

# Information geometry:

A short introduction  
with some recent advances

Frank Nielsen

Sony Computer Science Laboratories Inc

Tokyo, Japan



Sony CSL

# Talk outline

- **Information geometry from the pure viewpoint of geometry:**
  - Geometry of dual structures
- **Dual multivariate quasi-arithmetic averages:**
  - Information geometry yielding a generalization of quasi-arithmetic means
- **Chernoff information and its purely geometric counterpart:**
  - Geometry likelihood ratio exponential families
- **Duo Bregman pseudo-divergences:**
  - Application to KLD between truncated densities of an exponential family

Information geometry:

A short introduction to the geometry of dual structures

# Information geometry (IG): Rationale and scope

- IG field originally born by investigating **geometric structures** of statistical/probability models (e.g, space of Gaussians, space of multinomials)
- **Statistical models**: parametric vs nonparametric models, regular vs singular (ML) models, hierarchical (ML) or simple models, ...
- Define **statistical invariance**, use **language of geometry** (e.g., ball, projection, bisector) to design algorithms in statistics, information theory, statistical machine learning, etc.
- IG study **interplays** of **statistical/parameter divergences** with geometric structures
- Relationships between **many types of dualities** in IG: dual connections, reference duality (dual f-divergences), Legendre duality, duality of representations/monotone embeddings, etc

# Geometric science of information (GSI)

Further extend broadly the original scope of information geometry by unravelling **connections** of information geometry (IG) with **other domains of geometry** like:

- geometry of domains and cones (e.g., Siegel/Vinberg/Koszul)
- geometric mechanics for dynamic models (symplectic/contact geometry)
- thermodynamics/thermostatistics and deformed statistical models
- geometric statistics (eg, computational anatomy/medical imaging)
- shape space analysis and deformation (computer vision)
- algebraic statistics (manifolds versus algebraic surfaces/varieties)
- dynamics of learning (singularity, plateau)
- neurogeometry (neuroscience)
- etc.

[franknielsen.github.io/GSI/](http://franknielsen.github.io/GSI/)



# GSI: Biannual conference since 2013



**GSI'23 Conference**  
**FROM CLASSICAL TO QUANTUM INFORMATION GEOMETRY**  
6th Conference Edition  
Palais du Grand Large, Saint-Malo  
August 30th - September 1st, 2023



<https://gsi2023.org>

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



**Eva Miranda**

Polytechnic University of Catalonia, Spain  
From Alan Turing to Contact geometry:  
towards a "Fluid computer"



**Francis BACH**

Inria, Ecole Normale Supérieure, France  
Information Theory with Kernel Methods



**Bernd STURMFELS**

MPI-MiS Leipzig Germany  
Algebraic Statistics and Gibbs Manifolds



**Diarra FALL**

Institut Denis Poisson, Université d'Orléans & Université de Tours, France  
Statistics Methods for Medical Image Processing and Reconstruction



**Hervé SABOURIN**

Poitiers University, France  
Transverse Poisson Structures to adjoint orbits in a complex semi-simple Lie algebra



**Juan-Pablo ORTEGA**

Nanyang Technological University, Singapore  
Learning of Dynamic Processes

*Random ordering of keynote speakers*

<https://conference-gsi.org/>

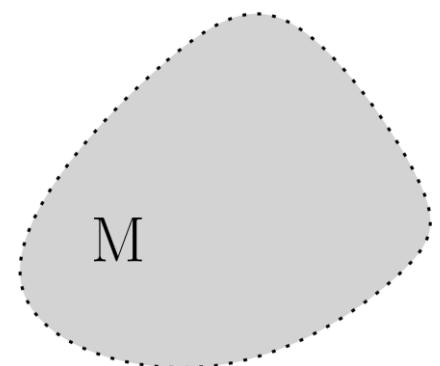
Include 500+ GSI video talks: [franknielsen.github.io/GSI/](https://franknielsen.github.io/GSI/)

# Information geometry: Geometry of dual structures

# Build your own information geometry in three steps

Choose

① manifold  $M$

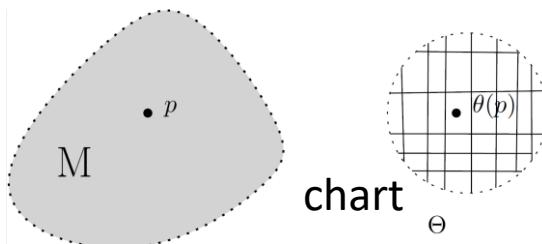


Examples:

Gaussians

SPD cone

Probability simplex

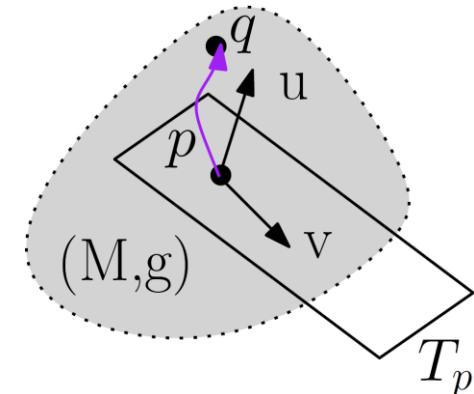


Concepts:

local coordinates

locally Euclidean

② metric tensor  $g$



Examples:

Fisher information metric  
metric  $g^D$  from divergence  
trace metric

Concepts:

vector length

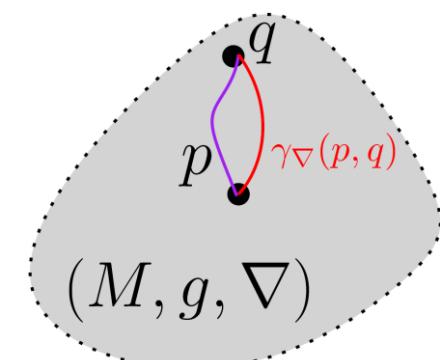
vector orthogonality

Riemannian geodesic

Riemannian distance

Levi-Civita connection  $\nabla^g$

③ affine connection  $\nabla$



Examples:

exponential connection

mixture connection

metric connection  $\nabla^g$

divergence connection  $\nabla^D$

$\alpha$ -connection

Concepts:

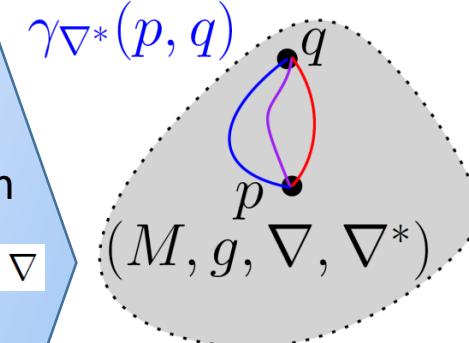
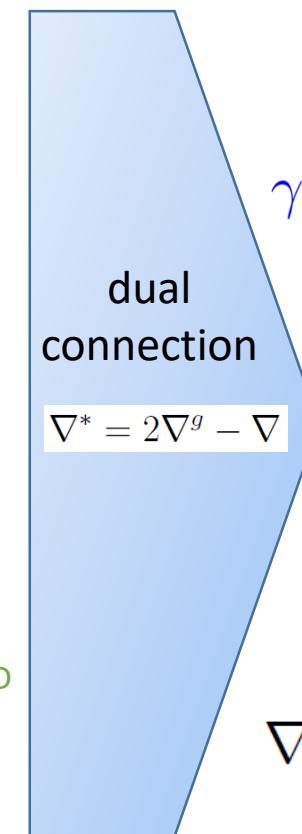
covariant derivative  $\nabla$

$\nabla$ -geodesic

$\nabla$ -parallel transport

curvature

Get dual IG manifold  $(M, g, \nabla, \nabla^*)$



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

Concepts:

dual connections coupled to metric  $g$

dual parallel transport preserve metric  $g$

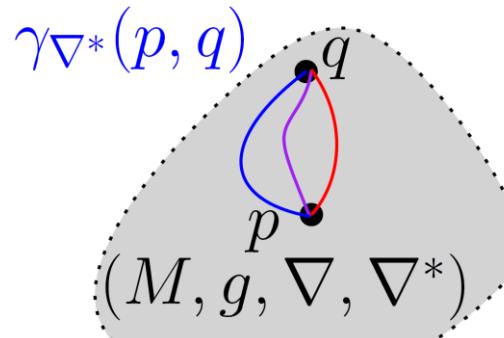
# From dual information geometry to $\pm\alpha$ -geometry, $\alpha \in \mathbb{R}$

Choose

- ① manifold  $M$
- ② metric tensor  $g$
- ③ affine connection  $\nabla$   
by defining Christoffel symbols  $\Gamma_{ijk}^\nabla$

Get dual IG manifold

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



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

$$T_{ijk} = \Gamma_{ijk}^* - \Gamma_{ijk}$$

$$T_{ijk} = \nabla_i g_{jk}$$

- ④ choose  $\alpha$

Examples:

Amari-Chentsov cubic tensor

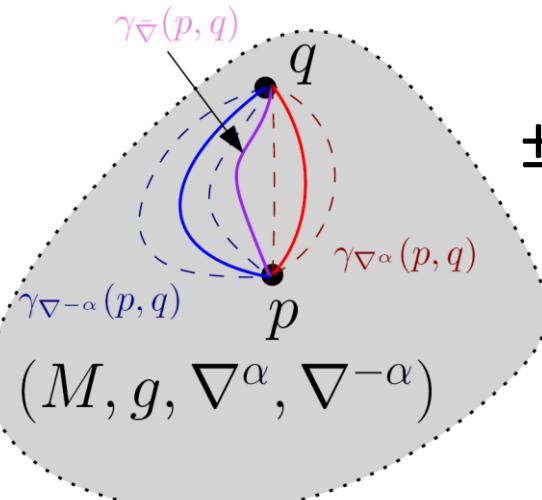
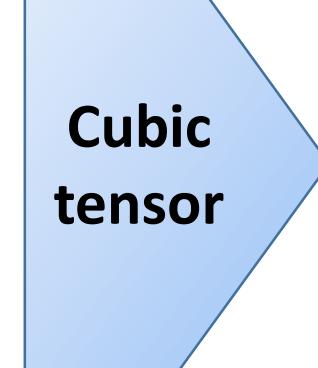
Cubic tensor from divergence

$$T_{ijk}(\theta) = E[\partial_i l \partial_j l \partial_k l]$$

$$T_{ijk}(\theta) = \partial_i \partial_j \partial_k F(\theta)$$

Get a family of dual connections/IG

$(M, g, \nabla^\alpha, \nabla^{-\alpha})$



$$\nabla^\alpha = \bar{\Gamma}_{ijk} - \frac{\alpha}{2} T_{ijk}$$

$$\nabla^{-\alpha} = \bar{\Gamma}_{ijk} + \frac{\alpha}{2} T_{ijk}$$

$\pm\alpha$ -geometry

$(M, g, \nabla^\alpha, \nabla^{-\alpha})$

0-geometry

= Riemannian geometry  
with geodesic distance

# Information geometry from statistical models: $(M, g^F, \nabla^{-\alpha}, \nabla^\alpha)$

- Consider a parametric **statistical/probability model**:  $\mathcal{P} := \{p_\theta(x)\}_{\theta \in \Theta}$
  - Define metric tensor  $g$  from **Fisher information** = **Fisher metric**  $g^F$

$$\mathcal{P}I(\theta) := E_\theta [\partial_i l \partial_j l]_{ij} \succeq 0 \quad \partial_i l := \frac{\partial}{\partial \theta_i} l(\theta; x) \quad l(\theta; x) := \log L(\theta; x) = \log p_\theta(x).$$

**covariance of the score**  $s_\theta = \nabla_\theta l = (\partial_i l)_i$       **log-likelihood**

- Model is **regular** if partial derivatives of  $I_\theta(x)$  smooth and Fisher metric is well-defined and positive-definite

- **Amari-Chentsov cubic tensor:**  $C_{ijk} := E_\theta [\partial_i l \partial_j l \partial_k l] \rightarrow \{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\}_{\alpha \in \mathbb{R}}$

•  **$\alpha$ -connections**  $\nabla^\alpha = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m$   $\alpha=1$  **exponential connection**

$$\begin{aligned} {}_P\Gamma^\alpha{}_{ij,k}(\theta) &:= E_\theta [\partial_i \partial_j l \partial_k l] + \frac{1-\alpha}{2} C_{ijk}(\theta), \\ &= E_\theta \left[ \left( \partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right] \end{aligned}$$


$$\begin{aligned} {}_P^e \nabla &:= E_\theta [(\partial_i \partial_j l)(\partial_k l)], \\ {}_P^m \nabla &:= E_\theta [(\partial_i \partial_j l + \partial_i l \partial_j l)(\partial_k l)] \end{aligned}$$

**mixture connection**

- Fisher-Rao geometry when  $\alpha=0$ , get geodesic distance called **Rao distance**

$$D_\rho(p, q) := \int_0^1 \|\gamma'(t)\|_{\gamma(t)} dt = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt$$

[Hotelling 1930] [Rao 1945] [Amari Nagaoka 1982]

# Rao distance on the Fisher-Rao manifold

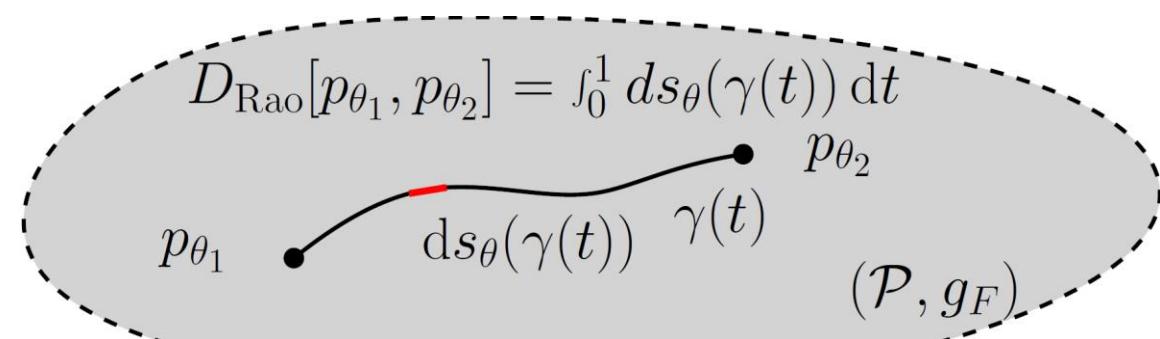
$$\begin{aligned} D_{\text{Rao}}[p_{\theta_1}, p_{\theta_2}] &= \rho_g(\theta_1, \theta_2) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt, \gamma(0) = \theta_1, \gamma(1) = \theta_2 \\ &= \int_0^1 ds_{\theta}(\gamma(t)) dt \end{aligned}$$

Here,  $\gamma$  is the Riemannian geodesic  
(or add a minimizer on all paths  $\gamma$ )

**Length element**

$$\dot{\theta}_k(t) = \frac{d}{dt}\theta_k(t)$$

$$ds_{\theta}^2(t) = \sum_{i=1}^D \sum_{j=1}^D g_{ij}(\theta) \dot{\theta}_i(t) \dot{\theta}_j(t)$$



