Introduction to Information Geometry

Frank NIELSEN

July 2022

Sony CSL

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

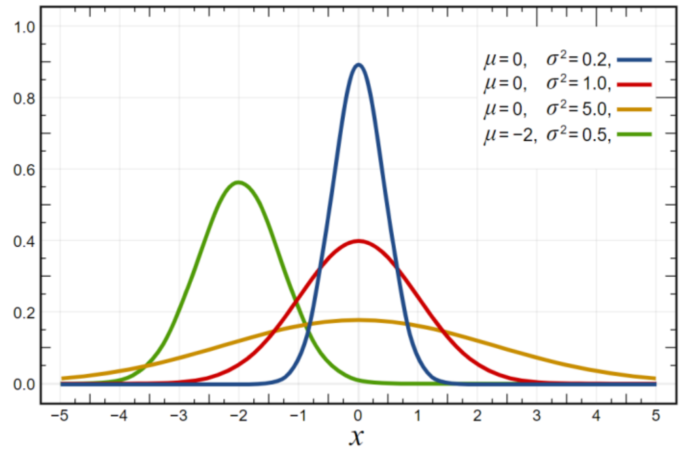
What is information geometry? (1/4)

Consider a set of parametric probability distributions called the **statistical model**

For example, the set of normal distributions with mean **μ** and variance **σ²**

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \lambda = (\mu, \sigma) \in \mathbb{H} \right\}$$

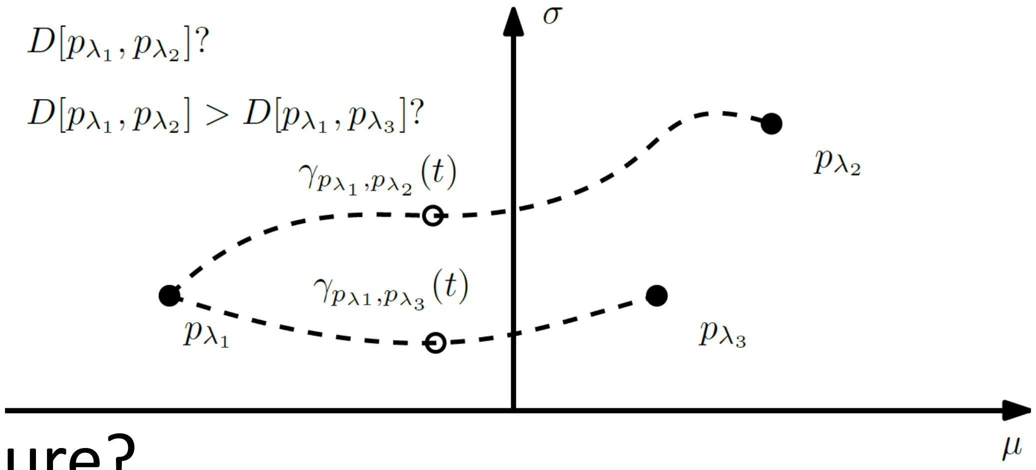
Parameter space Λ is the *upper plane* $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$



What kinds of **geometric structures** for this family of normal laws?

A few related questions:

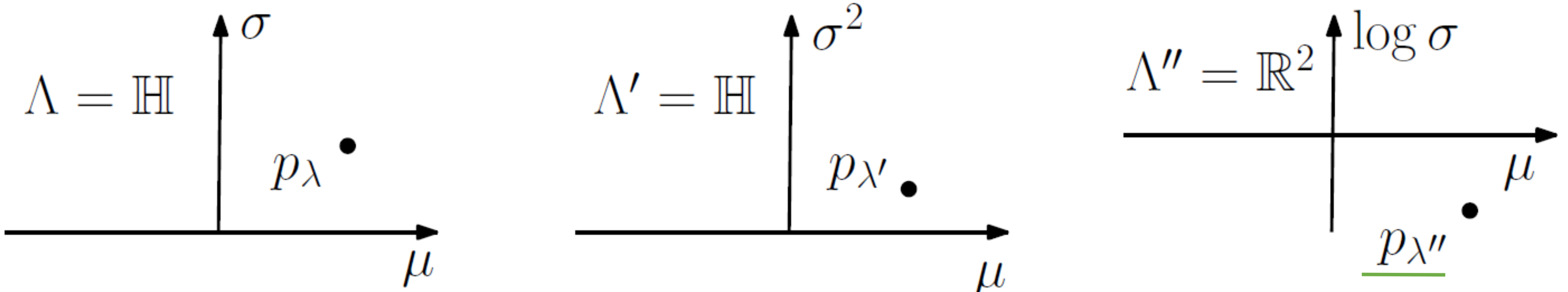
- How to **interpolate** between two normals?
- How to define **distances** between them?
- Are they **several ways** to proceed?
  If so why? And how to choose the right structure?



What is information geometry? (2/4)

- Which **invariance principles** shall be satisfied by the geometric structures and the distances between statistical models
- **First invariance principle**: If we parameter Gaussians by **(μ,σ^2)** or **(μ,log(σ))** instead of **(μ,σ)**, it should not change their distances nor the interpolating paths called ``**geodesics**''

Euclidean distance between parameters not invariant!



Same family of Gaussians $\boxed{\mathcal{P}} = \left\{ p_{\lambda''}(x) = \frac{1}{\sqrt{2\pi}\exp(\lambda''_2)} \exp\left(-\frac{(x-\lambda''_1)^2}{2\exp(2\lambda''_2)}\right), \quad \lambda'' = (\mu, \log\sigma) \in \mathbb{R}^2 \right\}$

Thus we need these **two invariance properties**:

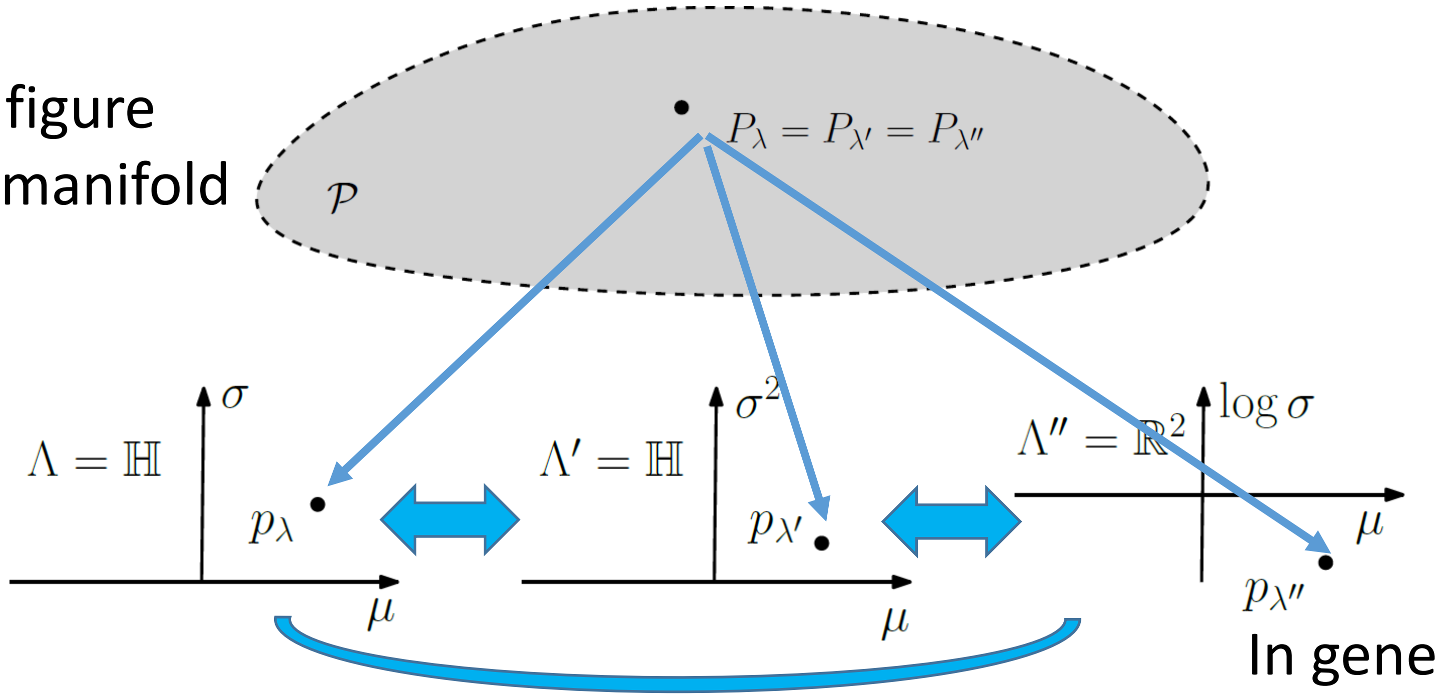
① $D[p_{\lambda_1}, p_{\lambda_2}] = D[p_{\lambda'_1}, p_{\lambda'_2}] = D[p_{\lambda''_1}, p_{\lambda''_2}]$

② $\gamma_{p_{\lambda_1}, p_{\lambda_2}}(t) = \gamma_{p_{\lambda'_1}, p_{\lambda'_2}}(t) = \gamma_{p_{\lambda''_1}, p_{\lambda''_2}}(t), \forall t \in [0,1]$

Differential geometry of statistical models

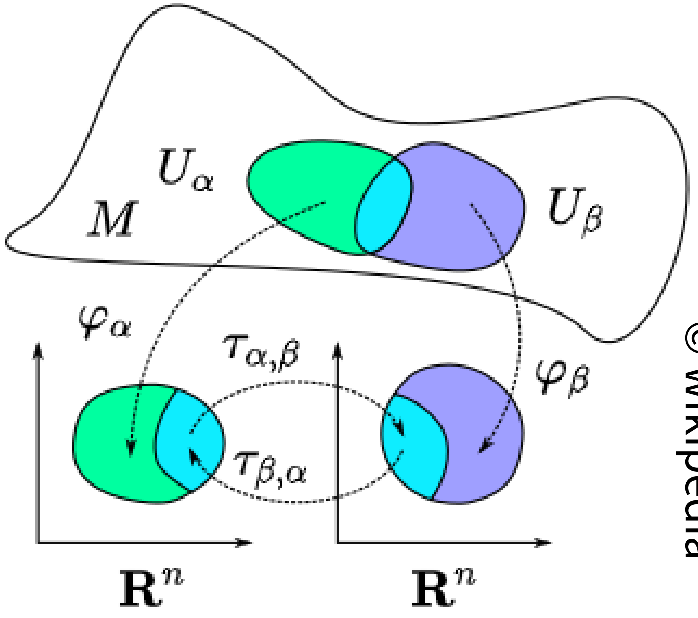
- To each point of the manifold corresponds a unique parametric distribution:
- Statistical model is **identifiable** when $\lambda \leftrightarrow P_\lambda$
- Often a single global <u>**chart**</u> = atlas which covers the parameter domain

Abstract figure depicting a manifold



Domains

Several global charts (atlas with a single chart)



In general, need several local charts to cover a manifold
(eg., two charts for the sphere)

What is information geometry? (3/4)

Information geometry: study geometric structures on the manifold induced by identifiable statistical models

- Use language of geometry: **geodesics**, **balls**, **information projection, statistical curvature** and the **tensor calculus**. This tensor calculus made possible to study the efficiency of statistical estimators to higher order.
- Study the principles of **invariance** in statisticcs
- The new dual geometric structure can also be used beyond the statistical scope (pure geometry). For example, the information-geometric structures have been used to analyzed interior point methods in optimization

- Information geometry was born from the mathematical consideration of the **Fisher metric** and its induced **geodesic distance** to solve classification and statistical hypothesis test tasks

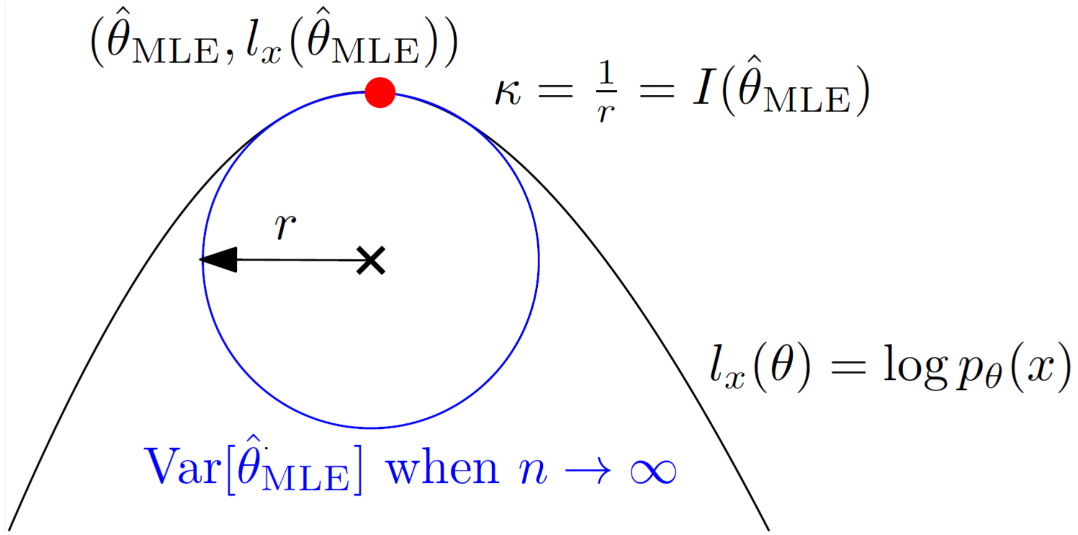
[Mahalanobis 1936] [Hotelling 1930] [Rao 1945]

