Information geometry:

Geometry of dual structures

Frank Nielsen

Sony Computer Science Laboratories Inc

Tokyo, Japan

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, regular vs singular (ML), hierarchical (ML) or not, ...
- 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, etc

Information geometry: Rationale and scope

- More generally, **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

- Because the information-geometric structures are **purely geometric** (i.e., 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)*

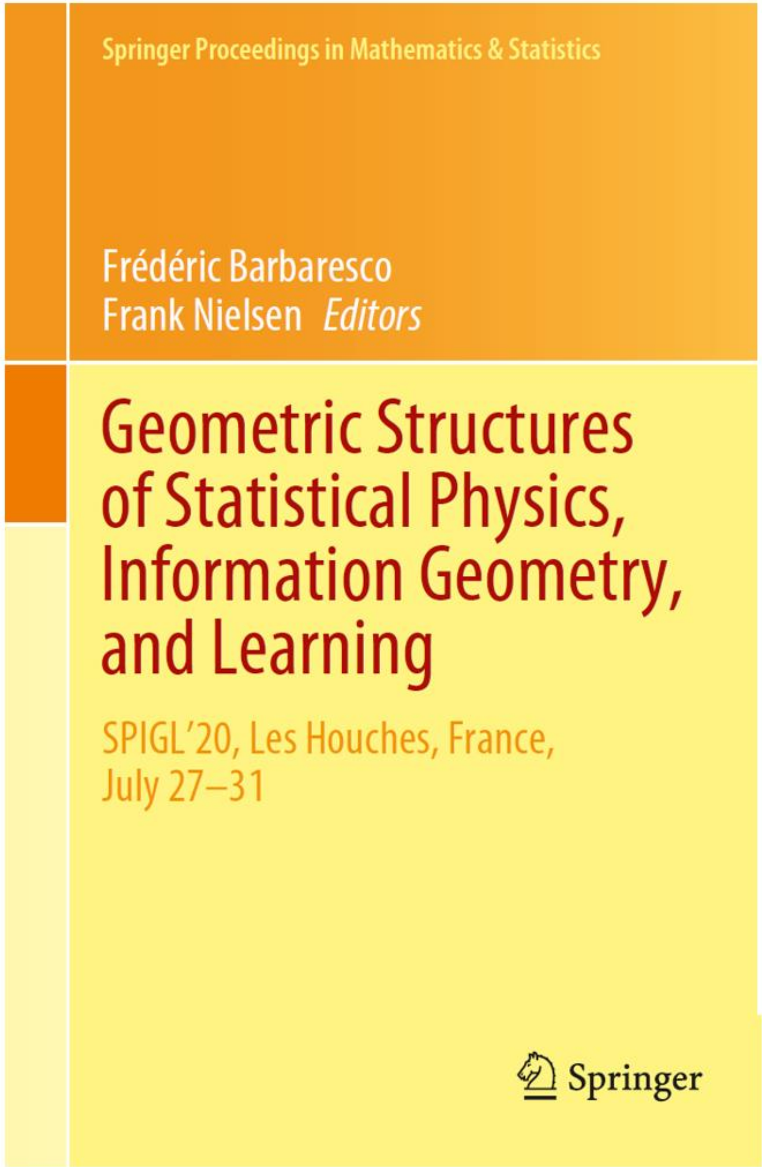

Shun-ichi Amari
Information Geometry and Its Applications
Springer
Information Geometry and Its Applications
Springer
Translations of MATHEMATICAL MONOGRAPHS
Statistical Decision Rules and Optimal Inference
N. N. Čencov
Signals and Communication Technology
Frank Nielsen
Frank Critchley
Christopher T.J. Dodson Editors
Computational Information Geometry
For Image and Signal Processing
Springer
Ovidiu Calin · Constantin Udrişte
Geometric Modeling in Probability and Statistics
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
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)

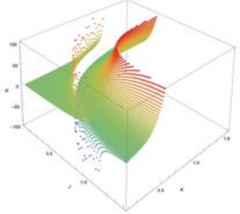
Extend broadly the original scope of information geometry by unravelling **connections** of information geometry (IG) with **other domains of geometry** like:

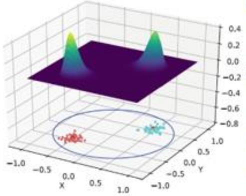
- geometry of domains and cones (e.g., Siegel/Vinberg/Koszul)
- geometric mechanics for dynamic models (symplectic/contact geometry)
- thermodynamics/thermostatistics and deformed statistical models
- geometric statistics (eg, computational anatomy/medical imaging)
- shape space analysis and deformation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

franknielsen.github.io/GSI/



Contents





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

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**

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: $[g_F]_\theta := I_X(\theta)$ with

