

# Some generalizations of Bregman divergences

- Biduality and Bregman manifolds -

Frank Nielsen

Sony Computer Science Laboratories Inc

**Bayes-Duality**  
14 June 2024



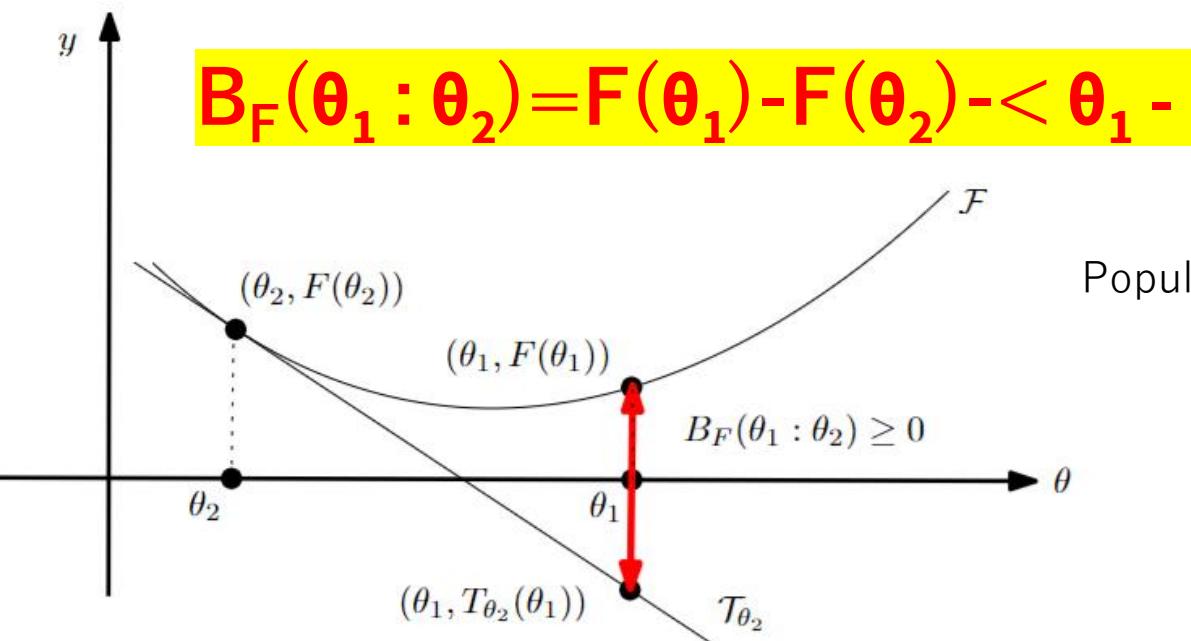
# Agenda

- Bregman divergences with one generator & **two generators**
- **Biduality** in Bregman manifolds [information geometry]
- Geometric computing on **Bregman manifolds**
  - Gain insights of problems from information geometry!
- Bregman divergence with respect to **comparative convexity**:
  - Build many Bregman divergences from one exponential family!

# Bregman divergence (1960's)

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

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



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



**Lev M. Bregman**

(1941 - 2023)

Photo: courtesy of Alexander Fradkov

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

# Bregman divergence: Taylor remainder

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

- BDs are non-negative because = remainder of Taylor 1<sup>st</sup> order expansion:

$$F(\theta_1) = F(\theta_2) + \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle + \text{Remainder}_F(\theta_1 : \theta_2)$$

BD = exact Lagrange remainder = a quadratic distance:

$$\text{LagrangeRemainder}_F(\theta_1 : \theta_2) = \frac{1}{2} (\theta_1 - \theta_2)^T \nabla^2 F(\xi) (\theta_1 - \theta_2) = B_F(\theta_1 : \theta_2)$$

for some  $\xi \in [\theta_1 \ \theta_2]$

- BDs are *never metrics* and
- *BDs are only symmetric for generalized quadratic distance* (proof:  $\nabla^2 F$  constant)

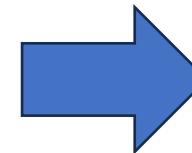
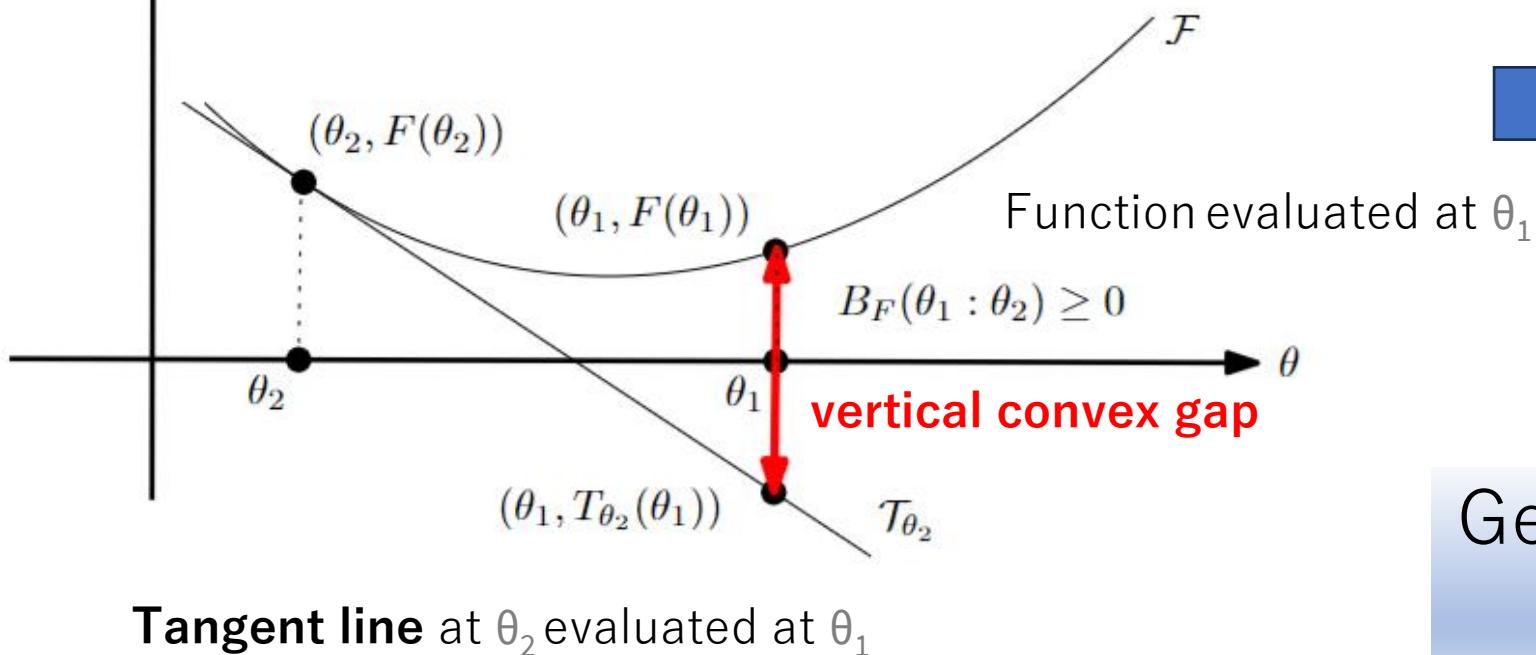
$$F_Q(\theta) = \frac{1}{2} \theta^T Q \theta, \text{ where } Q \text{ is symmetric positive-definite matrix}$$

$Q=I$ , squared Euclidean distance

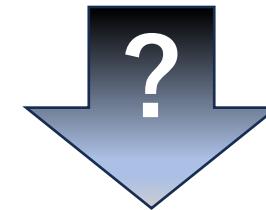
# How can we measure other convexity gaps in graphs?

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

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - T_{\theta_2}(\theta_1)$$



Bregman divergence  
as a **vertical convex gap**

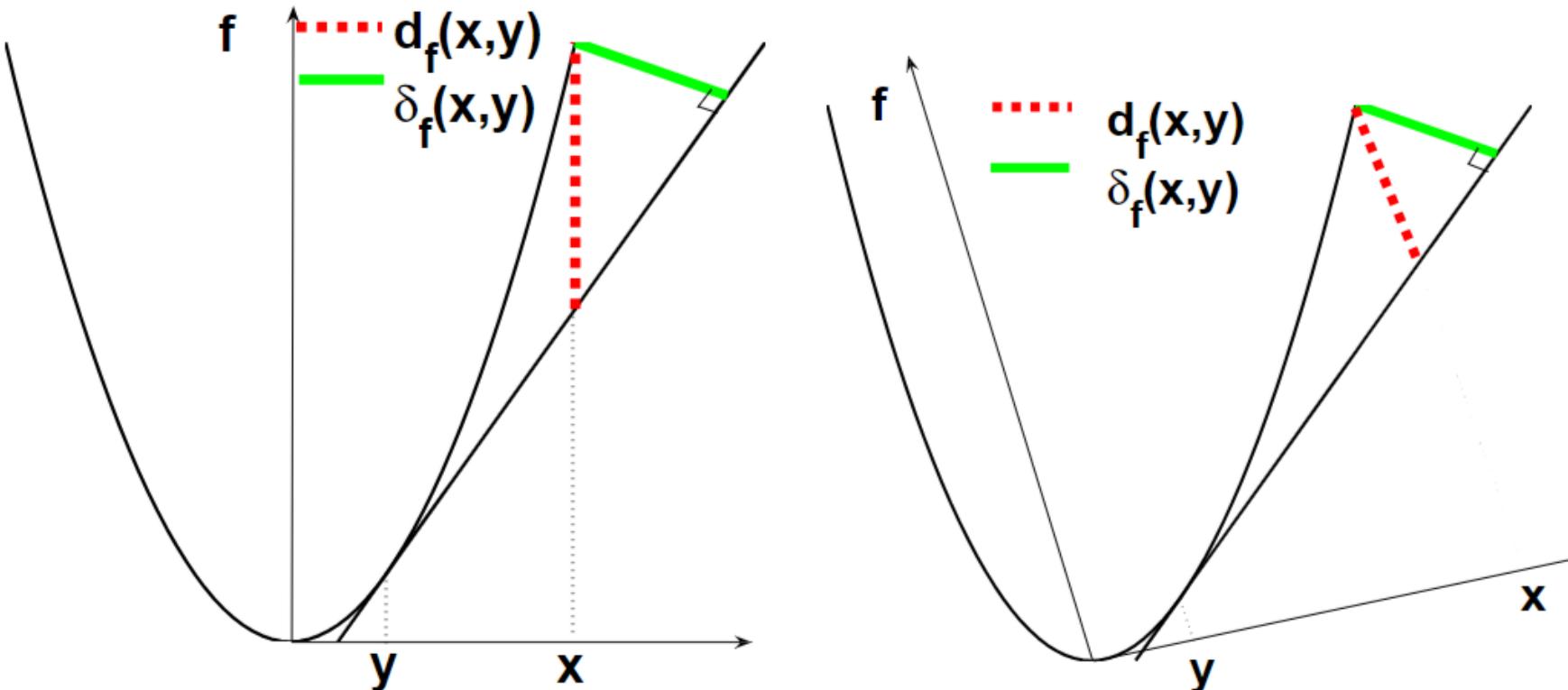


Generalize BDs by measuring  
various convex gaps?

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

# Total Bregman divergences

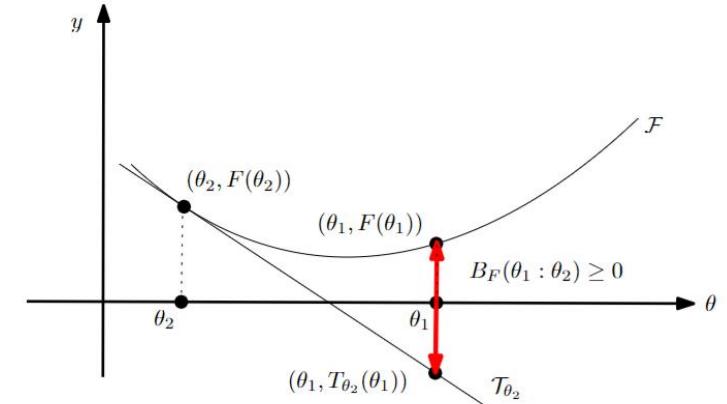
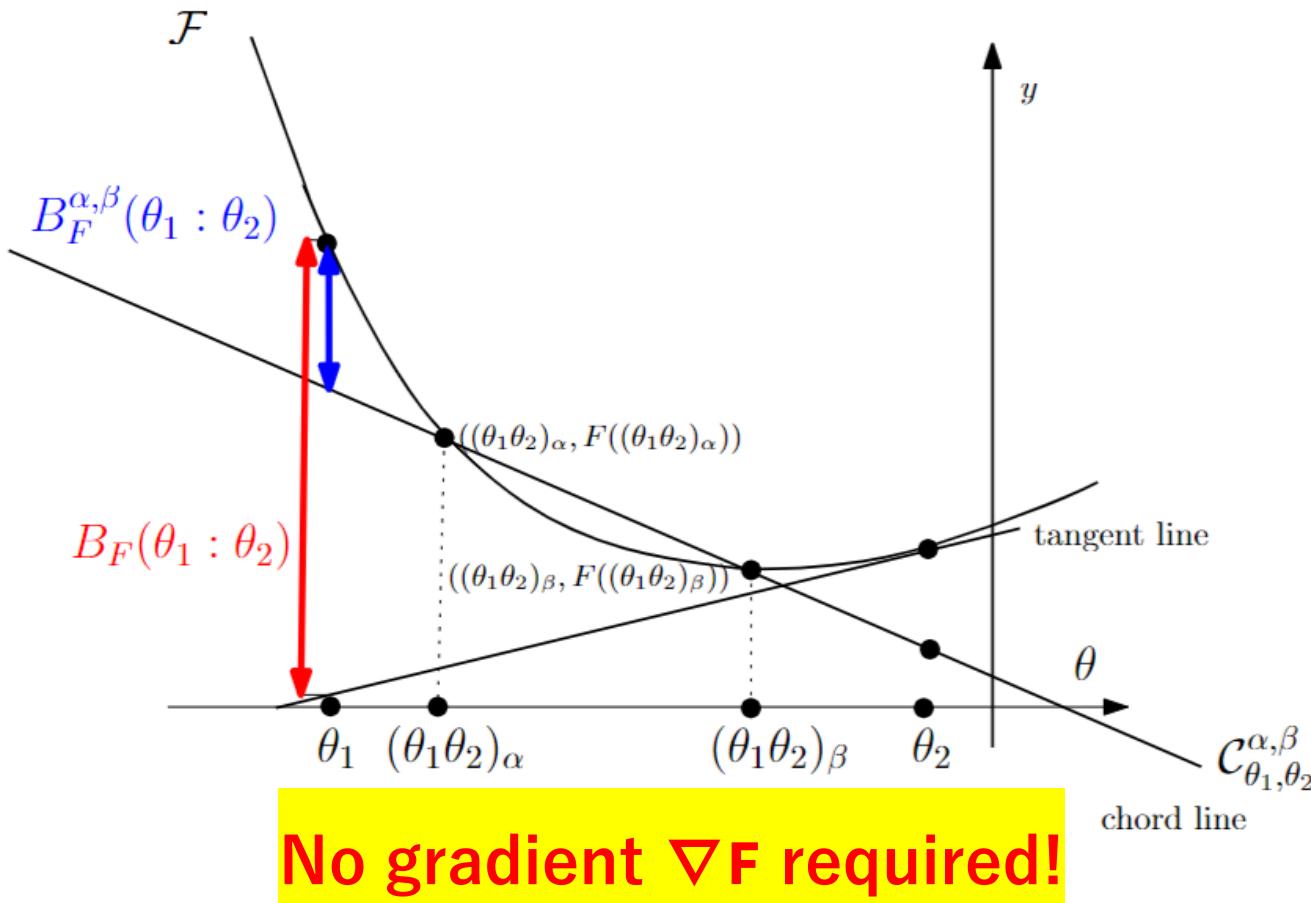
$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}}$$



- Invariant to rotation
- Robust clustering with respect outliers
- Conformal Bregman divergence

$$\begin{aligned} tB(p : q) &= \frac{B(p : q)}{\sqrt{1 + \langle \nabla F(q), \nabla F(q) \rangle}} = \rho_B(q) B(p : q) \\ \rho_B(q) &= \frac{1}{\sqrt{1 + \langle \nabla F(q), \nabla F(q) \rangle}}. \end{aligned}$$

# Bregman-type divergences measuring convexity gaps



When  $\alpha \rightarrow \beta$ , **Bregman chord divergences**  $\rightarrow$  **Bregman tangent divergences**  
or **Bregman secant divergences**

# Bregman divergences in machine learning

- Kullback-Leibler divergence between two probability densities:

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

difficult to calculate in closed form because of the integral  $\int \dots$

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

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

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

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

Bypass the  $\int, \nabla F$  easy!