In practice:

- Need to calculate geodesics which are curves locally minimizing the length linking two endpoints (equivalently minimize the energy of squared length elements)
- Finding Fisher-Rao geodesics is a non-trivial task.
- **Good news 2023:** closed-form geodesics with boundary conditions for **MultiVariate Normals**

# Information geometry from divergences: $(M, g^D, \nabla^D, \nabla^{D^*})$

- A **statistical divergence** like the Kullback-Leibler divergence is a smooth non-metric distance between probability measures

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

- A statistical divergence between two densities of a statistical model is a **parametric divergence** (e.g., KLD between two normal distributions)

$$D_{\text{KL}}^P(\theta_1 : \theta_2) := D_{\text{KL}}[p_{\theta_1} : p_{\theta_2}]$$

- Construction of *dual geometry from asymmetric parametric divergence*  $D(\theta_1 : \theta_2)$
- **Dual divergence** is  $D^*(\theta_1 : \theta_2) = D(\theta_2 : \theta_1)$ , *reverse divergence* [Eguchi 1983]

Dual structure:

$$\begin{aligned} {}^D g &:= -\partial_{i,j} D(\theta : \theta')|_{\theta=\theta'} = {}^{D^*} g, \\ {}^D \Gamma_{ijk} &:= -\partial_{ij,k} D(\theta : \theta')|_{\theta=\theta'}, \\ {}^{D^*} \Gamma_{ijk} &:= -\partial_{k,ij} D(\theta : \theta')|_{\theta=\theta'}. \end{aligned}$$

Cubic tensor:

$$\begin{aligned} {}^D C_{ijk} &= {}^{D^*} \Gamma_{ijk} - {}^D \Gamma_{ijk}, \\ \partial_{i,jk} f(x, y) &= \frac{\partial}{\partial x^i} \frac{\partial^2}{\partial y^j \partial y^k} f(x, y), \\ \partial_{i,.} f(x, y) &= \frac{\partial}{\partial x^i} f(x, y), \quad \partial_{.,j} f(x, y) = \frac{\partial}{\partial y^j} f(x, y), \quad \partial_{ij,k} f(x, y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial}{\partial y^k} f(x, y) \end{aligned}$$

# Realizations of dual information geometry (stat mfd)

- Realize  $(M, g, \nabla, \nabla)$  as a divergence information geometry  $(M, g^D, \nabla^D, \nabla^{D*})$ :  
always exists a divergence  $D$  such that  $(M, g, \nabla, \nabla) = (M, g^D, \nabla^D, \nabla^{D*})$

Matumoto, "Any statistical manifold has a contrast function—On the C3-functions taking the minimum at the diagonal of the product manifold." *Hiroshima Math. J* 23.2 (1993)

- Realize  $(M, g, \nabla, \nabla)$  as a model information geometry  $(M, g^F, \nabla^{-\alpha}, \nabla^\alpha)$   
always exists a statistical model  $M$  such that  $(M, g, \nabla, \nabla) = (M, {}_P g^F, {}_P \nabla^{-\alpha}, {}_P \nabla^\alpha)$

Lê, Hồng Vân. "Statistical manifolds are statistical models." *Journal of Geometry* 84 (2006): 83-93.

# Equivalence: model $\alpha$ -IG $\leftrightarrow$ divergence IG for f-divergences

- Let  $P=\{p_\theta\}$  be a statistical model of probability distributions dominated by  $\mu$
- Consider the **f-divergence** for a convex generator  $f(u)$  with  $f(1)=0$ ,  $f'(1)=1$ ,  $f''(1)=1 \leftarrow$  standard f-divergence (can always rescale  $g(u)=f(u)/f''(1)$ )

$$I_f[p(x; \theta) : p(x; \theta')] = \int_{\mathcal{X}} p(x; \theta) f\left(\frac{p(x; \theta')}{p(x; \theta)}\right) d\mu(x) \quad I_f^*[p(x; \theta) : p(x; \theta')] = I_f[p(x; \theta') : p(x; \theta)] = I_{f^\diamond}[p(x; \theta) : p(x; \theta')]$$

Dual reverse f-divergence is a f-divergence for  $f^\diamond(u) := u f\left(\frac{1}{u}\right)$

- The f-divergence between  $p_{\theta_1}$  and  $p_{\theta_2}$  is a parameter divergence  $D(\theta_1 : \theta_2)$

$$D_{\mathcal{P}}(\theta_1 : \theta_2) := I_f[p_{\theta_1} : p_{\theta_2}]$$

from which we can build the divergence information geometry  $(M, g^D, \nabla^D, \nabla^{D*})$

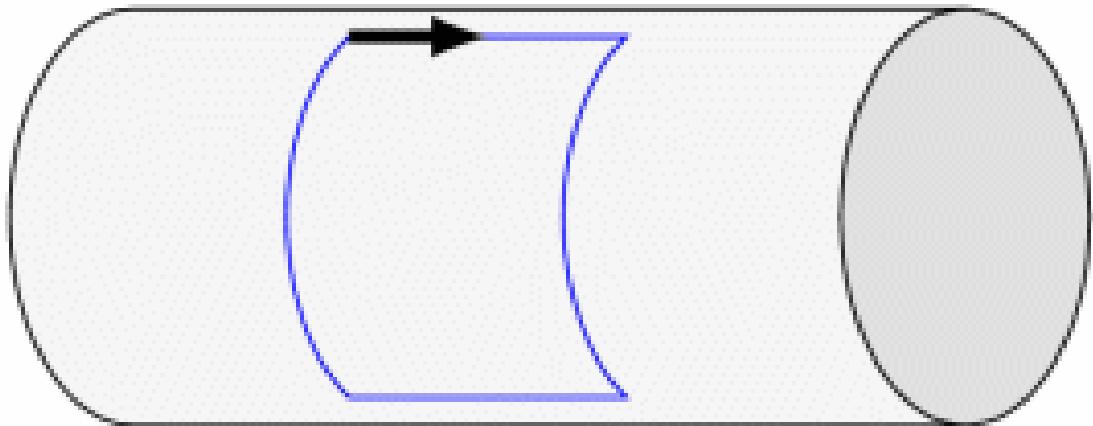
- Then **model  $\alpha$ -geometry** for  $\alpha=2$   $f'''(1)+3$  coincide with **divergence IG**:

$$(M, g^D, \nabla^D, \nabla^{D*}) = (M, g^F, \nabla^{-\alpha}, \nabla^\alpha) \text{ for } \alpha=2 \text{ } f'''(1)+3$$

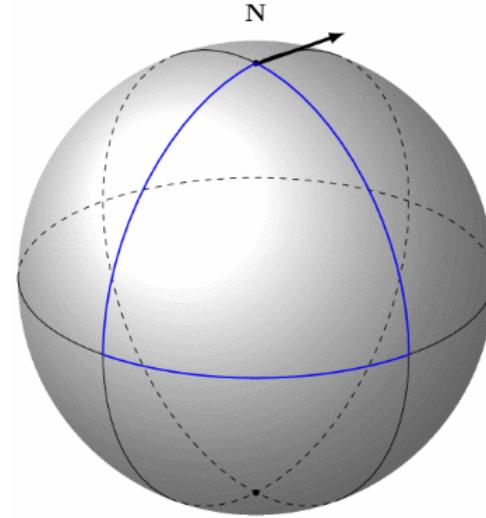
metric tensor  $g^D$  and cubic tensor  $T^D$  coincides with Fisher metric  $g^F$  and Amari-Chentsov tensor  $T$

# Curvature is associated to affine connection $\nabla$

- For Riemannian structure  $(M,g)$ , use default **Levi-Civita connection**  $\nabla=\nabla^g$
- Riemannian manifolds of dim  $d$  can always be embedded into Euclidean spaces  $E^D$  of dim  $D=O(d^2)$
- Euclidean spaces have a natural affine connection  $\nabla=\nabla^E$



Cylinder is flat, 0 curvature:  
Parallel transport along a loop of a  
vector preserves the orientation



Sphere has positive constant curvature:  
Parallel transport along a loop exhibits  
an angle defect related to curvature

© CNRS

# Dually flat spaces $(M, g, \nabla, \nabla^*)$

- **Fundamental theorem of information geometry:** If torsion-free affine connection  $\nabla$  is of constant curvature  $\kappa$ , then curvature of dual torsion-free affine connection  $\nabla^*$  is also constant  $\kappa$
- Corollary: if  $\nabla$  is flat ( $\kappa=0$ ) then  $\nabla^*$  is flat: **Dually flat space  $(M, g, \nabla, \nabla^*)$**
- A connection  $\nabla$  is flat if there exists a local coordinate system  $\theta$  such that  $\Gamma(\theta)=0$
- In  $\nabla$ -affine coordinate system  $\theta(\cdot)$ ,  $\nabla$ -geodesics are visualized as line segments

$$\Gamma(\theta)=0$$
$$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^p \sum_{j=1}^p \Gamma_{ij}^k \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = 0, \quad k = 1, \dots, p,$$

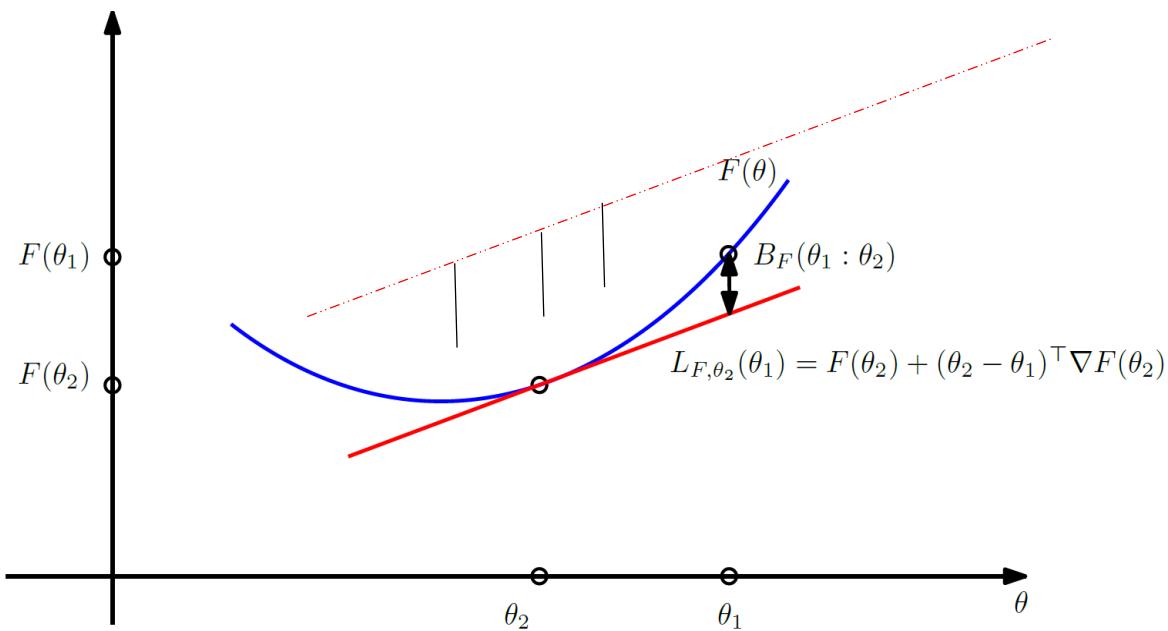
geodesics=line segments in  $\theta$

# Canonical divergences of DFSs: Bregman divergences

- Dually flat structure  $(M, g, \nabla, \nabla^*)$  can be realized by a Bregman divergence

$$(M, g, \nabla, \nabla^*) \leftarrow (M, g^{B_F}, \nabla^{B_F}, \nabla^{B_F^*})$$

- Let  $F(\theta)$  be a strictly convex and differentiable function defined on an open convex domain  $\Theta$
- Bregman divergence interpreted as the vertical gap between point  $(\theta_1, F(\theta_1))$  and the linear approximation of  $F(\theta)$  at  $\theta_2$  evaluated at  $\theta_1$ :



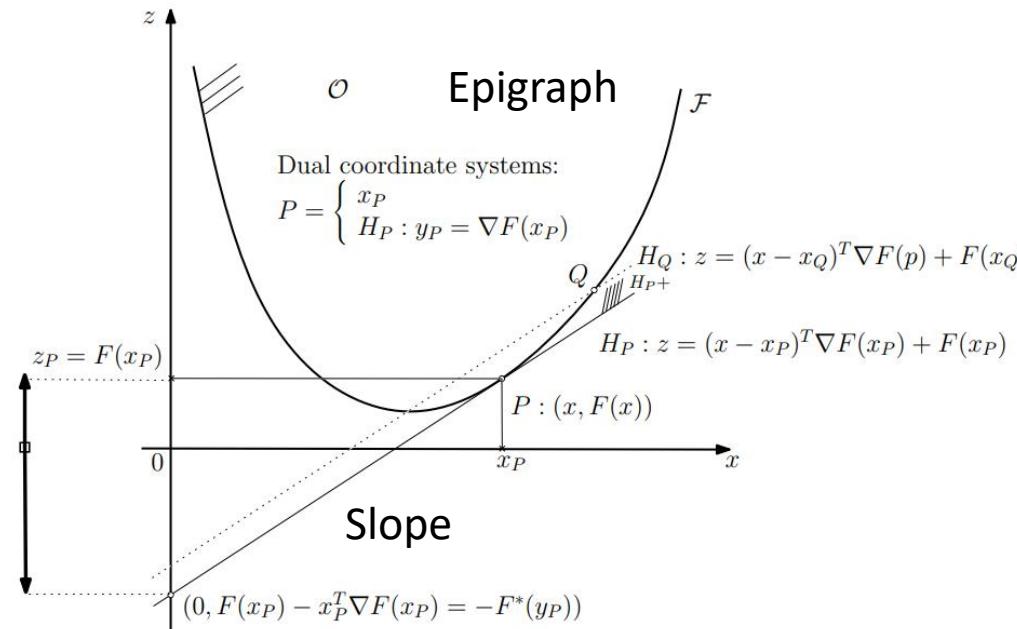
$$\begin{aligned} B_F(\theta_1 : \theta_2) &= F(\theta_1) - \underbrace{\left( F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2) \right)}_{L_{F,\theta_2}(\theta_1)} \\ &= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \end{aligned}$$

[Bregman 1967]

# Legendre-Fenchel transformation: Slope transformation

- Consider a Bregman generator of **Legendre-type** (proper, lower semi-continuous). Then its **convex conjugate** obtained from the **Legendre-Fenchel transformation** is a Bregman generator of Legendre type.

$$\begin{aligned} F^*(\eta) &= \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\} \\ &= - \inf_{\theta \in \Theta} \{F(\theta) - \theta^\top \eta\} \end{aligned}$$



Concave programming:

$$F^*(\eta) = \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\} = \sup_{\theta \in \Theta} \{E(\theta)\}$$

$$\nabla E(\theta) = \eta - \nabla F(\theta) = 0 \Rightarrow \eta = \nabla F(\theta)$$

- Analogy of the Halfspace/Vertex representation of the **epigraph** of  $F$
- Fenchel-Moreau's **biconjugation theorem** for  $F$  of Legendre-type:  $F = (F^*)^*$

