Computational information geometry on Bregman manifolds and submanifolds

Frank Nielsen

Sony Computer Science Laboratories, Inc.

Applied geometry for data sciences Part II
Singapore
2nd June 2025

Sony CSL

Outline of the talk

- Bregman divergences with some extensions
- Geometry of Bregman balls
- Two applications on Bregman manifolds:
  - Jensen-Shannon centroid on a mixture family manifold
  - Chernoff information/point on an exponential family manifold

Bregman divergences (1960's)

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

- $F: \Theta \subseteq \mathbb{R}^m \to \mathbb{R}$ a strictly convex and smooth

real-valued function on a finite dim. Hilbert space $\langle .,. \rangle$

Bregman divergence $B_F: \Theta \times \mathrm{RelInt}(\Theta) \to \mathbb{R}_{\geq 0}$

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

Smooth measure of discrepancy, not a metric distance because it violates the triangle inequality, and is asymmetric when F is not quadratic function. Hence the delimiter notation ":" instead of $B_F(\theta_1, \theta_2)$

BD interpreted as **remainder** of a first order Taylor expression of $F(\theta_1)$ around θ_2:

$$F(\theta_1) = F(\theta_2) + \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle + \underbrace{B_F(\theta_1 : \theta_2)}_{\text{Taylor remainder}}$$

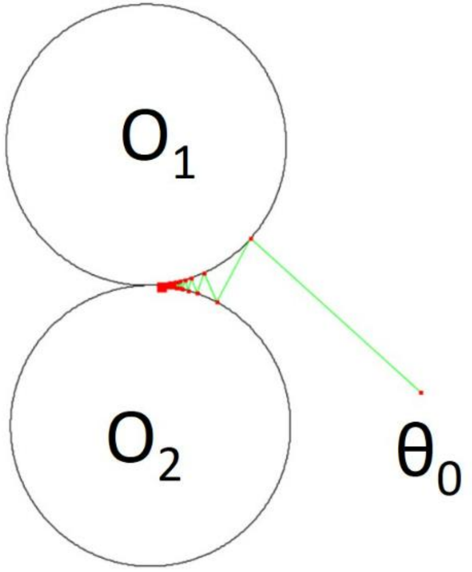
Example of remainder: **Lagrange remainder** (smooth C^2 generators): $\nabla^2 F$ **SPD** $\Rightarrow B_F(\theta_1 : \theta_2) \geq 0$

$$B_F(\theta_1 : \theta_2) = \tfrac{1}{2} (\theta_2 - \theta_1)^T \nabla^2 F(\theta) (\theta_2 - \theta_1) \geq 0, \ \theta \in [\theta_1 \theta_2]$$

BDs: Versatile and popular in OR, ML, IT, signal processing

Originally motivated for finding an intersection point in a set of convex objects using **Bregman projections**.
(ex. of convex objects: halfspaces, balls, etc.)

convex feasibility of
Bregman cyclic projections



$\theta_0 \in \Theta, t \leftarrow 0$

$$\boxed{\theta_{t+1} = \arg \min_{\theta \in O_{1+(t \mod n)}} B_F(\theta_t : \theta)}$$

BDs unify:

- *squared Euclidean divergence* $F(\theta) = \frac{1}{2} \sum_i \langle \theta, \theta \rangle$
- *Kullback-Leibler divergence* $F(\theta) = \sum_i \theta_i \log(\theta_i)$ (relative Shannon entropy)
- *Itakura-Saito divergence* $F(\theta) = \sum_i -\log(\theta_i)$ (relative Burg entropy)

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

L22 (β =2), KLD ($\beta \to 0$), ISD (β =1), belong to a *family* of **β-divergences**, learn ad hoc $\beta \geq 0$

x,y>0, $\beta \geq 0$

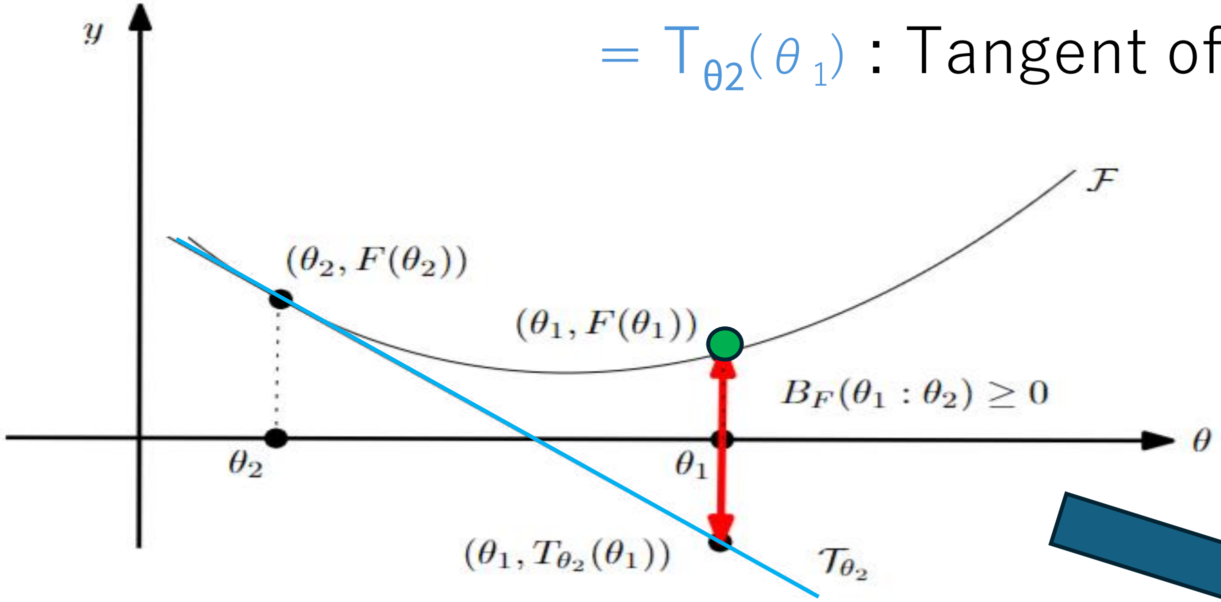
$$d_\beta(x|y) = \begin{cases} \frac{x}{y} - \log(\frac{x}{y}) - 1 & \beta = 0 \\ x(\log x - \log y) + (y - x) & \beta = 1 \\ \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)} & \beta \in \mathbb{R}\backslash\{0,1\} \end{cases}$$

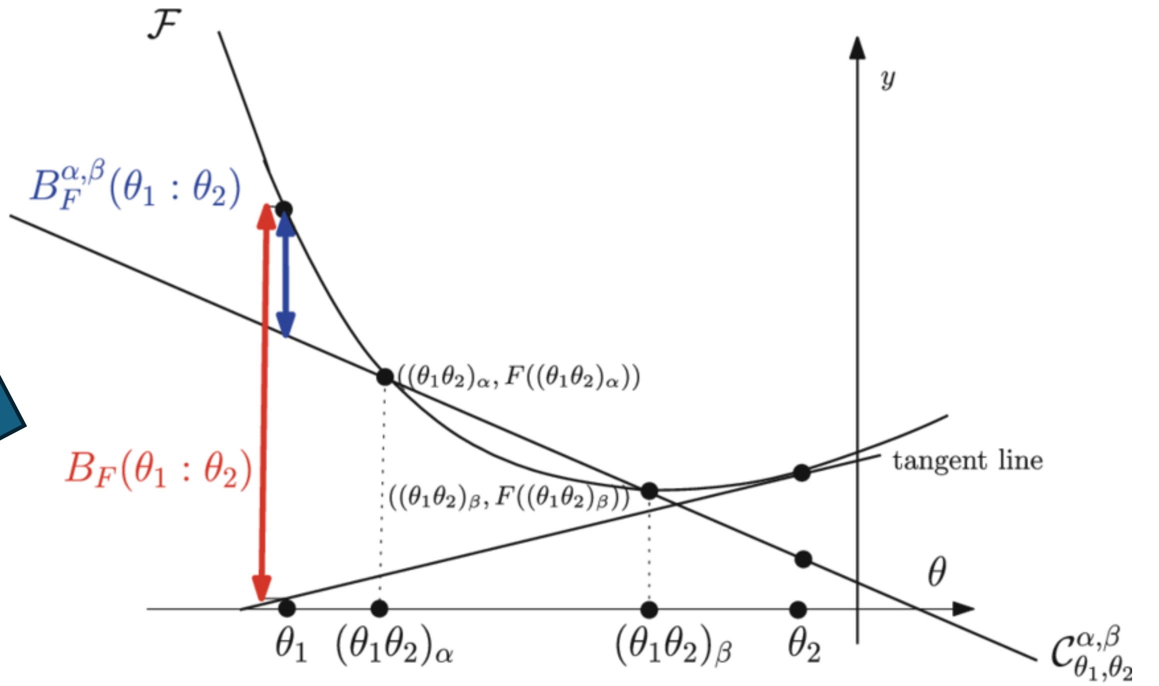
Bregman Generator: $$\phi_\beta(x) = \begin{cases} -\log x + x - 1 & \beta = 0 \\ x \log x - x + 1 & \beta = 1 \\ \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta} & \text{otherwise.} \end{cases}$$

Geometric interpretation as a **vertical gap** using the graph $(\theta, F(\theta))$:

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

$= T_{\theta_2}(\theta_1)$: Tangent of the function graph at θ_2 evaluated at θ_1





Design novel divergences from graph of convex functions···

Example: **Bregman chord divergence**, **application**: zero-order optimization in ML

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

Bregman divergences in machine learning…

- Kullback-Leibler divergence between two probability densities:

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

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

- But Kullback-Leibler divergence between two probability densities of a **natural exponential family** with densities $p(x|\theta) \propto \exp(\langle x, \theta \rangle)$ amount to a **reverse Bregman divergence** $B_F^{rev}(\theta_1 : \theta_2) := B_F(\theta_2 : \theta_1)$

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

Bypass the ∫, ∇F in BD easy to calculate! ⇒ Easy calculations of KLDs

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)

Representational Bregman divergences (2009)

- Use a **representation function** R :

$$B_{F,R}(\lambda_1 : \lambda_2) := B_F(R(\lambda_1) : R(\lambda_2))$$
$$= F(R(\lambda_1)) - F(R(\lambda_2)) - \langle R(\lambda_1) - R(\lambda_2), \nabla F(R(\lambda_2)) \rangle$$

Note that $F \circ R$ may not be a Bregman generator, i.e., not be strictly convex.

For example, consider the KLD between two densities of a **generic exponential family (natural parameter from representation function)**

$$p_\lambda(x) \propto \tilde{p}_\lambda(x) = \exp(\langle \theta(\lambda), t(x) \rangle)\, h(x)$$ include normal, Gamma/Beta, Wishart, Poisson, etc.