- Notice divergence between parameters  $B_F$  vs divergence between functions  $KL$

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

# My first encounter with Bregman divergences



## Clustering with Bregman Divergences

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

### Abstract

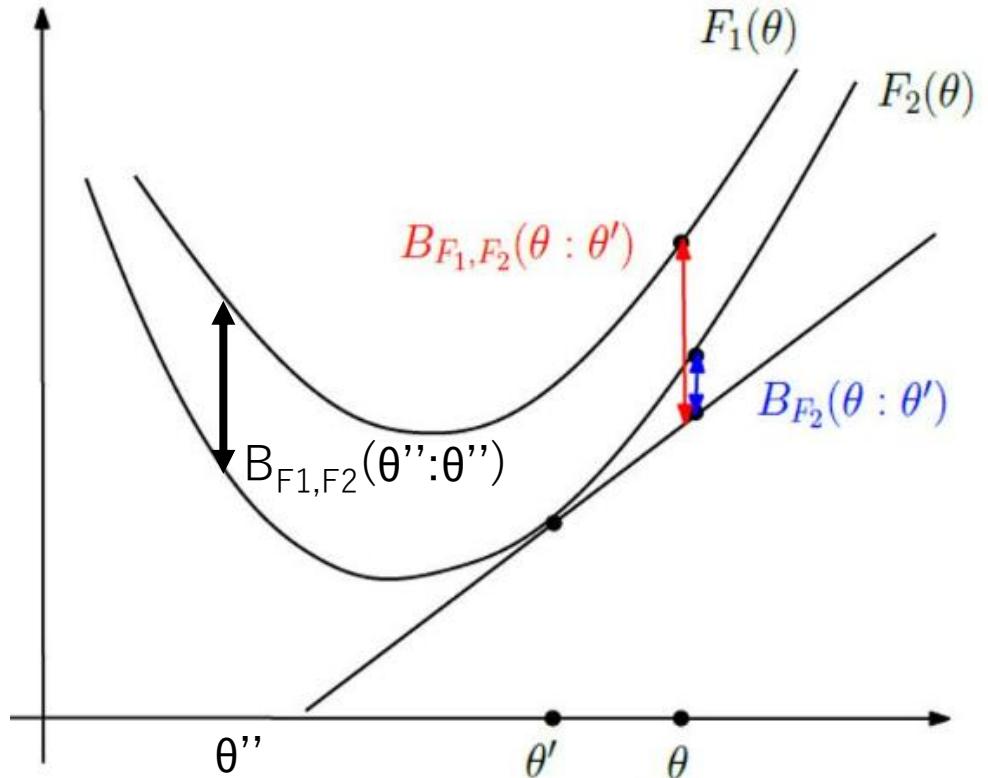
A wide variety of distortion functions are used for clustering, e.g., squared Euclidean distance, Mahalanobis distance and relative entropy. In this paper, we propose and analyze parametric hard and soft clustering algorithms based on a large class of distortion functions known as Bregman divergences. 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**

# Duo Bregman divergences: Generalize BDs with a pair of generators



One generator majorizes the other one:

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

Then

$$\begin{aligned} B_{F_1,F_2}(\theta:\theta') &= F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta') \\ &\geq B_F(\theta:\theta') \end{aligned}$$

- Recover Bregman divergence when  $F_1(\theta) = F_2(\theta) = F(\theta)$   
 $B_F(\theta_1:\theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle$
- Only **pseudo-divergence** because  $B_{F_1,F_2}(\theta'':\theta'')$  may be positive, not zero

# KLD between nested exponential families amount to duo Bregman divergences

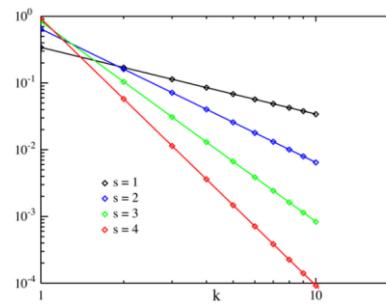
$$\frac{\int_{X_1} \frac{q(x|\theta) \gg p(x|\theta)}{p(x|\theta)} d\mu(x)}{\int_{X_2} q(x|\theta) d\mu(x)}$$

- Consider an exponential family on support  $X_1$ :  $D_{KL}[p(x):q(x)] = \int p(x) \log(p(x)/q(x)) d\mu(x)$   
 $p(x|\theta) = \exp(\langle x, \theta \rangle - F_1(\theta)) d\mu(x)$        $0 \log(0/0) = 0$   
with cumulant function  $F_1(\theta) = \log \int_{X_1} \exp(\langle x, \theta \rangle) d\mu(x)$
- Another exponential family with **nested supports:  $X_1 \subseteq X_2$**   
 $q(x|\theta) = \exp(\langle x, \theta \rangle - F_2(\theta)) d\mu(x)$   
is an exponential family with  $F_2(\theta) = \log \int_{X_2} \exp(\langle x, \theta \rangle) d\mu(x) \geq F_1(\theta)$
- 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)

# Application of duo Bregman divergences: Clustering distributions with different supports



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

$$f(k; N, s) = \frac{1}{H_{N,s}} \frac{1}{k^s} \quad H_{N,s} = \sum_{k=1}^N \frac{1}{k^s}$$

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

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s} \quad \text{This ensures that } q(x|\lambda_j) \gg p(x|\theta_i)$$

- Objective is a la “k-means”: minimize  $\sum_i D_{KL}[p(x|\theta_i) : \{q(x|\lambda_j) : j \text{ in } \{1, \dots, k\}\}]$
- Duo Bregman k-means** algorithm as an extension of Bregman k-means
- Example: Cluster Zipf's distributions from word frequencies in a collection of translations of a renown book. Find similarities of word frequency in languages

# Convex duality via Legendre-Fenchel transform

- Legendre-Fenchel transform of a convex function  $F$ :

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

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

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

Then we get:

- reciprocal gradient maps  $\eta = \nabla F(\theta)$  and  $\theta = \nabla F^*(\eta)$ ,  $\nabla F^* = (\nabla F)^{-1}$
- conjugation yields  $(H, F^*(\eta))$  of Legendre type
- biconjugation is an involution:  $(H, F^*(\eta))^* = (H^* = \Theta, F^{**} = F(\theta))$

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

# Solo and duo Fenchel-Young divergences

- Young inequality:  $F(\theta_1) + F^*(\eta_2) \geq \langle \theta_1, \eta_2 \rangle$  with equality when  $\eta_2 = \nabla F(\theta_1)$
- Build the Fenchel-Young divergence from the inequality: lhs-rhs  $\geq 0$

$$Y_{F, F^*}(\theta_1, \eta_2) = F(\theta_1) + F^*(\eta_2) - \langle \theta_1, \eta_2 \rangle \geq 0$$



- Legendre transform reverses majorization order:

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

- **Duo Fenchel-Young divergence:**

$$Y_{F_1, F_2^*}(\theta, \eta') := F_1(\theta) + F_2^*(\eta') - \theta^\top \eta',$$

$$\geq F_1(\theta) + F_1^*(\eta') - \theta^\top \eta' = Y_{F_1, F_1^*}(\theta, \eta') \geq 0$$

# Outline

- Bregman divergences with a single generator and a pair of generators
- **Biduality and Bregman manifolds**
- Information geometry in action!
- Bregman divergence with respect to comparative convexity

# Biduality reference/representation

Legendre-Fenchel transformation \* induces two functions:

- ① a convex conjugate function  $F^*(\cdot) : \eta \rightarrow F^*(\eta)$
- ② a gradient map:  $\eta = \nabla F(\cdot) : \theta \rightarrow \eta(\theta) = \nabla F(\theta)$

This yields the interplay of two dualities:

- ① **Reference duality** of divergences:  $D^*(p_1:p_2) := D(p_2: p_1)$ ,  $(D^*)^* = D$
- ② **Representation duality**:  $\theta^* = \nabla F(\theta)$ , and  $(\theta^*)^* = \theta$

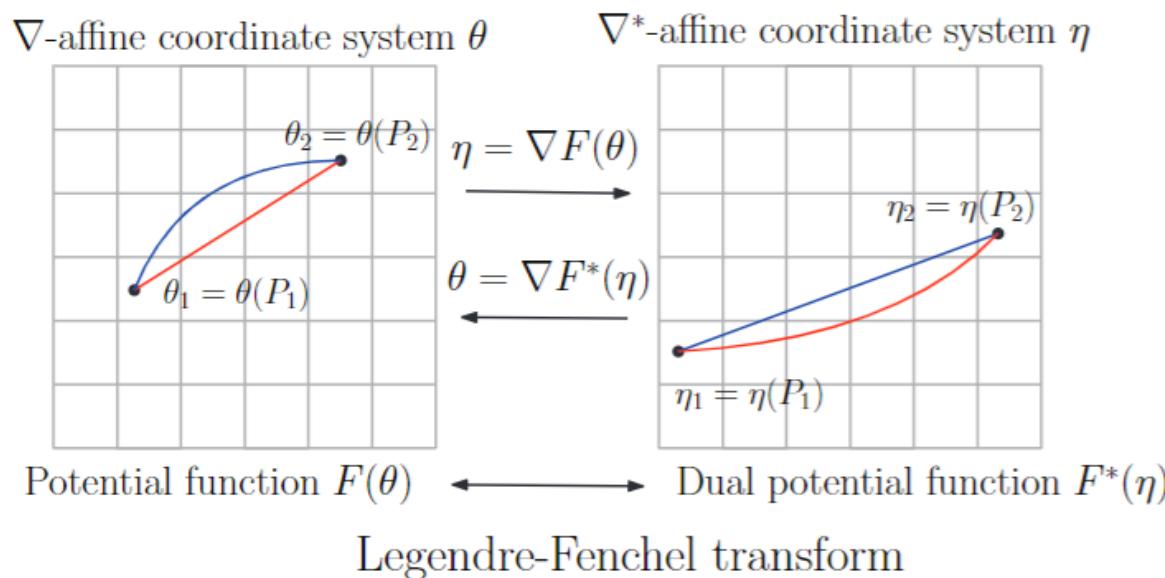
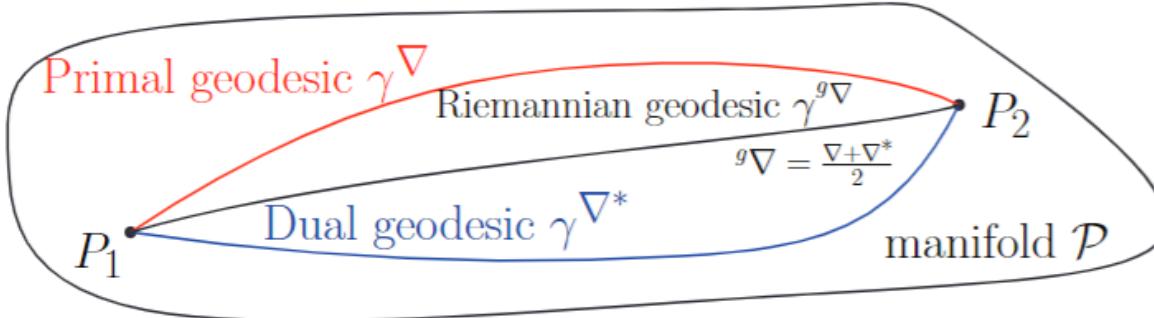
Fenchel-Young divergence demonstrates the biduality interaction:

$$(Y_F)^*(\theta_1: \theta_2^*) = Y_F(\theta_2^*: \theta_1) = Y_{F^*}(\theta_2^*: \theta_1)$$

Functional divergences on densities as pointwise scalar BD can use monotone  $(\rho, \tau)$ -embedding dualities

# Bregman manifolds: Convex conjugates $(F, F^*)$ yield dual flat connections

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



- A connection  $\nabla$  is **flat** if there exists a coordinate system  $\theta$  such that all Christoffel symbols vanish:  $\Gamma(\theta) = 0$ .

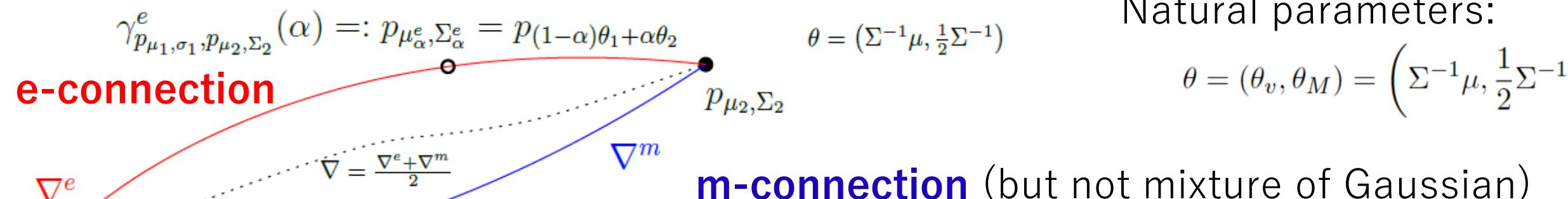
- $\theta$  is called  **$\nabla$ -affine coordinate system**
- **$\nabla$ -geodesic** solves as **line segments**

