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

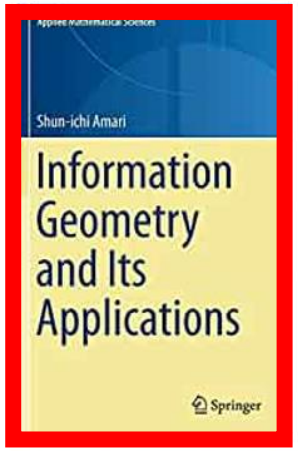
Information geometry: Rationale and scope

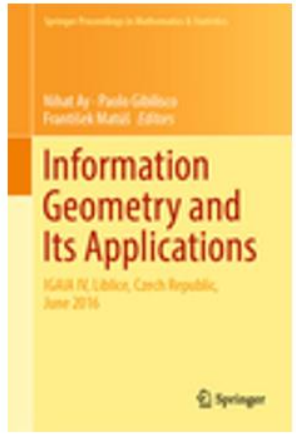
- More generally, **Information Geometry = Dual geometry of models**:

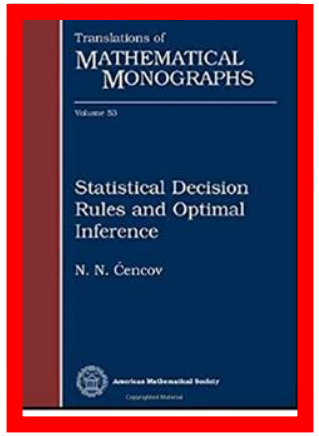
quantum information geometry of quantum models (space of density matrices with unit trace for modeling quantum states)

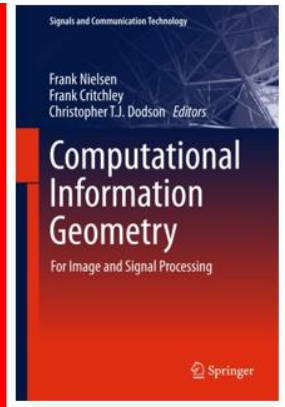
- **Geometric objects** are defined **globally** and can be expressed **locally in *any* convenient coordinate systems** to ease computations, change of coordinate systems (atlas of manifolds)

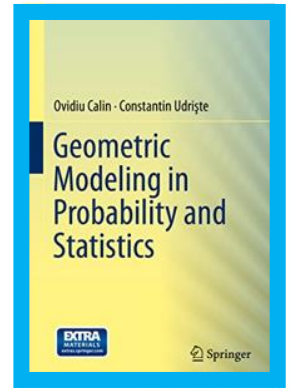
- Because the information-geometric structures are **purely geometric** (i.e., there is no attached meanings to objects), information-geometric structures can also be used in **non-statistical contexts** too, like *mathematical programming (e.g., IG of barrier functions)*

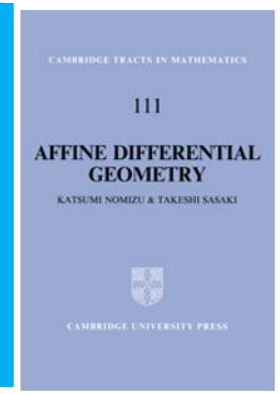

Shun-ichi Amari
Information Geometry and Its Applications
Springer

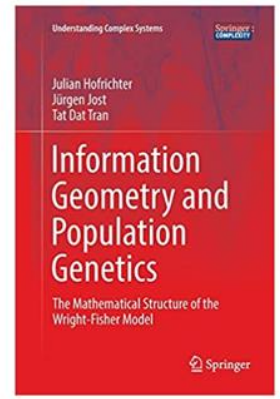

Information Geometry and Its Applications

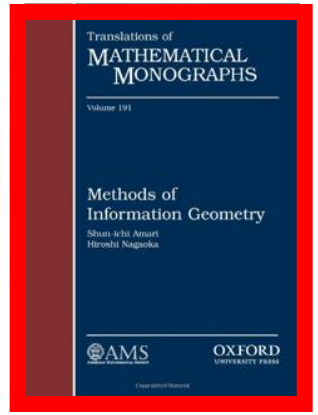

Translations of
MATHEMATICAL MONOGRAPHS
Statistical Decision Rules and Optimal Inference
N. N. Čencov

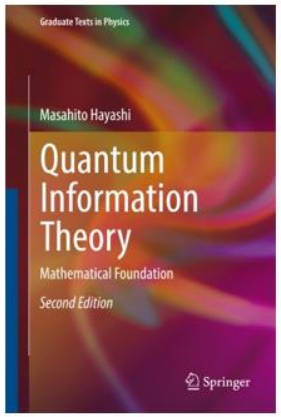

Signals and Communication Technology
Frank Nielsen
Frank Critchley
Christopher T.J. Dodson Editors
Computational Information Geometry
For Image and Signal Processing
Springer


Ovidiu Calin · Constantin Udrişte
Geometric Modeling in Probability and Statistics
EXTRA MATERIALS
Springer

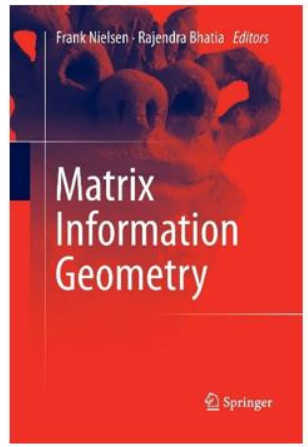

CAMBRIDGE TRACTS IN MATHEMATICS
111
AFFINE DIFFERENTIAL GEOMETRY
KATSUMI NOMIZU & TAKESHI SASAKI
CAMBRIDGE UNIVERSITY PRESS

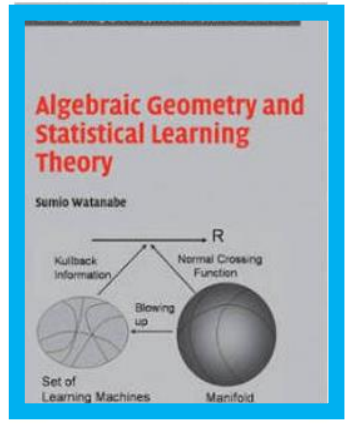

Understanding Complex Systems
Julian Hofrichter
Jürgen Jost
Tat Dat Tran
Information Geometry and Population Genetics
The Mathematical Structure of the Wright-Fisher Model
Springer

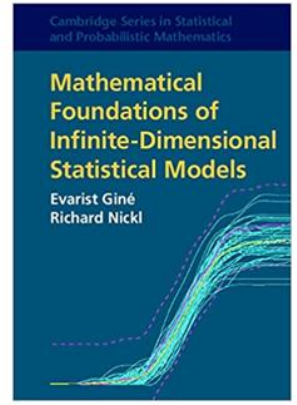

Translations of
MATHEMATICAL MONOGRAPHS
Volume 191
Methods of Information Geometry
Shun-ichi Amari
Hiroshi Nagaoka
AMS
OXFORD UNIVERSITY PRESS

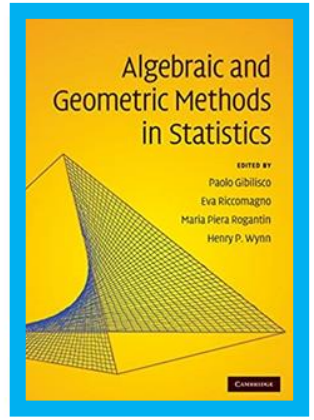

Graduate Texts in Physics
Masahito Hayashi
Quantum Information Theory
Mathematical Foundation
Second Edition
Springer

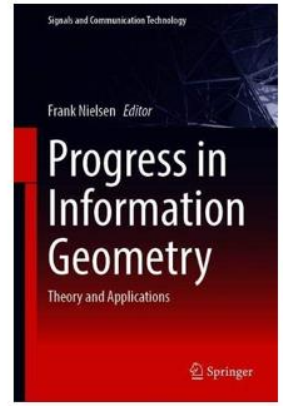

HIROHIKO SHIMA
THE GEOMETRY OF
HESSIAN STRUCTURES

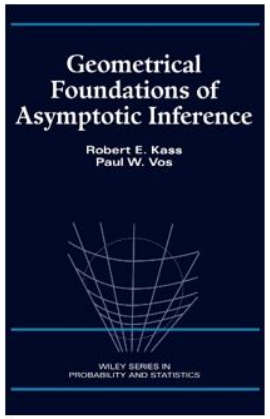

Frank Nielsen · Rajendra Bhatia Editors
Matrix Information Geometry
Springer

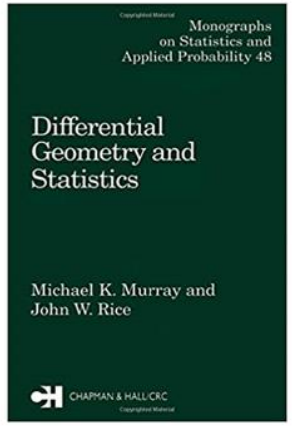

Algebraic Geometry and Statistical Learning Theory
Sumio Watanabe
R
Kullback Information
Normal Crossing Function
Blowing up
Set of Learning Machines
Manifold

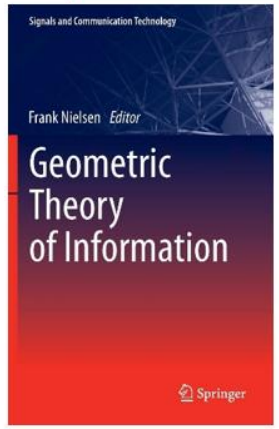

Cambridge Series in Statistical and Probabilistic Mathematics
Mathematical Foundations of Infinite-Dimensional Statistical Models
Evarist Giné
Richard Nickl

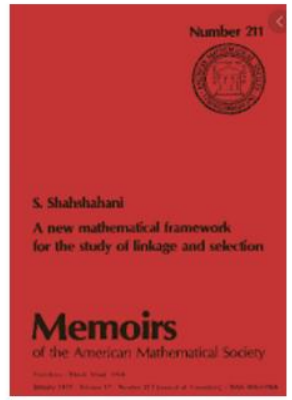

Algebraic and Geometric Methods in Statistics
EDITED BY
Paolo Gibilisco
Eva Riccomagno
Maria Piera Rogantin
Henry P. Wynn

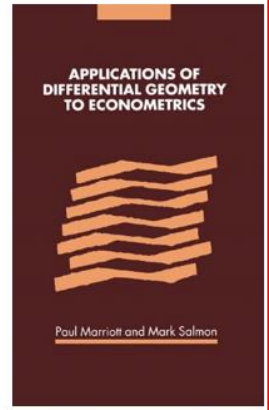

Signals and Communication Technology
Frank Nielsen Editor
Progress in Information Geometry
Theory and Applications
Springer

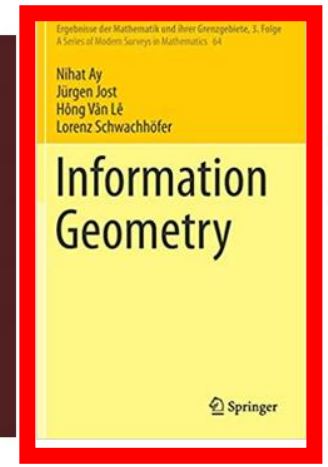

Geometrical Foundations of Asymptotic Inference
Robert E. Kass
Paul W. Vos
WILEY SERIES IN PROBABILITY AND STATISTICS

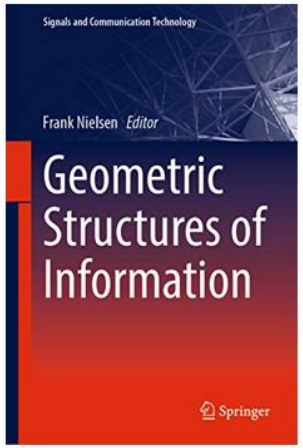

Monographs on Statistics and Applied Probability 48
Differential Geometry and Statistics
Michael K. Murray and John W. Rice
CHAPMAN & HALL/CRC

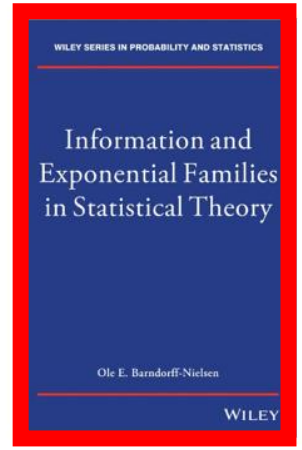

Signals and Communication Technology
Frank Nielsen Editor
Geometric Theory of Information
Springer


Number 211
S. Shahshahani
A new mathematical framework for the study of linkage and selection
Memoirs
of the American Mathematical Society


APPLICATIONS OF DIFFERENTIAL GEOMETRY TO ECONOMETRICS
Paul Marriott and Mark Salmon


Nihat Ay
Jürgen Jost
Hông Vân Lê
Lorenz Schwachhöfer
Information Geometry
Springer


Signals and Communication Technology
Frank Nielsen Editor
Geometric Structures of Information
Springer

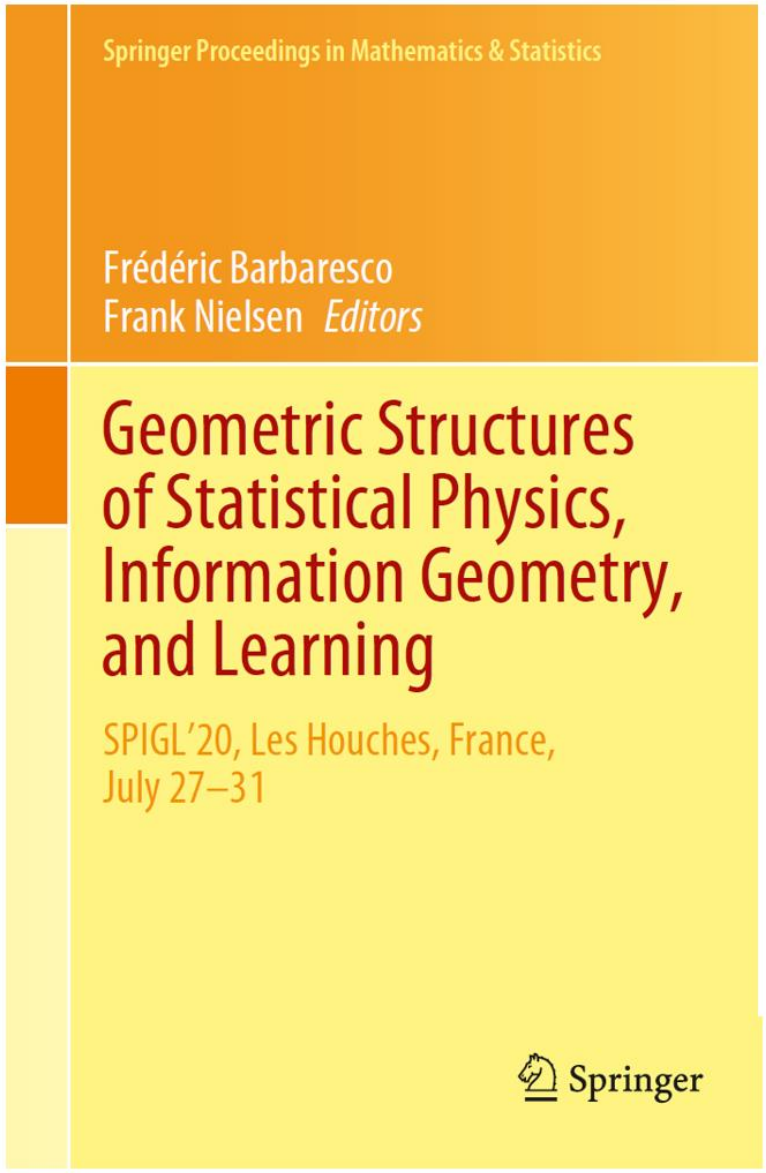

WILEY SERIES IN PROBABILITY AND STATISTICS
Information and Exponential Families in Statistical Theory
Ole E. Barndorff-Nielsen
WILEY

