Information geometry: Geometry of dual structures

- A very short introduction -

Frank Nielsen

Sony Computer Science Laboratories Inc

Tokyo, Japan

Sony CSL

2023

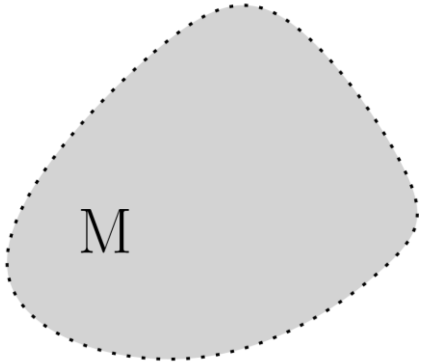
Information geometry (IG): Rationale and scope

- 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
- **Pure geometric dual structures** which can be used in many different contexts

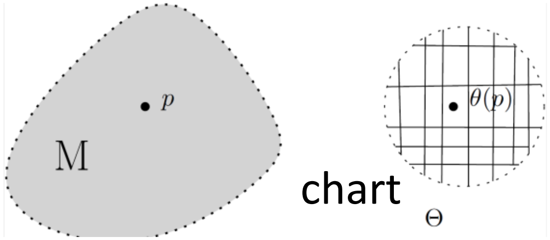
Build your own information geometry in three steps

Choose

① manifold M

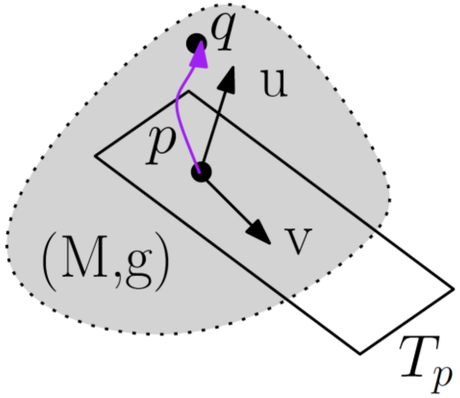


Examples:
Gaussians
SPD cone
Probability simplex



Concepts:
local coordinates
locally Euclidean

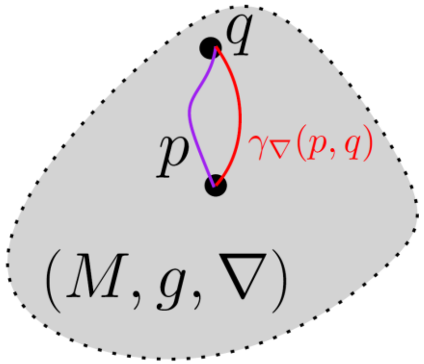
② metric tensor g



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

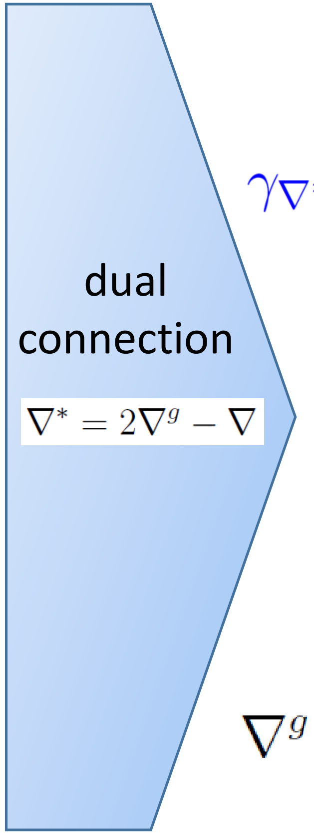
Concepts:
vector length
vector orthogonality
Riemannian geodesic
Riemannian distance
Levi-Civita connection ∇^g

③ affine connection ∇

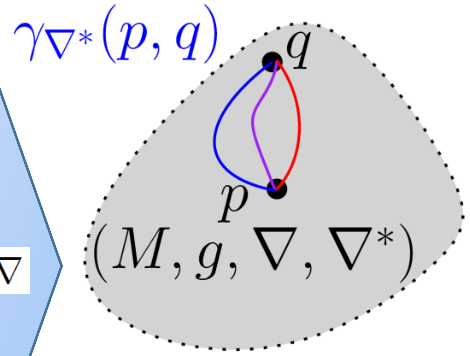


Examples:
exponential connection
mixture connection
metric connection ∇^g
divergence connection ∇^D
α-connection

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}

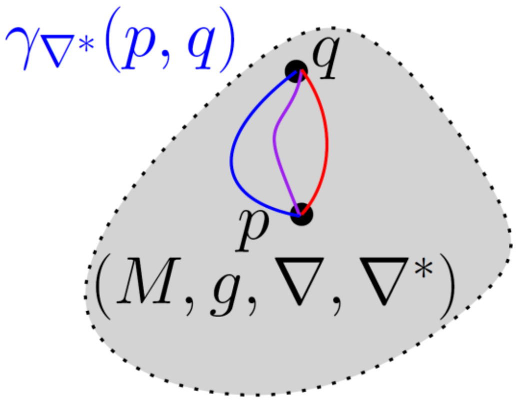
④ choose α

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

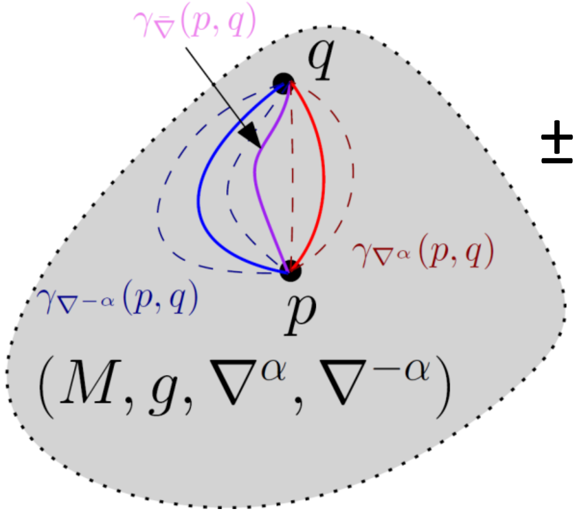
$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}$$

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

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

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

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



±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**: $C_{ijk} := E_\theta\left[\partial_i l \partial_j l \partial_k l\right] \Rightarrow \boxed{\{(\mathcal{P}, {_\mathcal{P}g}, {_\mathcal{P}\nabla^{-\alpha}}, {_\mathcal{P}\nabla^{+\alpha}})\}_{\alpha\in\mathbb{R}}}$
- **α-connections** $\nabla^\alpha = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m$

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

α=1 ← **exponential connection (e)**

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

α=-1 ← **mixture connection (m)**

- **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

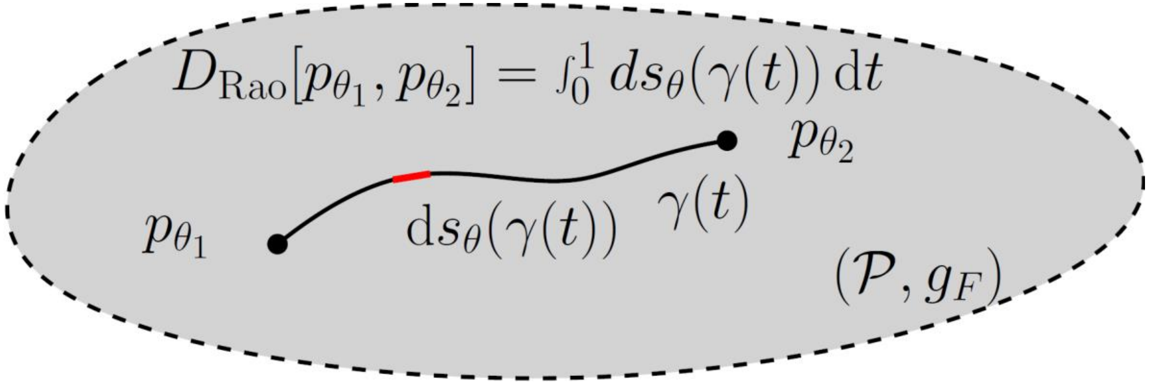
$$D_{\mathrm{Rao}}[p_{\theta_1}, p_{\theta_2}] = \rho_g(\theta_1, \theta_2) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, \mathrm{d}t, \gamma(0) = \theta_1, \gamma(1) = \theta_2$$
$$= \int_0^1 ds_\theta(\gamma(t))\, \mathrm{d}t$$

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

square 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 calculate geodesics which are curves locally minimizing the length linking two endpoints (or equivalently minimize the energy of squared length elements)
- Finding Fisher-Rao geodesics is a non-trivial tasks.
- **New in 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

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

Hellinger divergence, chi-square divergence
f-divergence, α-divergence, etc.

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

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

- 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 C or T:

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

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

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

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

Realizations of dual information geometry

- Consider a statistical manifold structure (M,g,C) or equivalently (M,g,∇,∇)
- 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 <u>diagonal of the product manifold</u>." *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 $(M,g^D,\nabla^D,\nabla^{D*})$

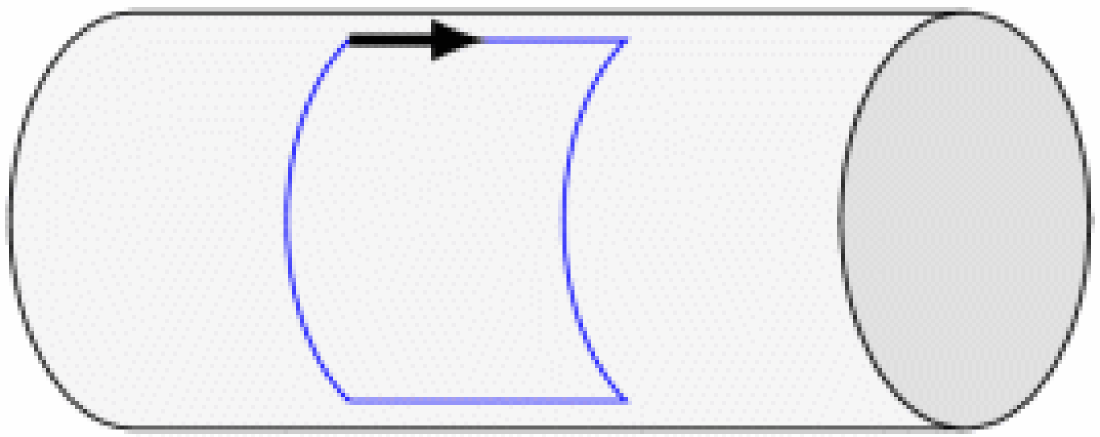
- Then **model α-geometry** for α=2 f'''(1)+3 coincide with **divergence IG**:

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

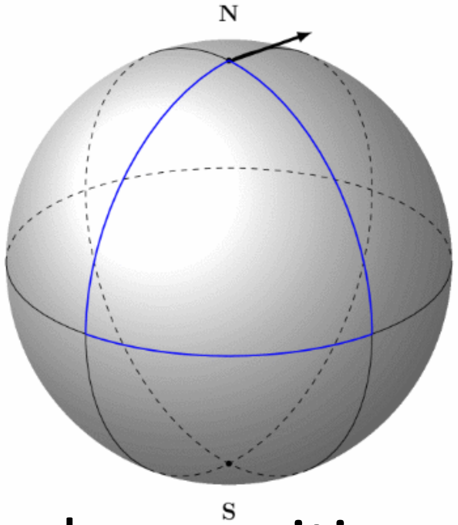
metric tensor g^D and cubic tensor T^D coincides with Fisher metric g^F and Amari-Chentsov tensor T

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
(PT of flat connection is path independent)



Sphere has positive constant curvature:
Parallel transport along a loop exhibits an angle defect related to curvature
(PT is path dependent)

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 ($\kappa=0$) then ∇^* is flat → **Dually flat space (M,g,∇,∇^*)**

