$. In order that it be possible to define an operation (f, g) in L, so that L with the original $\alpha \cdot f$, $f + g$ and this (f, g) is a linear space with an inner product, and that the original $\| f \|$ is the metric derived from this (f, g), the following condition is necessary and sufficient:*

(*) $$\| f + g \|^2 + \| f - g \|^2 = 2 \left(\| f \|^2 + \| g \|^2 \right) \text{ for all } f, g \in L.^6$$



When the parallelogram law holds for the norm, there exists an inner product that induces it:

Polarization identity

$$\langle x, y \rangle = \frac{1}{4} \left(\|x + y\|^2 - \|x - y\|^2 \right) \ \forall\, x, y \in V\,.$$

Jordan and von Neumann, "On inner products in linear, metric spaces." Annals of Mathematics (1935): 719-723.

Power chi expansions of f-divergences[2]

- Smooth convex $f(u)$ with $f(1) = f'(1) = 0$, strictly convex at 1, f-**divergence** for $p, q \ll \mu$: $I_f(p:q) := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$
- **Power chi pseudo-distances**: $\chi_i^{\pm}(p:q) = \int \frac{(q(x)-p(x))^i}{p(x)^{i-1}} \mathrm{d}\mu(x)$
- **Order-k power chi expansion of f-divergences** ("chinomial"):

$$I_{f,k}^{\chi}(p:q) := \sum_{i=2}^{k} \frac{f^{(i)}(1)}{i!} \chi_i^{\pm}(p:q)$$

I_f name	generator f	k-order power chi expansion
generic I_f	generic convex generator f	$I_{f,k}^{\chi}(p:q) = \sum_{i=2}^{k} \frac{f^{(i)}(1)}{i!} \chi_i^{\pm}(p:q)$
polynomial	$f_P(u) = \sum_{j=0}^{d} a_j u^j$	$I_{f_P,k}^{\chi}(p:q) = \sum_{i=2}^{d} \left(\sum_{j=i}^{d} a_j \binom{j}{i}\right) \chi_i^{\pm}(p:q)$
KL	$f_{\mathrm{KL}}(u) = -\log u$	$\mathrm{KL}_k^{\chi}(p:q) = \sum_{i=2}^{k} \frac{(-1)^i}{i} \chi_i^{\pm}(p:q)$
reverse KL	$f_{\mathrm{KL}^r}(u) = u \log u$	$\mathrm{KL}^r{}_k^{\chi}(p:q) = \sum_{i=2}^{k} (-1)^i \frac{1}{i(i-1)} \chi_i^{\pm}(p:q)$
α-divergence	$f_\alpha(u) = \frac{4}{1-\alpha^2}(1 - u^{\frac{1+\alpha}{2}})$	$I_{\alpha,k}^{\chi}(p:q) = \sum_{i=2}^{k} -\frac{2}{1-\alpha} \frac{1}{i!} \binom{\frac{1+\alpha}{2}}{i} \chi_i^{\pm}(p:q)$
Jeffreys	$f_J(u) = (u-1)\log u$	$J_k^{\chi}(p:q) = \sum_{i=2}^{k} (-1)^i \frac{1}{i-1} \chi_i^{\pm}(p:q)$
Jensen-Shannon	$f_{\mathrm{JS}}(u) = -(u+1)\log\frac{1+u}{2} + u\log u$	$\mathrm{JS}_k^{\chi}(p:q) = \sum_{i=2}^{k} \frac{(-1)^{i-2}}{i(i-1)} \left(1 - \frac{1}{2^{i-1}}\right) \chi_i^{\pm}(p:q)$
Harmonic	$f_H(u) = \frac{2u}{u+1}$	$H_k^{\chi}(p:q) = 1 + \sum_{i=2}^{k} (-1)^{i+1} \frac{1}{2^i} \chi_i^{\pm}(p:q)$
Exponential	$f_{\exp}(u) = e^x - x$	$E_k^{\chi}(p:q) = \sum_{i=2}^{k} \frac{e}{i!} \chi_i^{\pm}(p:q)$

[2] https://arxiv.org/abs/1903.05818

Powered Minkowski metric distances

1864-1909
(44 years old)

$$M_\alpha(x,y) := \|x-y\|_\alpha = \left(\sum_{i=1}^{d} |x_i - y_i|^\alpha\right)^{\frac{1}{\alpha}}$$

is a well-known **metric distance** for $\alpha \geq 1$

$$M_\alpha^\alpha(x,y) = \sum_{i=1}^{d} |x_i - y_i|^\alpha$$ is a **metric distance** for $\alpha \in (0,1]$

Since $|a+b|^\alpha \leq |a|^\alpha + |b|^\alpha, \alpha \in (0,1]$, we get the triangle inequality:

$$\begin{aligned} |(a-c)+(c-b)|^\alpha &\leq |a-c|^\alpha + |b-c|^\alpha \\ M_\alpha^\alpha(a,b) &\leq M_\alpha^\alpha(a,c) + M_\alpha^\alpha(b,c) \end{aligned}$$

x^β for $0 < \beta < 1$

Metric transform

M_α^β is a **metric distance** for $0 < \alpha \leq 1,\ 0 < \beta \leq \alpha$

The statistical Minkowski distances: Closed-form formula for Gaussian Mixture Models, Springer LNCS GSI 2019, https://arxiv.org/abs/1901.03732

Power r-Jensen-Shannon divergences

- **Statistical mixture wrt an abstract weighted mean** M_α between two mutually absolutely continuous distributions p(x) and q(x):

$$(pq)_{M_\alpha}(x) := \frac{M_\alpha(p(x), q(x))}{\underbrace{\int M_\alpha(p(x), q(x))\mathrm{d}\mu(x)}_{:=Z_\alpha[p:q]}}$$

- **Skewed Jensen-Shannon divergence wrt to two abstract means** M_α and N_β

$$\boxed{D_{\mathrm{JS}}^{M_\alpha, N_\beta}[p:q] := N_\beta\left(D_{\mathrm{KL}}\left[p:(pq)_{M_\alpha}\right], D_{\mathrm{KL}}\left[q:(pq)_{M_\alpha}\right]\right)}$$

- Ordinary Jensen-Shannon divergence wrt arithmetic mean: $A(x,y) = A_{\frac{1}{2}}(x,y) = \frac{x+y}{2}$

$$\Rightarrow \quad D_{\mathrm{JS}}[p:q] = A(D_{\mathrm{KL}}[p:(pq)^A], D_{\mathrm{KL}}[q:(pq)^A]) \qquad A_\gamma(x,y) = (1-\gamma)x + \gamma y$$

- **Skewed power Jensen-Shannon divergences wrt weighted power means**:

$$\boxed{D_{\mathrm{JS}}^{P_{r,\alpha},\beta}[p:q] := D_{\mathrm{JS}}^{P_{r,\alpha},A_\beta}[p:q] = (1-\beta)D_{\mathrm{KL}}\left[p:(pq)_{P_{r,\alpha}}\right] + \beta D_{\mathrm{KL}}\left[q:(pq)_{P_{r,\alpha}}\right]}$$

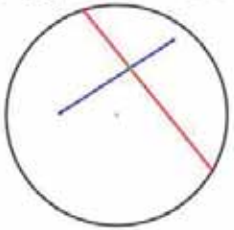
Weighted power mean: $$P_{r,\alpha}(x,y) = \begin{cases} ((1-\alpha)x^r + \alpha y^r)^{\frac{1}{r}}, & r \neq 0 \\ x^{1-\alpha}y^\alpha, & r = 0. \end{cases}$$

On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. Entropy (2019)

Hyperbolic pregeodesics & geodesics in Klein model

- Distance in the d-dimensional **Klein unit disk model** of hyperbolic geometry:

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



- **Geodesics** in Klein geodesic metric space (constant speed, arclength paameterization):

$$\gamma(p,q;\alpha) = (1-c(\alpha))p + c(\alpha)q \qquad d_K(\gamma(p,q;s),\gamma(p,q;t)) = |s-t|\; d_K(p,q)$$

- **Pregeodesics** are line segments: $\Gamma(p,q) := \{(1-\alpha)p + \alpha q \;:\; \alpha \in [0,1]\}$
- Solution:

$$\boxed{c(\alpha) = \frac{ad(\alpha)\sqrt{(ac+b^2)(d(\alpha)^2-1)} + ab(1-d(\alpha)^2)}{acd(\alpha)^2 + b^2}}$$

$$\begin{aligned} a &= 1 - p^\top p, \\ b &= p^\top(q-p), \\ c &= (q-p)^\top(q-p), \\ d(\alpha) &= \cosh(\alpha d_K(p,q)) \end{aligned}$$

Pattern recognition on statistical manifolds

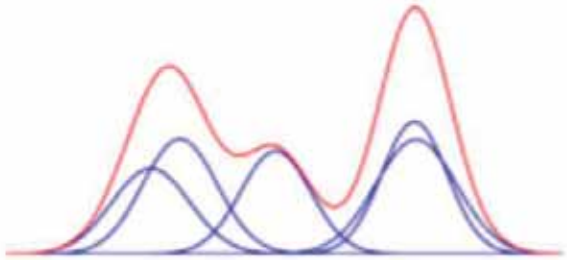
	Multinomial ($n = 1$ trial)	Multivariate Gaussian
density	$\frac{1}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$	$\frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right)$
support	$\{E_1, ..., E_k\}$	$\mathbb{R}^d$
base measure	Counting measure	Lebesgue measure
auxiliary carrier $k(x)$	$-\sum_{i=1}^k \log x_i!$	0
sufficient statistics $t(x)$	$(x_1, \cdots, x_{k-1})$	$(x, -xx^\top)$
$\theta(\lambda)$	$\left(\log\left(\frac{p_i}{p_k}\right)\right)_i$	$(\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$
Order D	$d-1$	$\frac{d}{d+3}2$
log-normalizer $F(\theta)$	$\log\left(1 + \sum_{i=1}^{k-1} \exp\theta_i\right)$	$\frac{1}{4}\mathrm{tr}(\theta_2^{-1}\theta_1\theta_1^\top) - \frac{1}{2}\log\lvert\theta_2\rvert + \frac{d}{2}\log\pi$
$\eta = \nabla F(\theta) = E[t(X)]$	$\left(\frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j}\right)_i$	$\left(\frac{1}{2}\theta_2^{-1}\theta_1, -\frac{1}{2}\theta_2^{-1} - \frac{1}{4}(\theta_2^{-1}\theta_1)(\theta_2^{-1}\theta_1)^\top\right)$
$\theta = \nabla F^*(\eta)$	$\left(\log\left(\frac{\eta_i}{1-\sum_{j=1}^{k-1}\eta_j}\right)\right)_i$	$\left(-(\eta_2 + \eta_1\eta_1^\top)^{-1}\eta, -\frac{1}{2}(\eta_2 + \eta_1\eta_1^\top)^{-1}\right)$
$F^*(\eta) = \langle\theta, \eta\rangle - F(\theta)$	$\left(\sum_{i=1}^{k-1}\eta_i\log\eta_i\right) + \left(1 - \sum_{i=1}^{k-1}\eta_i\right)\log\left(1 - \sum_{i=1}^{k-1}\eta_i\right)$	$-\frac{1}{2}\log\left(1 + \eta_1^\top\eta_2^{-1}\eta_1\right) - \frac{1}{2}\log\lvert-\eta_2\rvert - \frac{d}{2}\log(2\pi e)$
Kullback-Leibler divergence	$p_{1,k}\log\frac{p_{1,k}}{p_{2,k}} - \sum_{i=1}^{k-1} p_{1,i}\log\frac{p_{2,i}}{p_{1,i}}$	$\frac{1}{2}\left(\log\left(\frac{\lvert\Sigma_2\rvert}{\lvert\Sigma_1\rvert}\right) + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right)\right) + \frac{1}{2}\left((\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - d\right)$
Fisher information $I(\theta) = \nabla^2 F(\theta)$	$I_{ii} = \frac{1}{p_i} + \frac{1}{p_d}, \quad I_{ij} = \frac{1}{p_d}(i \neq j)$	$I_{ij} = \partial_i\mu^\top\Sigma^{-1}\partial_j\mu$
MLE $\hat{\eta} = \bar{t}$	$\hat{p}_i = \frac{n_i}{n}$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})\top$
$\lambda(\theta)$	$\begin{cases} p_i = \frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j} & \text{if } i < k \\ p_k = \frac{1}{1+\sum_{j=1}^{k-1}\exp\theta_j} \end{cases}$	$\left(\frac{1}{2}\theta_2^{-1}\theta, \frac{1}{2}\theta_2^{-1}\right)$
$\lambda(\eta)$	$\begin{cases} p_i = \eta_i & \text{if } i < k \\ p_k = 1 - \sum_{j=1}^{k-1}\eta_j \end{cases}$	$(\eta, -(\eta_2 + \eta\eta^\top))$

Pattern learning and recognition on statistical manifolds: an information-geometric review, SIMBAD, Springer 2013

pyMEF/jMEF: Libraries for statistical mixtures

Implement
Gaussian Mixture Models (GMMs)
Bernoulli Mixture Models (BMMs)
Rayleigh Mixture Models (RMMs)
Wishart Mixture Models (WMMs)
And **any mixtures of an exponential family**!

```
class MyEF(ExponentialFamily):
    ...
    def lambda2theta(self, l): ...
    def theta2lambda(self, t): ...
    def eta2lambda(self, e): ...
    def lambda2eta(self, l): ...
    def t(self, x): ...
    def k(self, x): ...
    def F(self, t): ...
    def gradF(self, t): ...
    def G(self, e): ... # F*
    def gradG(self, e): ... # ∇F*
    def density(self, x): ...
```



Java

http://vincentfpgarcia.github.io/jMEF/

http://www-connex.lip6.fr/~schwander/pyMEF/

@FrnkNlsn

PyMEF: A framework for exponential families in Python, IEEE SSP 2011.

Quasi-arithmetic means and divergences

- **Quasi-arithmetic means** for a strictly increasing smooth function g(u):

$$ {}^{g}\Sigma_{i=1}^{n}(w_i, x_i) := g^{-1}\left(\Sigma_{i=1}^{n} w_i g(x_i)\right) $$

$$ {}^{g}\Sigma(x_1, \ldots, x_n; w_1, \ldots, w_n) := g^{-1}\left(\Sigma_{i=1}^{n} w_i g(x_i)\right) $$

- Quasi-arithmetic means include the Pythagorean means (arithmetic, geometric and harmonic means) and the Holder power means

Example: Geometric mean $G(a,b) = \sqrt{ab}$ $\xrightarrow{g_G(u) = \log u}$ $G(a,b) = {}^{g_G}\Sigma(a, b; \frac{1}{2}, \frac{1}{2})$

- Quasi-arithmetic means used to defined novel **Bregman divergences**:

$$ J_{F,\alpha}(p:q) := (1-\alpha)F(p) + \alpha F(q) - F((1-\alpha)p + \alpha q) $$

$$ B_F(p:q) = \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(p:q) $$

$$ M_\rho(a,b) = {}^{\rho}\sum(a, b; \frac{1}{2}, \frac{1}{2}) $$

$$ M_\tau(a,b) = {}^{\tau}\sum(a, b; \frac{1}{2}, \frac{1}{2}) $$

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

Generalizing skew Jensen divergences and Bregman divergences with comparative convexity, IEEE SPL 2017

A generalization of the α-divergences based on comparable and distinct weighted means, arxiv 2001.09660

q-derivatives: Tsallis q-entropy and q-neurons in DNNs

Consider the function: $$E_\alpha(p) := \sum_{i=1}^{d} p_i^\alpha$$

Recover **Shannon entropy** for the **derivative**:

$$H(p) := -\left.\frac{dE_\alpha(p)}{d\alpha}\right|_{\alpha=1} = -\sum_{i=1}^{d} p_i \log p_i$$

Consider Jackson q-derivatives:

$$\frac{d_q f(x)}{d_q x} := \frac{f(qx) - f(x)}{qx - x}$$

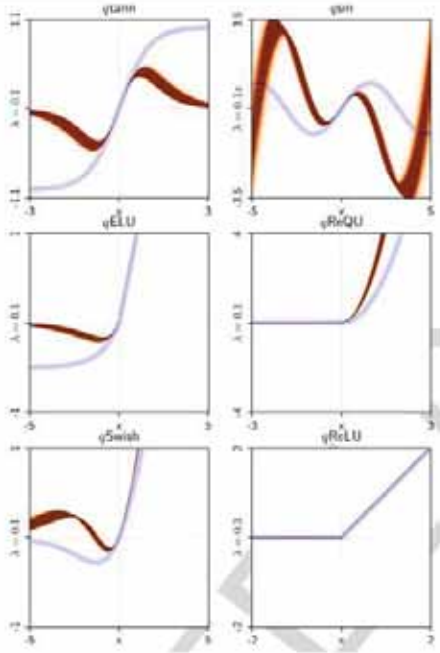
Recover **Tsallis entropy** for the **q-derivative**:

$$H_{\text{Tsallis}}(p) := -\left.\frac{d_q E_\alpha(p)}{d\alpha}\right|_{\alpha=1} = \frac{1 - \sum_{i=1}^{d} p_i^q}{q-1}$$

q-derivatives in neural networks

For an activation function f, build the **stochastic q-activation:**

$$g_q(x) := \frac{f(x) - f(qx)}{1-q} = \left(D_q f(x)\right)x$$



F. N., K. Sun, "q-Neurons: Neuron Activations Based on Stochastic Jackson's Derivative Operators"
IEEE Transactions on Neural Networks and Learning Systems.

Computing statistical divergence with sigma points

arXiv: 2003.02469

- Idea: Bypass the definite integral calculations of statistical divergences (e.g. Kullback-Leibler divergence) by replacing them by **exact finite discrete sums:**

$$D[p_{\theta_1} : p_{\theta_2}] := \int_{\mathcal{X}} d(p_{\theta_1}(x) : p_{\theta_2}(x)) \mathrm{d}\mu(x) = \frac{1}{s} \sum_{i=1}^{s} h(p_{\theta_1}(\omega_i), p_{\theta_2}(\omega_i))$$

Replace ∫ by Σ

- The **points ω_i** 's at which the function h is evaluated are called **sigma points**
- Example 1: **Kullback-Leibler divergence** between two densities of an exponential family of order D can be calculated using at most D+1 sigma points:

$$D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s} \sum_{i=1}^{s} \log \left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)} \right), \quad \frac{1}{s} \sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]$$

- Example 2: **q-divergences** between two densities of a q-exponential family, here the **q-Gaussians/q-normals** can be calculated using *two sigma points*:

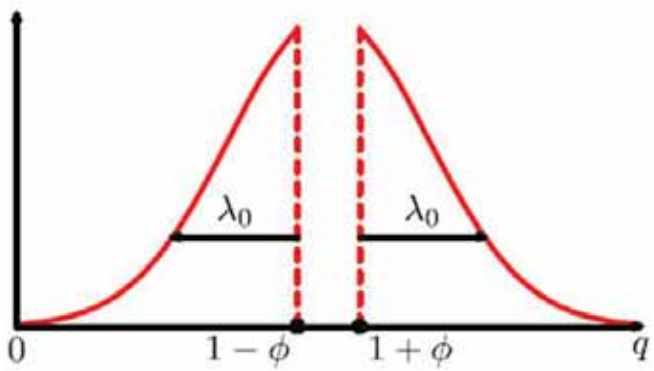
$$D_q[p_{\mu_1,\sigma_1} : p_{\mu_2,\sigma_2}] = \frac{1}{2} \left(\log_q p_{\lambda_1}(\mu_1 - \sigma_1) - \log_q p_{\lambda_2}(\mu_1 - \sigma_1) + \log_q p_{\lambda_1}(\mu_1 + \sigma_1) - \log_q p_{\lambda_2}(\mu_1 + \sigma_1) \right)$$

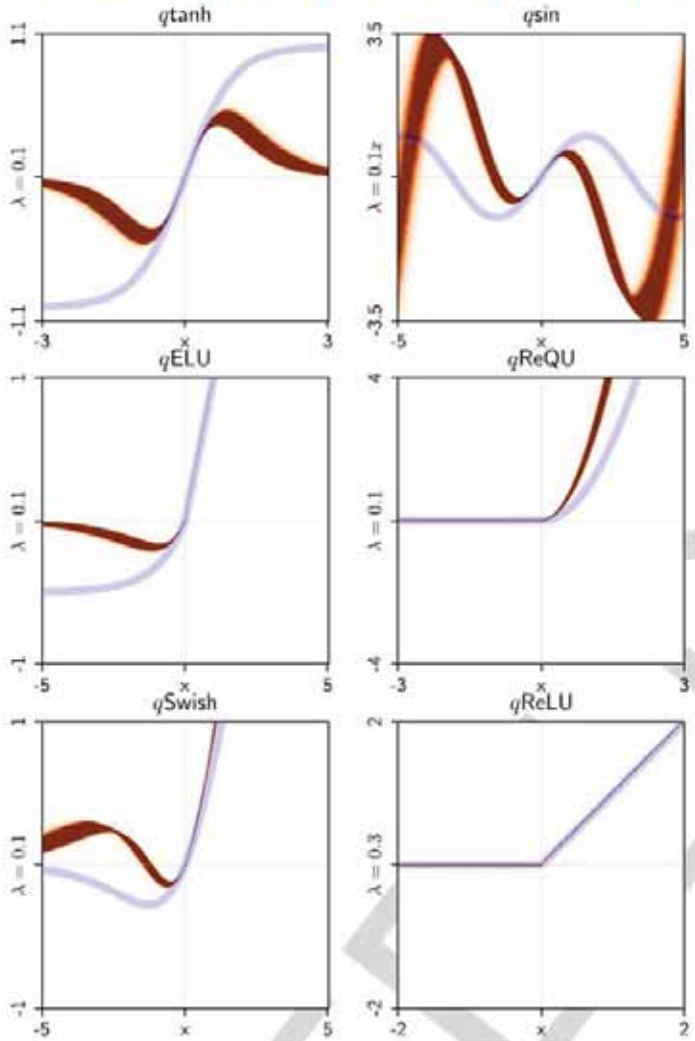
Q-neurons: Noise injection in DNNs via stochastic activation functions

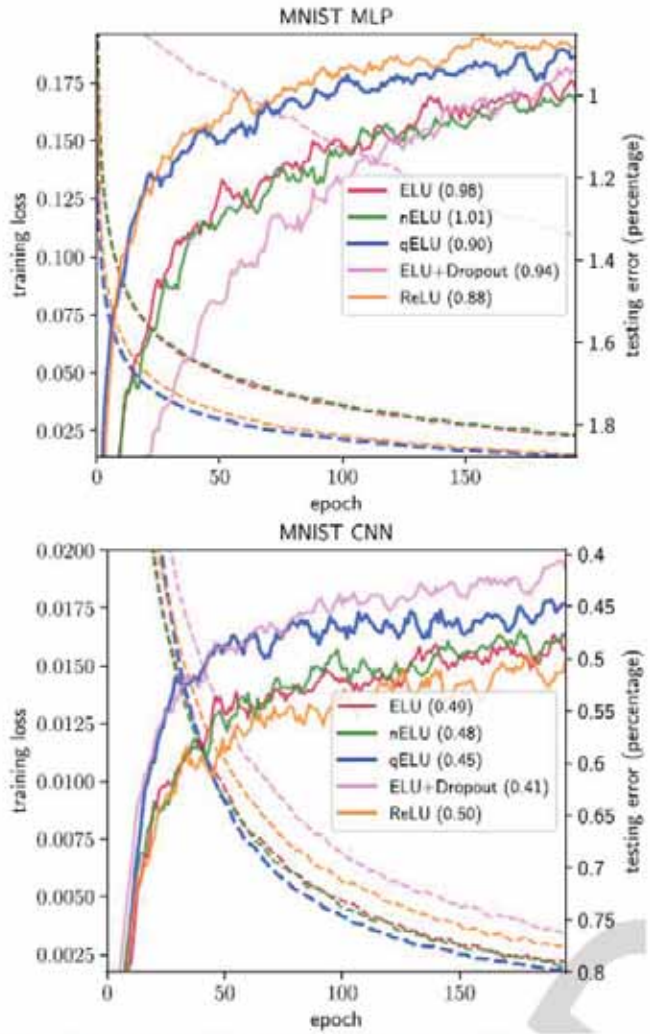
For an activation function f,
build the **stochastic q-activation:**

$$g_q(x) := \frac{f(x) - f(qx)}{1-q} = \left(D_q f(x)\right)x$$

q is a stochastic parameter:







FN and Ke Sun, "q-Neurons: Neuron Activations Based on Stochastic Jackson's Derivative Operators," *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2020.3005167.