[Touchette 2005] Legendre-Fenchel transforms in a nutshell  
 [2010] Legendre transformation and information geometry

# Mixed coordinates and the Legendre-Fenchel divergence

- Dual Legendre-type functions
- Convex conjugate of  $F$  is
- **Fenchel-Young inequality** :

$$\theta = \nabla F^*(\eta) \quad \longleftrightarrow \quad \eta = \nabla F(\theta)$$

$$F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta))$$

$$\underline{F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2}$$

with equality holding if and only if  $\eta_2 = \nabla F(\theta_1)$

$$\nabla F^* = (\nabla F)^{-1}$$

Gradient  
are inverse  
of each other

- **Fenchel-Young divergence** make use of the mixed coordinate systems  $\theta$  et  $\eta$  to express a Bregman divergence as  $B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2)$

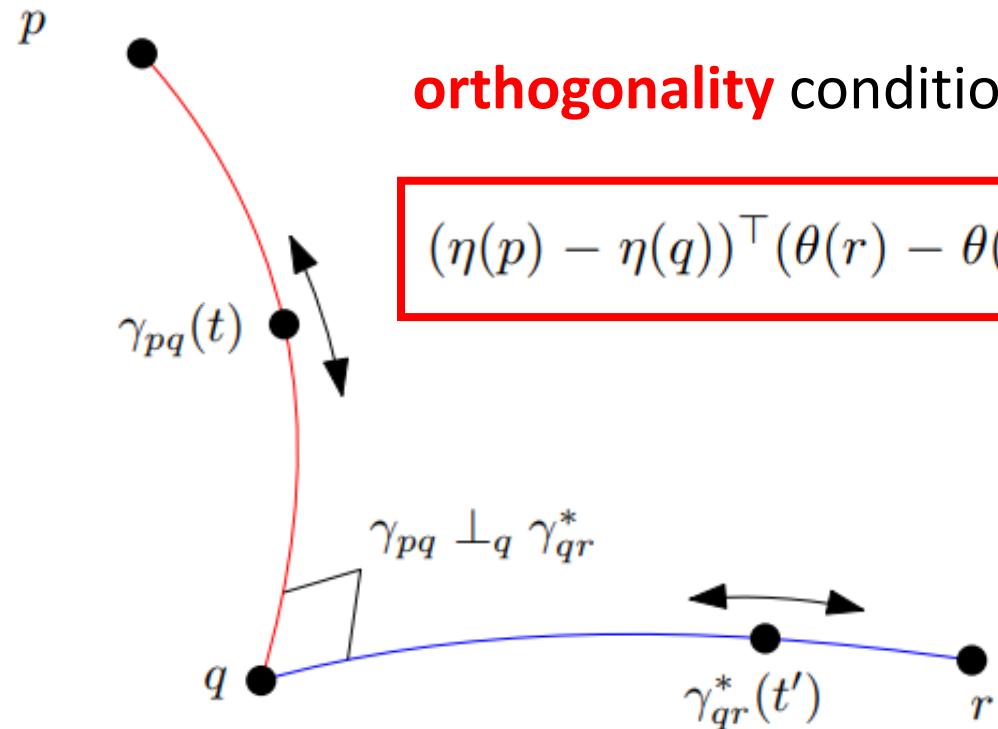
$$Y_{F,F^*}(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 = Y_{F^*,F}(\eta_2, \theta_1)$$

# Generalized Pythagoras theorem in dually flat spaces

In general, **Identity of Bregman divergence with three parameters** = law of cosines

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3)) \geq 0$$

Generalized Pythagoras' theorem

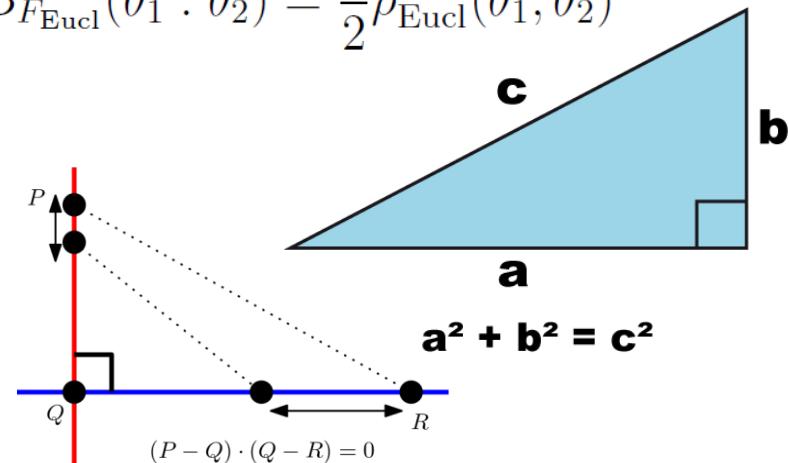


$$D_F(\gamma_{pq}(t) : \gamma_{qr}(t')) = D_F(\gamma_{pq}(t) : q) + D_F(q : \gamma_{qr}^*(t')), \quad \forall t, t' \in (0, 1).$$

Pythagoras' theorem in  
the Euclidian geometry  
(Self-dual)

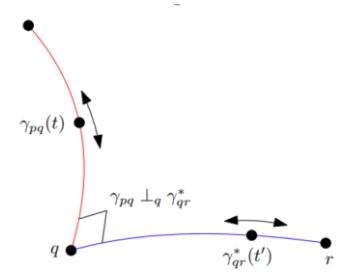
$$F_{\text{Eucl}}(\theta) = \frac{1}{2}\theta^\top \theta \quad g_{F_{\text{Euc}}} = I$$

$$B_{F_{\text{Eucl}}}(\theta_1 : \theta_2) = \frac{1}{2}\rho_{\text{Eucl}}^2(\theta_1, \theta_2)$$

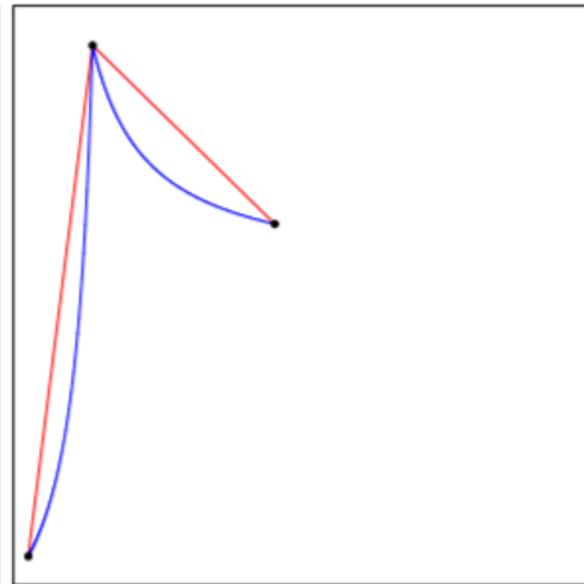
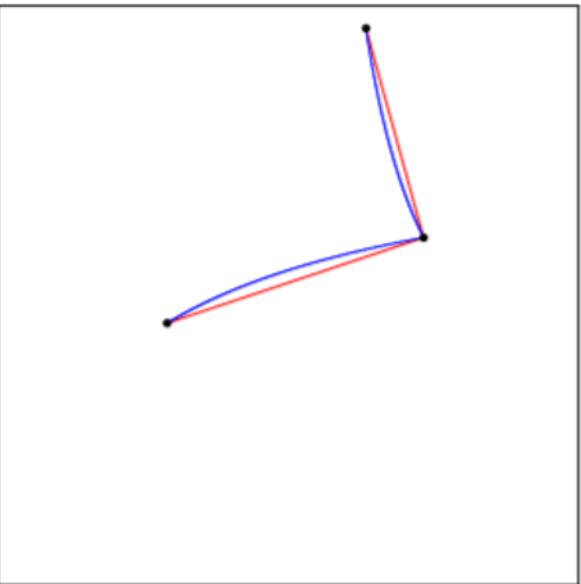
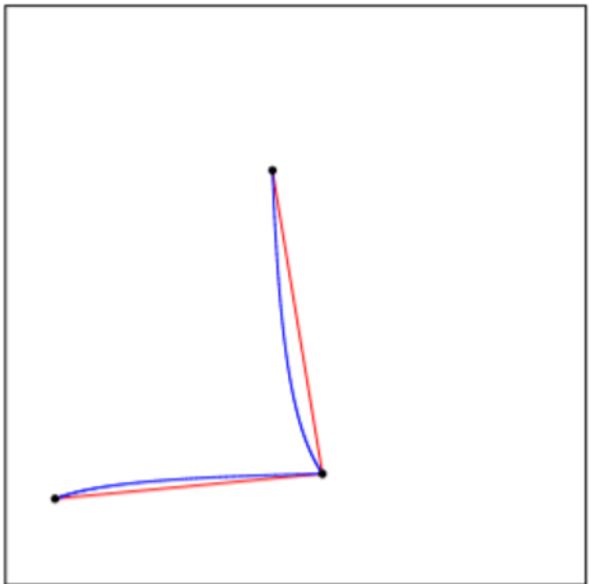


$$\|P - Q\|^2 + \|Q - R\|^2 = \|P - R\|^2$$

# Triples of points $(p, q, r)$ with dual Pythagorean theorems holding simultaneously at $q$



$$\gamma_{pq} \perp_q \gamma_{qr}^* \iff (\theta(p) - \theta(q))^\top (\eta(r) - \eta(q)) = 0 \iff D_F(p : q) + D_F(q : r) = D_F(p : r)$$
$$\gamma_{pq}^* \perp_q \gamma_{qr} \iff (\eta(p) - \eta(q))^\top (\theta(r) - \theta(q)) = 0 \iff D_F(r : q) + D_F(q : p) = D_F(r : p)$$



Itakura-Saito  
Manifold  
(solve quadratic system)

Two blue-red geodesic pairs orthogonal at q

<https://arxiv.org/abs/1910.03935>

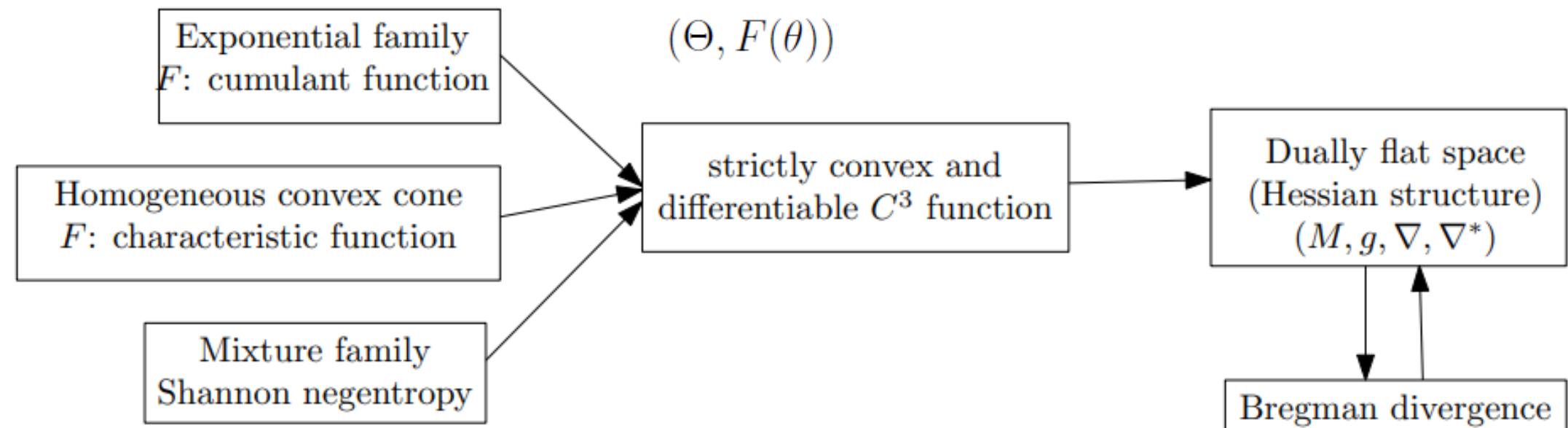
# Dually flat space from a smooth strictly convex function $F(\theta)$

- A smooth strictly convex function  $F(\theta)$  define a Bregman divergence and hence a dually flat space via Eguchi's divergence-based IG

$$(\Theta, F(\theta)) \longrightarrow (M, g^{B_F}, \nabla^{B_F}, \nabla^{B_F^*}) = (M, g^F, \nabla^F, \nabla^{F^*})$$

Domain                          dual Bregman divergences                           $(\nabla^F)^* = \nabla^{(F^*)}$

- Examples of DFSs induced by convex functions:



# Quasi-arithmetic centers, quasi-arithmetic mixtures, and the Jensen-Shannon $\nabla$ -divergences

# Outline and contributions

## Goals:

- I. Generalize scalar quasi-arithmetic means to multivariate cases
- II. Show that the dually flat spaces of information geometry yields a natural framework for defining and studying this generalization

# Weighted quasi-arithmetic means (QAMs)

Standard  $(n-1)$ -dimensional simplex:  $\Delta_{n-1} = \{(w_1, \dots, w_n) : w_i \geq 0, \sum_i w_i = 1\}$

**Definition** (Weighted quasi-arithmetic mean (1930's)). Let  $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$  be a strictly monotone and differentiable real-valued function. The weighted quasi-arithmetic mean (QAM)  $M_f(x_1, \dots, x_n; w)$  between  $n$  scalars  $x_1, \dots, x_n \in I \subset \mathbb{R}$  with respect to a normalized weight vector  $w \in \Delta_{n-1}$ , is defined by

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

QAMs enjoy the in-betweenness property:

$$\min\{x_1, \dots, x_n\} \leq M_f(x_1, \dots, x_n; w) \leq \max\{x_1, \dots, x_n\}$$

# Quasi-arithmetic means (QAMs)

- **Classes of generators**  $[f]=[g]$  with  $f \equiv g$  yieldings the same QAM:

$$M_g(x, y) = M_f(x, y) \text{ if and only if } g(t) = \lambda f(t) + c \text{ for } \lambda \in \mathbb{R} \setminus \{0\}$$

- So let us fix wlog. **strictly increasing and differentiable**  $f$  since we can always either consider either  $f$  or  $-f$  (i.e.,  $\lambda=-1$ ,  $c=0$ ).
- QAMs include **p-power means** for the smooth family of generators  $f_p(t)$ :

$$\dot{M}_p(x, y) := M_{f_p}(x, y) \quad f_p(t) = \begin{cases} \frac{t^p - 1}{p}, & p \in \mathbb{R} \setminus \{0\}, \\ \log(t), & p = 0. \end{cases}, \quad f_p^{-1}(t) = \begin{cases} (1 + tp)^{\frac{1}{p}}, & p \in \mathbb{R} \setminus \{0\}, \\ \exp(t), & p = 0. \end{cases}$$

- Pythagoras means: Harmonic ( $p=-1$ ), Geometric ( $p=0$ ), Arithmetic ( $p=1$ )
- **Homogeneous QAMs**  $M_f(\lambda x, \lambda y) = \lambda \dot{M}_f(x, y)$  for all  $\lambda > 0$  are exactly p-power means

# Quasi-Arithmetic Centers (QACs) = Multivariate QAMs:

Univariate QAMs:  $M_f(x_1, \dots, x_n; w) := f^{-1} \left( \sum_{i=1}^n w_i f(x_i) \right)$

**Two problems** we face when going from univariate to multivariate cases:

1. Define the proper notion of "*multivariate increasing*" function  $F$  and its equivalent class of functions
2. In general, the **implicit function theorem** only proves locally and inverse function  $F^{-1}$  of  $F: \mathbb{R}^d \rightarrow \mathbb{R}^d$  provided its Jacobian matrix is not singular

**Information geometry** provides the right framework to generalize QAMs to quasi-arithmetic centers (QACs) and study their properties.