Fisher information $I_X(\theta)$

- Consider a parametric family of laws indexed by D parameters **Fisher Information Matrix** (FIM) = covariance matrix of the **score**

$$X \sim p_\theta(x) \qquad s_X(\theta) = \nabla_\theta \log p_\theta(x) \qquad \boxed{I_X(\theta) = \mathrm{Cov}[s_\theta]}$$

- FIM is symmetric and positive semi-definite (could be undefined too)
- FIM is said **regular** when positive-definite (yields the Fisher metric on manifolds)
- Interpreted as the **curvature** of the log-likelihood function:



Asymptotic normality of the MLE (Cramér-Rao lower bound)

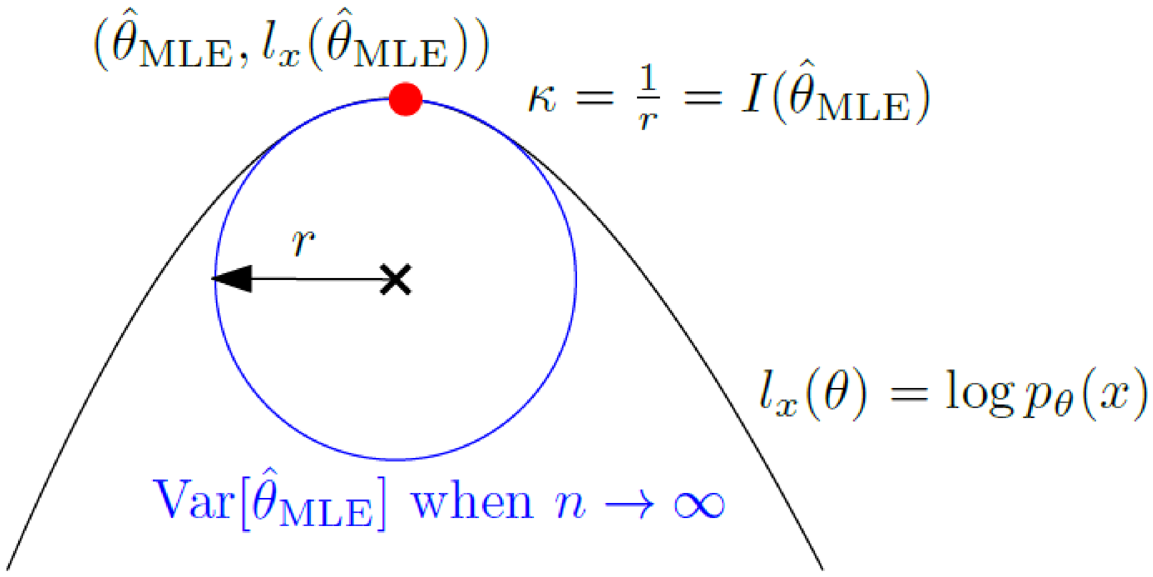
For a general function f(x), the curvature is defined as

$$\kappa(x) = \frac{f''(x)}{(1+(f'(x))^2)^{\frac{3}{2}}}, \qquad r(x) = \frac{(1+(f'(x))^2)^{\frac{3}{2}}}{|f''(x)|}$$

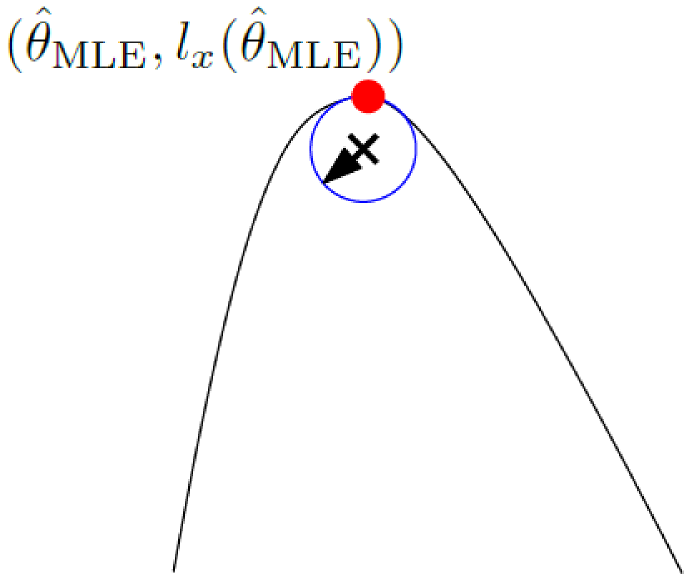
MLE: Maximum Likelihood Estimator

Radius of the **osculating circle** corresponds
To the inverse of the absolute curvature

Fisher information $I_x(\theta)$



Fisher information small:
Likelihood curvature is small (flat peak)
Variance is large, low accuracy



Fisher information large:
Likelihood curvature is large (sharp peak)
Variance is small, good accuracy

Taylor 2nd-order with $l'(\hat{\theta}_{\mathrm{MLE}}) = 0$: $l_x(\theta) \approx l_x(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{2}(\theta - \hat{\theta}_{\mathrm{MLE}})^2 l''(\hat{\theta}_{\mathrm{MLE}})$

$$-l''(\hat{\theta}_{\mathrm{MLE}}) = -E_{\hat{\theta}_{\mathrm{MLE}}}[l''(\theta)] = I(\hat{\theta}_{\mathrm{MLE}}) \Rightarrow l_x(\theta) \approx l_x(\hat{\theta}_{\mathrm{MLE}}) - \frac{1}{2}(\theta - \hat{\theta}_{\mathrm{MLE}})^2 I(\hat{\theta}_{\mathrm{MLE}})$$

Two usual expressions of the Fisher information

- Using the first two **Bartlett identity** under the regularity condition that we can exchange k times the differentiation with the integration operations, we get

(Bartlett k) $$\nabla^k E_\theta[\exp(l_x(\theta))] = E_\theta\left[\nabla^k \exp(l_x(\theta))\right] = \nabla^k E_\theta[p_\theta] = \nabla^k 1 = 0$$

- Allows to rewrite the FIM under it is most famous forms:

$I_X(\theta) = \mathrm{Cov}[s_\theta]$ $\quad \mathrm{Cov}[s_\theta] = E[s_\theta s_\theta^\top] - E[s_\theta]E[s_\theta]^\top$

① **First form**: $$E[\nabla l_x(\theta)] = 0 \Rightarrow I_X(\theta) = E_\theta\left[\nabla \log p_\theta(x)(\nabla \log p_\theta(x))^\top\right]$$

(Bartlett k=1)

② **Second form** (negative of the Hessian of the log-likelihood) :

$$E_\theta\left[\nabla \log p_\theta(x)(\nabla \log p_\theta(x))^\top\right] + E\left[\nabla^2 \log p_\theta(x)\right] = 0 \Rightarrow I_X(\theta) = -E[\nabla^2 \log p_\theta(x)]$$

(Bartlett k=2)

$$I_X(\theta) = -E_{p_\theta}\left[\nabla^2 l_x(\theta)\right] = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l_x(\theta) l_x(\theta)\right]_{ij}$$

Mahalanobis and his generalized distance

- Motivated by the statistical analysis of human skulls collected in various regions. Each skull is characterized by d.

- Mahalanobis (1928, 1936) introduced the following D2 statistics and **divergence** entre deux groupes S_1 et S_2 :

$$\Delta^2[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] = (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)$$

Precision matrix = inverse covariance

- Nowadays the **metric Mahalanobis distance**:

$$\Delta[p_{\mu_1,\Sigma}, p_{\mu_2,\Sigma}] = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)}$$

Generalized the Euclidian distance when Σ =I, the identity matrix

Divergence = smooth dissimilarity which may be asymmetric and may not satisfy the triangular inequality of metric

© wikipedia

P. C. Mahalanobis
(1893-1972)
Found of the
Indian Statistical Institute (ISI)

Vol. VIII. APRIL & SEPT. 1928. Nos. 2 & 3.

I.—A STATISTICAL STUDY OF THE CHINESE HEAD

BY

P. C. MAHALANOBIS

ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

Mahalanobis distances: Vector spaces equipped with an inner product

- Mahalanobis distance rewritten as

$$\Delta_\Sigma(\mu,\mu') = \|\mu - \mu'\|_{\Sigma^{-1}}$$

$$\|x\|_{\Sigma^{-1}} = \sqrt{x^\top \Sigma^{-1} x}$$

- For a symmetric positive-definite matrix (SPD) Q, we define the inner product by the following bilinear form:

$$\langle v_1, v_2\rangle_Q = v_1^\top Q v_2$$

- Inner product induces a **norm** which in turn induces a metric **distance**:

$$\langle v_1, v_2\rangle_E \rightarrow \|v\|_E = \sqrt{\langle v, v\rangle} \rightarrow D_E(v_1, v_2) = \|v_1 - v_2\|_E$$

- Inner product allows us to define the **orthogonality** between two vectors (and their subtended angle) and the vector **lengths**:

$$v_1 \perp v_2 \leftrightarrow \langle v_1, v_2\rangle_Q = 0 \qquad \|v\| = \sqrt{\langle v, v\rangle}$$

- This geometry corresponds to the extrinsic geometry of tangent spaces of manifolds

Riemannian Fisher metric tensor field aka Fisher metric

- Consider the manifold $\mathcal{P} = \{p_\theta(x) \; : \theta \in \Theta\}$

g_F : smooth fields of inner products on tangent planes

$$g_F(u,v) = [u]_B^\top I(\theta)\,[v]_B$$

- vector **components** $[v]_B = (v^1, \ldots, v^D)$

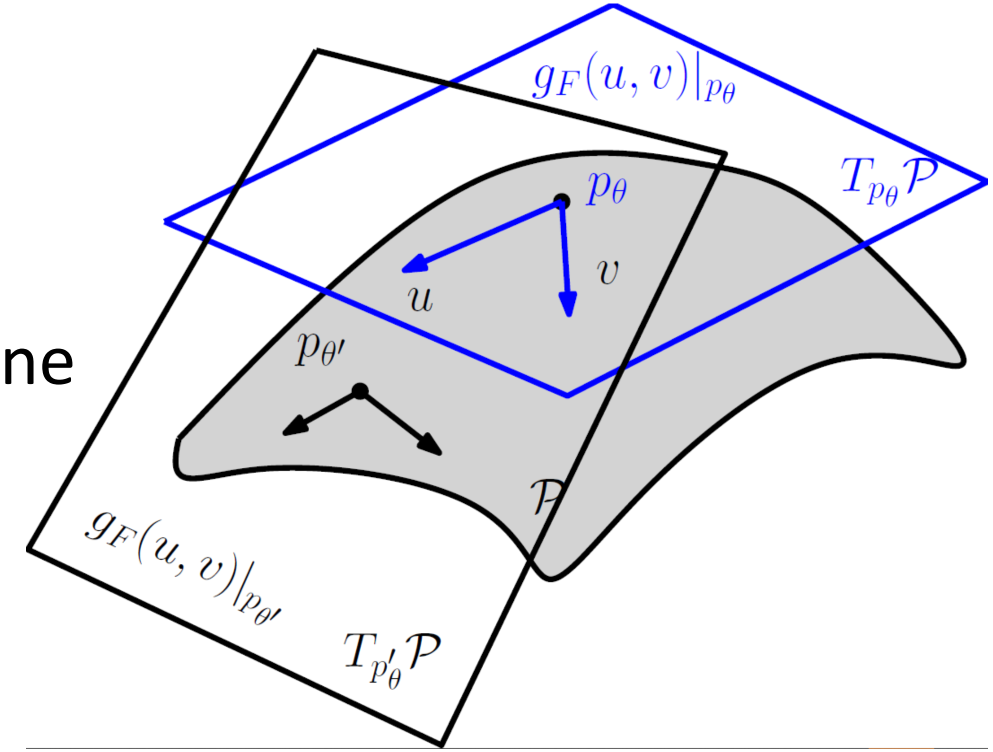
expressed in the natural basis of the tangent plane

Induced by the (local) chart θ(.) $\partial_i = \frac{\partial}{\partial \theta^i}$

$$B = \{e_1 = \partial_1, \ldots, e_D = \partial_D\}$$

$$g_{ij} = g_F(\partial_i, \partial_j) = I_{ij}(\theta)$$

$$g_F(u,v) = \sum_{i,j} g_{ij} u^i v^j = u^\top I(\theta)\, v$$

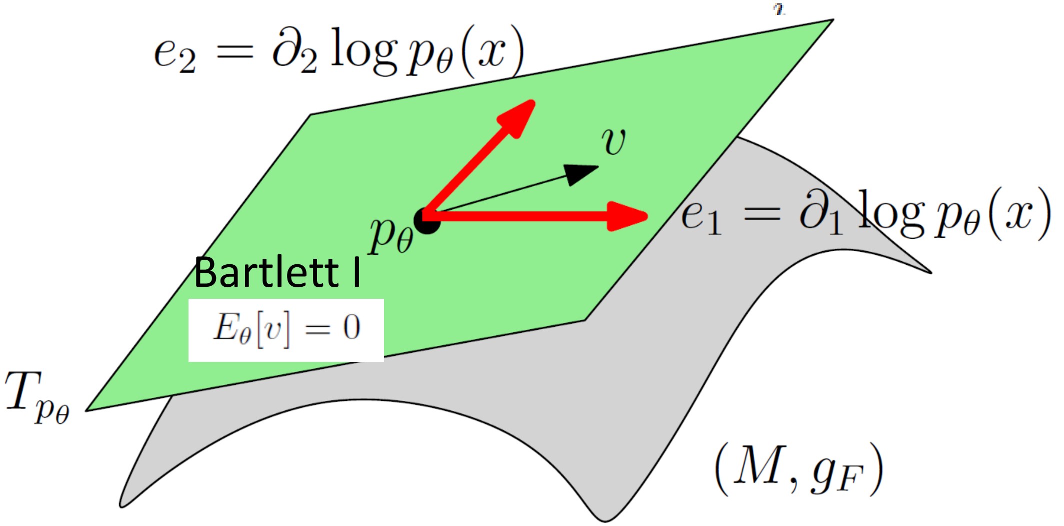


Tangent plane representation for a manifold induced by a statistical model: Reinterpret the inner product

- 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**:

$$T_\theta = T_{p_\theta} = \{\sum_i v^i \partial_i l_x(\theta)\}$$

$$B = \{e_1 = \partial_1 l_x(\theta), \ldots, e_D = \partial_D l_x(\theta)\}$$



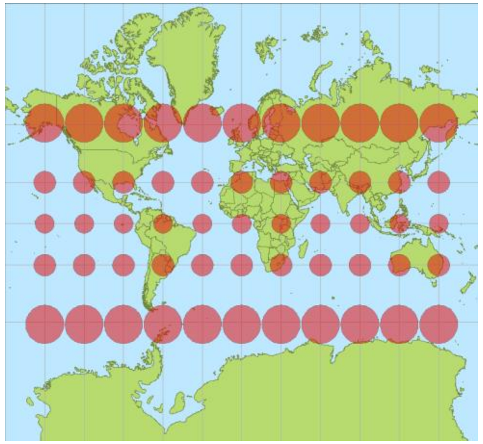
- The inner product can be reinterpreted as:

$$g_F(u, v) = E_\theta[u(x)v(x)] = \mathrm{Cov}(u(x), v(x))$$

$$\underbrace{g_F(\partial_i, \partial_j) = E_\theta[\partial_i l_x(\theta)\partial_j l_x(\theta)]}_{\text{Expectation}}$$

Visualizing the Fisher metric and the Cramér–Rao bound

- Fisher metric: $$g_F(u,v) = [u]_B^\top I(\theta)\,[v]_B$$
- Visualize I(θ) by an **ellipsoid**
- Visualise the metric tensor field by **Tissot indicatrix**



Mercator projection

© wikipedia

Visualizing the Cramer-Rao lower bound :

- For each grid position (μ,σ) :

- Sample iid N (μ,σ)
- Maximum likelihood estimator of (μ,σ)
- Repeat k times to get an empirical estimator of the covariance matrix of the 2D parameters.

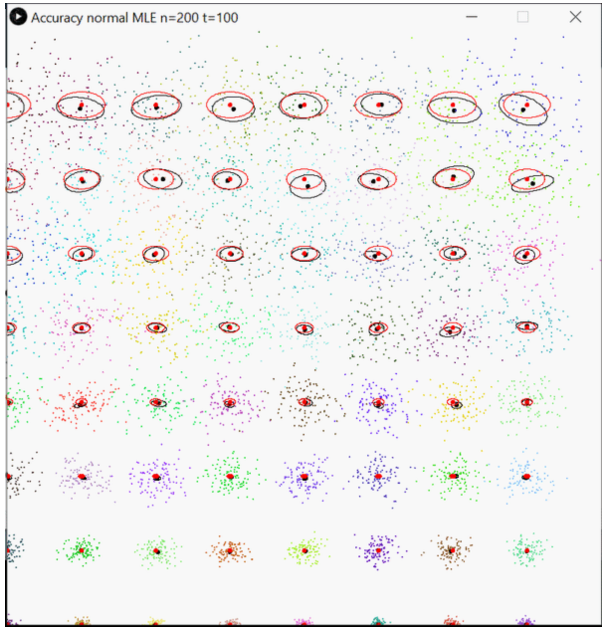
Converge to the scaled inverse FIM

Inverse of the FIM displayed with Tissot indicatrix

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta)^{-1}$$

$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

$$\mathrm{Cov}[\hat{\theta}_n]$$



Upper plane (μ,σ) $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$

Rao distance on the Fisher-Rao manifold

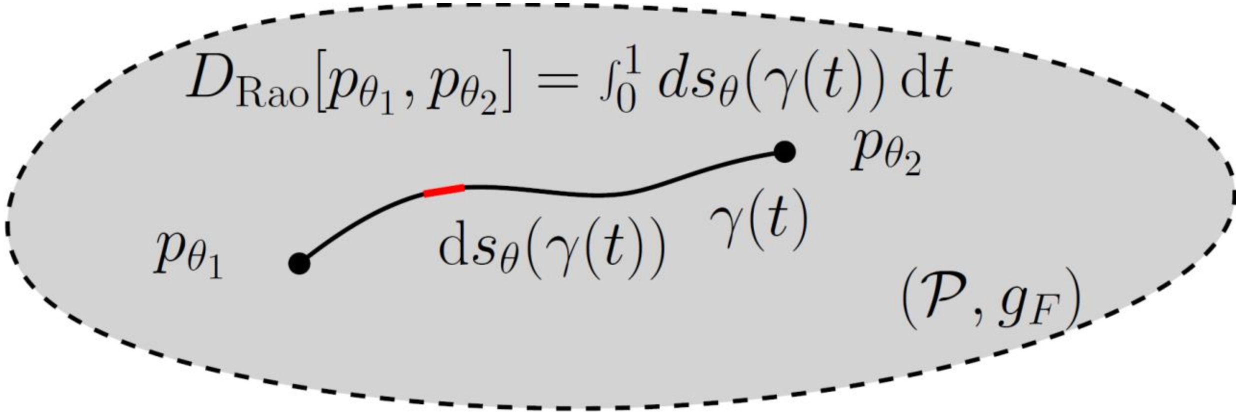
$$
\begin{aligned}
D_{\mathrm{Rao}}[p_{\theta_1}, p_{\theta_2}] &= \rho_g(\theta_1, \theta_2) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\,\mathrm{d}t, \gamma(0) = \theta_1, \gamma(1) = \theta_2 \\
&= \int_0^1 ds_\theta(\gamma(t))\,\mathrm{d}t
\end{aligned}
$$

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: No-known closed-form for the Fisher-Rao geodesic/distance between **multivariate Gaussians**!

Invariance under reparameterization of Rao's distance

Consider two different parameterizations of a statistical model:

$$\mathcal{P} = \{p_\theta \ : \ \theta \in \Theta\} = \{p_\eta \ : \ \eta \in H\}$$

Covariance transformation of the FIM under reparameterization

$$I_\theta(\theta) \xrightarrow{\eta=\eta(\theta)} I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial \theta_i}{\partial \eta_j}\right]$$

... However the length element is **invariant** : $\mathrm{d}s_\theta = \mathrm{d}s_\eta$

So that the Fisher-Rao distance is **invariant** : $\boxed{\rho_{\mathrm{Rao}}(p_{\eta_1}, p_{\eta_2}) = \rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})}$

- This is **the first principle of invariance of information geometry**

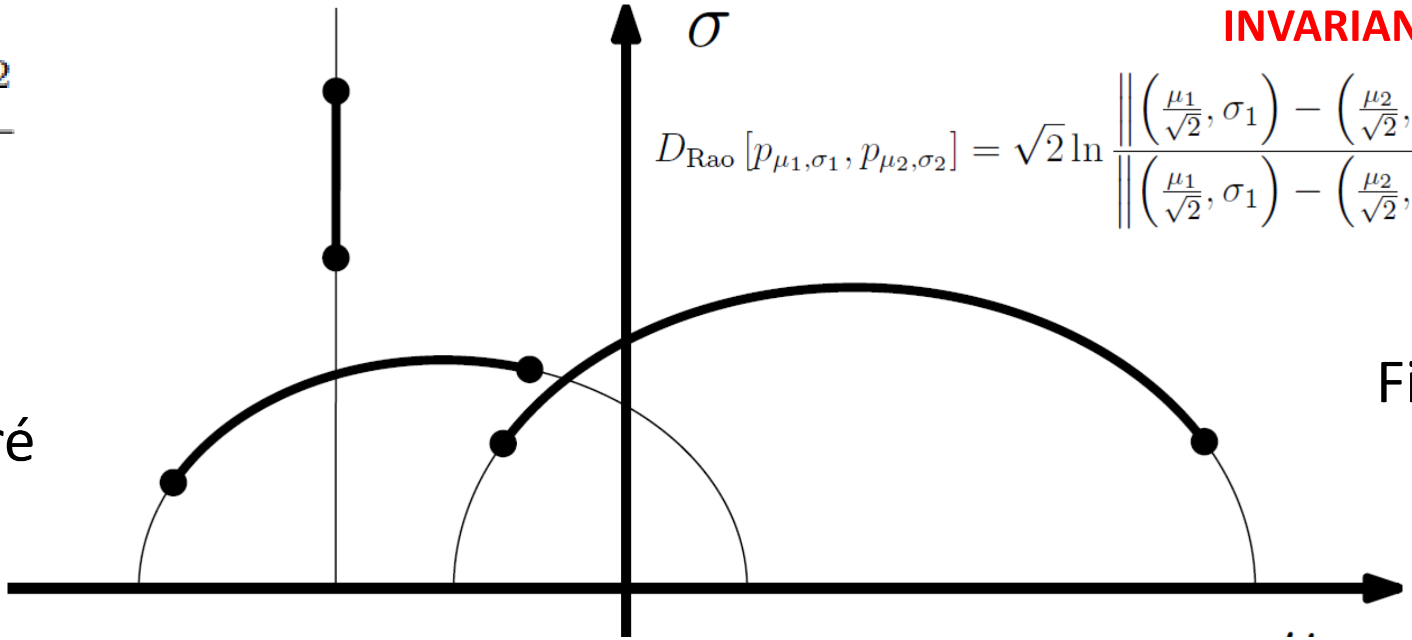
Fisher-Rao geometry of univariate normal distributions

INVARIANT

$$ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$

Fisher metric

stretched Poincaré half-plane

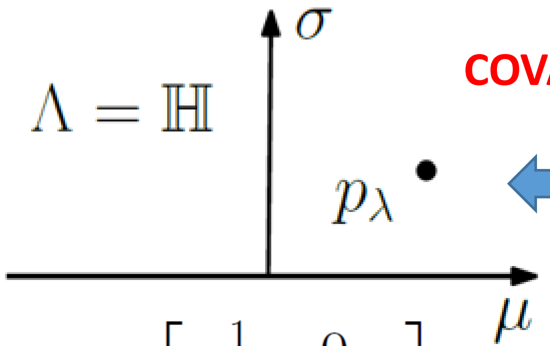


INVARIANT

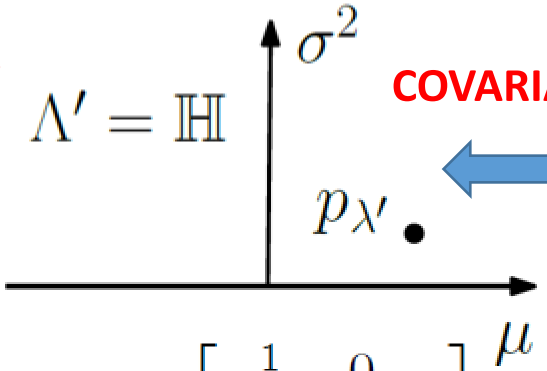
$$D_{\mathrm{Rao}}\left[p_{\mu_1,\sigma_1}, p_{\mu_2,\sigma_2}\right] = \sqrt{2}\ln \frac{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| + \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\| - \left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|}$$