Quasiconvex Bregman divergences[1]

- **Quasiconvex Jensen divergence** for $\alpha \in (0,1)$ and a strictly quasiconvex generator Q:

$$ {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta : \theta') \quad := \quad \max\{Q(\theta), Q(\theta')\} - Q((1-\alpha)\theta + \alpha\theta')). $$

- **Quasiconvex Bregman divergence** (only a pseudo-divergence at countably many inflection points of Q):

$$ \begin{aligned} {}^{\mathrm{qcvx}}\mathrm{B}_Q(\theta : \theta') \quad &:= \quad \lim_{\alpha \to 1} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta : \theta'), \\ &= \quad \begin{cases} -(\theta - \theta')^{\top} \nabla Q(\theta') & \text{if } Q(\theta) \leqslant Q(\theta') \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} $$

- δ-averaged quasiconvex Bregman divergence for $\delta > 0$:

$$ {}^{\mathrm{qcvx}}\mathrm{B}_Q^{\delta}(\theta, \theta') = (\theta' - \theta) \left(\frac{Q(\theta' + \delta) - Q(\theta')}{\delta} \right) $$

- Related to the **Kullback-Leibler divergence** between parametric densities with **nested supports**.

[1] https://arxiv.org/abs/1909.08857

Quasiconvex Jensen and Bregman divergences

- **Quasiconvex Jensen divergence** for a generator Q and α in (0,1):

$$^{\mathrm{qcvx}}J_Q^{\alpha}(\theta:\theta') \quad := \quad \max\{Q(\theta), Q(\theta')\} - Q((1-\alpha)\theta + \alpha\theta')).$$

Quasiconvex Bregman pseudodivergence (not separable divergence):

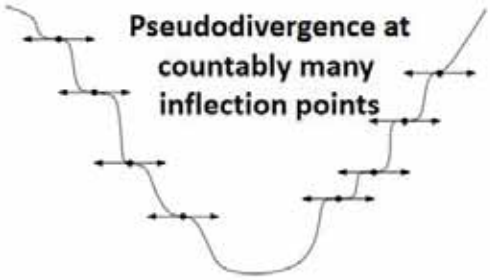
Quasiconvex functions

$$^{\mathrm{qcvx}}\mathrm{B}_Q(\theta:\theta') \quad := \quad \lim_{\alpha\to 1} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta:\theta'),$$

$$= \begin{cases} -(\theta-\theta')^\top \nabla Q(\theta') & \text{if } Q(\theta) \leqslant Q(\theta') \\ +\infty & \text{otherwise.} \end{cases}$$

- **δ-quasiconvex Bregman divergence** for δ>0:

$$^{\mathrm{qcvx}}B_Q^{\delta}(\theta,\theta') := \begin{cases} \frac{1}{\delta}\left(Q\left(\theta' + \delta(\theta'-\theta)\right) - Q(\theta')\right) & \textit{if } Q(\theta') \geq Q(\theta) \\ +\infty & \textit{otherwise} \end{cases}$$



- Qcvx Bregman pseudodivergence related to **Kullback-Leibler divergence** for distributions with **nested supports**

$$\mathrm{KL}[p:q] = \begin{cases} \int_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x), & \mathrm{supp}(p)\subset\mathrm{supp}(q), \\ +\infty, & \mathrm{supp}(p)\not\subset\mathrm{supp}(q). \end{cases}$$

A note on the quasiconvex Jensen divergences and the quasiconvex Bregman divergences derived thereof, arXiv 1909.08857

Exponential family: Raw moments and MGF derivatives

- Uniorder exponential family (eg., exponential distributions, Rayleigh, Poisson, etc.): $p_\theta(x) = \exp(\theta t(x) - F(\theta) + k(x))$
- **Moment generating function** is: $m_\theta(u) = E_{p_\theta}[e^{uX}] = \exp(F(\theta + u) - F(\theta))$
- **Raw moment** (or geometric/power moment) from higher order derivatives:

$$\boxed{\mu_l(\theta) = E_{p_\theta}[t(x)^l] = m_\theta^{(l)}(0)}$$

- Example: The exponential family of **exponential distributions**

$$p_\lambda(x) = \lambda \exp(-\lambda x) \qquad \mathcal{X} = (0, \infty)$$

$$\theta = -\lambda \qquad t(x) = x \qquad F(\theta) = -\log(-\theta)$$

$$m_\theta(u) = \exp(F(\theta + u) - F(\theta)) = \frac{\theta}{\theta - u} \Rightarrow \mu_l = E_{p_\lambda}[X^l] = m_\theta^{(l)}(0) = \frac{l!}{\lambda^l}$$

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

Some recently proposed distances:

- **Siegel-Klein distance** (Hilbert distance of the Siegel disk) 10.3390/e22091019
- **Bregman chord divergence** (No need of gradient!) arXiv:1810.09113
- **Bregman-Riemannian distance** (Riemannian distance from Hessian metric) arXiv:1808.01638
- **Statistical Minkowski distances** (derived from Minkowski inequalities) arXiv:1901.03732
- **Total Jensen divergence** (invariant by rotation) arXiv:1309.7109
- **Total Bregman divergence** (invariant by rotation) 10.1109/CVPR.2010.5539979
- **Quasiconvex Jensen divergence** and **quasiconvex Bregman divergence** arXiv:1909.08857
- **Comparative Convex Jensen divergences** and Bregman divergences 10.1109/LSP.2017.2712195
- **Chain Rule Optimal Transport** (CROT, mixture distance using joint convexity) arXiv:1812.08113
- **Hölder Projective Divergence** (do not need normalized densities!) 10.3390/e19030122
- **Optimal copula transport** arXiv:1509.08144
- **Curved Mahalanobis distances** (Cayley-Klein distances) 10.1109/ICIP.2016.7532355
- **Mixed α-Divergences** (composite divergence for clustering) 10.3390/e16063273

Some recently proposed distances II:

- **α-divergences based on comparable weighted means**. arXiv: 2001.09660

→ Obtain also generalized Kullback-Leibler divergences

- **Vector-skewed Jensen-Shannon divergence/Jensen-Shannon** Centroid. 10.3390/e22020221

→ Obtain also vector-skewed Jensen-Bregman divergences

- **(M,N)-Jensen-Shannon divergences..** 10.3390/e21050485

→ Examples: Geometric JSD for exponential families, Power JSDs for t-Student distributions

- **Conformal Divergences** and their centroids 10.1109/TIT.2015.2448072

→ Generalize the dually flat (u,v)-structures of information geometry

- **Bhattacharyya/Chernoff using quasi-arithmetic means.** 10.1016/j.patrec.2014.01.002
- **Matrix Bregman divergences.** ICML 2011

→ Spectral matrix distances

- **Higher-order chi divergences** and f-divergences. 10.1109/LSP.2013.2288355

→ Taylor series for f-divergences based on k-order chi divergences

Dually flat mixture family manifolds: $\mathcal{M} = \left\{ m_\theta = \sum_{i=1}^{D} \theta_i p_i(x) + (1 - \sum_{i=1}^{D} \theta_i) p_0(x) \right\}$

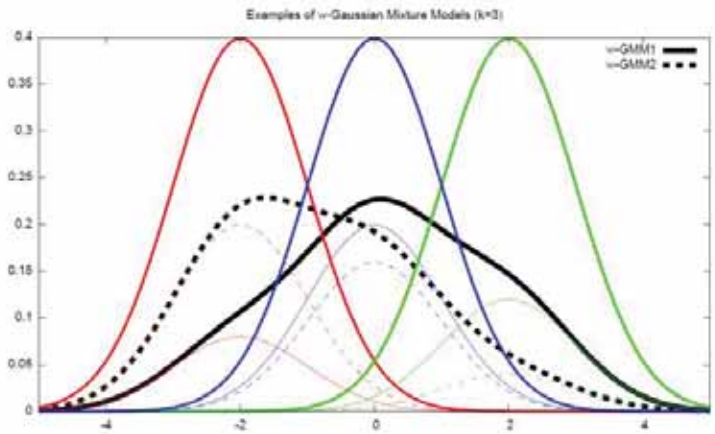
Recovering the Kullback-Leibler divergence from the canonical Legendre-Fenchel divergence

Shannon negentropy (Bregman generator): $F(\theta) = F_{\mathcal{M}}(m_\theta) = -h(m_\theta).$

Convex conjugate (cross-entropy): $F^*(\eta) = -\int p_0(x) \log m_\theta(x) \mathrm{d}\mu(x) = h^\times(p_0 : m_\theta)$

Legendre-Fenchel divergence: $L_F(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2$

$$
\begin{aligned}
D_{F,\mathcal{M}}[m_{\theta_1} : m_{\theta_2}] &:= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2, && (63)\\
&= -h(m_{\theta_1}) - \int p_0(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x) - \int \sum_i \theta_1(i) p_i(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x)\\
&\quad + \sum_i \theta_1(i) p_0(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (64)\\
&= -h(m_{\theta_1}) - \int ((1 - \sum_i \theta_1(i)) p_0(x) + \sum_i \theta_1(i) p_i(x)) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (65)\\
&= -h(m_{\theta_1}) - \int m_{\theta_1}(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (66)\\
&= \int m_{\theta_1}(x) \log \frac{m_{\theta_1}(x)}{m_{\theta_2}(x)} \mathrm{d}\mu(x), && (67)\\
&= D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}]. && (68)
\end{aligned}
$$



$\Rightarrow$ $D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}] = B_F(\theta_1 : \theta_2)$

On w-mixtures: Finite convex combinations of prescribed component distributions, arXiv:1708.00568 V2

Dually flat exponential family manifolds:

Recovering the reverse Kullback-Leibler divergence from the canonical Legendre-Fenchel divergence

- It is well-known that the *KL divergence* between two densities of an exponential family amount to a *Bregman divergence* for the *cumulant function* on **swapped** parameters [AW 2001]: $D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1) = L_F(\theta_2 : \eta_1)$
- However, the **reverse KL divergence** can be **reconstructed** from the dually flat structure of an exponential family: $\mathcal{E} = \{p_\theta(x) = \exp(\theta^\top t(x) - F(\theta)) : \theta \in \Theta\}$,

$$F(\theta) = F_{\mathcal{E}}(p_\theta) = \log\left(\int \exp(\theta^\top t(x))\mathrm{d}\mu(x)\right) \Longrightarrow$$

Convex conjugate: $F^*(\eta) = -h(p_\theta) = \int p(x)\log p(x)\mathrm{d}\mu(x)$

Legendre-Fenchel divergence: $L_F(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2$

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

$$\theta_1^\top \eta_2 = E_{p_{\theta_2}}[\log p_{\theta_1} + F(\theta_1)] = F(\theta_1) + E_{p_{\theta_2}}[\log p_{\theta_1}] \Longrightarrow$$

$$\boxed{\begin{aligned} D_{F,\mathcal{E}}[p_{\theta_1} : p_{\theta_2}] &= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2, \\ &= F(\theta_1) - h(p_{\theta_2}) - E_{p_{\theta_2}}[\log p_{\theta_1}] - F(\theta_1) \\ &= E_{p_{\theta_2}}\left[\log \frac{p_{\theta_2}}{p_{\theta_1}}\right] =: D_{\mathrm{KL}^*}[p_{\theta_1} : p_{\theta_2}], \end{aligned}}$$

Azoury, Warmuth, Relative loss bounds for on-line density estimation with the exponential family of distributions, Machine Learning 43.3 (2001)
On w-mixtures: Finite convex combinations of prescribed component distributions, arXiv:1708.00568 V2

Reconstructing *statistical divergences* from Bregman divergences with statistical generators

- **Statistical model**: $\theta \Leftrightarrow m_\theta$
- **Statistical Bregman generator**: $F(\theta) = \int f(m_\theta(x))\mathrm{d}\mu(x)$

  (strictly convex and differentiable) (integral-based generator)
- Reconstruction process:

$$B_F(\theta_1 : \theta_2) \Rightarrow D_F(m_{\theta_1}(x) : m_{\theta_2}(x)) \Rightarrow D_F(m_1(x) : m_2(x))$$

Bregman divergence (Legendre-Fenchel divergence) → **Statistical divergence between parametric probability models** → **Statistical divergence between arbitrary probability densities**

Example: Mixture family

$$\mathcal{M} = \left\{ m_\theta = \sum_{i=1}^{D} \theta_i p_i(x) + (1 - \sum_{i=1}^{D} \theta_i) p_0(x) \right\}$$

Statistical Bregman generator

$$F(\theta) = F_{\mathcal{M}}(m_\theta) = -h(m_\theta) = \int m_\theta(x) \log m_\theta(x) \mathrm{d}\mu(x) \qquad B_{F,\mathcal{M}}[m_{\theta_1} : m_{\theta_2}] = D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}].$$

Recover the **Kullback-Leibler divergence**: $D_{\mathrm{KL}}(m_{\theta_1}(x) : m_{\theta_2}(x)) \Rightarrow D_{\mathrm{KL}}(m_1(x) : m_2(x))$

An Elementary Introduction to Information Geometry, Entropy 2020

Fisher relative information (OT)
= Fisher divergence or Hyvärinen divergence (IG)

Hyvärinen divergence or **Fisher divergence** (Information Geometry):

$$D_H[p : p_\theta] := \frac{1}{2} \int \|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2 p(x) \mathrm{d}x$$

also called (half of) **relative Fisher information** in Optimal Transport community:

$$I[\mu : \nu] = \int_{\mathcal{X}} \left\| \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|^2 \mathrm{d}\mu = 4 \int_{\mathcal{X}} \left\| \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} \right\|^2 \mathrm{d}\nu.$$

Hyvärinen divergence or order α: $D_{H,\alpha}[p : q] := \int p(x)^\alpha \left(\nabla_x \log p(x) - \nabla_x \log q(x)\right)^2 \mathrm{d}x, \quad \alpha > 0$

(closed-form for 2-Hyvärinen divergence between a GMM and an exponential-polynomial density)

Otto, Villani, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality, Journal of Functional Analysis 173.2 (2000)

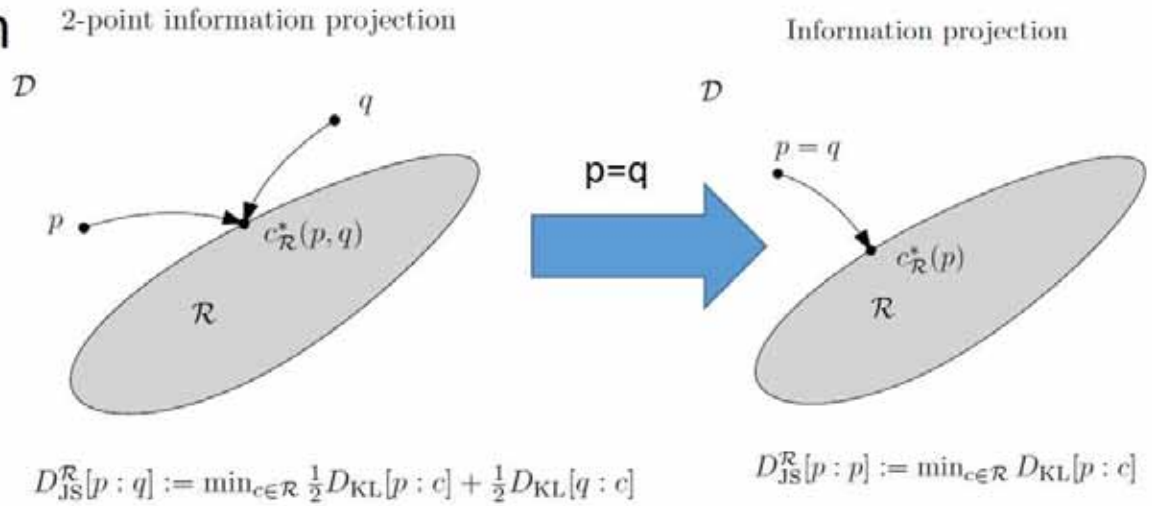
Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities (2021)

Relative Jensen-Shannon divergence & 2-point projections

- Constrain the variational definition of the Jensen-Shannon divergence to belong to **a prescribed family R**:

$$D_{\mathrm{JS}}^{\mathcal{R}}[p:q] := \min_{c\in\mathcal{R}} \tfrac{1}{2} D_{\mathrm{KL}}[p:c] + \tfrac{1}{2} D_{\mathrm{KL}}[q:c]$$

- When R is all space of densities of a measurable space, recover the ordinary Jensen-Shannon divergence
- Relative JSD interpreted as a 2-point projection. When p=q, information projection



Not a proper divergence because:

$$\forall p, q \notin \mathcal{R}, \quad D_{\mathrm{JS}}^{\mathcal{R}}[p:q] > 0$$

Ordinary JSD when $\forall p, q \in \mathcal{R}$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius

Entropy from divergence and uniform distribution

- **Discrete distribution** P with probability mass function $p_1, \ldots, p_d$
- Sample space (alphabet) X of finite size |X|= d
- Shannon entropy H(P) derived from the Kullback-Leibler divergence between P and the uniform distribution U:

$$H(P) = -D_{\mathrm{KL}}[P : U] + \log |X|$$

- Similarly, Rényi entropy H α(P) derived from the Rényi information of order α commonly called "Rényi divergence":

$$\boxed{H_\alpha^R(P) = -D_\alpha^R[P : U] + \log |X|}$$

Uniform distribution:

$$u_i = u = \frac{1}{d} = \frac{1}{|X|}$$

$$\begin{aligned} D_{\mathrm{KL}}[P : U] &:= \sum_{i=1}^{d} p_i \log \frac{p_i}{u_i} \\ &= -H(P) + \log \frac{1}{u} \end{aligned}$$

$$\begin{aligned} D_\alpha^R[P : U] &:= \frac{1}{\alpha - 1} \log \sum_{i=1}^{d} p_i^\alpha u_i^{1-\alpha} \\ &= \frac{1}{\alpha - 1} \log \sum_{i=1}^{d} p_i^\alpha \left(\frac{1}{d}\right)^{1-\alpha} \\ &= \underbrace{\frac{1}{\alpha - 1} \log \sum_{i=1}^{d} p_i^\alpha}_{=:-H_\alpha^R(P)} + \log |X| \end{aligned}$$

On Rényi and Tsallis entropies and divergences for exponential families, arXiv:1105.3259

Representational Bregman divergences

- α-divergences, $\alpha \in \mathbb{R}$:

$$I_\alpha^A(p:q) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^d \left(\alpha p_i + (1-\alpha) q_i - p_i^\alpha q_i^{1-\alpha}\right), \alpha \notin \{0,1\} \\ \sum_{i=1}^d p_i \log \frac{p_i}{q_i} + q_i - p_i = \mathrm{KL}_+(p:q), \alpha = 1 \\ \sum_{i=1}^d q_i \log \frac{q_i}{p_i} + p_i - q_i = \mathrm{KL}_+^r(p:q), \alpha = 0 \end{cases}$$

- β-divergences, $\beta \in \mathbb{R}_+$

$$I_\beta^B(p:q) = \begin{cases} \frac{\beta}{\beta+1} \sum_{i=1}^d (p_i^{\beta+1} + \beta q_i^{\beta+1} - (\beta+1) p_i q_i^\beta), \\ \sum_{i=1}^d p_i \log \frac{p_i}{q_i} + q_i - p_i = \mathrm{KL}(p:q), \beta = 0 \\ \sum_{i=1}^d \left(\frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1\right) = \mathrm{IS}(p:q), \beta = -1 \end{cases}$$

- Conversion by a **non-linear transformation** $I_\alpha^A \leftrightarrow I_\beta^B$:

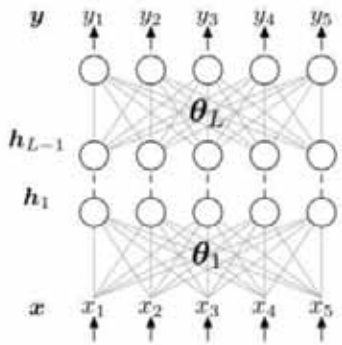
$$I_\alpha^A(p:q) = I_{\frac{1}{\alpha}-1}^B \left(\left(\frac{p}{\alpha^2}\right)^\alpha : \left(\frac{q}{\alpha^2}\right)^\alpha\right)$$

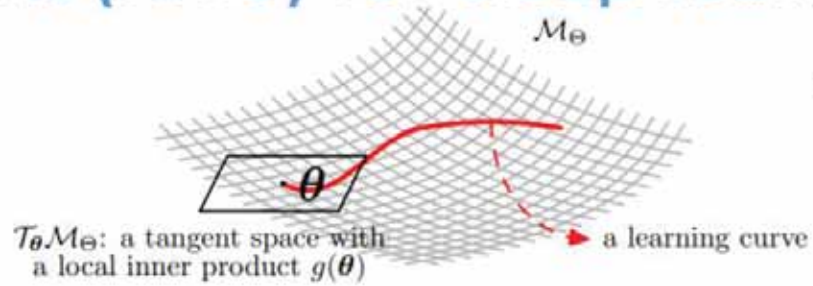
- α-divergences and β-divergences are **representational** Bregman divergences[1] on positive arrays $\mathbb{R}_+^d$.