Consider the **dually flat spaces** of information geometry

# Legendre-type functions

$\Gamma_0(E)$ : Cone of lower semi-continuous (lsc) convex functions from  $E$  into  $\mathbb{R} \cup \{+\infty\}$

**Legendre-Fenchel transformation** of a convex function:  $F^*(\eta) := \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\}$

Problem: Domain  $H$  of  $\eta$  may not be convex...  $F^* \in \Gamma_0(E)$   $F^{**} = F$

counterexample with  $h(\xi_1, \xi_2) = [(\xi_1^2/\xi_2) + \xi_1^2 + \xi_2^2]/4$  [Rockafeller 1967]

To bypass this problem:

**Definition Legendre type function**.  $(\Theta, F)$  is of Legendre type if the function  $F : \Theta \subset \mathbb{X} \rightarrow \mathbb{R}$  is strictly convex and differentiable with  $\Theta \neq \emptyset$  an open convex set and

$$\lim_{\lambda \rightarrow 0} \frac{d}{d\lambda} F(\lambda\theta + (1 - \lambda)\bar{\theta}) = -\infty, \quad \forall \theta \in \Theta, \forall \bar{\theta} \in \partial\Theta. \quad (1)$$

Convex conjugate of a Legendre-type function  $(\Theta, F(\theta))$  is of Legendre-type:

Given by the **Legendre function**:  $F^*(\eta) = \langle \nabla F^{-1}(\eta), \eta \rangle - F(\nabla F^{-1}(\eta))$   
Gradient map  $\nabla F$  is globally invertible:  $\nabla F^{-1}$

# Comonotone functions in inner product spaces

- **Comonotone functions:**  $\forall \theta_1, \theta_2 \in \mathbb{X}, \theta_1 \neq \theta_2, \quad \langle \theta_1 - \theta_2, G(\theta_1) - G(\theta_2) \rangle > 0$   
(i.e., **co**monotone = monotone with respect to the **identity function**)

**Proposition (Gradient co-monotonicity).** *The gradient functions  $\nabla F(\theta)$  and  $\nabla F^*(\eta)$  of the Legendre-type convex conjugates  $F$  and  $F^*$  in  $\mathcal{F}$  are strictly increasing co-monotone functions.*

Proof using symmetrization of Bregman divergences = Jeffreys-Bregman divergence:

$$B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) = \langle \theta_2 - \theta_1, \nabla F(\theta_2) - \nabla F(\theta_1) \rangle > 0, \quad \forall \theta_1 \neq \theta_2$$

$$B_{F^*}(\eta_1 : \eta_2) + B_{F^*}(\eta_2 : \eta_1) = \langle \eta_2 - \eta_1, \nabla F^*(\eta_2) - \nabla F^*(\eta_1) \rangle > 0, \quad \forall \eta_1 \neq \eta_2$$

because Bregman divergences(and sums thereof) are always non-negative

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle \geq 0,$$

$$B_{F^*}(\eta_1 : \eta_2) = F^*(\eta_1) - F^*(\eta_2) - \langle \eta_1 - \eta_2, \nabla F^*(\eta_2) \rangle \geq 0.$$

Remark: **Generalization of monotonicity** because when  $d=1$ ,  $f(x)$  is strictly monotone iff  $f(x_1)-f(x_2)$  is of same sign of  $x_1-x_2$  that is,  $(f(x_1)-f(x_2)) (x_1-x_2) > 0$

# Quasi-arithmetic centers: Definition generalizing QAMs

**Definition (Quasi-arithmetic centers, QACs).** Let  $F : \Theta \rightarrow \mathbb{R}$  be a strictly convex and smooth real-valued function of Legendre-type in  $\mathcal{F}$ . The weighted quasi-arithmetic average of  $\theta_1, \dots, \theta_n$  and  $w \in \Delta_{n-1}$  is defined by the gradient map  $\nabla F$  as follows:

$$\begin{aligned} M_{\nabla F}(\theta_1, \dots, \theta_n; w) &:= \nabla F^{-1} \left( \sum_i w_i \nabla F(\theta_i) \right), \\ &= \nabla F^* \left( \sum_i w_i \nabla F(\theta_i) \right), \end{aligned}$$

where  $\nabla F^* = (\nabla F)^{-1}$  is the gradient map of the Legendre transform  $F^*$  of  $F$ .

This definition generalizes univariate quasi-arithmetic means :  $M_f(x_1, \dots, x_n; w) := f^{-1} \left( \sum_{i=1}^n w_i f(x_i) \right)$   
Let  $F(t) = \int_a^t f(u) du$

Then we have  $M_f = M_{F'}$

# An illustrating example: The matrix harmonic mean

- Consider the real-value minus **logdet function**  $F(\theta) = -\log \det(\theta)$
- Domain  $F: \text{Sym}_{++}(d) \rightarrow \mathbb{R}$  the cone of symmetric positive-definite matrices
- Inner product:  $\langle A, B \rangle := \text{tr}(AB^\top)$
- We have:
  - $F(\theta) = -\log \det(\theta), \quad \leftarrow \text{Legendre-type function}$
  - $\nabla F(\theta) = -\theta^{-1} =: \eta(\theta),$
  - $\nabla F^{-1}(\eta) = -\eta^{-1} =: \theta(\eta)$
  - $F^*(\eta) = \langle \theta(\eta), \eta \rangle - F(\theta(\eta)) = -d - \log \det(-\eta) \quad \leftarrow \text{Legendre-type function}$

The quasi-arithmetic center with respect to  $F: M_{\nabla F}(\theta_1, \theta_2) = 2(\theta_1^{-1} + \theta_2^{-1})^{-1}$

The quasi-arithmetic center with respect to  $F^*: M_{\nabla F^*}(\eta_1, \eta_2) = 2(\eta_1^{-1} + \eta_2^{-1})^{-1}$

Generalize univariate harmonic mean with  $F(x) = \log x, f(x) = F'(x) = 1/x: H(a, b) = \frac{2ab}{a+b}$  for  $a, b > 0$

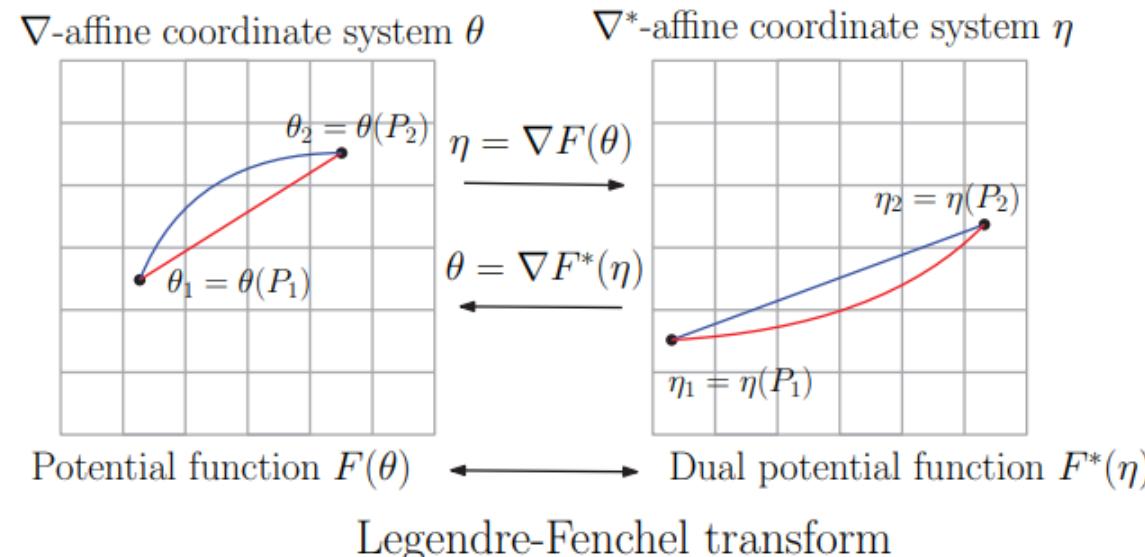
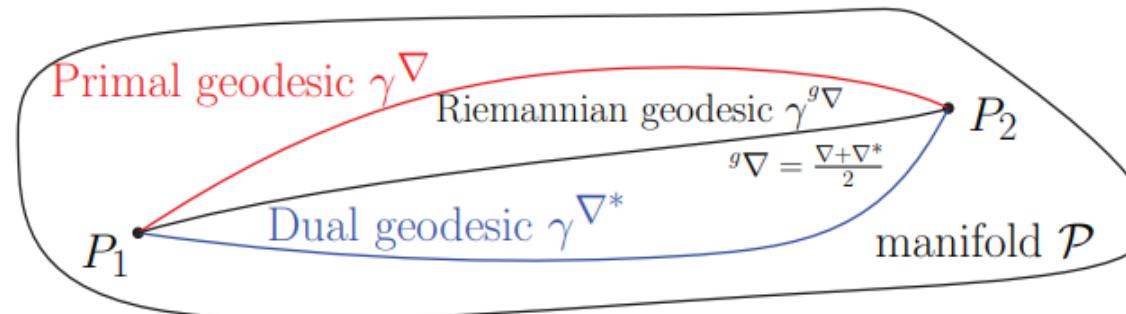
**A Legendre-type function  $F$  gives rise to a pair of dual quasi-arithmetic centers**  
 **$M_{\nabla F}$  and  $M_{\nabla F^*}$  : dual operators**

# Dually flat structures of information geometry

- A Legendre-type Bregman generator  $F()$  induces a **dually flat space structure**:

$$(\Theta, g(\theta) = \nabla_\theta^2 F(\theta), \nabla, \nabla^*)$$

- A point  $P$  can be either parameterized by  $\theta$ -coordinate and dual  $\eta$ -coordinate



# Quasi-arithmetic barycenters and dual geodesics

- The **dual geodesics** induced by the dual flat connections can be expressed using **dual weighted quasi-arithmetic centers**:

$$\nabla\text{-geodesic } \gamma_{\nabla}(P, Q; t) = (PQ)^{\nabla}(t)$$

$$(PQ)^{\nabla}(t) = \begin{pmatrix} M_{\text{id}}(\theta(P), \theta(Q); 1-t, t) \\ M_{\nabla F^*}(\eta(P), \eta(Q); 1-t, t) \end{pmatrix} \quad \leftarrow \text{dual QAC } M_{\nabla F^*}$$



$$\nabla^*\text{-geodesic } \gamma_{\nabla^*}(P, Q; t) = (PQ)^{\nabla^*}(t)$$

$$(PQ)^{\nabla^*}(t) = \begin{pmatrix} M_{\nabla F}(\theta(P), \theta(Q); 1-t, t) \\ M_{\text{id}}(\eta(P), \eta(Q); 1-t, t) \end{pmatrix} \quad \leftarrow \text{primal QAC } M_{\nabla F}$$

# n-Variable Quasi-arithmetic centers as centroids in dually flat spaces

Consider  $n$  points  $P_1, \dots, P_n$  on the DFS  $(M, g, \nabla, \nabla^*)$  (canonical divergence = Bregman divergence)

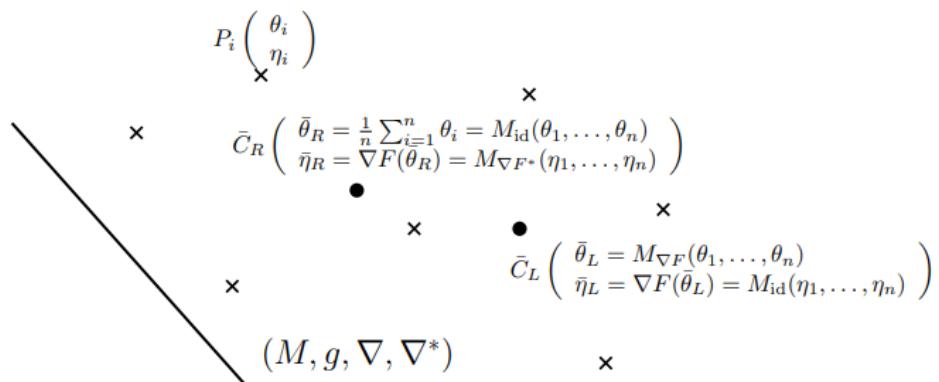
Right-sided centroid:

$$\bar{C}_R = \arg \min_{P \in M} \sum_{i=1}^n \frac{1}{n} D_{\nabla, \nabla^*}(P_i : P)$$

$$\bar{\theta}_R = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n B_F(\theta_i : \theta)$$

$$\bar{\theta}_R = \theta(\bar{C}_R) = \frac{1}{n} \sum_{i=1}^n \theta_i = M_{\text{id}}(\theta_1, \dots, \theta_n)$$

$$\bar{\eta}_R = \nabla F(\bar{\theta}_R) = M_{\nabla F^*}(\eta_1, \dots, \eta_n). \quad \leftarrow \text{dual QAC}$$



Reference duality

Left-sided centroid:

$$\bar{C}_L = \arg \min_{P \in M} \sum_{i=1}^n \frac{1}{n} D_{\nabla, \nabla^*}(P : P_i)$$

$$\bar{\theta}_L = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n B_F(\theta : \theta_i)$$

$$\bar{\theta}_L = M_{\nabla F}(\theta_1, \dots, \theta_n), \quad \leftarrow \text{primal QAC}$$

$$\bar{\eta}_L = \nabla F(\bar{\theta}_L) = M_{\text{id}}(\eta_1, \dots, \eta_n)$$

Notice that when  $n=2$ , weighted dual quasi-arithmetic barycenters define the dual geodesics

# Invariance/equivariance of quasi-arithmetic centers

Information geometry is well-suited to study the **properties of QACs**:

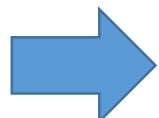
A dually flat space (DFS) can be **realized** by a class of Bregman generators:

$$(M, g, \nabla, \nabla^*) \leftarrow \text{DFS}([\theta, F(\theta); \eta, F^*(\eta)])$$

## Affine Legendre invariance of dually flat spaces:

- By adding an affine term...

Same DFS with  $\bar{F}(\theta) = F(\theta) + \langle c, \theta \rangle + d$ .



### Invariance of quasi-arithmetic center:

$$M_{\nabla \bar{F}}(\theta_1, \dots; \theta_n; w) = M_{\nabla F}(\theta_1, \dots; \theta_n; w)$$

- By an affine change of coordinate...