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

# Example: 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:

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

**m-connection** (but not mixture of Gaussian)

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

Kullback-Leibler divergence = reverse Bregman div

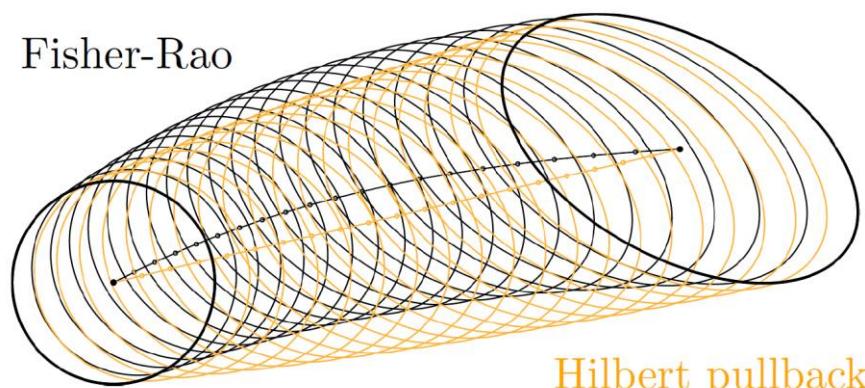
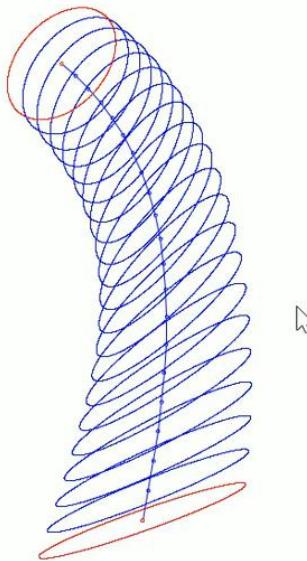
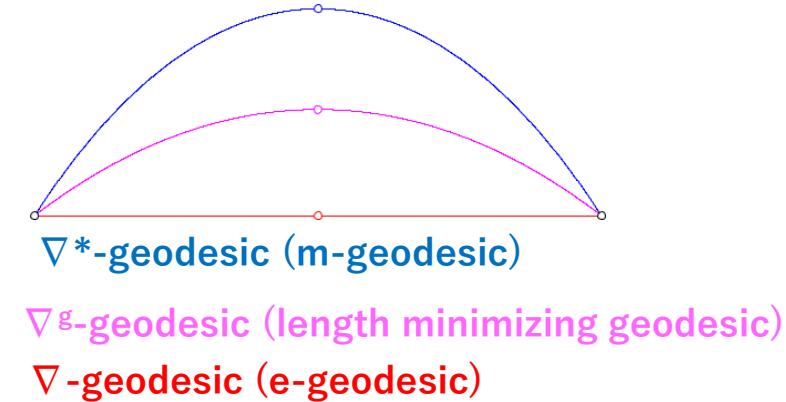
(M,g,  $\nabla$ ,  $\nabla^*$ )

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

Geometric Jensen-Shannon divergence, 2019

# Fisher-Rao geodesics for d-variate normals

- When  $d > 1$ , some sectional curvatures of  $(M, g_{\text{Fisher}})$  are positive, Fisher-Rao is not Hadamard manifold.  
But centered normal submanifold is Hadamard manifold



Geodesic in closed form same-mean centered normal:

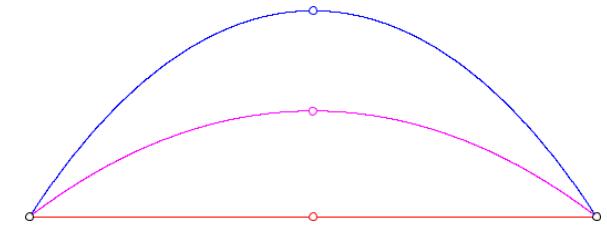
$$\gamma_{\mathcal{P}}(P, Q; t) = P^{\frac{1}{2}} \left( P^{-\frac{1}{2}} Q^{\frac{1}{2}} P^{-\frac{1}{2}} \right)^t P^{\frac{1}{2}}, \quad t \in [0, 1]$$

$$t=1/2 \text{ yields matrix geometric mean } G(X, Y) = X^{\frac{1}{2}} (X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})^{\frac{1}{2}} X^{\frac{1}{2}}$$

# Bregman manifolds and Bregman divergences

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

For example:



- The **cumulant functions** of exponential families  $F(\theta) = \log \int \exp(\langle x, \theta \rangle) d\mu(x)$ . In that case, the Bregman divergence amounts to a reverse Kullback-Leibler divergence
- The **partition functions**  $Z(\theta) = \int \exp(\langle x, \theta \rangle) d\mu(x) = \exp(F(\theta))$  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  $B_Z$ ?**

# Bregman divergences and Jensen divergences

## Cumulant functions/Partition functions

$$F(\theta) = \log Z(\theta) \Leftrightarrow Z(\theta) = \exp(F(\theta))$$

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

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

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

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

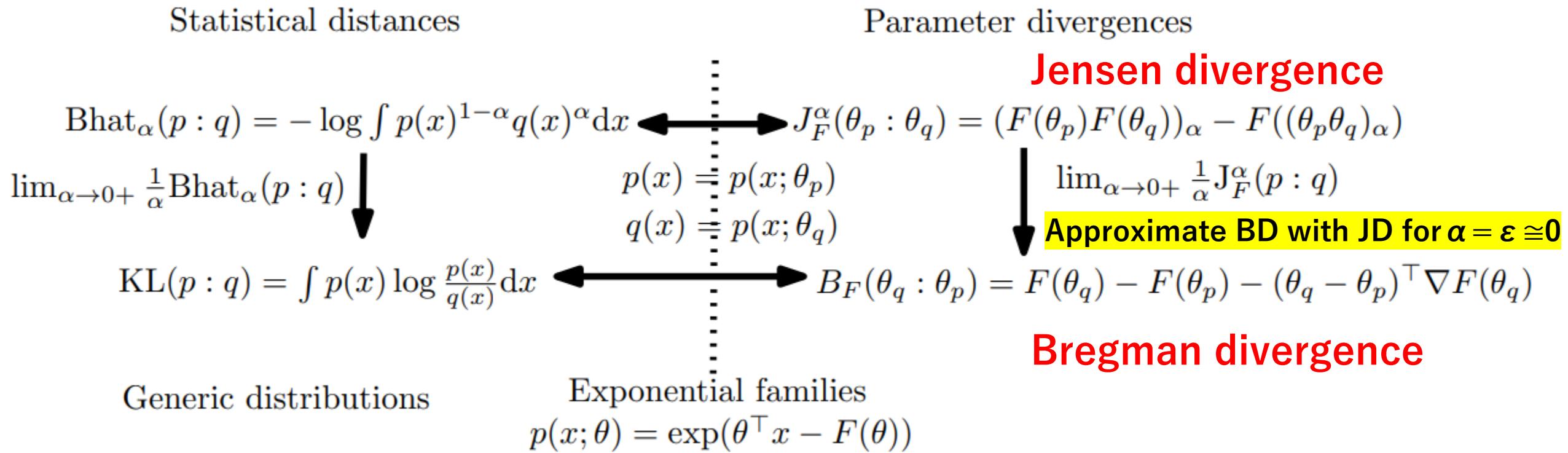
$$\textcircled{1} \quad J_{Z,\alpha}(\theta_1 : \theta_2) = \alpha Z(\theta_1) + (1 - \alpha)Z(\theta_2) - Z(\alpha\theta_1 + (1 - \alpha)\theta_2) \geq 0,$$

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

Including the **symmetric Jensen divergence** when  $\alpha=1/2$ :

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

# KLD/ $\alpha$ -Bhattacharyya $\Leftrightarrow$ Bregman/Jensen divergences when considering exponential families, F cumulant function



Question: What are the reconstructed statistical divergences from Jensen-Bregman divergence  $J_Z$  /  $B_Z$ ?  
 $Z$ : partition function

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  $B_Z$ ?

$$J_{Z,\alpha}^s(\theta_1 : \theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{Z,\alpha}(\theta_1 : \theta_2), & \alpha \in \setminus\{0,1\}, \\ B_Z(\theta_1 : \theta_2), & \alpha = 0, \\ 4 J_Z(\theta_1, \theta_2), & \alpha = \frac{1}{2}, \\ B_Z^*(\theta_1 : \theta_2) = B_Z(\theta_2 : \theta_1), & \alpha = 1. \end{cases} \quad \longleftrightarrow \quad D_\alpha(\tilde{p} : \tilde{q}) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \int (\alpha \tilde{p} + (1-\alpha)\tilde{q} - \tilde{p}^\alpha \tilde{q}^{1-\alpha}) d\mu, & \alpha \notin \{0,1\} \\ D_{\text{KL}}^*(\tilde{p} : \tilde{q}) = D_{\text{KL}}(\tilde{q} : \tilde{p}) & \alpha = 0, \\ 4 D_H^2(\tilde{p}, \tilde{q}) & \alpha = \frac{1}{2}, \\ D_{\text{KL}}(\tilde{p} : \tilde{q}) & \alpha = 1. \end{cases}$$

Scaled skewed Jensen divergence  
for partition function Z



Amari  $\alpha$ -divergences extended to  
positive measures

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

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

# Outline

- Bregman divergences with a single generator and a pair of generators
- Bregman manifolds
  - Normalized/unnormalized exponential families and Bregman divergences
  - Fisher-Rao distance and arithmetic-harmonic mean on the SPD manifold
- **Further information geometry in action!**
  - on a mixture family manifold
  - on an exponential family manifold
- Bregman divergence with respect to comparative convexity
- Conclusion: Dualities and conjugacies in ML

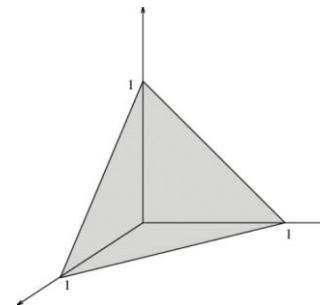
# Information geometry in action! (1/2)

- Set of categorical distributions form a **mixture family  $\mathbf{M}$** , a Bregman manifold for the negentropy generator

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

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

A mixture family is closed under mixture operations



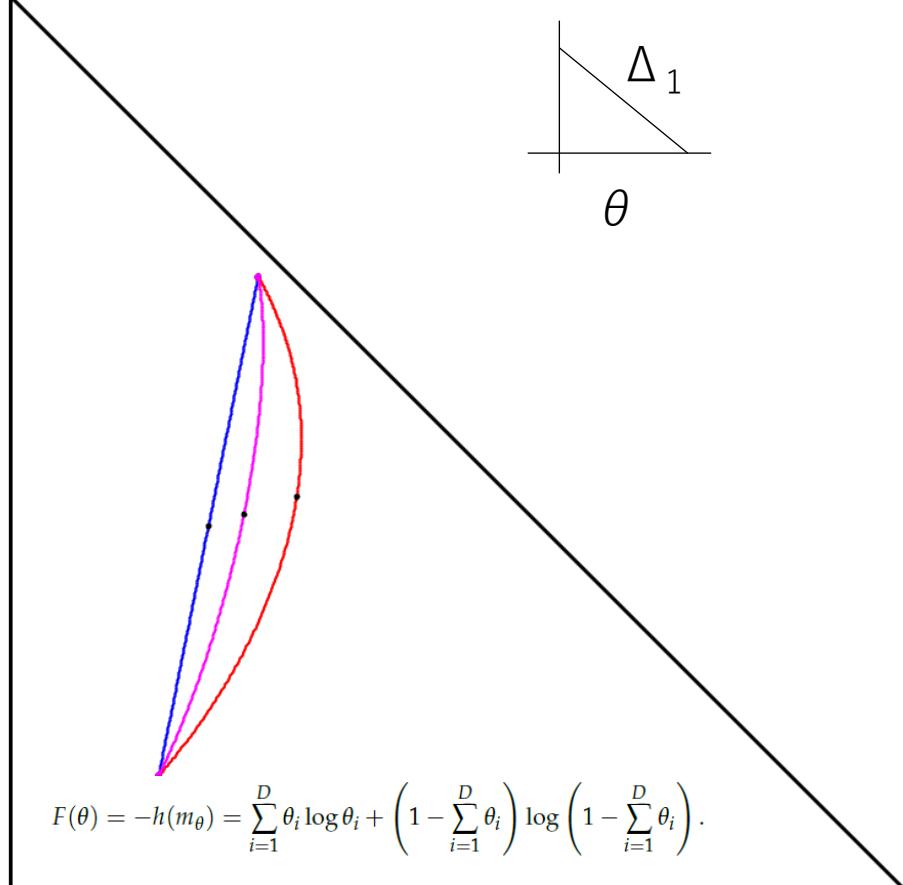
- Given a set of  $n$  discrete distributions (categorical distributions, normalized histograms), calculate its **Jensen-Shannon centroid**

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

$$\begin{aligned} \text{JS}(p, q) &= h\left(\frac{p+q}{2}\right) - \frac{h(p) + h(q)}{2} \\ h(p) &= -\int p \log p d\mu \end{aligned}$$

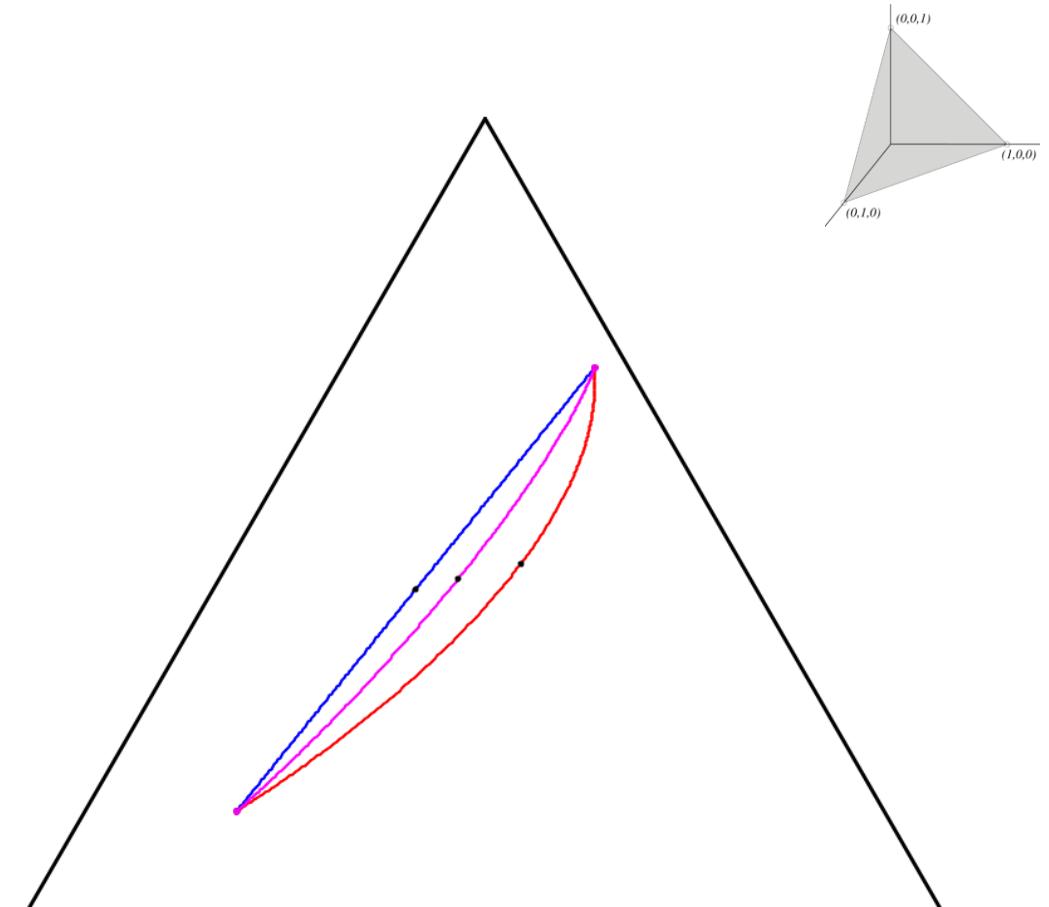
# Dual geodesics and Fisher-Rao geodesics on the categorical manifold