[1] The dual Voronoi diagrams with respect to representational Bregman divergences, 2009.

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

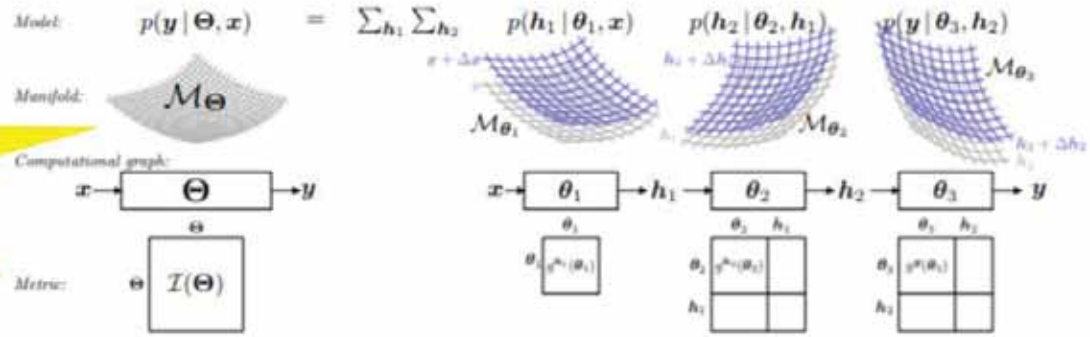
$$p(\boldsymbol{y}\,|\,\boldsymbol{x},\Theta) = \sum_{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_{L-1}} p(\boldsymbol{y}\,|\,\boldsymbol{h}_{L-1},\boldsymbol{\theta}_L)\cdots p(\boldsymbol{h}_2\,|\,\boldsymbol{h}_1,\boldsymbol{\theta}_2)p(\boldsymbol{h}_1\,|\,\boldsymbol{x},\boldsymbol{\theta}_1),$$





$$g(\Theta) = E_{\boldsymbol{x}\sim\hat{p}(X_n),\,\boldsymbol{y}\sim p(\boldsymbol{y}\,|\,\boldsymbol{x},\Theta)}\left[\frac{\partial l}{\partial\Theta}\frac{\partial l}{\partial\Theta^\top}\right] = \frac{1}{n}\sum_{i=1}^{n} E_{p(\boldsymbol{y}\,|\,\boldsymbol{x}_i,\Theta)}\left[\frac{\partial l_i}{\partial\Theta}\frac{\partial l_i}{\partial\Theta^\top}\right]$$

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



Dynamic geometry

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

@FrnkNlsn

Relative Fisher Information and Natural Gradient for Learning Large Modular Models (ICML'17)

Riemannian geodesics versus affine geodesics

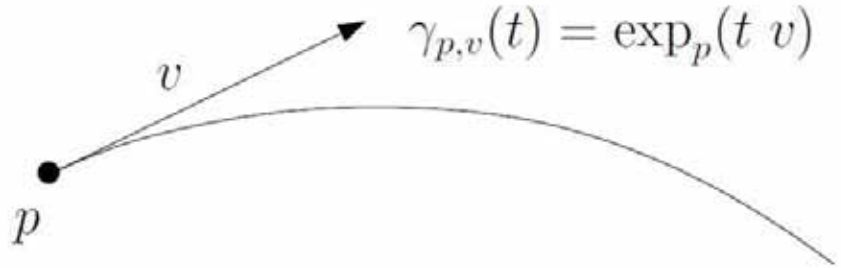
- In Riemannian geometry, **geodesics** are **locally shortest length minimizing curves** parameterized by **arclengths**.
- In information geometry, geodesics induced by an affine connection ∇ are **autoparallel curves** parameterized by an affine parameter.
- **Autoparallel** means that the tangent vector keeps the **same direction** along the curve. **Parallel-transported** means the transported vector keeping the **same scale**.
- Keeping same direction: $\nabla_{\dot\gamma(t)}\dot\gamma(t) \propto \dot\gamma(t)$ Same scale: $\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0$

- The affine geodesic parameter is said **affine** because if t is a parameterization then t'=at+b yields also a valid parameterization so that $\nabla_{\dot\gamma(t')}\dot\gamma(t') = 0$
- **Pregeodesics** are geodesics curve shape (without parameterization).
- Affines connections differing only by **torsions** yield the same geodesics
- **Levi-Civita connection**: unique torsion-free affine connection **induced by the metric g** such that ${}^{\mathrm{LC}}\nabla g = 0$. LC affine connection yields Riemannian geodesics.

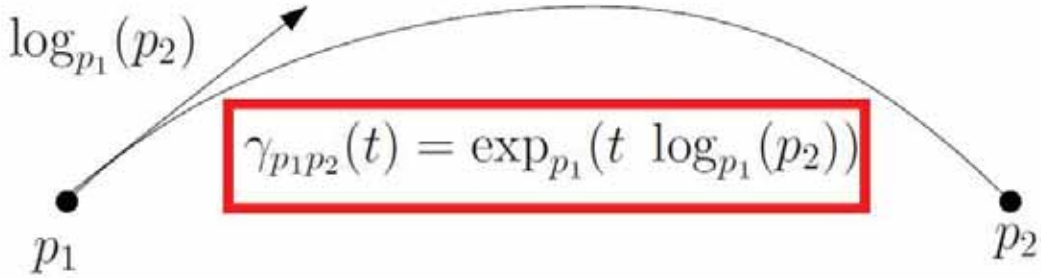
An elementary introduction to information geometry, *arXiv:1808.08271*

Geodesics on Hadamard manifolds

- **Hadamard manifolds** = simply connected complete Riemannian manifold M with non-positive sectional curvature. Example: Cone of symmetric positive-definite matrices (SPD cone)
- For any p in M, Riemannian **$\exp_p$** is a **global diffeomorphism**
- Geodesics wrt boundary conditions in closed-form wrt Riemannian exp/log:



Geodesic wrt. initial conditions



Geodesic wrt. boundary conditions

- Example SPD cone:

$$\mathrm{Exp}_P(M) = P^{\frac{1}{2}}\mathrm{Exp}(P^{-\frac{1}{2}}MP^{-\frac{1}{2}})P^{\frac{1}{2}}$$

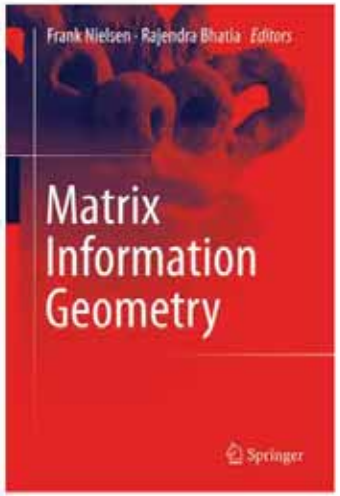
$$\exp(M) = \sum_{i=0}^{\infty} \frac{1}{i!}M^i$$

$$\mathrm{Log}_P(M) = P^{\frac{1}{2}}\mathrm{Log}\left(P^{-\frac{1}{2}}MP^{-\frac{1}{2}}\right)P^{\frac{1}{2}}$$

$$P(t) = P_1^{\frac{1}{2}}\exp(t\mathrm{Log}(P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}}))P_1^{\frac{1}{2}}$$

On Approximating the Riemannian 1-Center, Computational Geometry 46(1), 2012

The Riemannian mean of positive matrices



- **Riemannian mean** of two positive-definite matrices:

$$\boxed{P_1 \# P_2 = P_1^{\frac{1}{2}} \left(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}} \right)^{\frac{1}{2}} P_1^{\frac{1}{2}}}$$

- Solution of the Ricatti equation: $X P_1^{-1} X = P_2$
- Riemannian geodesic equation with respect to the **trace metric**:

$$\gamma_{P_1 P_2}(t) = P_1 \#_t P_2, \quad \gamma_{P_1 P_2}(0) = P_1, \gamma_{P_1 P_2}(1) = P_2 \qquad P_1 \#_t P_2 = P_1^{\frac{1}{2}} (P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})^t P_1^{\frac{1}{2}}$$

- Invariance by inversion: $(P_1 \# P_2)^{-1} = P_1^{-1} \# P_2^{-1}$
- Harmonic-Geometric-Arithmetic inequality: $\left(\frac{P_1^{-1} + P_2^{-1}}{2} \right)^{-1} \preceq P_1 \# P_2 \preceq \frac{P_1 + P_2}{2}$

  Lowner partial ordering
- For scalars, = **geometric mean**: $G(p_1, \ldots, p_n) = \left(\prod_{i=1}^{n} p_i \right)^{\frac{1}{n}}$
- **Inductive mean** yields the geometric Riemannian mean in the limit:

$$\boxed{S_1 = P_1, \quad S_i = S_{i-1} \#_{\frac{1}{i}} P_{((i-1) \mod (n))+1} \qquad \lim_{n \to \infty} S_n = G(P_1, \ldots, P_n)}$$

Matrix information geometry, Springer, 2013.

Riemannian minimum enclosing ball

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

Algorithm `GeoA`

$c_1 \leftarrow$ choose randomly a point in $\mathcal{P}$;

for $i = 2$ **to** l **do**

  `// farthest point from` c_i

  $s_i \leftarrow \arg\max_{j=1}^{n} \rho(c_i, p_j)$;

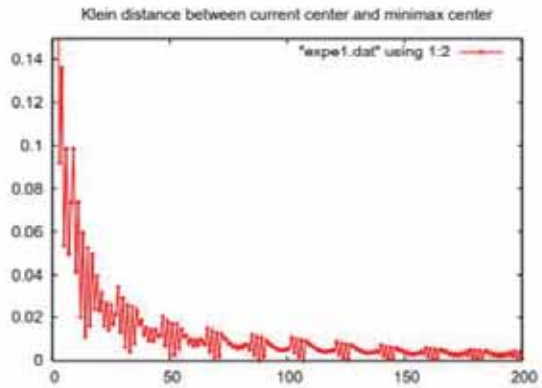
  `// update the center: walk on the geodesic line segment` $[c_i, p_{s_i}]$

  $c_{i+1} \leftarrow c_i \#_{\frac{1}{i+1}}^M p_{s_i}$;

end

`// Return the SEB approximation`

return $\mathrm{Ball}(c_l, r_l = \rho(c_l, \mathcal{P}))$;



Hyperbolic geometry:

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

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

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

Positive-definite matrices:

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

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

@FrnkNlsn

On Approximating the Riemannian 1-Center, Comp. Geom. 2013

Approximating Covering and Minimum Enclosing Balls in Hyperbolic Geometry, GSI, 2015

Optimization on Riemannian manifolds $(M, g, \nabla^{\mathrm{LC}})$

- <u>**Directional derivative**</u>:

$$Df(x)[v] = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}$$

- <u>**Riemannian gradient**</u>: unique $\mathrm{grad}_p f(x) \in T_p M$ such that

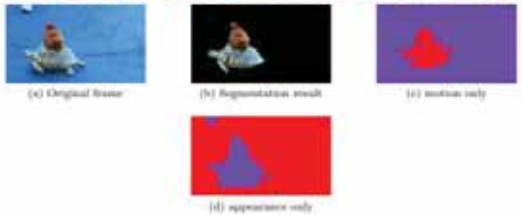
$$\langle \mathrm{grad}_p f(x), v \rangle = Df(x)[v], \quad \forall v \in T_p M$$

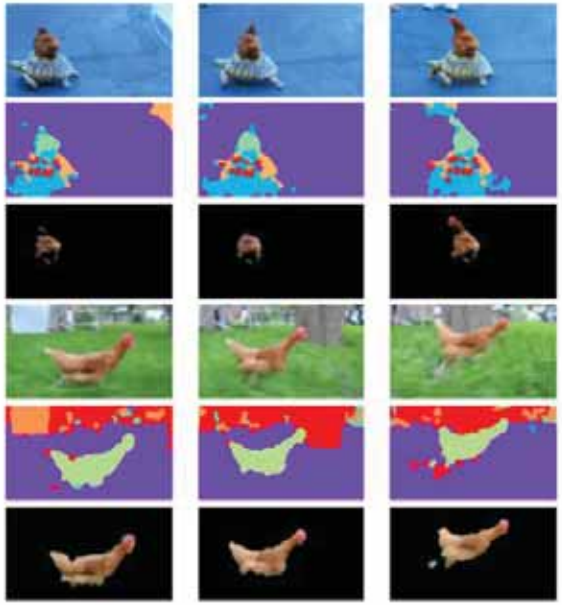
- <u>**Riemannian Hessian**</u>: $\mathrm{hess}_p f(x) : T_p M \to T_p M$ such that

$$\mathrm{hess}_p f(x)[v] = \nabla_v^{\mathrm{LC}} \mathrm{grad}_p f(x), \quad \forall v \in T_p M$$

  ∇^{LC}: Riemannian metric connection (torsion-free)
  Property: $\langle \mathrm{hess}_p f(v)[v], w \rangle_p = \langle v, \mathrm{hess}_p f(x)[w] \rangle_p$ for all $v, w \in T_p M$.
- <u>**Retraction**</u>: $R : TM \to M$ with R_p its restriction to $T_p M$: (i) centering: $R_p(0) = p$, (ii) local rigidity: curve $\gamma_v(t) = R_p(tv)$ satisfies $\frac{d\gamma_v(t)}{dt}|_{t=0} = v$ for all $v \in T_p M$.

SSSC-AM: A Unified Framework for Video Co-Segmentation by Structured Sparse Subspace Clustering with Appearance and Motion Features





Unified optimization framework

We propose the following unified optimization framework (4). In order to make it work on affine subspaces, we add an affine subspace constraint $Z_k^\top \mathbf{1} = \mathbf{1}$ to the original S^{3}C framework (1), and formulate the optimization problem to solve as follows:

$$C \min_{\substack{Z_1,Z_2,\ldots,Z_K \\ E_1,E_2,\ldots,E_K \\ Q_1,Q_2,\ldots,Q_K}} \sum_{k=1}^{K} (\|Z_k\|_{1,Q_k} + \lambda_k \|E_k\|_1) + \beta \|Z\|_{2,1}$$

$$s.t. \quad X_k = X_k Z_k + E_k, \ \mathrm{diag}(Z_k) = \mathbf{0}, \quad Z_k^\top \mathbf{1} = \mathbf{1}, \ Q_k \in \mathcal{Q}$$

where

$$Z = \begin{pmatrix} \mathrm{vec}(Z_1^\top) & \mathrm{vec}(Z_2^\top) & \cdots & \mathrm{vec}(Z_K^\top) \end{pmatrix}^\top = \begin{bmatrix} (Z_1)_{11} & (Z_1)_{12} & \cdots & (Z_1)_{NN} \\ (Z_2)_{11} & (Z_2)_{12} & \cdots & (Z_2)_{NN} \\ \vdots & \vdots & & \vdots \\ (Z_K)_{11} & (Z_K)_{12} & \cdots & (Z_K)_{NN} \end{bmatrix}_{K \times N^2}$$

and $\|Z_k\|_{1,Q_k} = \|Z_k\|_1 + \alpha \|\Theta_k \odot Z_k\|_1$. $\mathbf{0}$ and $\mathbf{1}$ are two vectors of appropriate dimensions filled with 0 and 1, respectively. X_k refers to different feature data points.

No. motions	2		3	
Error(%)	Ave.	Med.	Ave.	Med.
S^{3}C [11]	1.94	**0**	4.92	0.89
S^{3}C-AM	**1.53**	**0**	**4.40**	**0.56**

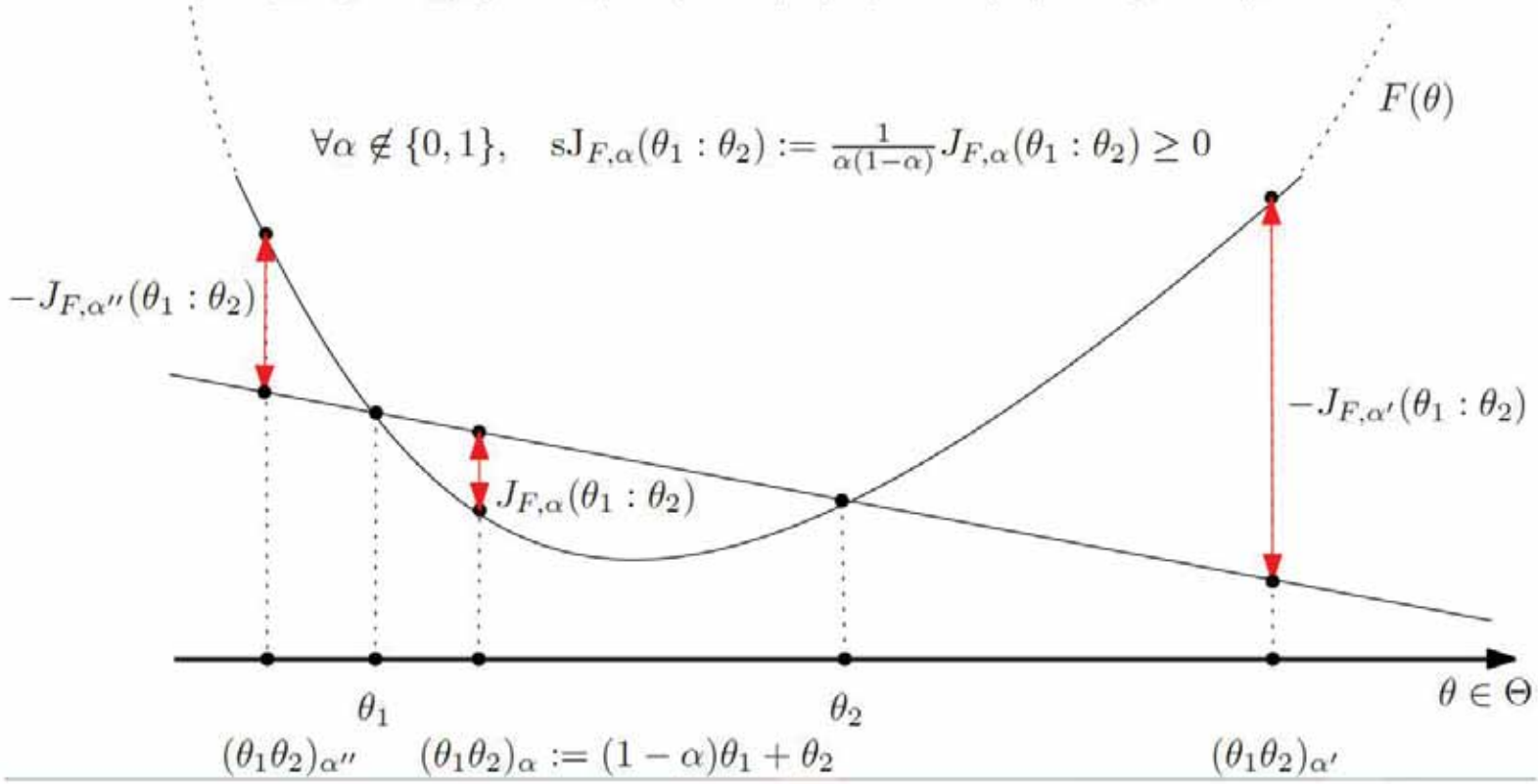
@FrnkNlsn

IEEE ICIP 2016, https://arxiv.org/abs/1603.04139

Scaled skewed Jensen divergences & Bregman divergences

$$\forall \alpha \in (0,1), \quad J_{F,\alpha}(\theta_1 : \theta_2) := (1-\alpha)F(\theta_1) + \alpha F(\theta_2) - F((1-\alpha)\theta_1 + \alpha\theta_2)$$

$$\forall \alpha \notin \{0,1\}, \quad \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) := \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1 : \theta_2) \geq 0$$