$\theta(\lambda)$: natural parameter corresponding to λ, representation function $R(.) = \theta(.)$

$$D_{KL}[p(x|\lambda_1) : p(x|\lambda_2)] = B_F^{rev}(\theta(\lambda_1) : \theta(\lambda_2)) = B_F(\theta(\lambda_2) : \theta(\lambda_1))$$

NEF density $p(x|\theta) \propto \exp(\langle x, \theta \rangle)$ $\quad D_{KL}[p(x|\theta_1) : p(x|\theta_2)] = B_F^{rev}(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$

Extended α -divergences are representational BDs

α -divergences extended to m-dimensional positive measures are **representational Bregman divergences**:

$$D_\alpha^+(q_1:q_2) = \begin{cases} \frac{4}{1-\alpha^2}\sum_{i=1}^m\left(\frac{1-\alpha}{2}q_1+\frac{1+\alpha}{2}q_2-q_1^{\frac{1-\alpha}{2}}q_2^{\frac{1+\alpha}{2}}\right), & \alpha\in\mathbb{R}\backslash\{-1,1\} \\ {D_{\mathrm{KL}}^*}^+(q_1:q_2) = D_{\mathrm{KL}}^+(q_2:q_1) = \sum_{i=1}^m q_2^i\log\frac{q_2^i}{q_1^i}+q_1^i-q_2^i & \alpha=1 \\ D_{\mathrm{KL}}^+(q_1:q_2) = \sum_{i=1}^m q_1^i\log\frac{q_1^i}{q_2^i}+q_2^i-q_1^i & \alpha=-1. \end{cases}$$

$$\boxed{D_\alpha^+(q_1:q_2) = B_{F_\alpha}(R_\alpha(q_1):R_\alpha(q_2))}$$

Bregman generator: $F_\alpha(r)=\sum_{i=1}^m f_\alpha(r_i), \quad f_\alpha(x)=\begin{cases}\frac{2}{1+\alpha}\left(\frac{1-\alpha}{2}x\right)^{\frac{2}{1-\alpha}}, & \alpha\neq 1\\ \log x, & \alpha=1.\end{cases}$

Representation function: $R_\alpha(q)=(r_\alpha(q_1),\ldots,r_\alpha(q_m)), \quad r_\alpha(x)=\frac{2}{1-\alpha}x^{\frac{1-\alpha}{2}}$

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

"The dual Voronoi diagrams with respect to representational Bregman divergences." IEEE ISVD *2009*

Convex duality via Legendre-Fenchel transform

- Legendre-Fenchel transform of a convex function F:

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

- Problem: some *tricky functions* with gradient map ∇F domain not convex…

  Example: $h(\xi_1,\xi_2)=[(\xi_1^2/\xi_2)+\xi_1^2+\xi_2^2]/4$ on upper plane domain $\Xi=(\xi_1,\xi_2)$

- Thus, we consider "**nice convex functions**" = **Legendre-type functions $(\Theta, F(\theta))$**

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

Then we get:

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

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

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

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

Fenchel-Young divergences & convex duality

- 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 ≥0

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

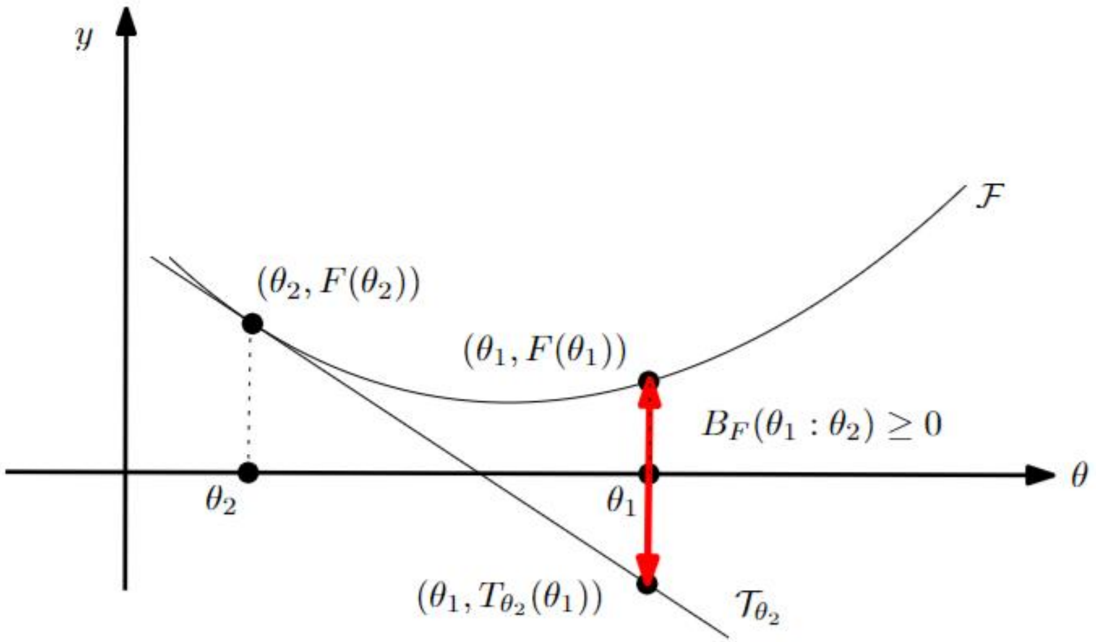
- **Mixed parameterizations** θ and η : $B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1, \eta_2)$
- Duality: $\mathbf{B_F(\theta_1 : \theta_2)} = \mathbf{Y_{F,F^*}(\theta_1, \eta_2)} = \mathbf{Y_{F^*,F}(\eta_2, \theta_1)} = \mathbf{B_{F^*}(\eta_2, \eta_1)}$
- Dual BDs + Dual FYs from involution $F^{**}=F$
- Note : $B_F(\theta_1 : \theta_2)=0 \Leftrightarrow \theta_1 = \theta_2 \Leftrightarrow \eta_1 = \eta_2$ i.e., $\nabla F(\theta_1) = \nabla F(\theta_2)$

(FY initially called Legendre-Fenchel divergences…)

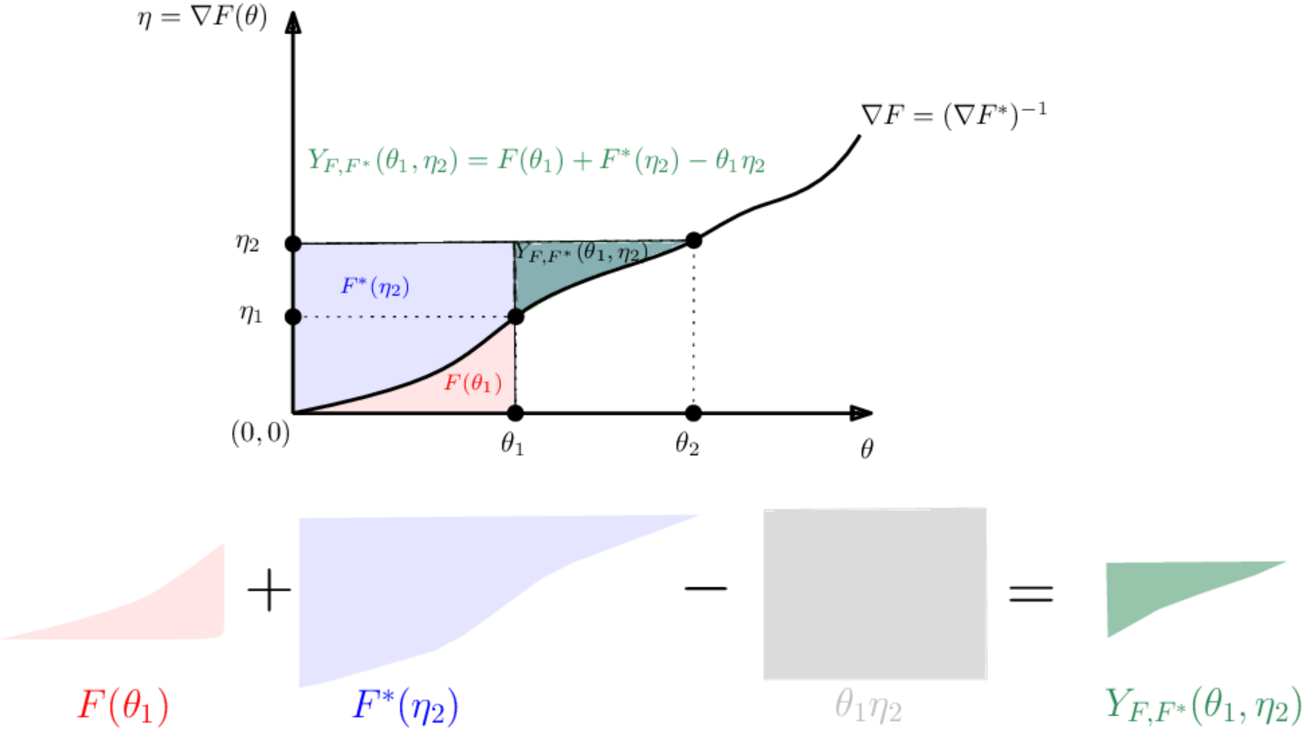
Bregman divergence vs Fenchel-Young divergence

Same parameterization $B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1, \eta_2)$ **mixed parameterization**

F strictly convex and differentiable



F' ↗ strictly increasing



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

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

Kullback-Leibler divergence between non-normalized exponential family densities

- Kullback-Leibler divergence between two **positive measures**:

$$D_{KL}^{+}[p_1(x):p_2(x)] = \int \{ p_1(x) \log (p_1(x)/p_2(x)) + p_2(x) - p_1(x) \} \, d\mu(x)$$

- Exponential family density:
  - Normalized: $p(x|\theta) = \exp(\langle x, \theta\rangle - F(\theta)) \, d\mu(x)$
  - Non-normalized: $q(x|\theta) = \exp(\langle x, \theta\rangle) \, d\mu(x)$
- Hence, $p(x|\theta) = q(x|\theta)/Z(\theta)$ with **partition function** $Z(\theta)=\exp(F(\theta))$ and **cumulant function** $F(\theta)=\log Z(\theta)$
- When F is convex, Z=exp(F) is log-convex
- log-convex functions are convex functions: So **both F and Z are convex functions**

- KLD between normalized densities = reverse Bregman wrt F:

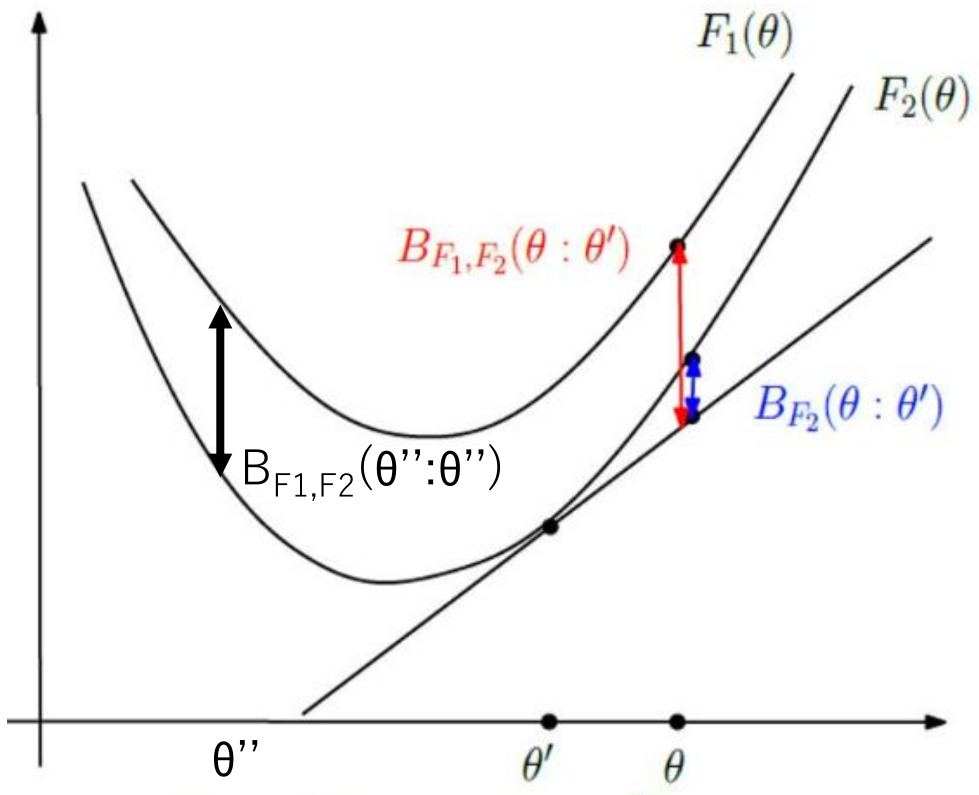
$$D_{KL}[p_{\theta_1}(x):p_{\theta_2}(x)] = B_F^*[\theta_1:\theta_2] = B_F[\theta_2:\theta_1]$$