Mixture parameter space

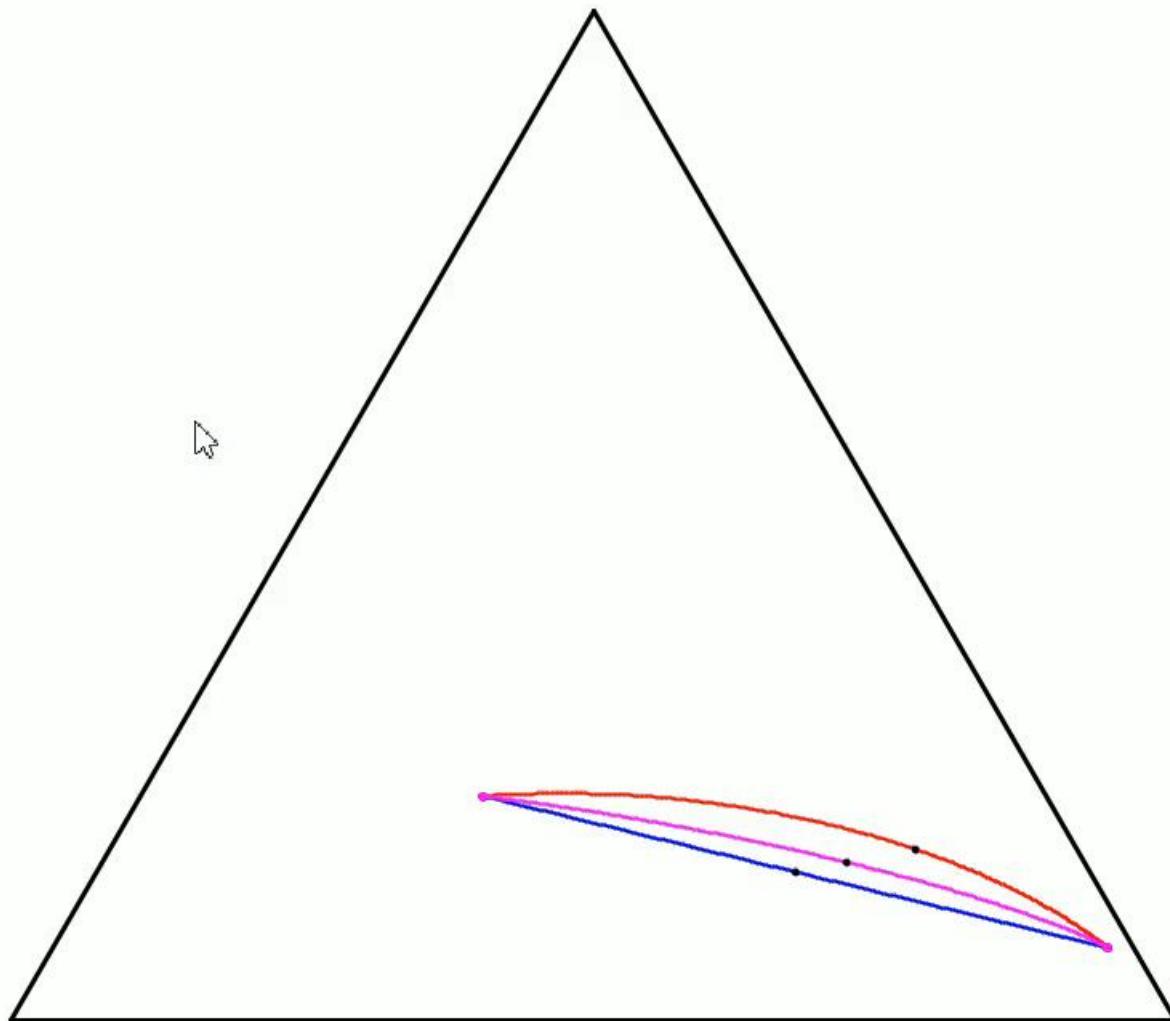


Exponential  $\nabla$ -geodesic  
Mixture  $\nabla^*$ -geodesic  
Fisher-Rao  $\nabla^g$ -geodesic (Levi-Civita )

Probability simplex/Categorical manifold

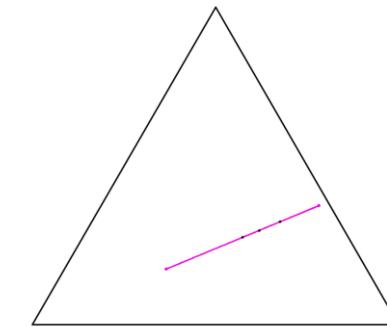


## Probability simplex/Categorical manifold

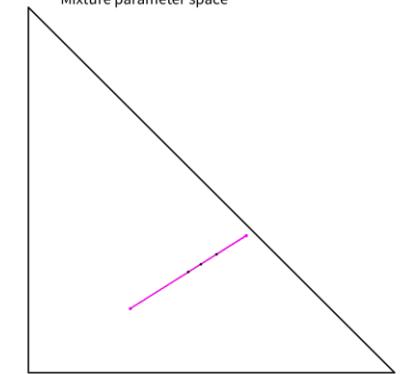


Geodesics coincide when passing through a simplex vertex but not the midpoints

Probability simplex/Categorical manifold



Mixture parameter space



Exponential  $\nabla$ -geodesic

Mixture  $\nabla^*$ -geodesic

Fisher-Rao  $\nabla^g$ -geodesic  
(Levi-Civita )

# Jensen-Shannon centroid for mixtures

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

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

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

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

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

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

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

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

# Jensen-Bregman divergence

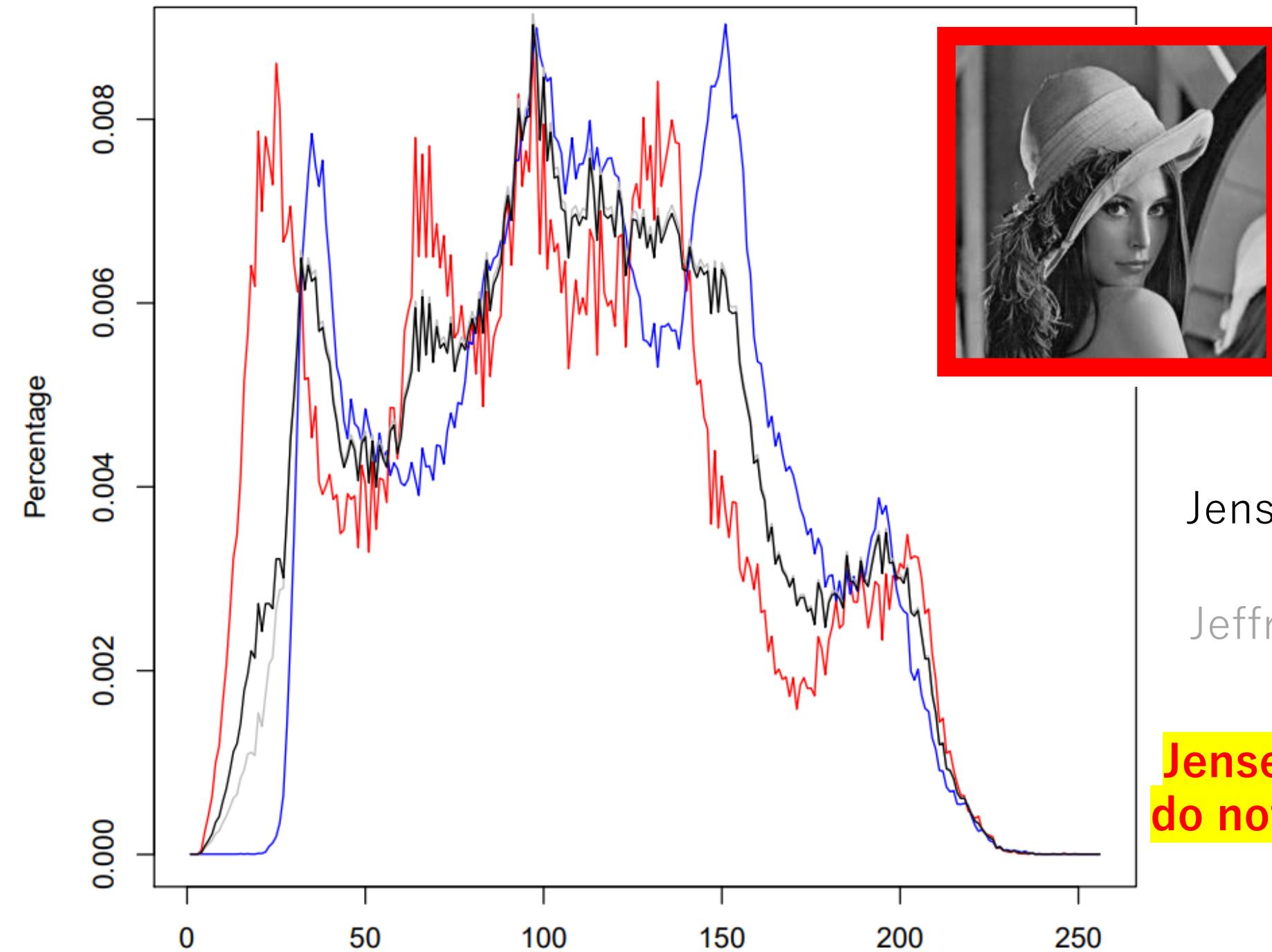
$$\begin{aligned} & \min_p \sum_i \text{JS}(p_i, p), \\ & \min_{\theta} \sum_i J_F(\theta_i, \theta), \\ & \min_{\theta} \sum_i \frac{F(\theta_i) + F(\theta)}{2} - F\left(\frac{\theta_i + \theta}{2}\right) \end{aligned}$$

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

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

amounts to a Jensen divergence (also called Burbea-Rao divergence).

- Jensen-Shannon centroid of a mixture family = Jensen-Bregman centroid = Jensen centroid



Jensen–Shannon centroid

Jeffreys SKL centroid

**Jensen–Shannon centroid  
do not require same support**

# Information geometry in action! (2/2)

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

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

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

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

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

- $\alpha$ -Bhattacharrya distances related to **Renyi  $\alpha$ -divergences**:

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

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

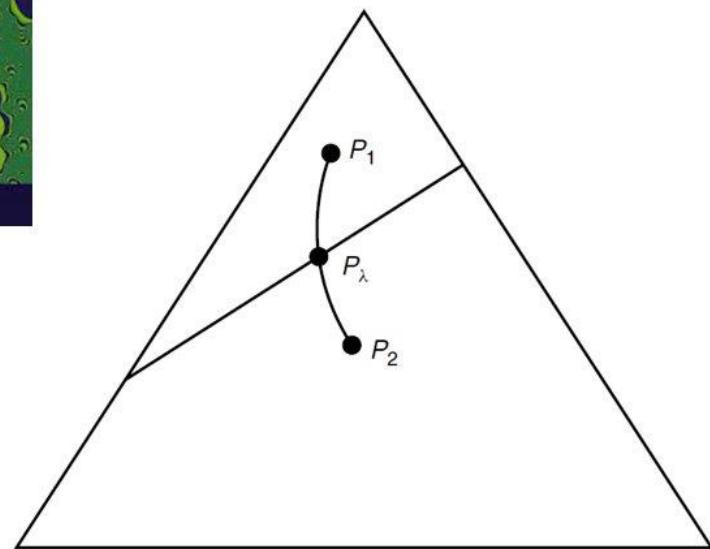
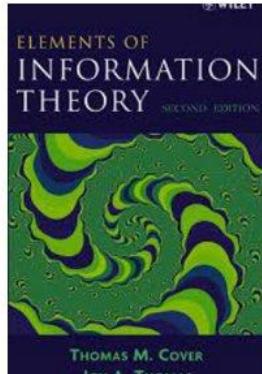
- CI is often used in **Bayesian hypothesis testing & information fusion**

- An information-geometric characterization of Chernoff information, *IEEE Signal Processing Letters* (2013)
- Revisiting Chernoff information with likelihood ratio exponential families, *Entropy* 24.10 (2022)
- Julier, An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models , *IEEE Information Fusion* 2006.

# Chernoff information: A geometric characterization

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

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



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

Probability simplex

Generalized  
to  
exponential family  
manifold

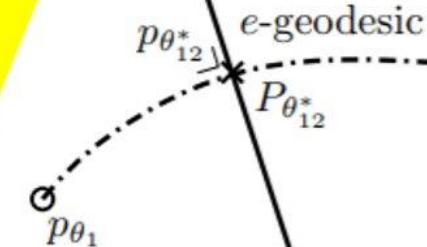
$$C(P, Q) = - \log \min_{\alpha \in (0,1)} \int p^\alpha(x) q^{1-\alpha}(x) d\nu(x).$$

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

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

$\eta$ -coordinate system

$m$ -bisector  
 $\text{Bi}_m(P_{\theta_1}, P_{\theta_2})$



Example:  
Gaussian manifold

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

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

Exponential family manifold

# Chernoff-Bregman divergence

- Chernoff information:  $C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})$
- Chernoff-Bregman is another way to symmetrize a Bregman divergence by maximizing the skew Jensen divergence:

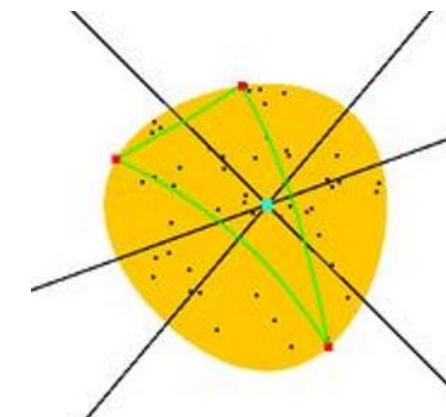
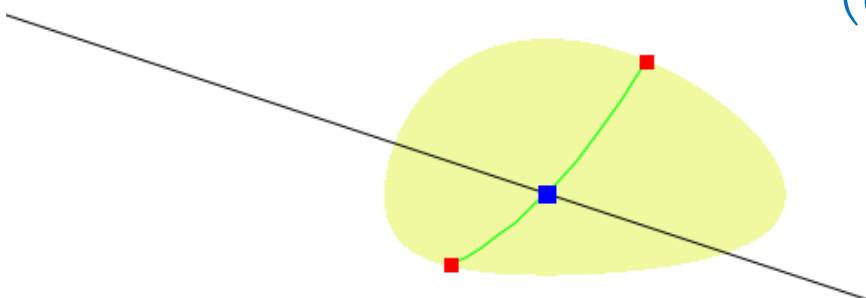
$$C_F(\theta_1, \theta_2) = \max_{\alpha \in (0,1)} J_{F,\alpha}(\theta_1 : \theta_2)$$

$$C_F(\theta_1, \theta_2) = \min_{\theta} \{B_F(\theta_1 : \theta), B_F(\theta_2 : \theta)\}$$

- This amounts to

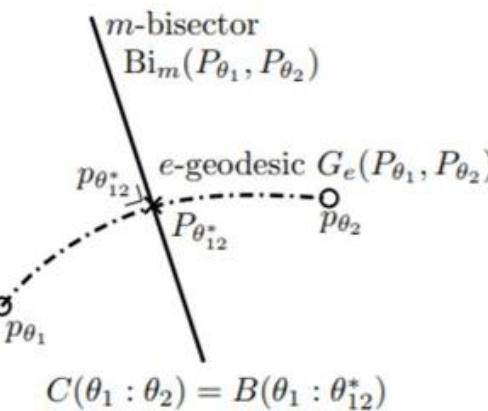
- Chernoff-Bregman divergence is the radius of a right-sided Bregman ball of two points

Chernoff point  $r^*$   
(eKLD)