Jensen divergences

measures the vertical gap induced by a strictly convex function

$\lim_{\alpha \to 0} \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_2)$ (Bregman divergence) $\lim_{\alpha \to 1} \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$ (reverse BD)

The Burbea-Rao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57.8 (2011)

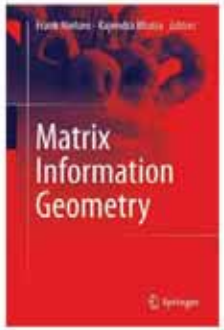
Gauge functions and Schatten-von Neumann matrix norms

M: a complex square matrix, M* its conjugate transpose.
Eigenvalues λ and singular values: $s_i(M) = \sqrt{\lambda_i(M^*M)}$
Unitary matrix: $U^{-1} = U^*$

Schatten-von Neumann matrix norm:

Symmetric gauge function Φ = d-variate function invariant under permutation and sign changes:

$$\boxed{\|M\|_\Phi = \Phi(s_1(M), \ldots, s_d(M))}$$

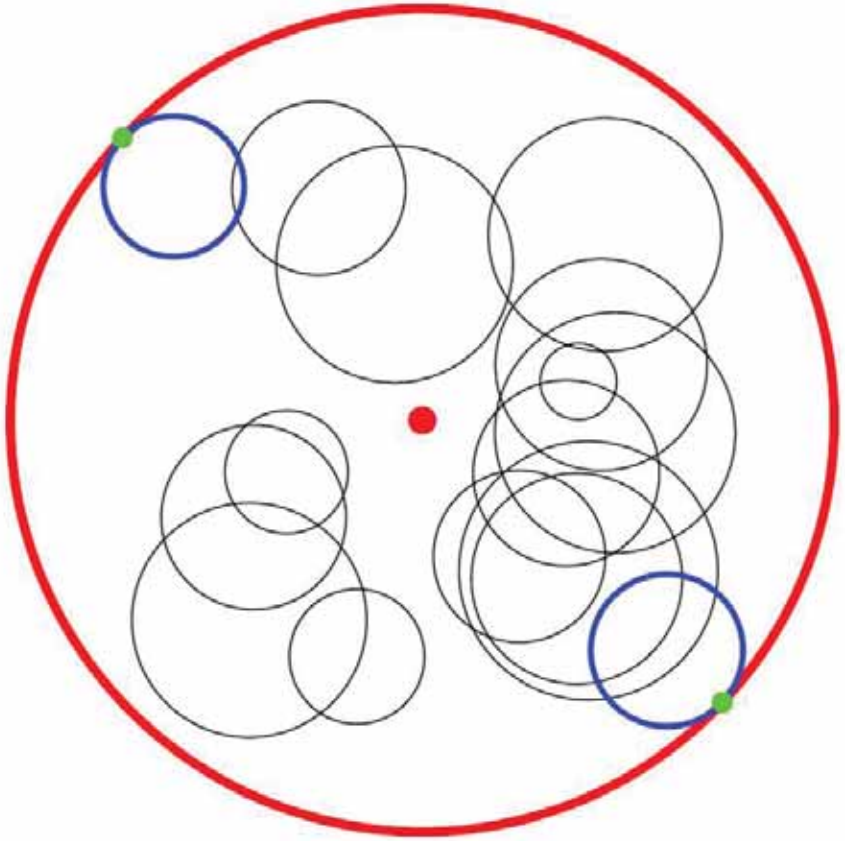


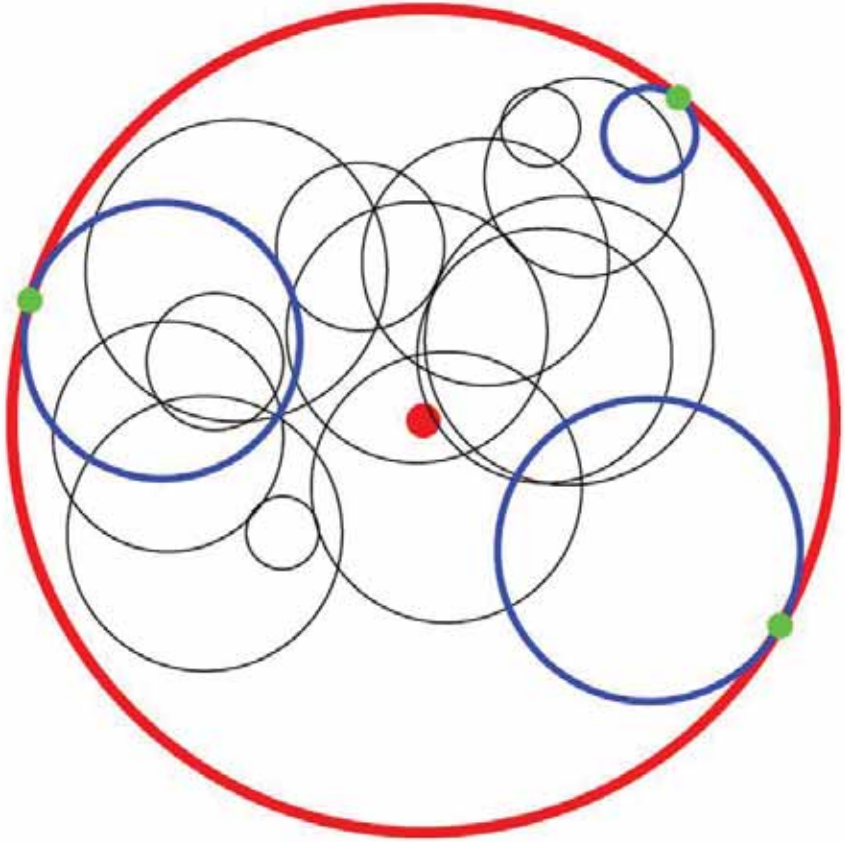
Property: these norms are unitarily invariant: $\|UMV\|_\Phi = \|M\|_\Phi$

Example: **Schatten p-norms** for $\Phi_p(x) = (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$

Mining matrix data with Bregman matrix divergences for portfolio selection, Matrix Information Geometry, 2013.

Smallest enclosing ball of balls (SEBB)





Approximating smallest enclosing balls, ICCSA (2004)
Approximating smallest enclosing balls with applications to machine learning, IJCGA (2009)
A fast deterministic smallest enclosing disk approximation algorithm, IPL (2005)

Sectional curvatures from affine connections

- **Riemann curvature (1,3)-tensor** R induced by an affine connection ∇ :

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

- **Sectional curvatures** are defined for *tangent 2-planes* which are 2-dimensional subspaces of the tangent plane TpM:

$$\boxed{\sec_p(\pi) := \frac{\langle R(X,Y)Y, X\rangle}{Q_p(X,Y)}}$$

$$Q_p(X,Y) = \langle X,X\rangle_p \langle Y,Y\rangle_p - \langle X,Y\rangle_p^2$$

Squared area of the **parallelogram** induced by vectors X and Y

Sec is **independent** of the choice of the basis X, Y

- In local coordinates: $\kappa_{ij} := \sec(\mathrm{span}\{\partial_i, \partial_j\})$ generalize the **Gauss curvature** of surfaces:

Local basis $\{\partial_1, \partial_2\}$ ➡ $\sec(T_pM) = \dfrac{\mathrm{R}_{2112}}{(g_{11}g_{22} - g_{12}^2)} = \dfrac{\langle(\nabla_{\partial_2}(\nabla_{\partial_1}\partial_1) - \nabla_{\partial_1}(\nabla_{\partial_2}\partial_1)), \partial_2\rangle}{\det g}$

- ***Sectional curvatures fully define R*** (ie., R can be reconstructed from sec).
- A manifold is of constant curvature κ iff all sectional curvatures are equal to κ.
- A manifold is flat if it is of constant curvature κ=0.

arxiv:1808.08271

Semi-closed-form formula for the Kullback-Leibler divergence

- KLD between an **arbitrary density p** with moment $E_p[x] = m$ and covariance matrix $\mathrm{Cov}_p[x] = S$ and a Gaussian density q=N(μ, Σ) is:

$$D_{\mathrm{KL}}[p : q_{\mu,\Sigma}] = \frac{1}{2}\left(\log|2\pi\Sigma| + (\mu - m)^\top \Sigma^{-1}(\mu - m) + \mathrm{tr}(\Sigma^{-1}S)\right) - h(p)$$

- In general, KLD between an **arbitrary density** p with moment $m = E_p[t(x)]$ and entropy h[p] and a density $q_\theta \sim \exp(\theta^\top t(x) - F(\theta))$ of a exponential family is:

$$D_{\mathrm{KL}}[p : q_\theta] = F(\theta) - m^\top \theta - h[p]$$

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances based on the Information Radius

Entropy power inequality (EPI)

Claude Elwood Shannon
1916-2001

- For a continuous random variable X of R^d, **entropy power** N(X) where h(X) denotes the **differential entropy**:

$$N(X) := \frac{1}{2\pi e} \exp\left(\frac{2}{d} h(X)\right)$$

- When $X \sim N(\mu, \Sigma)$ is a d-dimensional normal distribution: $N(X) = |\Sigma|^{\frac{1}{d}}$

- The entropy power is a **superadditive function**: For X,Y independent RVs with densities in L^p space

$$\boxed{N(X+Y) \geq N(X) + N(Y)}$$

with **equality iff X and Y are independent normal distributions**

Shannon, A mathematical theory of communication (1948)

Sharma-Mittal entropy[1] (Rényi/Tsallis/Shannon)

- **Sharma-Mittal entropy** $\alpha, \beta > 0, \alpha, \beta \neq 1$:

$$\boxed{h_{\alpha,\beta}(p) = \frac{1}{1-\beta}\left(\left(\int p(x)^\alpha \mathrm{d}x\right)^{\frac{1-\beta}{1-\alpha}} - 1\right)}$$

  generalizes **Rényi entropies** $R_\alpha(p) = \frac{1}{1-\alpha}\log\int p(x)^\alpha \mathrm{d}x$ $(\beta \to 1)$, **Tsallis entropies** $T_\alpha(p) = \frac{1}{1-\alpha}(\int p(x)^\alpha \mathrm{d}x - 1)$ $(\beta = \alpha)$, and **Shannon entropy** $h(p) = -\int p(x)\log p(x)\mathrm{d}x$ $(\alpha, \beta \to 1)$.
- For multivariate normal/Gaussian distributions $p_{\mu,\Sigma} \sim \mathcal{N}(\mu, \Sigma)$:

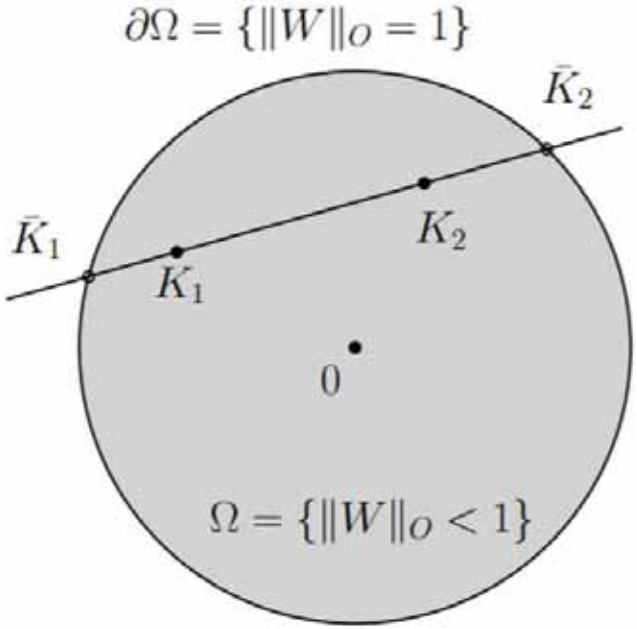
$$h_{\alpha,\beta}(p_{\mu,\Sigma}) = \frac{1}{1-\beta}\left(\frac{\left((2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}\right)^{1-\beta}}{\alpha^{\frac{d(1-\beta)}{2(1-\alpha)}}} - 1\right), \quad h(p_{\mu,\Sigma}) = \log\sqrt{(2\pi e)^d|\Sigma|}$$

$$R_\alpha(p_{\mu,\Sigma}) = \log\frac{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}{\alpha^{\frac{d}{2(1-\alpha)}}}, \quad T_\alpha(p_{\mu,\Sigma}) = \frac{1}{1-\alpha}\left(\frac{\left((2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}\right)^{1-\alpha}}{\alpha^{\frac{d}{2}}} - 1\right)$$

- Extend to arbitrary **exponential families** with natural parameters being a cone.

[1] https://arxiv.org/abs/1112.4221

Siegel-Klein distance and geometry



Siegel-Klein distance: $\rho_K(K_1, K_2) = \frac{1}{2}\log\left(\frac{\alpha_+(1-\alpha_-)}{|\alpha_-|(\alpha_+ - 1)}\right)$

$$\bar{K}_1 = K_1 + \alpha_-(K_2 - K_1)$$

$$\bar{K}_2 = K_1 + \alpha_+(K_2 - K_1)$$

Siegel-Klein distance: from the disk origin 0

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

$$\mathbb{SD}(d) \;:=\; \left\{W \in \mathrm{Sym}(d, \mathbb{C}) \;:\; I - \overline{W}W \succ 0\right\}.$$

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

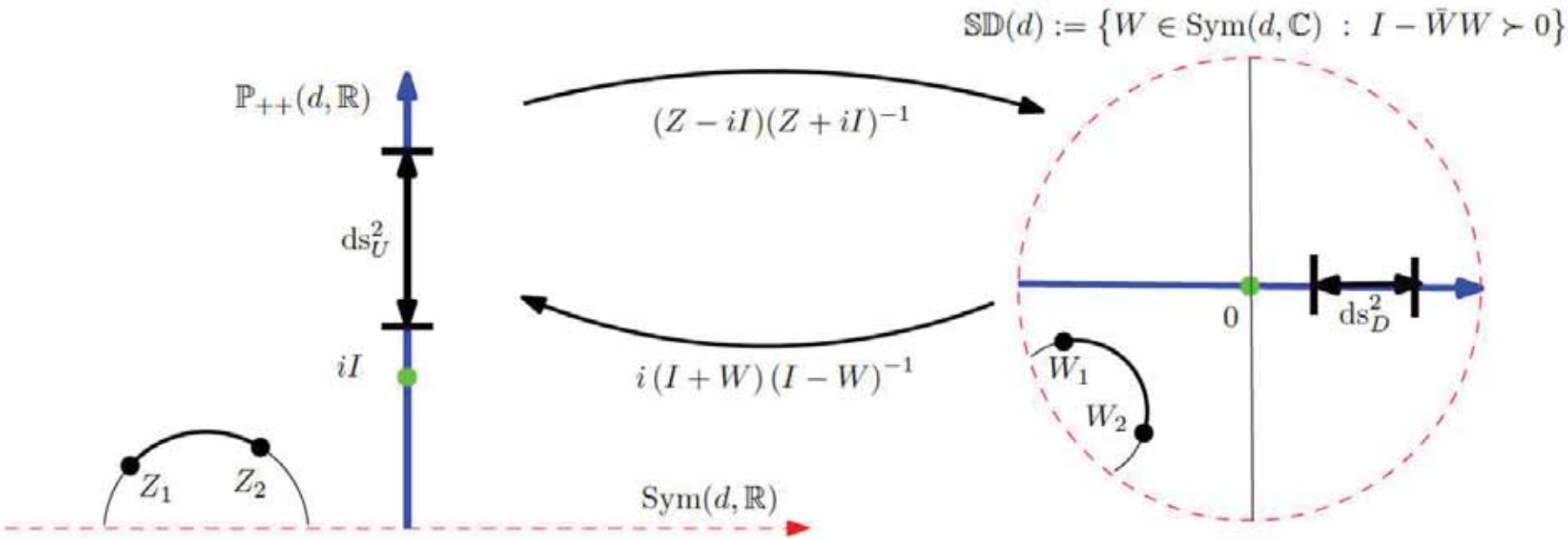
$$\begin{aligned}\|M\|_O &= \max_{x\neq 0}\frac{\|Mx\|_2}{\|x\|_2},\\ &= \sqrt{\lambda_{\max}(M^H M)},\\ &= \sigma_{\max}(M).\end{aligned}$$

Generalize Hilbert geometry:

- Symmetric positive-definite matrix manifold (SPD)
- Hyperbolic geometry (and polydisk)

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

Siegel-Klein distance and geometry



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

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

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

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

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

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

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

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

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 Poincare upper plane, and the Siegel disk generalizes the Poincare 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.

Summary and conclusion

- Since Bregman divergences are defined for generators up to an affine term, considered $F(\theta) = -\log p_\theta(\omega)$ for **any** ω in the convex hull of the support $\mathcal{X}$ of the exponential family
- Can express the KLD (as a reverse BD) between two densities of an exponential family as an **average sum of log density-ratio**:

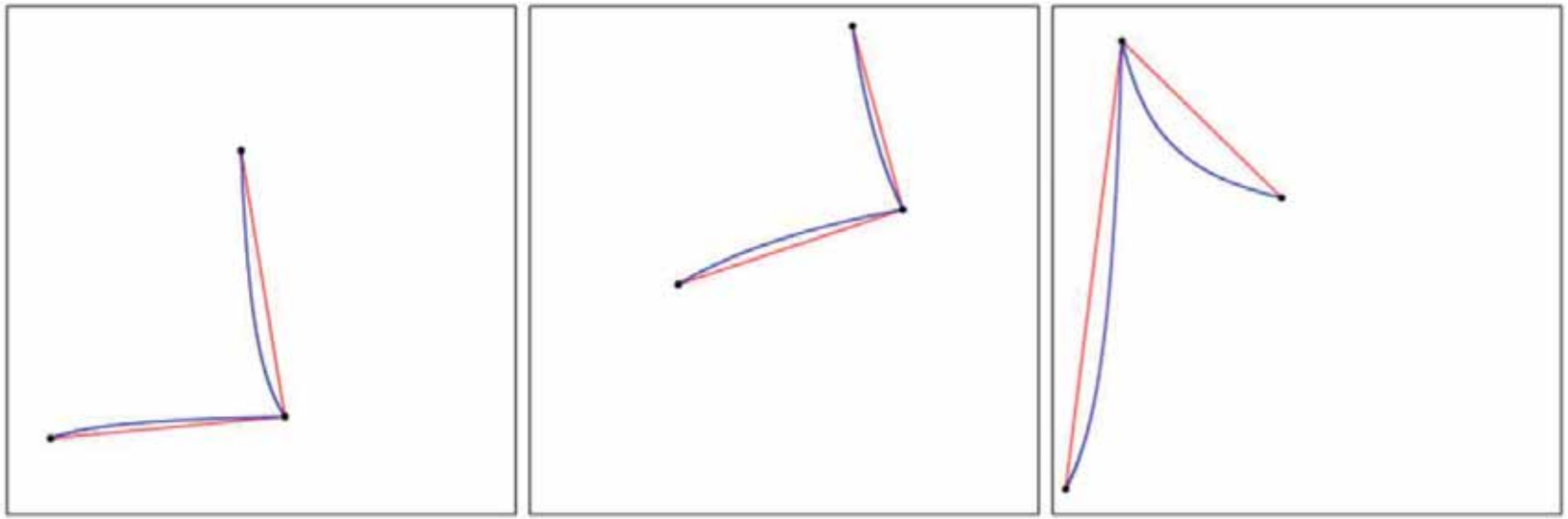
$$\boxed{D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s}\sum_{i=1}^{s} \log\left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)}\right)}$$

  where $\omega_1, \ldots, \omega_s$ are $s \leqslant D+1$ distinct points of $\mathcal{X}$ chosen such that $\frac{1}{s}\sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)] = \eta_1$.
- Allows to **assess the error in the Monte Carlo estimators of KLD**
- Approach works whenever a statistical divergence between two parametric densities amount to a Bregman divergence (eg., q-exponential families, χ-exponential families), etc.
- Can also considered other statistical divergences like Bhattacharyya, Hellinger, α-divergences between exponential families and report formula without the explicit use of the log-normalizer $F(\theta)$, see [13]

Triples of points (p,q,r) with dual Pythagorean theorems holding simultaneously at q

$$\gamma_{pq} \perp_q \gamma^*_{qr} \Leftrightarrow (\theta(p)-\theta(q))^\top(\eta(r)-\eta(q)) = 0 \Leftrightarrow D_F(p:q) + D_F(q:r) = D_F(p:r)$$

$$\gamma^*_{pq} \perp_q \gamma_{qr} \Leftrightarrow (\eta(p)-\eta(q))^\top(\theta(r)-\theta(q)) = 0 \Leftrightarrow D_F(r:q) + D_F(q:p) = D_F(r:p)$$



Itakura-Saito Manifold
(solve quadratic system)

Two blue-red geodesic pairs orthogonal at q

https://arxiv.org/abs/1910.03935

Skew Jensen-Shannon divergences & generalizations

f-divergence for a convex generator f with f(1)=0: $I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x).$

Ordinary skewed Jensen-Shannon divergence is a f-divergence:

$$\boxed{\begin{aligned} D_{\mathrm{JS}}^{\alpha,\beta}[p:q] &:= (1-\beta) D_{\mathrm{KL}}[p:m_\alpha] + \beta D_{\mathrm{KL}}[q:m_\alpha] \\ &= I_{f_{\mathrm{JS}}^{\alpha,\beta}}[p:q] \\ & f_{\mathrm{JS}}^{\alpha,\beta}(u) = -\left((1-\beta)\log\left(\alpha u + (1-\alpha)\right) + \beta u \log\left(\frac{1-\alpha}{u} + \alpha\right)\right) \end{aligned}}$$

Generalization of skewed Jensen-Shannon divergence wrt weighted means M and N

$$\boxed{D_{\mathrm{JS}}^{M_\alpha,N_\beta}[p:q] := N_\beta\left(D_{\mathrm{KL}}(p:(pq)_{M_\alpha}), D_{\mathrm{KL}}(q:(pq)_{M_\alpha})\right)}$$

Statistical M-mixtures: $(pq)_{M_\alpha}(x) = \frac{M_\alpha(p(x),q(x))}{\int M_\alpha(p(x),q(x))\mathrm{d}\mu(x)}$

On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. Entropy (2019)

Background for Information Geometry

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

Frank Nielsen

Sony CSL

An elementary introduction to information geometry

https://arxiv.org/abs/1808.08271

© Frank Nielsen

Some applications of information geometry:

- Natural gradient and deep learning
- Bayesian hypothesis testing
  geometry of the error exponent
- Clustering
  partition-based, soft mixtures and hierarchical dendrograms

Frank Nielsen

Sony CSL

Information-geometric structures:

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

Frank Nielsen

Sony CSL

© Frank Nielsen

When is a smooth function analytic?[1]

- A **smooth function** is a C^∞ function on an open interval I
- An **analytic function** is a function on an open interval D such that its *Taylor series* at any point $x_0 \in D$ $T_{f,x_0}(x) := \sum_{i=0}^{\infty} \frac{f^{(i)}(x)}{i!}(x - x_0)^i$ **converges** to $f(x)$ for $|x - x_0| < \rho(x_0)$ (and diverges otherwise). C^ω denotes the space of analytic functions.
- Example: $\exp(x) := \sum_{i=0}^{\infty} \frac{x^i}{i!}$ is an analytic function in $\mathbb{R}$
- Taylor series converge to f when the Taylor remainder converges to zero. Two kinds of **Taylor singularities**: (1) $T_{f,x_0}(x)$ diverges except at x_0, (2) $T_{f,x_0}(x)$ converges to a different function in a neighborhood of x_0.
- **Theorem**: Let $f \in C^\infty(I)$ and $\rho(x)$ denotes the radius of convergence of the Taylor series of f at x. When
  - $\rho(x) > 0$ for all $x \in I$, **and**
  - for all $x_0 \in I$ we have $\liminf_{x \to x_0} \frac{\rho(x)}{|x - x_0|} > 1$

  then $f \in C^\omega(I)$.