Rao's distance or Fisher-Rao distance

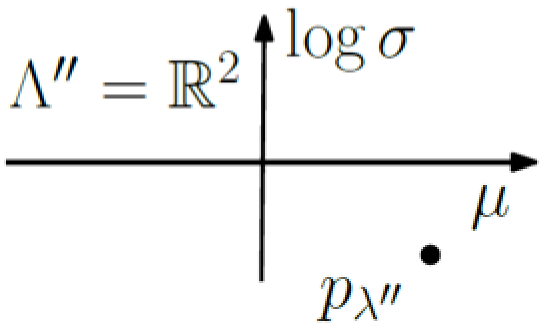
FIM and domains for various parameterizations



COVARIANT



COVARIANT



$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

$$I(\mu,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

$$I(\mu,\log\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2 \end{bmatrix}$$

In general, location-scale families yield a hyperbolic Fisher-Rao geometry

Fisher-Rao manifolds: Intrinsic vs extrinsic viewpoints

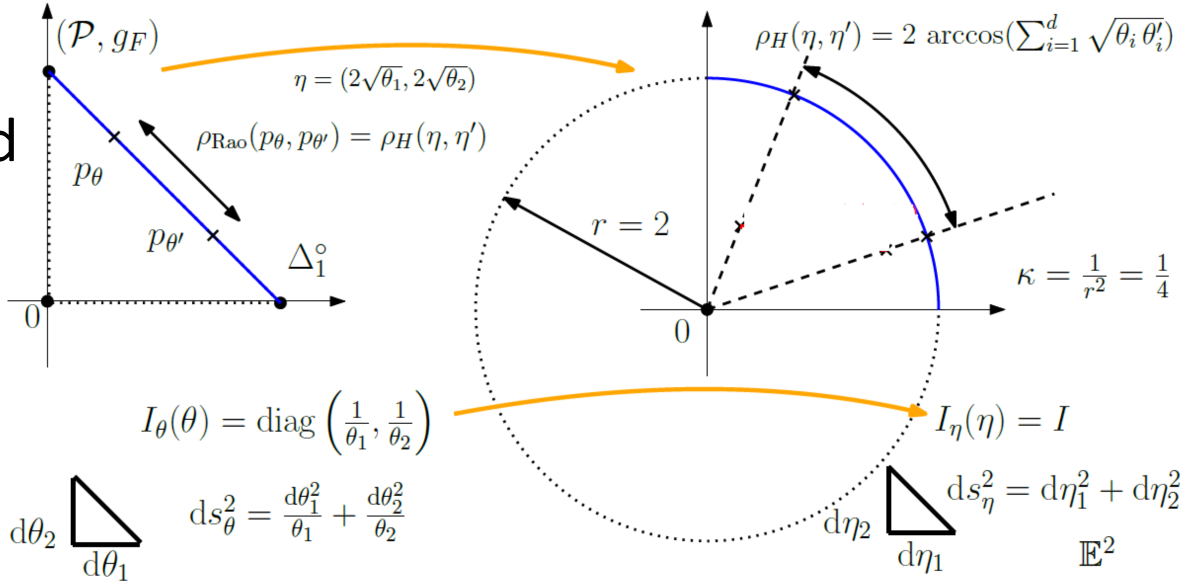
- A Riemannian manifold of dimension D can be embedded as a **surface** of Euclidian space in dimension $O(D^2)$:

  Isometric embedding of the manifold

- For example, Rao's distance between two Bernoulli distributions or categorial distributions can be easily found by embedding the standard simplex on the positive orthant of the 1D sphere of radius 2 in R2 by the 2 x square root transformation

Covariance transformation of the FIM

Parameter space
Intrinsic Fisher-Rao manifold
of dimension 1
(Bernoulli family)



Fisher-Rao manifold
Extrinsic
Embedded in R^2

$$\mathcal{P} = \{p^x(1-p)^{1-x}, p \in (0,1), x \in \{0,1\}\}$$

Neuromanifolds and deep learning

- A neural network (like the multilayer perceptrons, MLPs) is described by a feed-forward architecteur by means of a large number of parameters θ
- Consider **stochastic neural networks (SNNs) with noisy output**:

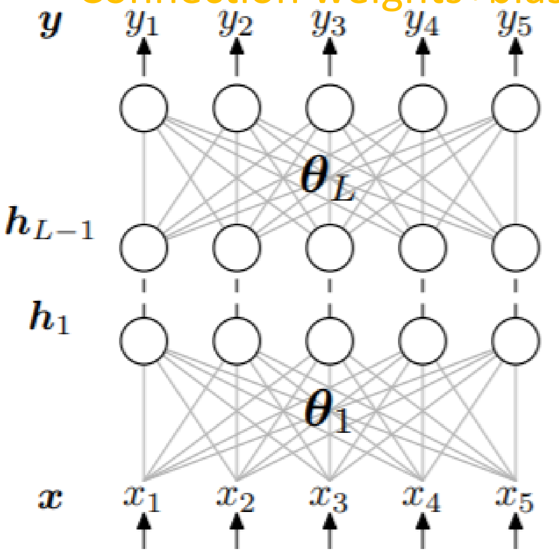
$$\underline{y = \mathrm{NN}_\theta(x) + \epsilon}$$

... Eg a Gaussian noise: $p_\theta(x,y) = p(x)p_\theta(y|x) = \frac{p(x)}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y-\mathrm{NN}_\theta(x))^2\right)$

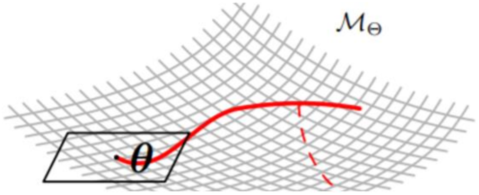
- Neuromanifold is $\boxed{\mathcal{P} = \{p_\theta(x,y) \ : \ \theta \in \Theta\}}$

- **Maximizing the likelihood** of a SNN with Gaussian noise amounts to **minimize the mean quadratic error**
- Given a training set, we learn the parameter of the NNs using gradient descent. We can visualize the learning process as a trajectory on the neuromanifold modelling the parameter space. We observe plateau phenomena when nearning a singularity on the manifold where the Fisher information matrix is rank deficient or close to (small eigenvalues of the FIM).

Parameter space θ :
Connection weights+bias



Artificial Neural Network Modelled by a neuromanifold



Learning trajectory

[SN 2017]

Natural gradient: Steepest Riemannian descent

Ordinary gradient descent (GD) method for minimizing a loss function E(.) :

$$\theta_{t+1} = \theta_t - \alpha \nabla E(\theta_t)$$

Learning step or rate

- Ordinary GD depends on the parameterization
- Plateau phenomena near singularities (almost degenerate Fisher information)

Natural gradient is invariant to reparameterization and avoids plateaus:

$$\boxed{\tilde{\nabla} E(\theta) := G(\theta)^{-1} \nabla_\theta E(\theta)}$$

$$\tilde{\nabla} E_\eta(\eta) = \tilde{\nabla} E_\theta(\theta)$$

Natural gradient descent (NGD)

$$\theta_{t+1} = \theta_t - \alpha \tilde{\nabla} E(\theta_t)$$

Where α = step size

Natural gradient descent is different from the **Riemannien gradient descent** which relies on the Riemannian exponential map which is time consuming (retraction)

What is information geometry? (4/4)

- A **dual structure** which allows to explain the duality between statistical inference like the maximum likelihood estimator and a family of statistical models obtained from the maximum entropy principle: Information geometry explains the link between Shannon entropy, the Kullback-Leibler divergence and exponential families in statistics.
- **Second principle of invariance** by **sufficient statistic** $p_\lambda(\underline{x})$

x ∈ sample space Ω

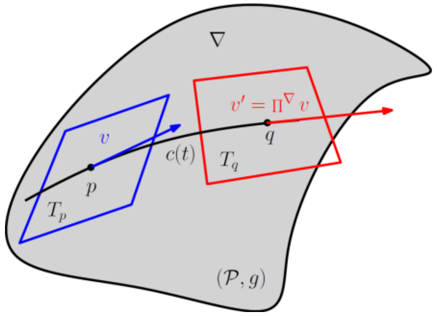
This core dual structure of information geometry:

- Open **new perspectives**: For example, non-extensive entropies like Tsallis entropy, complex systems, conformal geometry of deformed exponential families, etc.
- Many applications of information geometry ranging from signal processing (Radar, Brain-Machine interfances, etc.), to medical imaging, to machine learning and AI, etc.

Geodesics are defined according to affine connections

- In Riemannian geometry, geodesics are locally length minimizing curves

- Geodesics γ(t) defined by connection ∇ as **∇-autoparallel curves**:



$$\boxed{\nabla_{\dot\gamma}\dot\gamma = 0,} \qquad \dot\gamma = \frac{d}{dt}\gamma(t) \qquad \frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p}\Gamma_{ij}^{k}\frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0, \quad k = 1,\ldots,p,$$

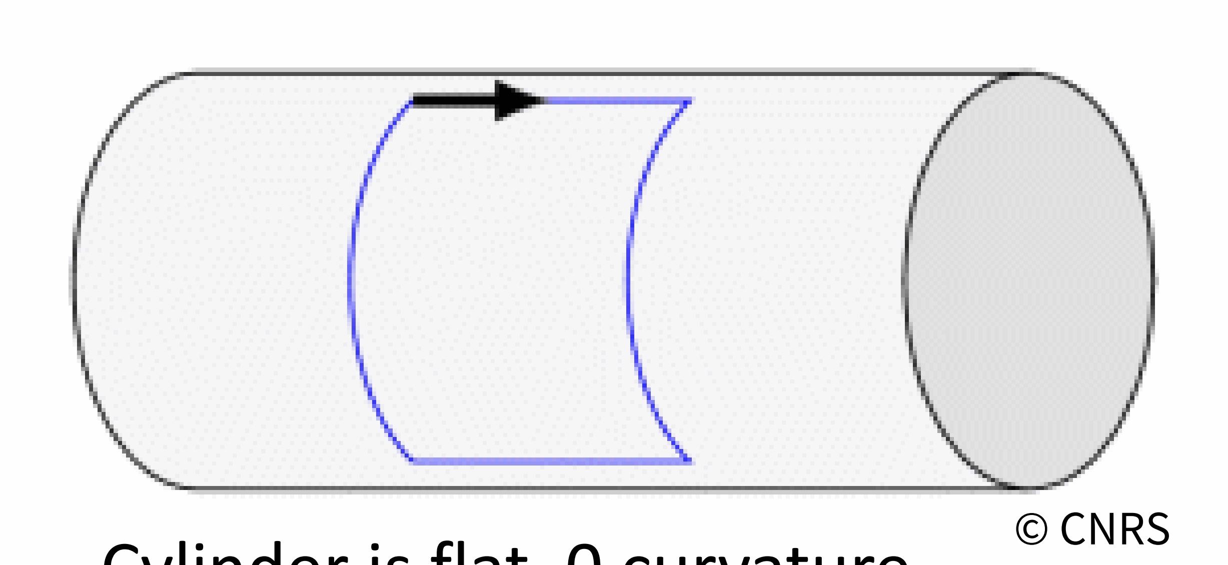
where $\nabla_X T$ is the **covariant differentiation operator** and X is a vector field

D^3 **Christoffel symbols** Γ which are functions characterizing the affine connection ∇ (covariant derivative)

- In Riemannian geometry, we use by default the **Levi-Civita connection** which is derived from the metric tensor field g (thus implicit in Rie. Geo.) :

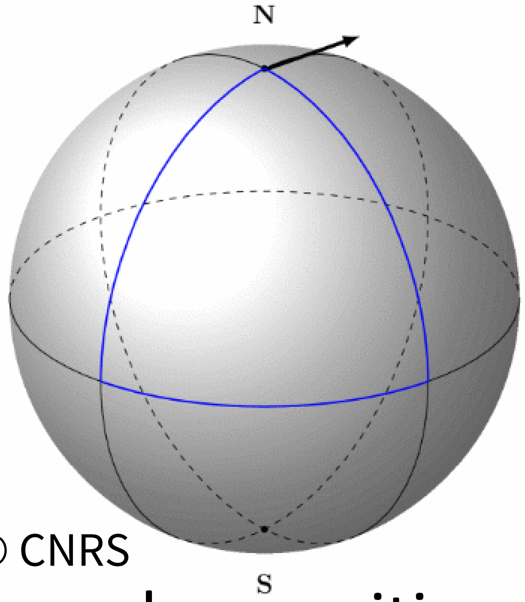
$$\boxed{\nabla = {}^g\nabla} \qquad \Gamma_{ij}^{k} = \frac{1}{2}\sum_{m=1}^{p}\left(\frac{\partial g_{im}(\boldsymbol\theta)}{\partial\theta_j} + \frac{\partial g_{jm}(\boldsymbol\theta)}{\partial\theta_i} - \frac{\partial g_{ij}(\boldsymbol\theta)}{\partial\theta_m}\right) g^{mk}(\boldsymbol\theta), \quad i,j,k = 1,\ldots,p,$$

Affine connection ∇ : Visualizing the curvature by the ∇-parallel transport along smooth loops



© CNRS

Cylinder is flat, 0 curvature



© CNRS

Sphere has positive constant curvature

© wikipedia

Élie Cartan
1869-1951

A connection is flat is there exists a coordinate system θ for which the Christoffel symbols all vanish: **Γ(θ)=0**

→ Geodesics are plotted as **line segments in the local chart** θ

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

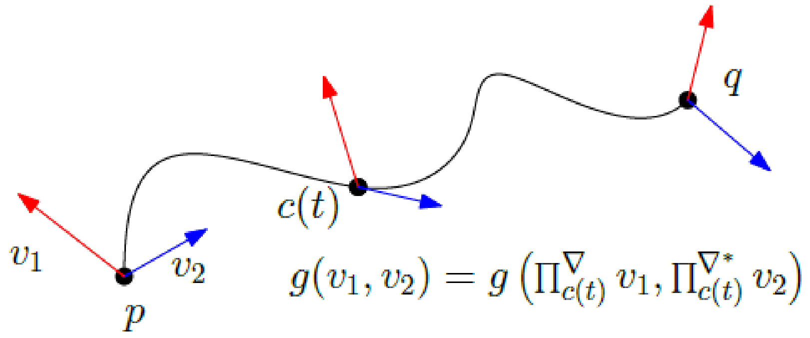
Geodesics of flat connection
=
Line segments

The key dual structure of information geometry

$$(M, g, \nabla, \nabla^*) \qquad \text{such that} \qquad \boxed{\frac{\nabla + \nabla^*}{2} = {}^g\nabla}$$

- Given a torsion-free affine connection ∇ and a metric tensor **g**, we can build a **unique dual torsion free connection ∇^*** such that the metric is preserved by the bi-parallel transport

$$\langle u, v\rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)}.$$



- The dual connection of the dual connection is the original connection $(\nabla^*)^* = \nabla.$
- Question: How to find meaningful dual connections?
  - Method of Amari-Nagaoka (1982) : the statistical expected **α-connexions** (Chentsov 1972)
  - Method of Eguchi (1983): Build dual connections from dual divergences (contrast functions)

The dual α-geometry of Amari and Nagaoka

Structure $(\mathcal{P}, g_F, \nabla^{\alpha}, \nabla^{-\alpha})$ Dual connections with respect to Fisher metric

∇^{α} Defined by the Christoffel symbols $\Gamma^{\alpha}_{ij,k} = E_{\theta}\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right]$

Some α-connections:

- **0-connection** = **Levi-Civita metric connection of Fisher metric : Fisher-Rao mfd**
- **1-connection** is called the **exponential connection** [Efron 1975]
- **-1 connection** is called the **mixture connection** [Dawid 1975]

$$(\nabla^e)^* = \nabla^m \qquad (\nabla^m)^* = \nabla^e$$

$$\boxed{\nabla^{\alpha} = \frac{1+\alpha}{2}\nabla^e + \frac{1-\alpha}{2}\nabla^m}$$

Dual geometry em used to study the duality between estimators/stat models

Eguchi's dual geometry induced by a divergence

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

Divergence information geometry

(self dual when divergence is symmetric)

- Get a **divergence** (contrast function) from a statistical divergence between parametric distributions. For example, the Kullback-Leibler divergence between two parametric distributions from a family P:

Parametric divergence

contrast divergence

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

Statistical divergence

- Eguchi Levi-Civita metric associated to D is ${}^D g_{ij}(\theta_1) = -\partial_{\theta_1}\partial_{\theta_2} D(\theta_1 : \theta_2)|_{\theta_1=\theta_2}$
- Eguchi connection associated to D ${}^D\Gamma_{ij,k} = g\left({}^D\nabla_{\partial_i}\partial_j, \partial_k\right) = -\partial_{\theta_1^i}\partial_{\theta_1^j}\partial_{\theta_2^k} D\left(\theta_1 \| \theta_2\right)\Big|_{\theta_1=\theta_2}$
- Define the **dual divergence** en by swapping the parameter order:

$$D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$$

- Get dual affine connections ${}^D\nabla^* = {}^{D^*}\nabla$

Levi-Civita connection is recovered from

$$\boxed{{}^{D_g}\nabla = \frac{1}{2}\left({}^D\nabla + {}^{D^*}\nabla\right)}$$

f-divergences and their induced connections

- Relative entropy or the Kullback-Leibler divergence belongs to a broader class of dissimilarities : **f-divergences** **[Csiszar'63] [Ali&Silvey'66]**

$$I_f[p:q] = \int p f(q/p) \mathrm{d}\mu \Rightarrow D_{\mathrm{KL}}[p:q] = \int p \log p/q \mathrm{d}\mu = I_{f_{\mathrm{KL}}}[p:q] \quad f_{\mathrm{KL}}(u) = -\log u$$

Separable divergence

- Generator f(.) is convex, strictly convex at 1.

WLOG, fix f'(1)=0 et f''(1)=1 to get a **standard f-divergence**

- **Dual f-divergence** $I_f^*[p:q] = I_f[q:p] = I_{f^*}[p:q]$ with $f^*(u) = u f(1/u)$
- The Eguchi induced metric tensor of std f-divergences = Fisher : ${}^{I_f^{\mathcal{P}}}g = g_F = {}^{I_{f^*}^{\mathcal{P}}}g$
- **Induced f-connections** wrt to f-divergences between distributions of a family P match with the **α-connections of** Amari and Nagaoka :

$$\boxed{{}^{I_f^{\mathcal{P}}}\nabla = {}^{\mathcal{P}}\nabla^{\alpha_f}} \qquad \alpha_f = 3 + 2\frac{f'''(1)}{f''(1)}$$

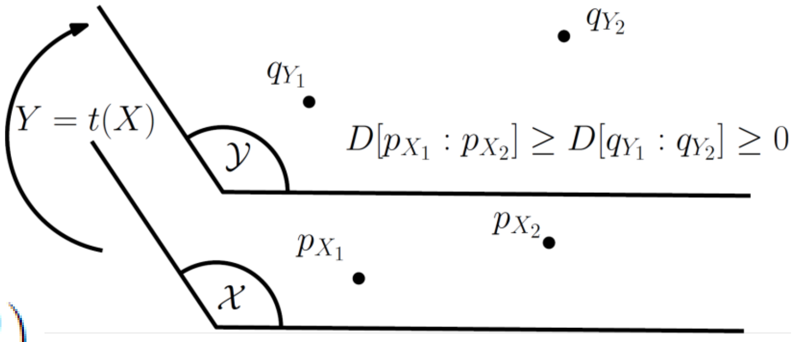
Statistical distances and information monotonicity

- Consider a transformation Y=t(X) on random variables between two measurable spaces (deterministic or stochastic, Markov kernel):

$$t: (\mathcal{X}, \Sigma) \rightarrow (\mathcal{Y}, \Sigma') \qquad Y_i = t(X_i)$$

- **Second principle of invariance**: We should not increase the power of discrimination of divergences by a transformation:

$$D[p_{X_1} : p_{X_2}] \geq D[q_{Y_1} : q_{Y_2}] \geq 0$$



- Fisher information monotonicity: $I_{t(X)}(\theta) \leq I_X(\theta)$
- Equality holds if and only if t(X) is a **sufficient statistic**
- A sufficient statistic summarizes all necessary information for inference on the parameter θ (statistical lossless compression): $\Pr(x|\theta) = \Pr(x|t)$
- **Theorem: f-divergences are the only separable monotone divergences**

Exponential families have finite dim. sufficient statistic vector

- An exponential family is a set of parametric distributions with density which can be expressed canonically as :

$$p_\theta(x) = \exp\left(\langle\theta, t(x)\rangle - F(\theta)\right)$$

μ (eg., Lebesgue or counting measure)

By default, scalar product = Euclidean inner product

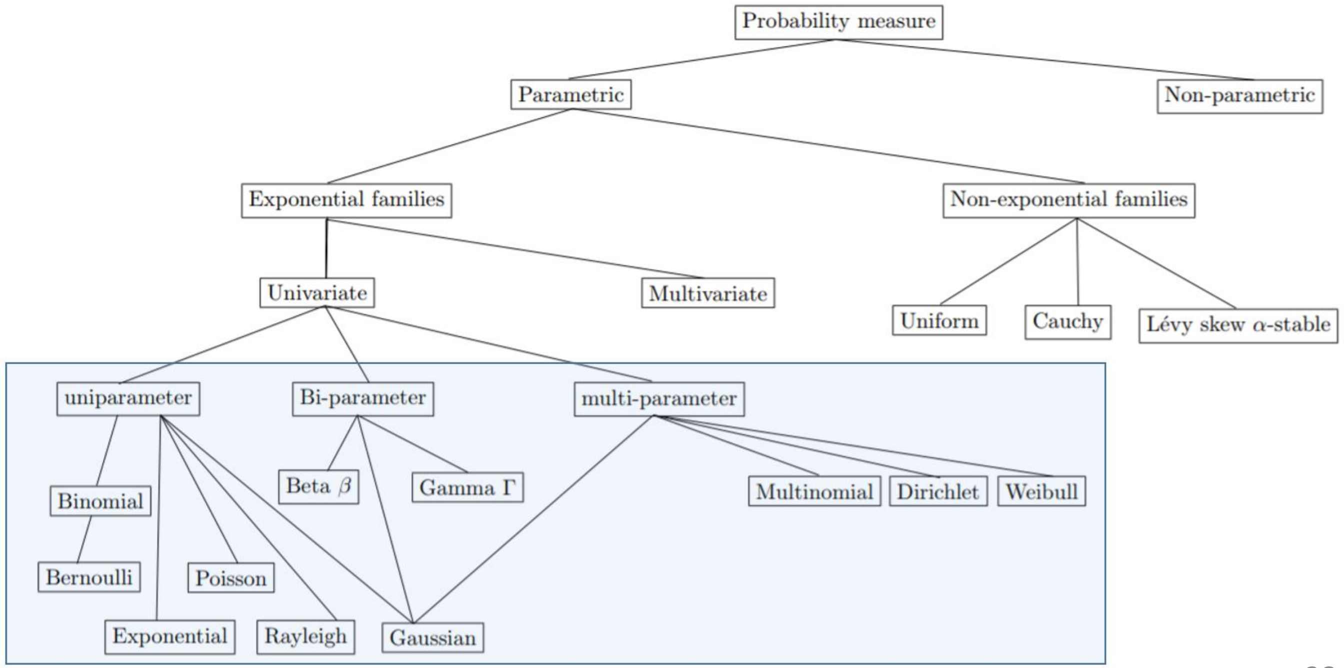
whereF is an analytic and **stricty convexe and differentiable function**:

$$F(\theta) = \log \int \exp(\theta x) \mathrm{d}\mu(x)$$

F : log partition function or cumulant function

Natural parameter space for full EFs

$$\Theta = \{\theta \; : \; \int \exp(\theta x) \mathrm{d}\mu(x) < \infty\}$$



Information geometry of exponential families: Dually flat

- Statistical model: natural exp. family $\mathcal{P} = \{p_\theta(x) = \exp(x^\top \theta - F(\theta))\}$

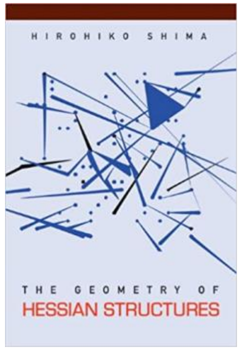
  or more generally $\mathcal{P} = \{\exp(\theta^\top t(x) - F(\theta))\, h(x)\}$ $\quad p_\theta(x) = p^\eta(x)$

- Exponential connection and dual mixture connection are both flat:

Dually flat spaces of exponential families : $(M, g_F, \nabla^e, \nabla^m)$

- Fisher information metric is a Hessian metric $g_F(\theta) = I(\theta) = \mathrm{Cov}[t(X)] = \nabla^2 F(\theta)$
- By using the **Legendre-Fenchel transformation,** we get a dual coordinate system eta $F^*(\eta) = \sup_\theta \theta^\top \eta - F(\theta), \eta = \nabla F(\theta) = E[t(X)]$
- Moment or mean parameterization:

$$\boxed{p_\theta(x) = p^\eta(x)}$$



- Fisher information matrix can be expressed in the moment parameter:

$$I(\eta) = \nabla^2 F^*(\eta) = g_F(\theta)^{-1}$$

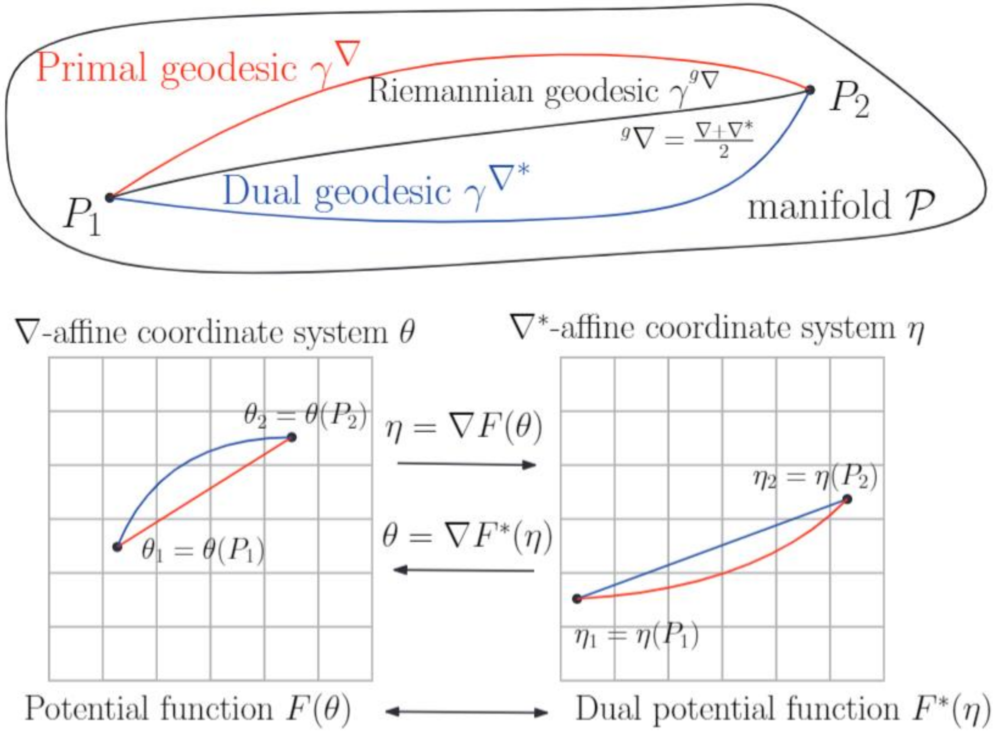
Dually flat spaces with Hessian structures

- The primal and dual geodesics are **line segments** in the affine theta and eta coordinate system:

$$p_\theta(x) = p^\eta(x)$$

Primal geodesic $\gamma_{p_{\theta_1} p_{\theta_2}}(t) = p_{(1-t)\theta_1 + t\theta_2}$

Dual geodesic $\gamma^*_{p_{\theta_1} p_{\theta_2}}(t) = p^{(1-t)\eta_1 + t\eta_2}$

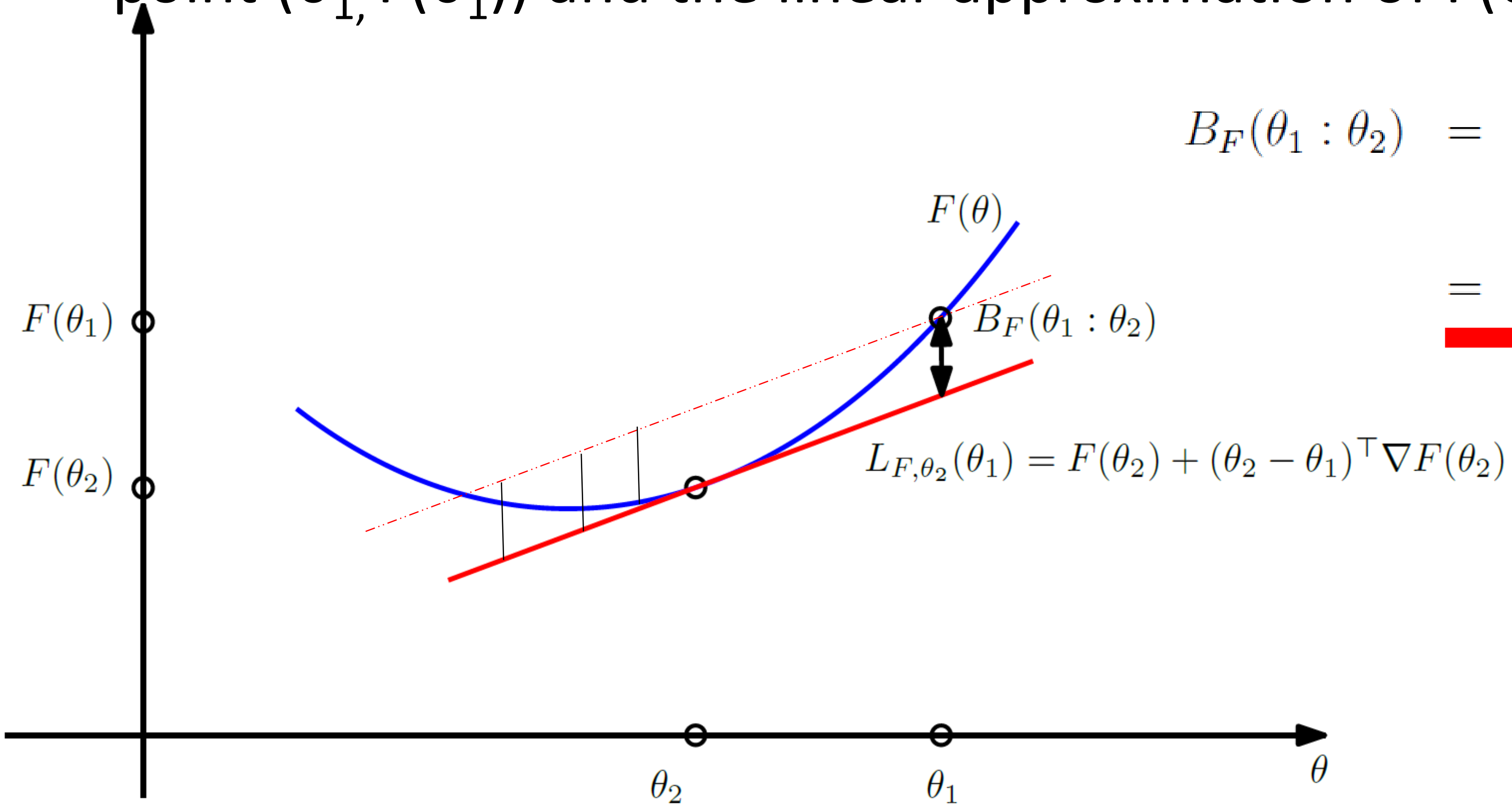


- In dually flat spaces, there is a **canonical divergence**: **The Bregman divergence.** For exponential families, this Bregman divergence amounts to the **dual divergence of the Kullback-Leibler divergence** (=reverse KLD) between corresponding densities :

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

Visualizing a Bregman divergence as a vertical gap

- 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, $F(\theta_1)$) 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]

Mixed coordinates and the Legendre-Fenchel divergence

- Dual **<u>Legendre-type</u>** 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** : $$F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2$$

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

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

Dual Bregman and dual Fenchel-Young divergences

- **Identity for dual Bregman divergences**: $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$

(The Bregman divergence coincides with the reverse Bregman divergence for the convex dual generator)

- By definition, dual divergence = divergence on swapped parameter order:

$$D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$$

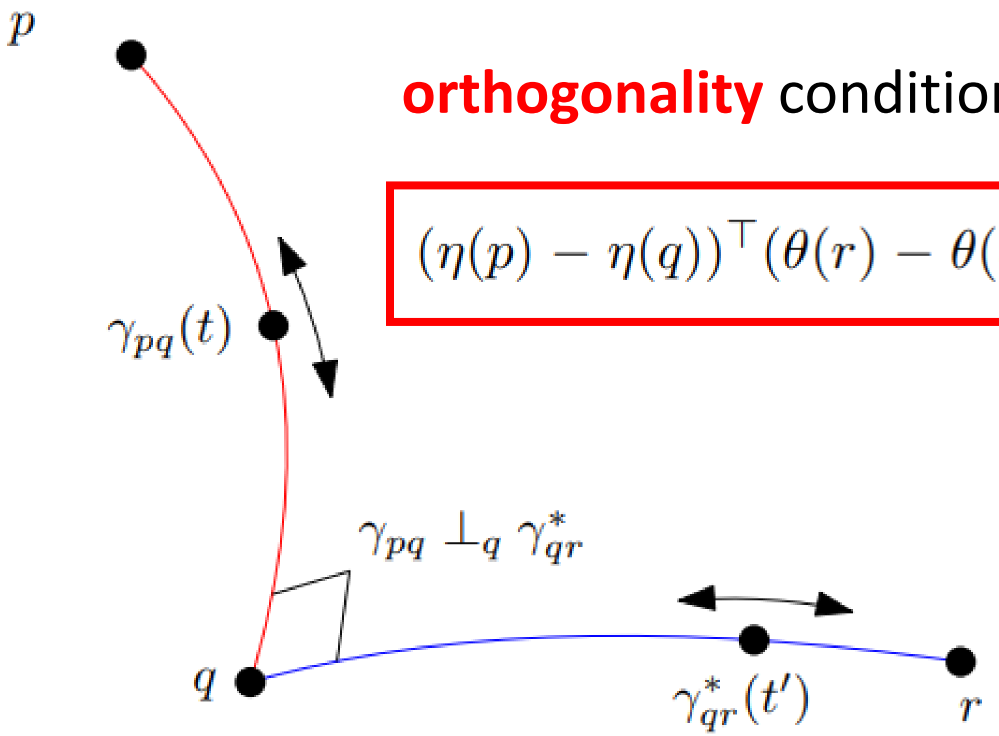
- Thus in a dually flat space, we can write the canonical divergence as :

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

On a Bregman manifold, we can thus get 2^n equivalent formula with n terms

Generalized Pythagoras theorem in dually flat spaces

Generalized Pythagoras' theorem



orthogonality condition:

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

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

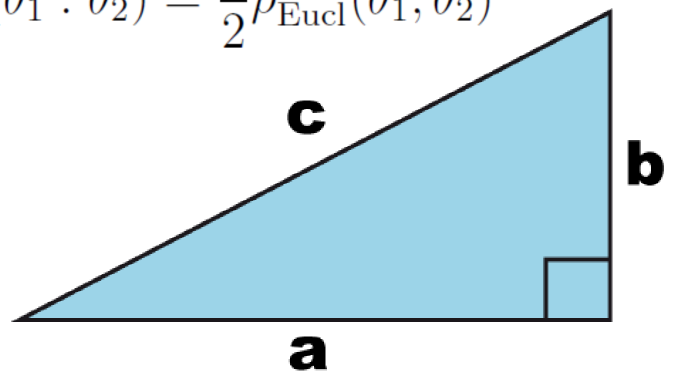
Pythagoras' theorem in

the Euclidian geometry

Self-dual

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

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



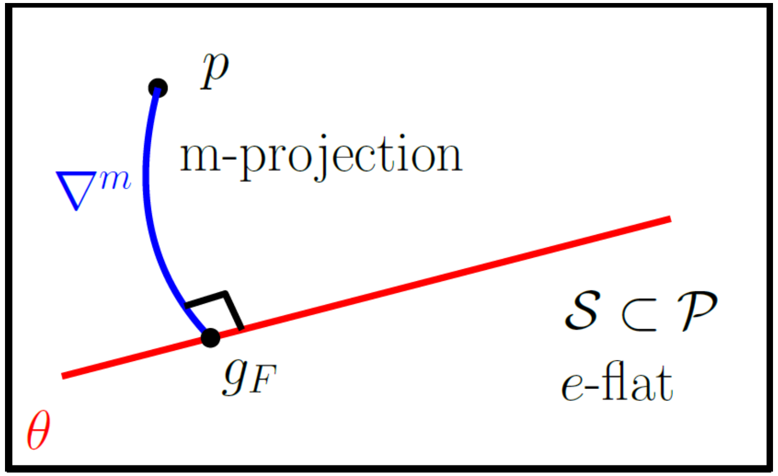
$a^2 + b^2 = c^2$

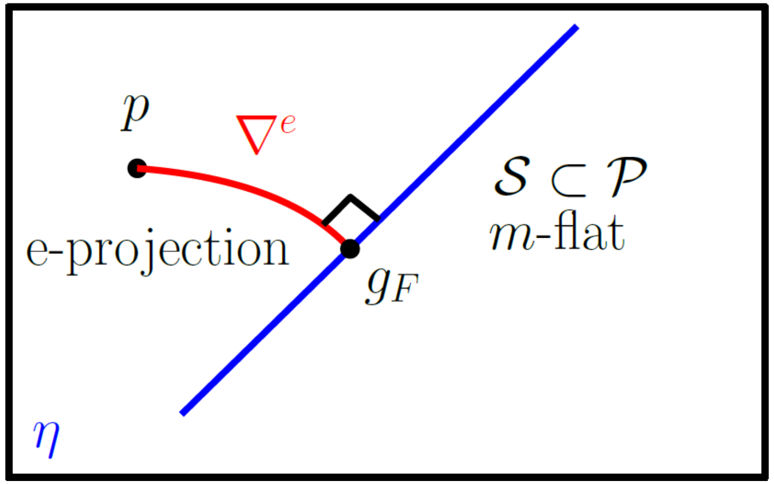
Identity of Bregman divergence with three parameters