Introduction to
Information Geometry

Frank NIELSEN
July 2022

Sony CSL

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

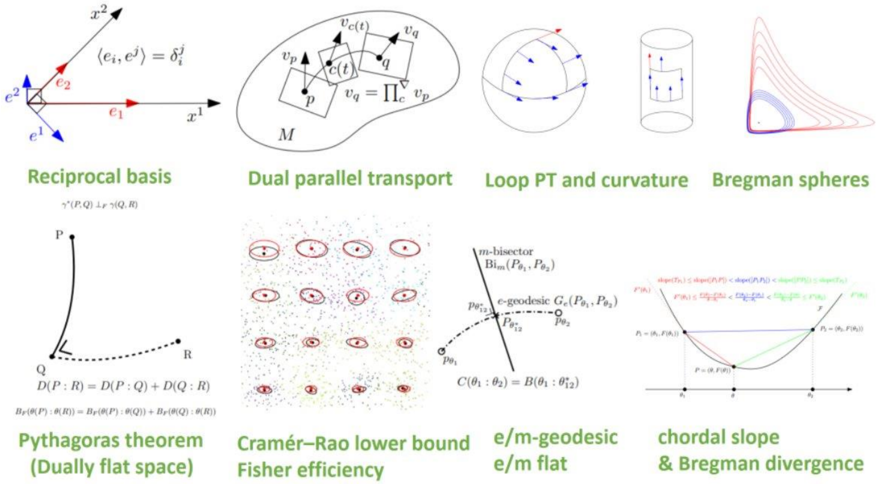
"Introduction to Information Geometry" by Frank Nielsen

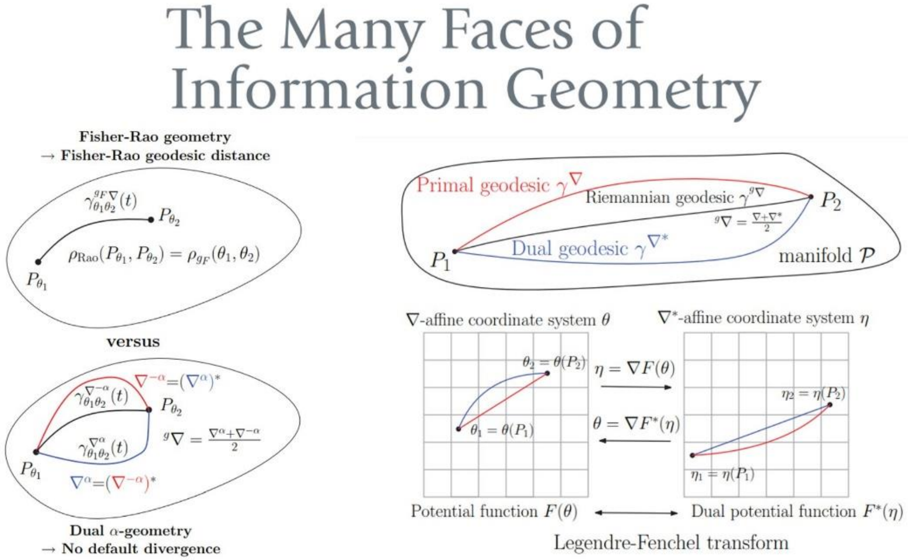
Frank Nielsen
628 subscribers
Analytics | Edit video | 295 | Share | Download

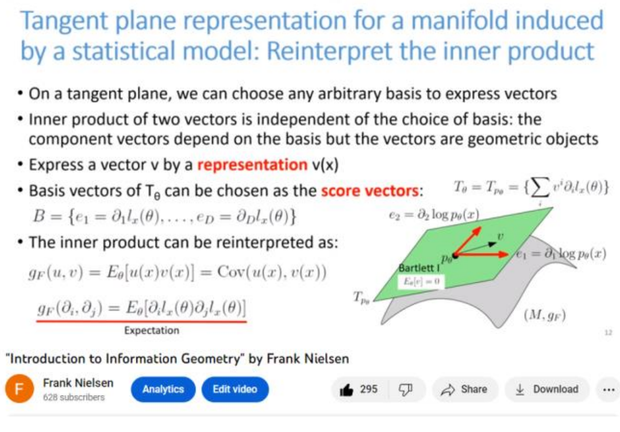
Tutorial 60+ pages
https://www.mdpi.com/1099-4300/22/10/1100