[1] Boas, R. P. "When is a C^∞ function analytic?", The mathematical Intelligencer 11.4 (1989): 34-37.

Softplus: A strictly convex LogSumExp

- **Softplus** is an activation function (smooth ReLU): $\mathrm{Softplus}(x) := \log\left(1 + e^x\right)$

  Softplus is strictly convex

- **LogSumExp** (LSE) is a convex function: $\mathrm{LSE}(a, b) := \log(e^a + e^b)$

  Not strictly convex because $\mathrm{LSE}(a, b) = \lambda + \mathrm{LSE}(a - \lambda, b - \lambda), \forall \lambda$

Related to the **exponential quasi-arithmetic mean** $M_f(a, b) := f^{-1}\left(\frac{f(a) + f(b)}{2}\right)$:

$$\mathrm{LSE}(a, b) = M_{\exp}\left(a + \log\frac{1}{2}, b + \log\frac{1}{2}\right) = \log_2 + M_{\exp}(a, b)$$

$$\boxed{\mathrm{Softplus}(x) = \mathrm{LSE}(0, x) =: \mathrm{LSE}_+(x)}$$

Monte Carlo information-geometric structures.,*Geometric Structures of Information*. Springer, 2019. 69-103.

Symmetric positive-definite matrix & correlation matrix

- Set of symmetric d x d matrices is a vector space which contains the **cone of symmetric positive-definite (SPD) matrices** satisfying: $P \in \mathbb{P}_{++}(d) \Rightarrow \forall x \neq 0, x^\top P x > 0$
- SPD matrix P satisfies $P_{ii} > 0, P_{ij} < \sqrt{P_{ii}P_{jj}}$ and has all *eigenvalues positive*: $\lambda_i(P) > 0$
- A *sufficient condition* to be SPD is to be **diagonally dominant**: $P_{ii} > \sum_{j \neq i} |P_{ij}|$
- The determinant of a SPD matrix satisfies (with equality for diagonal SPD matrices)

$$\boxed{0 < \det(P) \leq \prod_{i=1}^{d} P_{ii}}$$

- The inverse of a SPD matrix is SPD, and a SPD matrix has a unique Cholesky decomposition (lower triangular matrix), and a unique square root matrix
- A **correlation matrix** is a SPD matrix with all coefficients in [-1,1] and *diagonal unit elements (therefore trace =d)*: $C_{ij} \in [-1, 1], \quad \lambda_i(C) \in [0, d], \quad \sum_i \lambda_i(C) = d$
- Transform a covariance matrix (with nonzero diagonal elements) into a correlation matrix

$$\boxed{C = D^{-\frac{1}{2}} \Sigma D^{-\frac{1}{2}}, \quad D = \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{dd})}$$

arXiv:1604.01592 arXiv:1704.00454

Springer Proceedings in Mathematics & Statistics

Frédéric Barbaresco
Frank Nielsen *Editors*

Geometric Structures of Statistical Physics, Information Geometry, and Learning

SPIGL'20, Les Houches, France, July 27–31

Springer

Contents

Tribute to Jean-Marie Souriau Seminal Works

Structure des Systèmes Dynamiques Jean-Marie Souriau's Book 50th Birthday 3
Géry de Saxcé and Charles-Michel Marle

Jean-Marie Souriau's Symplectic Model of Statistical Physics: Seminal Papers on Lie Groups Thermodynamics - Quod Erat Demonstrandum 12
Frédéric Barbaresco

Lie Group Geometry and Diffeological Model of Statistical Physics and Information Geometry

Souriau-Casimir Lie Groups Thermodynamics and Machine Learning 53
Frédéric Barbaresco

An Exponential Family on the Upper Half Plane and Its Conjugate Prior 84
Koichi Tojo and Taro Yoshino

Wrapped Statistical Models on Manifolds: Motivations, The Case *SE* (*n*), and Generalization to Symmetric Spaces 96
Emmanuel Chevallier and Nicolas Guigui

Galilean Thermodynamics of Continua 107
Géry de Saxcé

Nonparametric Estimations and the Diffeological Fisher Metric 120
Hông Vân Lê and Alexey A. Tuzhilin

Advanced Geometrical Models of Statistical Manifolds in Information Geometry

Information Geometry and Integrable Hamiltonian Systems 141
J.-P. Françoise

Relevant Differential Topology in Statistical Manifolds 154
Michel Nguiffo-Boyom

A Lecture About the Use of Orlicz Spaces in Information Geometry . . . 179
Giovanni Pistone

Quasiconvex Jensen Divergences and Quasiconvex Bregman Divergences 196
Frank Nielsen and Gaëtan Hadjeres

Geometric Structures of Mechanics, Thermodynamics and Inference for Learning

Dirac Structures and Variational Formulation of Thermodynamics for Open Systems 221
Hiroaki Yoshimura and François Gay-Balmaz

The Geometry of Some Thermodynamic Systems 247
Alexandre Anahory Simoes, David Martin de Diego, Manuel Lainz Valcázar, and Manuel de León

Learning Physics from Data: A Thermodynamic Interpretation 276
Francisco Chinesta, Elias Cueto, Miroslav Grmela, Beatriz Moya, Michal Pavelka, and Martin Šípka

Computational Dynamics of Reduced Coupled Multibody-Fluid System in Lie Group Setting 298
Zdravko Terze, Viktor Pandža, Marijan Andrić, and Dario Zlatar

Material Modeling via Thermodynamics-Based Artificial Neural Networks 308
Filippo Masi, Ioannis Stefanou, Paolo Vannucci, and Victor Maffi-Berthier

Information Geometry and Quantum Fields 330
Kevin T. Grosvenor

Hamiltonian Monte Carlo, HMC Sampling and Learning on Manifolds

Geometric Integration of Measure-Preserving Flows for Sampling 345
Alessandro Barp

Bayesian Inference on Local Distributions of Functions and Multidimensional Curves with Spherical HMC Sampling 356
Anis Fradi, Ines Adouani, and Chafik Samir

Sampling and Statistical Physics via Symmetry 374
Steve Huntsman

A Practical Hands-on for Learning Graph Data Communities on Manifolds 428
Thomas Gerald, Hadi Zaatiti, and Hatem Hajri

Schoenberg-Rao distances:

Entropy-based and geometry-aware Hilbert distances

I. J. Schoenberg

C. R. Rao

Conditionally Negative Semi-Definite kernel (CNSD):

(e.g., squared Euclidean kernel) $\sum_{i=1}^n \sum_{j=1}^n c_i c_j d(x_i, x_j) \leq 0, \quad \sum_{i=1}^n c_i = 0$

Rao's quadratic entropy of a distribution: $d_{ij} := d(x_i, x_j)$

(concave for **CNSD kernels**) $H_d(p) := \sum_{i=1}^n \sum_{j=1}^n p_i p_j d_{ij}$

Schoenberg-Rao (pseudo-)divergence for distributions: $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$

(Jensen divergence for Rao quadratic entropy)

$$\boxed{\mathrm{SR}_d(p,q) := \sum_{i,j=1}^n p_i q_j d_{ij} - \frac{1}{2}\sum_{i,j=1}^n p_i p_j d_{ij} - \frac{1}{2}\sum_{i,j=1}^n q_i q_j d_{ij} \geq 0}$$

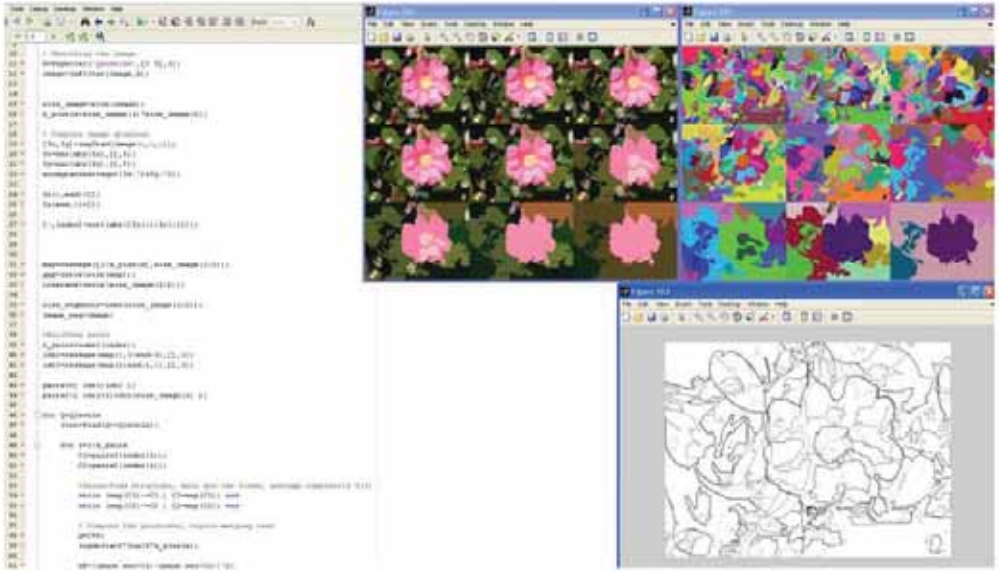
Schoenberg's embedding theorem (1930's): **Finite Hilbert embedding when d is CSND**

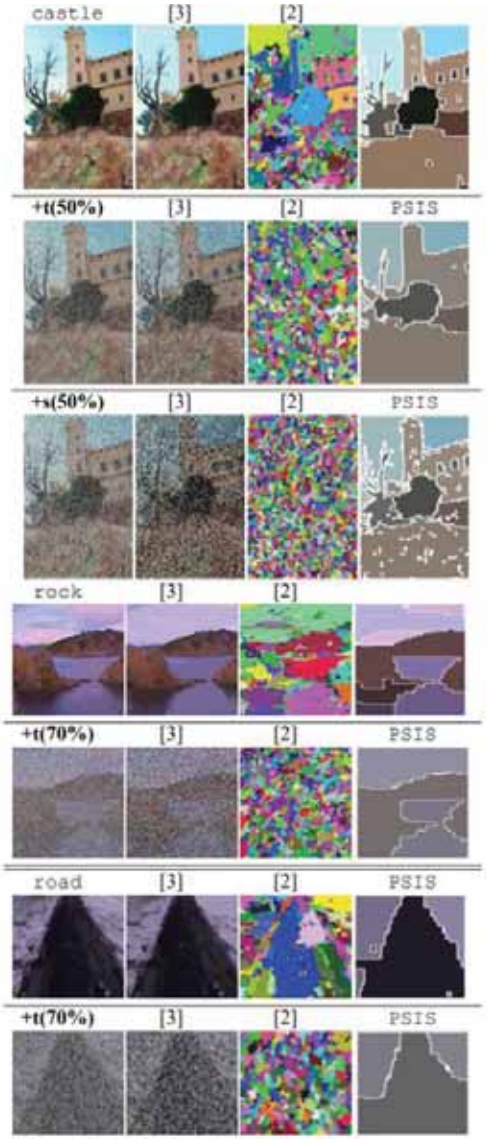
SR may potentially be an improper divergence: $\exists q \neq p, \mathrm{SR}_d(p,q) = 0$

(check proper/improper on CSND kernel/distribution families)

Statistical Region Merging (SRM)

- Simple+efficient image segmentation algorithm
- Robust to noise, many extensions (hierarchical, etc.)
- Single parameter Q to tune for region granularity
- Implementations online: C++, Java, Matlab, Python, …





On region merging: The statistical soundness of fast sorting, with applications, IEEE CVPR 2003
Statistical Region Merging, IEEE TPAMI 2004

Special Symmetric Positive-Definite Matrices (SSPD)

- SSPD(d,v) is the set of dxd dimensional symmetric positive-definite matrices with **prescribed determinant** v (= special). SSPD(d,v) are **totally geodesic submanifolds** of the cone manifold of SPD matrices
- Foliations and Riemannian product manifolds (de Rham decompositions):
  - **SSPD(2,v) is isometric to 2D hyperbolic geometry:** $\boxed{\mathrm{SPD}(2) = \mathrm{SSPD}(2,1) \times \mathbb{R}_{++}}$

$$\boxed{\rho_{\mathrm{SSPD}(2,1)}(P_1, P_2) = \sqrt{2}\rho_{H(2)}(z_1, z_2)} \qquad \rho_{H(2)}(z_1, z_2) = 2\mathrm{arctanh}\left(\frac{|z_1 - z_2|}{|1 - \bar{z}_1 z_2|}\right)$$

$$P = \begin{bmatrix} a & c \\ b & c \end{bmatrix}, ab - c^2 = 1 \mapsto z(P) := \frac{a-b}{2+a+b} + i\frac{c}{2+a+b} \qquad \mathrm{arctanh}(x) = \frac{1}{2}\log\frac{1+x}{1-x}$$

  - **SSPD(d,1):** $\boxed{\mathrm{SSPD}(d,1) \cong \mathrm{SL}(d)/\mathrm{SO}(d)}$

(irreducible symmetric space)

Hyperbolic Voronoi diagrams made easy, IEEE ICCSA 2010

Fast (1+ε)-Approximation of the Löwner Extremal Matrices of High-Dimensional Symmetric Matrices, Computational Information Geometry: For Image and Signal Processing, 2017. arxiv 1604.01592

Special symmetric positive-definite 2x2 matrices ($SSPD_2$) isometric to Klein hyperbolic geometry with κ=-1/2

- The open cone SPD_d of positive-definite dxd matrices equipped with the Riemannian trace metric yields **a scale-invariant** Riemannian distance:

$$\rho_{\mathrm{SPD}}(P_1, P_2) = \sqrt{\sum_{i=1}^{d} \log^2 \lambda_i(P_1 P_2^{-1})} \qquad \forall \lambda > 0, \quad \rho_{\mathrm{SPD}}(\lambda P_1, \lambda P_2) = \rho_{\mathrm{SPD}}(P_1, P_2)$$

- SPD_2 can be **foliated** by constant determinant matrix spaces ($SSPD_2$) . Each $SSPD_2(r)$: **totally geodesic submanifold of constant negative curvature κ=-1/2**. Riemannian SSPD2 distance computed from Klein hyperbolic distance:

$$\mathrm{SPD}_2 = \mathrm{SSPD}_2(r) \times \mathbb{R}_{++} \qquad \mathrm{SSPD}_2(r) = \left\{ \begin{bmatrix} a & c \\ c & b \end{bmatrix}, a > 0, ab - c^2 = r \right\}$$

$$\boxed{\forall P_1, P_2 \in \mathrm{SSPD}_2(r), \quad \rho_{\mathrm{SSPD}_2}(P_1, P_2) = \sqrt{2}\, \rho_K \left(k\left(\frac{P_1}{\sqrt{|P_1|}}\right), k\left(\frac{P_2}{\sqrt{|P_2|}}\right) \right)}$$

arXiv:1210.8234

$$\rho_K(k_1, k_2) = \operatorname{arccosh}\left(\frac{1 - k_1 \cdot k_2}{\sqrt{1 - k_1 \cdot k_1}\sqrt{1 - k_2 \cdot k_2}} \right) \qquad k(P) = \left(\frac{a^2 + 2a - b^2 - 2b}{a^2 + 2a + b^2 + 2b + 2 + 2c^2}, \frac{2c(a + b + 2)}{a^2 + 2a + b^2 + 2b + 2 + 2c^2} \right)$$

Statistical distances between mixtures: Exact formula/approximations

- **The Statistical Minkowski Distances: Closed-Form Formula for Gaussian Mixture Models**, GSI 2019: 359-367

→ Use Minkowski Lp-inequalities difference or log ratio gap distances with closed-form for p even integers

- **Monte Carlo information-geometric structures**, Geometric Structures of Information (Springer), 2019 (arxiv:1803.07225) **On the Geometry of Mixtures of Prescribed Distributions**, IEEE ICASSP 2018

→ For prescribed components, KL divergence amount to an untractable Bregman divergence+stochastic geometry

- **Guaranteed Deterministic Bounds on the total variation Distance between univariate mixtures**, IEEE MLSP 2018
- **MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions**, IEEE SPL 2017

→ Entropy of a mixture is necessarily less than entropy of MaxEnt distributions with prescribed moments

- **Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities**, Entropy 18(12): 442 (2016) **Guaranteed Bounds on the Kullback-Leibler Divergence of Univariate Mixtures**, IEEE SPL 2016

→ Partition the mixture support into elementary intervals with extremal components, then use LSE inequalities

- **Combinatorial bounds on the α-divergence of univariate mixture models**, IEEE ICASSP 2017
- **Closed-form information-theoretic divergences for statistical mixtures**, IAPR ICPR 2012,

  On Hölder Projective Divergences, Entropy, 2017

→ Cauchy-Schwarz divergence, Hölder divergence, Jensen-Rényi divergences for mixtures of exponential families

Tangent space on a statistical manifold: ∂ & RV identification

- **Tangent plane** Tp at point p is a *local linearization* of the manifold M
- Tangent space Tp at point p on a D-dimensional manifold M is a **linear space** spanned by D linearly independent vectors $e_1, \ldots, e_D$ (basis) where **e_i is the tangent vector along the coordinate curve $\theta^i(.)$**
- This vector e_i is notationally written ∂_i: $\boxed{\partial_i :=: \frac{\partial}{\partial \theta^i}}$

$$v = \sum_{i=1}^{D} v^i e_i = \sum_{i=1}^{D} v^i \partial_i$$ or using Einstein summation convention $\boxed{v = v^i \partial_i}$

- Infinitesimal change from p to p' identified as a vector: $\mathrm{d}p = \overrightarrow{pp'} = \sum_{i=1}^{D} \mathrm{d}\theta^i e_i$ or $\boxed{\mathrm{d}p = \mathrm{d}\theta^i \partial_i}$
- **Tangent bundle** is the collection of all tangent planes, and is of dimension **2D**
- On a statistical manifold, **identify** a tangent vector with a **random variable**= directional derivative of the log-probability: $\boxed{v(x) = v^i \partial_i l_\theta(x)}$

Inner product becomes **expectation** $\boxed{\langle v_1, v_2 \rangle_\theta = g_{ij}(\theta) v_1^i v_2^j = E_{p_\theta}[v_1(x) v_2(x)]}$

Fisher Information metric $g_{ij}(\theta) := E_{p_\theta}[\partial_i l_\theta(x) \partial_j l_\theta(x)]$

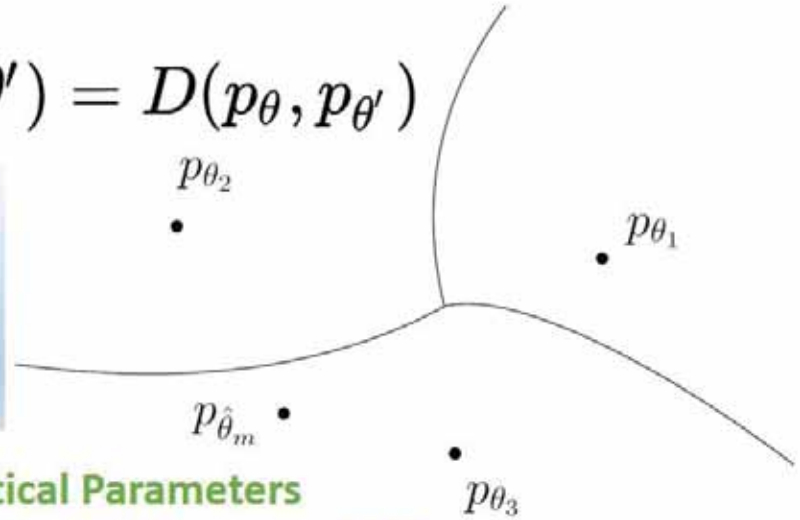
An Elementary Introduction to Information Geometry, Entropy 2020, 22(10), 1100

Statistical hypothesis testing as a classification problem

- Given m i.i.d. observations $x_1, \ldots,, x_m$, decide from which of the n given distributions $p_1, \ldots, p_n$ these observations emanate from
- Assume that all the distributions belong to a same **parametric family** like the Gaussian family, and consider a point set $\mathcal{P} = \{p_{\theta_1}, \ldots, p_{\theta_n}\}$ in the **parameter space**/**population space**
- Classify the **estimator** $\hat{\theta}_m$ with respect to a parameter/statistical "**distance**" : $D_{\text{param}}(\theta, \theta') = D(p_\theta, p_{\theta'})$

⇨ Statistical Voronoi diagrams

⇨ proximity queries



Rao, Information and the Accuracy Attainable in the Estimation of Statistical Parameters

N., Cramér-Rao lower bound and information geometry, Connected at Infinity II. Springer 2013

N., Hypothesis testing, information divergence and computational geometry, GSI 2013

Steepest descent in a Riemannian space

- The steepest descent direction of E(θ) in a Riemannian space is given by

$$-\tilde{\nabla} E(\theta) = -G^{-1}(\theta)\nabla E(\theta)$$

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

Contravariant form of the ordinary gradient

Learning rate

Computing the inverse of the Fisher information matrix is tricky

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

© Frank Nielsen

Stein's lemma for continuous exponential families

AC(R) : set of absolutely continuous functions on R.