- KLD between non-normalized densities = reverse Bregman wrt Z:

$$D_{KL}^{+}[q_{\theta_1}(x):q_{\theta_2}(x)] = B_Z^*[\theta_1:\theta_2] = B_Z[\theta_2:\theta_1]$$

2312.12849

Duo Bregman divergences: Generalize BDs with a pair of generators



One generator **majorizes** the other one:

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

Then

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

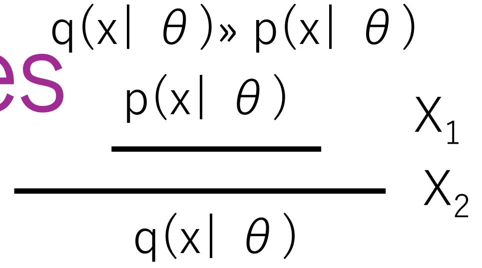
$$\geq B_{F2}(\theta:\theta')$$

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

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

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

KLD between nested exponential families amount to duo Bregman pseudo-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:

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

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**:

$$D_{KL}[p(x|\theta_1) : q(x|\theta_2)] = B_{F_2,F_1}^{rev}(\theta_1 : \theta_2) = B_{F_2,F_1}(\theta_2 : \theta_1)$$

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

Curved Bregman divergences

Consider a domain U which maps to a subset of Θ by $\theta = c(u)$ with $\dim(U) < \dim(\Theta)$:

$B_{F,u}(u_1 : u_2) := B_F(c(u_1) : c(u_2))$ is not Bregman when $\{c(u) \mid u \in U\}$ not convex

usually not a Bregman divergence unless c(.) is affine

Example: Symmetrized Bregman divergences (Jeffreys-Bregman div.) are curved Bregman divergences: $S_F(\theta_1, \theta_2) = \langle \theta_1 - \theta_2, \eta_1 - \eta_2 \rangle$

$$\begin{aligned} S_F(\theta_1 : \theta_2) &= B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1), \\ &= B_F(\theta_1 : \theta_2) + B_{F^*}(\nabla F(\theta_1) : \nabla F(\theta_2)) \\ &= \breve{B}_{F_\xi}(\xi(\theta_1) : \xi(\theta_2)), \end{aligned}$$

$$F^*(\eta) = \langle \theta, \eta \rangle - F(\theta) \qquad F_\xi(\theta, \eta) := F(\theta) + F^*(\eta) \qquad \xi(\theta) = (\theta, \nabla F(\theta))$$

$$\mathcal{U} = \{(\theta, \nabla F(\theta)) : \theta \in \Theta\}$$

m-dimensional submanifold in 2m-dimensional space

Curved Bregman centroid is the Bregman projection of the full Bregman centroid

Theorem:

$$\arg\min_{u\in\mathcal{U}} \sum_{i=1}^{n} w_i\, B_F(\theta_i : \theta(u)) = \arg\min_{u\in\mathcal{U}} B_F(\bar\theta : \theta(u))$$

[Bregman projection]

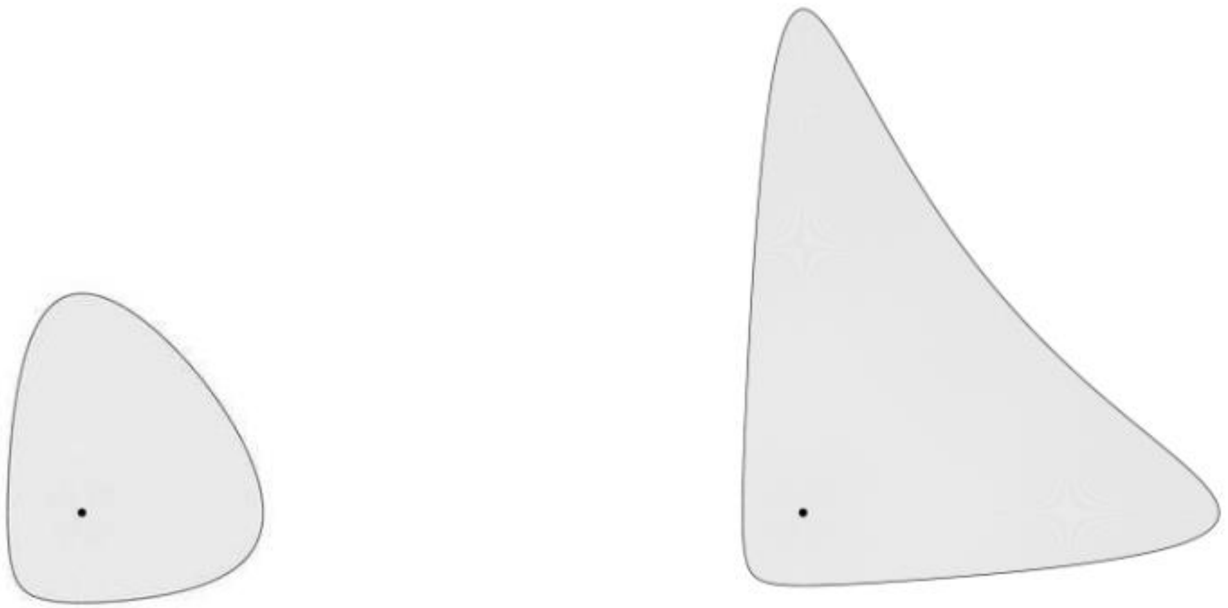
$$\theta_i = \theta(u_i) \qquad \bar\theta = \textstyle\sum_i w_i\theta_i$$

Proof.

$$\begin{aligned}
\min_{u\in\mathcal{U}} \sum_{i=1}^{n} w_i\, B_F(\theta_i : \theta(u)) &= \sum_{i=1}^{n} w_i(F(\theta_i) - F(\theta(u)) - \langle \theta_i - \theta(u), \nabla F(\theta(u))\rangle),\\
&\equiv -F(\theta(u)) - \langle \bar\theta - \theta(u), \nabla F(\theta(u))\rangle,\\
&\equiv F(\bar\theta) - F(\theta(u)) - \langle \bar\theta - \theta(u), \nabla F(\theta(u))\rangle\\
&= B_F(\bar\theta : \theta(u)).
\end{aligned}$$

"What is... an information projection?" Notices of the AMS 65.3 (2018): 321-324.

Space of Bregman balls



Example:
Itakura-Saito right and left spheres

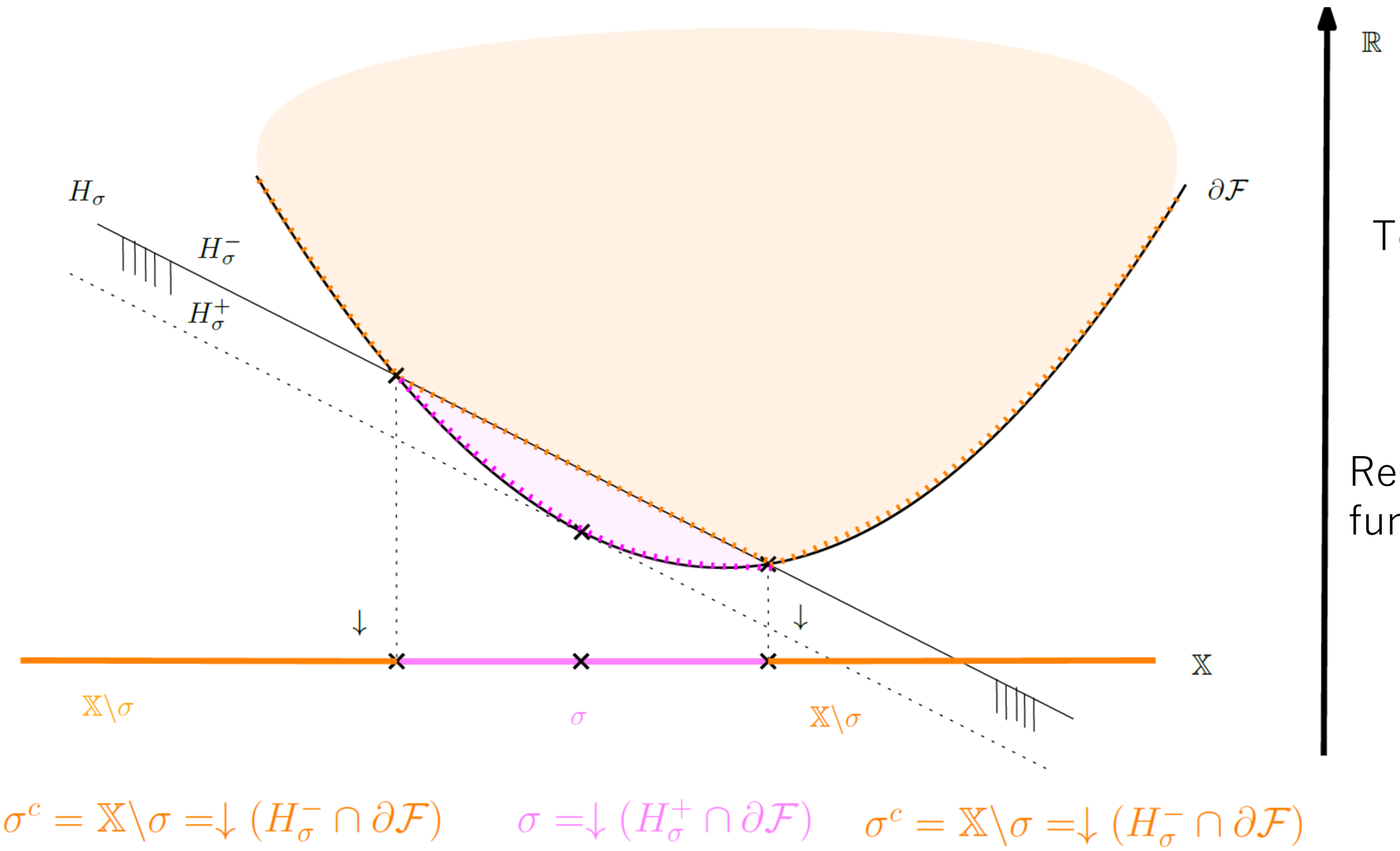
Right-sided Bregman ball: $$\sigma_F(\theta, r) = \{\theta' \in \Theta : B_F(\theta' : \theta) \leq r\}$$

Left-sided Bregman ball: $$\sigma_F^{\star}(\theta, r) = \{\theta' \in \Theta : B_F(\theta : \theta') \leq r\}$$

Application: Boolean algebra of unions & intersections of Bregman balls

Right Bregman ball and its complement

$$\mathcal{F} := \{(\theta, y \geq F(\theta)) \ : \ \theta \in \Theta \subset \mathbb{R}^m\} \subset \mathbb{R}^{m+1}$$



↓ means vertical projection

S^c: complement of set S

To any sphere, associate an hyperplane:

$$H_{\theta,r} : y = \langle \theta' - \theta, \nabla F(\theta) \rangle + F(\theta) + r$$

