Some generalizations of Bregman divergences:

Duality, geometry, and algorithms

Frank Nielsen

Sony Computer Science Laboratories Inc
Tokyo, Japan

Sony CSL

14 June 2024

Bregman divergence (1965, 1967)

- Let $F: \Theta \subseteq \mathbb{R}^m \rightarrow \mathbb{R}$ be a strictly convex and smooth real-valued function
- **Bregman divergence** $B_F: \Theta \times \mathrm{Int}(\Theta) \rightarrow \mathbb{R}$ defined by

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

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

- Non-negative because exact remainder of Taylor 1st order expansion:

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

= exact Lagrange remainder is a quadratic distance:

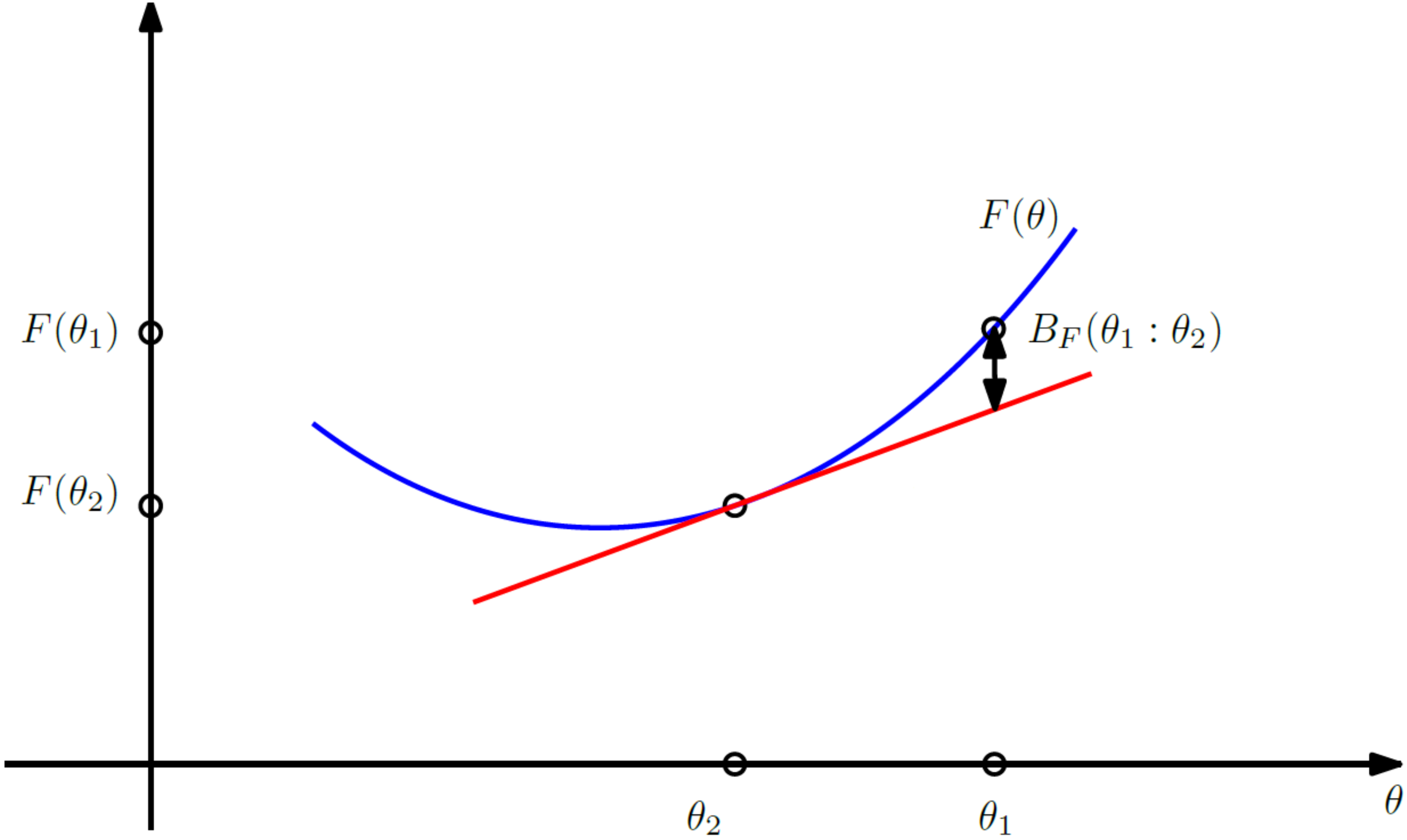
$L_F(\theta_1 : \theta_2) = \frac{1}{2} (\theta_1 - \theta_2)^\top \nabla^2 F(\xi) (\theta_1 - \theta_2) = B_F(\theta_1 : \theta_2), \ \xi \in [\theta_1 \ \theta_2]$

- BDs are never a metric and only symmetric for generalized quadratic distance $F_Q(\theta) = \frac{1}{2} \theta^\top Q \theta$ with symmetric positive-definite matrix Q

Q=I, squared Euclidean distance

- Unifies squared Euclidean divergence with Kullback-Leibler divergence $F(\theta) = \Sigma_i \theta_i \log(\theta_i)$ and Itakura-Saito divergence $F(\theta) = \Sigma_i -\log(\theta_i)$

Visualizing Bregman divergences: Function graph



Univariate: $B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)\, F'(\theta_2)$

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

Bregman divergences in machine learning

- Kullback-Leibler divergence between two probability densities:

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

In general, difficult to calculate in closed form because of the integral $\int$

- But the Kullback-Leibler divergence (relative entropy) between two probability densities of an exponential family (eg: Gaussians, Poisson, Dirichlet, Gamma/Beta, Wishart)

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

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

$$D_{KL}[p(x|\theta_1) : p(x|\theta_2)] = B_F{}^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$$ ⇒ Easy calculations

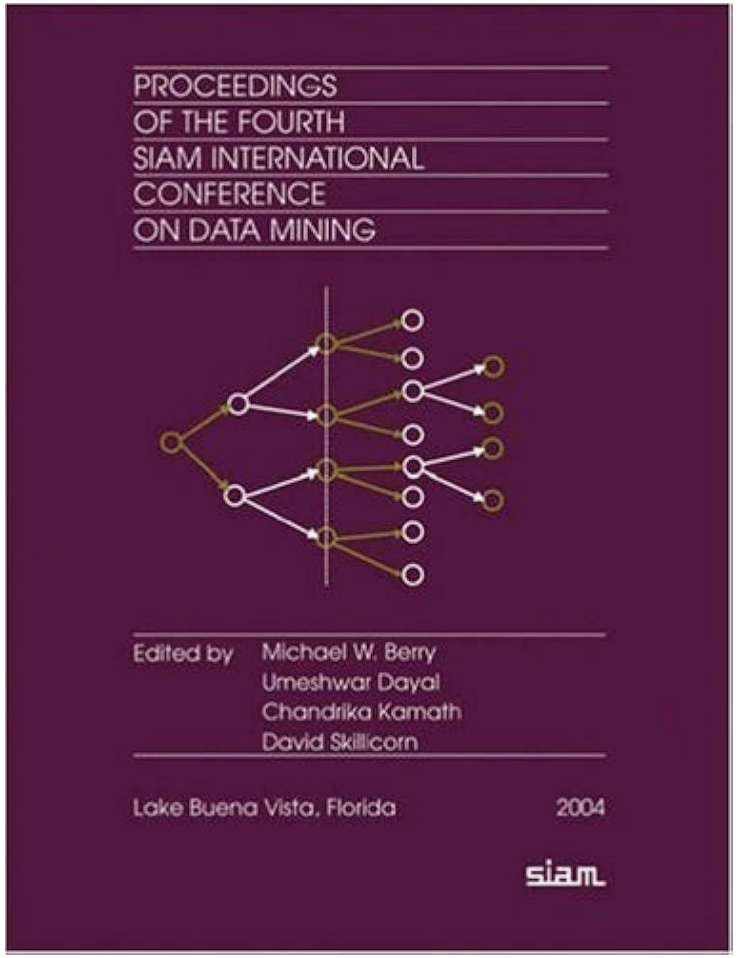
- Information geometry proves (⇐) that $B_F(\theta_1 : \theta_2) = D_{KL}{}^*[p(x|\theta_1) : p(x|\theta_2)]$

when $F(\theta) = \log \int \exp(<x, \theta>)\, d\mu(x)$ where $D_{KL}{}^*[p(x):q(x)] := D_{KL}[q(x):p(x)]$ is the **reverse Kullback-Leibler divergence**

- Notice divergence between parameters B_F vs divergence between functions KL

Azoury, Katy S., and Manfred K. Warmuth. "Relative loss bounds for on-line density estimation with the exponential family of distributions." *Machine learning* 43 (2001): 211-246.