Short overview 10 pages
https://www.ams.org/journals/notices/202201/rnoti-p36.pdf

40 min. video introduction
https://www.youtube.com/watch?v=w6r_jsEBlgU



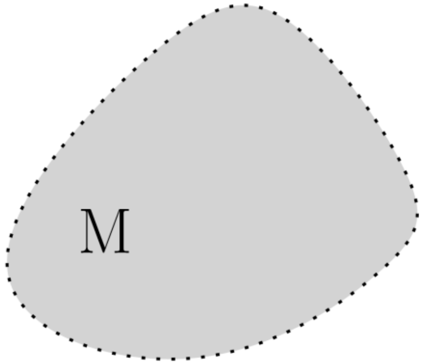




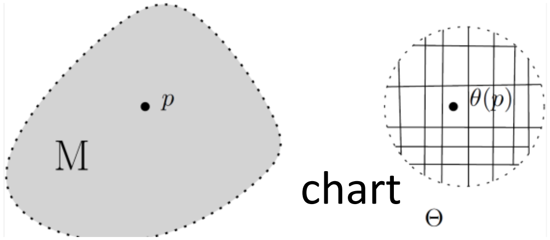
Build your own information geometry in three steps

Choose

① manifold M

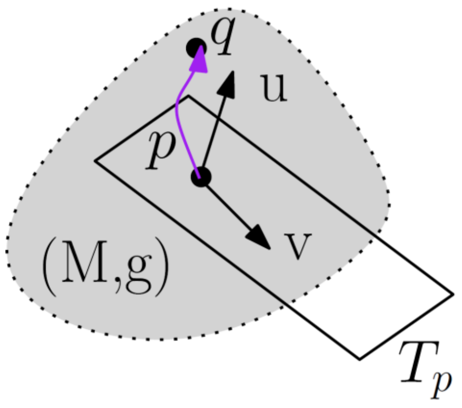


Examples:
Gaussians
SPD cone
Probability simplex



Concepts:
local coordinates
locally Euclidean

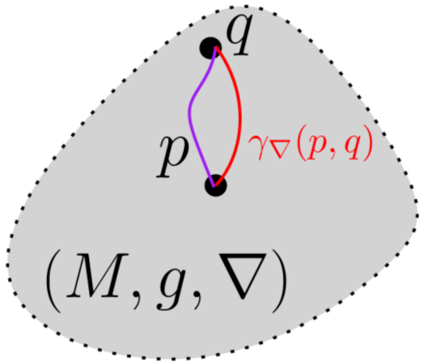
② metric tensor g



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 ∇



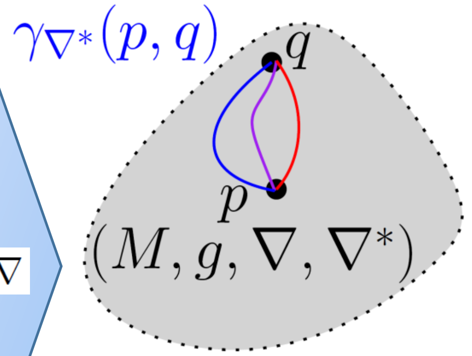
Examples:
exponential connection
mixture connection
metric connection ∇^g
divergence connection ∇^D
α-connection

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

dual connection

$\nabla^* = 2\nabla^g - \nabla$

Get dual IG manifold (M,g,∇,∇^*)



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

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

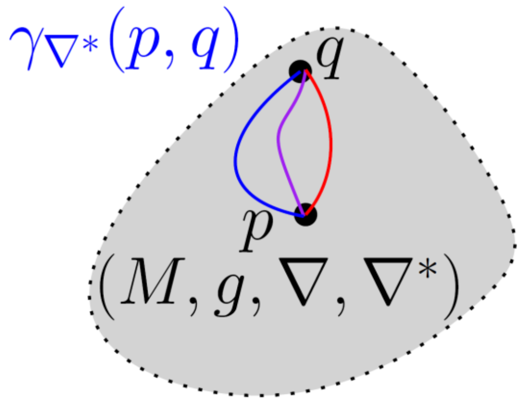
④ choose α

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,∇,∇*)