Reciprocally, to an hyperplane cutting the function graph, associate a sphere

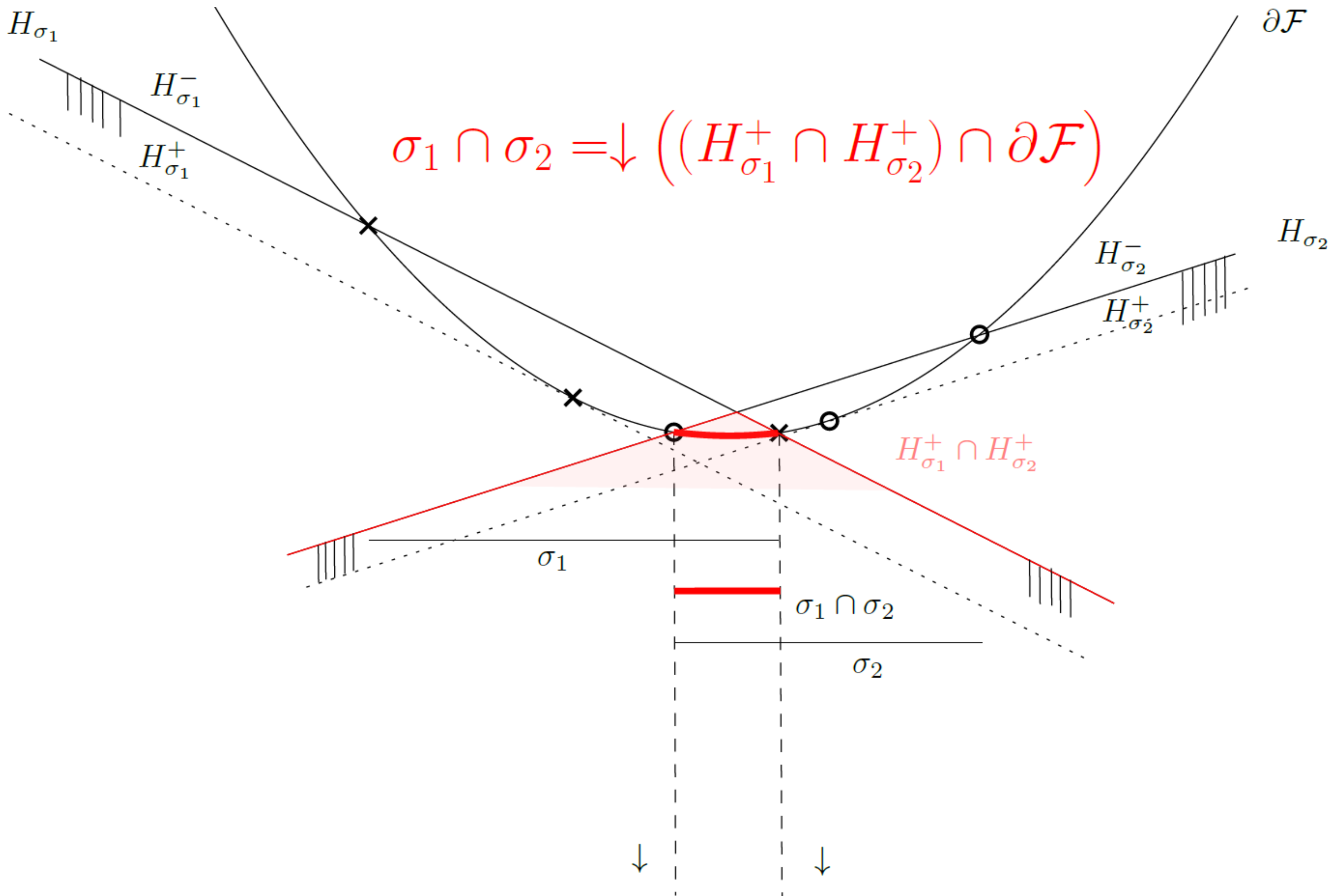
$$z = \langle \mathbf{x}, \mathbf{a} \rangle + b$$

Center: $\mathbf{c} = \nabla^{-1} F(\mathbf{a})$

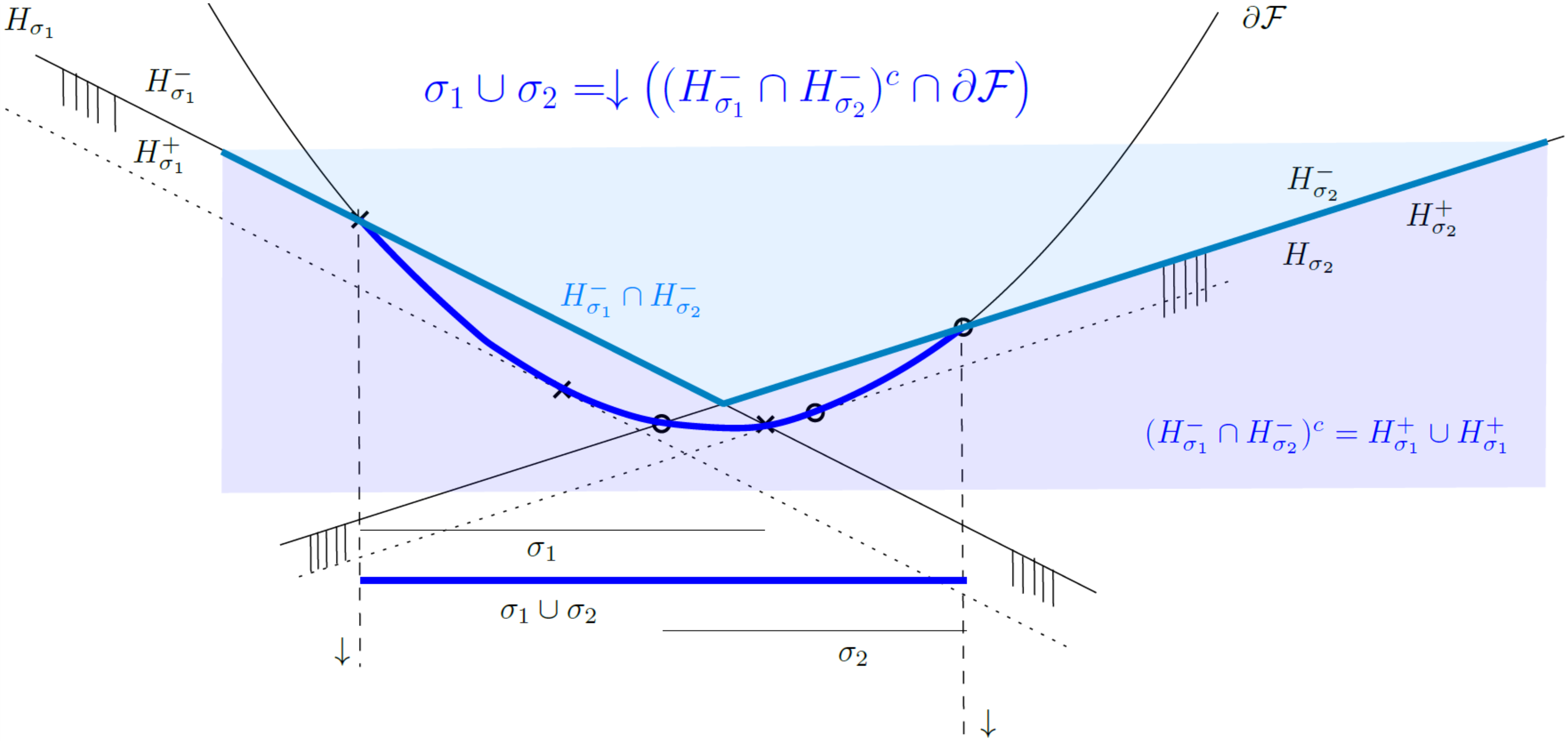
Radius: $\langle \mathbf{a}, \mathbf{c} \rangle - F(\mathbf{c}) + b$

Lifting to potential Bregman generator graph

Intersection of two right Bregman balls



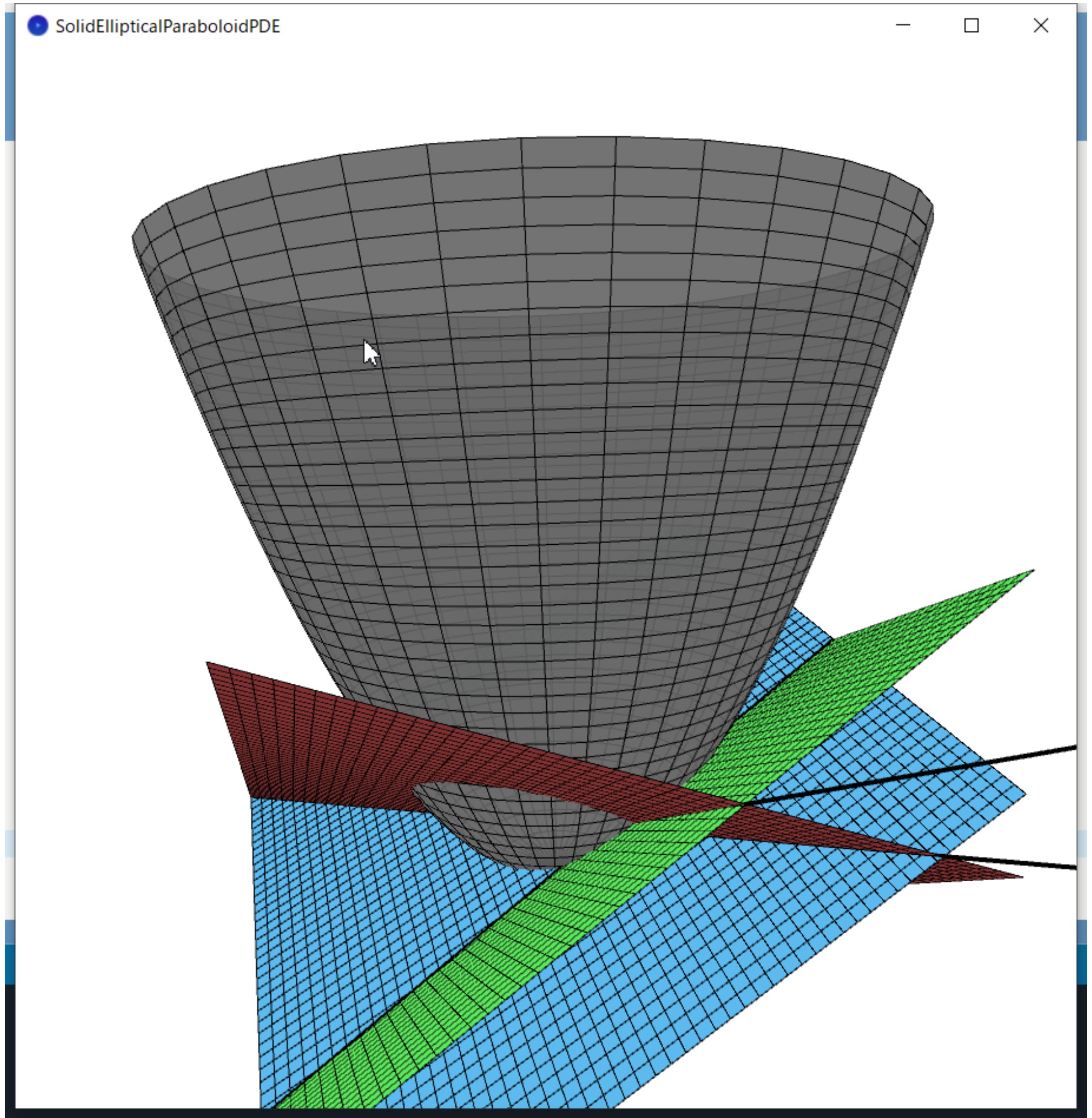
Union of two right Bregman balls



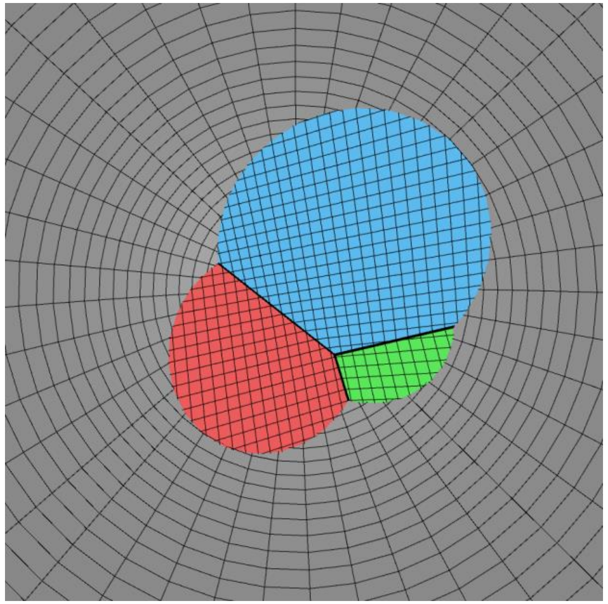
Set Morgan's law: $(A \cup B)^c = A^c \cap B^c$