My first encounter with Bregman divergences

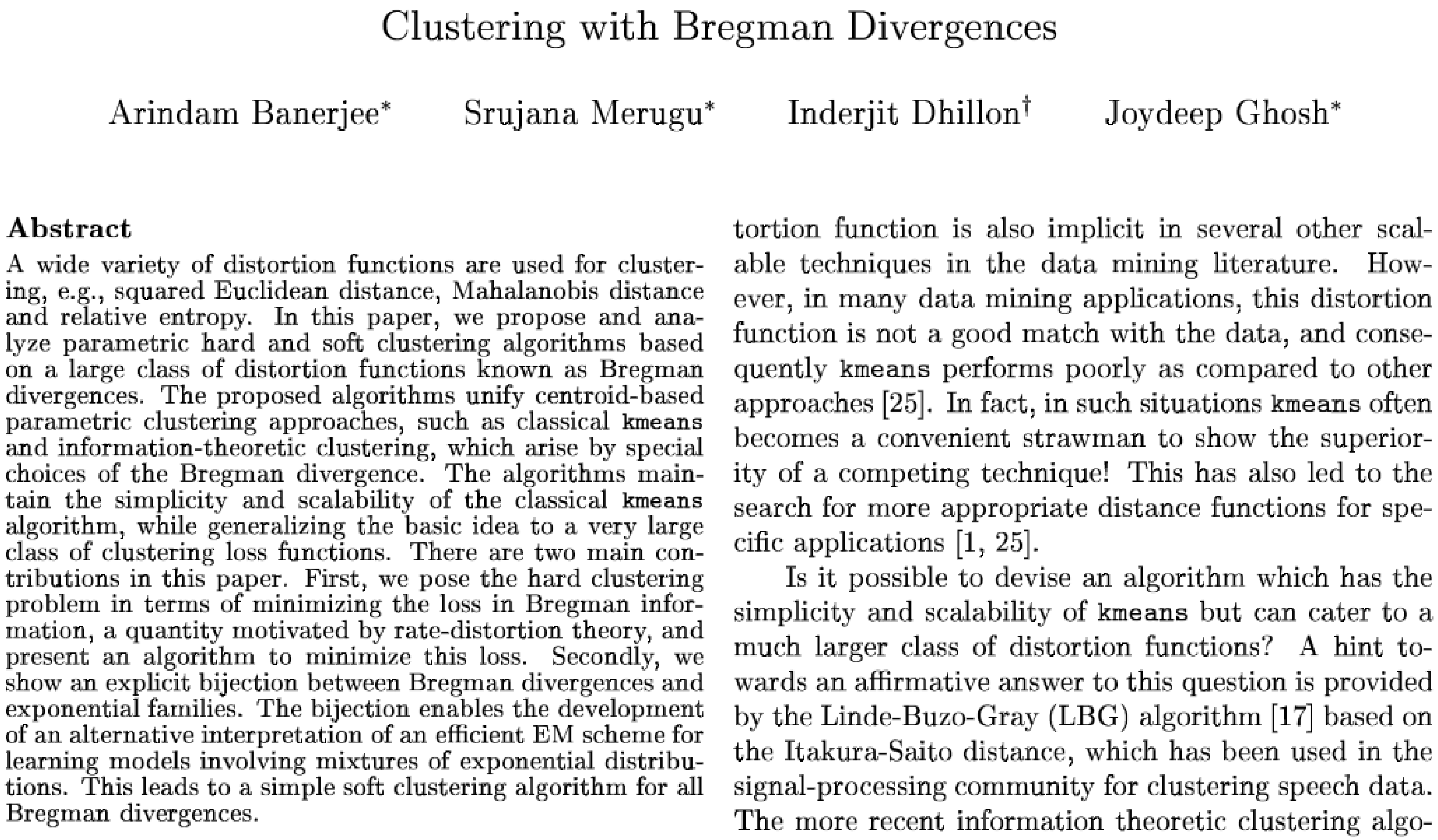


PROCEEDINGS
OF THE FOURTH
SIAM INTERNATIONAL
CONFERENCE
ON DATA MINING

Edited by Michael W. Berry
Umeshwar Dayal
Chandrika Kamath
David Skillicorn

Lake Buena Vista, Florida 2004

siam



Clustering with Bregman Divergences

Arindam Banerjee* Srujana Merugu* Inderjit Dhillon† Joydeep Ghosh*

Abstract

A wide variety of distortion functions are used for clustering, e.g., squared Euclidean distance, Mahalanobis distance and relative entropy. In this paper, we propose and analyze parametric hard and soft clustering algorithms based on a large class of distortion functions known as Bregman divergences. The proposed algorithms unify centroid-based parametric clustering approaches, such as classical `kmeans` and information-theoretic clustering, which arise by special choices of the Bregman divergence. The algorithms maintain the simplicity and scalability of the classical `kmeans` algorithm, while generalizing the basic idea to a very large class of clustering loss functions. There are two main contributions in this paper. First, we pose the hard clustering problem in terms of minimizing the loss in Bregman information, a quantity motivated by rate-distortion theory, and present an algorithm to minimize this loss. Secondly, we show an explicit bijection between Bregman divergences and exponential families. The bijection enables the development of an alternative interpretation of an efficient EM scheme for learning models involving mixtures of exponential distributions. This leads to a simple soft clustering algorithm for all Bregman divergences.

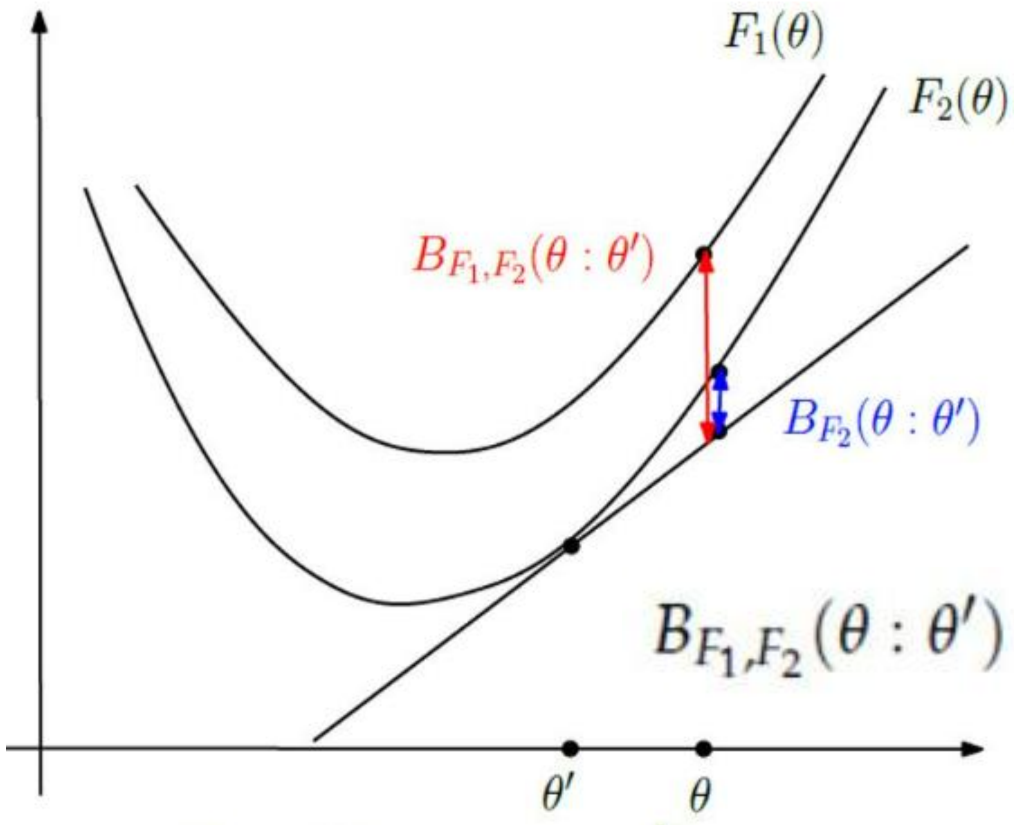
tortion function is also implicit in several other scalable techniques in the data mining literature. However, in many data mining applications, this distortion function is not a good match with the data, and consequently `kmeans` performs poorly as compared to other approaches [25]. In fact, in such situations `kmeans` often becomes a convenient strawman to show the superiority of a competing technique! This has also led to the search for more appropriate distance functions for specific applications [1, 25].

Is it possible to devise an algorithm which has the simplicity and scalability of `kmeans` but can cater to a much larger class of distortion functions? A hint towards an affirmative answer to this question is provided by the Linde-Buzo-Gray (LBG) algorithm [17] based on the Itakura-Saito distance, which has been used in the signal-processing community for clustering speech data. The more recent information theoretic clustering algo-

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh: Clustering with Bregman Divergences

SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004.

Duo BDs: Generalize BDs with a pair of generators

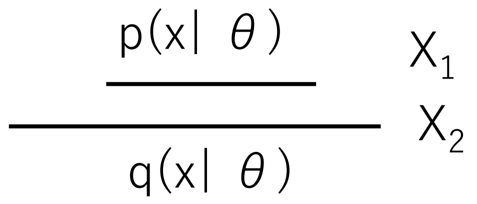


$F_1(\theta) \geq F_2(\theta)$

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

- Recover Bregman divergence when **$F_1(\theta) = F_2(\theta) = F(\theta)$**
- Only **pseudo-divergence** because $B_{F1,F2}(\theta : \theta)$ may be positive

KLD between nested exponential families amount to duo Bregman divergences



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

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

- Consider an exponential family on support X_1:

$$p(x|\ \theta) = \exp(<x,\ \theta>-F_1(\theta))\ d\mu(x)$$

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

- Truncated exponential family with **nested support $X_1 \subseteq X_2$**

$q(x|\ \theta)$» $p(x|\ \theta)$

$$q(x|\ \theta) = \exp(<x,\ \theta>-F_2(\theta))\ d\mu(x)$$

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

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

$$D_{KL}[p(x|\ \theta_1) : q(x|\ \theta_2)]= B_{F2F1}{}^*(\theta_{1:}\ \theta_2)=B_{F2F1}(\theta_2 : \theta_1)$$

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

Clustering distributions with different supports

- Consider n truncated densities $p(x| \theta_i) = \exp(<x, \theta_i >-F_i(\theta_i))\, d\mu(x)$ with supports $X_i \subseteq X$: e.g., Zipf's distributions

$q(x| \lambda_j)» p(x| \theta_i)$

- Cluster those distributions using full support X prototypes: e.g., Zeta distributions

- Objective is "k-means": minimize $\Sigma_i D_{KL}[p(x|\theta_i) : \{q(x|\lambda_j) : j \text{ in } \{1,\cdots,k\} \}]$

- Apply duo Bregman k-means algorithm

- Example: Cluster Zipf's distributions of words in a collection of translations of a famous book

arXiv:2104.10548

Convex duality via Legendre-Fenchel transform

- Legendre-Fenchel transform of a convex function:

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

- Consider **Legendre-type functions $(\Theta,F(\theta))$** : Θ open, and

  $$\lim_{\theta\to\partial\Theta}\|\nabla F(\theta)\|=\infty$$

- Reciprocal gradient maps $\eta=\nabla F(\theta)$ and $\theta=\nabla F^*(\eta)$

- $(H,F^*(\eta))$ is of Legendre type and biconjugation is involution: $(H,F^*(\eta))^*=(\Theta,F(\theta))$

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

Solo & duo Fenchel-Young divergences

- Young inequality: $F(\theta_1)+F^*(\eta_2)-<\theta_1, \eta_2>\geq 0$ with equality when $\eta_2=\nabla F(\theta_1)$
- Solo Fenchel-Young divergence:

$$\mathbf{Y_{F,\,F^*(}\boldsymbol{\theta}_1,\ \boldsymbol{\eta}_2)= F(\boldsymbol{\theta}_1)+F^*(\boldsymbol{\eta}_2)-<\boldsymbol{\theta}_1,\ \boldsymbol{\eta}_2>}$$