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/



Contents

Many slide decks online : https://franknielsen.github.io/SPIG-LesHouches2020/

GSI: Biannual conference since 2013

https://conference-gsi.org/

Include 500+ GSI video talks: franknielsen.github.io/GSI/

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

Information geometry:

Geometry of dual structures

Applications:

- Geometry of statistical models
- Geometry of divergences

Some resources

Open Access | Review

An Elementary Introduction to Information Geometry

by Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan

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

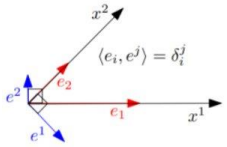
Received: 6 September 2020 / Revised: 25 September 2020 / Accepted: 27 September 2020 / Published: 29 September 2020

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

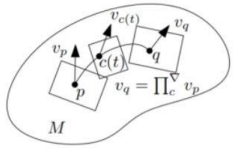
Download | Browse Figures | Versions Notes

Tutorial 60+ pages

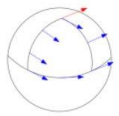
https://www.mdpi.com/1099-4300/22/10/1100



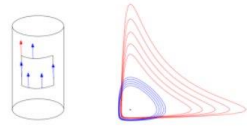
Reciprocal basis



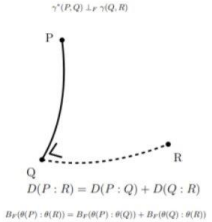
Dual parallel transport



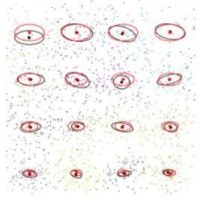
Loop PT and curvature



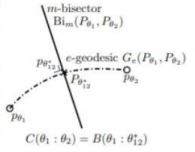
Bregman spheres



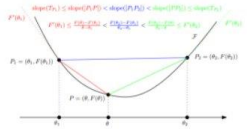
Pythagoras theorem (Dually flat space)



Cramér–Rao lower bound Fisher efficiency



e/m-geodesic e/m flat



chordal slope & Bregman divergence

The Many Faces of Information Geometry

Frank Nielsen

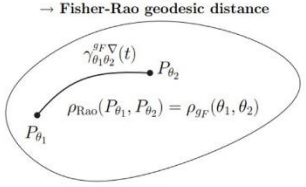
Information geometry [Ama16, AJLS17, Ama21] aims at unravelling the geometric structures of families of probability distributions and at studying their uses in information sciences. Information sciences is an umbrella term regrouping statistics, information theory, signal processing, machine learning and AI, etc. Information geometry was born independently from econometrician H. Hotelling (1930) and statistician C. R. Rao (1945) from the mathematical curiosity of considering a parametric family of μ, usually chosen as the Lebesgue mesure μ_L or the counting measure μ_c), and consider a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability distributions, all dominated by μ. Let $p_\theta(x) := \frac{dP_\theta(x)}{d\mu}$ denote the Radon-Nikodym derivative, the probability density function of random variable $X \sim p_\theta$. By definition, the Fisher Riemannian metric g_F expressed in the θ-coordinate system is the Fisher information matrix (FIM) of the random variable X: $[g_F]_\theta := I_X(\theta)$

Short overview 10 pages

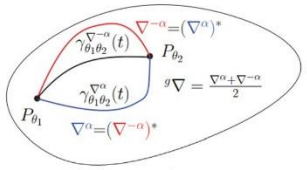
https://www.ams.org/journals/notices/202201/rnoti-p36.pdf

The Many Faces of Information Geometry

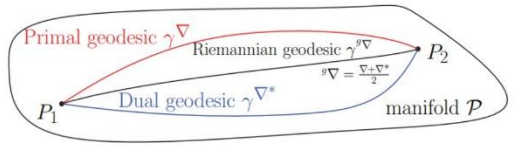
Fisher-Rao geometry
→ Fisher-Rao geodesic distance

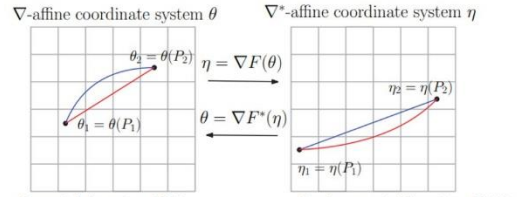


versus



Dual α-geometry
→ No default divergence





Potential function $F(\theta)$ ⟷ Dual potential function $F^*(\eta)$

Legendre-Fenchel transform

40 min. video introduction

https://www.youtube.com/watch?v=w6r_jsEBlgU

YouTube

Introduction to Information Geometry

Frank NIELSEN

July 2022

Sony CSL

https://franknielsen.github.io/IG/index.html

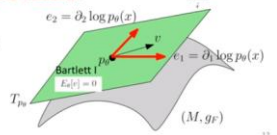
"Introduction to Information Geometry" by Frank Nielsen

Frank Nielsen 628 subscribers | Analytics | Edit video | 295 | Share | Download

Tangent plane representation for a manifold induced by a statistical model: Reinterpret the inner product

- On a tangent plane, we can choose any arbitrary basis to express vectors
- Inner product of two vectors is independent of the choice of basis: the component vectors depend on the basis but the vectors are geometric objects
- Express a vector v by a **representation** v(x)
- Basis vectors of T_θ can be chosen as the **score vectors**: $B = \{e_1 = \partial_1 l_x(\theta), \ldots, e_D = \partial_D l_x(\theta)\}$
- The inner product can be reinterpreted as: $g_F(u,v) = E_\theta[u(x)v(x)] = \mathrm{Cov}(u(x), v(x))$

$g_F(\partial_i, \partial_j) = E_\theta[\partial_i l_x(\theta) \partial_j l_x(\theta)]$ (Expectation)



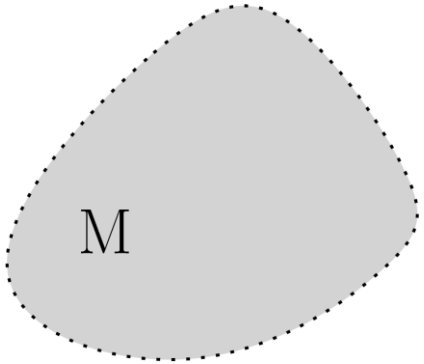
"Introduction to Information Geometry" by Frank Nielsen

Frank Nielsen 628 subscribers | Analytics | Edit video | 295 | Share | Download

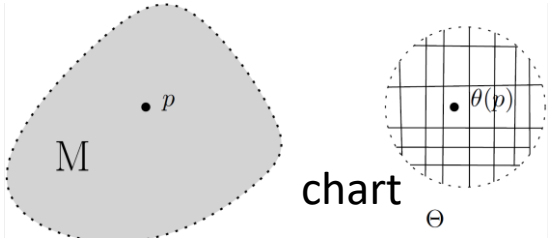
Build your own information geometry in three steps

Choose

① manifold M

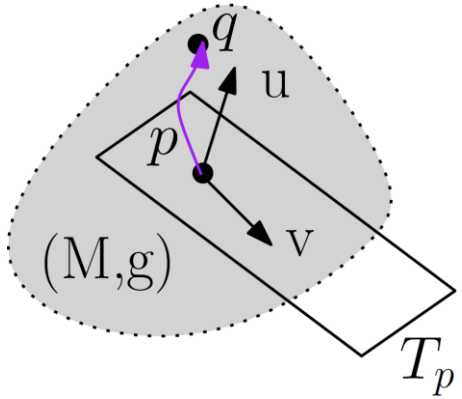


Examples:
Gaussians
SPD cone
Probability simplex



Concepts:
local coordinates
locally Euclidean

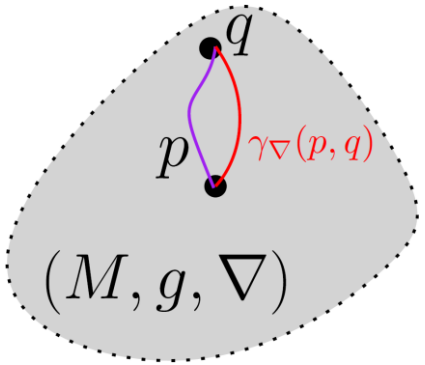
② metric tensor g



Examples:
Fisher information metric
metric g^D from divergence
trace metric

Concepts:
vector length
vector orthogonality
Riemannian geodesic
Riemannian distance
Levi-Civita connection ∇^g

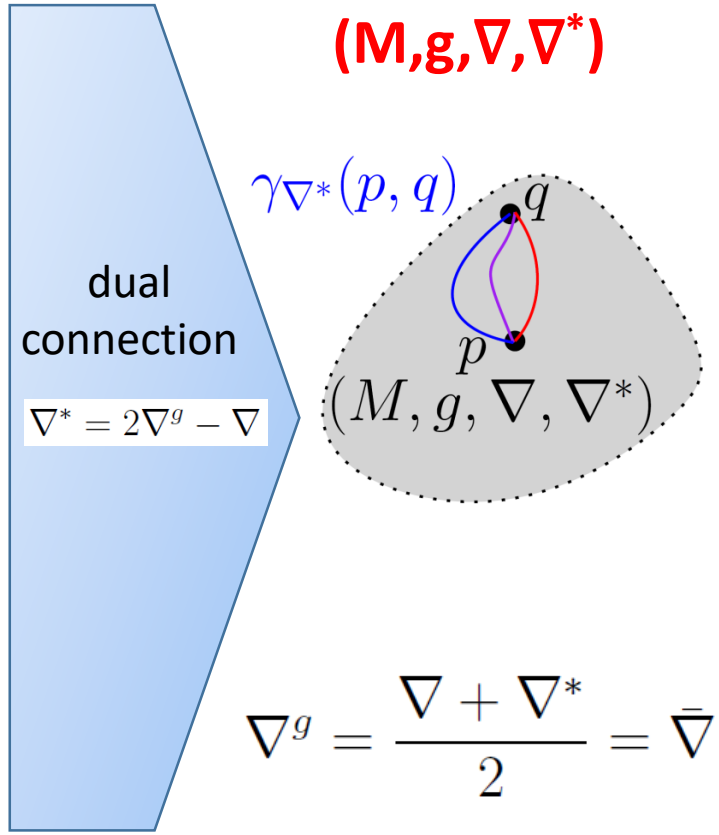
③ affine connection ∇



Examples:
exponential connection
mixture connection
metric connection ∇^g
divergence connection ∇^D
α-connection

Concepts:
covariant derivative ∇
∇-geodesic
∇-parallel transport
curvature

Get dual IG manifold (M,g,∇,∇^*)



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

Concepts:
dual connections coupled to metric g
dual parallel transport preserve metric g

From dual information geometry to ±α-geometry, α∈ℝ

Choose

① manifold M

② metric tensor g

③ affine connection ∇
by defining Christoffel symbols Γ^{∇}_{ijk}

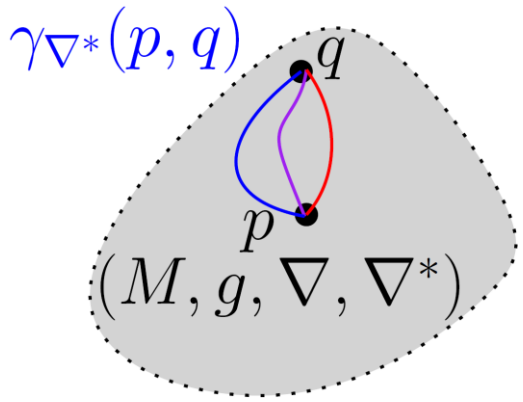
④ choose α

Examples:

Amari-Chentsov cubic tensor $T_{ijk}(\theta) = E[\partial_i l \partial_j l \partial_k l]$

Cubic tensor from divergence $T_{ijk}(\theta) = \partial_i \partial_j \partial_k F(\theta)$

Get dual IG manifold
(M,g,∇,∇^*)

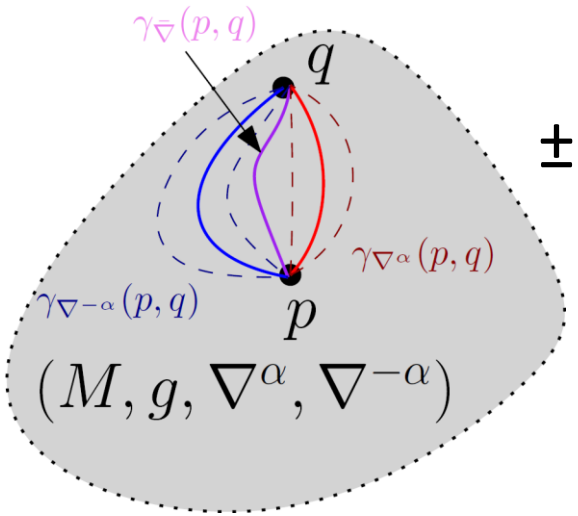


$$\nabla^g = \frac{\nabla + \nabla^*}{2} = \bar{\nabla}$$

Cubic tensor

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}$$
$$T_{ijk} = \nabla_i g_{jk}$$

Get a family of dual connections/IG
$(M,g,\nabla^\alpha,\nabla^{-\alpha})$



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

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

±α-geometry

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

0-geometry
= Riemannian geometry
with geodesic distance

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

Model Information Geometry (MIG)

- 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**

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

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

- Model is **regular** if partial derivatives of $l_\theta(x)$ smooth and Fisher metric is well-defined and positive-definite
- **Amari-Chentsov cubic tensor**: $C_{ijk} := E_\theta \left[\partial_i l \partial_j l \partial_k l\right]$ ⇒ $\left\{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\right\}_{\alpha \in \mathbb{R}}$
- **α-connections** $\nabla^\alpha = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m$

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

⇒

α=1 — **exponential connection**

$$ {}^{e}_{\mathcal{P}}\nabla := E_\theta \left[(\partial_i \partial_j l)(\partial_k l)\right], $$

$$ {}^{m}_{\mathcal{P}}\nabla := E_\theta \left[(\partial_i \partial_j l + \partial_i l \partial_j l)(\partial_k l)\right] $$

α=-1 — **mixture connection**

- **Fisher-Rao geometry when α=0,** get geodesic distance called **Rao distance**

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

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

Rao distance on the Fisher-Rao manifold

$$D_{\mathrm{Rao}}[p_{\theta_1}, p_{\theta_2}] = \rho_g(\theta_1, \theta_2) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\,\mathrm{d}t, \gamma(0) = \theta_1, \gamma(1) = \theta_2$$