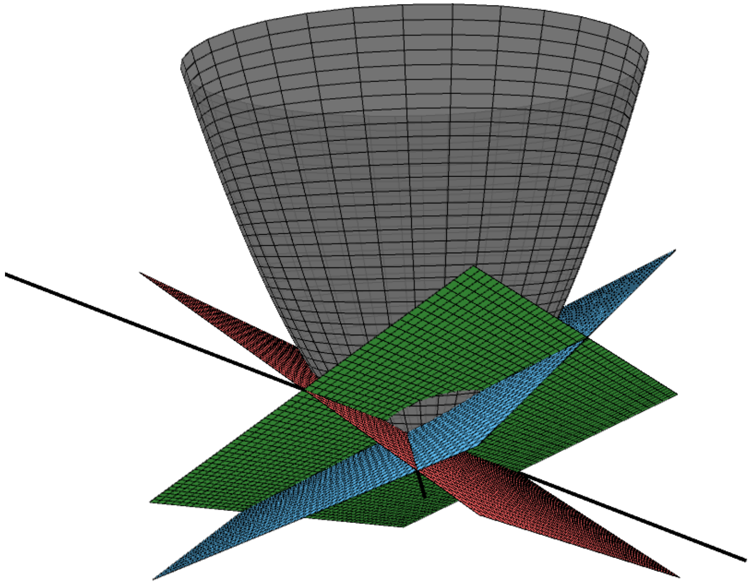
Complement of halfspace $(H^+)^c = H^-$

Example: Euclidean spheres potential function: Paraboloid, L22



Top view displays the union of disks





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

Bregman manifolds:
Geometry of convex conjugates

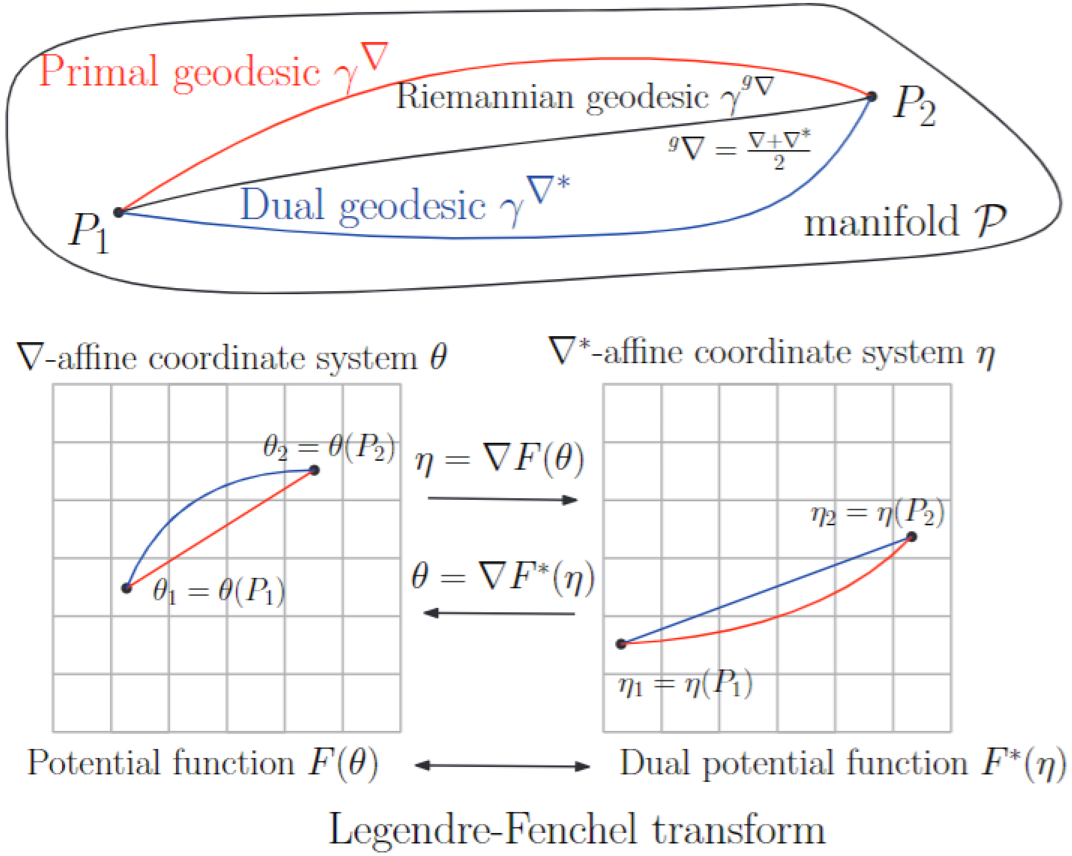
Dual Hessian geometry

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

On geodesic triangles with right angles in a dually flat space,
Progress in Information Geometry: Theory and Applications, Springer 2021

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

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



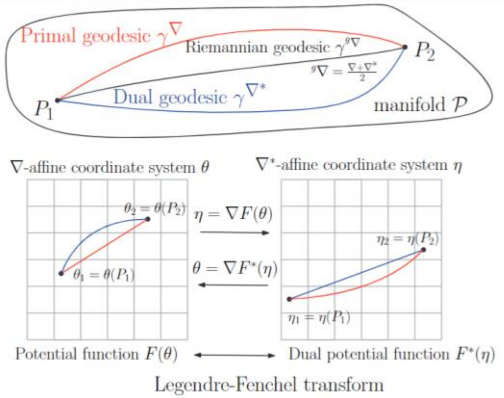
$D(P_1,P_2) = B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1, \eta_2) = Y_{F^*,F}(\eta_2, \theta_1) = B_{F^*}(\eta_2, \eta_1)$

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

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

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

Dual geometry of smooth Legendre-type functions



Legendre-Fenchel transform

$$(\Theta, F(\theta)) \longleftrightarrow (H, F^*(\eta))$$

$\eta = \nabla F(\theta), \eta = \nabla F^*(\theta)$

Riemannian Hessian metric g

$$g(\theta) = \nabla^2 F(\theta)$$

$$g(\eta) = \nabla^2 F^*(\eta)$$

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

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

flat ∇ torsion-free affine connection

$\Gamma_{ijk}(\theta) = 0, \Gamma^{ijk}(\eta) = \partial^i \partial^j \partial^k F^*(\eta)$

torsion-free affine connection ∇^* flat

$\Gamma^{*ijk}(\eta) = 0, \Gamma^{*ijk}(\theta) = \partial_i \partial_j \partial_k F(\theta)$

Levi-Civita connection

$$\bar{\nabla} = \frac{\nabla + \nabla^*}{2}$$

g-conjugate connections

$\bar{\Gamma}_{ijk}(\theta) = \frac{1}{2} \partial_i \partial_j \partial_k F(\theta)$ $\quad$ $\bar{\Gamma}^{ijk}(\eta) = \partial^i \partial^j \partial^k F^*(\eta)$

Example: Bregman manifold of multivariate Gaussians

(M, g, ∇, ∇^*)

Cumulant function is convex:

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

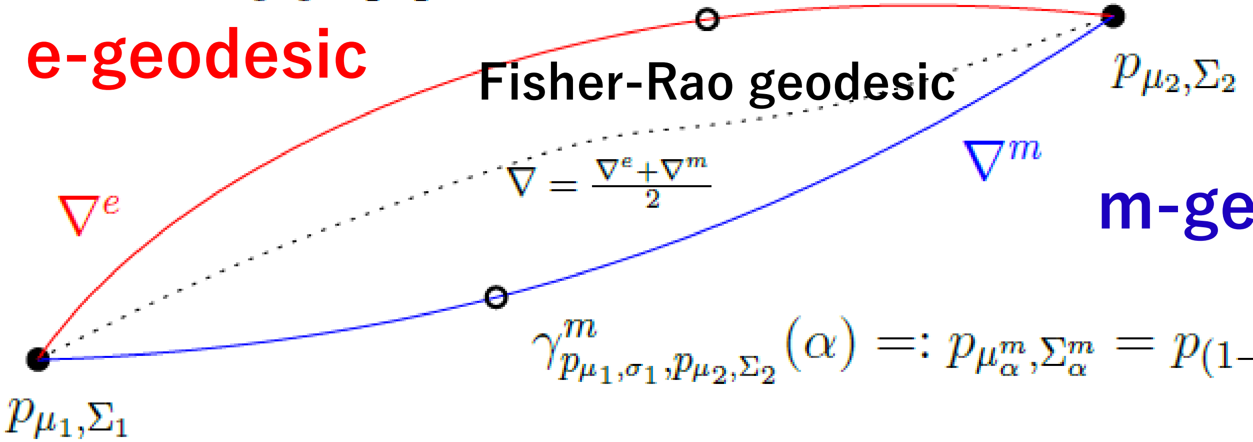
with respect to natural parameters:

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

…but not convex wrt（μ Σ）parameters

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

$$\gamma^e_{p_{\mu_1,\sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^e,\Sigma_\alpha^e} = p_{(1-\alpha)\theta_1+\alpha\theta_2} \qquad \theta = (\Sigma^{-1}\mu, \tfrac{1}{2}\Sigma^{-1})$$



m-geodesic beware not mixture of Gaussians!

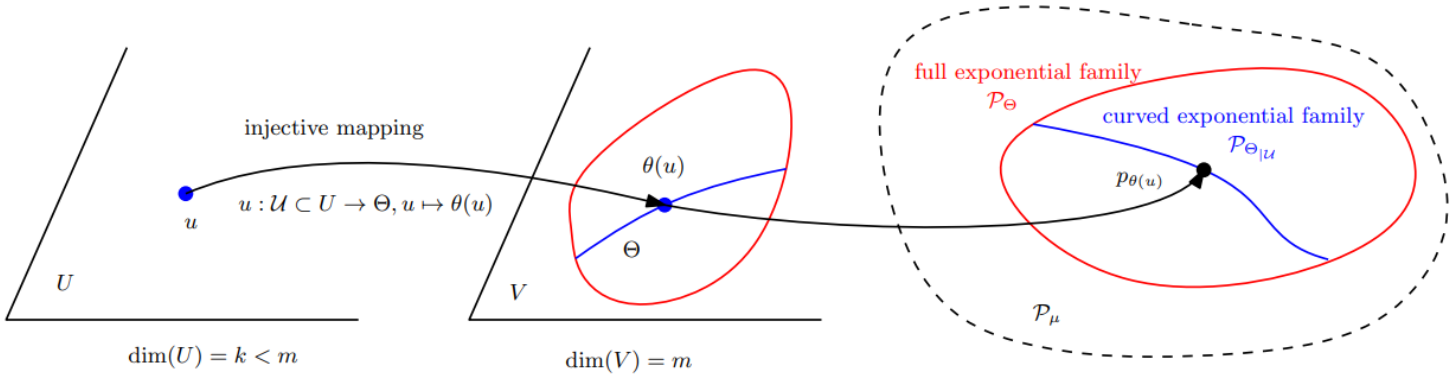
$$\gamma^m_{p_{\mu_1,\sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^m,\Sigma_\alpha^m} = p_{(1-\alpha)\eta_1+\alpha\eta_2} \qquad \eta = (\mu, -\Sigma - \mu\mu^\top)$$

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

Bregman divergence = reverse Kullback-Leibler divergence

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

Curved exponential families: Submanifolds



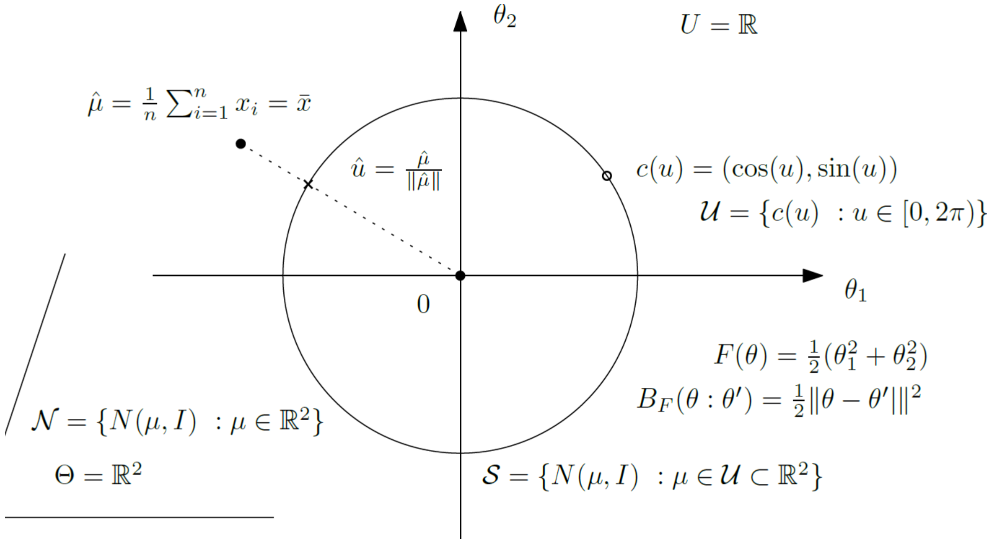
Constraint space

Parameter space

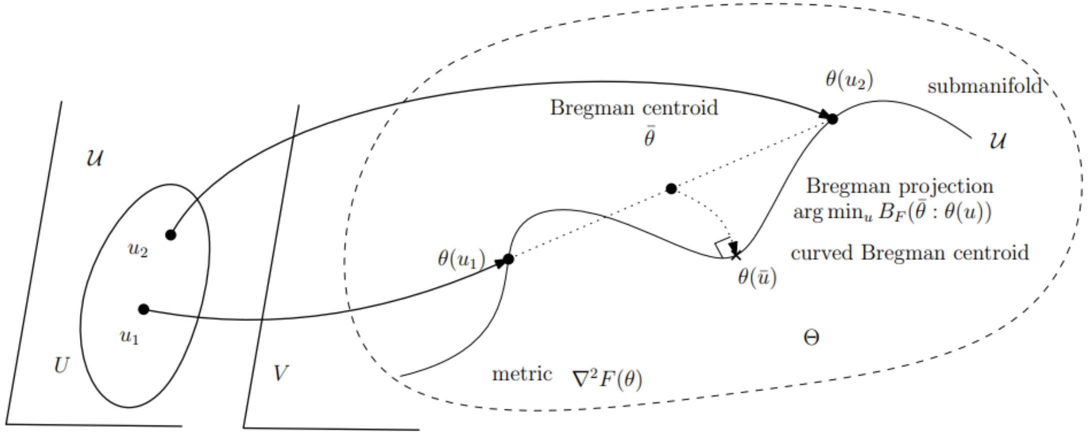
Theorem (Curved Bregman centroid/barycenter) *Let $\theta_i = \theta(u_i)$'s be n weighted parameters of $\mathcal{U}$ with weight vector $w \in \Delta_{n-1}$ (the $(n-1)$-dimensional standard simplex). Then the barycenter in $\mathcal{U}$ with respect to the curved Bregman divergence amounts to the Bregman projection of the center of mass $\bar{\theta} = \sum_i w_i \theta_i$ (right Bregman barycenter) onto $\mathcal{U}$:*

$$\arg\min_{u \in \mathcal{U}} \sum_{i=1}^{n} w_i B_F(\theta_i : \theta(u)) = \arg\min_{u \in \mathcal{U}} B_F(\bar{\theta} : \theta(u)).$$

Example: Fisher circle model



Note: submanifold topology can be non-trivial



Bijection between regular exponential families and regular Bregman divergences:

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

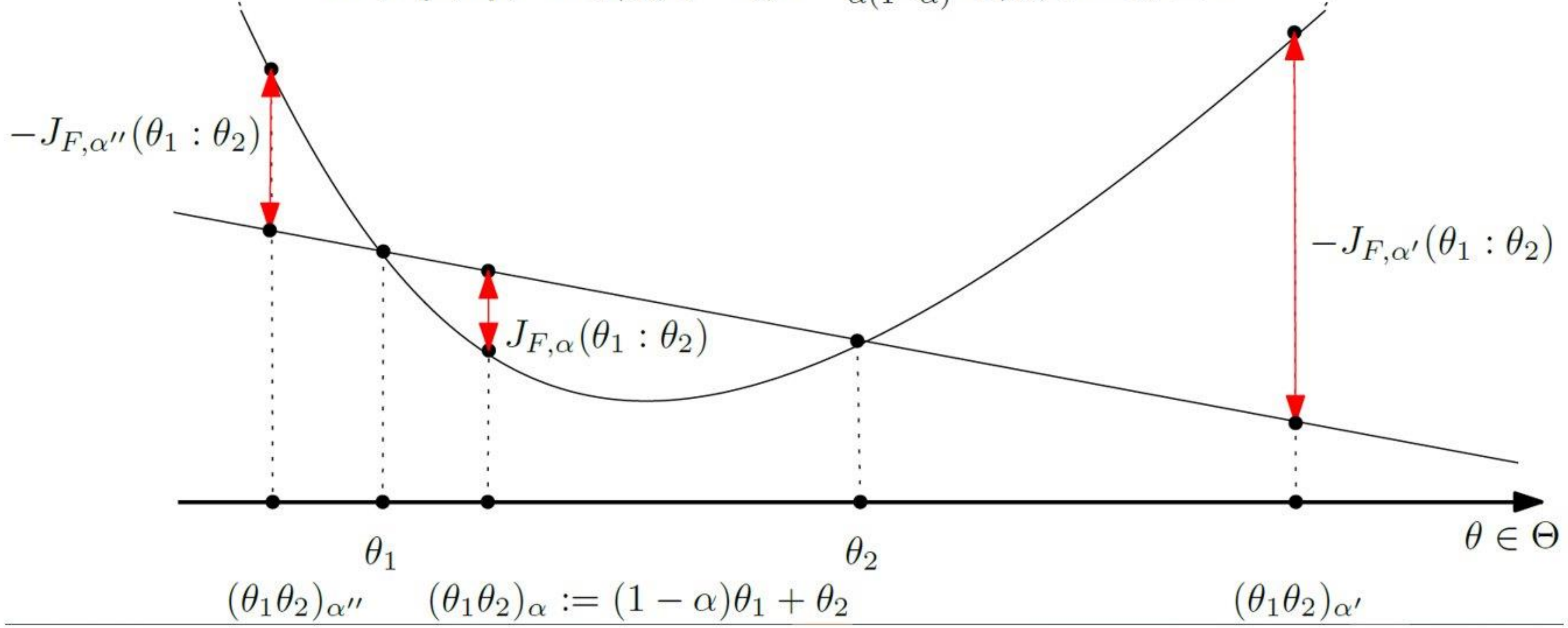
Curved BD centroid $\leftrightarrow$ MLE of curved exp. fam.

k-MLE: A fast algorithm for learning statistical mixture models, IEEE ICASSP 2012

Scaled skewed Jensen divergences & Bregman divergences

$$\forall \alpha \in (0,1), \quad J_{F,\alpha}(\theta_1 : \theta_2) := (1-\alpha)F(\theta_1) + \alpha F(\theta_2) - F((1-\alpha)\theta_1 + \alpha\theta_2)$$

Rescale: $$\forall \alpha \notin \{0,1\}, \quad \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) := \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1 : \theta_2) \geq 0$$



Jensen divergences measures the vertical gap induced by a strictly convex function

$$\lim_{\alpha \to 0} \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_2)$$ (Bregman divergence) $$\lim_{\alpha \to 1} \mathrm{sJ}_{F,\alpha}(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$$ (reverse BD)

The Burbea-Rao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57.8 (2011)

Example 1 of Bregman manifolds:

Mixture family manifolds
(F=-S is Shannon negentropy)