Same DFS with  $\bar{\theta} = A\theta + b$  such that  $\bar{F}(\bar{\theta}) = F(\theta)$

$$\nabla \bar{F}(x) = (A^{-1})^\top \nabla F(A^{-1}(x - b))$$

$$B_{\bar{F}(\bar{\theta}_1 : \bar{\theta}_2)} = B_F(\theta_1 : \theta_2)$$

### Equivariance of quasi-arithmetic center:

$$M_{\nabla \bar{F}}(\bar{\theta}_1, \dots, \bar{\theta}_n; w) = A M_{\nabla F}(\theta_1, \dots, \theta_n; w) + b$$

Same canonical divergence of the DFS  
(= contrast function on the diagonal of the product manifold)

# Canonical divergence versus Legendre-Fenchel/Bregman divergences

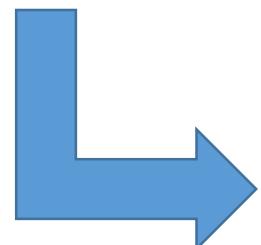
- Canonical divergence induced by dual flat connections is between **points**
- dual Bregman divergences  $B_F$  and  $B_{F^*}$  between **dual coordinates**
- Legendre-Fenchel divergence  $Y_F$  between **mixed coordinates**

$$F(\theta) + F^*(\eta) - \langle \theta, \eta \rangle = 0 \quad \eta = \nabla F(\theta)$$

$$\begin{aligned} B_F(\theta_1 : \theta_2) &:= F(\theta_1) - \underbrace{F(\theta_2)}_{= \langle \theta_2, \eta_2 \rangle - F^*(\eta_2)} - \langle \theta_1 - \theta_2, \nabla F(\eta_2) \rangle \\ &= F(\theta_1) + F^*(\eta_2) - \langle \theta_1, \eta_2 \rangle =: Y_F(\theta_1 : \eta_2) \end{aligned}$$

$$(M, g, \nabla, \nabla^*) \leftarrow \text{DFS}([\Theta, F(\theta), H, F^*(\eta)])$$

$$\leftarrow \text{DFS}([\bar{\Theta}, \bar{F}(\bar{\theta}), \bar{H}, \bar{F}^*(\bar{\eta})])$$



$$\begin{aligned} D_{\nabla, \nabla^*}(P_1 : P_2) &= B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_1, \eta_2) = Y_F(\theta_1 : \eta_2) = Y_{F^*}(\eta_2 : \theta_1) \\ &= B_{\bar{F}}(\bar{\theta}_1 : \bar{\theta}_2) = B_{\bar{F}^*}(\bar{\eta}_1, \bar{\eta}_2) = Y_F(\bar{\theta}_1 : \bar{\eta}_2) = Y_{F^*}(\bar{\eta}_2 : \bar{\theta}_1) \end{aligned}$$

# Affine Legendre invariance of dually flat spaces plus setting the unit scale of divergences

- Affine Legendre invariance:

$$\bar{F}(\bar{\theta}) = F(A\theta + b) + \langle c, \theta \rangle + d$$

$$\bar{F}^*(\bar{\eta}) = F^*(A^*\eta + b^*) + \langle c^*, \eta \rangle + d^*$$

- Set the unit scale of canonical divergence (DFS differ here, rescaled):

(does not change the quasi-arithmetic center)  $D_{\lambda, \nabla, \nabla^*} := \lambda D_{\nabla, \nabla^*}$

amount to scale the potential function  $\lambda F(\theta)$  vs  $F(\theta)$

**Proposition (Invariance and equivariance of QACs).** Let  $F(\theta)$  be a function of Legendre type. Then  $\bar{F}(\bar{\theta}) := \lambda(F(A\theta + b) + \langle c, \theta \rangle + d)$  for  $A \in \mathrm{GL}(d)$ ,  $b, c \in \mathbb{R}^d$ ,  $d \in \mathbb{R}^d$  and  $\lambda \in \mathbb{R}_{>0}$  is a Legendre-type function, and we have

$$M_{\nabla \bar{F}} = A M_{\nabla F} + b.$$

# Illustrating example: Mahalanobis divergence

- **Mahalanobis divergence** = squared Mahalanobis metric distance

$$\Delta^2(\theta_1, \theta_2) = B_{F_Q}(\theta_1 : \theta_2) = \frac{1}{2}(\theta_2 - \theta_1)^\top Q (\theta_2 - \theta_1) \quad \text{fails triangle inequality of metric distances}$$

Primal potential function:  $F_Q(\theta) = \frac{1}{2}\theta^\top Q\theta + c\theta + \kappa$

Dual potential function:  $F^*(\eta) = \frac{1}{2}\eta^\top Q^{-1}\eta = F_{Q^{-1}}(\eta),$

- The dual QACs induced by the dual Mahalanobis generators  $F$  and  $F^*$  coincide to **weighted arithmetic mean**  $M_{\text{id}}$ :

$$M_{\nabla F_Q}(\theta_1, \dots, \theta_n; w) = Q^{-1} \left( \sum_{i=1}^n w_i Q \theta_i \right) = \sum_{i=1}^n w_i \theta_i = M_{\text{id}}(\theta_1, \dots, \theta_n; w),$$

$$M_{\nabla F_Q^*}(\eta_1, \dots, \eta_n; w) = Q \left( \sum_{i=1}^n w_i Q^{-1} \eta_i \right) = M_{\text{id}}(\eta_1, \dots, \eta_n; w).$$

# Quasi-arithmetic mixtures (QAMixs), and $\alpha$ -mixtures

**Definition** . The  $M_f$ -mixture of  $n$  densities  $p_1, \dots, p_n$  weighted by  $w \in \Delta_n^\circ$  is defined by

$$(p_1, \dots, p_n; w)^{M_f}(x) := \frac{M_f(p_1(x), \dots, p_n(x); w)}{\int M_f(p_1(x), \dots, p_n(x); w) d\mu(x)}.$$

**Centroid** of  $n$  densities with respect to the  $\alpha$ -divergences yields a QAMix:

$$(p_1, \dots, p_n; w)^{M_\alpha} = \arg \min_p \sum_i w_i D_\alpha(p_i, p)$$

$D_\alpha$  denotes the  $\alpha$ -divergences:

$$D_\alpha [m(s) : l(s)] = \begin{cases} \int m(s) ds - \int l(s) ds + \int m(s) \log \frac{m(s)}{l(s)} ds & \alpha = -1 \\ \int l(s) ds - \int m(s) ds + \int l(s) \log \frac{l(s)}{m(s)} ds + \int l(s) \log \frac{l(s)}{m(s)} ds & \alpha = 1 \\ \frac{2}{1+\alpha} \int m(s) ds + \frac{2}{1-\alpha} \int l(s) ds - \frac{4}{1-\alpha^2} \int m(s)^{\frac{1-\alpha}{2}} l(s)^{\frac{1+\alpha}{2}} ds, & \alpha \neq \pm 1. \end{cases}$$

# $k=2$ QAMixs and the $\nabla$ -Jensen-Shannon divergence

- **Jensen-Shannon divergence** is bounded symmetrization of KL divergence:

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

- Interpret arithmetic mixture as the **midpoint of a mixture geodesic** (wrt to the flat non-parametric mixture connection  $\nabla^m$  in information geometry).
- Generalize Jensen-Shannon divergence with **arbitrary  $\nabla$ -connections**:

**Definition (Affine connection-based  $\nabla$ -Jensen-Shannon divergence).**

Let  $\nabla$  be an affine connection on the space of densities  $\mathcal{P}$ , and  $\gamma_\nabla(p, q; t)$  the geodesic linking density  $p = \gamma_\nabla(p, q; 0)$  to density  $q = \gamma_\nabla(p, q; 1)$ . Then the  $\nabla$ -Jensen-Shannon divergence is defined by:

$$D_{\nabla}^{\text{JS}}(p, q) := \frac{1}{2} \left( D_{\text{KL}} \left( p : \gamma_\nabla \left( p, q; \frac{1}{2} \right) \right) + D_{\text{KL}} \left( q : \gamma_\nabla \left( p, q; \frac{1}{2} \right) \right) \right).$$

# Inductive Means: Geodesics/quasi-arithmetic centers

- Gauss and Lagrange independently studied the following convergence of pairs of iterations:

$$\begin{aligned} a_{t+1} &= \frac{a_t + b_t}{2} \\ b_{t+1} &= \sqrt{a_t b_t} \end{aligned}$$

and proves quadratic convergence to the **arithmetic-geometric mean AGM**

$$\text{AGM}(a_0, b_0) = \frac{\pi}{4} \frac{a_0 + b_0}{K\left(\frac{a_0 - b_0}{a_0 + b_0}\right)}$$

where K is complete elliptic integral of the first kind  
AGM also used to approximate ellipse perimeter and  $\pi$

- In general, choosing two strict means M and M' with interness property will converge but difficult to *analytically express the common limits of iterations*
- When M=Arithmetic and M'=Harmonic, the **arithmetic-harmonic mean AHM** yields the geometric mean:

$$\begin{aligned} a_{t+1} &= A(a_t, h_t) \\ h_{t+1} &= H(a_t, h_t) \end{aligned}$$

$$\text{AHM}(x, y) = \lim_{t \rightarrow \infty} a_t = \lim_{t \rightarrow \infty} h_t = \sqrt{xy} = G(x, y)$$

# Inductive matrix arithmetic-harmonic mean

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

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

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

- Then the sequences converge quadratically to the **matrix geometric mean**:

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

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

which is also the **Riemannian center of mass** with respect to the trace metric:

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

$$g_P(V_1, V_2) = \text{tr}(P^{-1} V_1 P^{-1} V_2)$$

[Nakamura 2001, Atteia-Raissouli 2001 ]

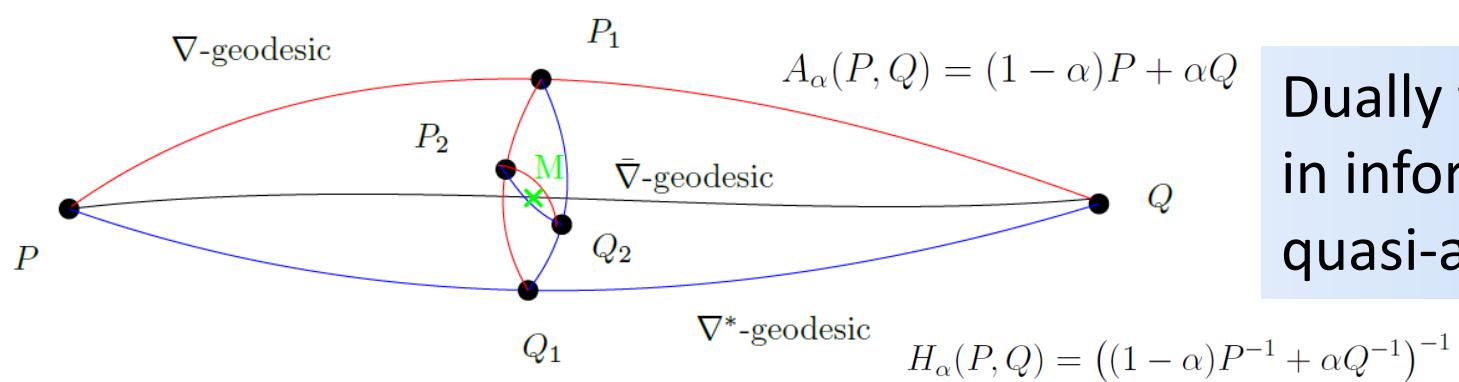
# Geometric interpretation of the AHM matrix mean

$$A_{t+1} = \frac{A_t + H_t}{2} = A(A_t, H_t)$$

$$H_{t+1} = 2(A_t^{-1} + H_t^{-1})^{-1} = H(A_t, H_t)$$

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

**(SPD,  $g^G$ ,  $\nabla^A$ ,  $\nabla^H$ ) 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,  $g^G$ ,  $\nabla^A$ ,  $\nabla^H$ ) in information geometry defines quasi-arithmetic centers as geodesic midpoints

Primal geodesic midpoint is the arithmetic center wrt Euclidean metric  $g_P^A(X, Y) = \text{tr}(X^\top Y)$

Dual geodesic midpoint = harmonic center wrt an isometric Eucl. metric  $g_P^H(X, Y) = \text{tr}(P^{-2} X P^{-2} Y)$

Levi-Civita geodesic midpoint is geometric Karcher mean (not QAC)  $g_P^G(X, Y) = \text{tr}(P^{-1} X P^{-1} Y)$

$$g_P(V_1, V_2) = \text{tr}(P^{-1} V_1 P^{-1} V_2)$$

A balanced metric

[Nakamura 2001, Thanwerdas & Pennec 2019]

# Revisiting Chernoff information with Likelihood Ratio Exponential Families

# Chernoff information: Definition & Background

## A symmetric statistical divergence

- Originally introduced by Chernoff (1952) to *upper bound the probability of error* (Bayes' error) in statistical hypothesis testing.

Definition:

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

$$\rho_\alpha[P : Q] := \int p^\alpha q^{1-\alpha} d\mu = \rho_{1-\alpha}[Q : P] \quad 0 < \rho_\alpha[P : Q] \leq 1.$$

(via Hölder inequality)



Herman Chernoff  
(1923-)

- **skewed Bhattacharyya coefficient  $\rho_\alpha$**  (similarity coefficient)
- Synonyms: Chernoff divergence, Chernoff information number, Chernoff index...
- Found later many applications in information fusion, radar target detection, generative adversarial networks (GANs), etc. due to its empirical robustness

# Chernoff information =

## Maximally skewed Bhattacharyya distance

- **skewed Bhattacharyya distance** (a Ali-Silvey **f-divergence**):