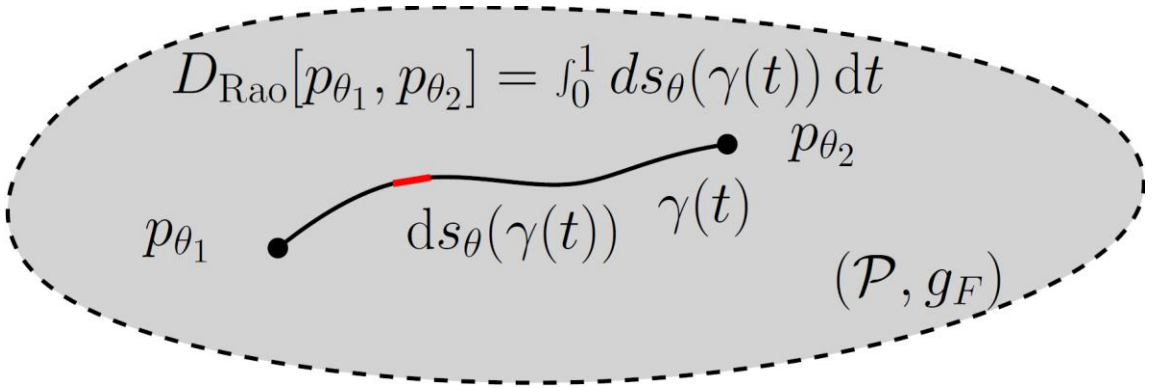
$$= \int_0^1 ds_\theta(\gamma(t))\,\mathrm{d}t$$

Here, γ is the Riemannian geodesic
(or add a minimizer on all paths γ)

Length element

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

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



In practice:

- Need to calculated 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 tasks.
- **Good news 2023**: closed-form geodesics with boundary conditions for **MultiVariate Normals**

Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures, ICML ws TAGML 2023

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

Divergence Information Geometry (DIG)

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

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

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

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

- 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:

$${}^{D}C_{ijk} = {}^{D^*}\Gamma_{ijk} - {}^{D}\Gamma_{ijk}$$

$${}^{D}\nabla^* = {}^{D^*}\nabla$$

$$\partial_{i,jk} f(x,y) = \frac{\partial}{\partial x^i} \frac{\partial^2}{\partial y^j \partial y^k} f(x,y)$$

$$\partial_{i,\cdot} f(x,y) = \frac{\partial}{\partial x^i} f(x,y), \quad \partial_{\cdot,j} f(x,y) = \frac{\partial}{\partial y^j} f(x,y), \partial_{ij,k} f(x,y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial}{\partial y^k} f(x,y)$$

Realizations of dual information geometry (stat mfd)

- Realize (M,g,∇,∇) 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,∇,∇) as a model information geometry $(M,g^F,\nabla^{-\alpha},\nabla^{\alpha})$

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

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

Equivalence: model α-IG ↔ divergence IG for f-divergences

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

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

$$I_f{}^*[p(x;\theta):p(x;\theta')] = I_f[p(x;\theta'):p(x;\theta)] = I_{f^\diamond}[p(x;\theta):p(x;\theta')]$$

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

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

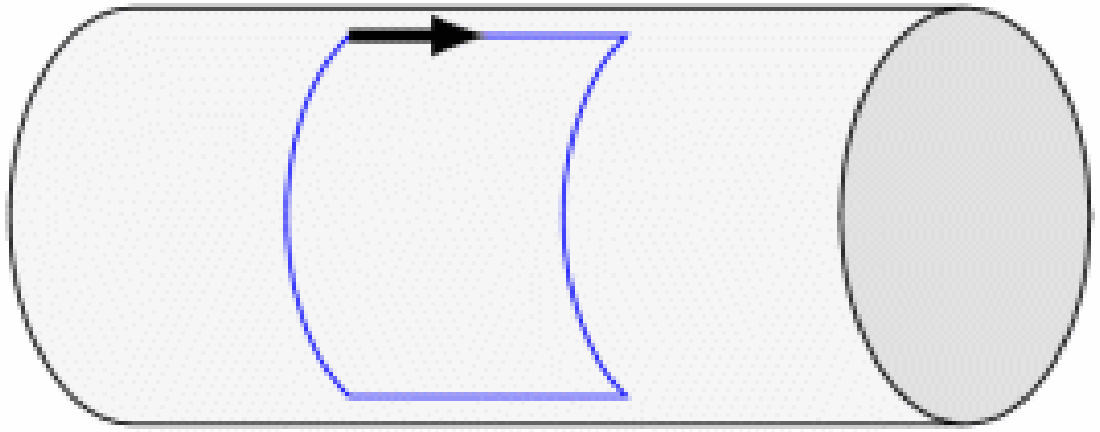
- Then **model α-geometry** for α=2 f'''(1)+3 coincide with **divergence IG**:

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

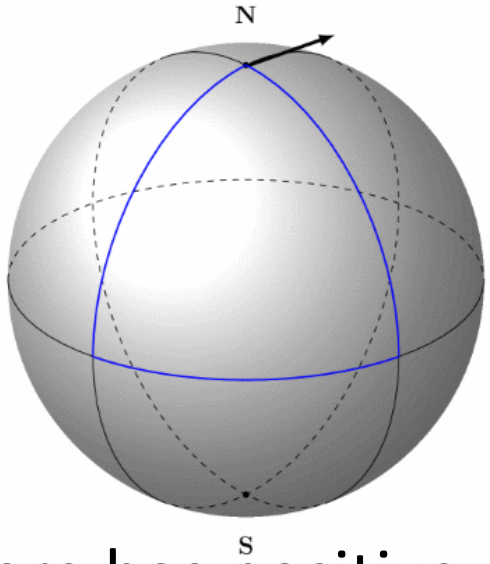
metric tensor g^D and cubic tensor T^D coincides with Fisher metric g^F and Amari-Chentsov tensor T

Curvature is associated to affine connection ∇

- 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

Dually flat spaces (M,g,∇,∇*)

- **Fundamental theorem of information geometry**: If torsion-free affine connection ∇ is of constant curvature κ, then curvature of dual torsion-free affine connection ∇* is also constant κ

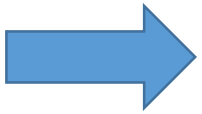
- Corollary: if ∇ is flat (κ=0) then ∇* is flat: **Dually flat space (M,g,∇,∇*)**

- A connection ∇ is flat if there exists a local coordinate system θ such that Γ(θ)=0

- In ∇-affine coordinate system θ(.), ∇-geodesics are visualized as line segments

Γ(θ)=0

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



geodesics=line segments in θ

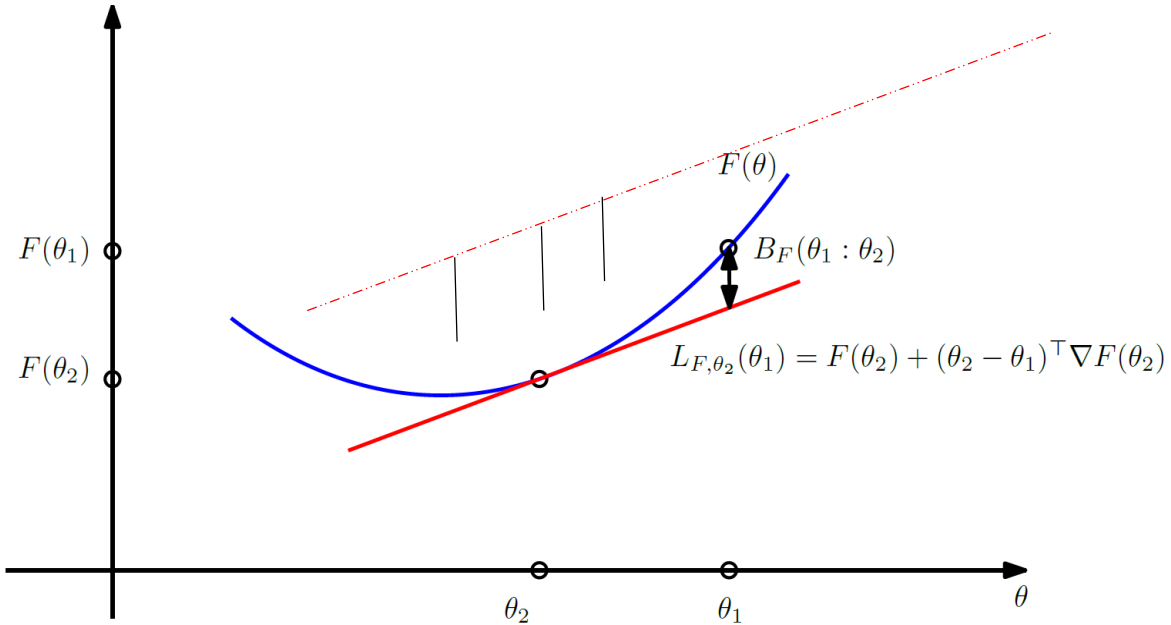
Canonical divergences of DFSs: Bregman divergences

- Dually flat structure (M,g,∇,∇*) can be realized by a Bregman divergence

$$(M, g, \nabla, \nabla^*) \longleftarrow (M, g^{B_F}, \nabla^{B_F}, \nabla^{B_{F^*}})$$

- Let F(θ) be a strictly convex and differentiable function defined on an open convex domain Θ
- Bregman divergence interpreted as the vertical gap between

point $(\theta_1, F(\theta_1))$ and the linear approximation of $F(\theta)$ at θ_2 evaluated at θ_1 :



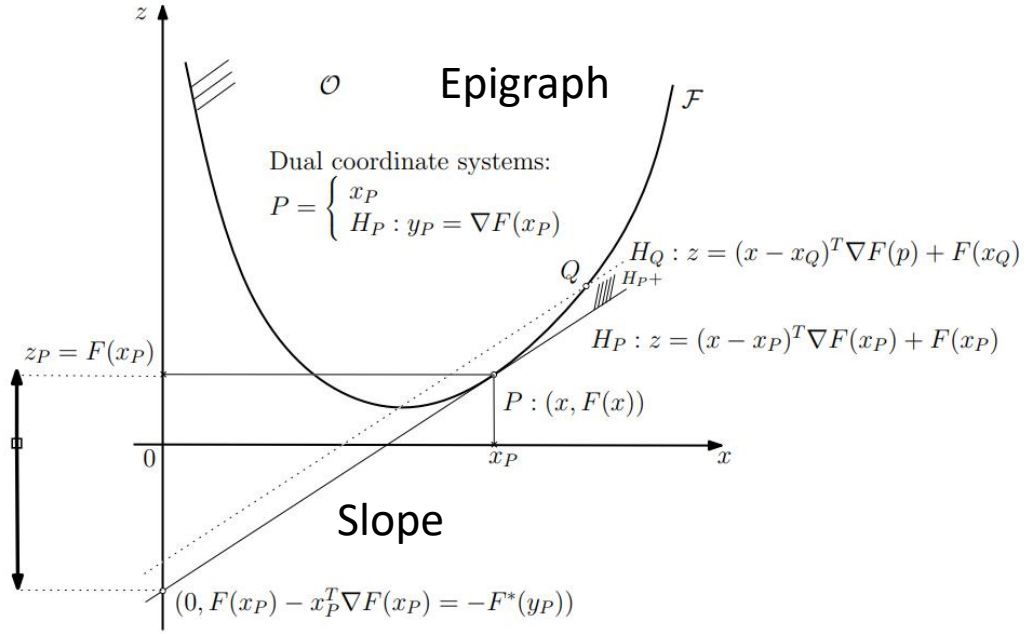
$$\begin{aligned} B_F(\theta_1 : \theta_2) &= F(\theta_1) - \underbrace{\left(F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2)\right)}_{L_{F,\theta_2}(\theta_1)} \\ &= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \end{aligned}$$

[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 $\theta = \nabla F^*(\eta) \iff \eta = \nabla F(\theta)$
- Convex conjugate of F is $F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta))$
- **Fenchel-Young inequality** :

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

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 θ et η to express a Bregman divergence as $B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2)$

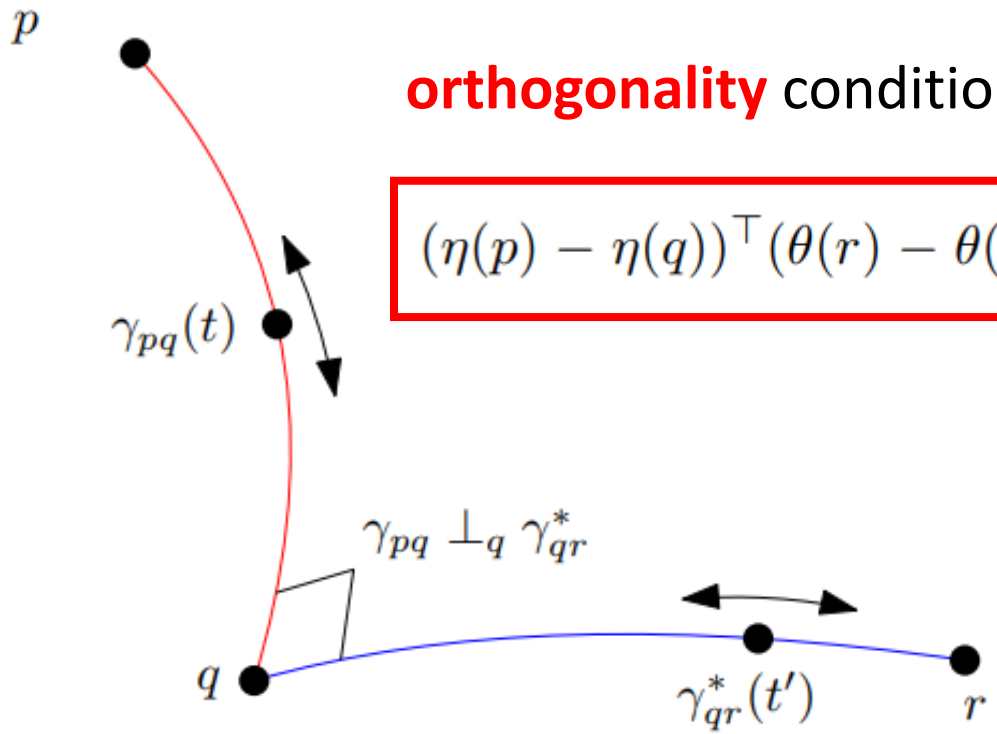
$$Y_{F,F^*}(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 = Y_{F^*,F}(\eta_2, \theta_1)$$

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



orthogonality condition:

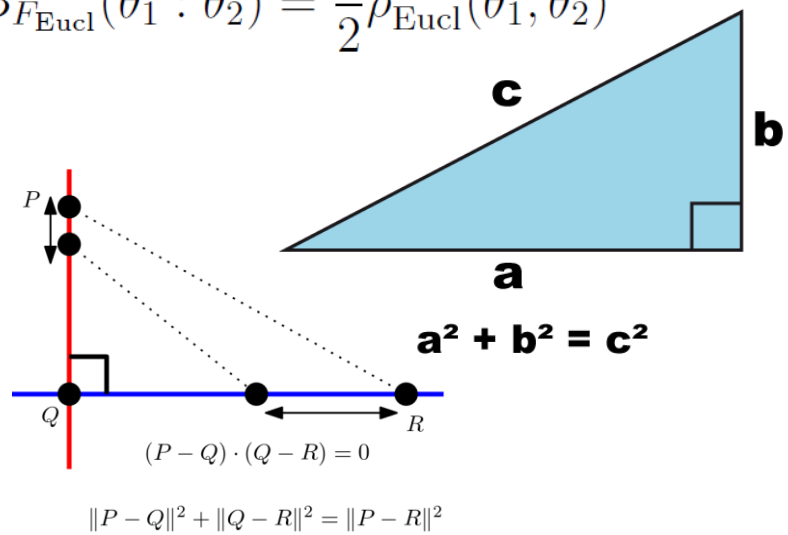
$$(\eta(p) - \eta(q))^\top (\theta(r) - \theta(q)) = 0$$