Exact smallest enclosing  
Bregman ball of n points

Fitting the smallest enclosing Bregman ball, ECML'05

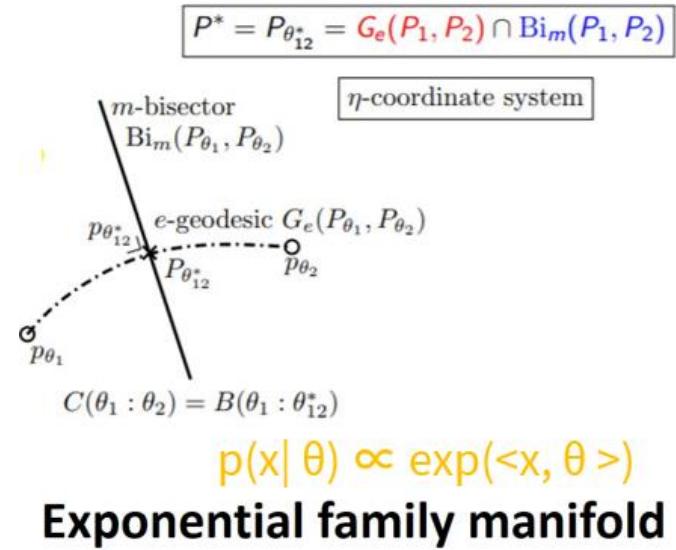


# Categorical distributions: Both an exponential and a mixture family!

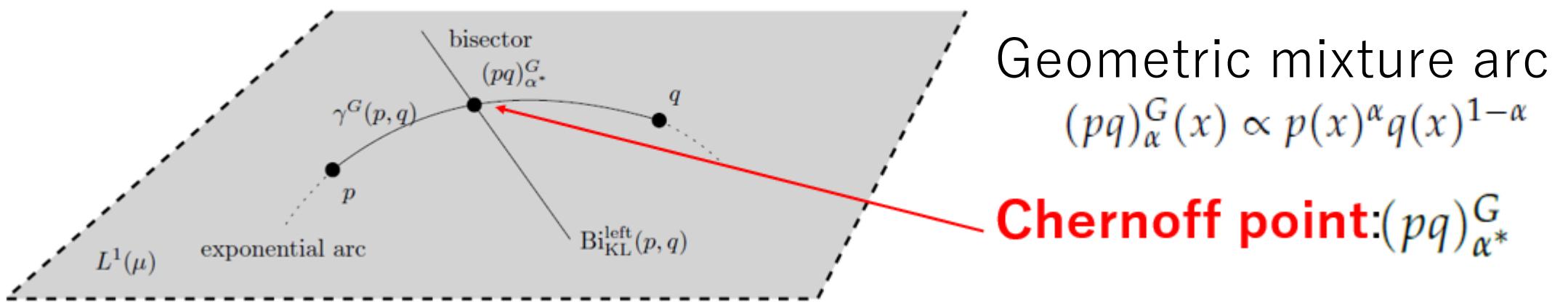
	<b>Exponential Family</b>	*	<b>Mixture Family</b>
pdf	$p_\theta(x) = \prod_{i=1}^d p_i^{t_i(x)}$ , $p_i = \Pr(x = e_i)$ , $t_i(x) \in \{0, 1\}$ , $\sum_{i=1}^d t_i(x) = 1$		$m_\theta(x) = \sum_{i=1}^d p_i \delta_{e_i}(x)$
primal $\theta$	$\theta_i = \log \frac{p_i}{p_d}$		$\theta_i = p_i$
$F(\theta)$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

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

# Revisiting Chernoff information/Point



Consider p and q arbitrary probability densities:



# Geometric mixture arc is a 1D likelihood ratio exponential family

- Geometric mixture Bhattacharyya / exponential arc )  $(pq)_{\alpha}^G(x) \propto p(x)^{\alpha}q(x)^{1-\alpha}$   
between two densities  $p, q$  of Lebesgue Banach space  $L_1(\mu)$
- Set of **geometric mixtures**:  
with **normalization factor**:  

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

$$Z_{pq}(\alpha) = \int_{\mathcal{X}} p(x)^{\alpha}q(x)^{1-\alpha} d\mu(x) = \underline{\rho_{\alpha}[p : q]}$$
- geometric mixture interpreted as a **1D exponential family**: LREF

$$(pq)_{\alpha}^G(x) = \exp \left( \alpha \log \frac{p(x)}{q(x)} - \log Z_{pq}(\alpha) \right) q(x),$$

$$=: \exp(\alpha t(x) - F_{pq}(\alpha) + k(x)).$$

Natural parameter space:

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

# LREFs: EF cumulant function is always analytic

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

- Cumulant function is neg-Bhattacharvva distance:

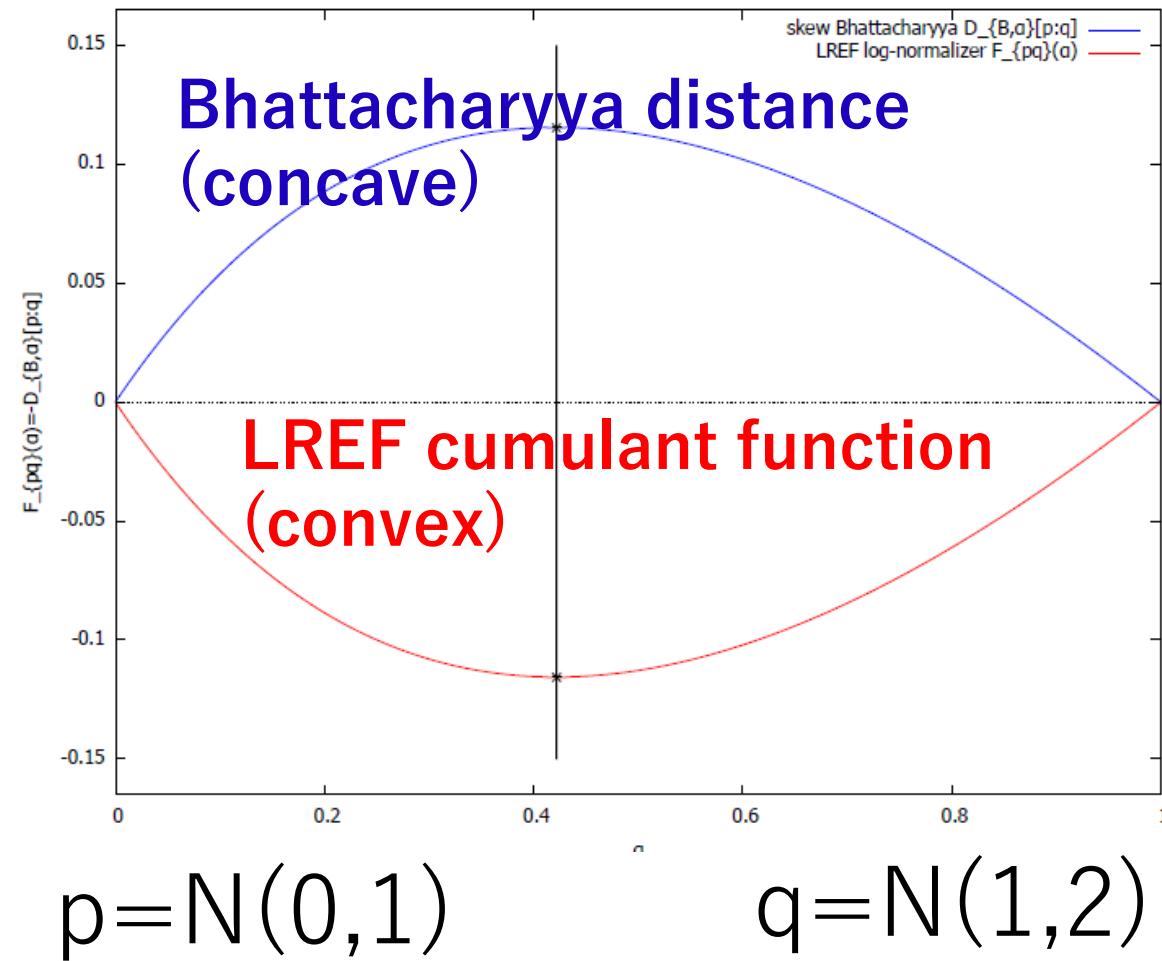
$$F_{pq}(\alpha) = \log Z_{pq}(\alpha) = -D_{B,\alpha}[p : q] < 0$$

⇒ Bhattacharyya. distance is **strictly concave**

- Theorem:

**Chernoff exponent exists and is unique**

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



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

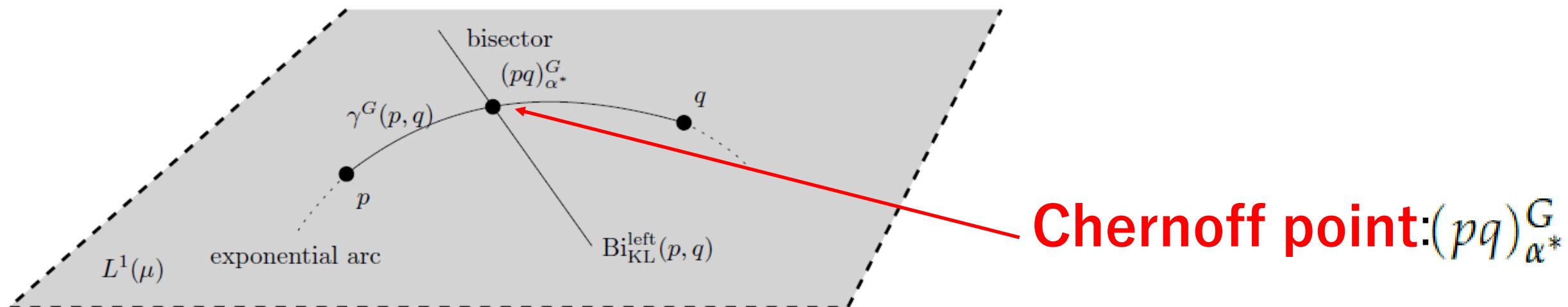
# Geometric interpretation for densities on $L_1(\mu)$

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

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

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

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



# Outline

- Bregman divergences with a single generator and a pair of generators
- Bregman manifolds
  - Normalized/unnormalized exponential families and Bregman divergences
  - Fisher-Rao distance and arithmetic-harmonic mean on the SPD manifold
- Further information geometry in action!
  - Jensen-Shannon centroid on a mixture family manifold
  - Chernoff information and 1D exponential family geometric arc manifolds
- **Bregman divergence with respect to comparative convexity**

# 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  $\alpha$  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

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

for all  $t$  in  $[0,1]$

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

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

for all  $t$  in  $[0,1]$

# Comparative convexity wrt quasi-arithmetic means

- **quasi-arithmetic mean** for a strictly monotone generator  $h(u)$ :

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

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

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

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

Include the **geometric mean** in the limit case  $p \rightarrow 0$

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

# Generalizing Bregman divergences with (M,N)-convexity: (M,N)-Bregman divergences

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

Definition:

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

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

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

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

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

# Generalizing Bregman divergences with quasi-arithmetic mean convexity

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

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

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

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

**Conformal factor**

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

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

# $(M,N)$ -convexity for convex-preserving deformations

- Recall that for exponential families, we had two convex functions:

Convex  $F(\theta) = \log Z(\theta) \leftrightarrow$  Log-convex  $Z(\theta) = \exp(F(\theta))$ , hence convex  
Cumulant function  $\leftrightarrow$  Partition function

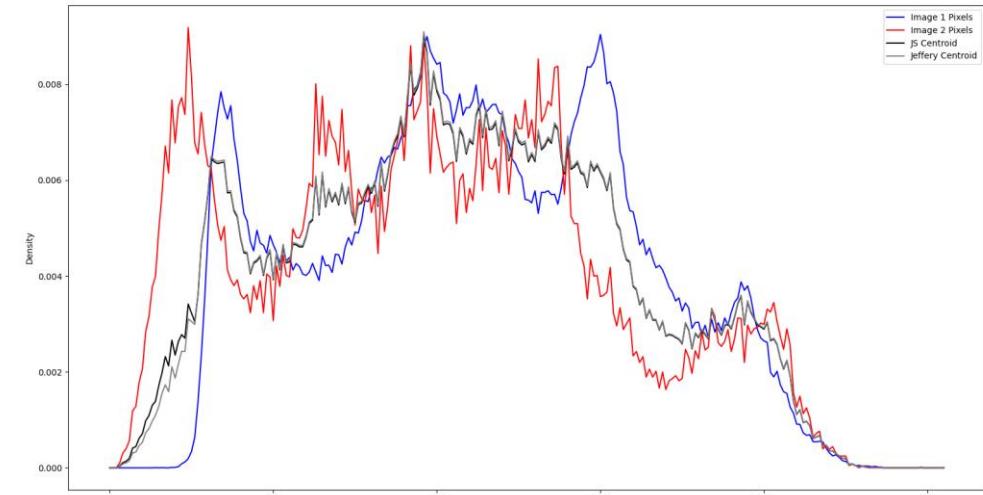
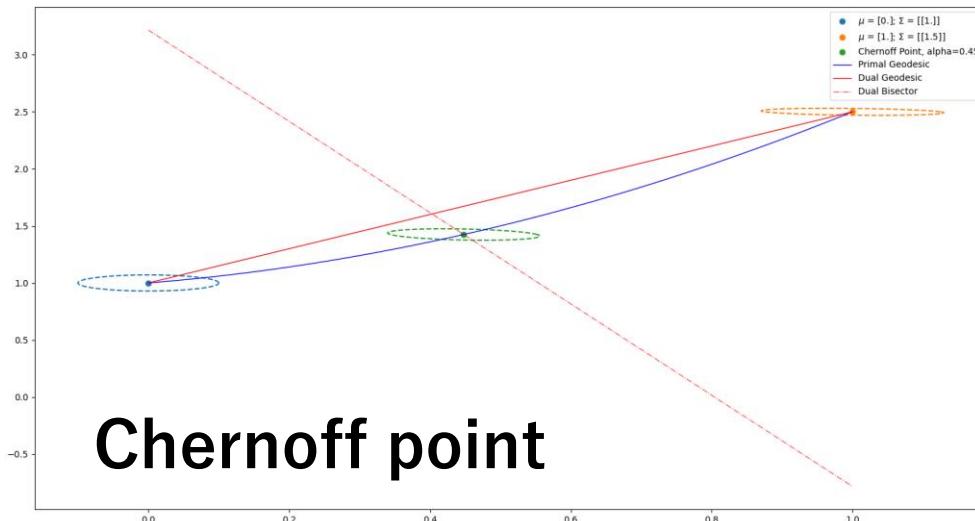
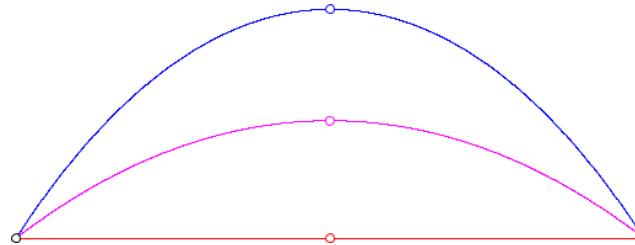
**Proposition 6** ([1, 34]). *A function  $Z(\theta)$  is strictly  $(M_\rho, M_\tau)$ -convex with respect to two strictly increasing smooth functions  $\rho$  and  $\tau$  if and only if the function  $F = \tau \circ Z \circ \rho^{-1}$  is strictly convex.*

- We may thus **deform F with  $(\rho, \tau)$ -quasi-arithmetic means** and seek the  $(\rho, \tau)$ -convex function which also remains ordinary convex
- For those convex-preserving  $(\rho, \tau)$ -deformations, we get new Bregman divergences and Bregman manifolds to play with!