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

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

$\Gamma(\theta)=0$

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

→ geodesics=line segments in θ

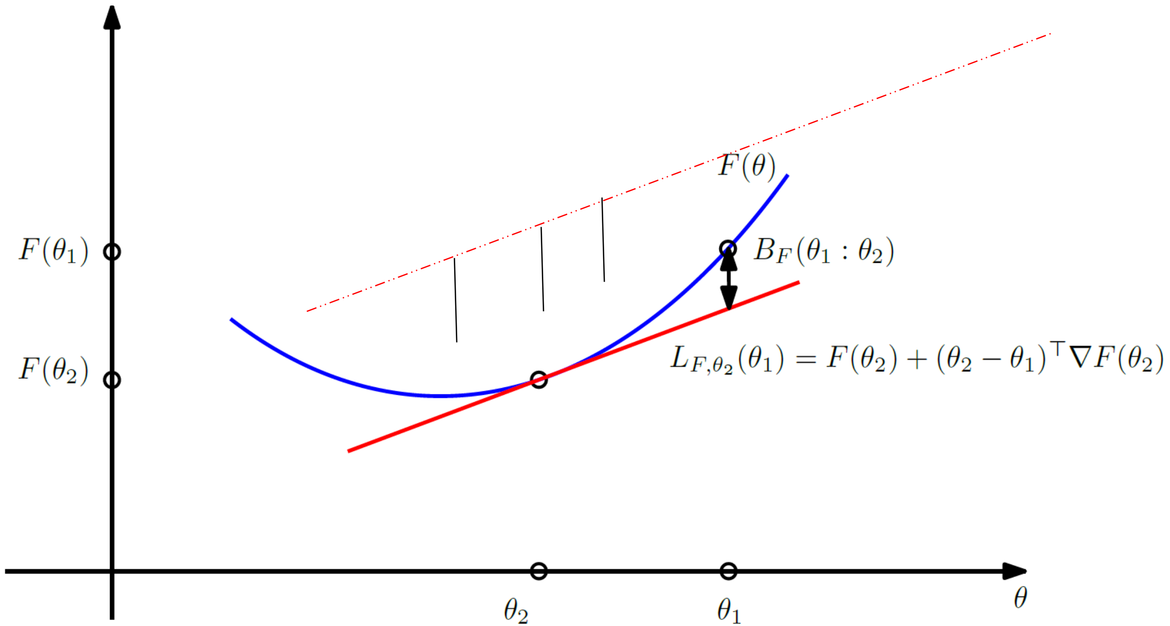
Canonical divergences of DFSs: Bregman divergences

- 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 :



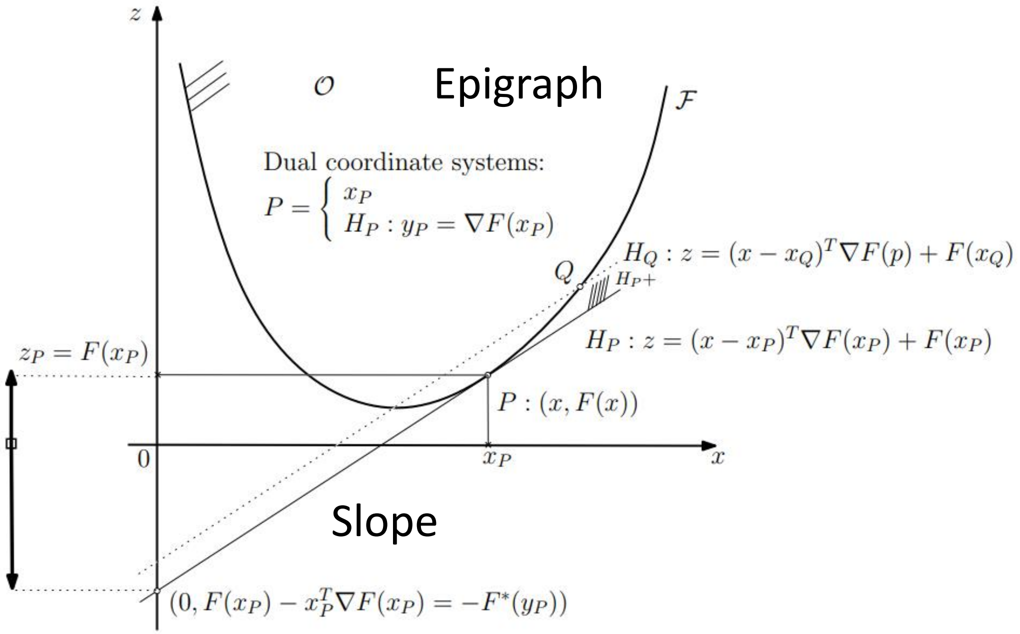
$$\begin{aligned} B_F(\theta_1 : \theta_2) &= F(\theta_1) - \underbrace{\left(F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2)\right)}_{L_{F,\theta_2}(\theta_1)} \\ &= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \end{aligned}$$