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

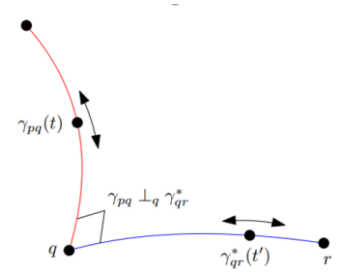
Pythagoras' theorem in the Euclidian geometry (Self-dual)

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

$$B_{F_{\mathrm{Eucl}}}(\theta_1 : \theta_2) = \frac{1}{2}\rho^2_{\mathrm{Eucl}}(\theta_1, \theta_2)$$

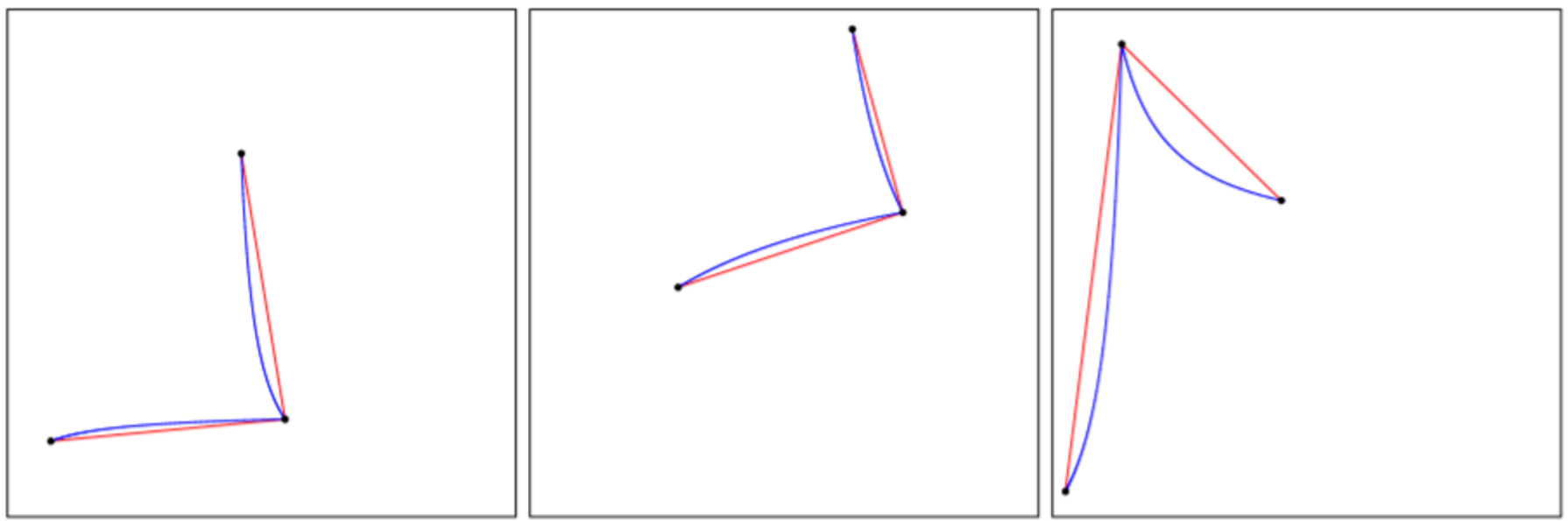


Triples of points (p,q,r) with dual Pythagorean theorems holding simultaneously at q



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

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



Itakura-Saito Manifold
(solve quadratic system)

Two blue-red geodesic pairs orthogonal at q

https://arxiv.org/abs/1910.03935

Dually flat space from a smooth strictly convex function F(θ)

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

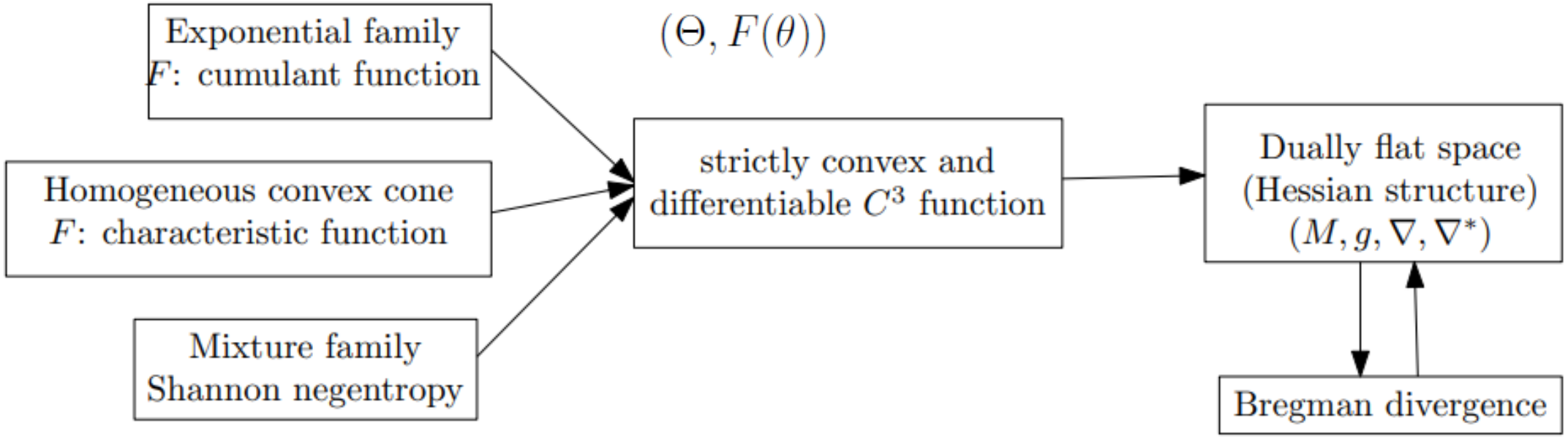
$$(\Theta, F(\theta)) \longrightarrow (M, g^{B_F}, \nabla^{B_F}, \nabla^{B_F{}^*}) = (M, g^F, \nabla^F, \nabla^{F*})$$

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 ∇-divergences

arXiv:2301.10980

Outline and contributions

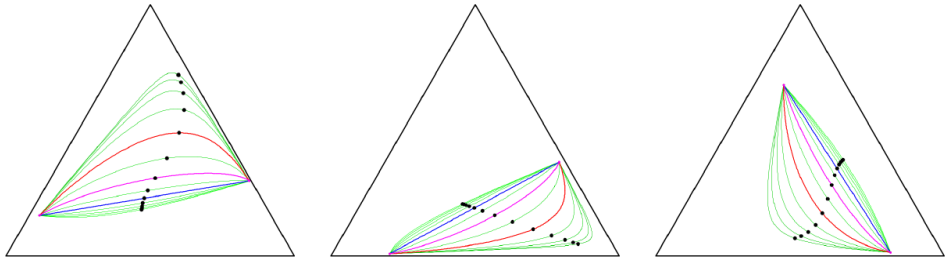
Goals:

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

Outline of the talk:

1. Weighted quasi-arithmetic means
2. Quasi-arithmetic centers and their invariance and equivariance properties
3. Quasi-arithmetic mixtures
4. Jensen-Shannon ∇-divergences



examples of α-geodesics with midpoints in the probability simplex

Weighted quasi-arithmetic means (QAMs)

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

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

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

QAMs enjoy the in-betweenness property:

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

[Kolmogorov 1930] [Nagumo 1930] [De Finetti 1931]

Quasi-arithmetic means (QAMs)

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

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

- So let us fix wlog. strictly increasing and differentiable f since we can always either consider either f or -f (i.e., λ=-1, c=0).

- QAMs include **p-power means** for the smooth family of generators $f_p(t)$:

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

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

Quasi-Arithmetic Centers (QACs) = Multivariate QAMs:

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

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 $F: R^d \rightarrow R^d$ 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 η may not be convex... $F^* \in \Gamma_0(E) \qquad F^{**} = F$

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

To by pass this problem:

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

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

Convex conjugate of a Legendre-type function (Θ,F(θ)) 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 ∇F is globally invertible: ∇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 $F : \Theta \to \mathbb{R}$ be a strictly convex and smooth real-valued function of Legendre-type in $\mathcal{F}$. The weighted quasi-arithmetic average of $\theta_1, \ldots, \theta_n$ and $w \in \Delta_{n-1}$ is defined by the gradient map ∇F as follows:*

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

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

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

Let $F(t) = \int_a^t f(u)\mathrm{d}u$

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: $\mathrm{Sym}_{++}(d) \to \mathbb{R}$ the cone of symmetric positive-definite matrices
- Inner product: $\langle A, B\rangle := \mathrm{tr}(AB^\top)$
- We have:

$$F(\theta) = -\log\det(\theta),$$

←Legendre-type function

$$\nabla F(\theta) = -\theta^{-1} =: \eta(\theta),$$

$$\nabla F^{-1}(\eta) = -\eta^{-1} =: \theta(\eta)$$

$$F^*(\eta) = \langle\theta(\eta), \eta\rangle - F(\theta(\eta)) = -d - \log\det(-\eta)$$

←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*: $M_{\nabla F^*}(\eta_1, \eta_2) = 2\left(\eta_1^{-1} + \eta_2^{-1}\right)^{-1}$

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

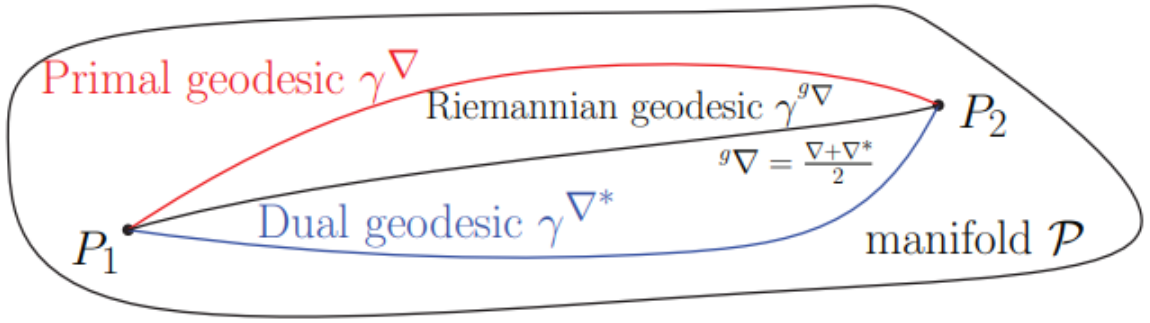
$M_{\nabla F}$ and $M_{\nabla F^*}$: dual operators

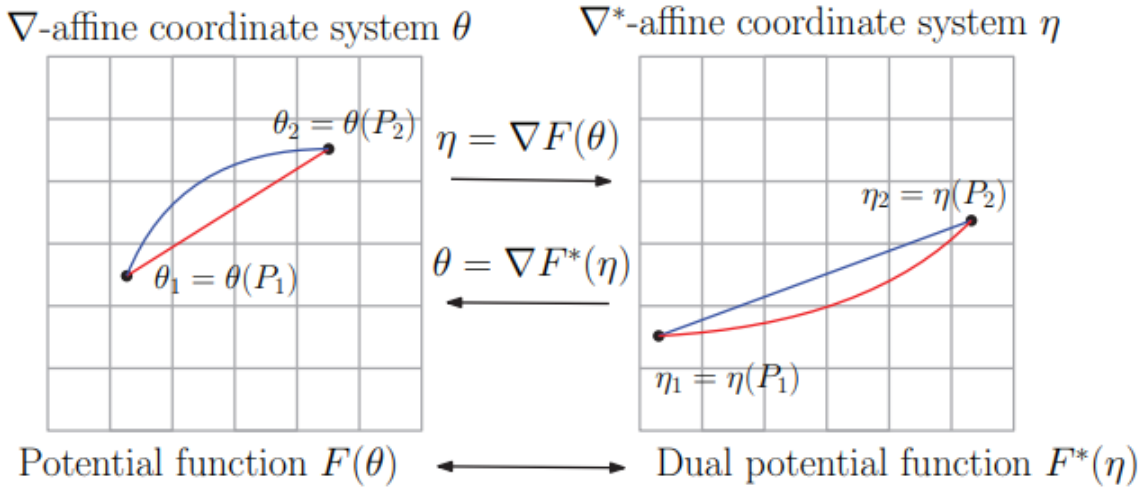
Dually flat structures of information geometry

- 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 θ-coordinate and dual η-coordinate





[AMS 2022]

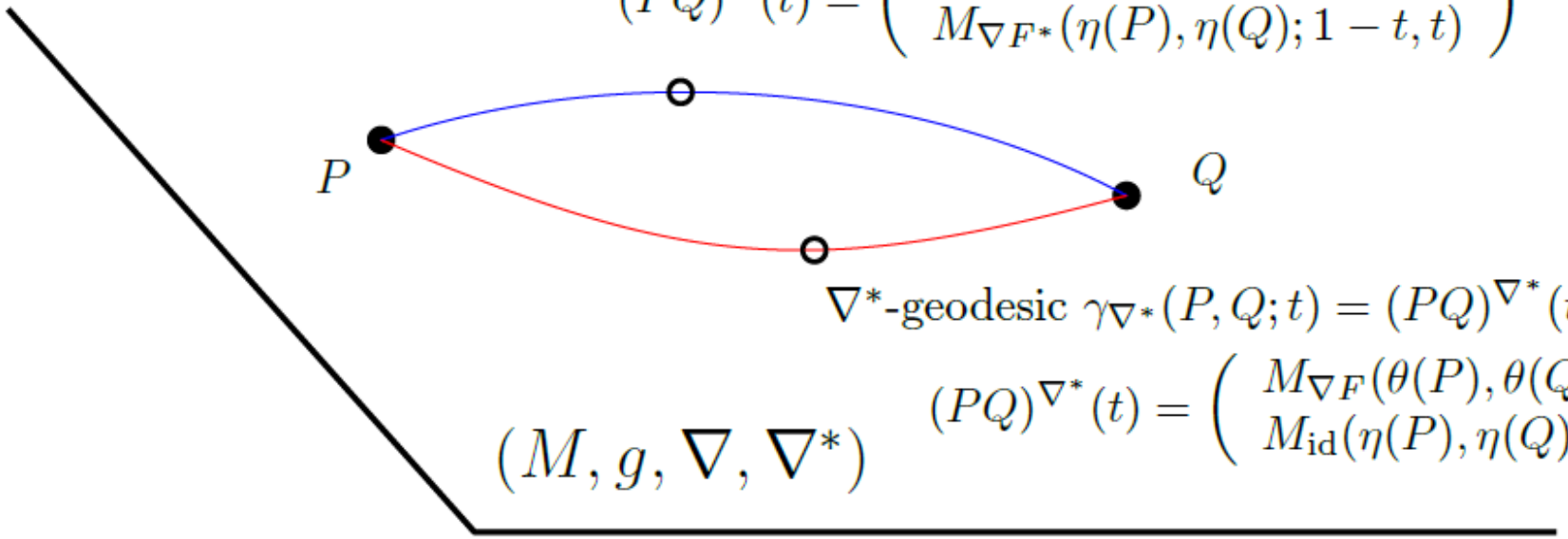
Quasi-arithmetic barycenters and dual geodesics

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

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

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

← dual QAC $M_{\nabla F^*}$



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

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

← primal QAC $M_{\nabla F}$

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

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

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

Right-sided centroid:

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

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

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

$$\bar{\eta}_R = \nabla F(\bar{\theta}_R) = M_{\nabla F^*}(\eta_1, \ldots, \eta_n).$$ ←dual QAC

Reference duality

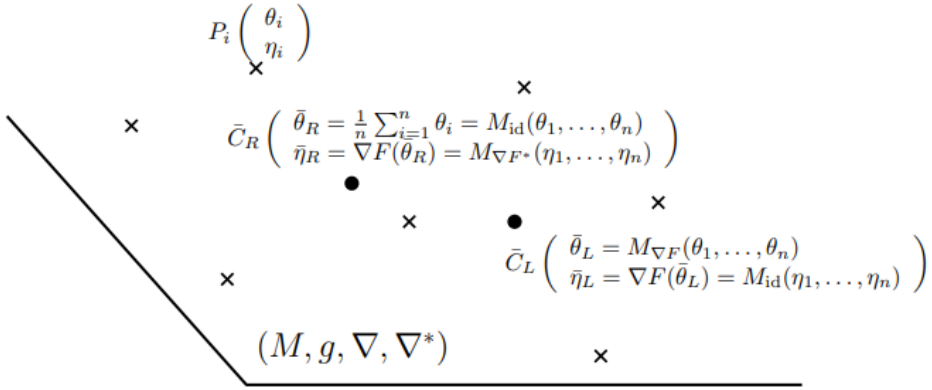
Left-sided centroid:

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

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

$$\bar{\theta}_L = M_{\nabla F}(\theta_1, \ldots, \theta_n),$$ ←primal QAC

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



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*:

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

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:

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

- By an affine change of coordinate...