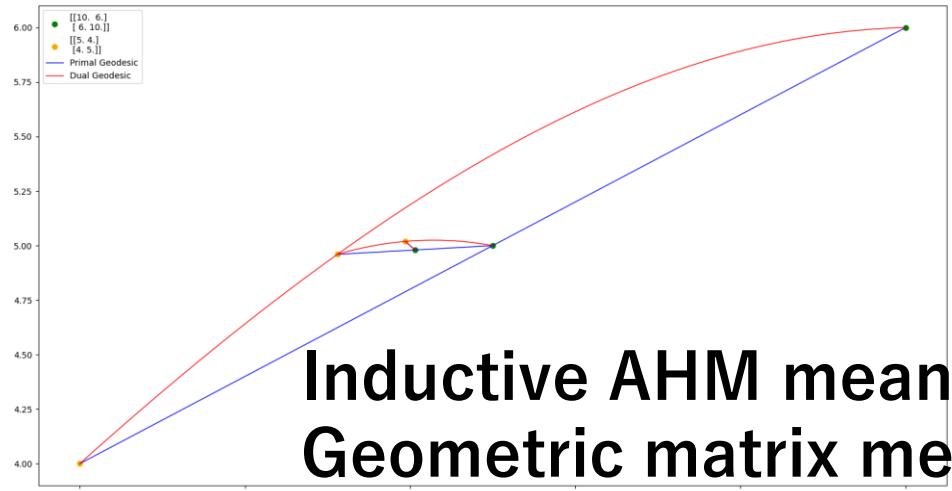
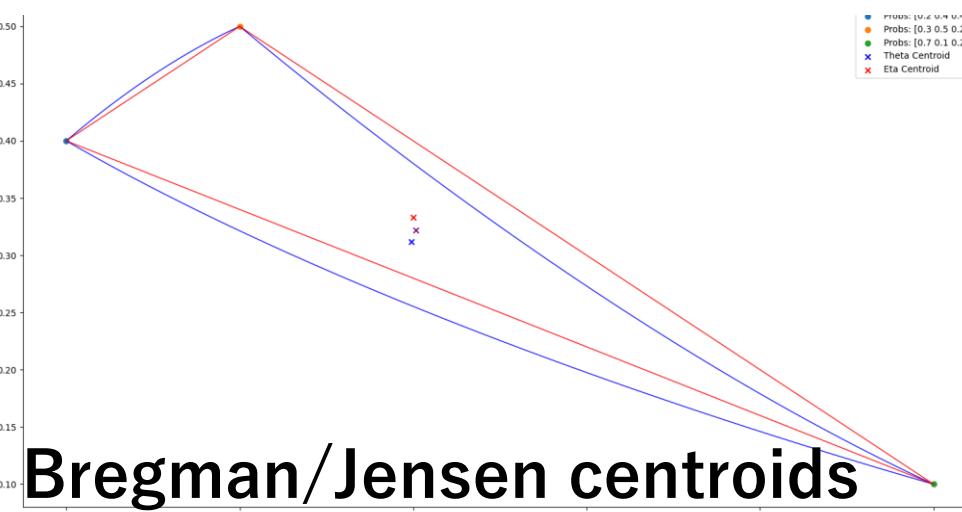
# A Python library for geometric computing on Bregman Manifolds

# pyBregMan

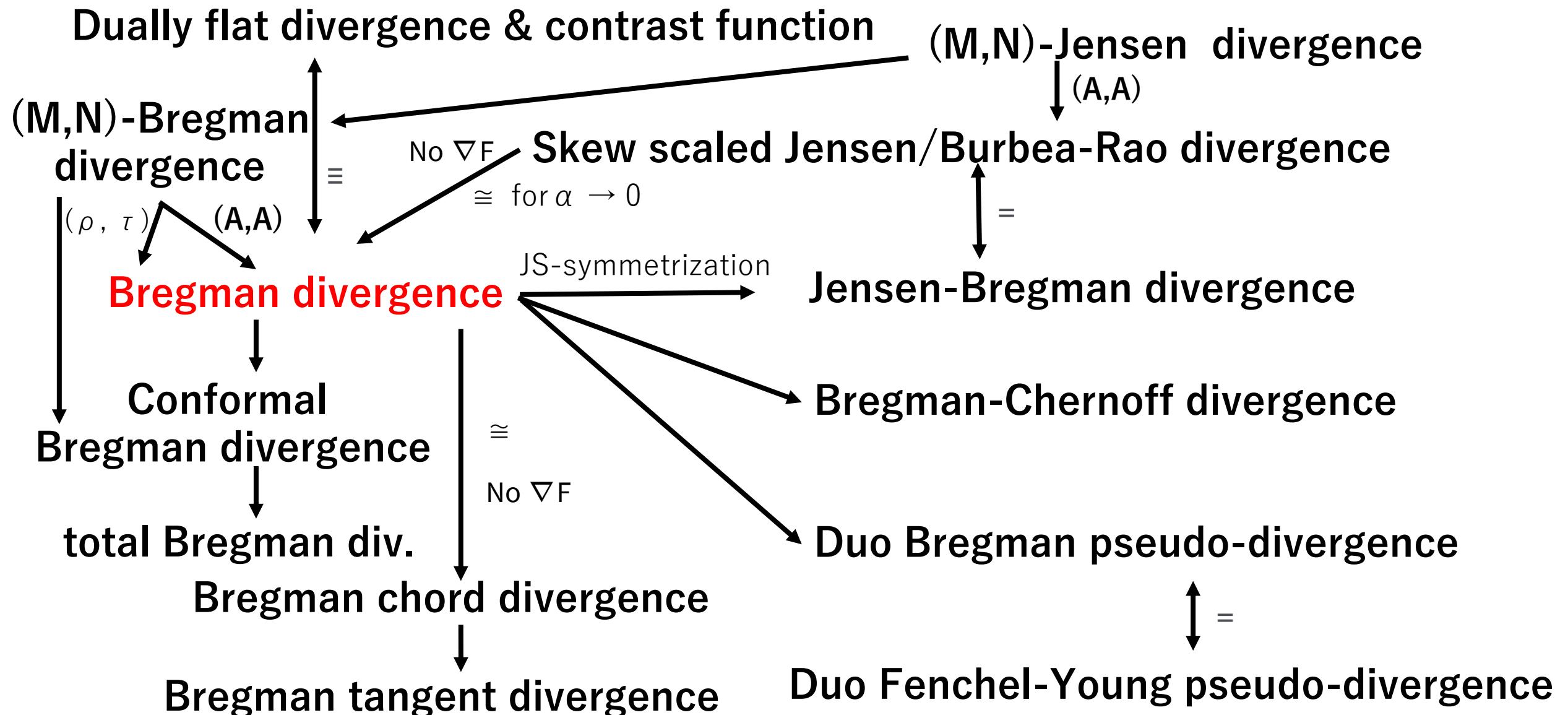
<https://franknielsen.github.io/pyBregMan/>



Jensen-Shannon centroid



# Some generalizations of Bregman divergences in this talk

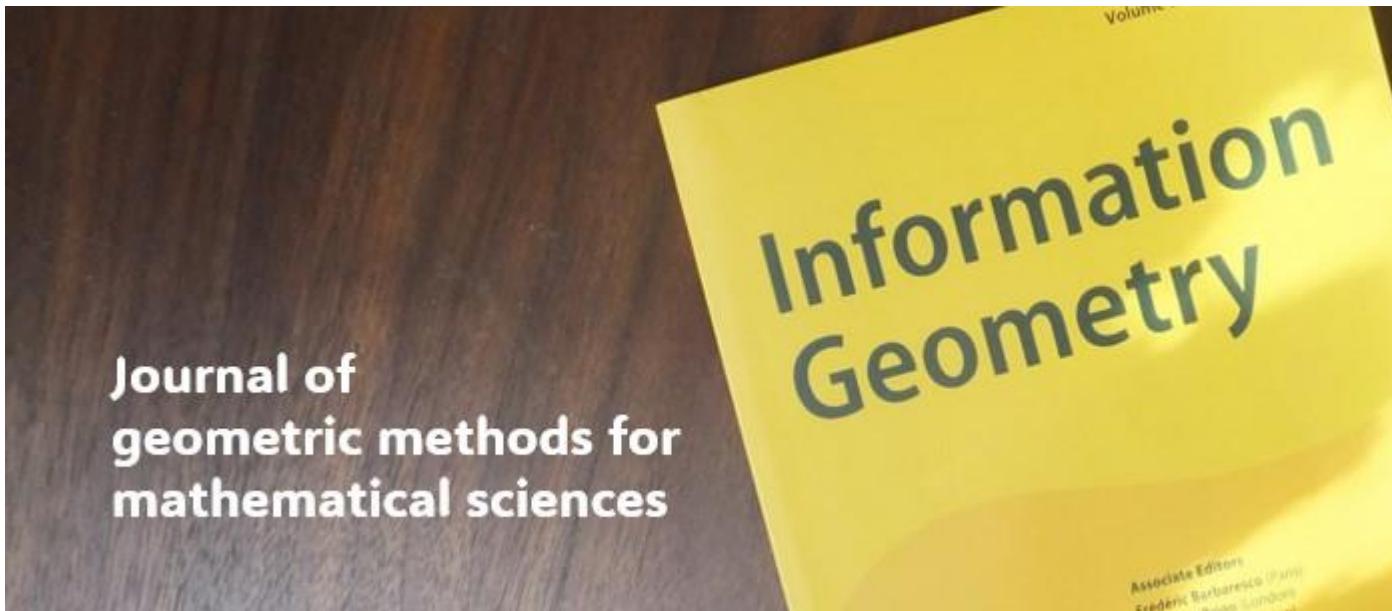


But also matrix Bregman divergence, functional Bregman divergence, etc.

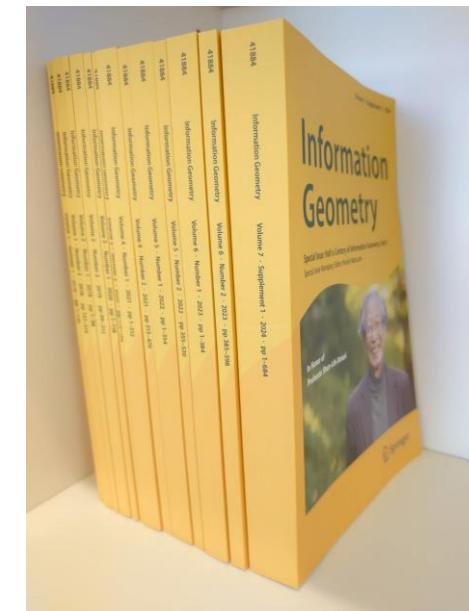
# Thank you!

Many thanks to all my inspiring collaborators.

Special thanks to Richard Nock, Ke Sun, Ehsan Amid, and Alexander Soen



<https://link.springer.com/journal/41884>



<https://franknielsen.github.io/>

# Summary

- Bregman divergences: arbitrarily well approximated by **Bregman chord divergences** or **skewed Jensen divergences** without using gradient  $\nabla F$
- **Jensen-Shannon centroid** on **mixture family manifold** using ConCave-Convex procedure
- 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:  
 **$(M,N)$ -Bregman divergences** as **conformal Bregman divergences**  
**Convex-preserving  $(\rho, \tau)$ -deformations**
- **Duality:** biduality reference/representation from convex duality

# Multivariate Bregman divergence as families of univariate Bregman divergences

- A d-variate function  $F(\theta)$  can be equivalently handled 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  $\nabla_{\theta_2 - \theta_1} F_{\theta_1, \theta_2}(u) = (\mathbf{F}_{\theta, \theta})'(\mathbf{u})$

$$\lim_{\epsilon \rightarrow 0} \frac{F(\theta_1 + (\epsilon + u)(\theta_2 - \theta_1)) - F(\theta_1 + u(\theta_2 - \theta_1))}{\epsilon}$$
$$= (\theta_2 - \theta_1)^\top \nabla F(\theta_1 + u(\theta_2 - \theta_1)).$$

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**

# Symmetrized Bregman divergences are not Bregman divergences except for quadratic distances

- **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:

$$S_F(\theta_1, \theta_2) = B_{\hat{F}}(\theta_1^\uparrow : \theta_2^\uparrow) \quad \xi = \begin{bmatrix} \theta \\ \eta \end{bmatrix} \quad \hat{F}(\xi) = F(\theta) + F^*(\eta) \quad \theta^\uparrow = \begin{bmatrix} \theta \\ \nabla F(\theta) \end{bmatrix}$$

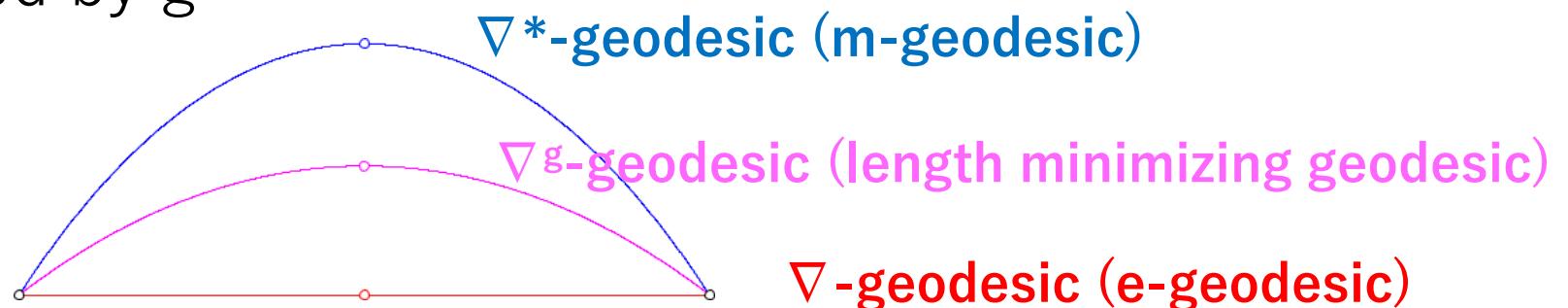
- parameter space  $\Theta^\uparrow$  is **not convex** in general !

$$\Theta^\uparrow = \left\{ \theta^\uparrow = \begin{bmatrix} \theta \\ \nabla F(\theta) \end{bmatrix} : \theta \in \Theta \right\} \subset \Xi \quad \Xi = \left\{ \xi = \begin{bmatrix} \theta \\ \eta \end{bmatrix} : (\theta, \eta) \in \Theta \times H \right\}$$

- Except for generalized quadratic distances (Mahalanobis), SBDs are not BDs. SBDs are **curved Bregman divergences**. Update  $B_F: \theta \times \text{Int}(\theta)$  to  $\theta \times \text{RellInt}(\theta)$
- Bregman divergence restricted to a linear subspace is **sub-dimensional Bregman divergence** : For example, extended KLD vs KLD on simplex  $\Delta$

# Bregman manifolds have Hessian metrics

- The metric  $g$  of a Bregman manifold  $(M, g, \nabla, \nabla^*)$  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:  
$$g(e_i, e^{*j}) = \delta_{ij}.$$
  
Crouzeix identity:  $\nabla^2 F(\theta) \nabla^2 F^*(\eta(\theta)) = \nabla^2 F(\theta(\eta)) \nabla^2 F^*(\eta) = I$
- Riemannian manifold  $(M, g)$  is not flat with respect to the Levi-Civita connection  $\nabla^g$  induced by  $g$

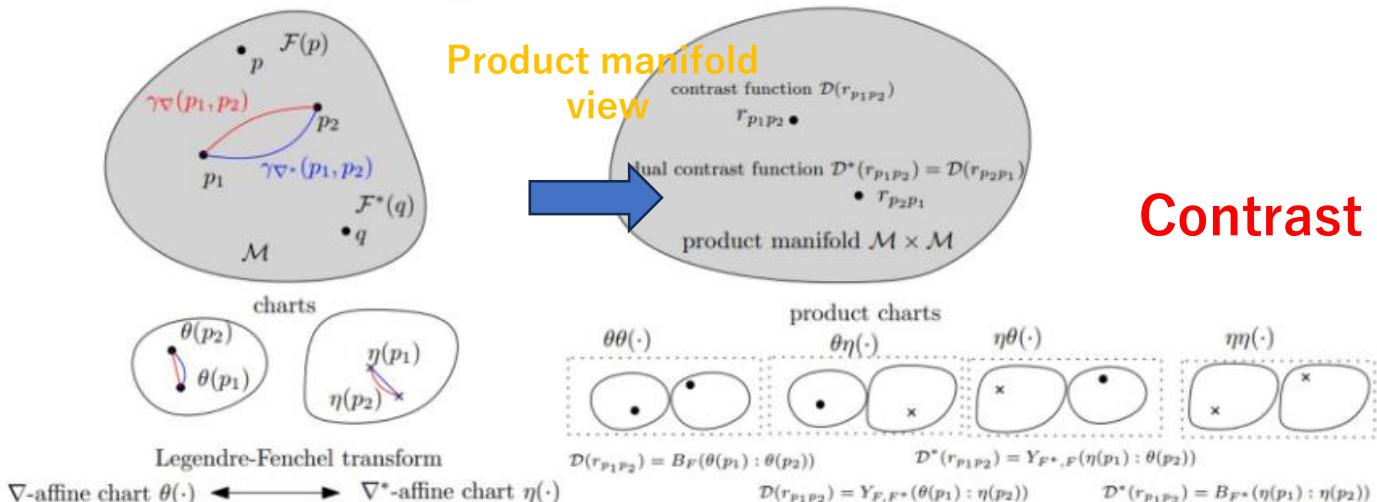


# Bregman manifolds and contrast functions

- A strictly convex and smooth Legendre-type function induces a **dually flat space** also called a global Hessian manifold in diff. geo.