- Legendre transform reverse order:

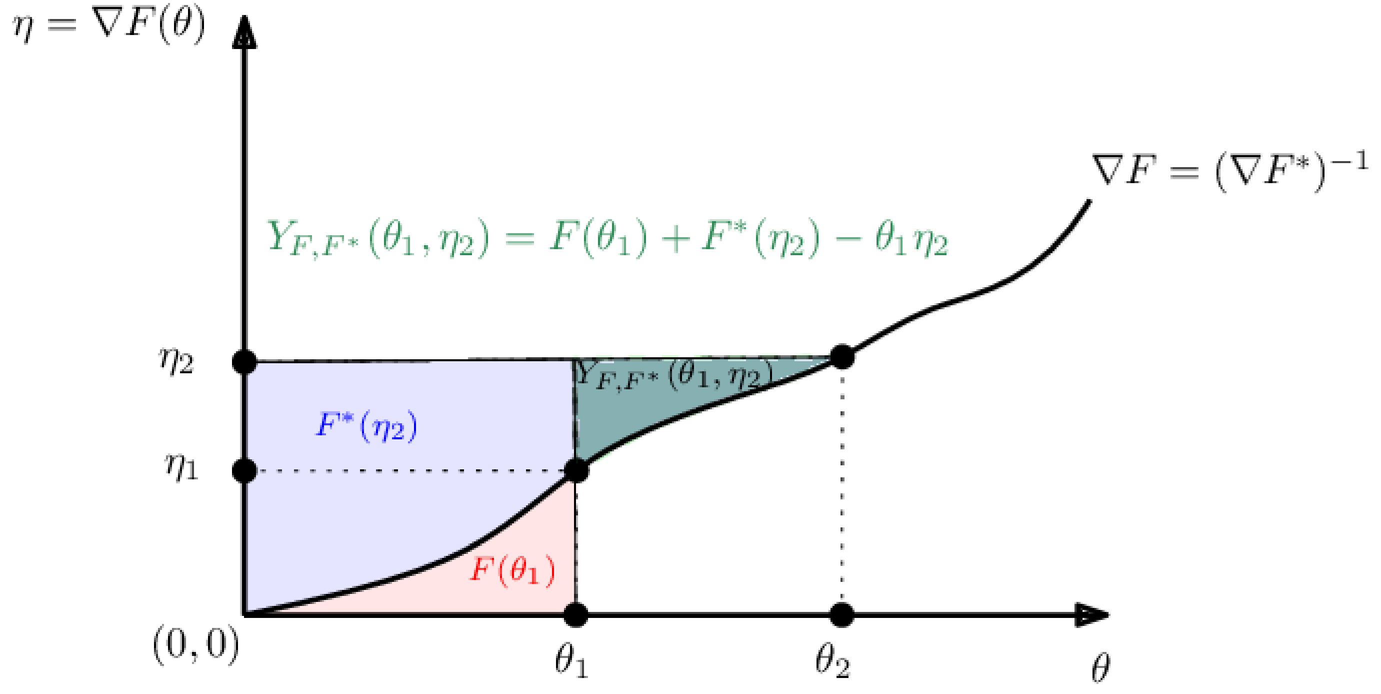
$$F_1(\theta)\geq F_2(\theta) \Leftrightarrow F_1^{*(}\eta)\leq F_2{}^*(\eta)$$

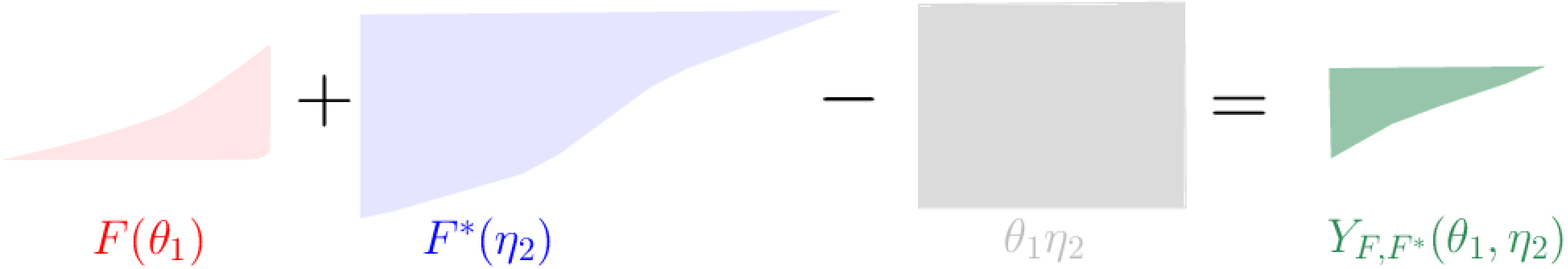
- Duo Fenchel-Young divergence:

$$\begin{aligned} Y_{F_1,F_2^*}(\theta,\eta') &:= F_1(\theta)+F_2^*(\eta')-\theta^\top\eta', \\ &\geq F_1(\theta)+F_1^*(\eta')-\theta^\top\eta' = Y_{F_1,F_1^*}(\theta,\eta')\geq 0 \end{aligned}$$

Amari, Shun-ichi. *Differential-geometrical methods in statistics*. Vol. 28. Springer, 1985

Visual interpretation of Fenchel-Young divergence

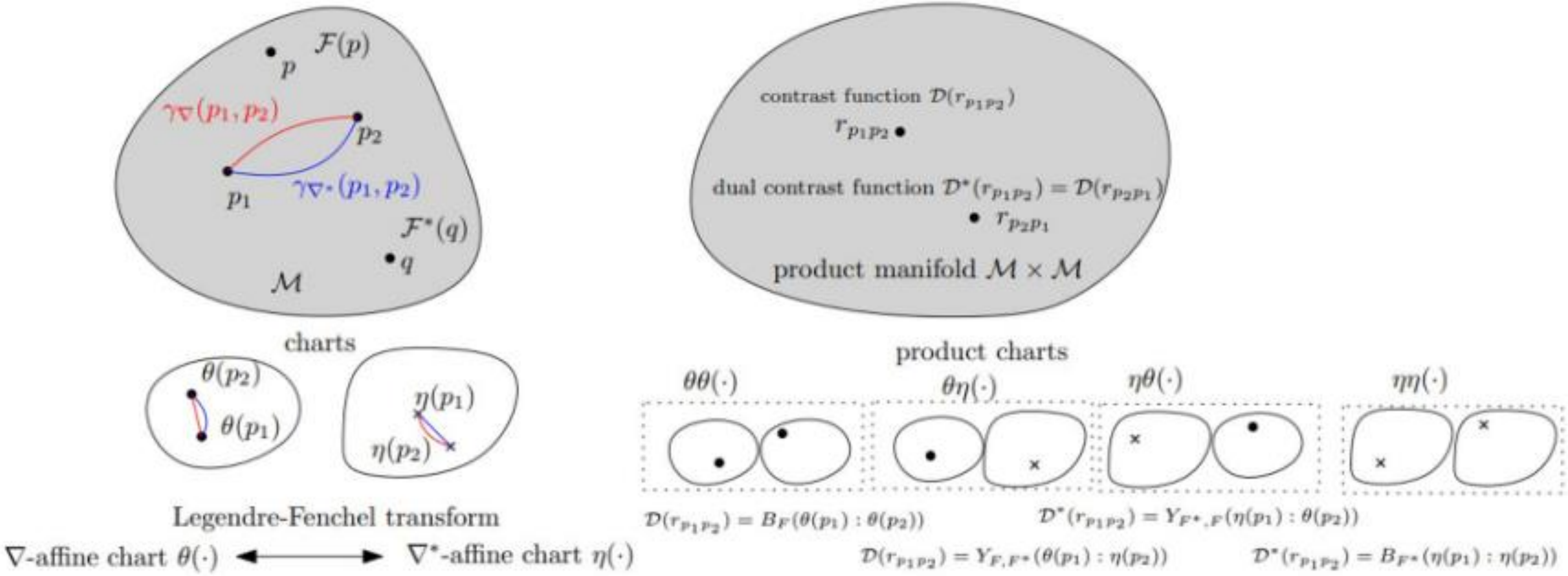

$\eta = \nabla F(\theta)$
$\nabla F = (\nabla F^*)^{-1}$
$Y_{F,F^*}(\theta_1, \eta_2) = F(\theta_1) + F^*(\eta_2) - \theta_1 \eta_2$
η_2
$Y_{F,F^*}(\theta_1, \eta_2)$
$F^*(\eta_2)$
η_1
$F(\theta_1)$
$(0, 0)$
θ_1
θ_2
θ



Bregman manifolds

- A strictly convex and smooth Legendre-type function induces a dually flat space (global Hessian manifold)