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

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

$$B_{\bar{F}(\overline{\theta_1}:\overline{\theta_2})} = B_F(\theta_1 : \theta_2)$$

Equivariance of quasi-arithmetic center:

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

Same canonical divergence of the DFS
(= constrast 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**

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

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

$$\begin{aligned} (M, g, \nabla, \nabla^*) &\leftarrow \mathrm{DFS}([\Theta, F(\theta), H, F^*(\eta)]) \\ &\leftarrow \mathrm{DFS}([\bar{\Theta}, \bar{F}(\bar{\theta}), \bar{H}, \bar{F}^*(\bar{\eta})]) \end{aligned}$$

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

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 λF(θ) vs F(θ)

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

$$\Delta^2(\theta_1, \theta_2) = B_{F_Q}(\theta_1 : \theta_2) = \frac{1}{2}(\theta_2 - \theta_1)^\top Q\,(\theta_2 - \theta_1)$$

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_{id}:

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

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

Quasi-arithmetic mixtures (QAMixs), and α-mixtures

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

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

Centroid of n densities with respect to the α-divergences yields a QAMix:

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

D_α denotes the α-divergences:

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

α-families of probability distributions [Amari 2007]

[arXiv:2209.07481]

k=2 QAMixs and the ∇-Jensen-Shannon divergence

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

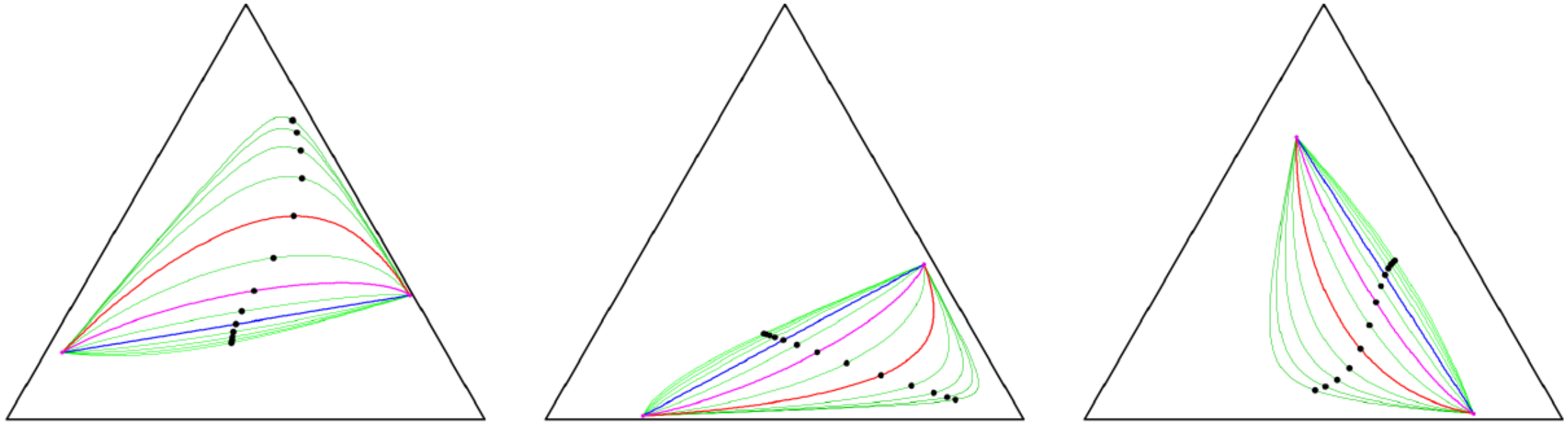
$$D_{\mathrm{JS}}(p,q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(p : \frac{p+q}{2}\right) + D_{\mathrm{KL}}\left(q : \frac{p+q}{2}\right)\right) \quad \leq \log(2)$$

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

Definition (Affine connection-based ∇-Jensen-Shannon divergence). *Let ∇ 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 ∇-Jensen-Shannon divergence is defined by:*

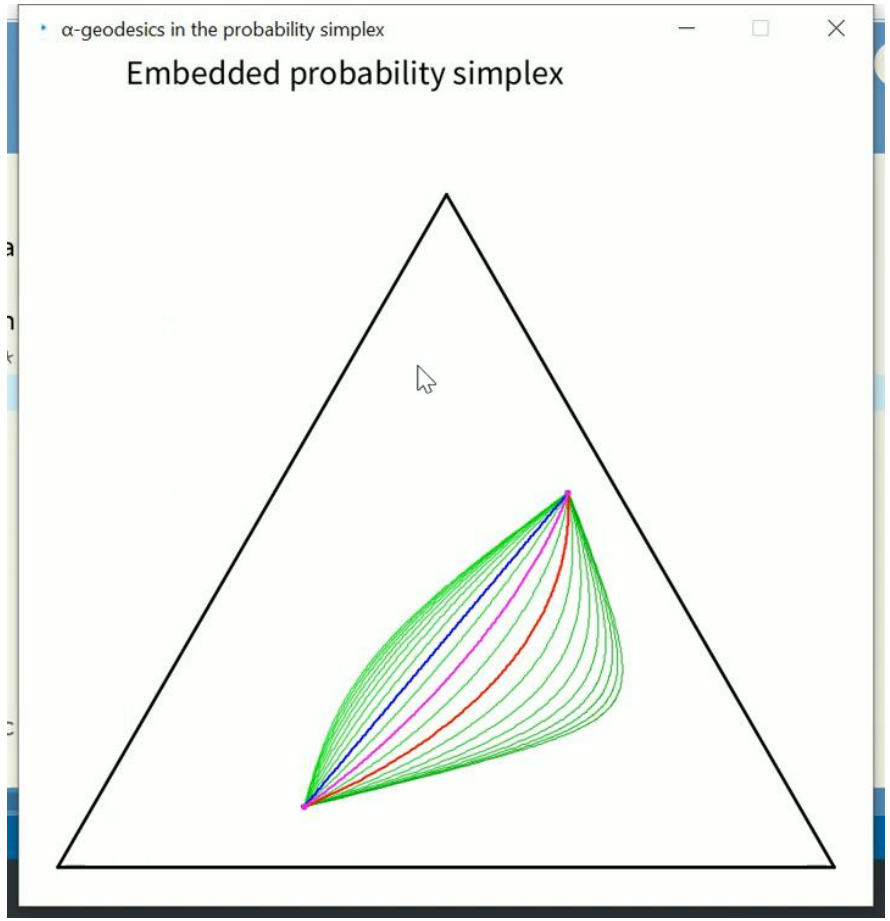
$$D_\nabla^{\mathrm{JS}}(p,q) := \frac{1}{2}\left(D_{\mathrm{KL}}\left(p : \gamma_\nabla\left(p,q;\frac{1}{2}\right)\right) + D_{\mathrm{KL}}\left(q : \gamma_\nabla\left(p,q;\frac{1}{2}\right)\right)\right).$$

∇^{α}-connections and geodesics in the probability simplex, ∇^{α}-Jensen-Shannon divergence

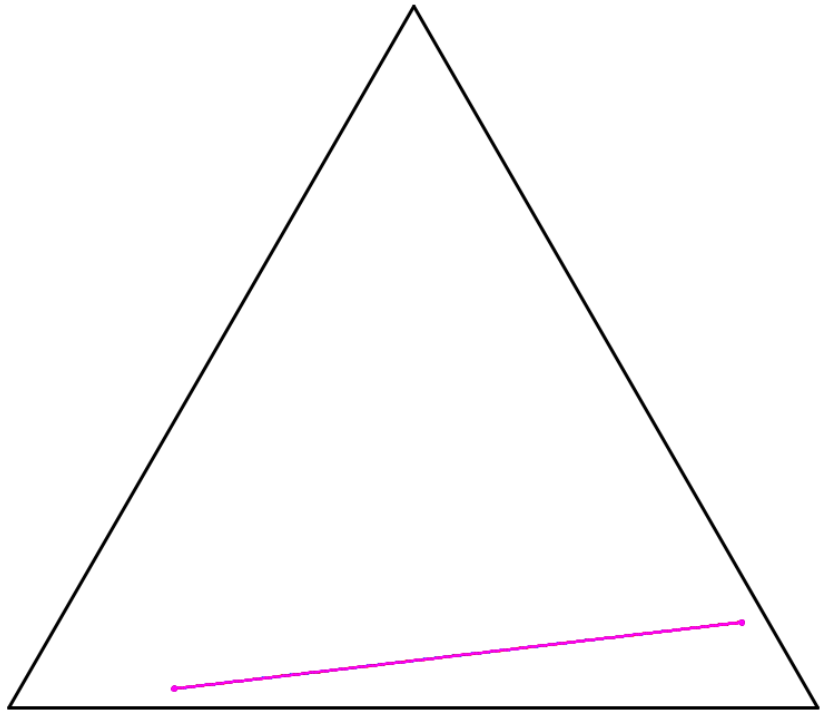


$$D_{\nabla^{\alpha}}^{\mathrm{JS}}(p,q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(p : \gamma_{\nabla^{\alpha}}\left(p,q;\frac{1}{2}\right)\right) + D_{\mathrm{KL}}\left(q : \gamma_{\nabla^{\alpha}}\left(p,q;\frac{1}{2}\right)\right)\right)$$

α-geodesics coincide when they pass through a standard simplex vertex



non-degenerate



degenerate

grateful for fruitful discussions with Fábio Meneghetti and Sueli Costa

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**

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

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

- 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:

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

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

Inductive matrix arithmetic-harmonic mean

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

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

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

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

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

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

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

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

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

$$g_P(V_1,V_2) = \mathrm{tr}\left(P^{-1}V_1P^{-1}V_2\right)$$

[Nakamura 2001, Atteia-Raissouli 2001]

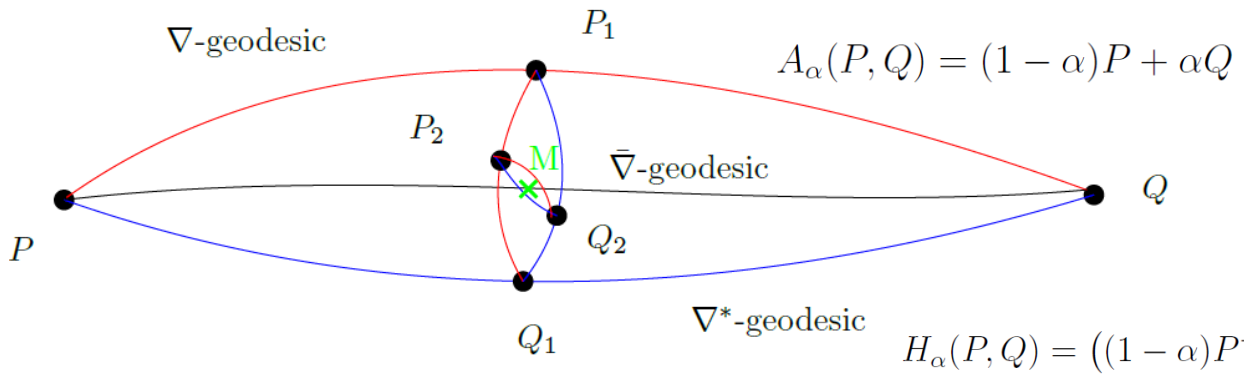
Geometric interpretation of the AHM matrix mean

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

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

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

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



$$H_\alpha(P,Q) = \left((1-\alpha)P^{-1} + \alpha Q^{-1}\right)^{-1}$$

Dually flat space (SPD, g^G, ∇^A, ∇^H) in information geometry defines quasi-arithmetic centers as geodesic midpoints

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

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

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

[Nakamura 2001]

Summary: Beyond scalar quasi-arithmetic means

Information geometry of dually flat spaces yields a generalization of quasi-arithmetic means:

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

- 1d monotone function generalize to gradient map of a Legendre-type multivatiate function (comonotone)

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

dual quasi-arithmetic centers induced by a Legendre-type function

Applications of QACs:

- dual centers of mass of n≥2 points expressed using weighted quasi-arithmetic centers
- dual geodesics expressed in coordinate systems as weighted quasi-arithmetic centers (n=2)
- invariance/equivariance analyzed from the viewpoint of information geometry

$$\bar{F}(\bar{\theta}) := \lambda(F(A\theta+b)+\langle c,\theta\rangle+d) \Rightarrow M_{\nabla \bar{F}} = A\, M_{\nabla F} + b.$$

- define quasi-arithmetic mixtures which provides a way to integrate density components
- define ∇-Jensen-Shannon divergences
- Inductive arithmetic-harmonic geometric matrix mean expressed using QACs

Revisiting Chernoff information with Likelihood Ratio Exponential Families

Entropy 2022
[2207.03745]

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:

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

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

(via Hölder inequality)

Herman Chernoff
(1923-)

- **skewed Bhattacharyya coefficient ρ_α** (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)