**Dually flat divergence**  $D_{\nabla, \nabla^*}(p : q) = B_F(\theta(p) : \theta(q)) = Y_{F, F^*}(\theta(p) : \eta(q))$

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



**Divergence =  
Contrast function on product manifold**

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

# Monte Carlo Bregman divergences

- In some cases, integral Bregman generators are not available in closed-form or **computationally intractable**. Examples:  
 $F(\theta) = \log \int \exp(\langle \text{Polynomial}(x), \theta \rangle) d\mu(x)$  or  $F(\theta) = \log \sum \exp(\langle x, \theta \rangle)$
- By MC importance sampling on the integral, get with high probability a convex function approximating the generator. Perform algorithms on this randomized Bregman manifolds to get **consistent algorithms**.

$$G(\eta) = -h(m(x; \eta)) = \int_{x \in \mathcal{X}} m(x; \eta) \log m(x; \eta) d\mu(x).$$



$$G(\eta) \simeq \tilde{G}_{\mathcal{S}}(\eta) := \frac{1}{m} \sum_{i=1}^m \frac{1}{q(x_i)} m(x_i; \eta) \log m(x_i; \eta).$$

**Stochastic Bregman generator**

Monte Carlo information-geometric structures

Geometric Structures of Information (2019): 69-103.

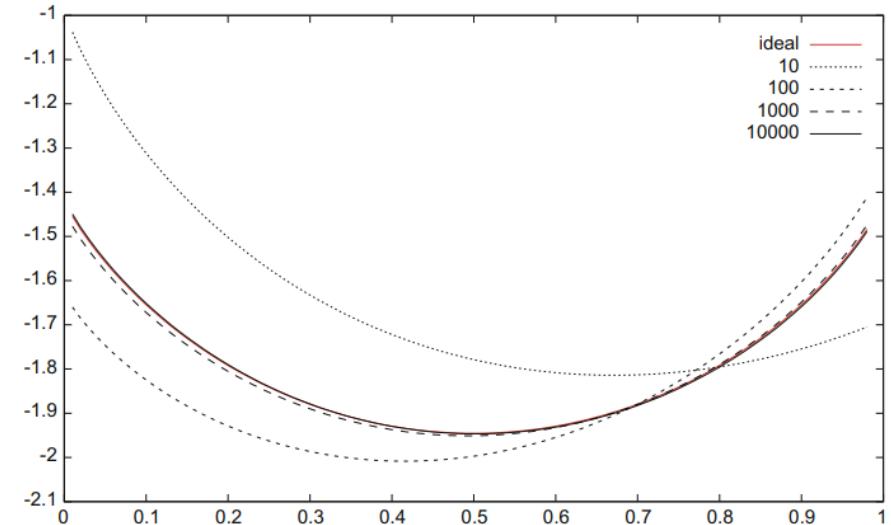
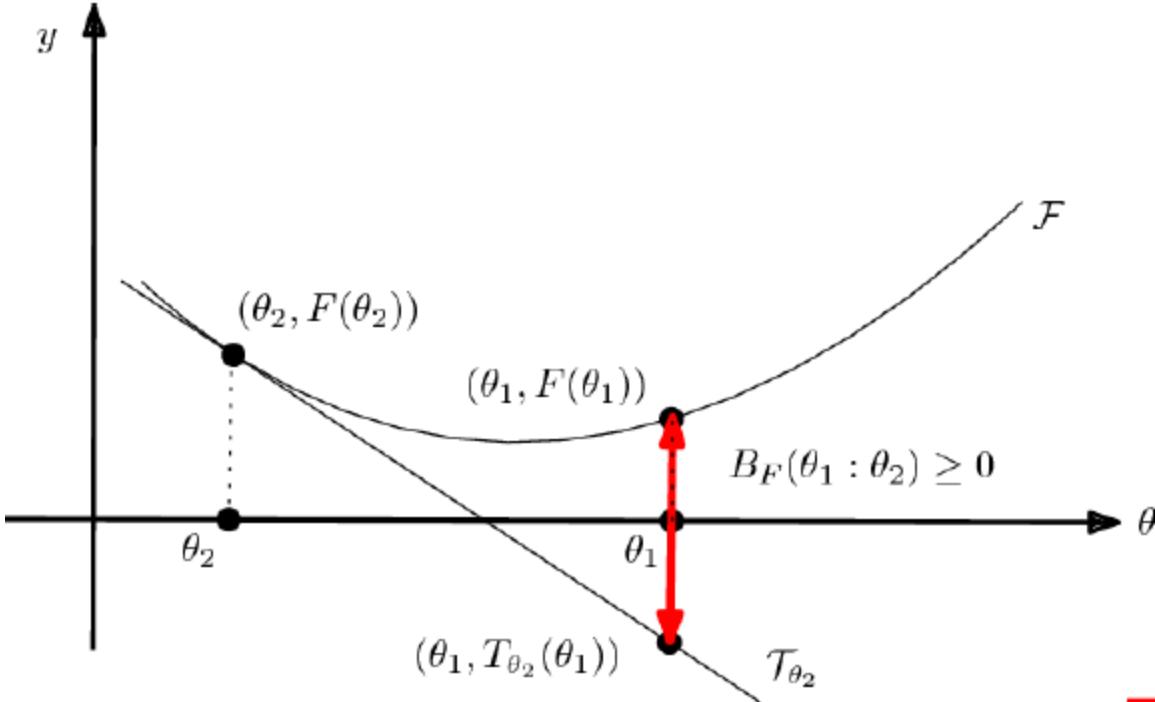


Fig. 2 A series  $G_{\mathcal{S}}(\eta)$  of Bregman Monte Carlo Mixture Family generators (for  $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$ ) approximating the untractable ideal negentropy generator  $G(\eta) = -h(m(x; \eta))$  (red) of a mixture family with prescribed Gaussian distributions  $m(x; \eta) = (1 - \eta)p(x; 0, 3) + \eta p(x; 2, 1)$  for the proposal distribution  $q(x) = m(x; \frac{1}{2})$

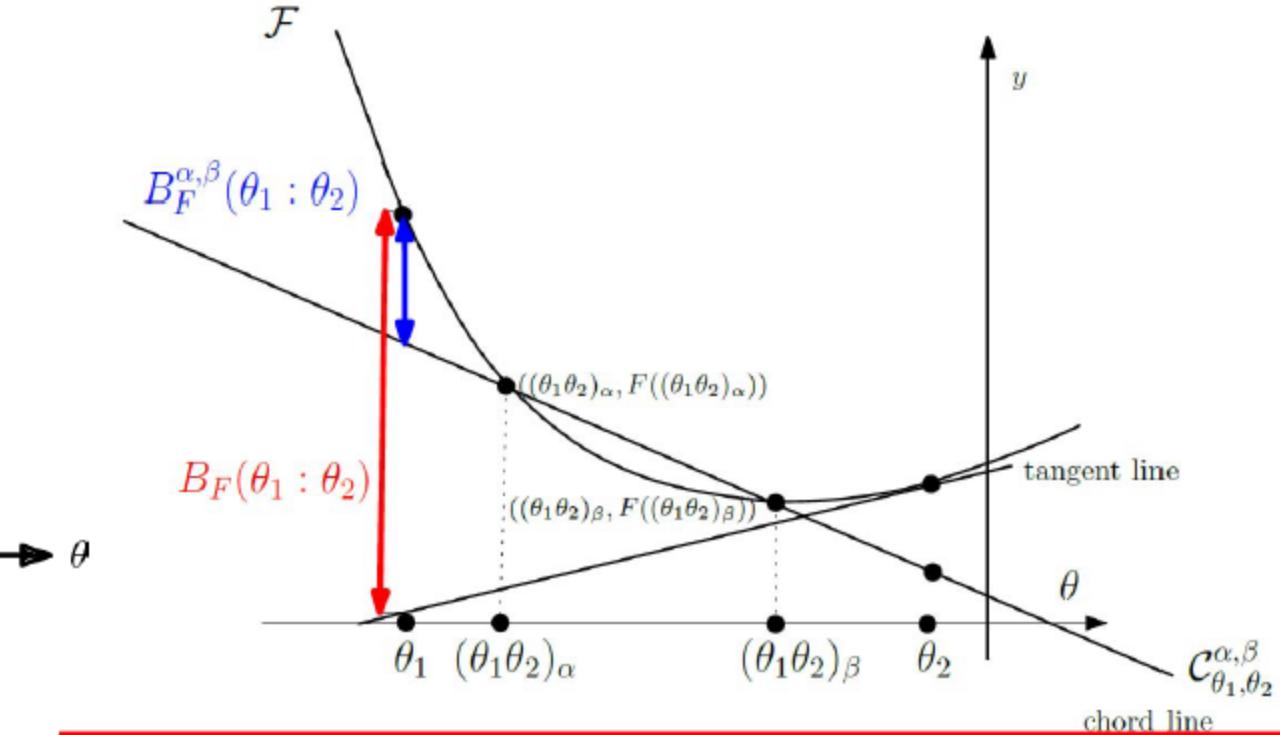
# Scalar Bregman chord divergences

Idea: *Get rid of the gradient* in the Bregman formula (yet approximate BD)



$$B_F(\theta_1 : \theta_2) = F(\theta_1) - T_{\theta_2}(\theta_1)$$

$$T_\theta(\omega) := F(\theta) + (\omega - \theta)F'(\theta)$$



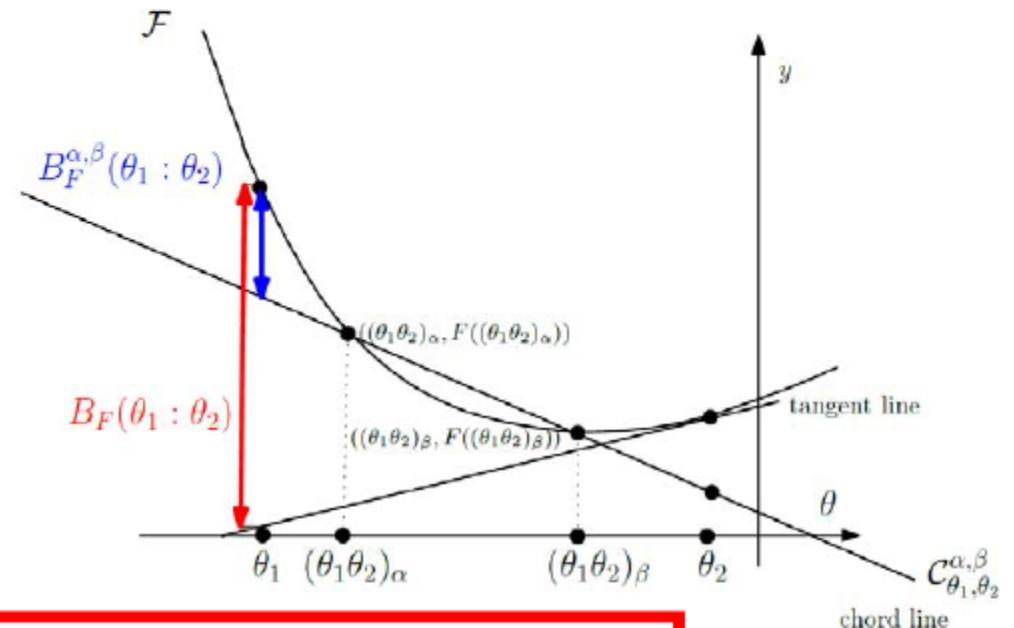
$$B_F^{\alpha, \beta}(\theta_1 : \theta_2) := F(\theta_1) - C_F^{(\theta_1 \theta_2)_\alpha, (\theta_1 \theta_2)_\beta}(\theta_1)$$

$$\rightarrow B_F^{\alpha, \beta}(\theta_1 : \theta_2) \leq B_F(\theta_1 : \theta_2)$$

Linear interpolation (LERP):  $(pq)_\lambda := (1 - \lambda)p + \lambda q$

# Scalar Bregman chord divergences

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



$$\begin{aligned} B_F^{\alpha,\beta}(\theta_1 : \theta_2) &:= F(\theta_1) - \Delta_F^{\alpha,\beta}(\theta_1, \theta_2)(\theta_1 - (\theta_1\theta_2)_\alpha) - F((\theta_1\theta_2)_\alpha), \\ &= F(\theta_1) - F((\theta_1\theta_2)_\alpha) + \frac{\alpha \{F((\theta_1\theta_2)_\alpha) - F((\theta_1\theta_2)_\beta)\}}{\beta - \alpha} \end{aligned}$$

where the *slope of the chord* is

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

# Scalar Bregman chord divergences: Properties

- Symmetry:

$$B_F^{\alpha,\beta}(\theta_1 : \theta_2) = B_F^{\beta,\alpha}(\theta_1 : \theta_2)$$

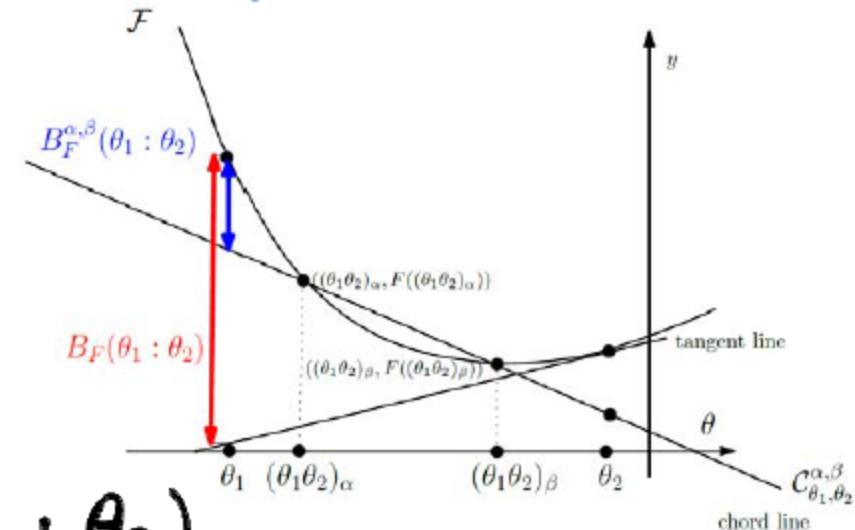
- Generalization of BD:

$$\lim_{\alpha \rightarrow 0, \beta \rightarrow 1} B_F^{\alpha,\beta}(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_2).$$

- Subfamily of **Bregman tangent divergences**:

$$B_F^\alpha(\theta_1 : \theta_2) = \lim_{\beta \rightarrow \alpha} B_F^{\alpha,\beta}(\theta_1 : \theta_2) \leq B_F(\theta_1 : \theta_2)$$

$$\begin{aligned} B_F^\alpha(\theta_1 : \theta_2) &:= F(\theta_1) - F((\theta_1 \theta_2)_\alpha) - (\theta_1 - (\theta_1 \theta_2)_\alpha)^\top \nabla F((\theta_1 \theta_2)_\alpha), \\ &= F(\theta_1) - F((\theta_1 \theta_2)_\alpha) - \alpha(\theta_1 - \theta_2)^\top \nabla F((\theta_1 \theta_2)_\alpha), \end{aligned}$$



# 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**:  
**(M,N)- Bregman divergences as conformal Bregman divergences**
- **Duality and conjugacies** (convex duality, reference/representation biduality, connection duality) are at the heart of information geometry!