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

Information projection uniqueness theorems

- Define the **e-projection** and the **m-projection** of a point onto a submanifold with respect to the affine connections ∇^e (∇^{+1}) and ∇^m (∇^{-1}) where the orthogonality is given by the Fisher information g_F
- A submanifold is **e-flat** if and only if when expressed in the θ-coordinate system, we get an affine subspace.
- Similar definition for a **m-flat** submanifold wrt η





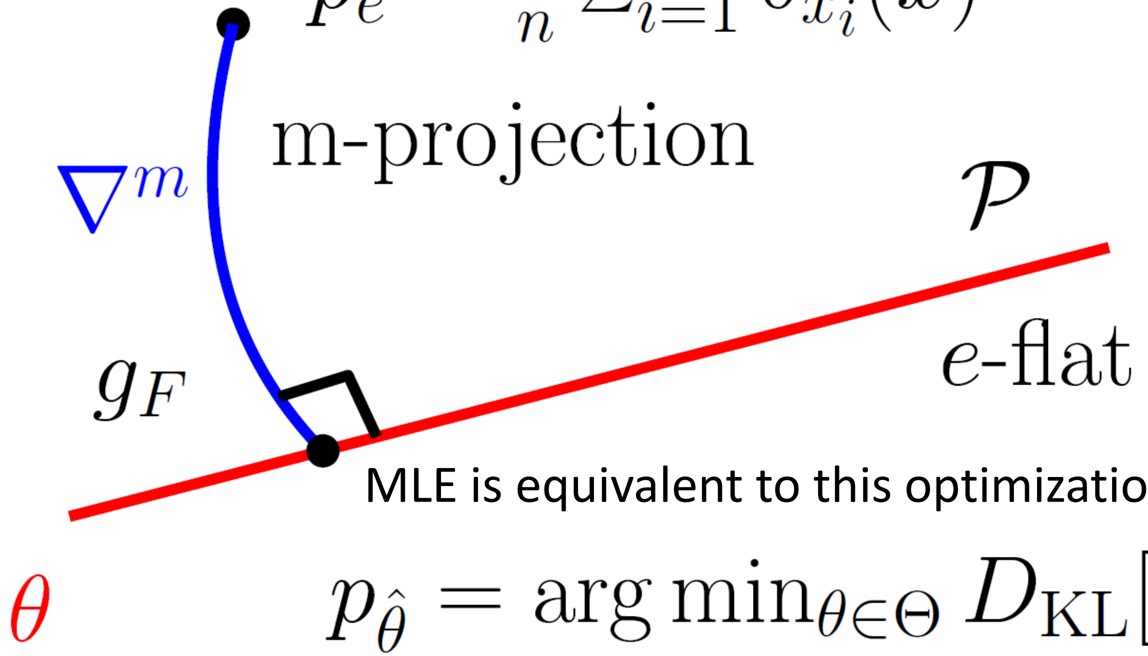
- Generalized Pythagoras' theorem allows to prove that the e-projection of a point onto a n-flat submanifold is **unique** and corresponds the minimization of a Bregman divergence. Similarly, m-projection of a point onto a e-flat submanifold is unique and can be obtained as the minimization wrt to the dual Bregman divergence

Maximum likelihood estimator as a m-projection

Let $\{x_1, \ldots, x_n\}$ be i.i.d variates of an **exponential family** P (**e-flat**)

The empirical distribution is called the **observed point**

$$p_e = \frac{1}{n} \Sigma_{i=1}^{n} \delta_{x_i}(x)$$



MLE is equivalent to this optimization problem:

$$p_{\hat\theta} = \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}[p_e : \boxed{p_\theta}]$$

Maximum likelihood estimator:

$$\boxed{p_{\hat\theta} = \mathrm{Proj}_{\mathcal{P}}^{\nabla^m}(p_e)}$$

Canonical divergence:

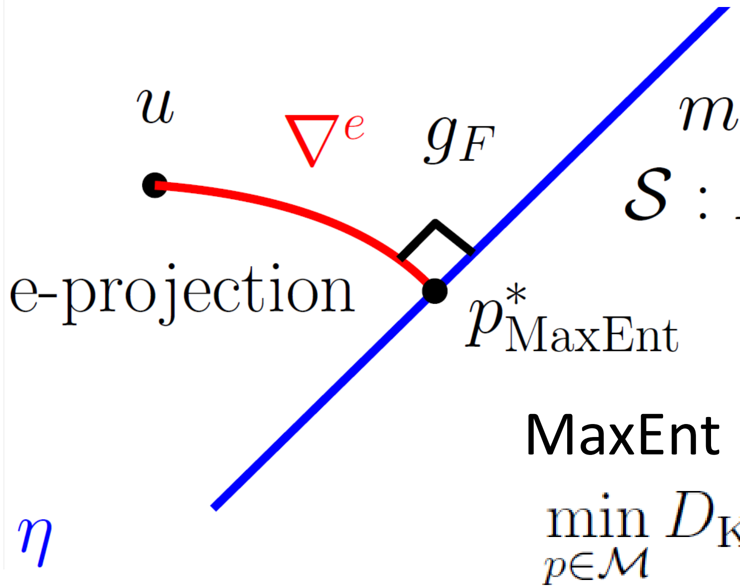
$$D_{\nabla^m} = D_{\mathrm{KL}}$$

By considering an arbitrary divergence D[.:.] instead of the Kullback-Leibler divergence we get **D-estimators**. MLE is interpreted as the KLD-estimator

Maximum entropy and e-projection

- Given **observations** with $E[t_i(x)]=m_i$, the **maximum entropy principle** of Jaynes estimate *the distribution* which maximizes Shannon entropy under the moment contraints $E_p[t_1(X)] = m_1, \ldots, E_p[t_D(X)] = m_D$.

$$t(x) = (t_1(x), \ldots, t_D(x))$$
$$m = (m_1, \ldots, m_D)$$



m-flat

$$\mathcal{S} : E[t(x)] = m$$

$$\boxed{p^*_{\mathrm{MaxEnt}} = \mathrm{Proj}_{\mathcal{S}}^{\nabla^e}(u)}$$

MaxEnt problem amounts to

$$\min_{p \in \mathcal{M}} D_{\mathrm{KL}}[p : u] = \min_{p} D^*_{\mathrm{KL}}[u : \boxed{p}]$$

Canonical dual divergence:

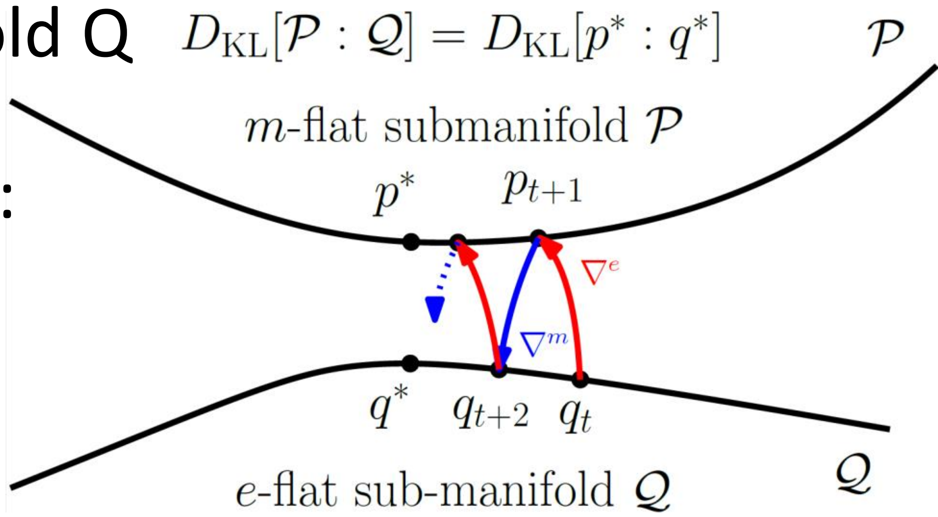
$$D_{\nabla^e} = D^*_{\mathrm{KL}}$$

- Set of distributions maximizing entropy under the constraints $E[t(x)]=\eta$ for all η form an **exponential family** $p^* \in \mathcal{E} := \{p_\theta(x) = \exp(\sum t_i(x)\theta_i - F(\theta))\}$
- For example, the MaxEnt distributions for $E[x]=\eta_1$ et $E[x^2]= \eta_2$ yield the family of normal distributions (univariate of order 2, dim. of natural parameter space)

Alternating projections: The em algorithm $(= \nabla^e \nabla^m)$

- Find the **minimal distance between two submanifolds**
  = Solve jointly the following minimization:
  $$\min_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} D_{\mathrm{KL}}[p : q]$$

- When a submanifold P is **m-flat** and the submanifold Q is **e-flat** then we get unique sequence of e/m alternating projections. Starting from q_1 we repeat:



e-projection : $p_{t+1} = \arg\min_{p \in \mathcal{P}} D_{\mathrm{KL}}[p : q_t]$

m-projection : $q_{t+2} = \arg\min_{q \in \mathcal{Q}} D_{\mathrm{KL}}[p_{t+1} : q]$

- converge towards the pair of points which minimize the Kullback-Leibler divergence between P and Q :

$$\boxed{D_{\mathrm{KL}}[p^* : q^*] = \min_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} D_{\mathrm{KL}}[p : q] = \lim_{t \to \infty} D_{\mathrm{KL}}[p_{t+1} : q_t]}$$

The em algorithm is useful for :
- Interpreting the EM alg. in statistics
- To analyze generative models in deep learning like the VAEs or GANs

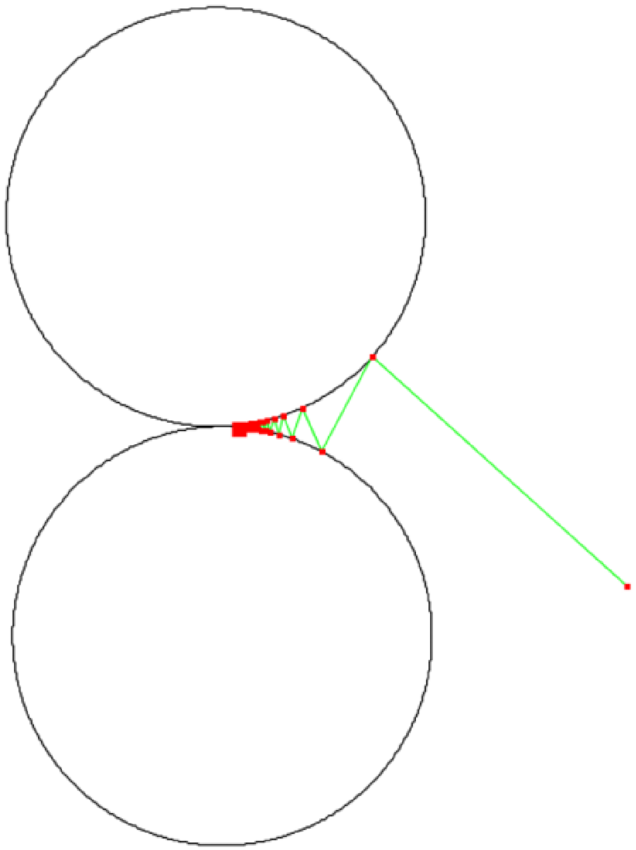
Bregman cyclic projections (in a chart)

- Let n **convex objects** $O_1, \ldots, O_n$ be defined in a θ-coordinate system on a convex Θ
- Goal: find a common point in the intersection of these objects if intersection is non-void
- Repeat cyclically the **Bregman projections:**