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

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

Bhattacharyya distance when α=1/2

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

α=1/2

$$D_B[p:q] = -\log \int \sqrt{pq} \mathrm{d}\mu = D_{B,\frac{1}{2}}[p:q]$$

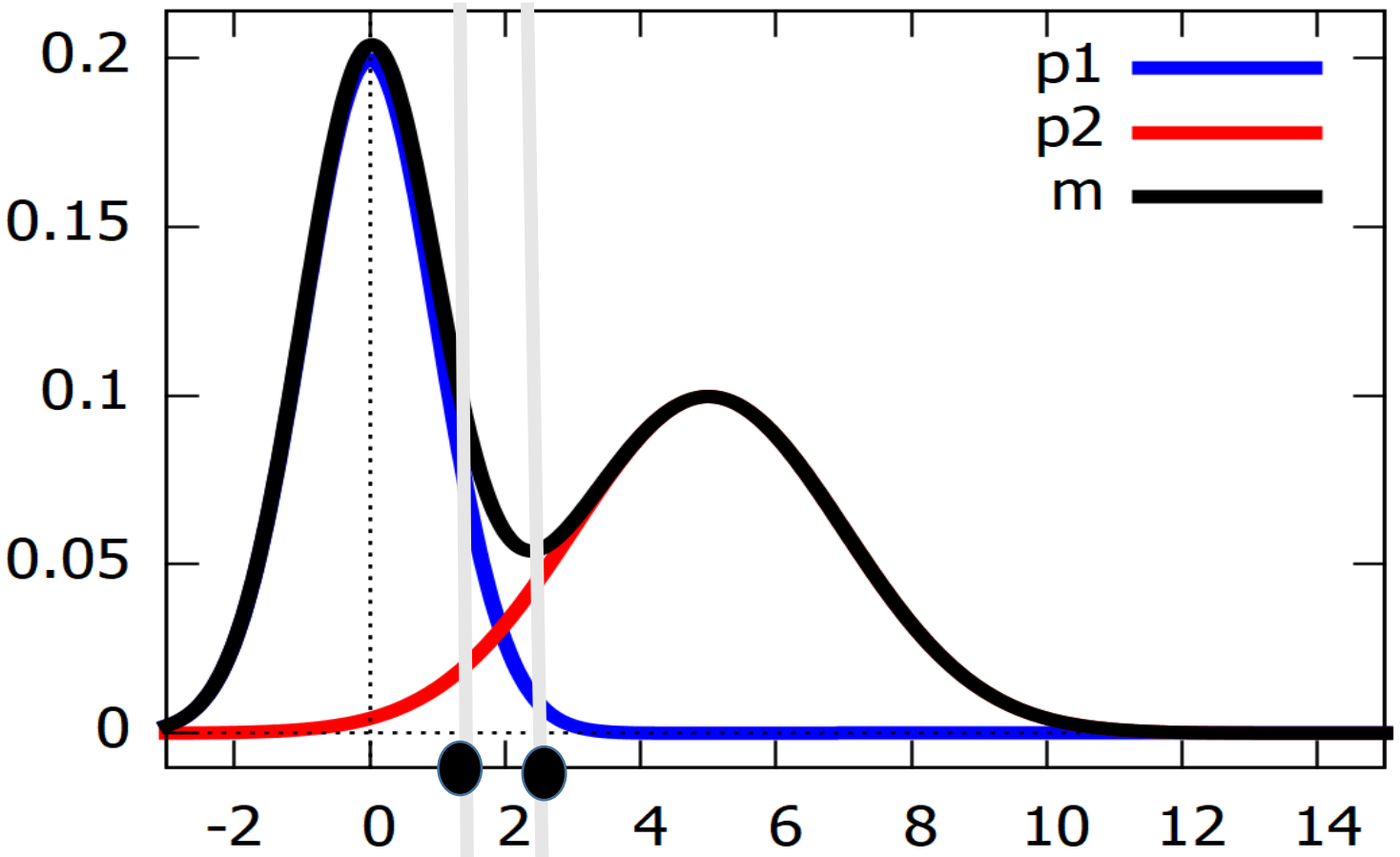
- **Bhattacharyya distance** does not satisfy the triangle inequality: not a metric
- Chernoff information tunes/learns the skewed Bhattacharyya distance
- Information = variational divergence (computed from an optimization procedure)
- Limit scaled skewed Bhattacharyya distance = Kullback-Leibler divergence

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

when α → 1, get KLD

Bhattacharyya, Anil. "On a measure of divergence between two statistical populations defined by their probability distributions." *Bull. Calcutta Math. Soc.* 35 (1943): 99-109.

Rationale for CI: Statistical hypothesis testing



x classified as p1

x classified as p2

Statistical mixture:

m(x)=0.5*N(0,1)+0.5*N(5,2)

Hypothesis task:

Decides whether x emanates from p1 or p2?

Classification rule:

Maximum a posteriori (MAP)

if p1(x)>p2(x) classify as p1

else classify as p2

Error at x: **min(p1(x),p2(x))**

Histogram intersection similarity:

$$P_e = \int \min(p_1(x), p_2(x))\mathrm{d}x$$

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**:

$$P_e = \int \min(p_1(x), p_2(x))\mathrm{d}x \quad \Rightarrow \quad \boxed{P_e = \frac{1}{2}\left(1 - D_{\mathrm{TV}}[p_1, p_2]\right)}$$

$$D_{\mathrm{TV}}[p_1, p_2] = \frac{1}{2}\int (p_1(x) - p_2(x))\mathrm{d}x$$

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

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

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

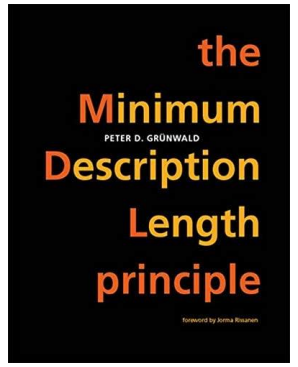
"Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means." *Pattern Recognition Letters* 42 (2014): 25-34.

Outline

- **Interplay of non-parametric with parametric study** of Chernoff information via the concept of **lilelihood-ratio exponential families** (LREFS)

- Derive various **optimality conditions** for the Chernoff exponent α^*
- Give some **geometric interpretations** on Bregman manifolds which yield *fast approximation algorithms*
- **novel closed-form solutions** for the Chernoff information between univariate Gaussians, centered scaled covariance matrices, etc.

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**:

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

  with **normalization factor**:

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

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

Sufficient statistics: log likelihood ratio

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

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

k(x)=log q(x)

- $D_{B,\alpha}[p:q]$

Natural parameter space:

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

LREFs: EF cumulant function is always analytic C^{ω}

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

  (and smooth for regular EFs)

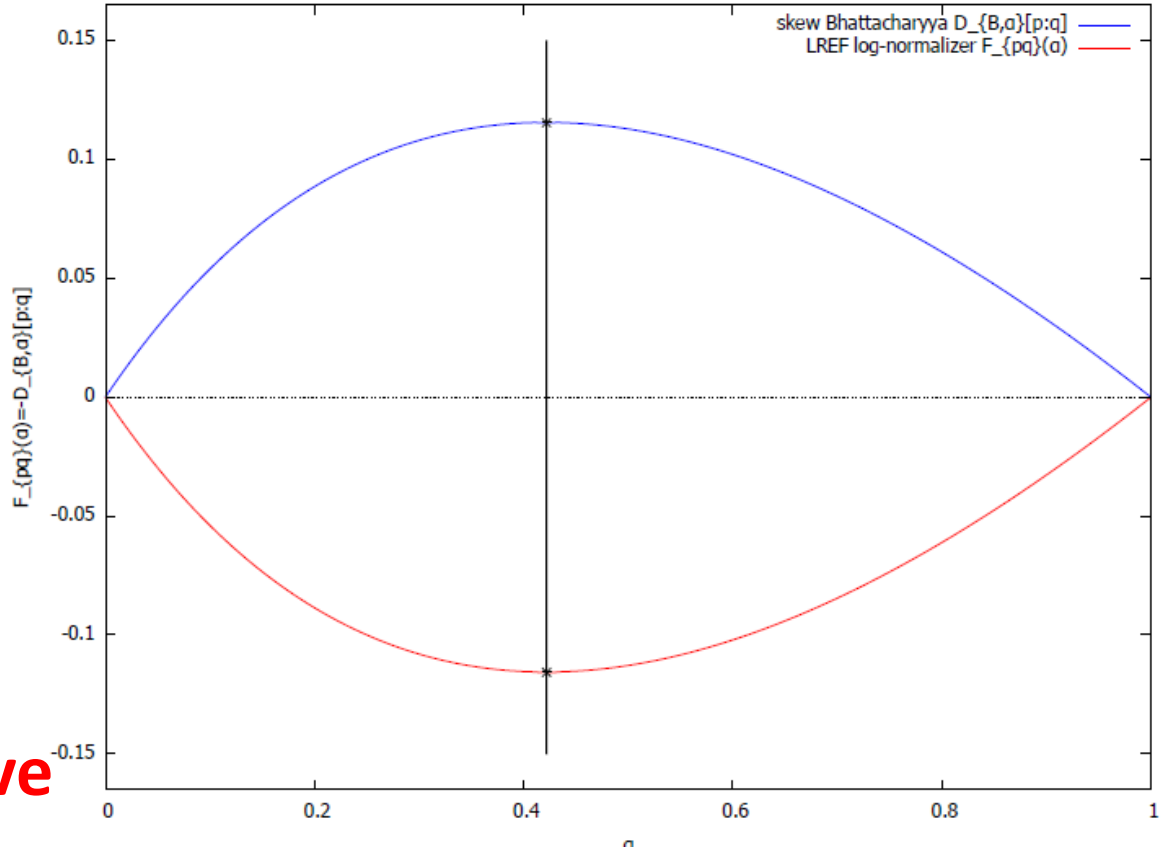
- Cumulant function is neg-Bhattacharyya distance:

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

⇒ Bhattacharyya. distance is **strictly concave**

- Theorem:

  Chernoff exponent exists and is unique



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

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

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

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

Geometric 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):

$$\mathrm{supp}\left((pq)_\alpha^G\right) = \begin{cases} \mathrm{supp}(p) \cap \mathrm{supp}(q), & \alpha \in \Theta_{pq} \setminus \{0,1\} \\ \mathrm{supp}(p), & \alpha = 1 \\ \mathrm{supp}(q), & \alpha = 0. \end{cases}$$

- Exponential family is said **regular** when the natural parameter space Θ is **open** (e.g., normal family, Dirichlet family, Wishart family, etc.)

Definition: regular EF $\Longleftrightarrow$ $\Theta = \Theta^\circ$

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 $D_{\mathrm{KL}}[p:q]$ and the reverse Kullback-Leibler divergence $D_{\mathrm{KL}}[q:p]$ are finite.*

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

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

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

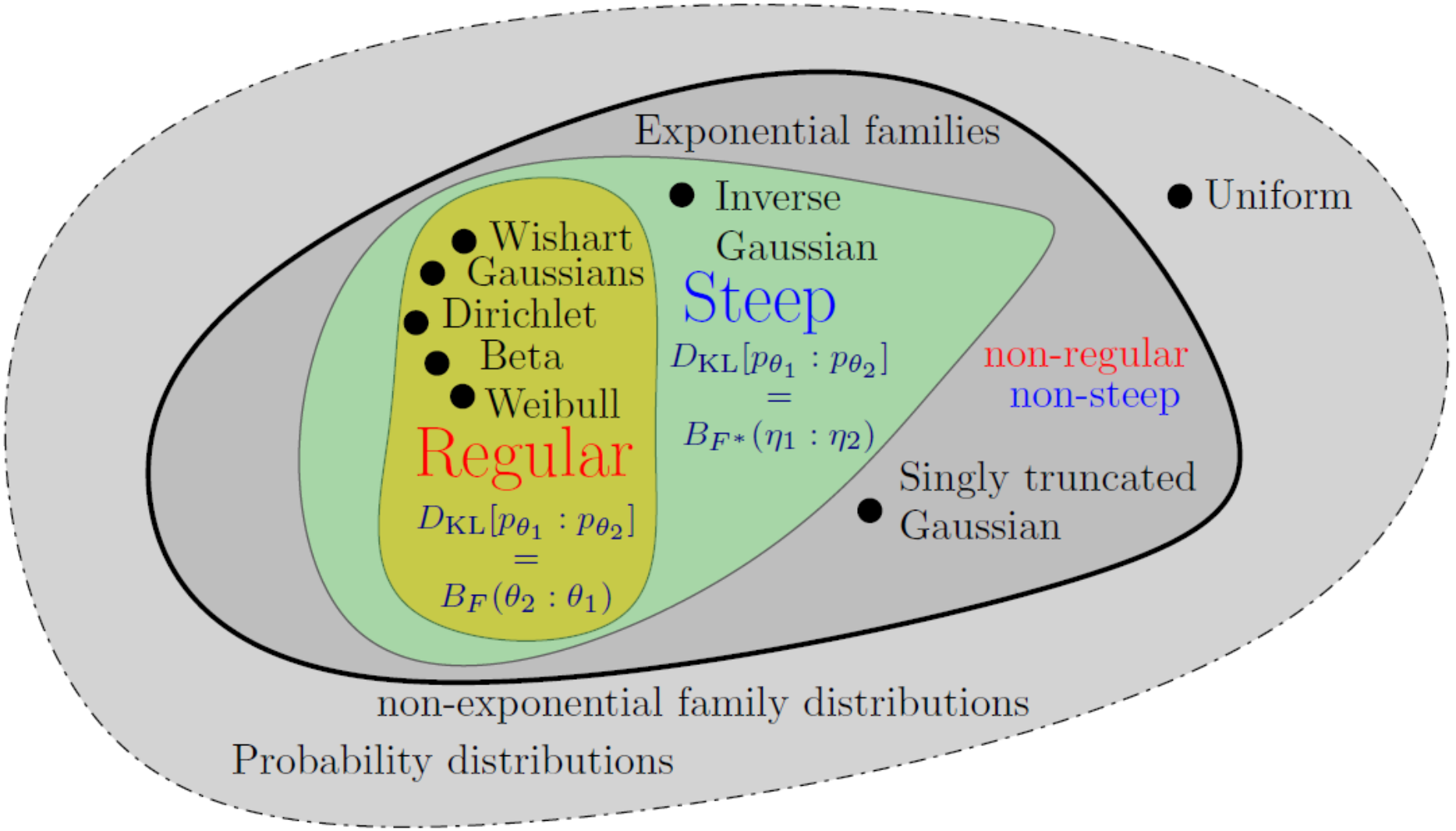
steep $\Rightarrow$ $E_{p_{\theta_1}}[t(x)] = \nabla F(\theta_1)$

regular EF $\Rightarrow$ steep EF

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

Venn diagram: Regular & steepness of (LR)EFs

- Steepness implies **duality between natural θ and moment η 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 $D_{\mathrm{KL}}[p:q]$ and the reverse Kullback-Leibler divergence $D_{\mathrm{KL}}[q:p]$ are finite.*

PROOF

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

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

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

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

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

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

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

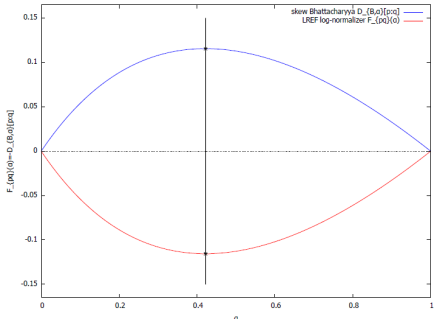
Chernoff information (for densities of a LREF)

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

PROOF

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

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



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

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

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

① $D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : p] = B_{F_{pq}}(1 : \alpha^*) = -F(\alpha^*)$ ② $D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : q] = B_{F_{pq}}(0 : \alpha^*) = -F(\alpha^*)$

③ $D_C[p:q] = -\log \rho_{\alpha^*}(p:q) = J_{F_{pq},\alpha^*}(1:0) = -F_{pq}(\alpha^*)$

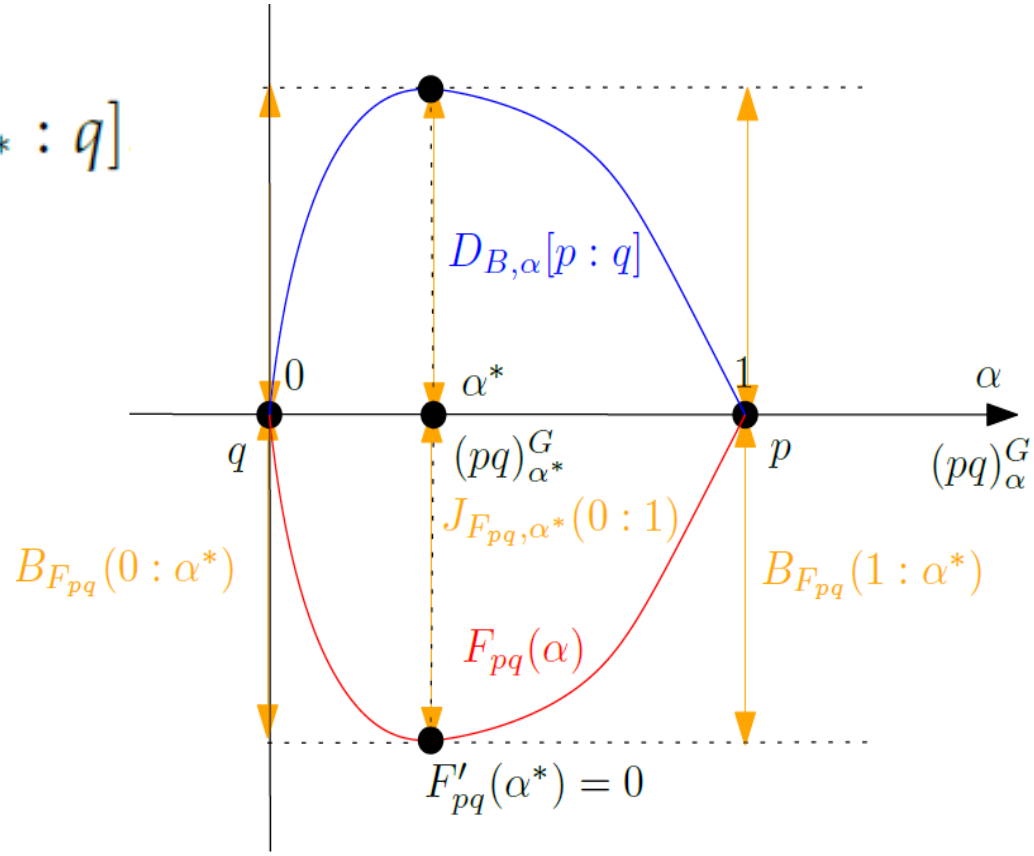
Jensen-Chernoff divergence

$$D_C[p:q] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : p] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : q]$$

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 $L_1(\mu)$

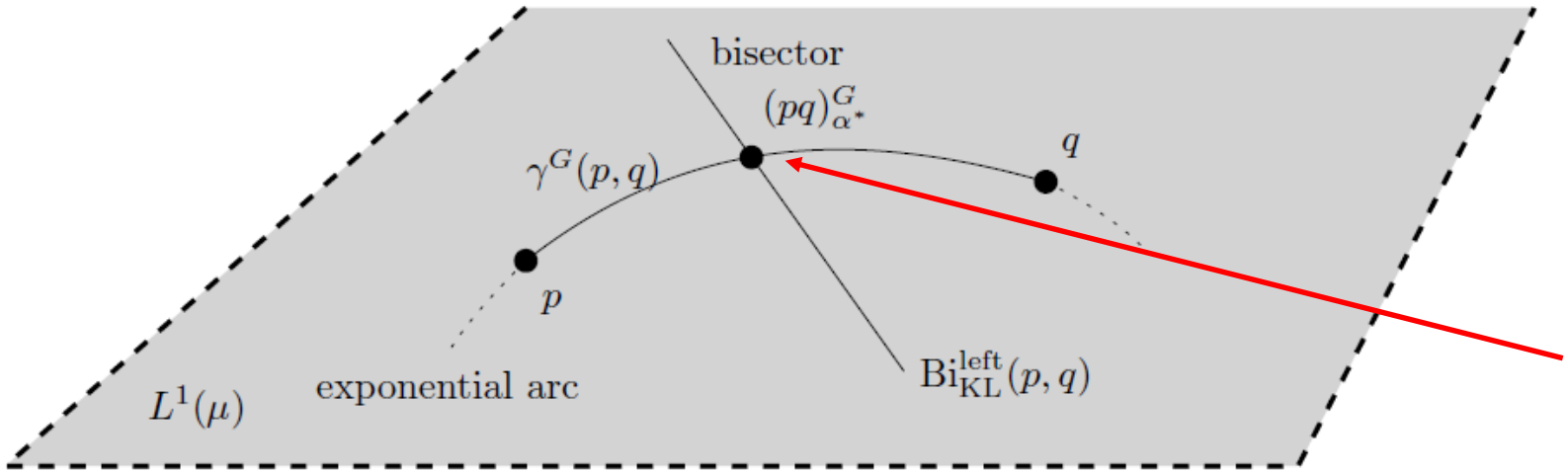
Proposition (Geometric characterization of the Chernoff information). *On the vector space $L^1(\mu)$, the Chernoff information distribution is the unique distribution*

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

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

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

2209.07481



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

Chernoff information viewed as a symmetrization of KLD

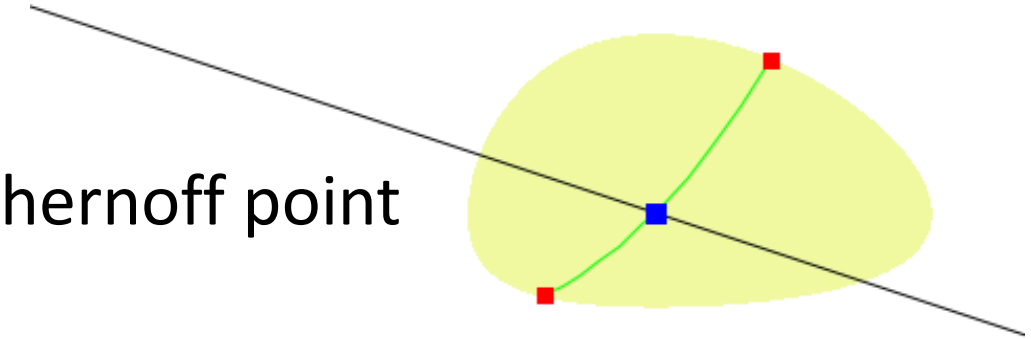
Rewrite $D_C[p:q] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : p] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : q]$

as $$D_C[p:q] = \min_{r \in \mathcal{E}_{\overline{pq}}} \{D_{\mathrm{KL}}[r:p], D_{\mathrm{KL}}[r:q]\}.$$

Chernoff information as the radius of

a **minimum enclosing left-sided Kullback–Leibler ball**

circumcenter = Chernoff point



Chernoff point r*
(eKLD)

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

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

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

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

$$(p_{\theta_1} p_{\theta_2})^G_\alpha = p_{\alpha\theta_1+(1-\alpha)\theta_2} \qquad D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1)$$

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

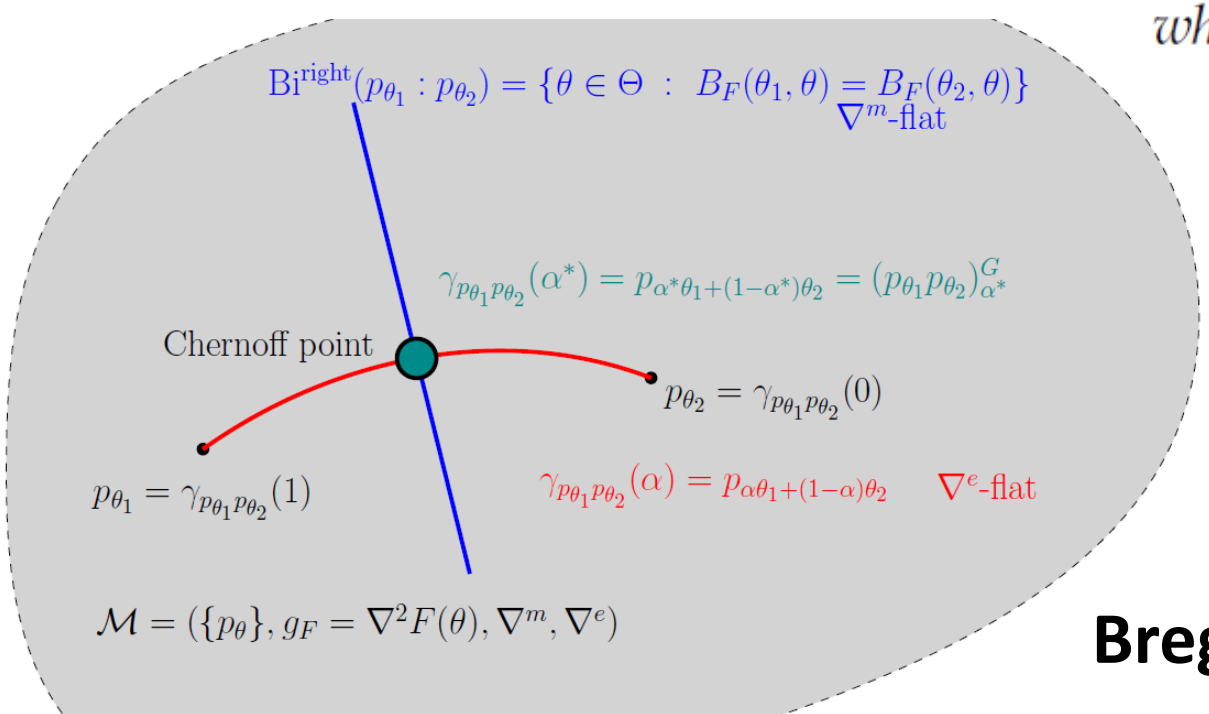
Proposition *Let p_{λ_1} and p_{λ_2} be two densities of a regular exponential family $\mathcal{E}$ with natural parameter $\theta(\lambda)$ and log-normalizer $F(\theta)$. Then the Chernoff information is*

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

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

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

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 $\gamma^e_{pq}(\alpha)$ and $\gamma^m_{pq}(\alpha)$ be a e-geodesic and m-geodesic passing through the points p and q of $\mathcal{M}$, respectively. Let $\mathrm{Bi}^m(p,q)$ and $\mathrm{Bi}^e(p,q)$ be the right-sided ∇^m-flat and left-sided ∇^e-flat Bregman bisectors, respectively. Then the intersection of $\gamma^e_{pq}(\alpha)$ with $\mathrm{Bi}^m(p,q)$ and the intersection of $\gamma^m_{pq}(\alpha)$ with $\mathrm{Bi}^e(p,q)$ are unique. The point $\gamma^e_{pq}(\alpha) \cap \mathrm{Bi}^m(p,q)$ is called the Chernoff point and the point $\gamma^m_{pq}(\alpha) \cap \mathrm{Bi}^e(p,q)$ 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.

New result: Exact closed-form for Chernoff between univariate Gaussian distributions

- Optimality condition amounts to solve a quadratic equation for α

$$\langle \theta_2 - \theta_1, \eta_{\alpha^*} \rangle = F(\theta_2) - F(\theta_1)$$

$$\mathrm{OC}_{\mathrm{Gaussian}}: \quad \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) m_\alpha - \left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right) v_\alpha = \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}.$$

- Use symbolic computing to get very long closed-form formula by solving a quadratic equation

Maxima

A Computer Algebra System

Symbolic computing

```
(%i13) varalpha(v1,v2,alpha):=(v1·v2)/((1−alpha)·v1+alpha·v2)$
       mualpha(mu1,v1,mu2,v2,alpha):=(alpha·mu1·v2+(1−alpha)·mu2·v1)/((1−alpha)·v1+alpha·v2)$
       KLD(mu1,v1,mu2,v2):=(1/2)·((((mu2−mu1)··2)/v2)+(v1/v2)−log(v1/v2)−1)$
       assume(alpha>0)$assume(alpha<1)$
       assume(v1>0)$assume(v2>0)$
       theta1(mu,v):=mu/v$
       theta2(mu,v):=−1/(2·v)$;
       F(theta1,theta2):=−theta1··2/(4·theta2)+0.5·log(−1/theta2)$

       eq: (theta1(mu1,v1)−theta1(mu2,v2))·mualpha(mu1,v1,mu2,v2,alpha)+(theta2(mu1,v1)
         −theta2(mu2,v2))·(mualpha(mu1,v1,mu2,v2,alpha)··2+varalpha(v1,v2,alpha))−F(theta1(mu1,v1),theta2(mu1,v1))+F(theta1(mu2,v2),theta2(mu2,v2));
       solalpha: solve(eq,alpha)$
       alphastar:rhs(solalpha[1]);
```

$$(\%o11)\quad 0.5\log(2\,v2)+\left(\frac{1}{2\,v2}-\frac{1}{2\,v1}\right)\left(\frac{\left(alpha\,mu1\,v2+(1-alpha)\,mu2\,v1\right)^2}{\left(alpha\,v2+(1-alpha)\,v1\right)^2}+\frac{v1\,v2}{alpha\,v2+(1-alpha)\,v1}\right)+\frac{\left(\frac{mu1}{v1}-\frac{mu2}{v2}\right)\left(alpha\,mu1\,v2+(1-alpha)\,mu2\,v1\right)}{alpha\,v2+(1-alpha)\,v1}+\frac{mu2^2}{2\,v2}-0.5\log(2\,v1)-\frac{mu1^2}{2\,v1}$$

```
rat: replaced −0.5 by −1/2 = −0.5
rat: replaced 0.5 by 1/2 = 0.5
```

$$(\%o13)\quad (\mathrm{sqrt}(\left(\left(4\,mu2^2-8\,mu1\,mu2+4\,mu1^2\right)v1\,v2^2+\left(-4\,mu2^2+8\,mu1\,mu2-4\,mu1^2\right)v1^2\,v2\right)\log(2\,v2)+v2^4-4\,v1\,v2^3+\left(\left(-4\,mu2^2+8\,mu1\,mu2-4\,mu1^2\right)v1\log(2\,v1)+6\,v1^2\right)v2^2+$$
$$\left(\left(4\,mu2^2-8\,mu1\,mu2+4\,mu1^2\right)v1^2\log(2\,v1)-4\,v1^3+\left(4\,mu2^4-16\,mu1\,mu2^3+24\,mu1^2\,mu2^2-16\,mu1^3\,mu2+4\,mu1^4\right)v1\right)v2+v1^4)+\left(2\,v1^2-2\,v1\,v2\right)\log(2\,v2)+v2^2+$$
$$(2\,v1\log(2\,v1)-2\,v1)\,v2-2\,v1^2\log(2\,v1)+v1^2+\left(-2\,mu2^2+4\,mu1\,mu2-2\,mu1^2\right)v1)/$$
$$\left(\left(2\,v2^2-4\,v1\,v2+2\,v1^2\right)\log(2\,v2)-2\log(2\,v1)\,v2^2+\left(4\,v1\log(2\,v1)+2\,mu2^2-4\,mu1\,mu2+2\,mu1^2\right)v2-2\,v1^2\log(2\,v1)+\left(-2\,mu2^2+4\,mu1\,mu2-2\,mu1^2\right)v1\right)$$

General multivariate Gaussian case: Approximation

input : Two normal densities p_{μ_1,Σ_1} and p_{μ_2,Σ_2}, and a numerical precision threshold $\epsilon > 0$

$\alpha_m = 0;$
$\alpha_M = 1;$
while $|\alpha_M - \alpha_m| > \epsilon$ **do**
  $\alpha = \frac{\alpha_m + \alpha_M}{2};$
  $\Sigma_\alpha^e = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1};$
  $\mu_\alpha^e = \Sigma_\alpha^e\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right);$
  //
  if $D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_1,\Sigma_1}] > D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_2,\Sigma_2}]$ **then**
    $\alpha_m = \alpha;$
    //
  end
  else
    $\alpha_M = \alpha;$
  end
end
return $D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_1,\Sigma_1}];$

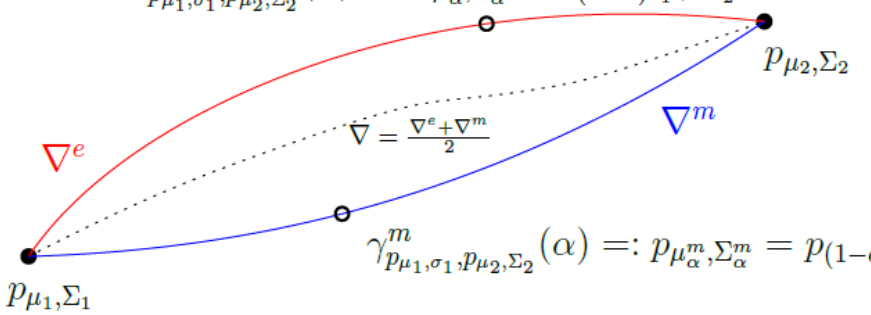
Kullback-Leibler divergence (= rev. Bregman div):

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

Efficient algorithm

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

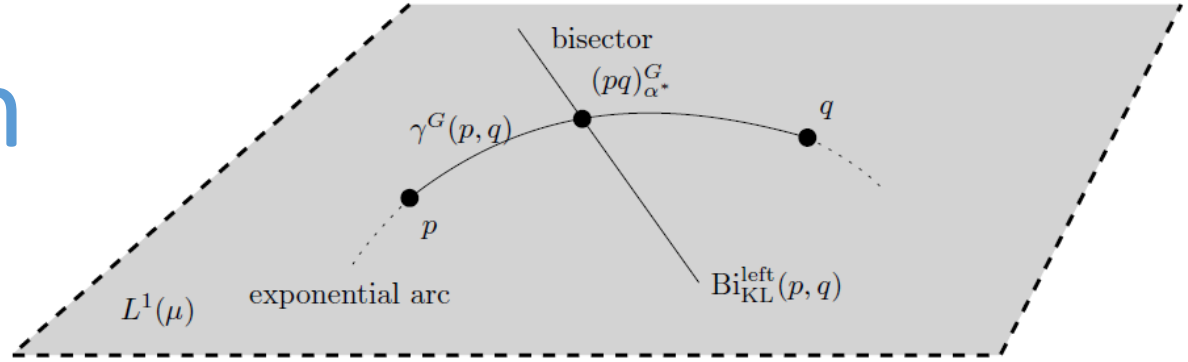
$\gamma^e_{p_{\mu_1,\sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^e,\Sigma_\alpha^e} = p_{(1-\alpha)\theta_1+\alpha\theta_2}$ $\quad \theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$

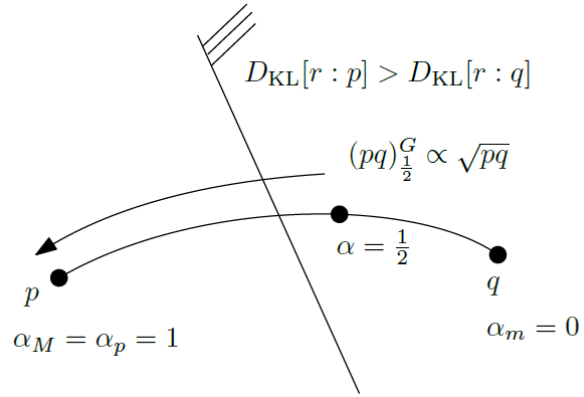


$\gamma^m_{p_{\mu_1,\sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) =: p_{\mu_\alpha^m,\Sigma_\alpha^m} = p_{(1-\alpha)\eta_1+\alpha\eta_2}$ $\quad \eta = (\mu, -\Sigma - \mu\mu^\top)$

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

Conclusion





- We revisited **Chernoff information** of two probability distributions under the umbrella of special exponential families
- The geometric mixture is a 1D **log-ratio exponential family**:
  - NEW • Chernoff exponent is unique (from the convexity of log-normalizer)
  - Geometrically, Chernoff point = intersection of a Voronoi bisector with a geodesic
  - Approximate the Chernoff information by **bisection search on exponential arc**
  - Express the **optimality condition** of Chernoff error exponent in various ways
- Consider Chernoff information between Gaussian densities:
  - NEW • exact closed-form for **univariate Gaussians** (solve quadratic eq. using symbolic computing)
  - NEW • exact closed-form for **centered scaled covariance matrices**
  - Practical bisection search on the exponential arc with numerical experiments

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

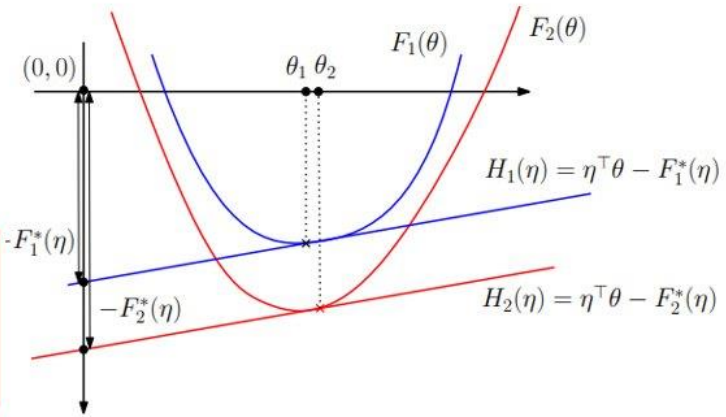
Legendre transformation reverses majorization order

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

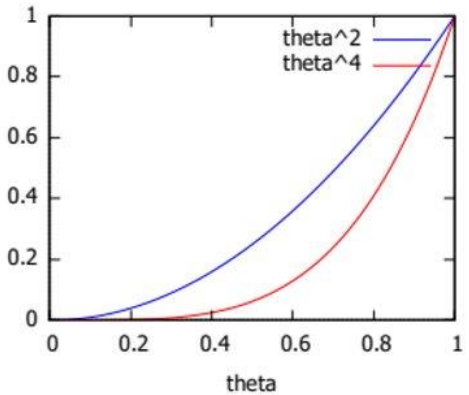
Legendre-Fenchel transform **reverses ordering**:

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

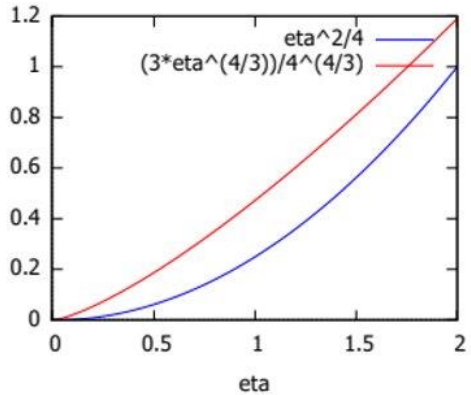


Proof:

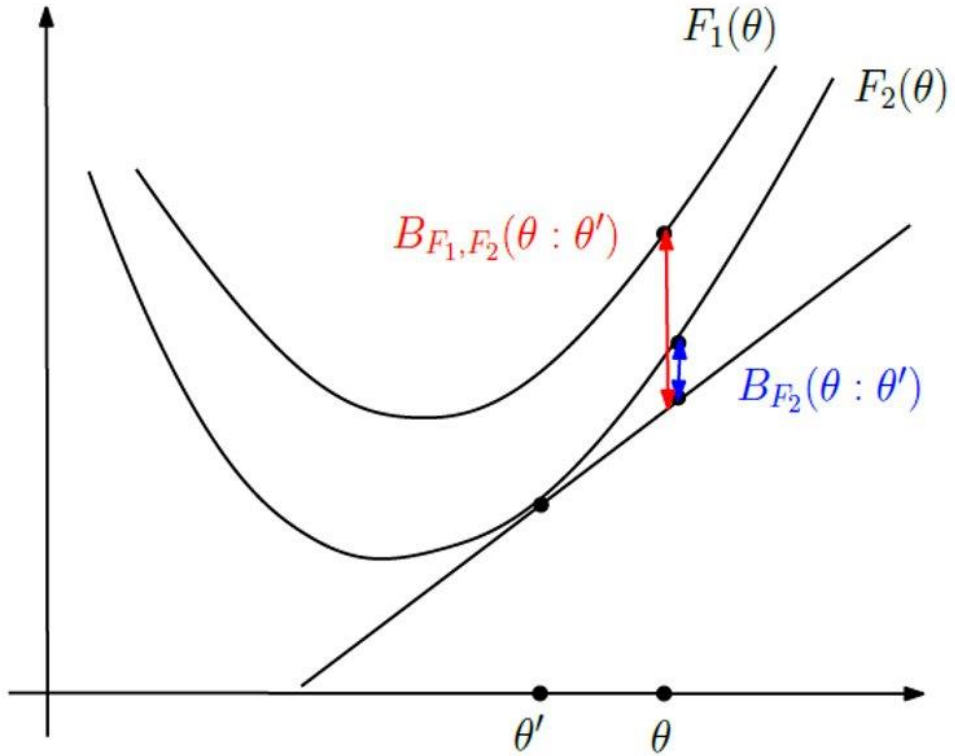
$$\begin{aligned} F_1^*(\eta) &:= \sup_{\theta \in \Theta} \{\eta^\top \theta - F_1(\theta)\}, \\ &= \eta^\top \theta_1 - F_1(\theta_1) \qquad (\text{with } \eta = \nabla F_1(\theta_1)) \\ &\leq \eta^\top \theta_1 - F_2(\theta_1), \\ &\leq \sup_{\theta \in \Theta} \{\eta^\top \theta - F_2(\theta)\} =: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



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

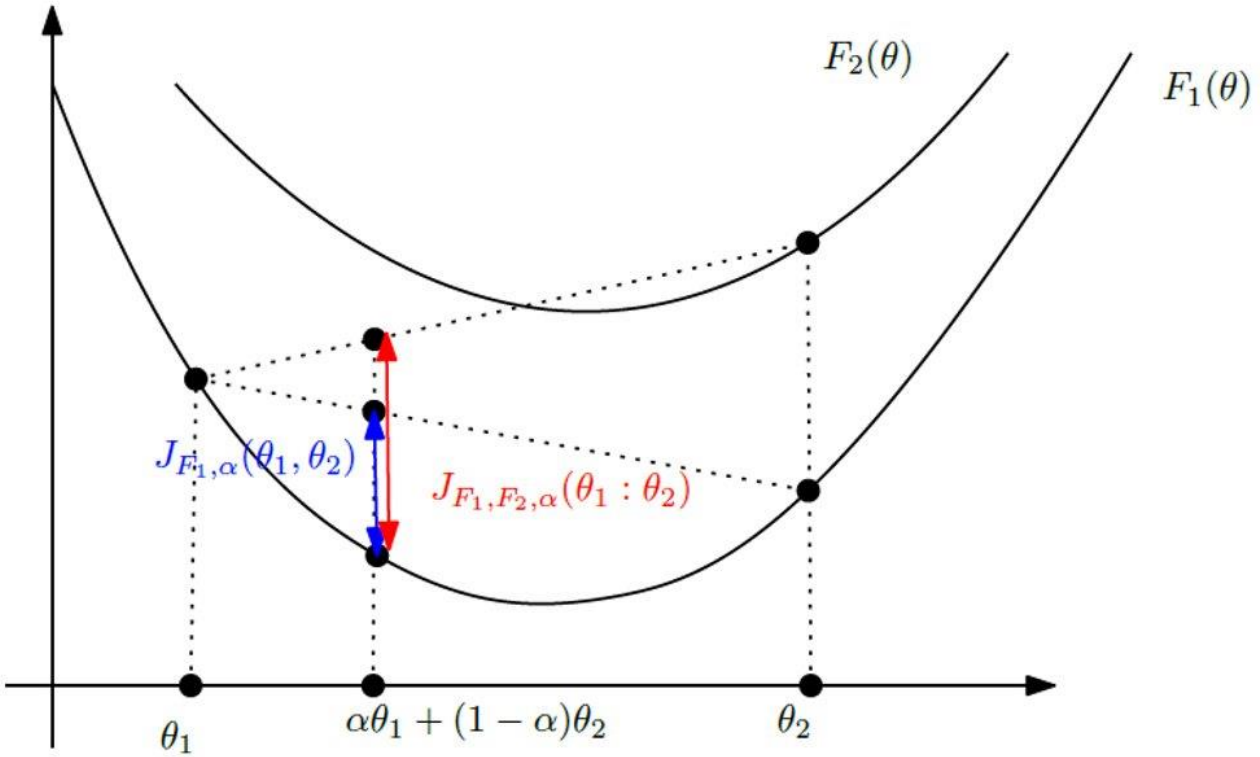


Conjugate functions $F_1^*(\eta) \leq F_2^*(\eta)$



Duo Bregman divergence

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



Duo Jensen divergence

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

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:

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

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

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

Kullback-Leibler divergence between exponential family densities

$$D_{\mathrm{KL}}[P:Q] = \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}P = E_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]$$

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

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

Duo

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

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

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

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

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

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

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

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

KL divergence between truncated normal densities

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

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

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

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

Moment parameters and mean & variance:

$$\begin{aligned}
\eta_1(m,s;a,b) &= E_{p_{m,s}^{a,b}}[x] = \mu(m,s;a,b), \\
\eta_2(m,s;a,b) &= E_{p_{m,s}^{a,b}}[x^2] = \sigma^2(m,s;a,b) + \mu^2(m,s;a,b).
\end{aligned}$$

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

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

Kullback-Leibler divergence between nested truncated normal distributions:

$$\begin{aligned}
D_{\mathrm{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = {} & \frac{m_2}{2s_2^2} - \frac{m_1}{2s_1^2} + \log\frac{Z_{a_2,b_2}(m_2,s_2)}{Z_{a_1,b_1}(m_1,s_1)} - \left(\frac{m_2}{s_2^2} - \frac{m_1}{s_1^2}\right)\eta_1(m_1,s_1;a_1,b_1) \\
& - \left(\frac{1}{2s_1^2} - \frac{1}{2s_2^2}\right)\eta_2(m_1,s_1;a_1,b_1)
\end{aligned}$$

if nested distributions $(a_1,b_1) \subseteq (a_2,b_2)$

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