[Bregman 1967]

Legendre-Fenchel transformation: Slope transformation

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

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



<u>Concave programming:</u>

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

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

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

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

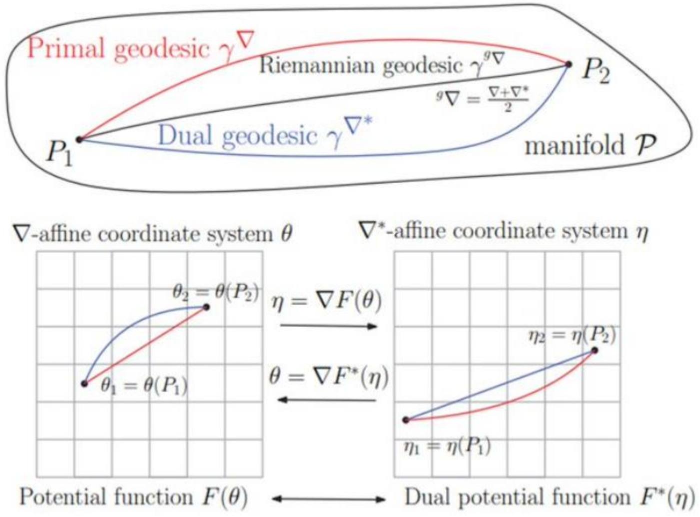
with equality holding if and only if $\eta_2 = \nabla F(\theta_1)$

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

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

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

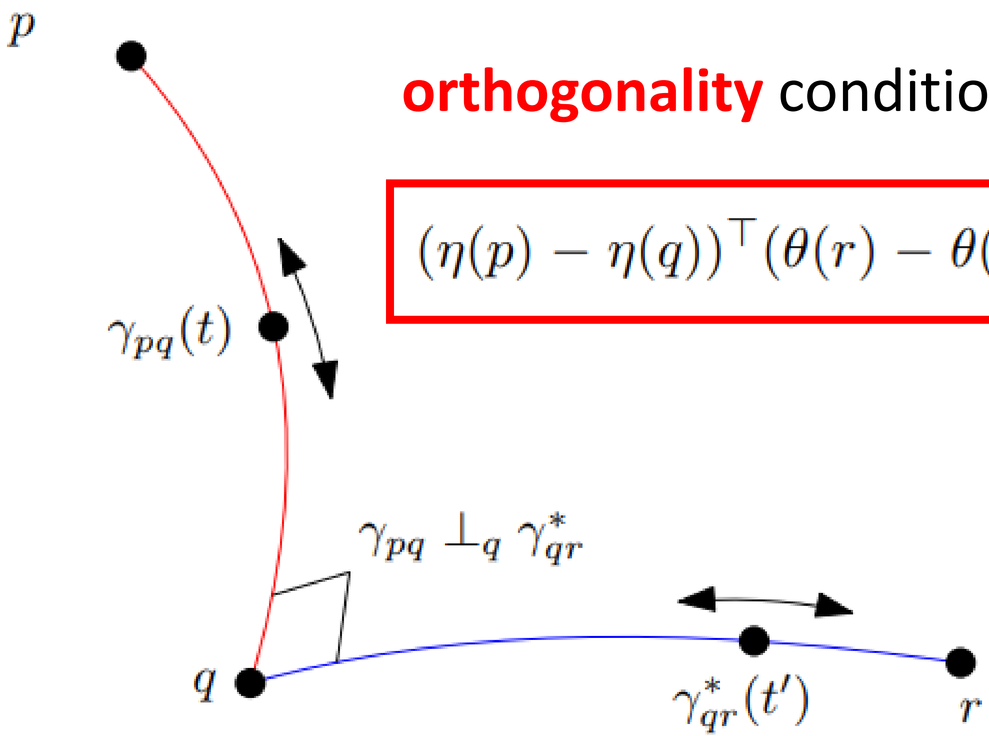


Generalized Pythagoras theorem in dually flat spaces

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:

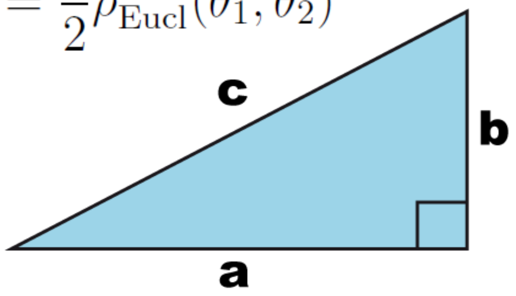
$$\boxed{(\eta(p)-\eta(q))^\top(\theta(r)-\theta(q))=0}$$

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

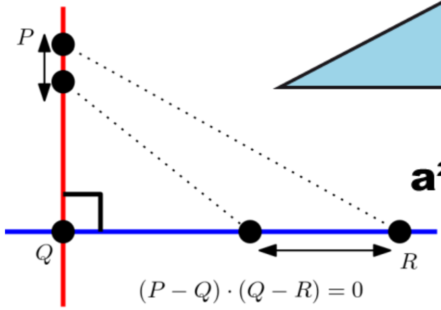
Pythagoras' theorem in **the Euclidian geometry (Self-dual)**

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

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



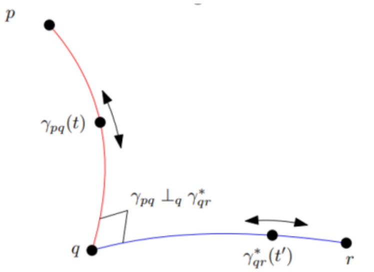
a² + b² = c²



$(P-Q)\cdot(Q-R)=0$

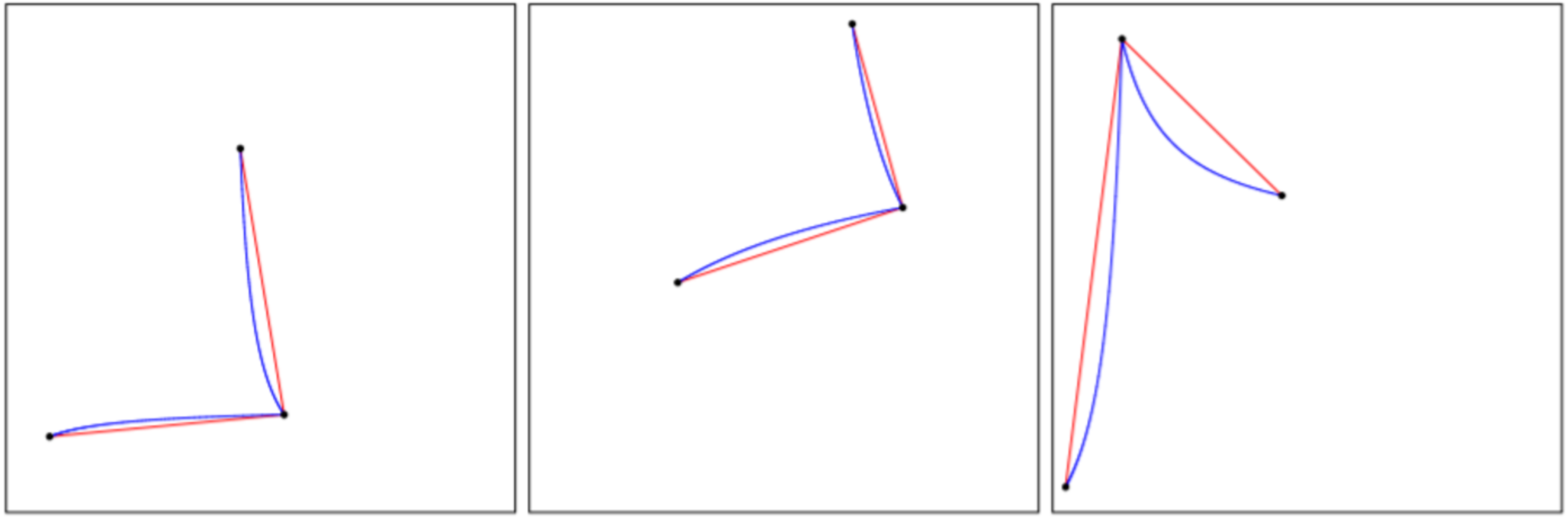
$\|P-Q\|^2+\|Q-R\|^2=\|P-R\|^2$

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



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

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



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

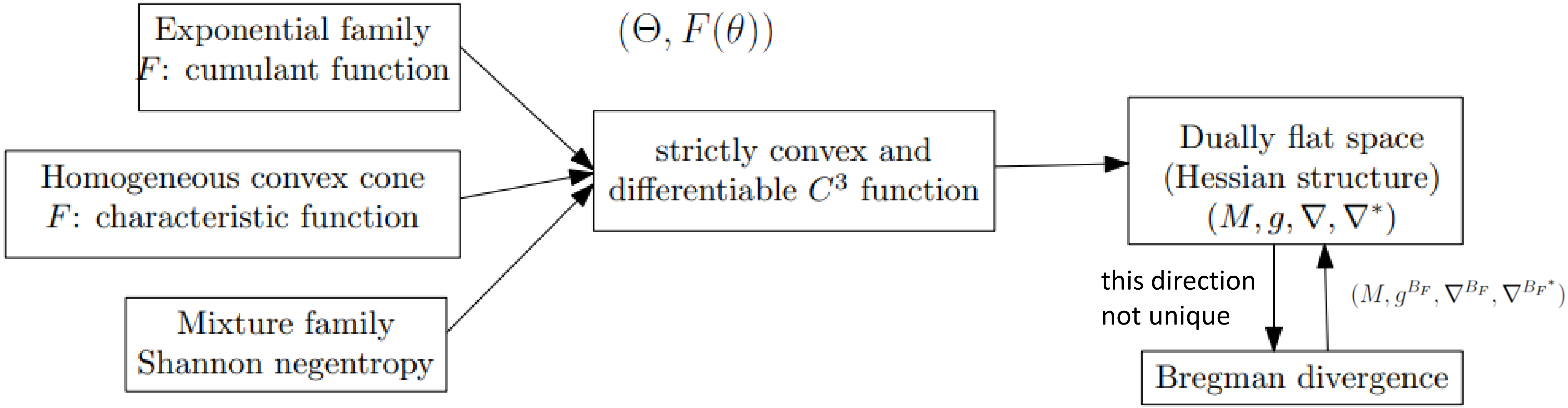
$$(\Theta, F(\theta)) \longrightarrow (M, g^{B_F}, \nabla^{B_F}, \nabla^{{B_F}^*}) = (M, g^F, \nabla^F, \nabla^{F^*})$$

Domain

dual Bregman divergences

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

- Examples of DFSs induced by convex functions:



Some references

- Amari, Shun-ichi, and Hiroshi Nagaoka. ***Methods of information geometry***. Vol. 191. American Mathematical Soc., 2000.
- Amari, Shun-ichi. ***Information geometry and its applications***. Vol. 194. Springer, 2016.
- Calin, Ovidiu, and Constantin Udrişte. ***Geometric modeling in probability and statistics***. Vol. 121. Berlin, Germany:: Springer, 2014.
- Nielsen, Frank. "An elementary introduction to information geometry." *Entropy* 22.10 (2020): 1100.
- Nielsen, Frank. "Legendre transformation and information geometry." *no. CIG-MEMO2* (2010).
- Nielsen, Frank. "On geodesic triangles with right angles in a dually flat space." *Progress in Information Geometry: Theory and Applications*. Springer, 2021. 153-190.