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

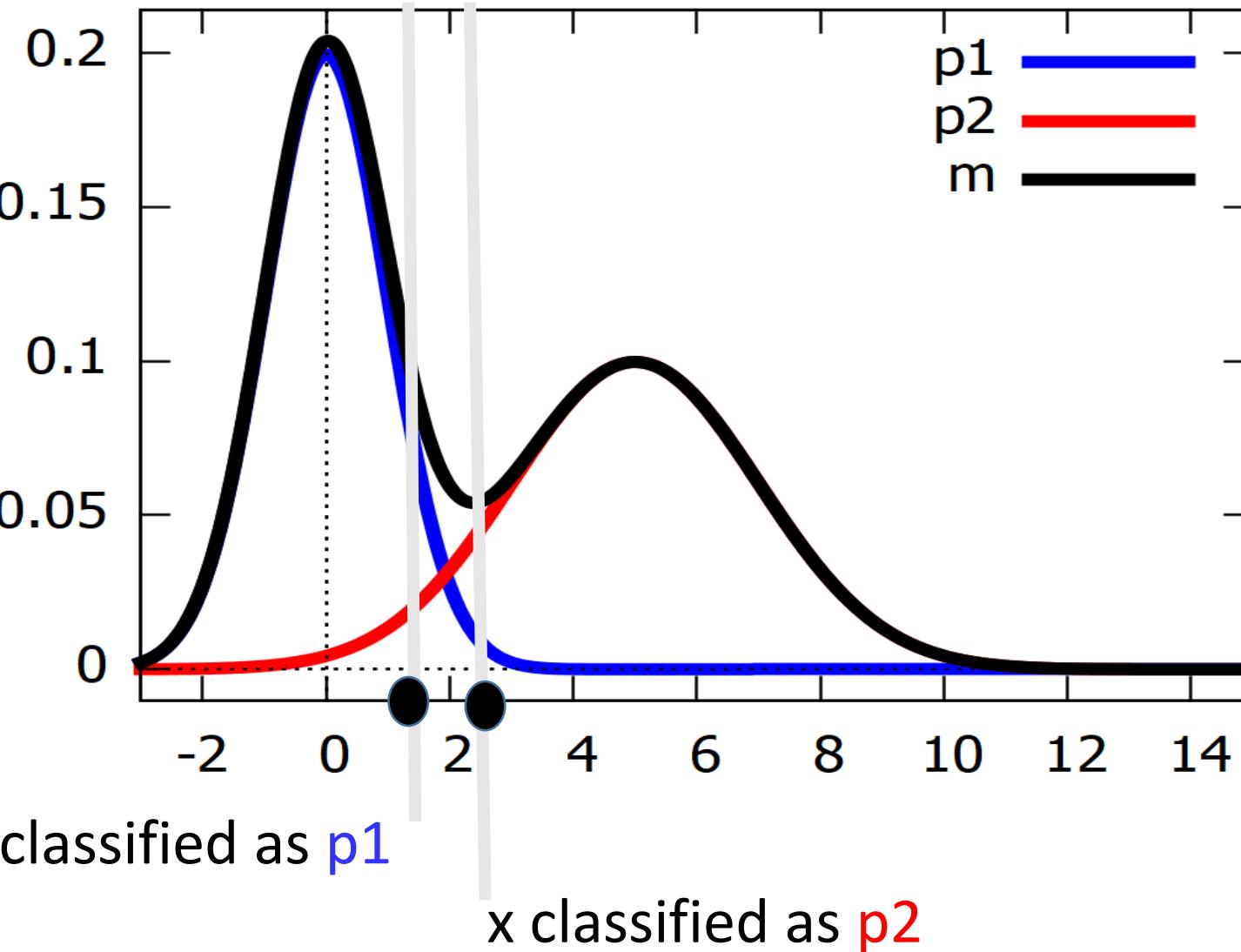
- **Chernoff information:**  $D_C[p, q] = \max_{\alpha \in (0,1)} D_{B,\alpha}[p : q].$

- **scaled skewed Bhattacharyya distance = Rényi divergence** (extends KLD)

$$D_{R,\alpha}[P : Q] = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu = \frac{1}{1-\alpha} D_{B,\alpha}[P : Q] \quad \alpha \in [0, \infty] \setminus \{1\}$$

- Optimal values of  $\alpha$  is called ``**Chernoff (error) exponent**'' (due to its seminal use in statistical hypothesis testing)

# Rationale for CI: Statistical hypothesis testing



Statistical mixture:

$$m(x) = 0.5 \cdot N(0, 1) + 0.5 \cdot N(5, 2)$$

Hypothesis task:

Decides whether  $x$  emanates from  $p_1$  or  $p_2$ ?

Classification rule:

**Maximum a posteriori** (MAP)

if  $p_1(x) > p_2(x)$  classify as  $p_1$   
else classify as  $p_2$

Error at  $x$ :  $\min(p_1(x), p_2(x))$

**Histogram intersection similarity:**

$$P_e = \int \min(p_1(x), p_2(x)) dx$$

# Rewriting and bounding the probability of error

- Use **rewriting trick**  $\min(a,b) = (a+b)/2 + |b-a|/2$  for  $a,b>0$   
express the probability of error using the **total variation distance**:

$$P_e = \int \min(p_1(x), p_2(x))dx \quad \longrightarrow \quad P_e = \frac{1}{2} (1 - D_{\text{TV}}[p_1, p_2])$$
$$D_{\text{TV}}[p_1, p_2] = \frac{1}{2} \int (p_1(x) - p_2(x))dx$$

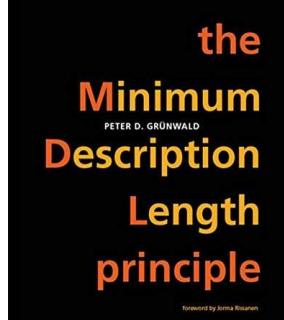
- Use a **generic (weighted) mean** which necessarily falls inbetween its extrema (e.g., **geometric mean**):

$$\min(a, b) \leq M(a, b) \leq \max(a, b) \quad \longrightarrow \quad \min(a, b) \leq M_\alpha(a, b) \leq \max(a, b), \forall \alpha \in [0, 1]$$

$$P_e = \int \min(p_1(x), p_2(x))dx \leq \min_{\alpha \in [0,1]} \int M_\alpha(p_1(x), p_2(x))dx \quad \xrightarrow{\substack{M_\alpha(a, b) = a^\alpha b^{1-\alpha} \\ \text{geometric weighted mean}}} \quad P_e \leq \rho_\alpha(p_1, p_2)$$

"Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means." *Pattern Recognition Letters* 42 (2014): 25-34.

# Likelihood ratio exponential families (LREFs)



- **Geometric mixture** (Bhattacharyya /exponential arc )

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

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

- 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

Sufficient statistics: log likelihood ratio

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

Natural parameter space:

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

$k(x) = \log q(x)$

# LREFs: EF cumulant function is always analytic $C^\omega$

- Cumulant function of EF is **strictly convex**  
(and smooth for regular EFs)

- Cumulant function is neg-Bhattacharyya distance:

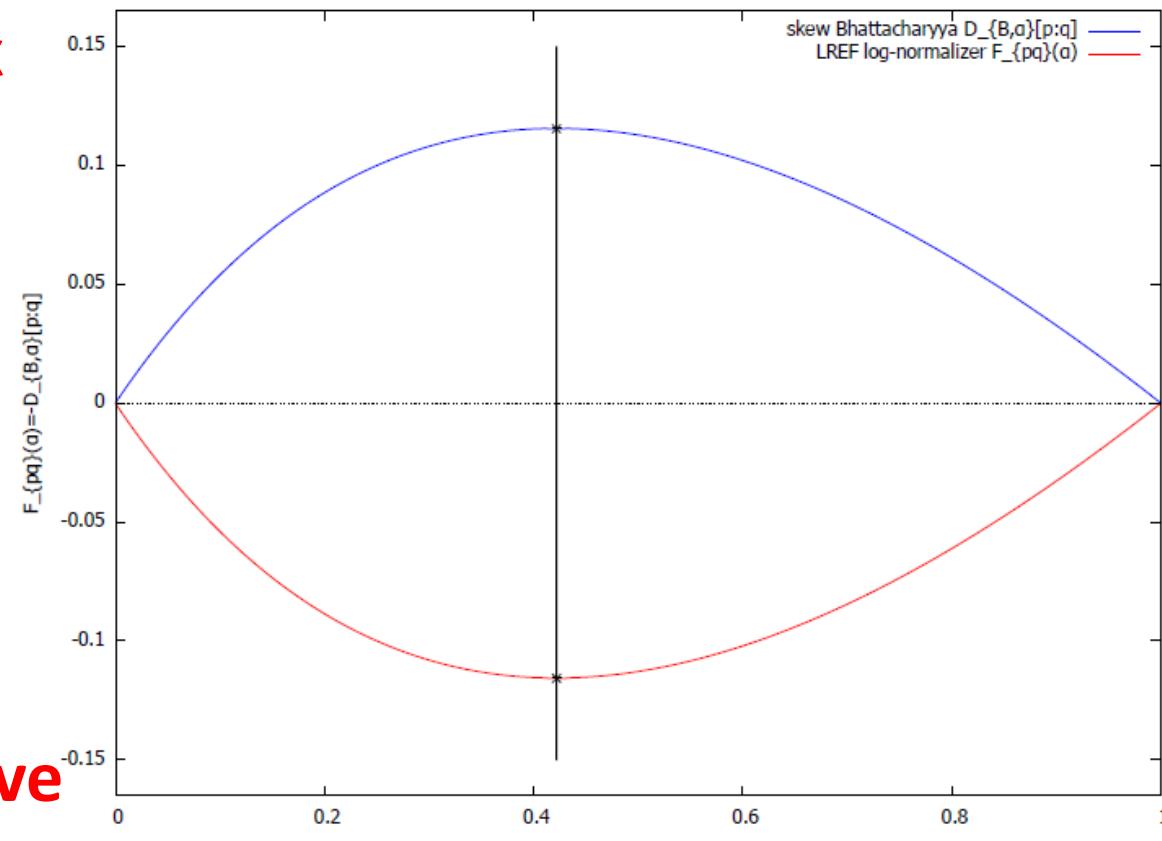
$$F_{pq}(\alpha) = \log Z_{pq}(\alpha) = -D_{B,\alpha}[p : q] < 0$$

⇒ Bhattacharyya. distance is **strictly concave**

- Theorem:

**Chernoff exponent exists and is unique**

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



$$p=N(0,1)$$

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

$$q=N(1,2)$$

$$\alpha^*(q : p) = 1 - \alpha^*(p : q)$$

# Geometric mixtures and LREFs: Regular EFs

- Natural parameter space:  $\Theta_{pq} = \{\alpha \in \mathbb{R} : \rho_\alpha(p : q) < +\infty\}$   
**always contains (0,1)** since  $0 < \rho_\alpha[P : Q] \leq 1$ .
- What happens at extremities and when extrapolating (depends on support):  
$$\text{supp}\left((pq)_\alpha^G\right) = \begin{cases} \text{supp}(p) \cap \text{supp}(q), & \alpha \in \Theta_{pq} \setminus \{0, 1\} \\ \text{supp}(p), & \alpha = 1 \\ \text{supp}(q), & \alpha = 0. \end{cases}$$
- Exponential family is said **regular** when the natural parameter space  $\Theta$  is **open** (e.g., normal family, Dirichlet family, Wishart family, etc.)

Definition:

regular EF



$\Theta = \Theta^\circ$

# When $(0,1)$ is strictly included in regular LREFs

**Proposition** (Finite sided Kullback-Leibler divergences). *When the LREF  $\mathcal{E}_{pq}$  is a regular exponential family with natural parameter space  $\Theta \supsetneq [0, 1]$ , both the forward Kullback-Leibler divergence  $D_{\text{KL}}[p : q]$  and the reverse Kullback-Leibler divergence  $D_{\text{KL}}[q : p]$  are finite.*

$$D_{\text{KL}}[P : Q] = D_{\text{KL}}[p : q] = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

- **KLD between two densities of a regular EF = reverse Bregman divergence:**

$$\begin{aligned} D_{\text{KL}}[p_{\theta_1} : p_{\theta_2}] &= E_{p_{\theta_1}} \left[ \log \frac{p_{\theta_1}}{p_{\theta_2}} \right], \\ &= F(\theta_2) - F(\theta_1) - (\theta_1 - \theta_2)^{\top} E_{p_{\theta_1}}[t(x)]. \end{aligned}$$

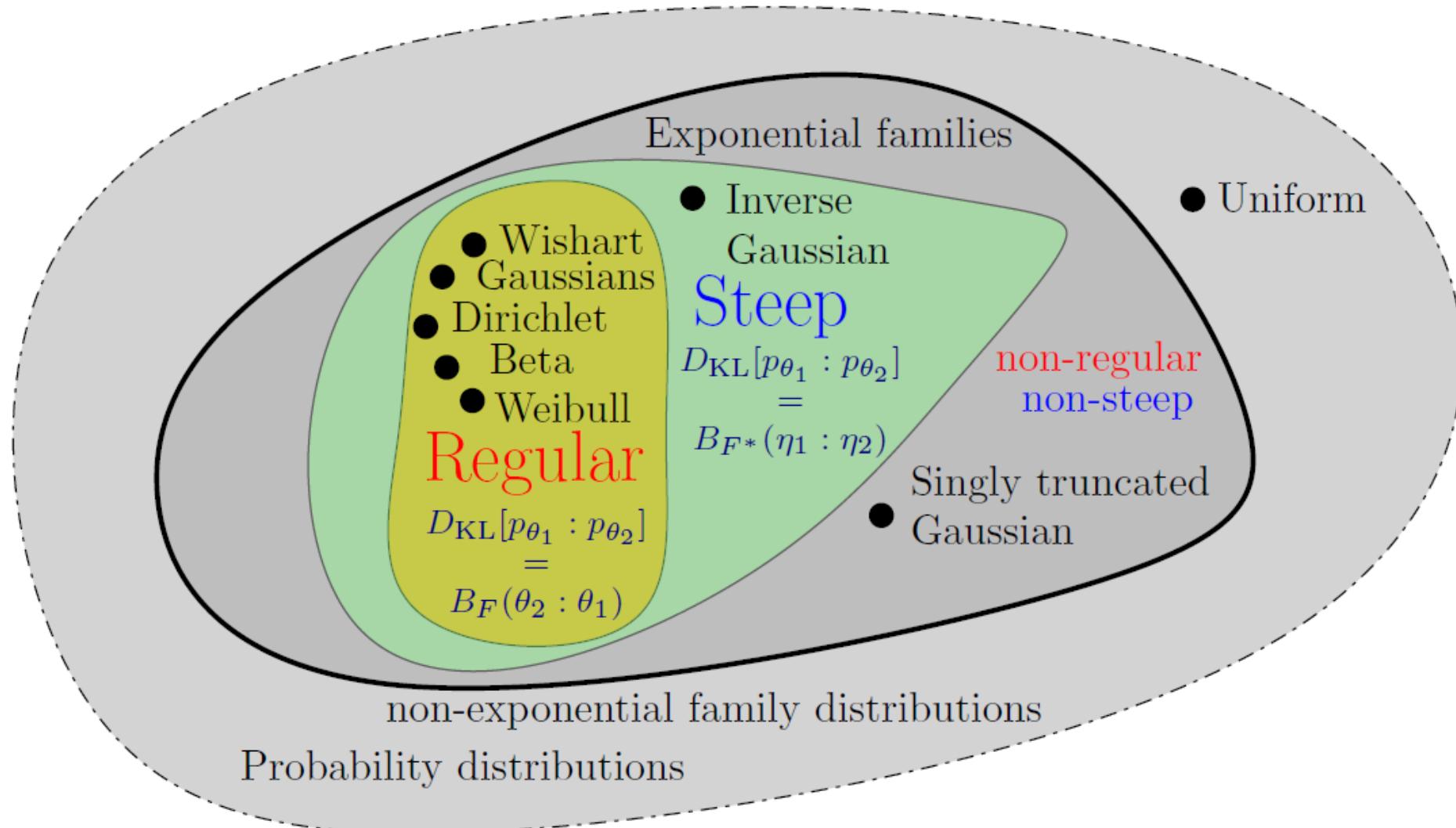
**steep**  $\Rightarrow E_{p_{\theta_1}}[t(x)] = \nabla F(\theta_1)$

**regular** EF  $\Rightarrow$  steep EF

$$D_{\text{KL}}[p_{\theta_1} : p_{\theta_2}] = F(\theta_2) - F(\theta_1) - (\theta_1 - \theta_2)^{\top} \nabla F(\theta_1) =: B_F(\theta_2 : \theta_1) = (B_F)^*(\theta_1 : \theta_2).$$

# Venn diagram: Regular & steepness of (LR)EFs

- Steepness implies **duality between natural  $\theta$  and moment  $\eta$  parameters**



**Proposition** (Finite sided Kullback-Leibler divergences). *When the LREF  $\mathcal{E}_{pq}$  is a regular exponential family with natural parameter space  $\Theta \supsetneq [0, 1]$ , both the forward Kullback-Leibler divergence  $D_{\text{KL}}[p : q]$  and the reverse Kullback-Leibler divergence  $D_{\text{KL}}[q : p]$  are finite.*

## PROOF

Remember KLD=Bregman divergence between densities of a **regular (LR)EF**

$$D_{\text{KL}}[p : q] = (B_F)^*(\alpha_p : \alpha_q) = B_{F_{pq}}(\alpha_q : \alpha_p) = B_{F_{pq}}(0 : 1)$$

Scalar Bregman divergence  $B_{F_{pq}} : \Theta \times \text{ri}(\Theta) \rightarrow [0, \infty)$

$$B_{F_{pq}}(\alpha_1 : \alpha_2) = F_{pq}(\alpha_1) - F_{pq}(\alpha_2) - (\alpha_1 - \alpha_2)F'_{pq}(\alpha_2).$$

$$F_{pq}(0) = F_{pq}(1) = 0$$

$$D_{\text{KL}}[p : q] = B_{F_{pq}}(\alpha_q : \alpha_p) = B_{F_{pq}}(0 : 1) = F'_{pq}(1) < \infty$$

idem for

$$D_{\text{KL}}[q : p] = B_{F_{pq}}(\alpha_p : \alpha_q) = B_{F_{pq}}(1 : 0) = -F'_{pq}(0) < \infty$$

# Chernoff information (for densities of a LREF)

- Proposition:  $D_C[p : q] = D_{\text{KL}}[(pq)_{\alpha^*}^G : p] = D_{\text{KL}}[(pq)_{\alpha^*}^G : q] = D_{B,\alpha^*}[p : q]$

## PROOF

First, **skew Bhattacharyya distance = skew Jensen divergence**

$$D_{B,\alpha}[p : q] := -\log \rho_\alpha[P : Q] \longrightarrow D_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J_{F,\alpha}(\theta_1 : \theta_2).$$

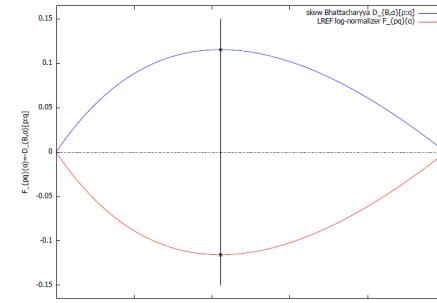
$$J_{F,\alpha}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1 - \alpha)F(\theta_2) - F(\alpha\theta_1 + (1 - \alpha)\theta_2).$$

Thus we have:

$$\begin{aligned} D_{B,\alpha}((pq)_{\alpha_1}^G : (pq)_{\alpha_2}^G) &= J_{F_{pq},\alpha}(\alpha_1 : \alpha_2), \\ &= \alpha F_{pq}(\alpha_1) + (1 - \alpha)F_{pq}(\alpha_2) - F_{pq}(\alpha\alpha_1 + (1 - \alpha)\alpha_2) \end{aligned}$$

At the optimal value  $\alpha^*$ , we have  $F'_{pq}(\alpha^*) = 0$

- ①  $D_{\text{KL}}[(pq)_{\alpha^*}^G : p] = B_{F_{pq}}(1 : \alpha^*) = -F(\alpha^*)$
- ②  $D_{\text{KL}}[(pq)_{\alpha^*}^G : q] = B_{F_{pq}}(0 : \alpha^*) = -F(\alpha^*)$
- ③  $D_C[p : q] = -\log \rho_{\alpha^*}(p : q) = J_{F_{pq},\alpha^*}(1 : 0) = -F_{pq}(\alpha^*)$



# Jensen-Chernoff divergence

$$D_C[p : q] = D_{\text{KL}}[(pq)_{\alpha^*}^G : p] = D_{\text{KL}}[(pq)_{\alpha^*}^G : q]$$

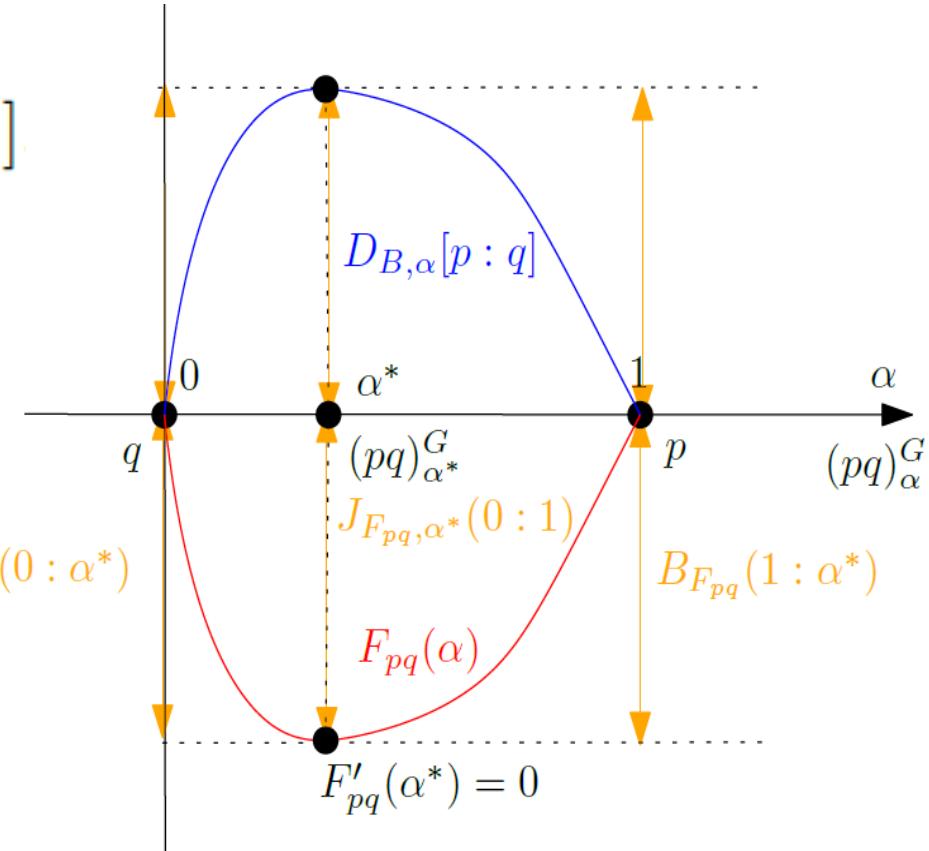
non-parametric arguments

$$\begin{aligned} D_C[p, q] &= B_{F_{pq}}(1 : \alpha^*) = B_{F_{pq}}(0 : \alpha^*) \\ &= J_{F_{pq}, \alpha^*}(0 : 1) \end{aligned}$$

scalar parametric arguments

In general, define **Jensen-Chernoff divergence**

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



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

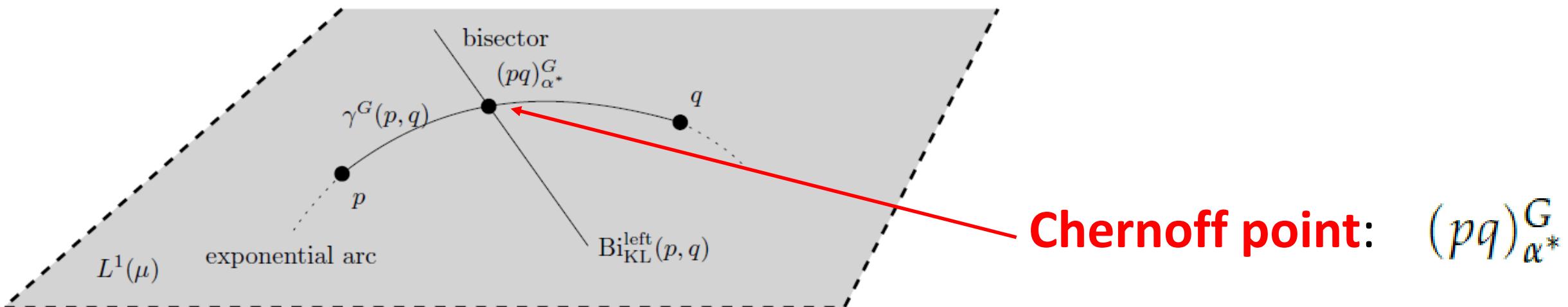
**Proposition** (Geometric characterization of the Chernoff information). *On the vector space  $L^1(\mu)$ , the Chernoff information distribution is the unique distribution*

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

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

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

2209.07481



# Special case of LREF: p,q are densities of a same EF!

EF includes Gaussians, Beta, Dirichlet, Wishart, etc.

$$\mathcal{E} = \left\{ P_\lambda : \frac{dP_\lambda}{d\mu} = p_\lambda(x) = \underline{\exp(\theta(\lambda)^\top t(x) - F(\theta(\lambda)))}, \quad \lambda \in \Lambda \right\}$$

$$\begin{aligned} p_{\theta_1}(x)^\alpha p_{\theta_2}(x)^{1-\alpha} &\propto \exp(\langle \alpha\theta_1 + (1-\alpha)\theta_2, t(x) \rangle - \alpha F(\theta_1) - (1-\alpha)F(\theta_2)), \\ &= p_{\alpha\theta_1+(1-\alpha)\theta_2}(x) \exp(F(\alpha\theta_1 + (1-\alpha)\theta_2) - \alpha F(\theta_1) - (1-\alpha)F(\theta_2)) \\ &= \underline{p_{\alpha\theta_1+(1-\alpha)\theta_2}(x) \exp(-J_{F,\alpha}(\theta_1 : \theta_2))}, \end{aligned}$$

→  $(p_{\theta_1} p_{\theta_2})_\alpha^G = p_{\alpha\theta_1+(1-\alpha)\theta_2}$        $D_{\text{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1)$ .

$\text{OC}_{\text{EF}} : \quad B_F(\theta_1 : \theta_{\alpha^*}) = B_F(\theta_2 : \theta_{\alpha^*})$

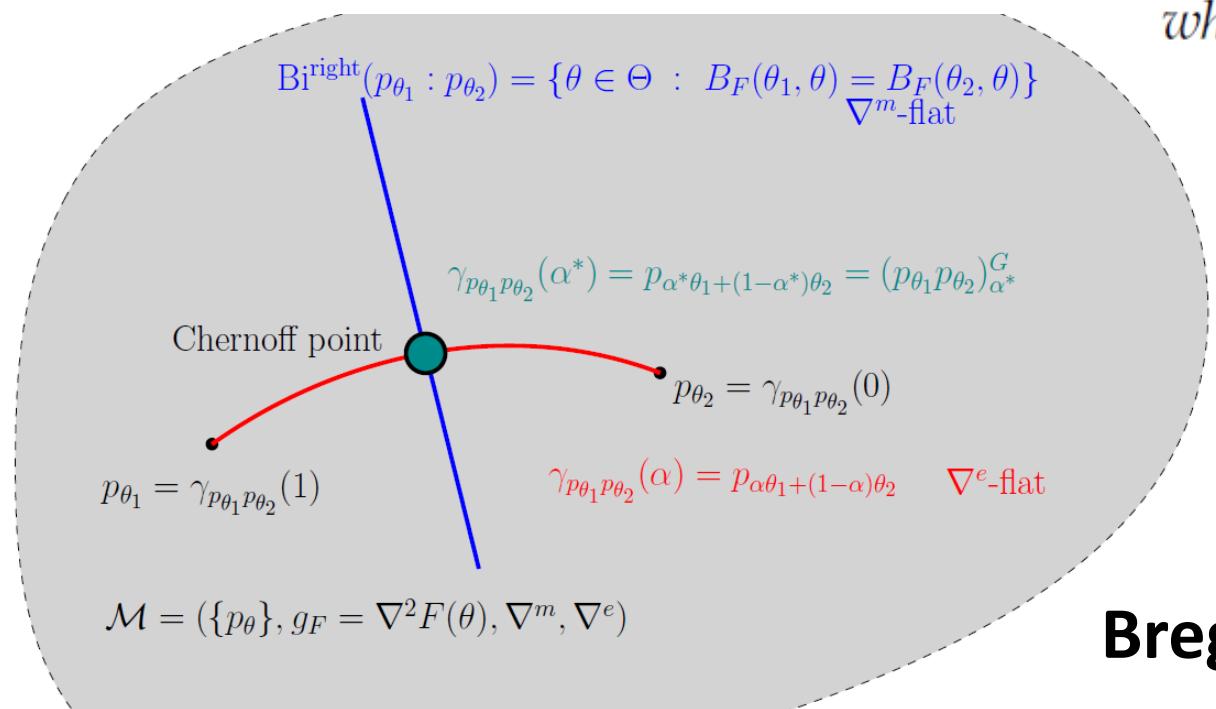
**Proposition** Let  $p_{\lambda_1}$  and  $p_{\lambda_2}$  be two densities of a regular exponential family  $\mathcal{E}$  with natural parameter  $\theta(\lambda)$  and log-normalizer  $F(\theta)$ . Then the Chernoff information is

$$D_C[p_{\lambda_1} : p_{\lambda_2}] = J_{F,\alpha^*}(\theta(\lambda_1) : \theta(\lambda_2)) = B_F(\theta_1 : \theta_{\alpha^*}) = B_F(\theta_2 : \theta_{\alpha^*}),$$

where  $\theta_1 = \theta(\lambda_1)$ ,  $\theta_2 = \theta(\lambda_2)$ , and the optimal skewing parameter  $\alpha^*$  is unique and satisfies the following optimality condition:

$$\text{OC}_{\text{EF}} : (\theta_2 - \theta_1)^\top \eta_{\alpha^*} = F(\theta_2) - F(\theta_1),$$

where  $\eta_{\alpha^*} = \nabla F(\alpha^*\theta_1 + (1 - \alpha^*)\theta_2) = E_{p_{\alpha^*\theta_1 + (1 - \alpha^*)\theta_2}}[t(x)]$ .



**Bregman manifold (= global Hessian manifold)**

# Interpreting the uniqueness of Chernoff exponent from pure information geometry point of view

- Since the Chernoff point is unique, we can also interpret more generally this property in a general dually flat space (not necessarily an EF) as known as a **Bregman manifold**

**Proposition** Let  $(\mathcal{M}, g, \nabla, \nabla^*)$  be a dually flat space with corresponding canonical divergence a Bregman divergence  $B_F$ . Let  $\gamma_{pq}^e(\alpha)$  and  $\gamma_{pq}^m(\alpha)$  be a  $e$ -geodesic and  $m$ -geodesic passing through the points  $p$  and  $q$  of  $\mathcal{M}$ , respectively. Let  $\text{Bi}^m(p, q)$  and  $\text{Bi}^e(p, q)$  be the right-sided  $\nabla^m$ -flat and left-sided  $\nabla^e$ -flat Bregman bisectors, respectively. Then the intersection of  $\gamma_{pq}^e(\alpha)$  with  $\text{Bi}^m(p, q)$  and the intersection of  $\gamma_{pq}^m(\alpha)$  with  $\text{Bi}^e(p, q)$  are unique. The point  $\gamma_{pq}^e(\alpha) \cap \text{Bi}^m(p, q)$  is called the Chernoff point and the point  $\gamma_{pq}^m(\alpha) \cap \text{Bi}^e(p, q)$  is termed the reverse or dual Chernoff point.

"On geodesic triangles with right angles in a dually flat space."  
Progress in Information Geometry. Springer, 2021. 153-190.

# Duo Bregman pseudo-divergences: Applications to the KL divergence between truncated densities

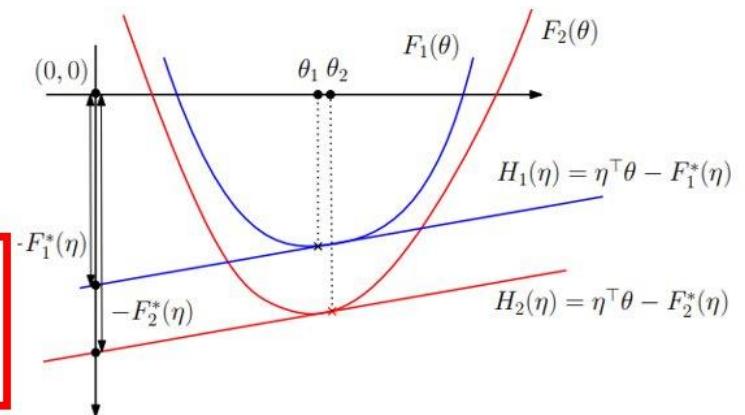
# Legendre transformation reverses majorization order

Legendre-Fenchel transformation:  $F^*(\eta) := \sup_{\theta \in \Theta} \{\eta^\top \theta - F(\theta)\}$

F Legendre-type function, Moreau **biconjugation theorem**:  $(F^*)^* = F$   
proper+lower semi-continuous+convex

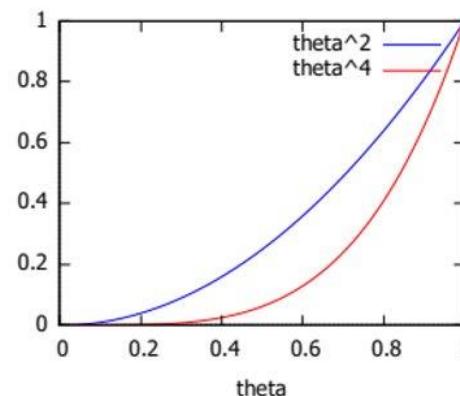
Legendre-Fenchel transform **reverses ordering**:

$$\forall \theta \in \Theta, \quad F_1(\theta) \geq F_2(\theta) \Leftrightarrow \forall \eta \in H, \quad F_1^*(\eta) \leq F_2^*(\eta)$$

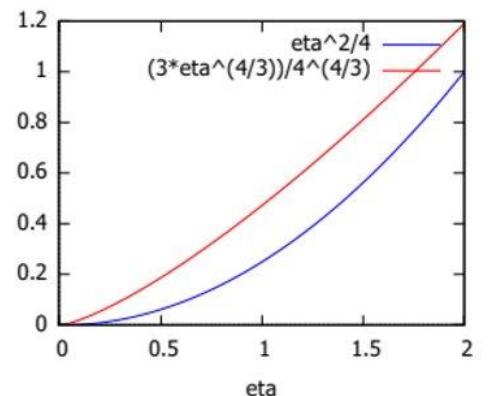


Proof:

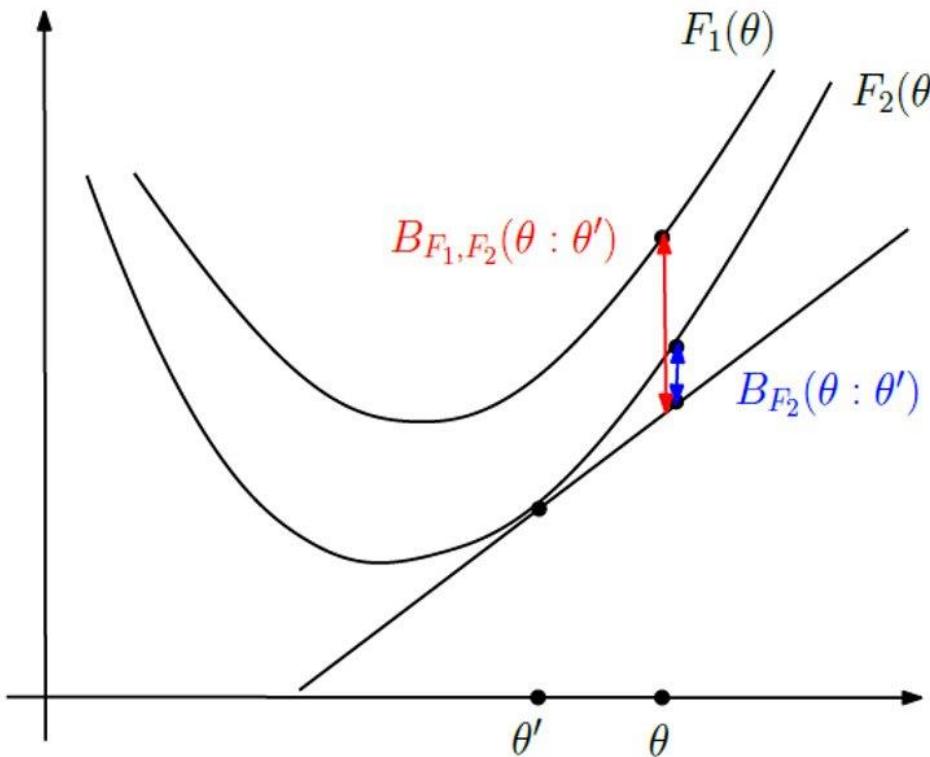
$$\begin{aligned} F_1^*(\eta) &:= \sup_{\theta \in \Theta} \{\eta^\top \theta - F_1(\theta)\}, \\ &= \eta^\top \theta_1 - F_1(\theta_1) \quad (\text{with } \eta = \nabla F_1(\theta_1)) \\ &\leq \eta^\top \theta_1 - F_2(\theta_1), \\ &\leq \sup_{\theta \in \Theta} \{\eta^\top \theta - F_2(\theta)\} =: F_2^*(\eta). \end{aligned}$$



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



Conjugate functions  $F_1^*(\eta) \leq F_2^*(\eta)$



### Duo Bregman divergence

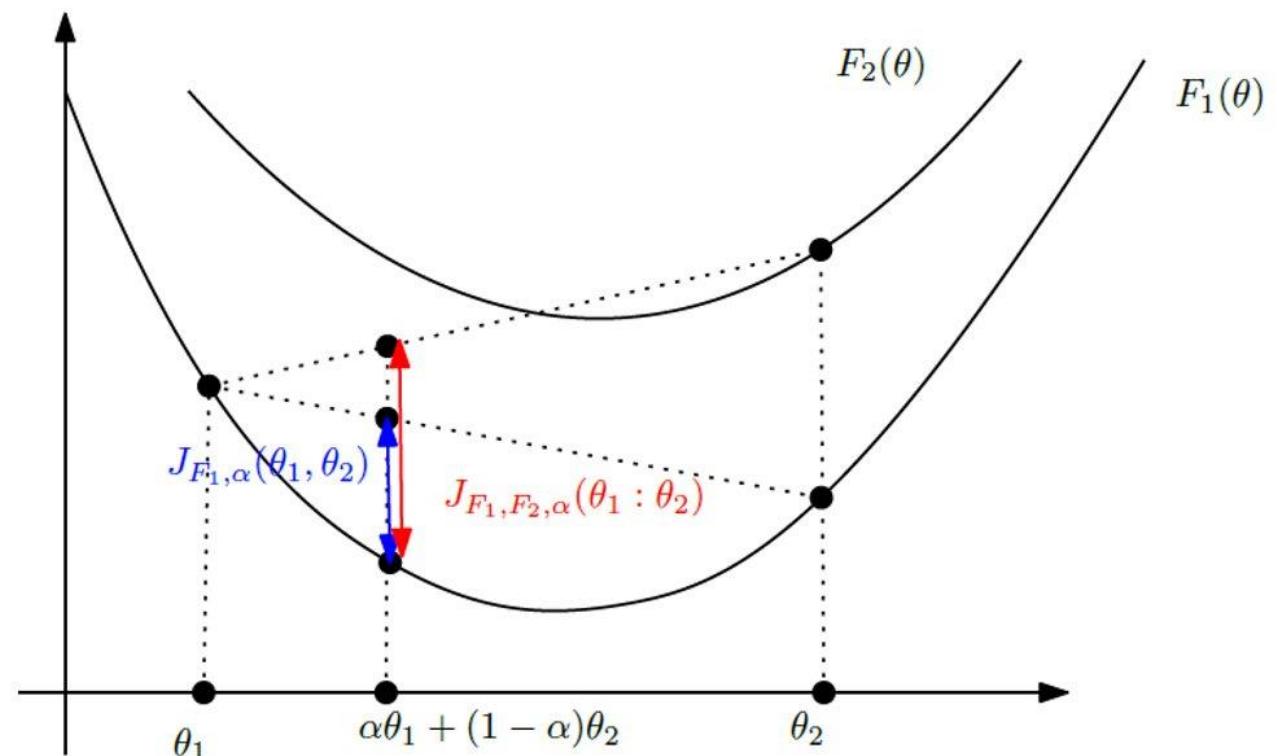
$$B_{F_1,F_2}(\theta : \theta') = F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta')$$

### Duo Fenchel-Young divergence

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

### Relationship with truncated exponential families with nested supports:

$$D_{\text{KL}}[p_{\theta_1} : q_{\theta_2}] = Y_{F_2,F_1^*}(\theta_2 : \eta_1) = B_{F_2,F_1}(\theta_2 : \theta_1)$$



### Duo Jensen divergence

$$J_{F_1,F_2,\alpha}(\theta_1 : \theta_2) = \alpha F_1(\theta_1) + (1 - \alpha) F_2(\theta_2) - F_1(\alpha \theta_1 + (1 - \alpha) \theta_2).$$

$$D_{\text{Bhat},\alpha}[p : q] := -\log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} d\mu(x)$$

$$D_{\text{Bhat},\alpha}[p_{\theta_1} : q_{\theta_2}] = J_{F_1,F_2,\alpha}(\theta_1 : \theta_2).$$

# Kullback-Leibler divergence between exponential family densities

$$D_{\text{KL}}[P : Q] = \int_{\mathcal{X}} \log \frac{dP}{dQ} dP = E_P \left[ \log \frac{dP}{dQ} \right].$$

$$\begin{aligned} B_F(\theta_1 : \theta_2) &:= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \\ Y_{F,F^*}(\theta_1, \eta_2) &:= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 \end{aligned} \quad \xrightarrow{\text{Duo}} \quad \begin{aligned} B_{F_1, F_2}(\theta : \theta') &:= Y_{F_1, F_2^*}(\theta, \eta') = F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta') \\ Y_{F_1, F_2^*}(\theta, \eta') &:= F_1(\theta) + F_2^*(\eta') - \theta^\top \eta'. \end{aligned}$$

- **Same exponential family:** KLD = reverse Bregman divergence or reverse Fenchel-Young divergence

$$D_{\text{KL}}[P_{\theta_1} : P_{\theta_2}] = Y_{F,F^*}(\theta_2 : \eta_1) = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2) = Y_{F^*,F}(\eta_1 : \eta_2).$$

- **Different exponential families** (mutually absolutely continuous):

$$D_{\text{KL}}[P_\theta : Q_{\theta'}] = F_Q(\theta') - F_P(\theta) + \theta^\top E_{P_\theta}[t_P(x)] - \theta'^\top E_{P_\theta}[t_Q(x)].$$

- **Same truncated exponential family:** reverse duo Bregman divergence or reverse duo Fenchel-Young divergence (nested supports)

$$D_{\text{KL}}[p_{\theta_1} : q_{\theta_2}] = Y_{F_2, F_1^*}(\theta_2 : \eta_1) = B_{F_2, F_1}(\theta_2 : \theta_1) = B_{F_1^*, F_2^*}(\eta_1 : \eta_2) = Y_{F_1^*, F_2}(\eta_1 : \theta_2).$$

# KL divergence between truncated normal densities

PDF of truncated normal on  $(a, b)$ :

$$p_{m,s}^{a,b}(x) = \frac{1}{\sqrt{2\pi}s (\Phi_{m,s}(b) - \Phi_{m,s}(a))} \exp\left(-\frac{(x-m)^2}{2s^2}\right)$$

$$\Phi_{m,s}(x) = \frac{1}{2} \left( 1 + \operatorname{erf}\left(\frac{x-m}{\sqrt{2}s}\right) \right), \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Truncated normal PDFs form an exponential family with log-normalizer :

$$F_{a,b}(m, s) = \frac{m^2}{2s^2} + \frac{1}{2} \log 2\pi s^2 + \log (\Phi_{m,s}(b) - \Phi_{m,s}(a))$$

Moment parameters and mean & variance:

$$\mu(m, s; a, b) = m - s \frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}, \quad \phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$\eta_1(m, s; a, b) = E_{p_{m,s}^{a,b}}[x] = \mu(m, s; a, b),$$

$$\eta_2(m, s; a, b) = E_{p_{m,s}^{a,b}}[x^2] = \sigma^2(m, s; a, b) + \mu^2(m, s; a, b).$$

$$\sigma^2(m, s; a, b) = s^2 \left( 1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left( \frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right),$$

**Kullback-Leibler divergence between nested truncated normal distributions:**

$$D_{\text{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = \frac{m_2}{2s_2^2} - \frac{m_1}{2s_1^2} + \log \frac{Z_{a_2,b_2}(m_2, s_2)}{Z_{a_1,b_1}(m_1, s_1)} - \left( \frac{m_2}{s_2^2} - \frac{m_1}{s_1^2} \right) \eta_1(m_1, s_1; a_1, b_1)$$

$$- \left( \frac{1}{2s_1^2} - \frac{1}{2s_2^2} \right) \eta_2(m_1, s_1; a_1, b_1) \quad \text{if nested distributions } (a_1, b_1) \subseteq (a_2, b_2)$$

$$D_{\text{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = +\infty, (a_1, b_1) \not\subseteq (a_2, b_2) \quad \text{otherwise}$$

# Paper references

- "An elementary introduction to information geometry." *Entropy* 22.10 (2020): 1100.
- "Beyond scalar quasi-arithmetic means: Quasi-arithmetic averages and quasi-arithmetic mixtures in information geometry." *arXiv preprint arXiv:2301.10980* (2023).
- "Revisiting Chernoff Information with Likelihood Ratio Exponential Families." *Entropy* 24.10 (2022): 1400.
- "Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences." *Entropy* 24.3 (2022): 421.