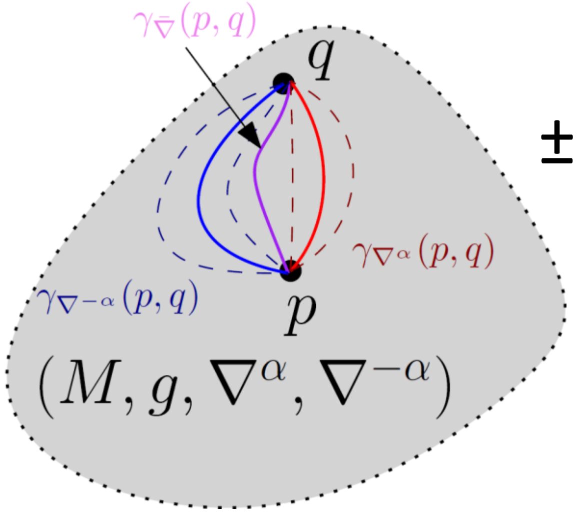
$$\nabla^g = \frac{\nabla + \nabla^*}{2} = \bar{\nabla}$$

Cubic tensor

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

Get a family of dual connections/IG
(M,g,∇^α,∇^-α)

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



±α-geometry

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

0-geometry
= Riemannian geometry
with geodesic distance

Information geometry from parametric statistical models

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

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

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

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

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

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

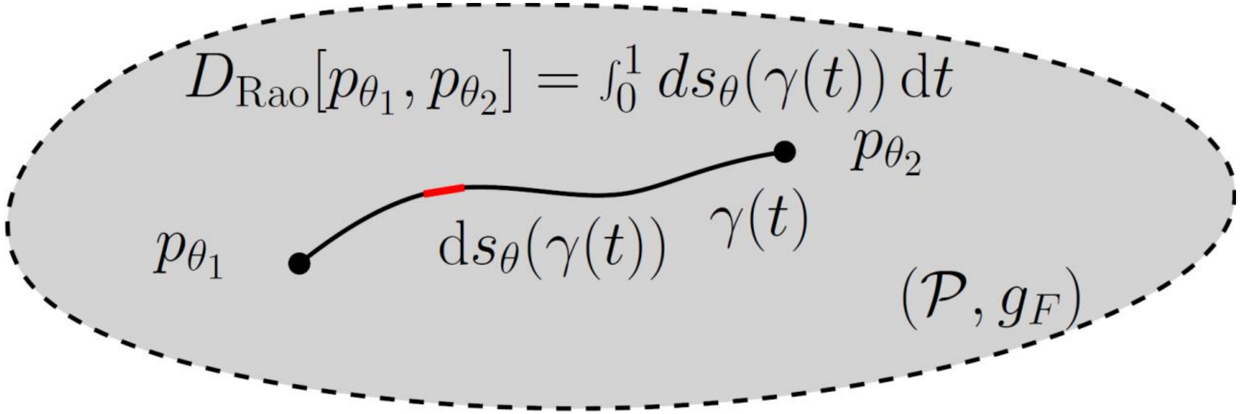
$$= \int_0^1 ds_\theta(\gamma(t))\,\mathrm{d}t$$

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

Length element

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

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



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 for MultiVariate Normals !

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

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

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

- 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:

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

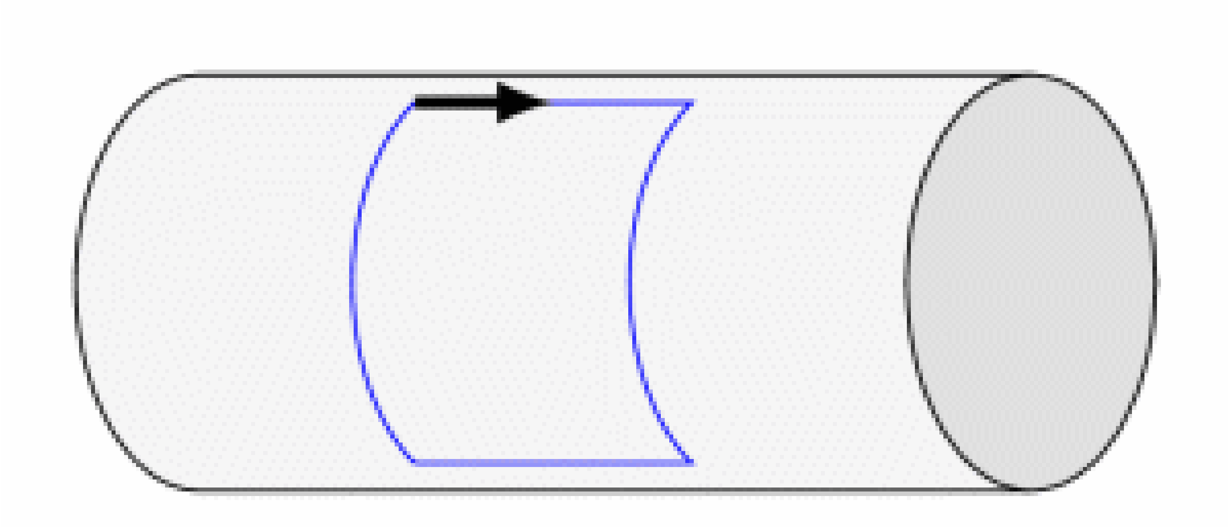
$${}^D\nabla^* = {}^{D^*}\nabla.$$

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

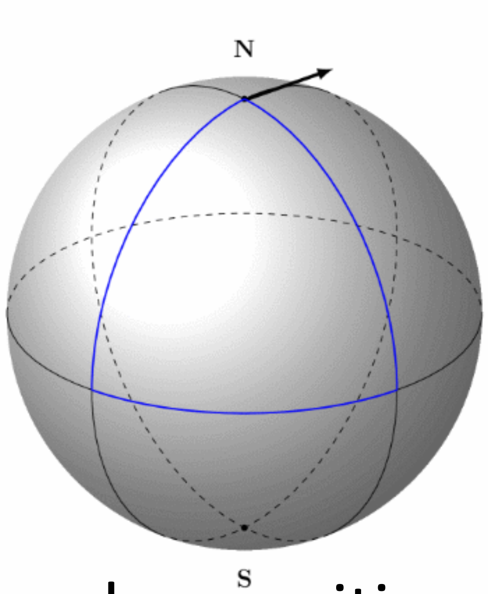
$$\partial_{i,\cdot} f(x,y) = \frac{\partial}{\partial x^i} f(x,y). \quad \partial_{\cdot,j} f(x,y) = \frac{\partial}{\partial y^j} f(x,y),\ \partial_{ij,k} f(x,y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial}{\partial y^k} f(x,y)$$

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 ED 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 constant κ

- <u>Corollary</u>: 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 line segments

Γ(θ)=0

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

geodesics=line segments in θ

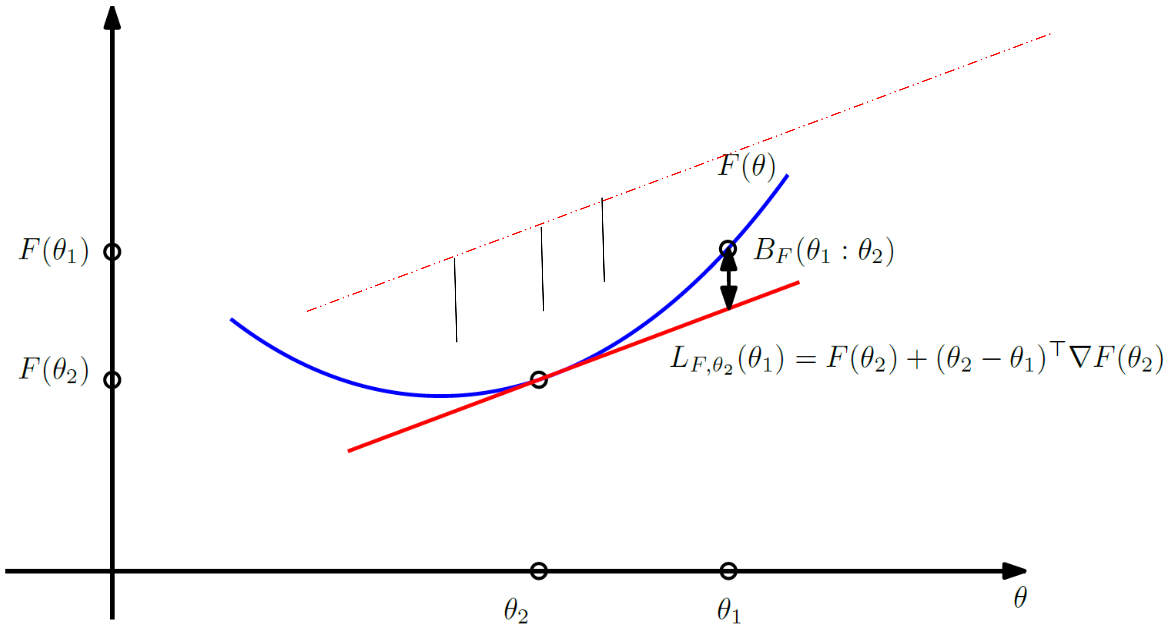
Canonical divergences of DFSs: Bregman divergences

- 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 (θ_1, F(θ_1)) and the linear approximation of F(θ) 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

- 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:

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

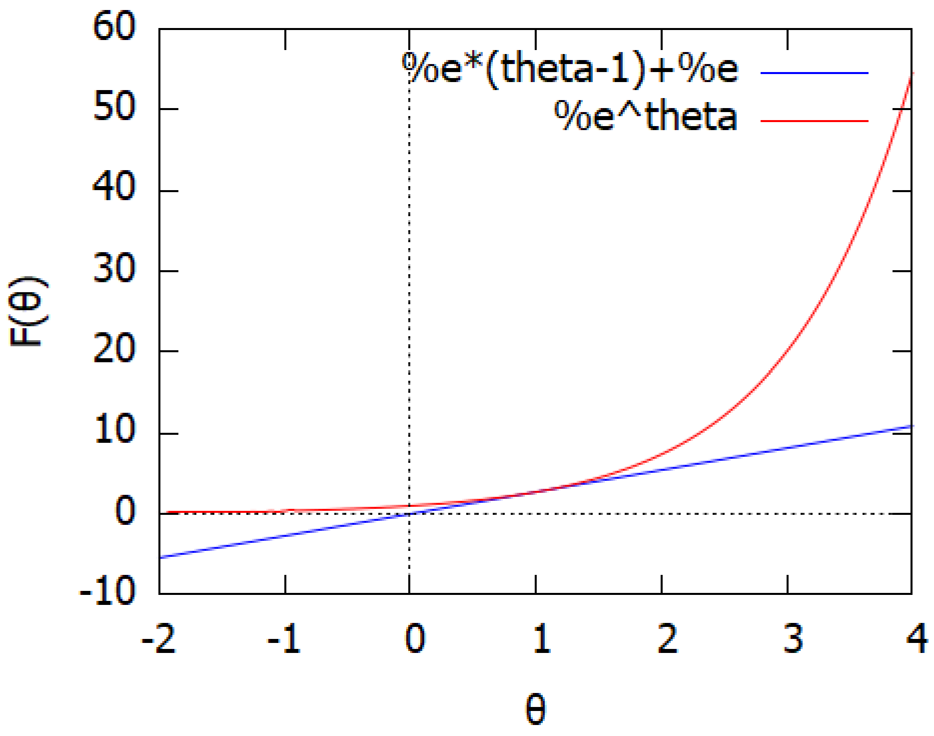
$$\nabla E(\theta) = \eta - \nabla F(\theta) = 0 \Rightarrow \eta = \nabla F(\theta)$$

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

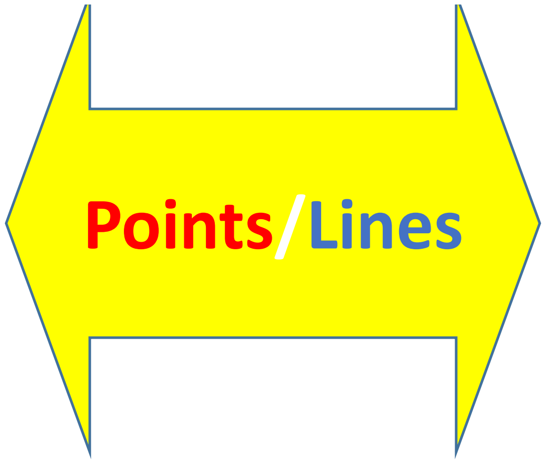
[Touchette 2005] Legendre-Fenchel transforms in a nutshell
[N 2010] Legendre transformation and information geometry

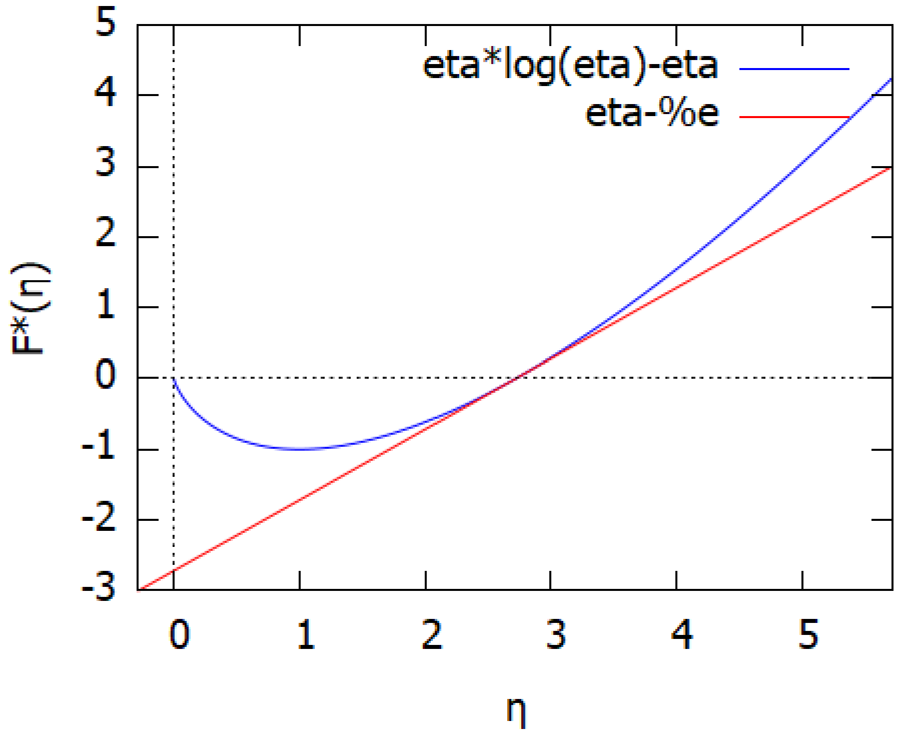
Reading the Legendre-Fenchel transformation