Jensen-Shannon centroid for mixture families

Jensen-Shannon divergence
Bounded symmetrization of KLD

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

- **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

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

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

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

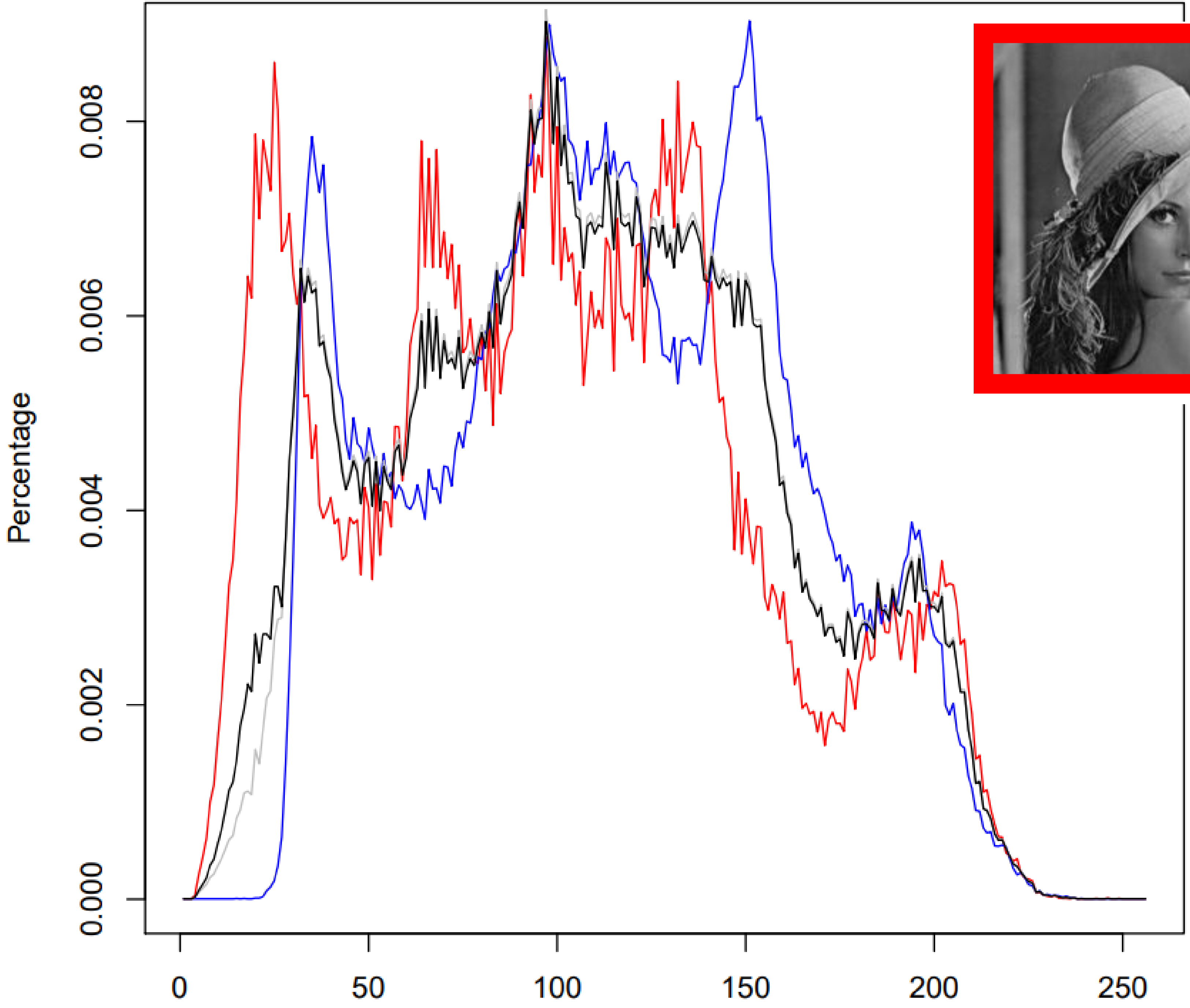
$$\min_\theta \sum_i J_F(\theta_i,\theta),$$

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

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

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

F is Shannon negentropy (convex)



Jensen–Shannon centroid

Jeffreys/SKL centroid

Jensen–Shannon centroid do not require same support

Example 2 of Bregman manifolds:

Exponential family manifolds

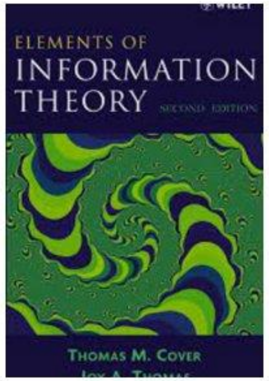
(F is cumulant function aka log-partition function)

Chernoff information: A geometric characterization

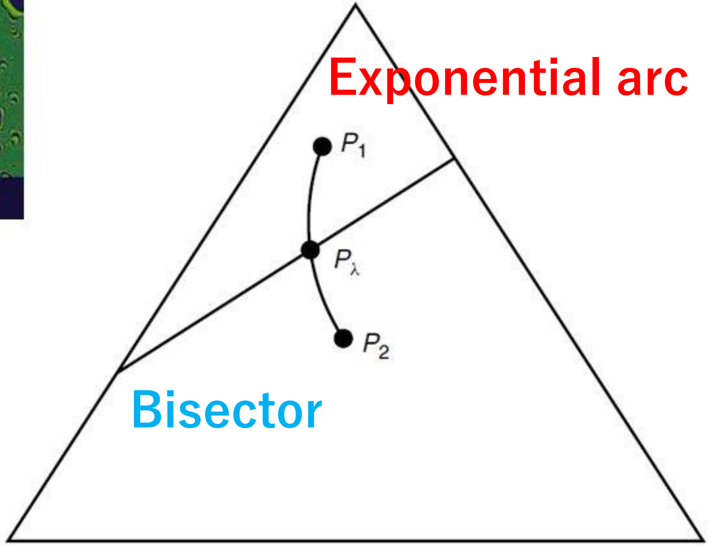
Exponential arc

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

Bisector



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



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

Probability simplex

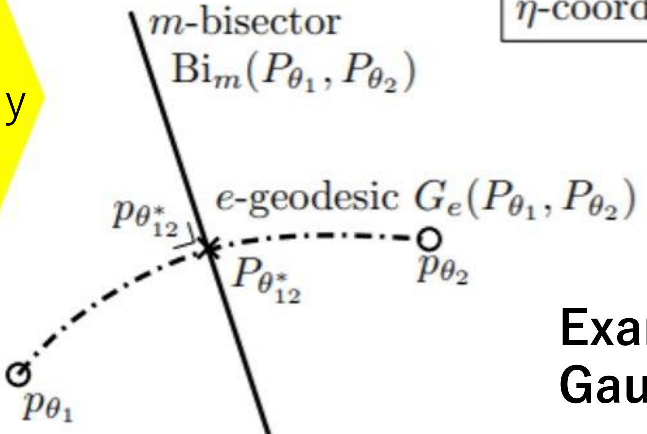
Generalized to exponential family manifold

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

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

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

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



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

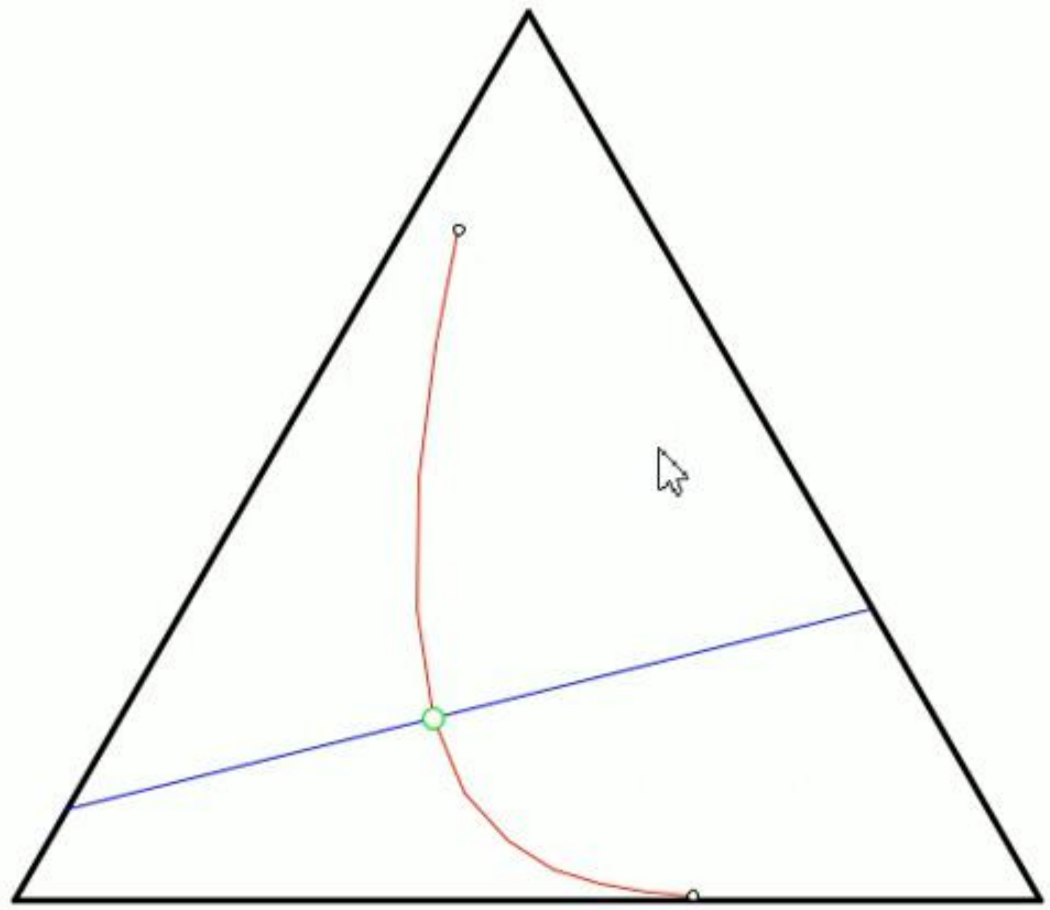
Example:
Gaussian manifold

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

Exponential family manifold

Chernoff point & information-geometry

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



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

In the beginning of IG···

[Hotelling 1930, Rao 1945]

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

Fisher-Rao manifolds

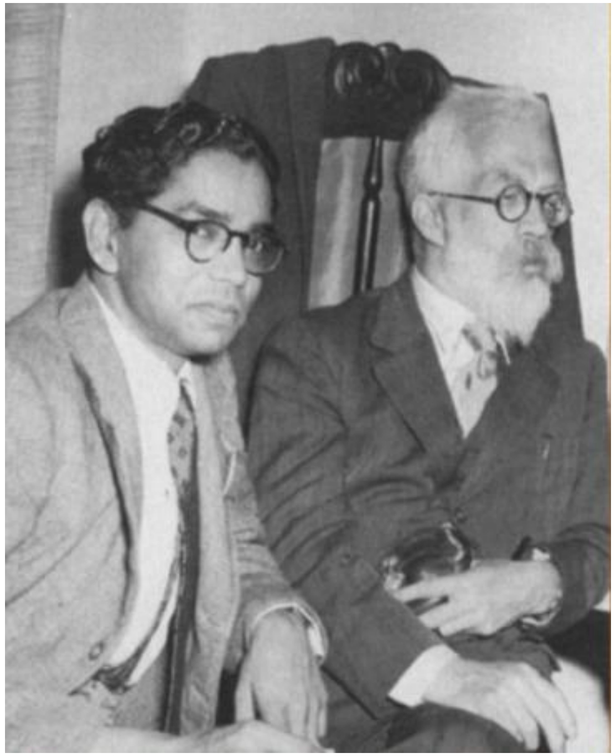
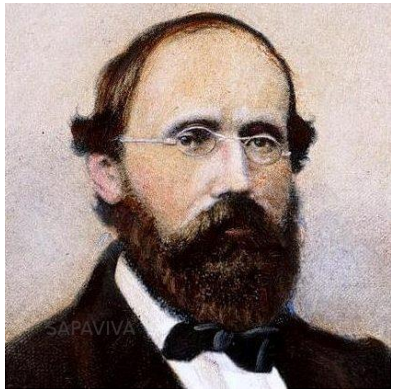


Photo 1956

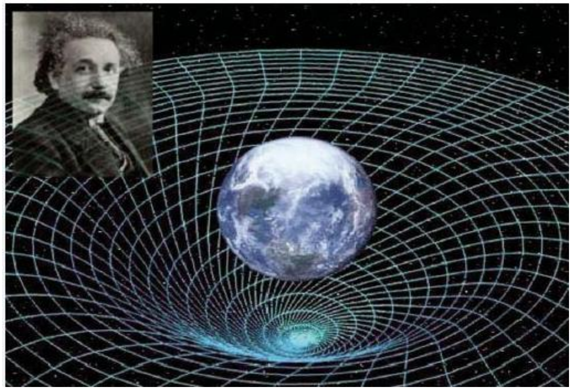
Riemannian geometry

Length element
ds

1854



``Killer''
application



1915, GR

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

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

Fisher length:

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

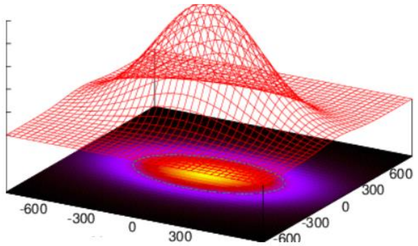
Geodesic ODE:

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

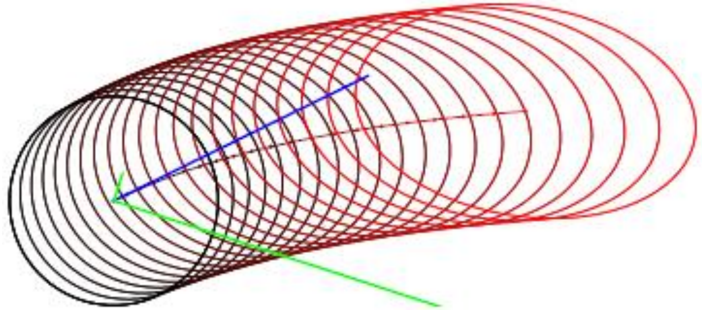
Solve ODE with initial values (IV) or boundary values (BV)

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

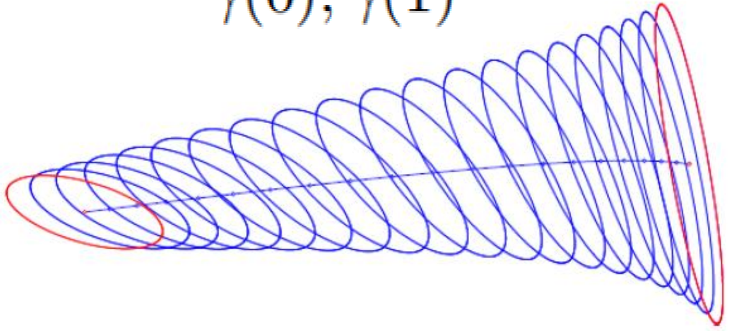
Bivariate normal (represented by ellipsoids)



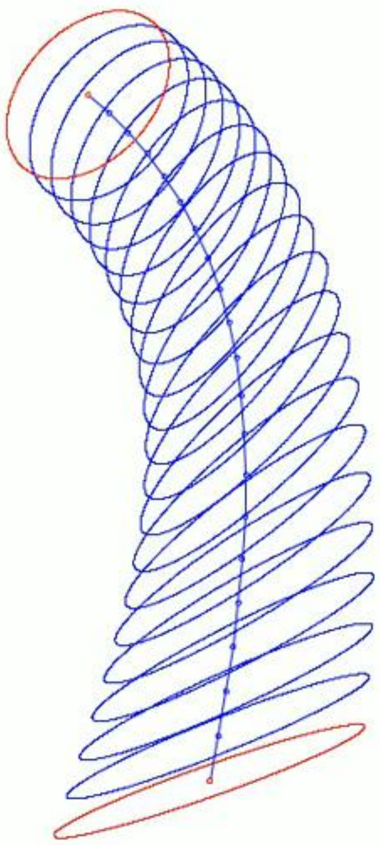
[IV: Eriksen 1987]
$\gamma(0), \dot{\gamma}(0) \in T_{\gamma(0)}$



[BV: Kobayashi 2023]
$\gamma(0), \gamma(1)$



Fisher-Rao geodesics with boundary



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

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

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

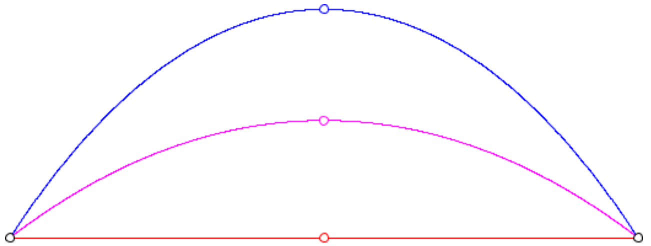
[BV: Kobayashi 2023]

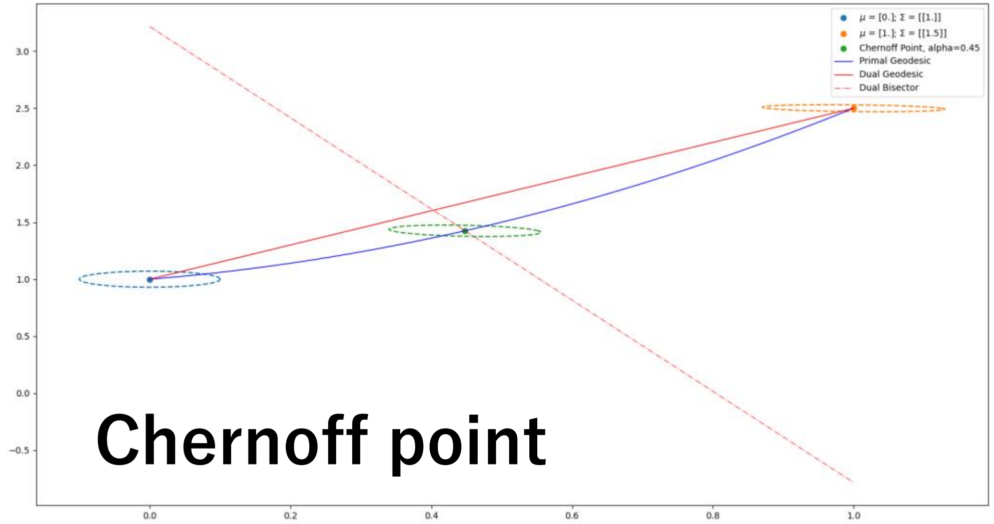
Technically, MVN Fisher-Rao geodesic:
Riemannian submersion of a horizontal geodesic
of a Riemannian symmetric space in 2d+1 dimension

A Python library for geometric computing on Bregman Manifolds

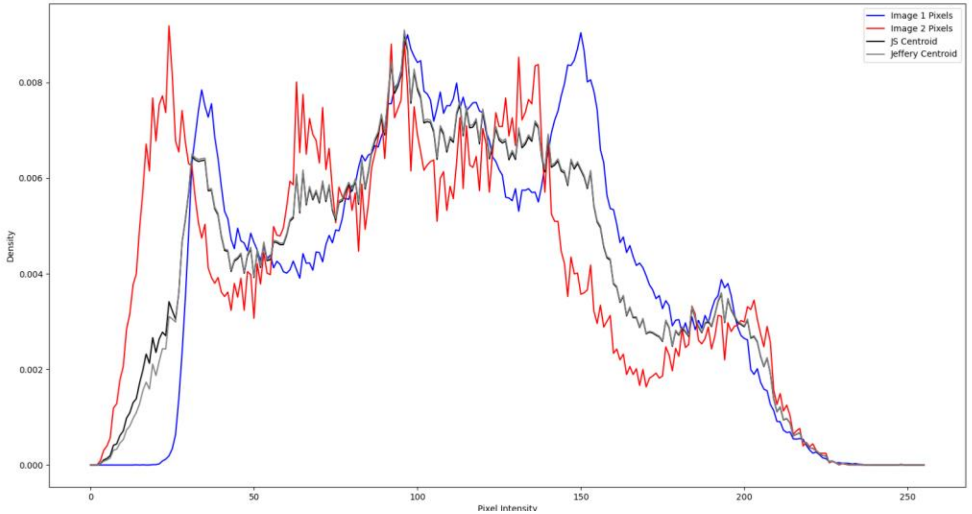
pyBregMan

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

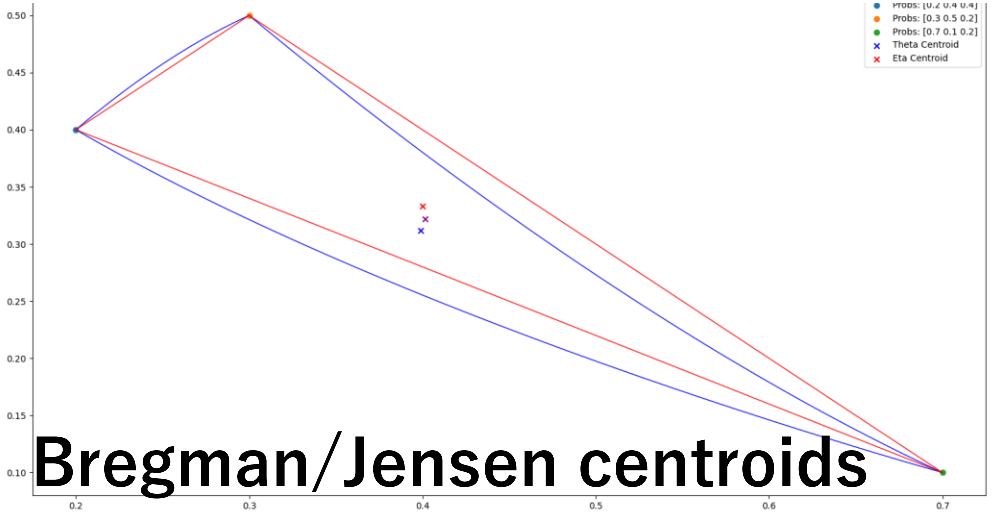




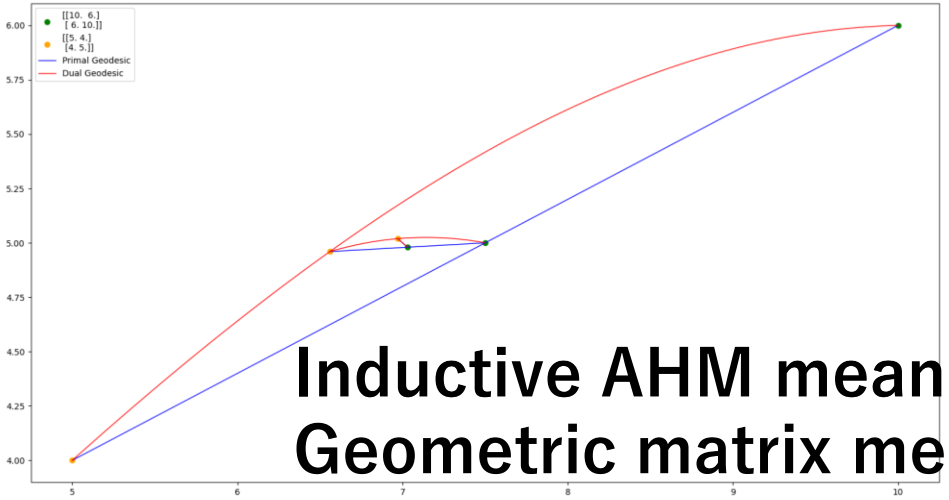
Chernoff point



Jensen-Shannon centroid



Bregman/Jensen centroids



Inductive AHM mean
Geometric matrix mean

Joint work of Frank Nielsen and Alexander Soen

Thank you! Some references

- NF and Richard Nock. "**The dual Voronoi diagrams with respect to representational Bregman divergences**." *Sixth International Symposium on Voronoi Diagrams*. IEEE, 2009.
- Boissonnat, Jean-Daniel, FN, and Richard Nock. "**Bregman Voronoi diagrams**." *Discrete & Computational Geometry* 44 (2010): 281-307.
- NF. "**Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences**." *Entropy* 24.3 (2022
- NF and Richard Nock. "**Generalizing skew Jensen divergences and Bregman divergences with comparative convexity**." *IEEE Signal Processing Letters* 24.8 (2017)
- NF. "**Curved representational Bregman divergences and their applications**." *arXiv preprint arXiv:2504.05654*