Theorem *Let X be a continuous random variable with exponential density $p_\theta(x) = \exp(\sum_{i=1}^D \theta_i t_i(x) - F(\theta) + k(x))$ with support $\mathcal{X} = \mathbb{R}$. For any $f \in \mathrm{AC}(\mathbb{R})$ with $E[f'(X)] < \infty$, we have the following Stein identity:*

$$E_{p_\theta}\left[\left(\sum_{i=1}^D \theta_i t_i'(x)\theta_i + k'(x)\right) f(x)\right] = -E_{p_\theta}[f'(x)].$$

Example: Stein' lemma for Polynomial exponential families with $f(x) = x^j$:

$$p_\theta(x) = \frac{q_\theta(x)}{Z(\theta)} \quad q_\theta(x) := \exp\left(\sum_{i=1}^D \theta_i x^i\right) \qquad E\left[\sum_{i=1}^D i\theta_i x^{i-1+j}\right] = -E[jx^{j-1}]. \xrightarrow{\mu_i = E[x^i]} \boxed{\sum_{i=1}^D i\theta_i\mu_{i+j} = -j\mu_{j-1}.}$$

Hudson. A natural identity for exponential families with applications in multiparameter estimation. The Annals of Statistics, 6(3), 1978.
Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities, arXiv:2107.05901

Stochastic neurons, Boltzmann machines & Information projections

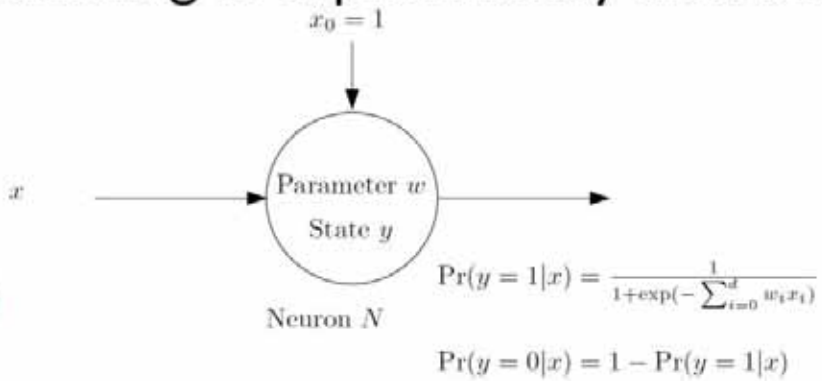
- A **stochastic neuron N** fires randomly according to a probability distribution:

Examples:

- **binary state neuron**

  using the *logistic function* (with x0 for bias)

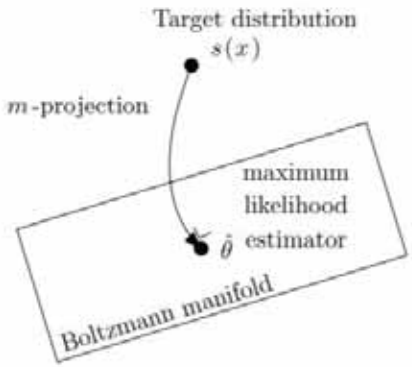
- **q-neuron with q-activation functions**



- A **Boltzmann machine (BM)** updates *asynchronously one-by-one* neuron states, and yields an *ergodic Markov process* on $\mathcal{X} = 2^d$ with exponential family stationary distribution (energy-based model):

Fully pairwise connected neurons
(graph is a clique)

$$p_W(x) = \exp\left(E(x) - F(W)\right)$$
$$E(x) = \sum_{i>j} w_{ij} x_i x_j$$



- Learning a BM by **information projections**

q-Neurons: Neuron Activations Based on Stochastic Jackson's Derivative Operators, IEEE TNNLS 2020
What is... an information projection?, Notices of the AMS 65.3 (2018): 321-324

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 Delaunay complex.

Matching image superpixels by Earth mover distance

Superpixels by image segmentation:

- Quickshift (mean shift)
- Statistical Region Merging (SRM)

Optimal transport between superpixels including topological constraints when a **segmentation tree** is available

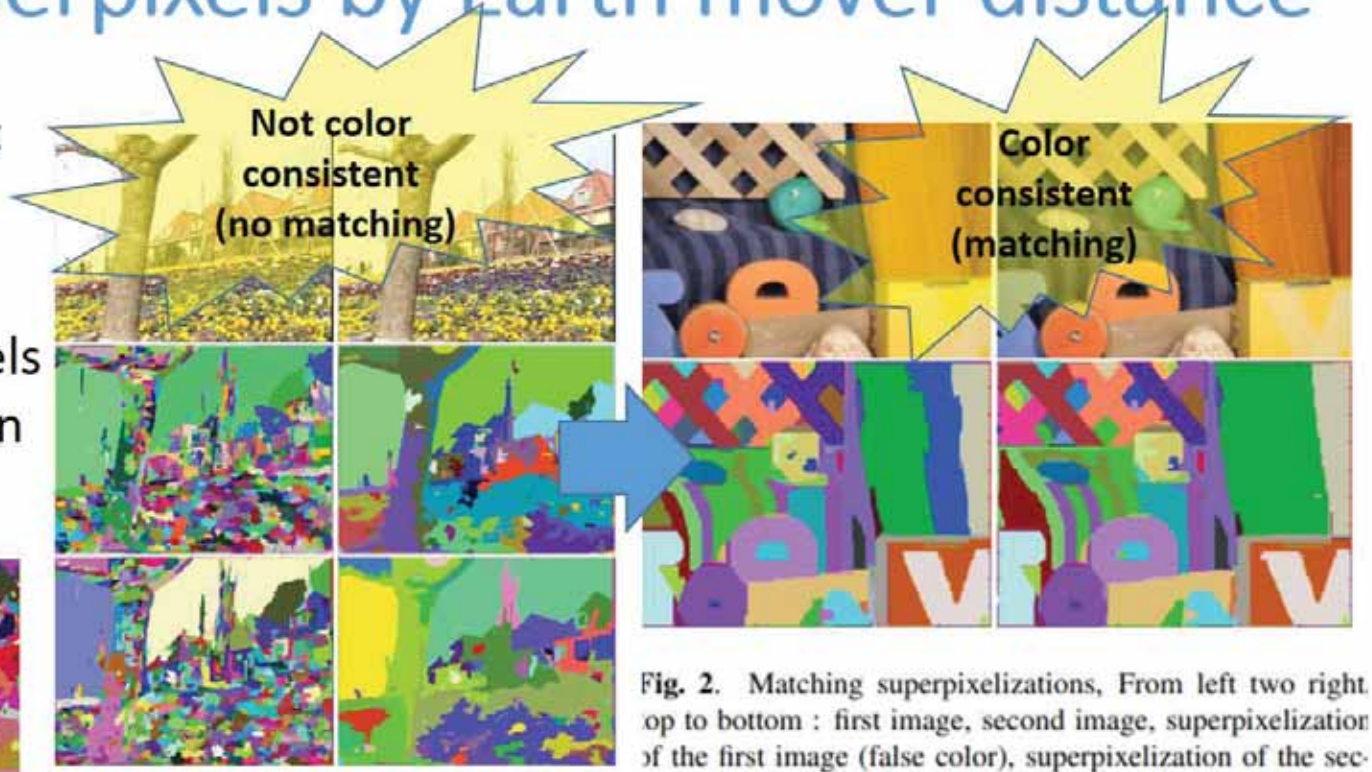


Fig. 1. Topological constraints for robust matching. From left to right, top to bottom: first image, second image, superpixelization of the first image at two different scales, superpixelization of the second image at two different scales. Toplogical constraints in EMD adds the cost of matching superpixelization at a coarse scale to the cost of matching superpixelization at a fine scale.

Fig. 2. Matching superpixelizations, From left two right, top to bottom : first image, second image, superpixelization of the first image (false color), superpixelization of the second image (false colors). Even between two images with small differences, the superpixelization, here in false colors, can be quite inconsistent. The matching of these two superpixel maps is in the color code: Superpixel i in image 2 have the color of the Superpixel in image 1 with label $j = \arg\max F(j, i)$.

@FrnkNlsn

Earth mover distance on superpixels, ICIP 2010

Which families yield all f-divergences symmetric?

- **f-divergences** for a convex generator f(u), strictly convex at 1, with f(1)=0:

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

- f-divergences between any pair of densities are symmetric when **f(u)=uf(1/u)** (ex: total variation with f(u)=1/2 |u-1|, Jensen-Shannon divergence, Hellinger)
- f-divergences on parametric families of densities = **parameter divergences**

$$\mathcal{F} := \{p_\theta \ : \ \theta \in \Theta\} \qquad\qquad I_f^{\mathcal{F}}(\theta_1 : \theta_2) := I_f(p_{\theta_1} : p_{\theta_2})$$

- **<u>Theorem I</u>: f-divergences between densities of a location family with even standard density are always symmetric** [arXiv:1904.10428]

- **<u>Theorem II</u>: f-divergences between Cauchy (location-scale) densities are always symmetric** [arXiv:2101.12459]

Skewed & reverse skewed versus symmetrized skewed Jensen-Shannon divergences

$$\begin{aligned}\mathrm{JS}_{\alpha,\beta}[p:q] &= (1-\beta)\mathrm{KL}[p:(1-\alpha)p+\alpha q]+\beta\mathrm{KL}[q:(1-\alpha)p+\alpha q]\\ &= (1-\beta)\mathrm{KL}_\alpha[p:q]+\beta\mathrm{KL}_{1-\alpha}[q:p]\end{aligned}$$

$$\mathrm{KL}[p:q]=\mathrm{KL}_1[p:q]$$

$$\mathrm{KL}_\alpha[p:q] := \mathrm{KL}[p:(1-\alpha)p+\alpha q] = \int p(x)\log\frac{p(x)}{(1-\alpha)p(x)+\alpha q(x)}\mathrm{d}\mu(x)$$

- Skewed Jensen-Shannon divergence: $\mathrm{JS}_\alpha[p:q]=\mathrm{JS}_{\alpha,\alpha}[p:q]$
- Jensen-Shannon divergence: $\mathrm{JS}[p:q]=\mathrm{JS}_{\frac{1}{2}}[p:q]$
- Reverse-skewed Jensen-Shannon divergence: $\mathrm{JS}^R_\alpha[p:q]=\mathrm{JS}_{\alpha,1-\alpha}[p:q]$
- Symmetrized-skewed Jensen-Shannon divergence:

$$\boxed{\mathrm{JS}^S_\alpha[p:q] = \frac{1}{2}\left(\mathrm{KL}_\alpha[p:q]+\mathrm{KL}_\alpha[q:p]\right)=\mathrm{JS}^S_\alpha[q:p]}$$

Jensen-Shannon divergence

$$\mathrm{JS}^S_{\frac{1}{2}}[p:q]=\mathrm{JS}[p:q]$$

$$\mathrm{JS}^S_1[p:q]=\frac{1}{2}\left(\mathrm{KL}[p:q]+\mathrm{KL}[q:p]\right)$$

Jeffreys divergence

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means, Entropy 2019
A family of statistical symmetric divergences based on Jensen's inequality, arXiv:1009.4004
On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid, Entropy 22.2 (2020)

Symplectic almost complex Riemannian manifolds

- **Symplectic geometry** born from classical mechanics (Lagrange and Poisson): **even-dimensional manifold equipped with a closed non-degenerate skew-symmetric 2-form ω** ("measuring oriented 2D areas").

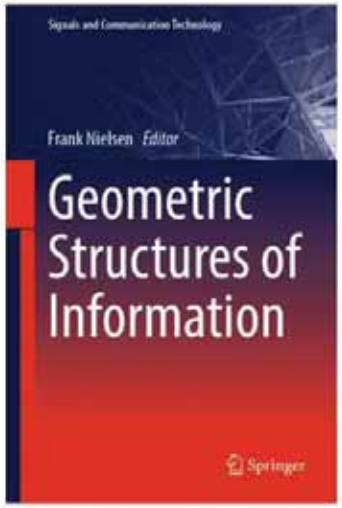
- **Almost complex structure** J: TM → TM such that J^2=-Id. Turn the tangent bundle TM as a **complex vector bundle**. Compatibility of J with symplectic form ω expressed by:

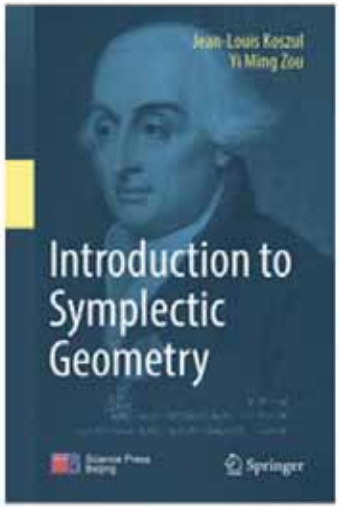
$$\omega(x,y)=\omega(Jx,Jy) \text{ and } \omega(x,Jx)>0 \text{ for all non-zero } x$$

- Build a **Riemannian metric tensor** from ω and J:

$$g(x,y)=\omega(x,Jy)$$

- **Complex Kahler manifold**: Build ω from (g,J): ω(x,y)= g(Jx,y)





Székely energy distance

- Generalize Cramér distance to **multivariate independent random variables**: $X_1, X_2 \sim_{\text{iid}} P \qquad Y_1, Y_2 \sim_{\text{iid}} Q$

$$\boxed{D^2_{\text{Székely}}[P,Q] := E\left[\|X_1 - Y_1\|_2\right] + E\left[\|X_2 - Y_2\|_2\right] - E\left[\|X_1 - X_2\|_2\right] - E\left[\|Y_1 - Y_2\|_2\right]}$$

Gábor J. Székely
1947-

- When distributions are univariate, coincide with twice **Cramér distance** :

$$D^2_{\text{Székely}}[P,Q] = 2D_{\text{Cramér}}[P,Q]$$

- Energy distance is a **metric distance** which can be extended to metric spaces other than Euclidean spaces
- Introduced in 1984 by Székely

Székely and Rizzo, The energy of data, Annual Review of Statistics and Its Application 4 (2017)

Taneja T-divergence between Cauchy distributions

- The **Taneja T-divergence** is a symmetric f-divergence: $D_T[p,q] = I_{f_T}[p:q]$

$$D_T[p,q] := \int \frac{p(x)+q(x)}{2} \log \frac{p(x)+q(x)}{2\sqrt{p(x)q(x)}} \mathrm{d}x. \qquad f_T(u) = \frac{u+1}{2} \log \frac{u+1}{2\sqrt{u}}$$

- Taneja T-divergence between Cauchy distributions is available in closed-form:

$$\boxed{D_T[p_{l_1,s_1}, p_{l_2,s_2}] = \log\left(\frac{1}{2}\left(1+\sqrt{\frac{(s_1+s_2)^2+(l_1-l_2)^2}{4s_1s_2}}\right)\right)}$$

$$D_T[p_{l_1,s_1}, p_{l_2,s_2}] = h_T(\chi[p_{l_1,s_1}, p_{l_2,s_2}])$$

$$h_T(u) = \log\left(\frac{1+\sqrt{1+\frac{u}{2}}}{2}\right)$$

Chi-squared divergence:

$$\chi[p_{l_1,s_1}, p_{l_2,s_2}] = \int \frac{(p_{l_1,s_1}(x) - p_{l_2,s_2}(x))^2}{p_{l_2,s_2}(x)} \mathrm{d}x = \frac{(l_2-l_1)^2+(s_2-s_1)^2)}{2s_1s_2}$$

On f-divergences between Cauchy distributions, arXiv 2101.12459 2021

Taneja T-divergence: Idendity with Jensen-Shannon and Jeffreys divergences

- The **Taneja T-divergence** is a symmetric f-divergence:

$$\boxed{D_T[p,q] := \int \frac{p(x)+q(x)}{2} \log \frac{p(x)+q(x)}{2\sqrt{p(x)q(x)}} \mathrm{d}x.}$$

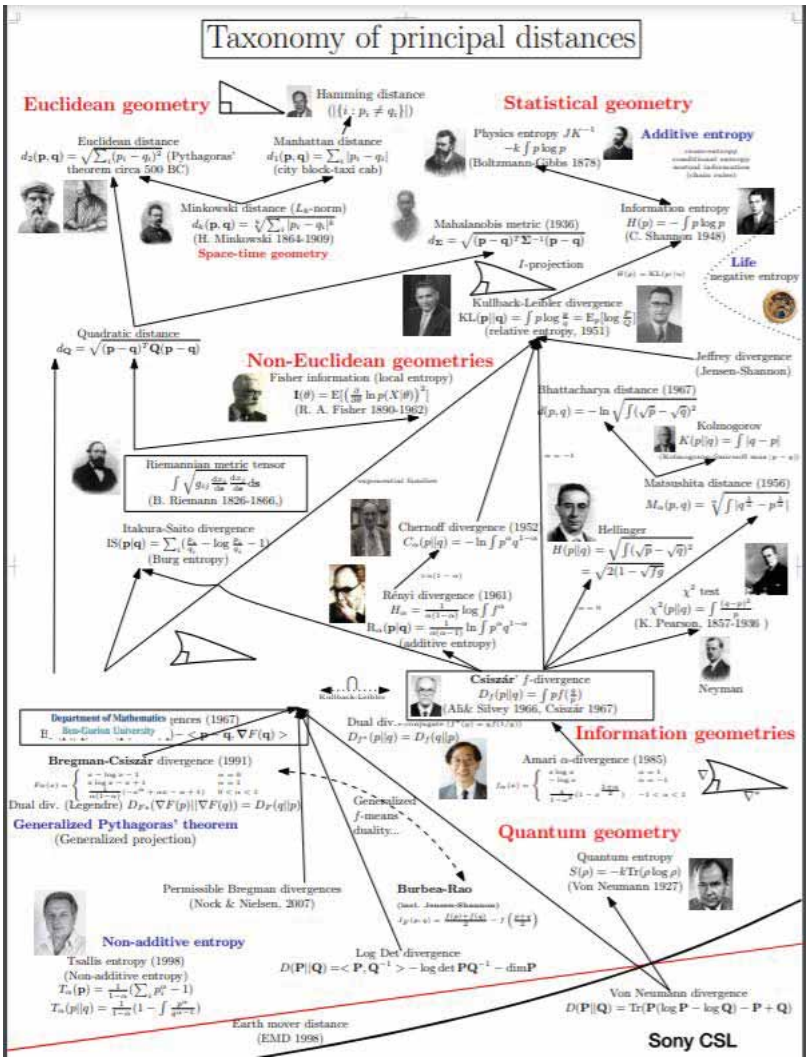
$$D_T[p,q] = I_{f_T}[p:q] \qquad f_T(u) = \frac{u+1}{2} \log \frac{u+1}{2\sqrt{u}}$$

- Taneja T-divergence is related to Jensen-Shannon divergence D_{JS} and Jeffreys divergence D_J as follows

$$\boxed{D_{\mathrm{JS}}(p,q) + D_T(p,q) = \frac{1}{4} D_J(p,q)}$$

Taneja, New developments in generalized information measures. In Advances in Imaging and Electron Physics, 1995.

Taxonomy of principal distances



A personal taxonomy of exponential families

- Exponential family[1] (EF) with log-density $\log p_\theta(x) = x^\top \theta - F(\theta)$. Log-normalizer F is real analytic convex function, **regular** when natural parameter space $\Theta = \{\theta : \int e^{x^\top \theta} \mathrm{d}\mu < \infty\}$ is open.
- Classification wrt natural parameter space Θ:
  - CEF: **Conic EF**[2] (Gaussian, Wishart, Laplacian, etc.) Some statistical distances between mixtures admit closed-form
  - AEF: **Affine EF** (IsoGaussian, Poisson, von Mises-Fisher, etc.), Closed-form formula for power chi-like pseudo-distances and f-divergences[3]
- Classification wrt sufficient statistics $t(x)$:
  - MEF: **Monomial EF** $t(x) = x^l$, closed-form MaxEnt distributions and entropy upper bounds for other distributions[4]
  - PEF: **Polynomial EF** $t(x) = (x^l, \ldots, x)$, Universal density approximations[5], projective distances

[1] https://arxiv.org/abs/0911.4863
[2] Closed-Form Information-Theoretic Divergences for Statistical Mixtures
[3] https://arxiv.org/abs/1903.05818
[4] https://arxiv.org/abs/1612.02954
[5] Patch Matching with Polynomial Exponential Families...

Thermodynamic Integration & path sampling

- Let $p(x,\theta) = \frac{q(x,\theta)}{Z(\theta)}$ where $q(x,\theta)$ is **unnormalized density** and $Z(\theta) = \int q(x,\theta)\mathrm{d}x$ the **normalizing constant** for $\theta \in \Theta$.
- Log-ratio of normalizing constants as the integral of an expectation of a **potential** $U(x,\theta) := \frac{d}{\mathrm{d}\theta} \log q(x,\theta)$:

$$\begin{aligned}
\frac{d}{\mathrm{d}\theta} \log Z(\theta) &= \frac{Z'(\theta)}{Z(\theta)} = \frac{1}{Z(\theta)} \int \frac{d}{\mathrm{d}\theta} q(x,\theta)\mathrm{d}x = \int \frac{q(x,\theta)}{Z(\theta)} \frac{\frac{d}{\mathrm{d}\theta} q(x,\theta)}{q(x,\theta)} \mathrm{d}x = E_{p(\theta)}\left[\frac{d}{\mathrm{d}\theta} \log q(x,\theta)\right] \\
\int_{\theta_0}^{\theta_1} E_{p(\theta)}[U(x,\theta)]\mathrm{d}\theta &= \log Z(\theta_1) - \log Z(\theta_0) = \log \frac{Z(\theta_1)}{Z(\theta_0)} = \lambda
\end{aligned}$$

- Estimate λ with n samples (x_i, θ_i) not necessarily independent: $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{U(x_i,\theta_i)}{p(\theta_i)}$ (in 1D). Minimum variance of the estimator in dD related to the <u>**Fisher-Rao geodesic between unnormalized densities**</u> (Gelman & Meng, 1998):

$$\boxed{\mathrm{var}(\hat{\lambda}) = \frac{1}{n}\left[\int_0^1 \left(\sum_{i,j=1}^{d} g_{ij}(\theta(t))\dot{\theta}_i(t)\dot{\theta}_j(t)\right) dt - \lambda^2\right]}$$

where $U_k(\omega,\theta) = \frac{\partial \log q(\omega,\theta)}{\partial \theta_k}$ and $g_{ij}(\theta) = E_\theta\left[U_i(\omega,\theta)U_j(\omega,\theta)\right]$.

Thompson metric: Fixed points in cones from contraction mappings

- A **closed cone** C⊂V in a vector space V: $\forall p, q \in C, \forall \lambda \geq 0, \quad p + \lambda q \in C$
- Pointed cone: $0 \in C$ Salient cone: $C \cap (-C) = \{0\}$
- **Partial order** on the vector space induced by the cone: $p \preceq_C q \Rightarrow q - p \in C$

  $p \in C \Rightarrow 0 \preceq_C p$
- **Thompson metric distance** on C:

$$\boxed{\rho_T(p,q) = \log \max\{M_C(p:q), M_C(q:p)\}}$$ where $M_C(p:q) = \inf_\lambda \{p \preceq_C \lambda q\}$

- Parts of a cone (equivalence classes): $C/\sim$ where $p \sim q \Leftrightarrow \rho_T(p,q) < \infty$

Example: Parts of orthant R3: interior of 3 faces and 3 edges (all excluding 0)

- **Thompson's fixed point theorem for a contraction mapping**:

Mapping f:V→V which *contracts* the distance $\rho_T(f(p), f(q)) < \rho_T(p,q)$ $\quad p \neq q$

Then there exists on **each part** a fixed point vector u which can be obtained by iterated mappings: Start from any point u0 of a part and iterate $u_{i+1} = f(u_i)$

$\lim_{n\to\infty} u_n = u$ where u is the **fixed point**: $\boxed{f(u) = u}$

A. C. Thompson. "On certain contraction mappings in a partially ordered vector space." *Proceedings of the AMS* 14.3 (1963)

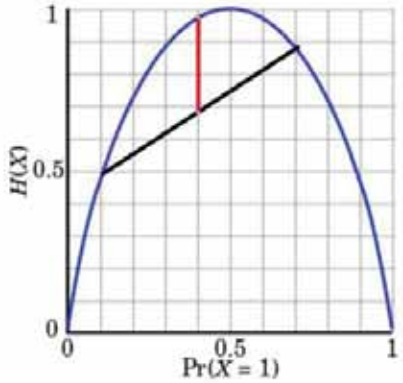
Three flavors of the Jensen-Shannon divergence

- Symmetrization of the relative entropy:

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

- Convexity gap of Shannon negentropy:

$$D_{\mathrm{JS}}[p,q] \quad := \quad h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2}$$



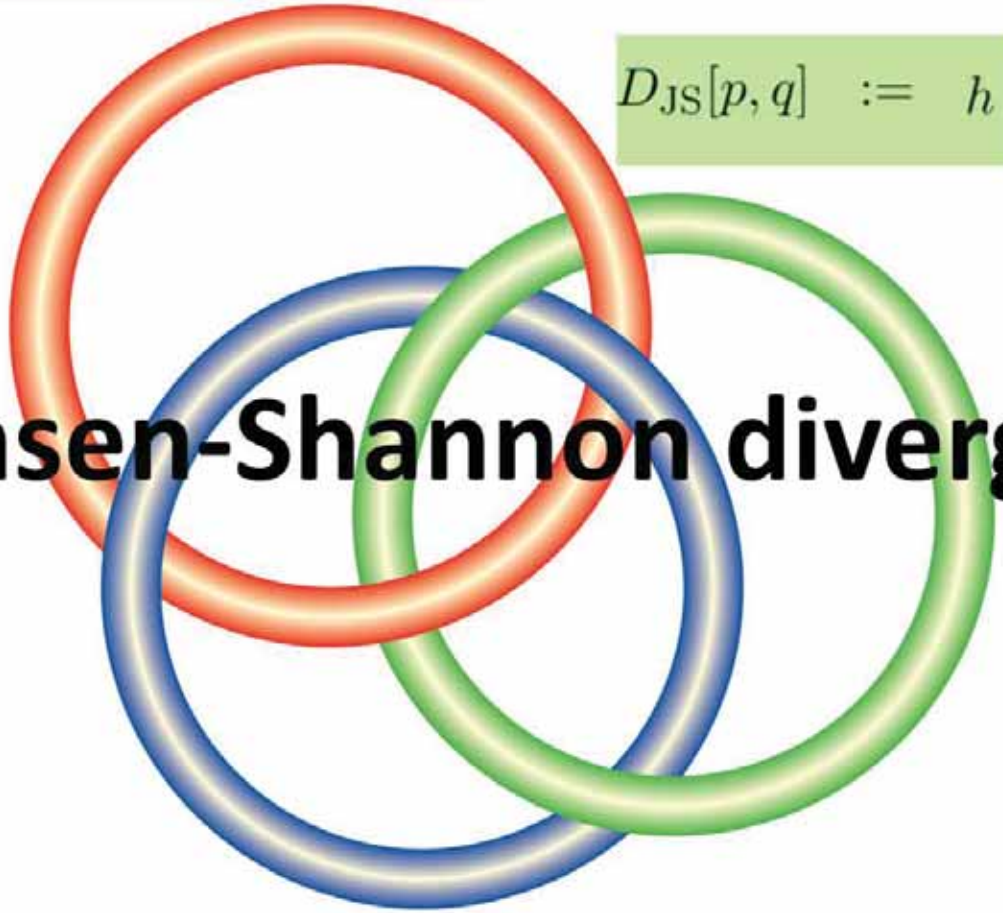
- Diversity index of two probability measures:

$$D_{\mathrm{JS}}[p,q] \quad := \quad \min_{c\in\mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right)$$

arXiv:2102.09728

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

$$D_{\mathrm{JS}}[p,q] \quad := \quad h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2}$$



Jensen-Shannon divergence

$$D_{\mathrm{JS}}[p,q] \quad := \quad \min_{c\in\mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right)$$

Total Bregman divergence and its applications to DTI analysis

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

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

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

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

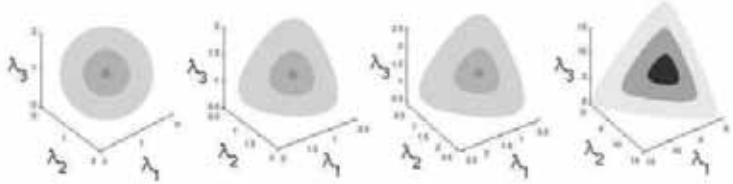
$$tKL(P,Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1 + \int (1 + \log q)^2 q dx}}$$

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

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

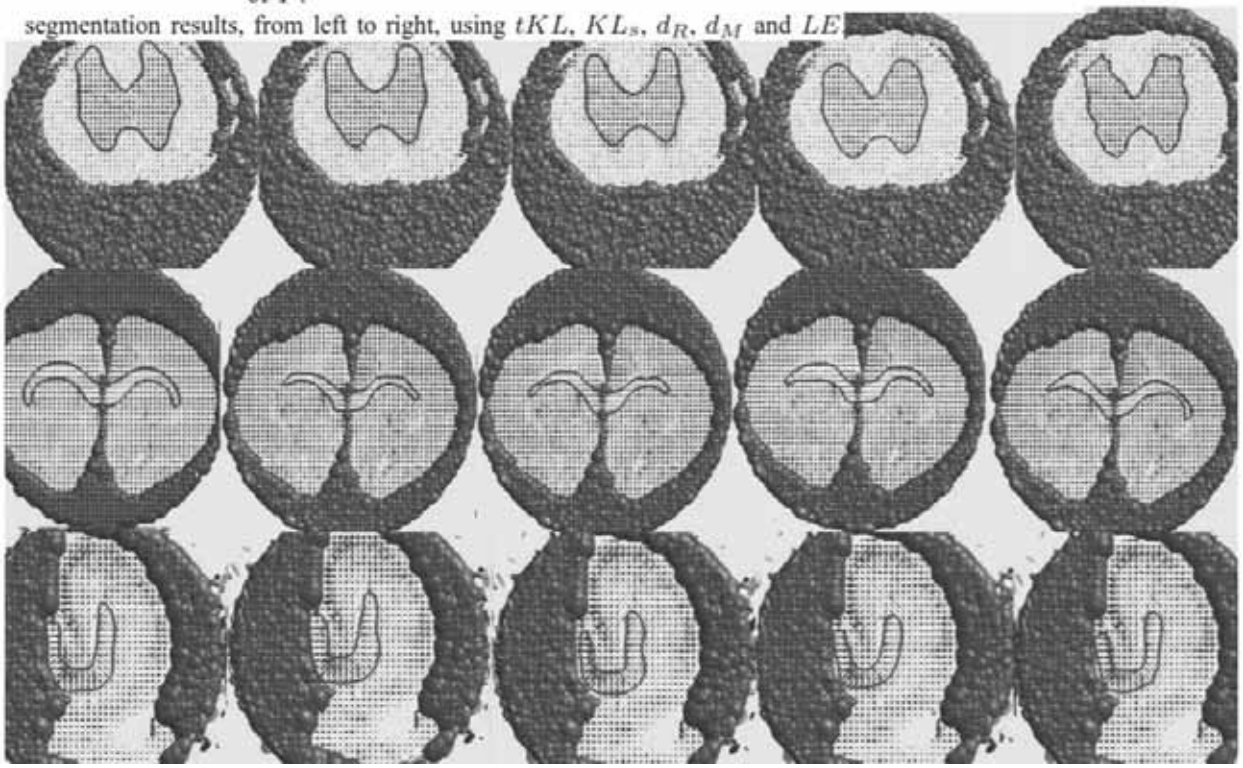
$$tSL(P,Q) = \frac{\int (p-q)^2 dx}{\sqrt{1 + \int (2q)^2 q dx}} =$$

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



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

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



@FrnkNlsn

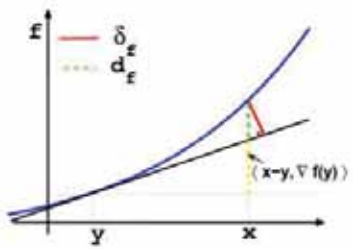
Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

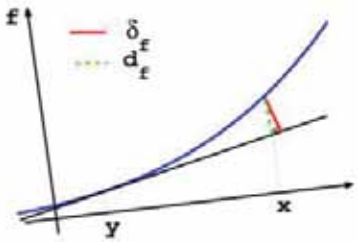
Definition *The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,*

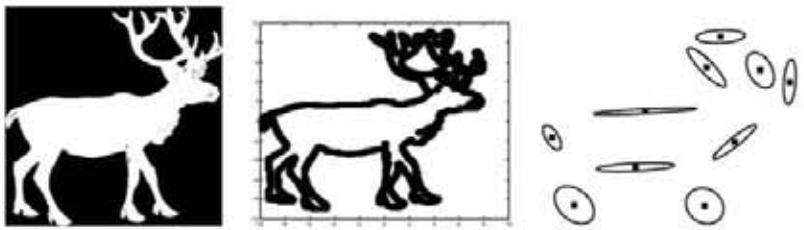
$$\delta_f(x,y) = \frac{f(x) - f(y) - \langle x-y, \nabla f(y)\rangle}{\sqrt{1+\|\nabla f(y)\|^2}},$$

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

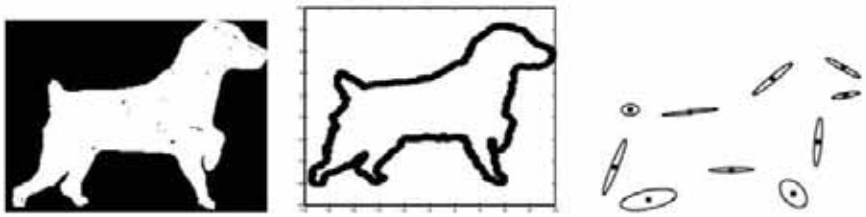
X	$f(x)$	$\delta_f(x,y)$	t-center	ℓ_1-norm BD center	Remark
$\mathbb{R}$	x^2	$\frac{(x-y)^2}{\sqrt{1+4y^2}}$	$\sum_i w_i x_i$	$\sum_i x_i$	total square loss (tSL)
$\mathbb{R}-\mathbb{R}_-$	$x\log x$	$\frac{x\log\frac{x}{y}+x\log\frac{x}{y}}{\sqrt{1+y(1+\log y)^2+y(1+\log y)^2}}$	$\prod_i (x_i)^{w_i}$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







(m) (n) (o)



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

@FrnkNlsn

IEEE TPAMI 34, 2012

Total variation distance between Cauchy densities

- Cauchy densities form a location-scale family with **even standard density** p(x)=p(-x):

$$p(x) := \frac{1}{\pi(1+x^2)} \qquad \mathcal{C} = \left\{ p_{l,s}(x) := \frac{1}{s} p\left(\frac{x-l}{s}\right) \ : \ (l,s) \in \mathbb{R} \times \mathbb{R}_{++} \right\}$$

- The **total variation distance** is the **unique metric f-divergence** (up to constant rescaling) :

$$D_{\mathrm{TV}}[p,q] = I_{f_{\mathrm{TV}}}[p:q] = \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| \mathrm{d}x$$

- The TVD between two Cauchy densities can be expressed using the **chi square divergence** (symmetric between Cauchy densities) :

$$h_{f_{\mathrm{TV}}}(u) = \frac{2}{\pi} \arctan\left(\sqrt{\frac{u}{2}}\right)$$

$$\boxed{\begin{aligned} D_{\mathrm{TV}}[p_{l_1,s_1}, p_{l_2,s_1}] &= \frac{2}{\pi} \arctan\left(\sqrt{\frac{\chi[p_{l_1,s_1}, p_{l_2,s_1}]}{2}}\right) \\ &= h_{f_{\mathrm{TV}}}\left(\chi[p_{l_1,s_1}, p_{l_2,s_1}]\right), \end{aligned}}$$

$$\begin{aligned} \chi(p_{l_1,s_1}, p_{l_2,s_2}) &= \int \frac{(p_{l_1,s_1}(x) - p_{l_2,s_2}(x))^2}{p_{l_2,s_2}(x)} \mathrm{d}x \\ &= \frac{(l_2 - l_1)^2 + (s_2 - s_1)^2)}{2 s_1 s_2} \end{aligned}$$

On f-divergences between Cauchy distributions, arXiv 2101.12459 2021

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)

Upper bounds for the f-divergence and the KL divergence

f-divergence, f convex, f(1)=0, strictly convex at 1:

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

When f'(1)=0 ⟹ $I_f(p:q) = \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) \mathrm{d}\mu(x)$

Bregman divergence: $B_f(a:b) = f(a) - f(b) - (a-b)f'(b)$

By analogy, let us define
for f strictly convex everywhere

$$\boxed{R_f(p:q) = \int p(x) B_f\left(1 : \frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x) \geq 0}$$

Expanding $R_f \geq 0$ yields

$$\boxed{I_f(p:q) \leq \int (p(x) - q(x)) f'\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)}$$

In particular, for f(u)=-log(u) ⟹ $\mathrm{KL}(p:q) \leq \int (p-q)\frac{p}{q}\mathrm{d}\mu = \int \frac{p^2}{q}\mathrm{d}\mu - 1$

On The Chain Rule Optimal Transport Distance, arXiv:1812.08113

Information geometry: Dually flat (u,v)- structures

- **Dually flat structure** = a pair of affine coordinate systems θ and η, a pair of dual potential functions F(θ) and F*(η) which defines a Legendre-Fenchel divergence (equivalent to dual Bregman divergences when using a single coordinate system, either θ or η)
- **Problem**: converting from/to θ and η may be difficult (not in closed form)
- **Solution**: Use a pair of increasing and differentiable **representation scalar functions** u and v to define a dually flat space on the discrete positive measure orthant: $m = (m_1, \ldots, m_D) \in \mathbb{R}^D$

$$\theta^i := u(m_i) \qquad \eta_i := v(m_i) \quad \Rightarrow \quad \boxed{\theta^i = u\left(v^{-1}(\eta_i)\right) \qquad \eta_i = v\left(u^{-1}(\theta^i)\right)}$$

$$f_{u,v}(\theta) := \int_0^{u^{-1}(\theta)} u'(m)v(m)\mathrm{d}m \quad \Rightarrow \quad F_{u,v}(\theta) := \sum_{i=1}^{D} f_{u,v}(\theta^i) \qquad g_{ij}(m) = \frac{v'(m_i)}{u'(m_i)}\delta_i^j \quad \Rightarrow \quad g = J^\top J$$

$$f^*_{u,v}(\eta) := \int_0^{v^{-1}(\eta)} v'(m)u(m)\mathrm{d}m \quad \Rightarrow \quad F^*_{u,v}(\eta) := \sum_{i=1}^{D} f^*_{u,v}(\eta_i) \qquad J = \mathrm{diag}\left(\sqrt{\frac{v'(m_1)}{u'(m_1)}}, \ldots, \sqrt{\frac{v'(m_D)}{u'(m_D)}}\right)$$

$$L_{F_{u,v}}(m, m') = F_{u,v}(\theta(m)) + F^*_{u,v}(\eta(m')) - \sum_{i=1}^{D} \theta(m_i)\eta(m'_i) \geq 0$$

$$\boxed{\gamma(m) = \int_0^m \sqrt{\frac{v'(m)}{u'(m)}}\mathrm{d}m \qquad g_\gamma(\gamma(m)) = I}$$

Euclidean metric!

- Includes the α-divergences, β-divergences, (α,β)-divergences, U-divergence

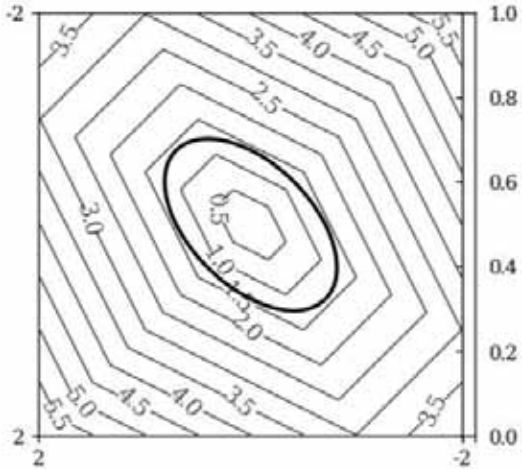
An Elementary Introduction to Information Geometry, Entropy 2020, 22(10), 1100

Geometry and Fixed-Rate Quantization in Riemannian Metric Spaces Induced by Separable Bregman Divergences, GSI 2019

On conformal divergences and their population minimizers, IEEE Information Theory, 2015

Variation norm and Hilbert projective distances

- **Variation norm**: $\boxed{\|x\|_{\mathrm{var}} := \max_{i\in[d]}\{x_i\} - \min_{i\in[d]}\{x_i\}}$



- Hilbert projective distance on the **probability simplex**:

$$D_{\mathrm{Hilbert}}(p,q) = \|\log\frac{p}{q}\|_{\mathrm{var}} = \log\frac{\max_i \frac{p_i}{q_i}}{\min_j \frac{p_j}{q_j}}$$

- Hilbert projective distance between two **correlation matrices**:

$$D_{\mathrm{Correlation}}(C_1,C_2) = \|\log\lambda(C_1^{-1}C_2)\|_{\mathrm{var}} := \log\frac{\lambda_{\max}(C_1^{-1}C_2)}{\lambda_{\min}(C_1^{-1}C_2)}$$

- Nielsen and Sun, Clustering in Hilbert simplex geometry, arXiv:1704.00454
- Nielsen and Sun, Clustering in Hilbert's projective geometry: The case studies of the probability simplex, Geom. Struct. Inf (2018): 297-331.

Sibson's information radius & Jensen-Shannon divergence

Sibson's obtained the **n-point Jensen-Shannon divergence** in 1969:

$$K_1\begin{pmatrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{pmatrix} = \int_X \sum_i \left\{ \frac{w_i p_i}{\sum_j w_j} \log_2 \frac{p_i \sum_j w_j}{\sum_j w_j p_j} \right\} d\lambda;$$

Robin Sibson
1944-2017

<u>Three definitions of the Jensen-Shannon divergence:</u>

Variational definition: $D_{\mathrm{JS}}[p,q] := \min_{c \in \mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right),$

Total KLD to the mid distribution: $= \frac{1}{2}\left(D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q : \frac{p+q}{2}\right]\right)$

Jensen divergence for Shannon Negentropy (=Jensen-Shannon) $= h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2}.$

Sibson, Robin. "Information radius." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* (1969)
[Generalization: arXiv 2102.09728]

Plücker geometry and coordinates of lines

- Plücker considered the space whose space elements (=points) are lines!
- 3D lines form a **6D vector space** (Plücker coordinates, special case of Grassmann coordinates), signed matrix minors

Julius Plücker
(1801–1868)

Two 3D
Homogeneous points

$$\mathbf{L}_{12} = [\mathbf{p}_1\ \mathbf{p}_2] = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \\ z_1 & z_2 \\ w_1 & w_2 \end{bmatrix}$$

$$\begin{aligned}
l_{41} &= -\det\begin{bmatrix} x_1 & x_2 \\ w_1 & w_2 \end{bmatrix} &&= x_2 w_1 - x_1 w_2 = x_2 - x_1, \\
l_{42} &= \det\begin{bmatrix} y_1 & y_2 \\ w_1 & w_2 \end{bmatrix} &&= y_2 w_1 - y_1 w_2 = y_2 - y_1, \\
l_{43} &= -\det\begin{bmatrix} z_1 & z_2 \\ w_1 & w_2 \end{bmatrix} &&= z_2 w_1 - z_1 w_2 = z_2 - z_1, \\
l_{23} &= -\det\begin{bmatrix} y_1 & y_2 \\ z_1 & z_2 \end{bmatrix} &&= y_1 z_2 - z_1 y_2, \\
l_{31} &= \det\begin{bmatrix} z_1 & z_2 \\ x_1 & x_2 \end{bmatrix} &&= x_2 z_1 - x_1 z_2, \\
l_{12} &= -\det\begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} &&= x_1 y_2 - x_2 y_1.
\end{aligned}$$

$$\mathbf{l} = \begin{bmatrix} l_{41} \\ l_{42} \\ l_{43} \\ l_{23} \\ l_{31} \\ l_{12} \end{bmatrix}$$

Plücker
6D coordinates

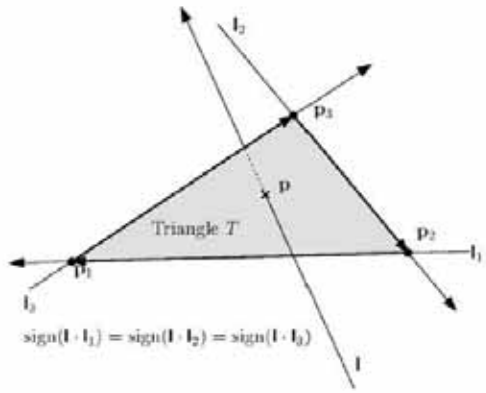
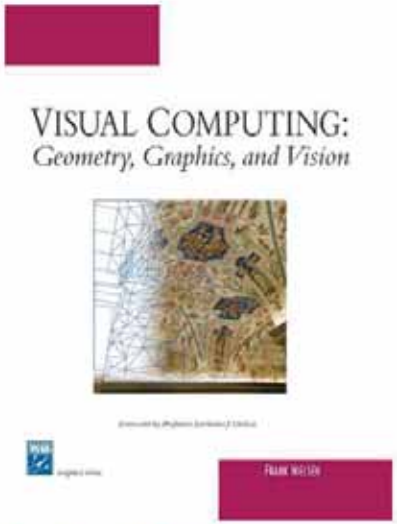


FIGURE 3.53 *Using Plücker coordinates to check whether a 3D line* **l** *intersects with a 3D triangle T.*

Easy line-triangle
intersection code



Visual computing: Geometry, graphics, and vision. **Hingham: Charles River Media, 2005.**

Vector-skew Jensen-Shannon divergence[1]

- For $\alpha \in [0,1]^k$ and unit vector $w \in \Delta_k$ (open standard simplex), $\bar{\alpha} = \sum_{i=1}^k w_i \alpha_i$, define the (α, w)-Jensen-Shannon divergence:

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

  where $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$ is Shannon's entropy.
- Generalize the ordinary Jensen-Shannon divergence
- Apply to densities with potentially different supports and is upper bounded by $0 \leqslant \mathrm{JS}^{\alpha,w}(p:q) \leqslant \log \frac{1}{\bar{\alpha}(1-\bar{\alpha})}$
- Compute the Jensen-Shannon centroids using the concave-convex procedure (CCCP)

[1] http://arxiv.org/abs/1912.00610

Summary: Vector-skewed Jensen-Shannon divergence

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

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

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

- 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 program** and solved using the **Convex-ConCave Procedure** (CCCP)

Video stippling/video pointillism (CG)

Figure *We extract two density maps from the original source image: one is the color map and the other the frequency map. Then we apply our stippling algorithm and finally add both contrast and color information. We end the process by summing these two contributions — the classical (3000 points) and frequency approaches (6000 points).*

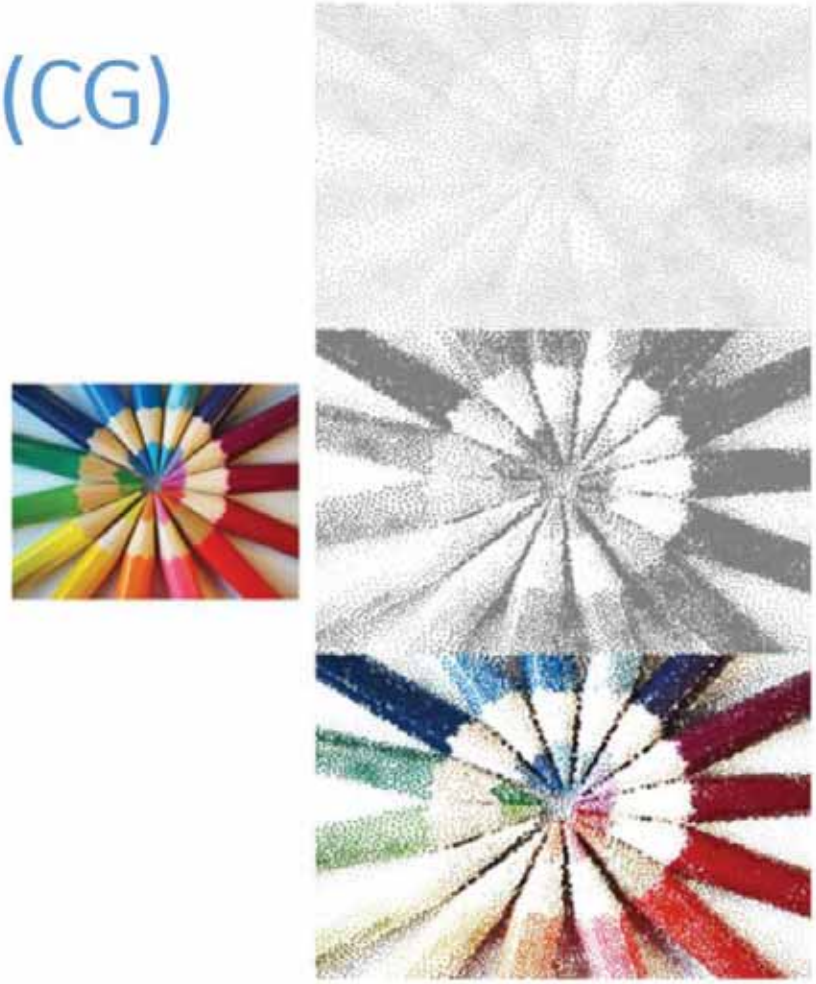


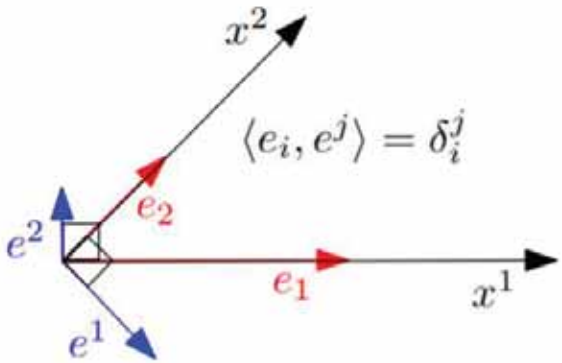
Figure *(Top) Stippling result without contrast and color. (Middle) stippling with contrast. (Bottom) stippling with color to enhance the rendering. (3000 points)*

Video

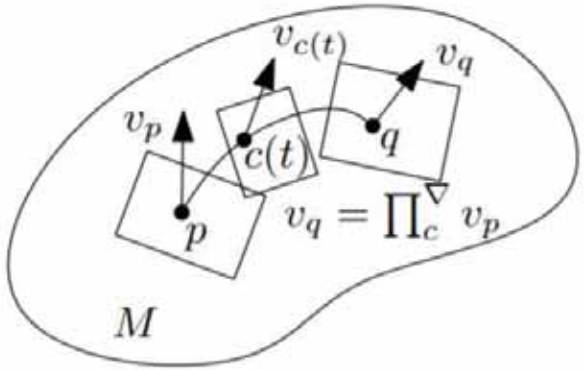
https://www.youtube.com/watch?v=O97MrPsISNk

@FrnkNlsn

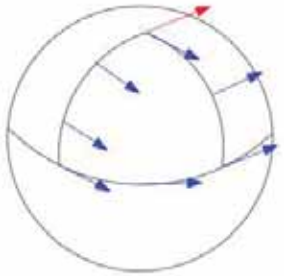
Video stippling, ACIVS 2011. arXiv:1011.6049

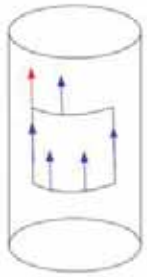


Reciprocal basis

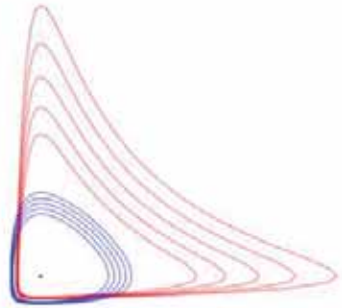


Dual parallel transport

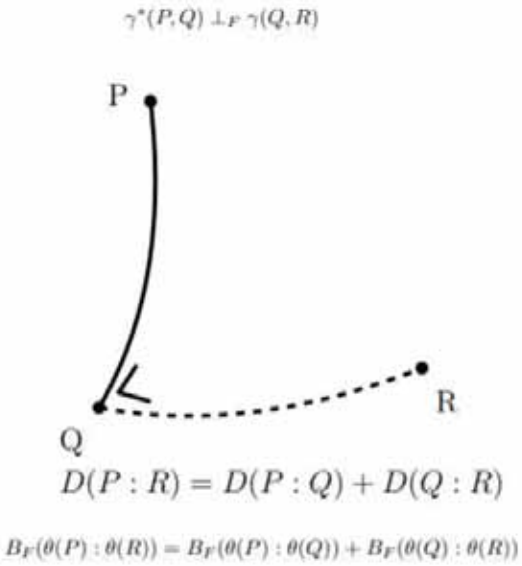




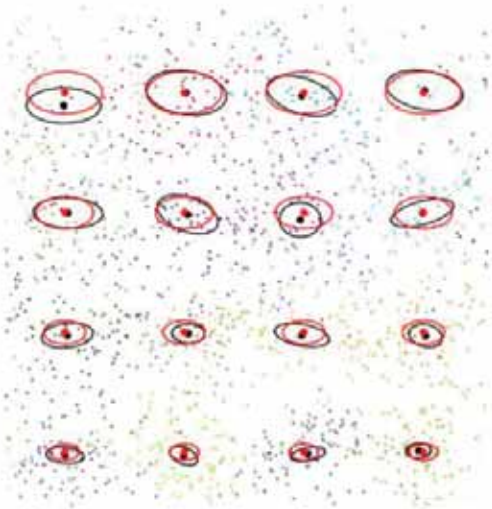
Loop PT and curvature



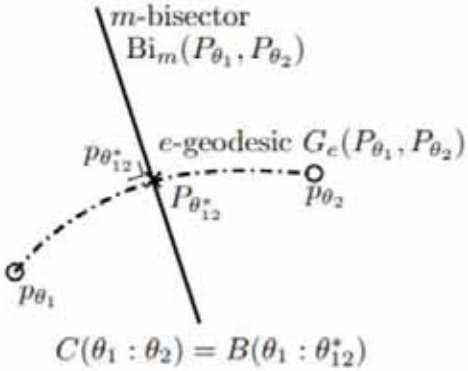
Bregman spheres



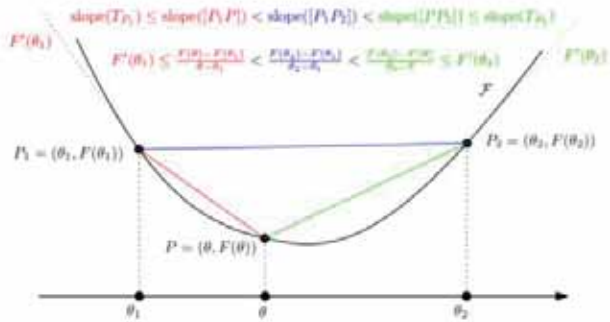
Pythagoras theorem
(Dually flat space)



Cramér–Rao lower bound
Fisher efficiency



e/m-geodesic
e/m flat



chordal slope
& Bregman divergence

Vector-skewed Jensen-Shannon divergences and centroids

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

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

$$\begin{aligned} \mathrm{JS}(p,q) &:= \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right), \\ &= \frac{1}{2}\int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}\right) \mathrm{d}\mu = \mathrm{JS}(q,p). \end{aligned}$$

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

$$\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$$

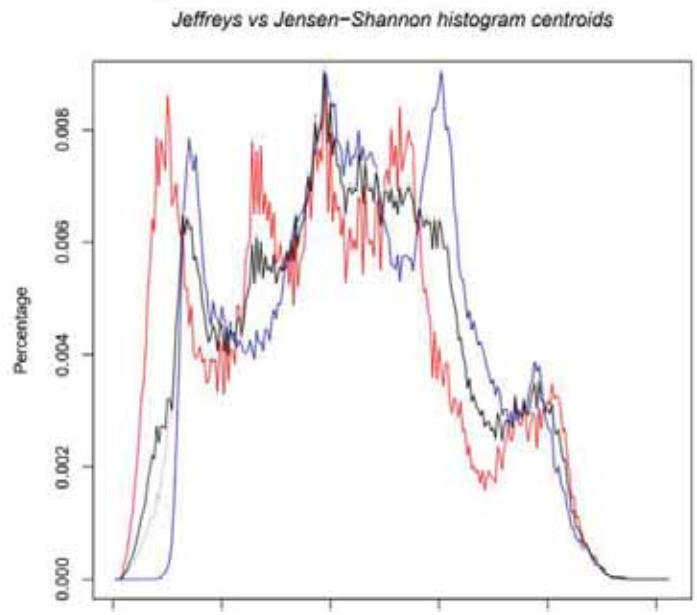
Jensen-Shannon centroids via CCCP:

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

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

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

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

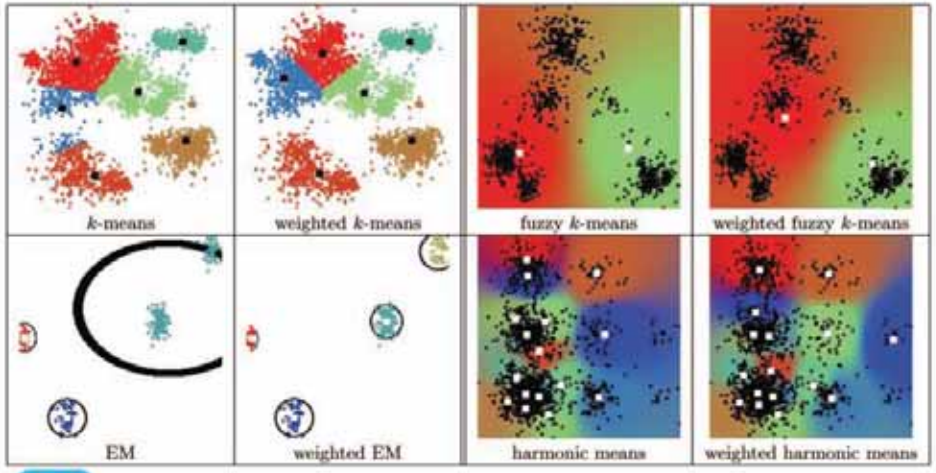


On a generalization of the Jensen-Shannon divergence
https://arxiv.org/abs/1912.00610

On weighting clustering: A generic boosting-inspired framework

Algorithm 1. `General Iterative Clustering` $\mathcal{X}, k$
Input: point set $\mathcal{X}$, integer $k > 0$
Initialize the models $\mathbf{p}_0^j, \forall 1 \leq j \leq k$ (plus eventual related parameters)
for $t = 0, 1, \ldots$ **do**
[1.] *Compute/update* the memberships:
$m_t(j,i), \forall 1 \leq i \leq n, \forall 1 \leq j \leq k;$
[2.] *Compute/update* the weights: $w_{t,i}, \forall 1 \leq i \leq n;$
[3.] *Update* the models: $\mathbf{p}_{t+1}^j, \forall 1 \leq j \leq k$ (plus eventual related parameters);

Algorithm	k-means [13]	Fuzzy k-means [14]	(Gaussian) EM [15]	Harmonic means [9]
Model parameters	Gaussians $\mathcal{N}(\boldsymbol{\mu}_j, \sigma I)$	Centers $\boldsymbol{\mu}_j$	Gaussians $\mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$	Centers $\boldsymbol{\mu}_j$
Loss:	$\ell_t(\mathbf{u})$ in eq. (3)			
	$\ell_{t,i}$: see eq. (4)	$\ell_{t,i}$: see eq. (5)	$\ell_{t,i}$: see eq. (4)	$\ell_{t,i}$: see eq. (6)
Membership	hard	soft	soft	soft
Step [1.]: $m_t(j,i) =$	1 for $\arg\min_{j'} \parallel \mathbf{x}_i - \boldsymbol{\mu}_{t,j'} \parallel^2$ 0 otherwise;	$1/\sum_{l=1}^k \lambda_{t,i,j,l}^{\left(\frac{2}{a-1}\right)}$; $\lambda_{t,i,j,l} = \frac{\parallel \mathbf{x}_i - \boldsymbol{\mu}_{t,j} \parallel}{\parallel \mathbf{x}_i - \boldsymbol{\mu}_{t,l} \parallel}$	$\lambda_{t,i,j}/\sum_{l=1}^k \lambda_{t,i,l}$; $\lambda_{t,i,j} = \pi_{t,j} p_{t,i}^j$	$\lambda_{t,i,j}/\sum_{l=1}^k \lambda_{t,i,l}$; $\lambda_{t,i,j} = \frac{1}{\parallel \mathbf{x}_i - \boldsymbol{\mu}_{t,j} \parallel^{a+2}}$
Step [2.]: $w_{t,i} =$	$1/n$			$\frac{\sum_{j=1}^k \lambda_{t,i,j}^{-2}}{z(\sum_{j=1}^k \lambda_{t,i,j})^2}$ $\lambda_{t,i,j} = \frac{1}{\parallel \mathbf{x}_i - \boldsymbol{\mu}_{t,j} \parallel^a}$
Step [3.]: $\boldsymbol{\mu}_{t+1,j} =$	see (7)	see (8)	see (7)	
Step [3.] (misc.)			$\boldsymbol{\pi}_{t+1}$, $\Sigma_{t+1,j}$: see eqs (9) and (10)	
Misc.		$a \in [1, +\infty)$		$a \in [2, +\infty)$ z=normalization coeff.



- Penalization for making clustering moving toward the hardest points to cluster: analogy with **boosting**
- Add weights to points, update based on the local variations of the expected complete log-likelihoods
- Clustering as a constrained minimization of a Bregman divergence

@FrnkNlsn

IEEE TPAMI 2006