# 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  $\sum_i D_{JS}(c, p_i)$

/\* 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

$t \leftarrow 0$ ;

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

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

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

$t \leftarrow t + 1$ ;

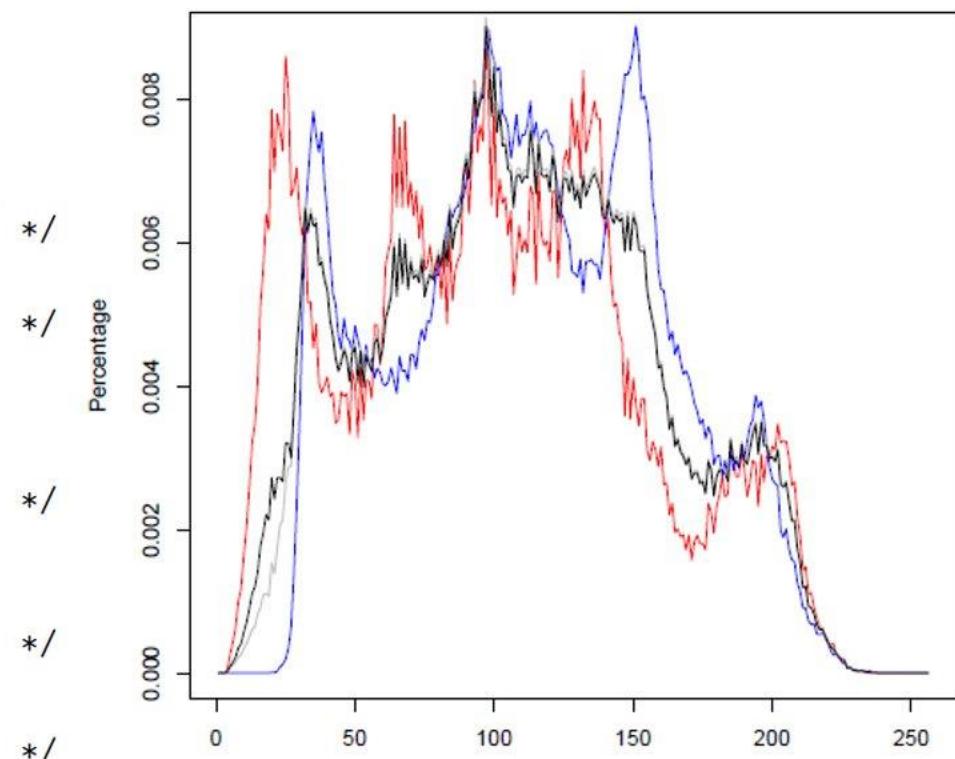
end

/\* 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**

# Inductive matrix arithmetic-harmonic mean (AHM)

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

[Nakamura 2001]

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

$$H_{t+1} = 2(A_t^{-1} + H_t^{-1})^{-1} = H(A_t, H_t) \quad \leftarrow \text{harmonic mean}$$

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

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

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

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

$$G(X, Y) = \arg \min_{M \in \mathbb{P}(d)} \frac{1}{2} \rho^2(X, M) + \frac{1}{2} \rho^2(Y, M). \quad \rho(P_1, P_2) = \sqrt{\sum_{i=1}^d \log^2 \lambda_i (P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})} \quad \text{Riemannian distance}$$

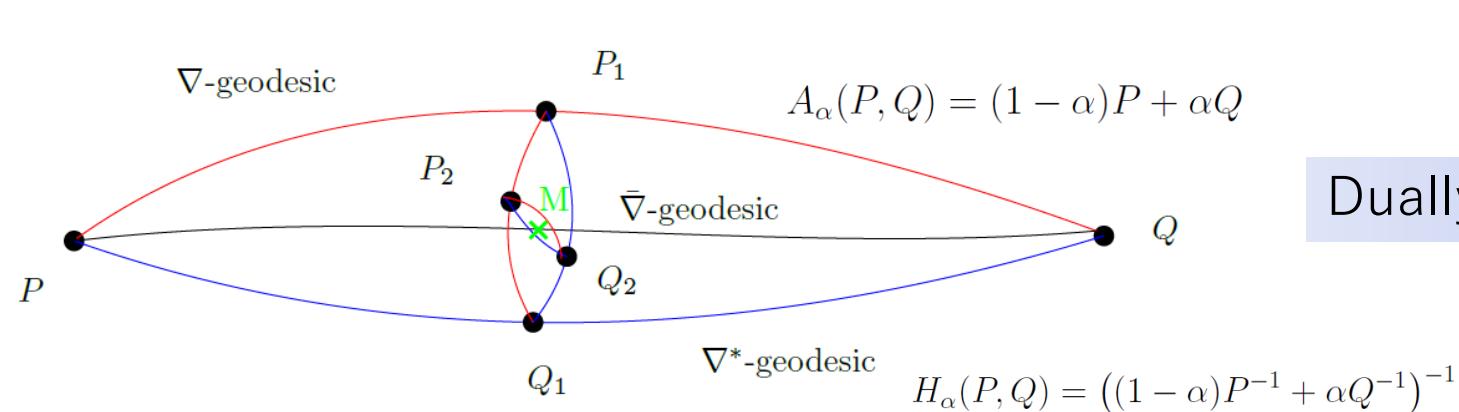
# Geometric interpretation of the AHM matrix mean

Repeat:

$$A_{t+1} = \frac{A_t + H_t}{2} = A(A_t, H_t) \quad P_{t+1} = \gamma \left( P_t, Q_t : \frac{1}{2} \right)$$

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

**(SPD,  $g^G$ ,  $\nabla^A$ ,  $\nabla^H$ ) is a dually flat space,  $\nabla^G$  is Levi-Civita connection**



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

Dually flat space (SPD,  $g^G$ ,  $\nabla^A$ ,  $\nabla^H$ )

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**

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

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

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

[Nakamura 2001]

MiniBBall: Smallest Enclosing Bregman Ball by F. Nielsen and R. Nock

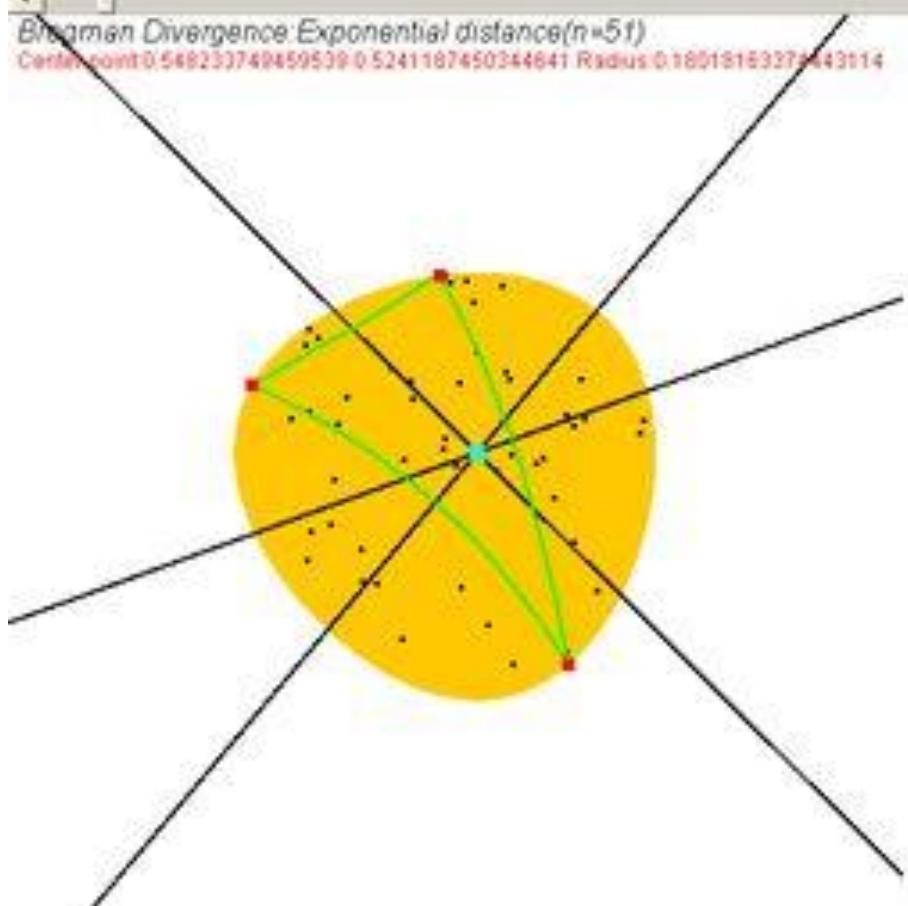
Refresh Draw a new Set

Mini Bregman Ball

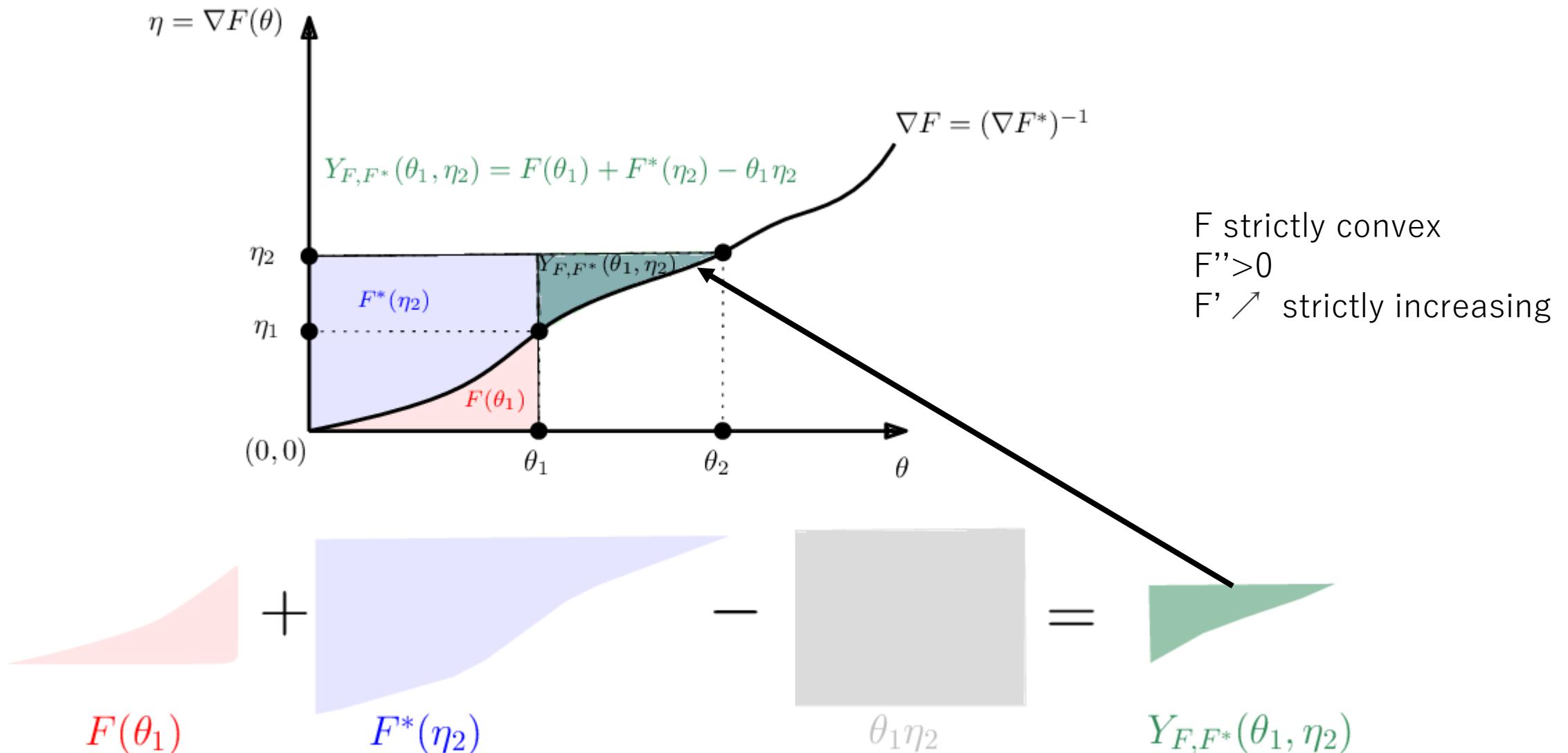
Bregman divergence:

- Squared Euclidean
- KullBack-Leibler
- Exponential
- Itakura-Saito
- Csiszar
- Mahalanobis
- Logistic Loss

Bregman Divergence: Exponential distance( $n=51$ )  
Center point 0.56823374945953905241187450344641 Radius 0.18018153379443114



# Visual interpretation of Fenchel-Young divergences



# Invariance of Bregman divergences

I ❤

BD

# Bregman divergence (1967)

- Let  $F: \Theta \subseteq \mathbb{R}^m \rightarrow \mathbb{R}$  be a strictly convex and smooth real-valued function on a Hilbert space  $\langle \cdot, \cdot \rangle$
- **Bregman divergence**  $B_F: \Theta \times \text{Int}(\Theta) \rightarrow \mathbb{R}$  defined by

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



**Lev M. Bregman**

(1941 - 2023)

Photo: courtesy of Alexander Fradkov

- BDs are non-negative because = remainder of Taylor 1<sup>st</sup> order expansion:  
$$F(\theta_1) = F(\theta_2) + \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle + \text{Remainder}_F(\theta_1 : \theta_2)$$
  
BD = exact Lagrange remainder = a quadratic distance:  
$$\text{LagrangeRemainder}_F(\theta_1 : \theta_2) = \frac{1}{2} (\theta_1 - \theta_2)^T \nabla^2 F(\xi) (\theta_1 - \theta_2) = B_F(\theta_1 : \theta_2)$$
 for some  $\xi \in [\theta_1, \theta_2]$
- BDs are *never metrics* and *only symmetric for generalized quadratic distance*  
$$F_Q(\theta) = \frac{1}{2} \theta^T Q \theta$$
, where  $Q$  is symmetric positive-definite matrix  
$$Q=I$$
, squared Euclidean distance
- BDs are popular! Unify *squared Euclidean divergence* with *Kullback-Leibler divergence*  $F(\theta) = \sum_i \theta_i \log(\theta_i)$  and *Itakura-Saito divergence*  $F(\theta) = \sum_i -\log(\theta_i)$