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

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

- This sequence converges **towards a common point for non-empty intersection**



Chernoff information and Bayesian hypothesis tests

- Let P_1 and P_2 be two distributions, and take n i. i. d. variates $x_1, \ldots, x_n$ from the statistical mixture model $1/2\ P_1 + 1/2\ P_2$
- **Which rule to classify these n samples with labels P1 or P2?**
- Best rule minimizing the probability of error is **maximum a posteriori** (MAP)
- **Probability of error** is bounded by $P_e^n = 2^{-nC(P_1,P_2)}$

  where C is the following **Chernoff divergence (or Chernoff information)**

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

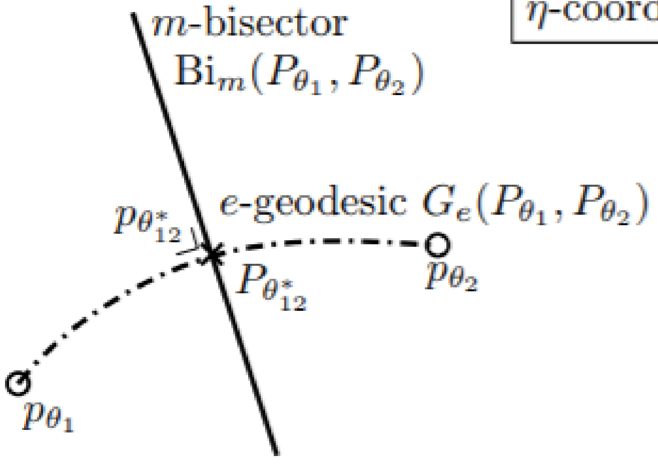
$$C(P,Q) = -\log \min_{\alpha \in (0,1)} \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}\nu(x).$$

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

When P_1 and P_2 are two densities of a same exponential family, we have:

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

where α^* is the optimal exponent in (0,1)



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

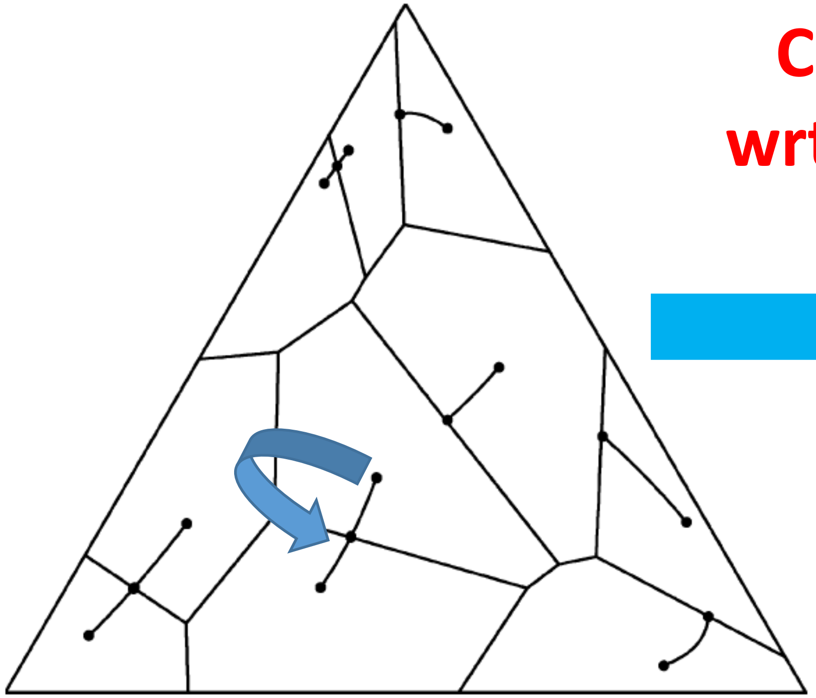
2207.03745

Chernoff information for multiple hypothesis

Probability of error: $P_e^n = 2^{-nC(P_i^*, P_j^*)}$

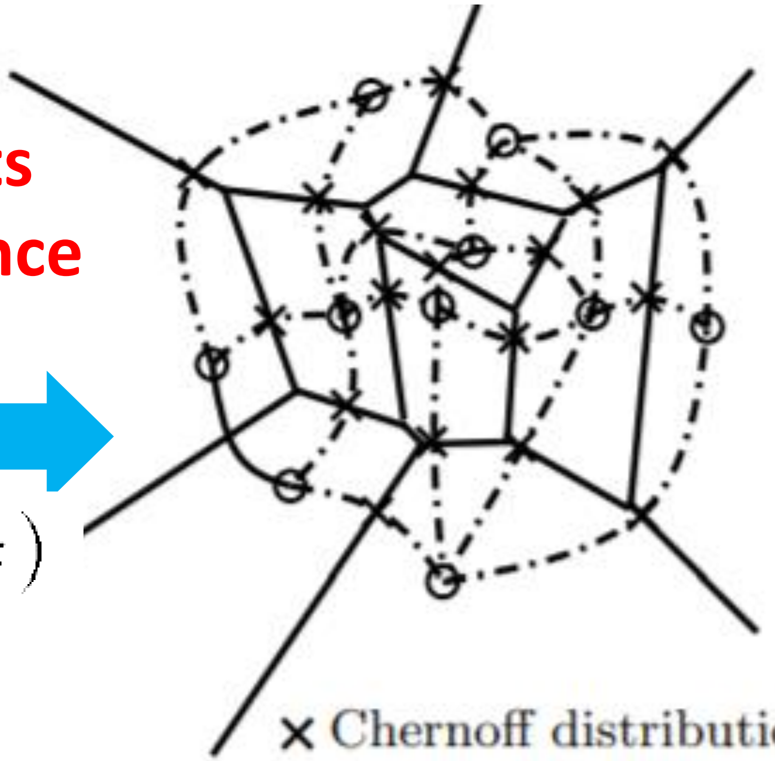
Closest pair of points wrt Chernoff divergence

$\operatorname{argmin}_{i \neq j} C(P_i, P_j)$



Standard simplex (categorial distributions)

Voronoi diagram wrt Kullback-Leibler divergence



Exponential family manifold (dually flat)

Bregman Voronoi diagram

Computational geometry to calculate the Bregman Voronoi diagrams

Natural gradient in dually flat spaces

On a global Hessian manifold (Bregman manifold induced by a convex function F), the Fisher information matrix can be expressed

$$I_\theta(\theta) = \nabla^2_\theta F(\theta) = \nabla_\theta \underline{\nabla_\theta F(\theta)} = \boxed{\nabla_\theta \eta}$$

Définition du paramètre moment

Natural gradient
wrt θ :

$$\tilde{\nabla}_\theta L_\theta(\theta) := I_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta)$$

$$= (\nabla_\theta \eta)^{-1} \underline{\nabla_\theta \eta \nabla_\eta L_\eta(\eta)}$$

<u>Ordinary gradient</u>
wrt η

$$= \underline{\nabla_\eta L_\eta(\eta)}$$

Chain rule derivation

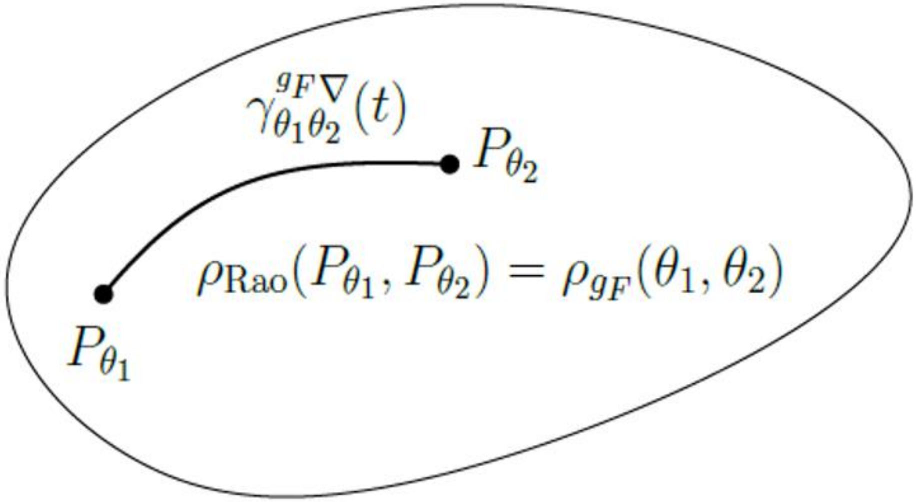
$$L_\theta(\theta) = L_\eta(\eta(\theta))$$

Find many applications in optimization in machine learning:

Natural evolution strategies (NES), Bayesian inference, etc.

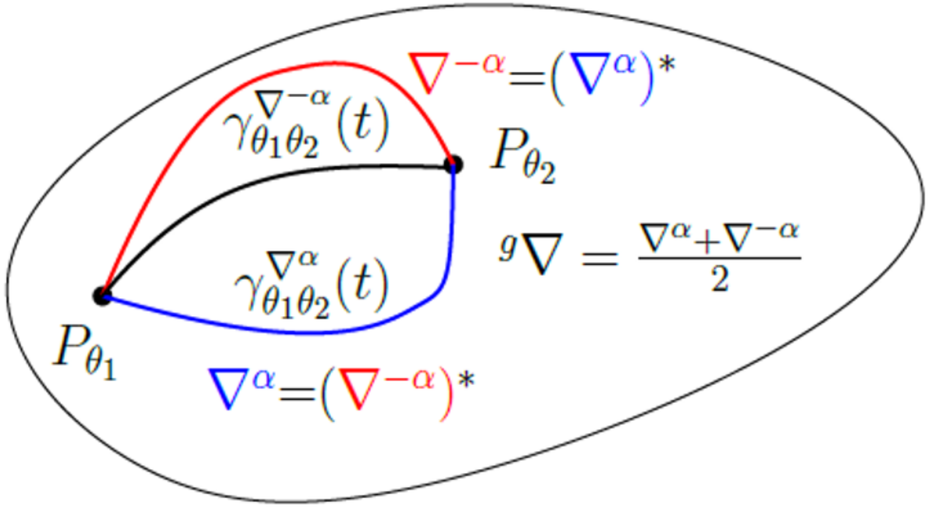
To summarize information geometry in 1 slide!

- **Geometric structures** for a parametric family of distributions: the **statistical model**
- **Invariance** wrt **distribution parameterizations** (θ) and **sufficient statistics** (on sample space Ω). Distance cannot increase by a measurable transformation Y=t(X), and does not change only if t is a sufficient transformation



Fisher-Rao Riemannian geometry

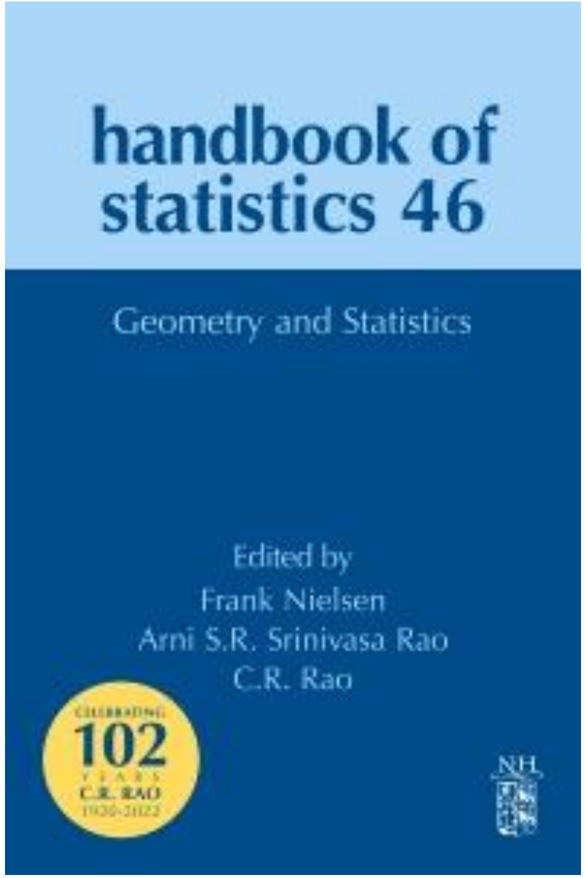
- **Fisher-Rao geometry** equipped with the Rao Riemannian geodesic length distance
- **Dual α-geometry** (they are not necessarily associated divergences, except when dually flat)
- Interpret **statistical estimator** (maximum likelihood estimator) and **statistical model** (maximum entropy): **Pythagoras' theorem** and **information projections in dually flat spaces** (e.g., exponential/mixture families)



Dual α-geometry

Thank you (1/2)

Fresh from the press (July 2022)



Some references

- Shun-ichi Amari, ***Information Geometry and Its Applications***, Springer (2016)
- Frank Nielsen, ***The Many Faces of Information Geometry***, Notices of the AMS, 69.1 (2022)
- Frank Nielsen, ***What is an information projection?,*** Notices of the AMS 65.3 (2018)
- Frank Nielsen, **An information-geometric characterization of Chernoff information**, *IEEE Signal Processing Letters* 20.3 (2013): 269-272.
- Vaden Masrani, Rob Brekelmans, Thang Bui, Frank Nielsen, Aram Galstyan, Greg Ver Steeg, Frank Wood, **q-Paths: Generalizing the geometric annealing path using power means**, UAI 2021.
- Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. **Bregman Voronoi diagrams**, *Discrete & Computational Geometry* 44.2 (2010): 281-307.

Natural gradient and deep learning:

- Ke Sun et Frank Nielsen, ***Relative Fisher information and natural gradient for learning large modular models***, ICML (2017)
- Wu Lin, Frank Nielsen, Emtiyaz Kahn, et Mark Schmidt, ***Tractable structured natural-gradient descent using local parameterizations***, ICML (2021)