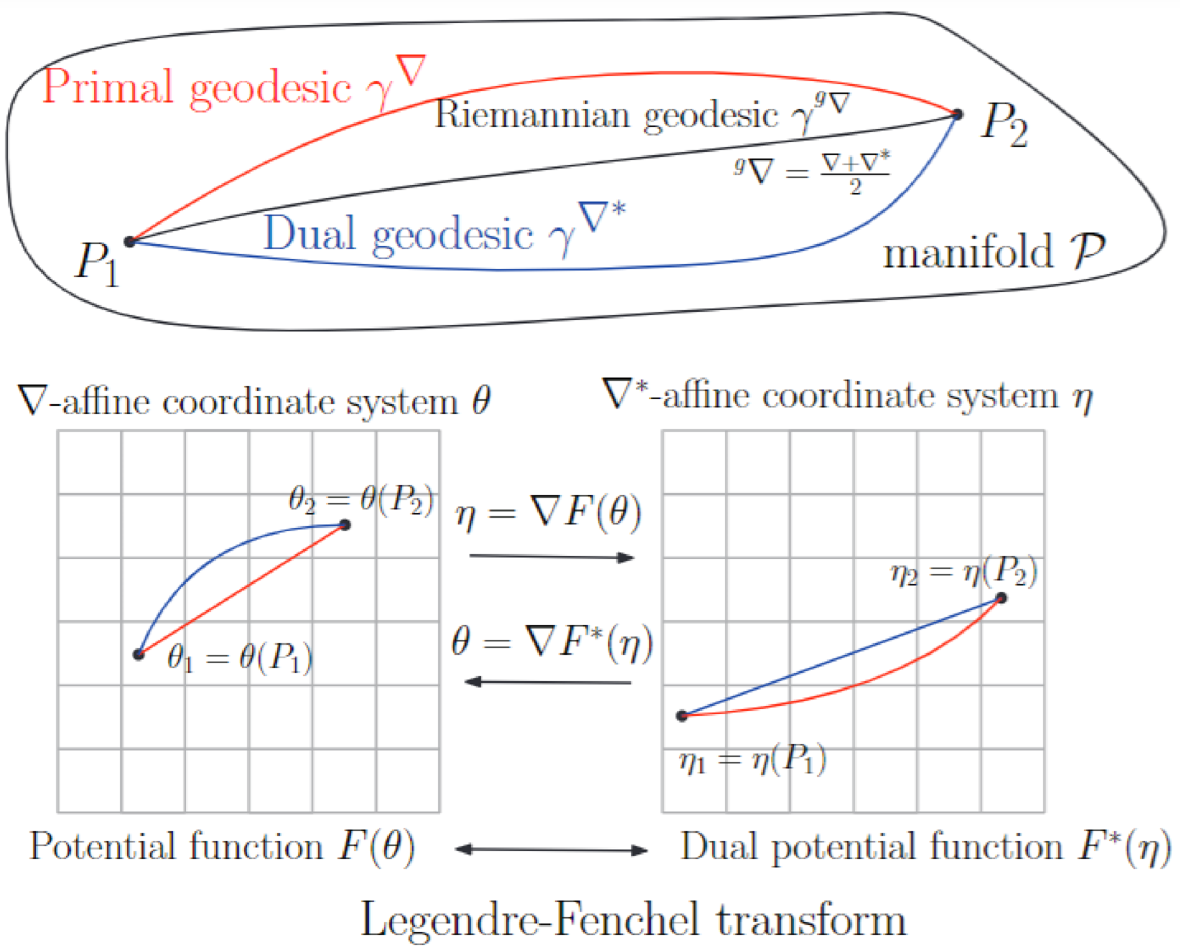
$$\begin{aligned}\mathcal{D}(r_{pq}) &= B_F(\theta(p):\theta(q)) = Y_{F,F^*}(\theta(p):\eta(q)),\\ &= \mathcal{D}^*(r_{qp}) = B_{F^*}(\eta(q):\eta(p)) = Y_{F^*,F}(\eta q:\theta(p))\end{aligned}$$


$\mathcal{F}(p)$
p
$\gamma_\nabla(p_1,p_2)$
p_2
p_1
$\gamma_{\nabla^*}(p_1,p_2)$
$\mathcal{F}^*(q)$
q
$\mathcal{M}$
charts
$\theta(p_2)$
$\theta(p_1)$
$\eta(p_1)$
$\eta(p_2)$
Legendre-Fenchel transform
∇-affine chart $\theta(\cdot)$
∇^*-affine chart $\eta(\cdot)$
contrast function $\mathcal{D}(r_{p_1p_2})$
$r_{p_1p_2}$
dual contrast function $\mathcal{D}^*(r_{p_1p_2}) = \mathcal{D}(r_{p_2p_1})$
$r_{p_2p_1}$
product manifold $\mathcal{M}\times\mathcal{M}$
product charts
$\theta\theta(\cdot)$
$\theta\eta(\cdot)$
$\eta\theta(\cdot)$
$\eta\eta(\cdot)$
$\mathcal{D}(r_{p_1p_2}) = B_F(\theta(p_1):\theta(p_2))$
$\mathcal{D}^*(r_{p_1p_2}) = Y_{F^*,F}(\eta(p_1):\theta(p_2))$
$\mathcal{D}(r_{p_1p_2}) = Y_{F,F^*}(\theta(p_1):\eta(p_2))$
$\mathcal{D}^*(r_{p_1p_2}) = B_{F^*}(\eta(p_1):\eta(p_2))$

- Reciprocally, a dually flat space induces a class of equivalent pair of Legendre-type functions with dual Bregman/Fenchel-Young divergences

Bregman manifolds

(M,g, ∇ , ∇ *)



Primal geodesic γ^{∇}
Riemannian geodesic $\gamma^{g\nabla}$
P_2
${}^g\nabla = \frac{\nabla+\nabla^*}{2}$
Dual geodesic γ^{∇^*}
P_1
manifold $\mathcal{P}$
∇-affine coordinate system θ
∇^*-affine coordinate system η
$\theta_2 = \theta(P_2)$
$\eta = \nabla F(\theta)$
$\theta_1 = \theta(P_1)$
$\theta = \nabla F^*(\eta)$
$\eta_2 = \eta(P_2)$
$\eta_1 = \eta(P_1)$
Potential function $F(\theta)$
Dual potential function $F^*(\eta)$
Legendre-Fenchel transform

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

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

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

Bregman manifold of multivariate normals

Cumulant function (convex):

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

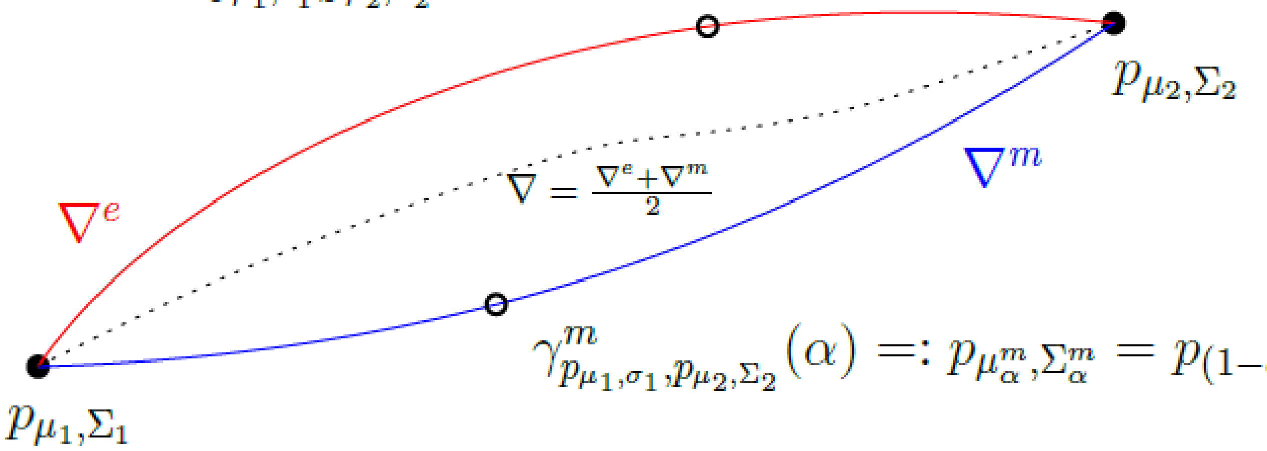
Natural parameters:

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

$$\begin{aligned} \mu^e_\alpha &= \Sigma^e_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right) \\ \Sigma^e_\alpha &= \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1} \end{aligned}$$

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

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



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

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

$$\begin{aligned} \mu^m_\alpha &= (1-\alpha)\mu_1 + \alpha\mu_2 =: \bar{\mu}_\alpha \\ \Sigma^m_\alpha &= (1-\alpha)\Sigma_1 + \alpha\Sigma_2 + (1-\alpha)\mu_1\mu_1^\top + \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha^\top \end{aligned}$$

Kullback-Leibler divergence = reverse Bregman div

$$\frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1)(5\right)$$

(M,g, ∇ , ∇*)

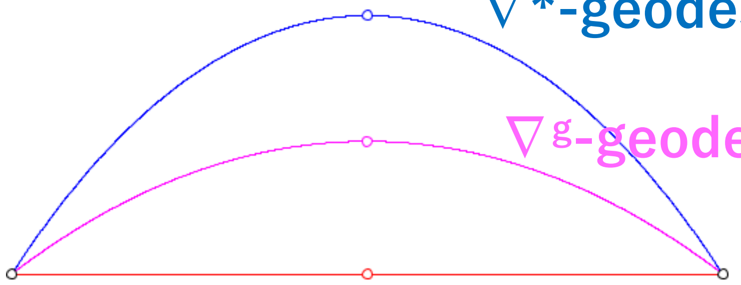
Bregman manifolds have Hessian metrics

- The metric g of a Bregman manifold (M,g,∇,∇^*) is Hessian: $g(\theta)=\nabla^2 F(\theta)$ and $g(\eta)=\nabla^2 F^*(\eta)$

Hessian $\nabla^2=\nabla\nabla^\top$

- The dual basis e(p) and e*(p) in tangent planes T_p are reciprocal: $g(e_i,e^{*j})=\delta_i^j$.
- Crouzeix identity: $\nabla^2 F(\theta)\,\nabla^2 F^*(\eta(\theta))=\nabla^2 F(\theta(\eta))\,\nabla^2 F^*(\eta)=I$

- (M,g) is not flat with respect to the Levi-Civita connection ∇^g induced by g



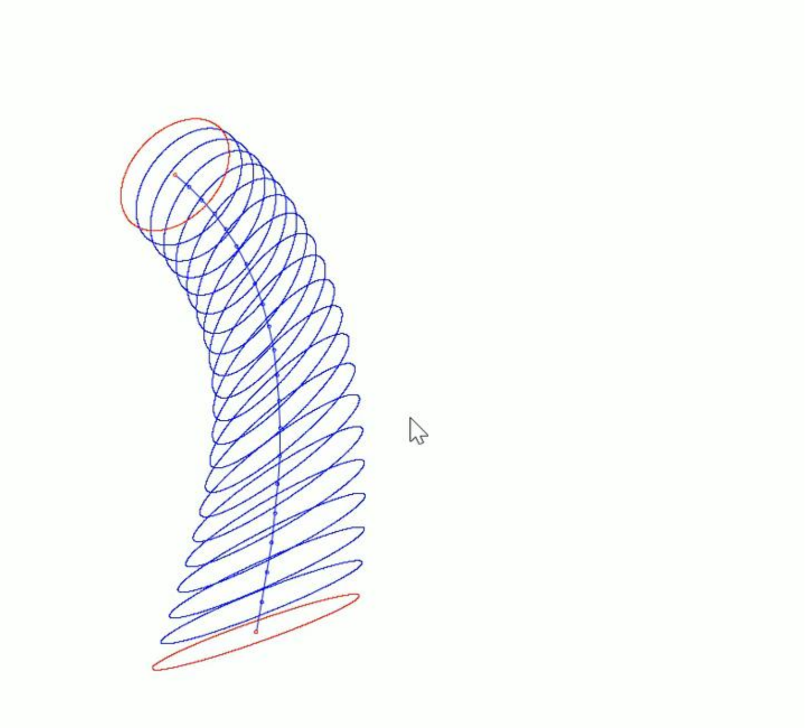
∇^*-geodesic (m-geodesic)

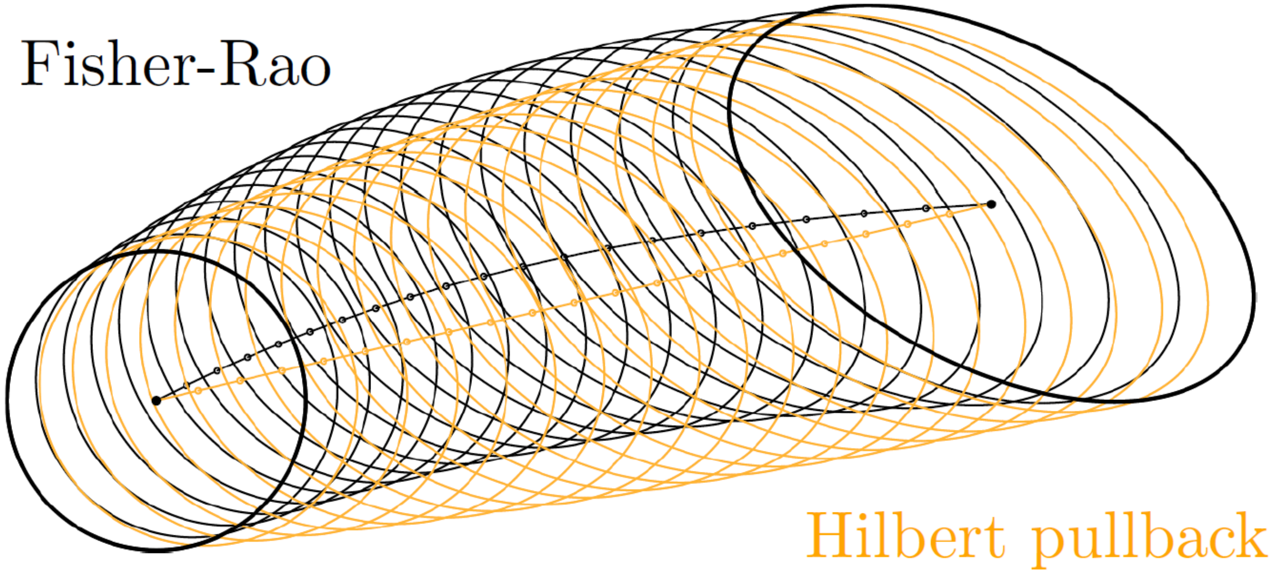
∇^g-geodesic (length minimizing geodesic)

∇-geodesic (e-geodesic)

Fisher-Rao geodesics for dD normals

- When d>1, some sectional curvatures of (M,g_{Fisher}) are positive, Not Hadamard manifold. But centered normal form Hadamard mfd





Kobayashi, Geodesics of multivariate normal distributions and a Toda lattice type Lax pair, Physica Scripta 98.11 (2023)
Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures, ICML TAG 2023.

Inductive matrix arithmetic-harmonic mean

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

[Nakamura 2001]

$$\begin{aligned} A_{t+1} &= \frac{A_t + H_t}{2} = A(A_t, H_t) \\ H_{t+1} &= 2\left(A_t^{-1} + H_t^{-1}\right)^{-1} = H(A_t, H_t) \end{aligned}$$

←arithmetic mean

←harmonic mean

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

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

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

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

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

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

Riemannian distance

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

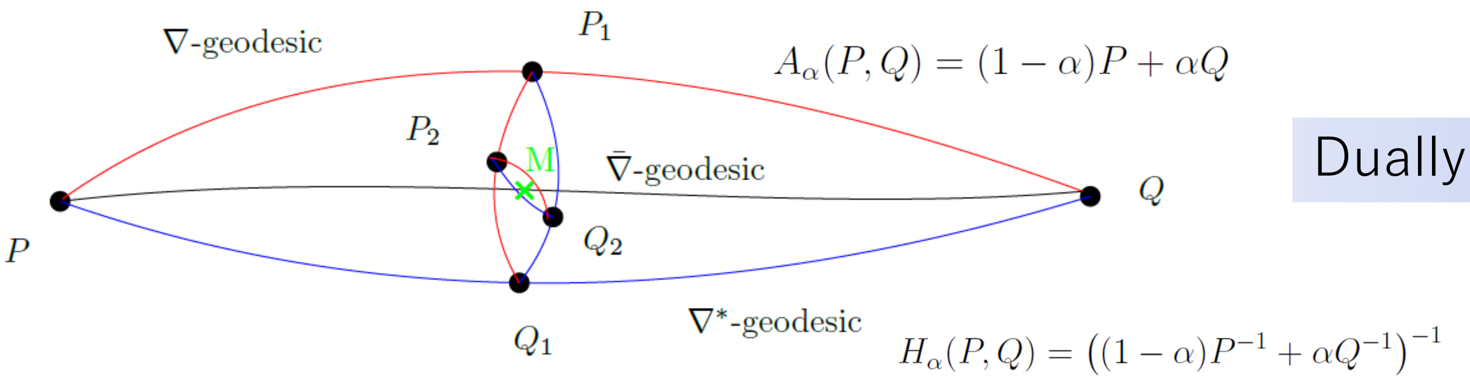
Geometric interpretation of the AHM matrix mean

$$\begin{aligned} A_{t+1} &= \frac{A_t + H_t}{2} = A(A_t, H_t) \\ H_{t+1} &= 2\left(A_t^{-1} + H_t^{-1}\right)^{-1} = H(A_t, H_t) \end{aligned}$$

$$\begin{aligned} P_{t+1} &= \gamma\left(P_t, Q_t : \frac{1}{2}\right) \\ Q_{t+1} &= \gamma^*\left(P_t, Q_t : \frac{1}{2}\right) \end{aligned}$$

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

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


∇-geodesic
P_1
$A_\alpha(P,Q) = (1-\alpha)P + \alpha Q$
P_2
M
$\bar{\nabla}$-geodesic
P
Q
Q_2
Q_1
∇^*-geodesic
$H_\alpha(P,Q) = \left((1-\alpha)P^{-1} + \alpha Q^{-1}\right)^{-1}$

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

Primal geodesic midpoint is the arithmetic center wrt Euclidean metric

Dual geodesic midpoint = harmonic center wrt an isometric Eucl. metric

Levi-Civita geodesic midpoint is geometric Karcher mean

Here, all 3 connections are metric connections!

$$g_P^A(X,Y) = \mathrm{tr}(X^\top Y)$$

$$g_P^H(X,Y) = \mathrm{tr}(P^{-2} X P^{-2} Y)$$

$$g_P^G(X,Y) = \mathrm{tr}(P^{-1} X P^{-1} Y)$$

[Nakamura 2001

Bregman manifolds and Bregman divergences

- Any Legendre-type function (⊖,F(θ)) generates (M,g, ∇ , ∇^*) where F(θ) defines flat connection ∇ via Christoffel symbols $\Gamma(\theta)=0$, and F* (η) defines flat connection ∇^* via Christoffel symbols $\Gamma^*(\eta)=0$
- Duality in information geometry: $(\nabla + \nabla^*)/2$ is Levi-Civita connection ∇^g

- For example, the cumulant functions of exponential families

$F(\theta)=\log \int \exp(<x, \theta>) \, d\mu(x)$. In that case, the Bregman divergence amounts to a reverse Kullback-Leibler divergence

- The partition function $Z(\theta)=\int \exp(<x, \theta>) \, d\mu(x)=\exp(F(\theta))$ is log-convex and log-convex functions are convex. Hence, we can build a Bregman manifold from $Z(\theta)$ too!

- **Question: What is the reconstructed statistical divergence from Bregman divergence B_Z?**

"An elementary introduction to information geometry." *Entropy* 22.10 (2020)

Bregman divergences and Jensen divergences Cumulant functions/Partition functions

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

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

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

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

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

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

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

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

KLD/ α-Bhat ⇔ Bregman/Jensen divergences when considering exponential families

Statistical distances

Parameter divergences

$$\begin{tikzcd}
\mathrm{Bhat}_\alpha(p:q) = -\log \int p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}x \arrow[r, leftrightarrow] \arrow[d, "\lim_{\alpha\to 0+} \frac{1}{\alpha}\mathrm{Bhat}_\alpha(p:q)"'] & J_F^\alpha(\theta_p:\theta_q) = (F(\theta_p)F(\theta_q))_\alpha - F((\theta_p\theta_q)_\alpha) \arrow[d, "\lim_{\alpha\to 0+} \frac{1}{\alpha}\mathrm{J}_F^\alpha(p:q)"] \\
\mathrm{KL}(p:q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x \arrow[r, leftrightarrow] & B_F(\theta_q:\theta_p) = F(\theta_q) - F(\theta_p) - (\theta_q-\theta_p)^\top \nabla F(\theta_q)
\end{tikzcd}$$

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

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

Generic distributions

Exponential families

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

Zhang, Divergence function, duality, and convex analysis, *Neural computation* 16.1 (2004)

N + Boltz. "The Burbea-Rao and Bhattacharyya centroids." *IEEE Transactions on Information Theory* (2011)

Bregman divergences corresponding to partition functions

Question: What is the reconstructed statistical divergence from Bregman divergence B_Z?

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

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

Amari α-divergences extended to positive measures

Scaled skewed Jensen divergence for partition function Z

So B_Z corresponds to the reverse extended Kullback-Leibler divergence:

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

"Divergences Induced by the Cumulant and Partition Functions of Exponential Families and Their Deformations Induced by Comparative Convexity." *Entropy* 26.3 (2024): 193.

Information geometry in action! (1/2)

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

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

Mixture family are closed under mixture operations

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

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

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

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

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

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

Jensen-Shannon centroid

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

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

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

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

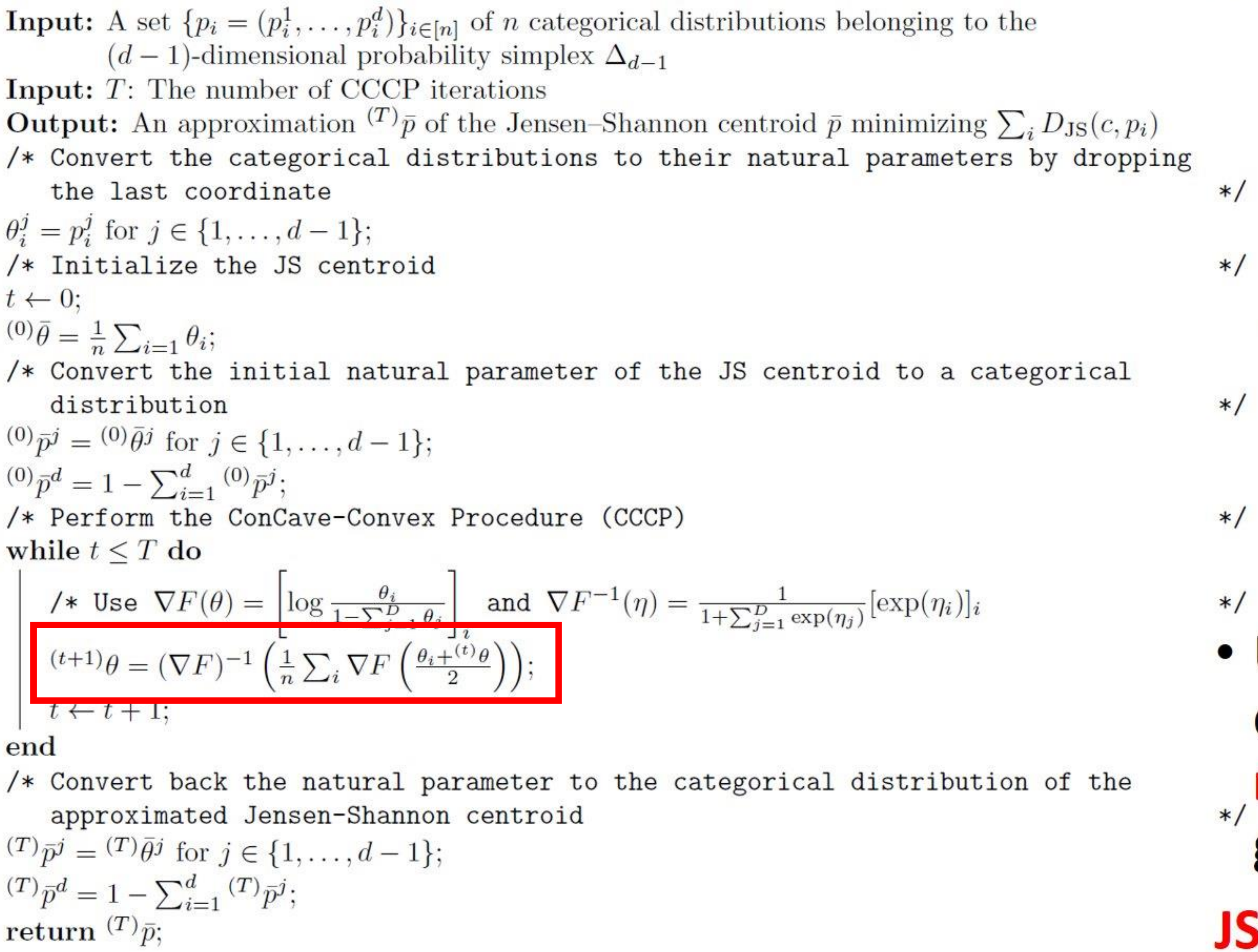
$$\min_\theta \sum_i J_F(\theta_i, \theta),$$

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

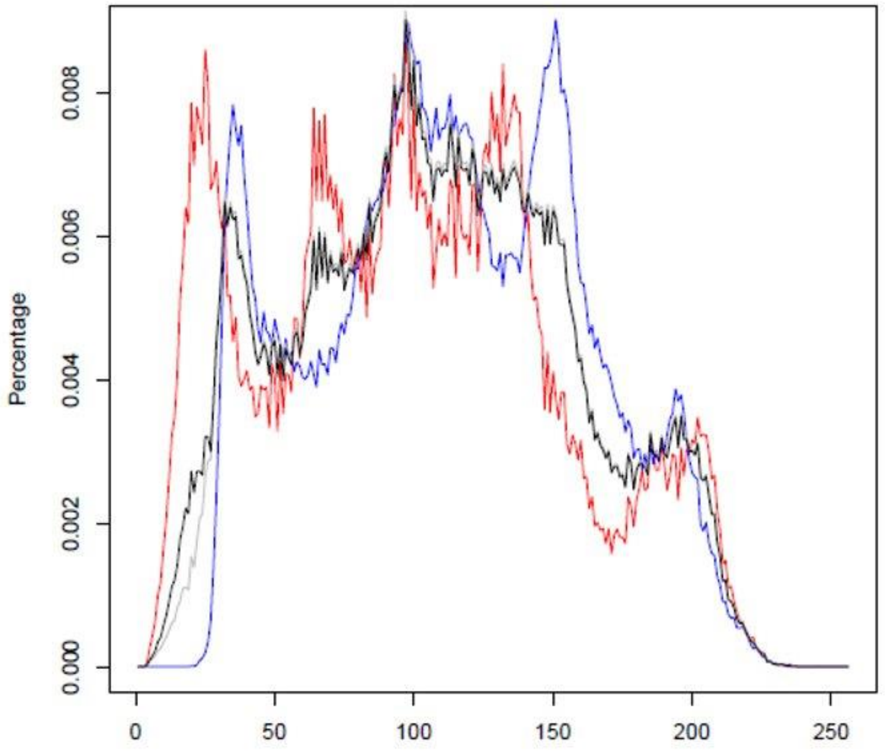
$$\equiv \min_\theta \frac{1}{2} F(\theta) - \frac{1}{n} \sum_i F\left(\frac{\theta_i + \theta}{2}\right) := E(\theta).$$

Need to minimize a difference of convex functions
DCA or CCCP algorithm!

Jensen-Shannon centroid of categorical distributions



```
Input: A set $\{p_i = (p_i^1, \dots, p_i^d)\}_{i\in[n]}$ of $n$ categorical distributions belonging to the
       $(d-1)$-dimensional probability simplex $\Delta_{d-1}$
Input: $T$: The number of CCCP iterations
Output: An approximation ${}^{(T)}\bar{p}$ of the Jensen–Shannon centroid $\bar{p}$ minimizing $\sum_i D_{\mathrm{JS}}(c, p_i)$
/* Convert the categorical distributions to their natural parameters by dropping
   the last coordinate                                                          */
$\theta_i^j = p_i^j$ for $j \in \{1, \dots, d-1\}$;
/* Initialize the JS centroid                                                   */
$t \leftarrow 0$;
${}^{(0)}\bar{\theta} = \frac{1}{n}\sum_{i=1} \theta_i$;
/* Convert the initial natural parameter of the JS centroid to a categorical
   distribution                                                                 */
${}^{(0)}\bar{p}^j = {}^{(0)}\bar{\theta}^j$ for $j \in \{1, \dots, d-1\}$;
${}^{(0)}\bar{p}^d = 1 - \sum_{i=1}^{d} {}^{(0)}\bar{p}^j$;
/* Perform the ConCave-Convex Procedure (CCCP)                                  */
while $t \leq T$ do
    /* Use $\nabla F(\theta) = \left[\log \frac{\theta_i}{1-\sum_{j=1}^{D}\theta_j}\right]_i$ and $\nabla F^{-1}(\eta) = \frac{1}{1+\sum_{j=1}^{D}\exp(\eta_j)}[\exp(\eta_i)]_i$ */
    $\boxed{{}^{(t+1)}\theta = (\nabla F)^{-1}\left(\frac{1}{n}\sum_i \nabla F\left(\frac{\theta_i + {}^{(t)}\theta}{2}\right)\right);}$
    $t \leftarrow t + 1$;
end
/* Convert back the natural parameter to the categorical distribution of the
   approximated Jensen-Shannon centroid                                         */
${}^{(T)}\bar{p}^j = {}^{(T)}\bar{\theta}^j$ for $j \in \{1, \dots, d-1\}$;
${}^{(T)}\bar{p}^d = 1 - \sum_{i=1}^{d} {}^{(T)}\bar{p}^j$;
return ${}^{(T)}\bar{p}$;
```



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

JSD centroid = Jensen centroid

Information geometry in action! (2/2)

- The **Chernoff information** between two distributions is defined by

$$D_C[P,Q] := \max_{\alpha\in(0,1)} -\log \rho_\alpha[P:Q]$$

$$\rho_\alpha[P:Q] := \int p^\alpha q^{1-\alpha}\mathrm{d}\mu = \rho_{1-\alpha}[Q:P],$$

- Chernoff information is the maximal skew Bhattacharrya distance (not metric!):

$$D_{B,\alpha}[p:q] := -\log \rho_\alpha[P:Q] = D_{B,1-\alpha}[q:p],$$

- α -Bhattacharrya distances related to Rényi α -divergences:

$$D_{R,\alpha}[P:Q] = \frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}\mathrm{d}\mu = \frac{1}{1-\alpha}D_{B,\alpha}[P:Q]$$

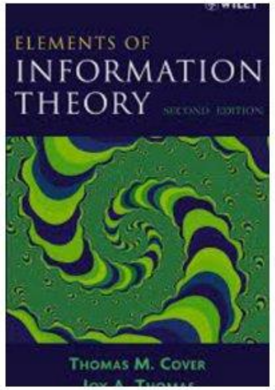
$$D_{B,\alpha}[P:Q] = (1-\alpha)\,D_{R,\alpha}[P:Q]$$

- CI is often used in Bayesian hypothesis testing, information fusion, etc.

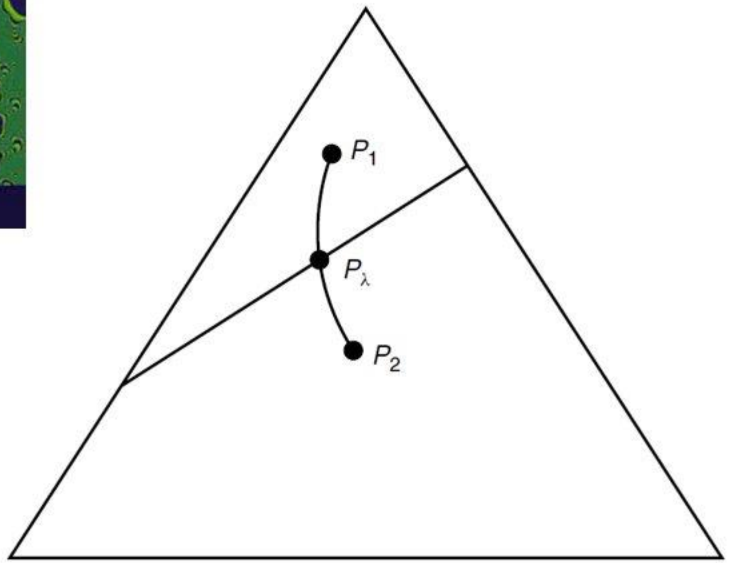
An information-geometric characterization of Chernoff information, *IEEE Signal Processing Letters* 20.3 (2013)
Revisiting Chernoff information with likelihood ratio exponential families, *Entropy* 24.10 (2022)

Chernoff information: A geometric characterization

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



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



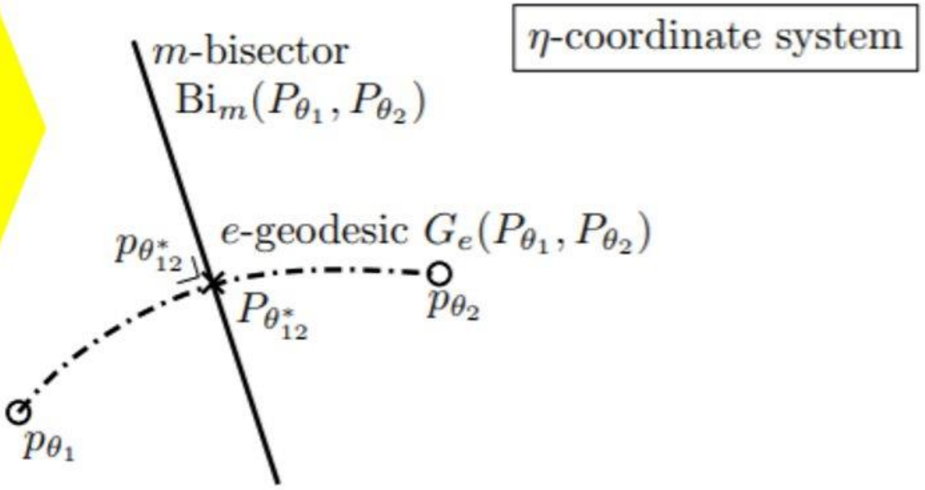
$$P_\lambda = \frac{P_1^{\lambda}(x) P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^{\lambda}(a) P_2^{1-\lambda}(a)}$$

Probability simplex

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

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

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



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

p(x| θ) ∝ exp(<x, θ >)

Exponential family manifold

Likelihood ratio exponential families (LREFs)

- Geometric mixture Bhattacharyya /exponential arc) $(pq)^G_\alpha(x) \propto p(x)^\alpha q(x)^{1-\alpha}$

  between two densities p, q of Lebesgue Banach space $L_1(\mu)$

- Set of **geometric mixtures**:

$$\mathcal{E}_{pq} := \left\{ (pq)^G_\alpha(x) := \frac{p(x)^\alpha q(x)^{1-\alpha}}{Z_{pq}(\alpha)} \ : \ \alpha \in \Theta \right\}$$

  with **normalization factor**:

$$Z_{pq}(\alpha) = \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}\mu(x) = \underline{\rho_\alpha[p:q]}$$

- geometric mixture interpreted as a **1D exponential family**: LREF

$$\begin{aligned} (pq)^G_\alpha(x) &= \exp\left(\alpha \log \frac{p(x)}{q(x)} - \log Z_{pq}(\alpha)\right) q(x), \\ &\overset{*}{=}: \exp(\alpha t(x) - \underbrace{F_{pq}(\alpha)}_{-D_{B,\alpha}[p:q]} + k(x)). \end{aligned}$$

k(x)=log q(x)

Natural parameter space:

$$\Theta := \{\alpha \in \mathbb{R} \ : \ Z_{pq}(\alpha) < \infty\}$$

LREFs: EF cumulant function is always analytic C^ω

- Cumulant function of EF is **strictly convex**

  (and smooth for regular EFs)

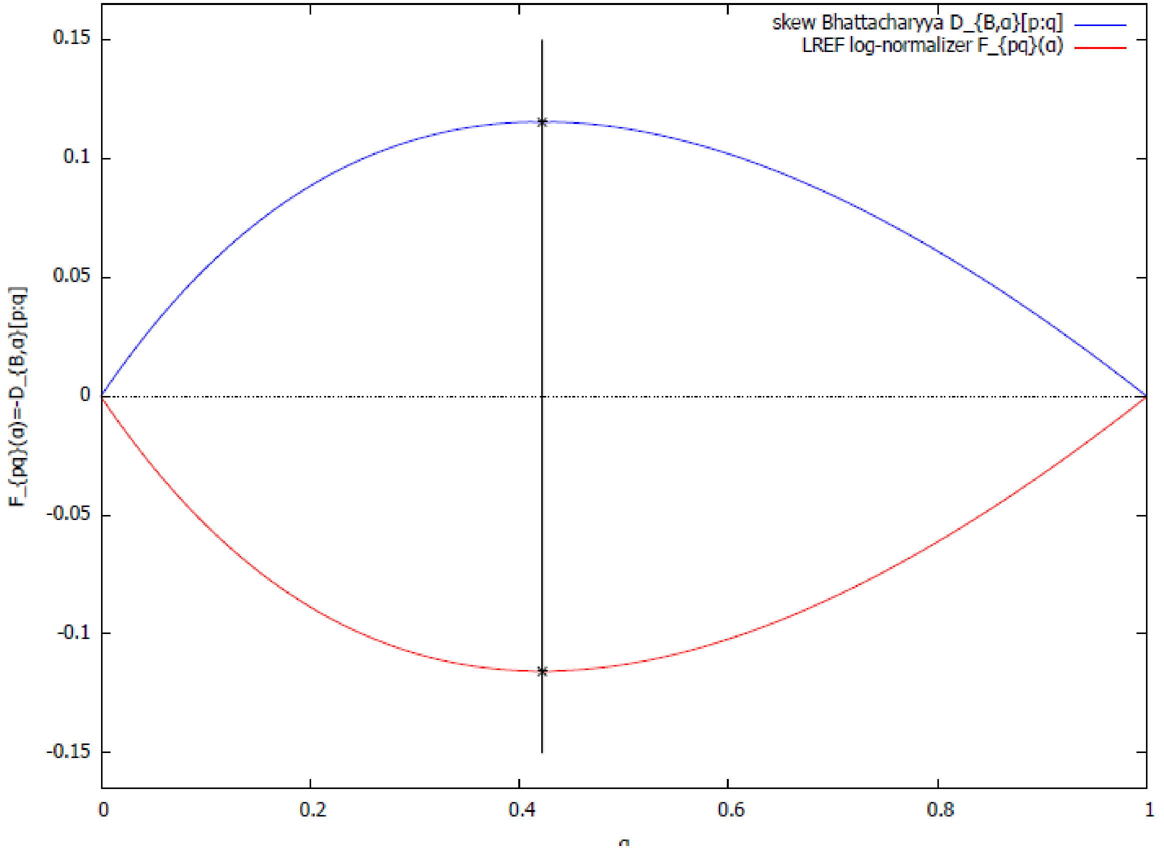
- Cumulant function is neg-Bhattacharyya distance:

$$\boxed{F_{pq}(\alpha) = \log Z_{pq}(\alpha) = -D_{B,\alpha}[p:q] < 0}$$

⇒ Bhattacharyya. distance is **strictly concave**

- <u>Theorem</u>:

  Chernoff exponent exists and is unique



p=N(0,1) q=N(1,2)

$$(pq)^G_\alpha(x) \propto p(x)^\alpha q(x)^{1-\alpha}$$

$$D_C[p,q] = D_{B,\alpha^*(p:q)}(p:q) = D_{B,\alpha^*(q:p)}(q:p) = D_C[q,p \; \alpha^*(q:p) = 1-\alpha^*(p:q)$$

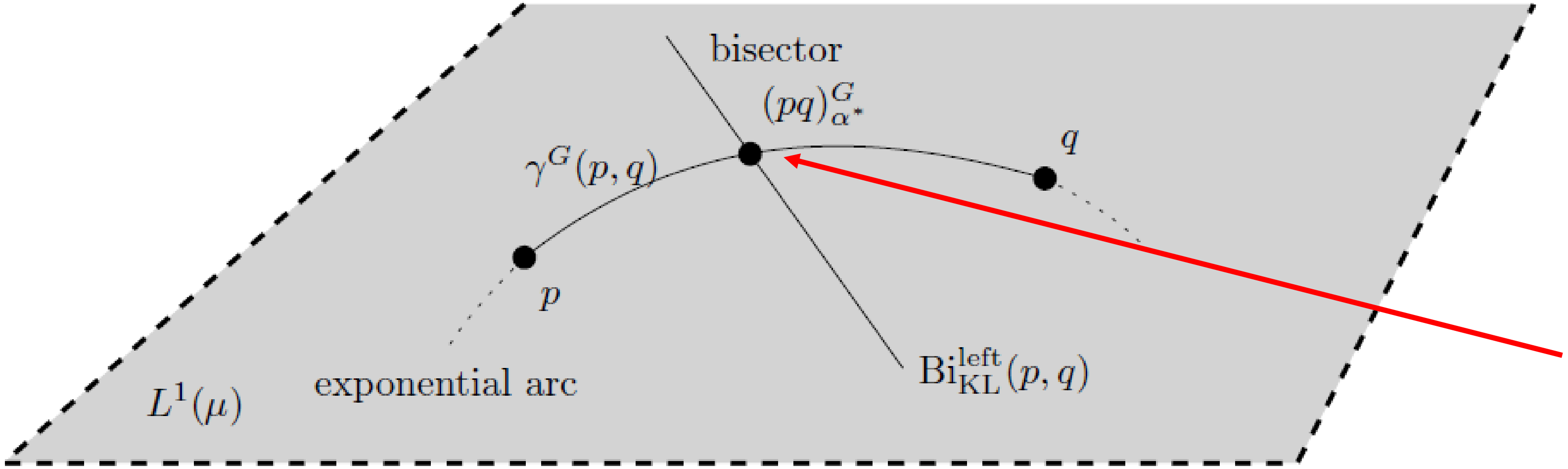
Geometric interpretation for densities on $L_1(\mu)$

Proposition (Geometric characterization of the Chernoff information). *On the vector space $L^1(\mu)$, the Chernoff information distribution is the unique distribution*

$$\boxed{(pq)^{G}_{\alpha^*} = \gamma^{G}(p,q) \cap \mathrm{Bi}^{\mathrm{left}}_{\mathrm{KL}}(p,q).}$$

Left KL Voronoi bisector: $\mathrm{Bi}^{\mathrm{left}}_{\mathrm{KL}}(p,q) := \left\{ r \in L^1(\mu) \ : \ D_{\mathrm{KL}}[\underline{r} : p] = D_{\mathrm{KL}}[\underline{r} : q] \right\}$

Geodesic = exponential arc: $\gamma^{G}(p,q) := \left\{ (pq)^{G}_{\alpha} \ : \ \alpha \in [0,1] \right\}$



Chernoff point: $(pq)^{G}_{\alpha^*}$

dD Bregman divergence as families of 1D BDs

- A d-variate function F(θ) can be equivalently handed as a family of 1D convex functions: **$F_{\theta,\theta'}(\alpha)=F((1-\alpha)\theta+\alpha\theta')$**

- A d-variate BD can be written as an equivalent 1D scalar BD:

Directional derivative

$$\nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(u) = \lim_{\epsilon\to 0}\frac{F(\theta_1+(\epsilon+u)(\theta_2-\theta_1))-F(\theta_1+u(\theta_2-\theta_1))}{\epsilon}$$

$$= (\theta_2-\theta_1)^\top \nabla F(\theta_1+u(\theta_2-\theta_1)).$$

Hence, write BD as equivalent scalar BDs:

$$B_F(\theta_1:\theta_2) := B_{F_{\theta_1,\theta_2}}(0:1)$$

write a BD wrt to anchor points as a **sub-dimensional Bregman divergence**

The chord gap divergence and a generalization of the Bhattacharyya distance." *IEEE ICASSP, 2018*

Symmetrized BD is not a BD in general

- **Symmetrized Bregman divergence**:

$$S_F(\theta_1;\theta_2) := B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) = (\theta_2 - \theta_1)^\top (\eta_2 - \eta_1) = S_{F^*}(\eta_1;\eta_2)$$

- We may double the dimension, and write:

$$S_F(\theta_1,\theta_2) = B_{\hat{F}}(\theta_1^\uparrow : \theta_2^\uparrow) \qquad \xi = \left[\begin{array}{c}\theta\\ \eta\end{array}\right] \quad \hat{F}(\xi) = F(\theta) + F^*(\eta) \qquad \theta^\uparrow = \left[\begin{array}{c}\theta\\ \nabla F(\theta)\end{array}\right]$$

- parameter space $\Theta^\uparrow$ not convex in general !

$$\Theta^\uparrow = \left\{\theta^\uparrow = \left[\begin{array}{c}\theta\\ \nabla F(\theta)\end{array}\right] : \theta \in \Theta\right\} \subset \Xi \qquad \Xi = \left\{\xi = \left[\begin{array}{c}\theta\\ \eta\end{array}\right] : (\theta,\eta) \in \Theta \times H\right\}$$

- Except for generalized quadratic distances (Mahalanobis), SBDs are not BDs. SBDs are **curved Bregman divergences**
- Bregman divergence restricted to a linear subspace is **sub-dimensional Bregman divergence** : For example, extended KLD vs KLD on simplex Δ

N+Nock., Sided and symmetrized Bregman centroids." *IEEE transactions on Information Theory* (2009)

Comparative convexity: (M,N)-convexity

Ordinary convexity of a function: $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

for all t in [0,1]

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

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

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

$A(x, y; \alpha, 1-\alpha) = \alpha x + (1-\alpha)y$ $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

for all t in [0,1]

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

$G(x, y; \alpha, 1-\alpha) = x^{\alpha} y^{1-\alpha}$ $f(tx_1 + (1-t)x_2) \leq f(x_1)^t f(x_2)^{1-t}$

for all t in [0,1]

Comparative convexity wrt quasi-arithmetic means

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

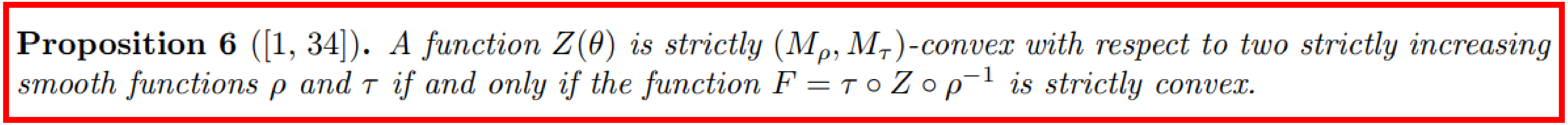
$$M_h(x,y;\alpha,1-\alpha)=h^{-1}(\alpha h(x)+(1-\alpha)h(x)).$$

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

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

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

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



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

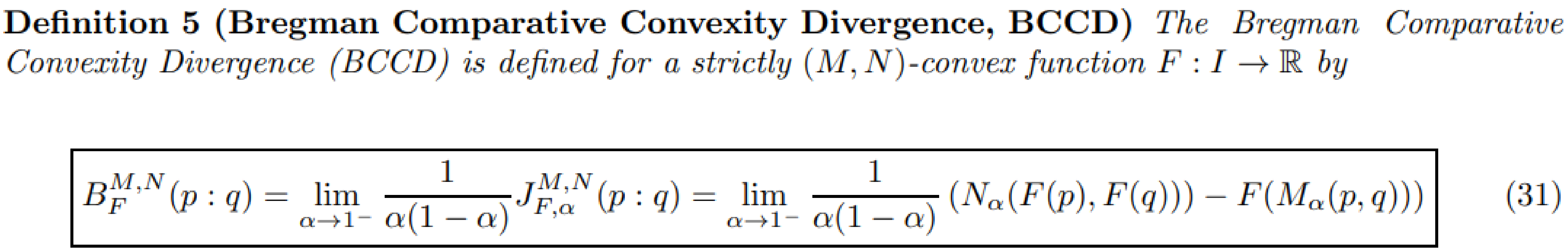
Generalizing Bregman divergences with (M,N)-convexity

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

<u>Definition:</u>

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

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

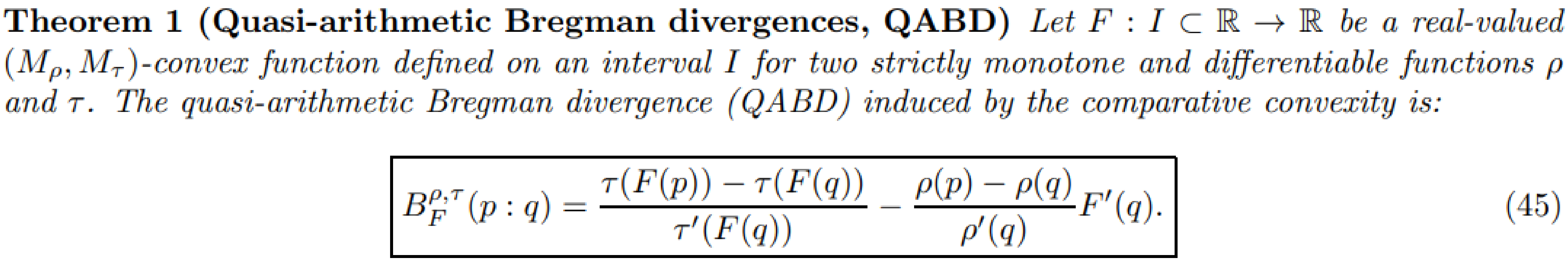


Definition 5 (Bregman Comparative Convexity Divergence, BCCD) *The Bregman Comparative Convexity Divergence (BCCD) is defined for a strictly (M,N)-convex function $F: I \to \mathbb{R}$ by*

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

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

Generalizing Bregman divergences with quasi-arithmetic mean convexity



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

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

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

$$\boxed{B_F^{\rho,\tau}(p:q) = \underbrace{\frac{1}{\tau'(F(q))}}_{\textbf{Conformal factor}} B_G(\rho(p):\rho(q))}$$

With generator:

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

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

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

Summary

- Bregman divergences induce dually flat spaces for **any** Legendre-type C^3 strictly convex generator
- When the generator is an integral from statistical models, we can **reconstruct a statistical divergence**:

  -Reverse KLD from cumulant function of exponential families, rev ext KLD from partition function

  -KLD from negentropy of mixture families

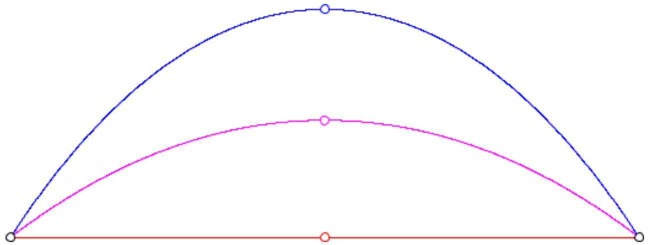
- Jensen-Shannon centroid on **mixture family manifold** using concave-convex algorithm
- Chernoff information on **exponential family manifold** using exact geometric characterization ``Chernoff point'' = unique intersection of primal geodesic with dual bisector

- Define Bregman divergences with respect to **(M,N)-convexity**: Get **conformal Bregman divergences**

- **Duality** is at the heart of information geometry!

Thank you!

Many thanks for joint works with my collaborators.
Special thanks to Richard Nock, Ke Sun, Ehsan Amid, and Alexander Soen



pyBregMan

A Python library for Bregman Manifolds with Applications

Joint work with Alexander Soen

https://franknielsen.github.io/