- Legendre-Fenchel transformation also called the **slope transform**



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





$$\begin{aligned} F(\theta) &= \exp(\theta) \\ \eta &= F'(\theta) = \exp(\theta) \\ \theta &= F'^{-1}(\eta) = \log\eta = F^{*\prime}(\eta) \\ F^*(\eta) &= \theta\eta - F(\theta) = \eta\log\eta - \eta \end{aligned}$$

(Here, F was chosen as the cumulant function of the Poisson distributions)

Mixed coordinates and the Legendre-Fenchel divergence

- Dual **Legendre-type** functions $\theta = \nabla F^*(\eta) \Longleftrightarrow \eta = \nabla F(\theta)$
- Convex conjugate of F is $F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta))$
- **Fenchel-Young inequality** :

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

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

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

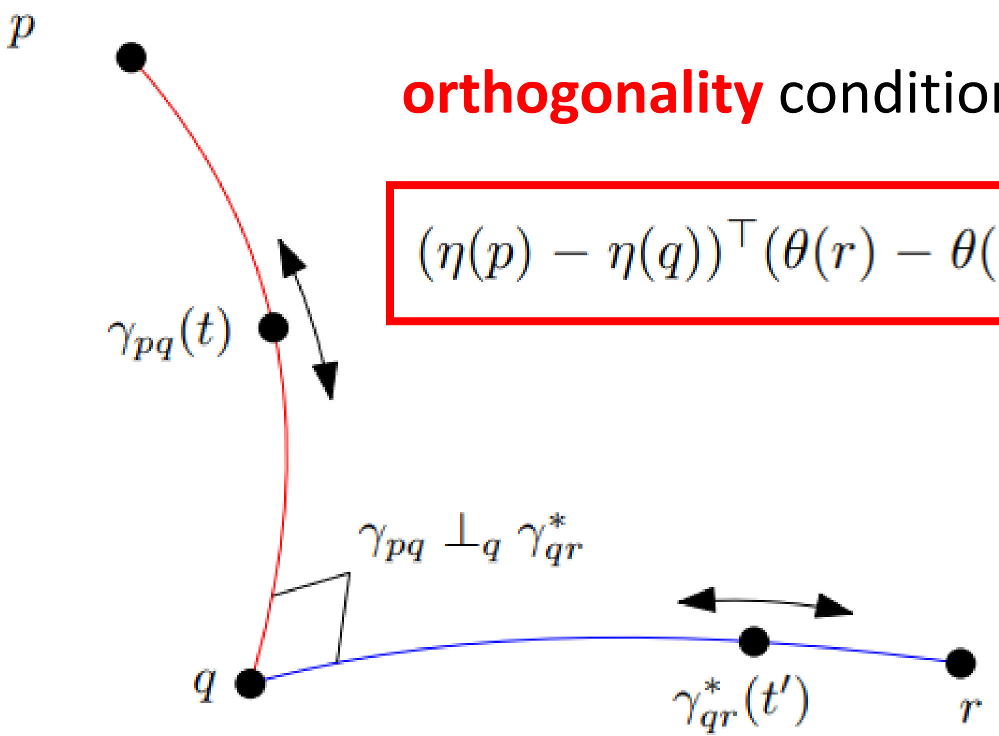
Gradient are inverse of each other

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

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

Generalized Pythagoras theorem in dually flat spaces

Generalized Pythagoras' theorem



orthogonality condition:

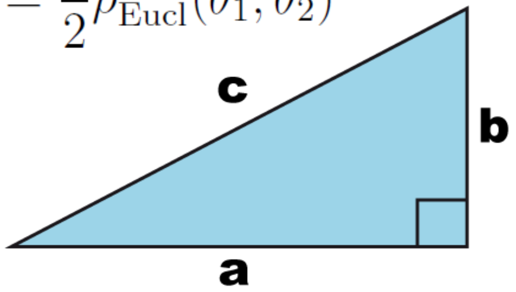
$$\boxed{(\eta(p) - \eta(q))^\top (\theta(r) - \theta(q)) = 0}$$

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

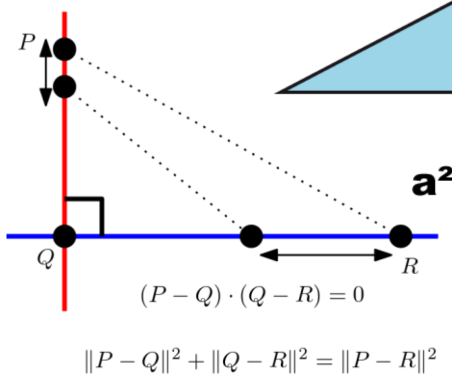
Pythagoras' theorem in the Euclidian geometry (Self-dual)

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

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



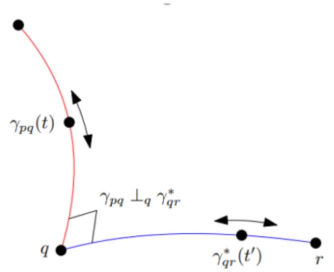
$a^2 + b^2 = c^2$



Identity of Bregman divergence with three parameters = law of cosines

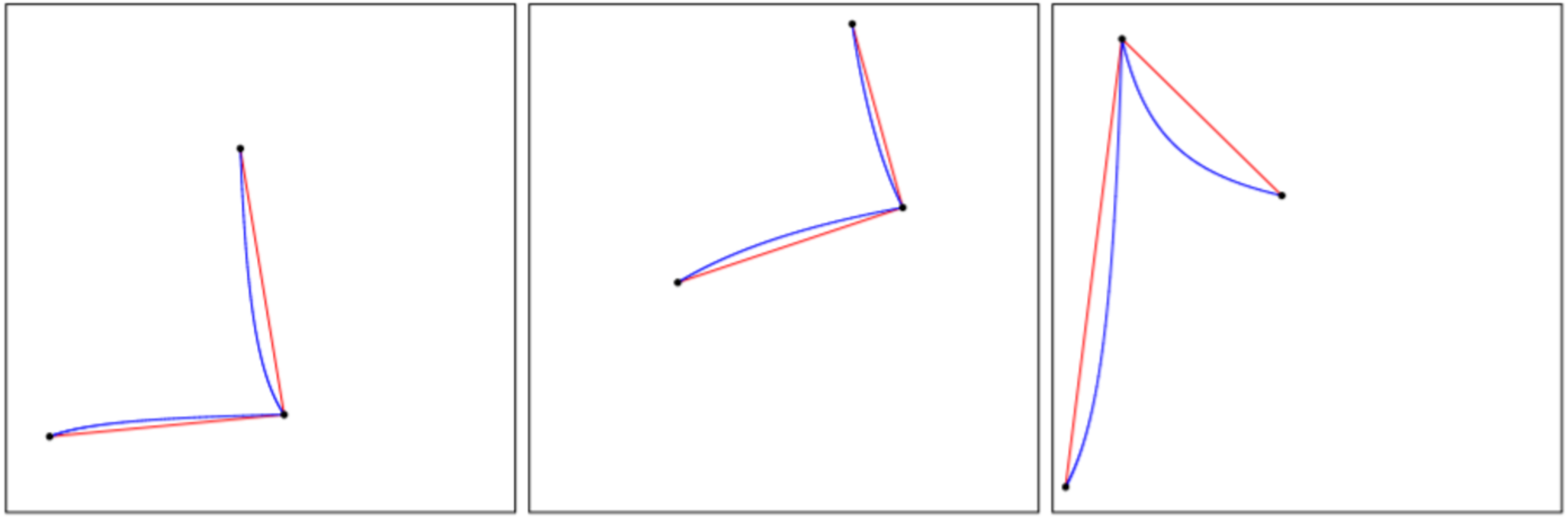
$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top(\nabla F(\theta_2) - \nabla F(\theta_3)) \geq 0$$

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

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

Domain

dual Bregman divergences

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

- Examples of DFSs induced by convex functions: