Information Geometry:
An *Invitation* for Machine Learning

Frank NIELSEN

Sony Computer Science Laboratories Inc.

Tokyo, Japan

https://franknielsen.github.io/

@FrnkNlsn

Sony CSL

NeurIPS meetup Japan
December 2021

Outline of this talk

Introduction

ML & **Computational Geometry**: A long and fruitful history!

1. Fisher-Rao information geometry

Natural-gradient descent

2. Bregman information geometry

Chernoff information on the exponential family manifold

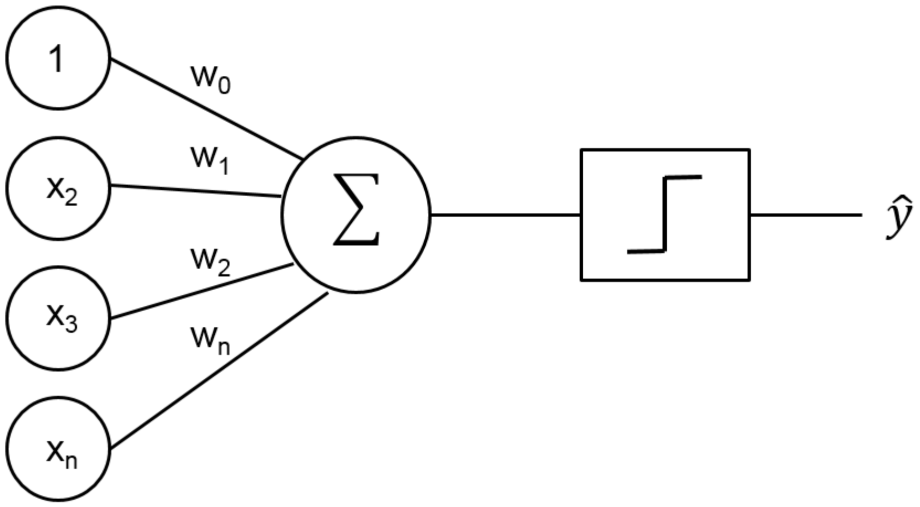
Some perspectives

Introduction:

Machine Learning & Computational Geometry:

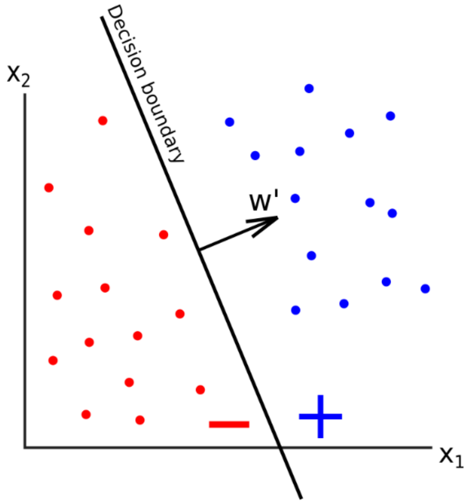
A long and fruitful cooperation from the start!

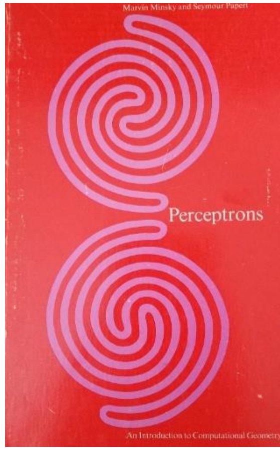
Learning machines: Perceptron & geometry (1960's)



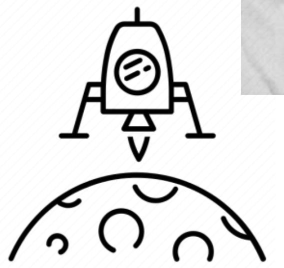
Decision boundary:

geometric hyperplane separator





MIT Press, 1969

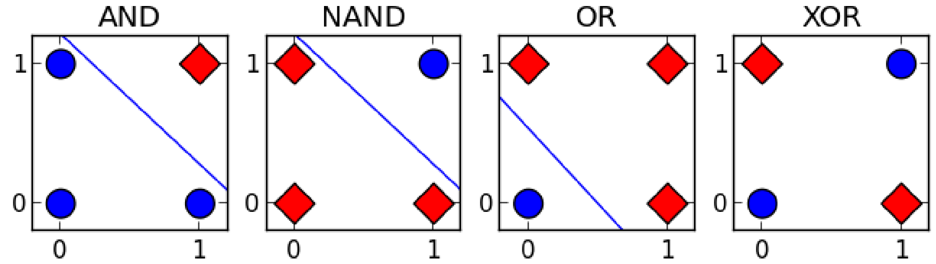


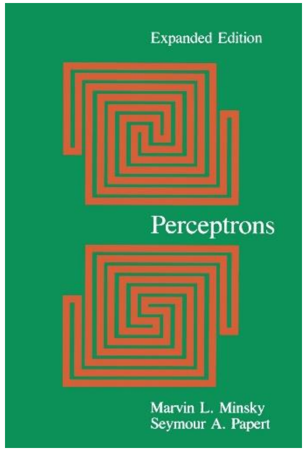
1969

Connectionism machine

≠ von Neumann machine

XOR cannot be learned… NN winter…





MIT Press, 3rd 1987

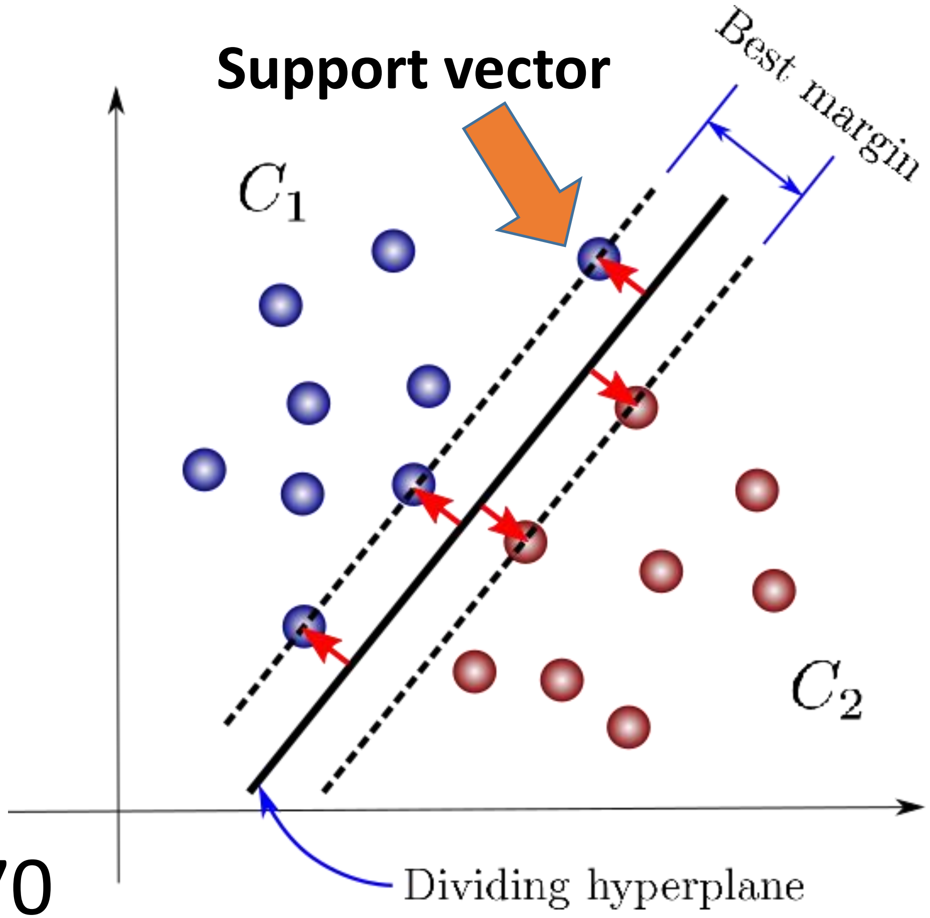
Connectedness…

Marvin Minsky and Seymour Papert:

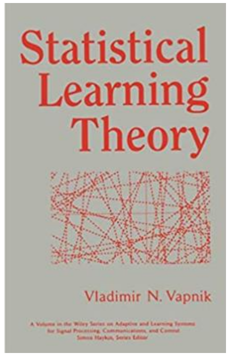
Perceptrons: An Introduction **to** **Computational Geometry***, 1969*

Geometric learning machines: SVMs (1970's/1992)

Linear separator



1970

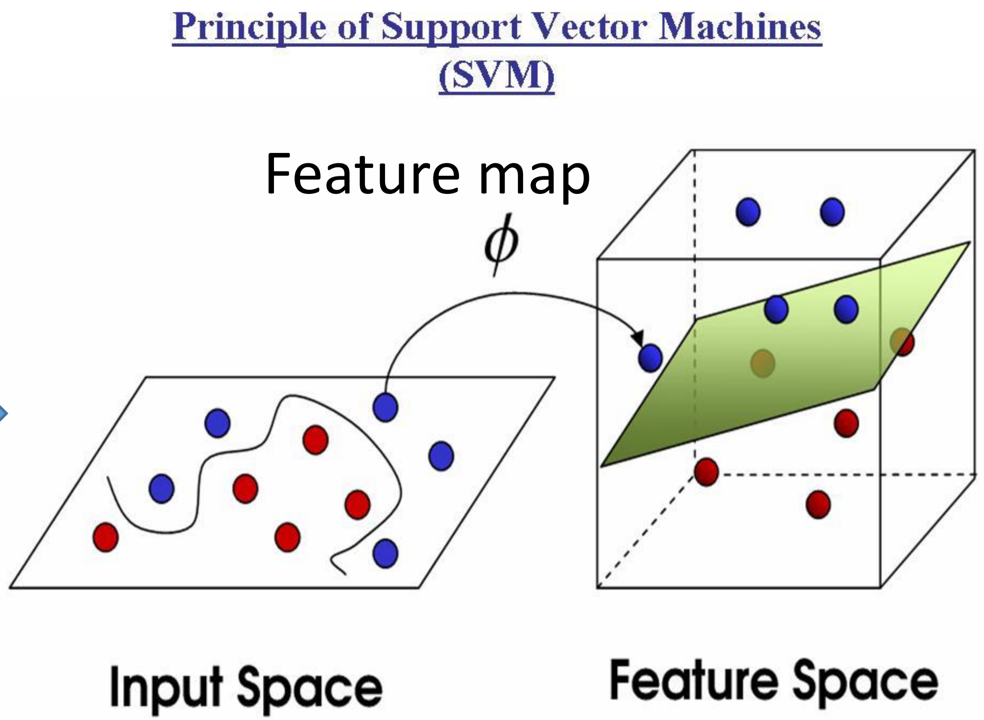


1998

Inner product (Hilbert space)

1992

Non-linear separator (Kernel trick, RKHS)

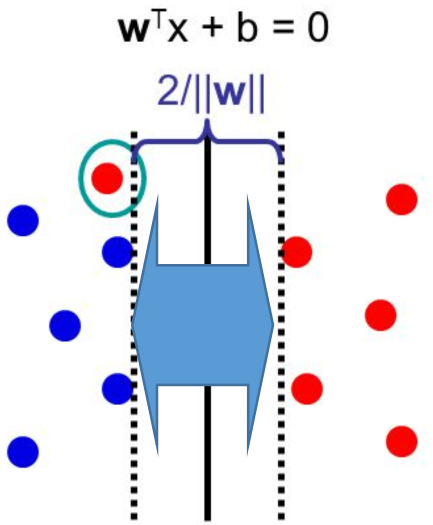


Kernel machines

Theory of VC-dimension = expressive power of (geometric) separators

Dual SVM quadratic program amounts to solve a Smallest Enclosing Ball (= SEB): Computational geometry !

The SVM Framework



Points $\mathbf{X} = \{\mathbf{x}_i\}$

Labels $\mathbf{y} = \{y_i\}$

$y_i \in \{-1, +1\}$

$$\min 1/2\ \mathbf{w}^T\mathbf{w} + C \sum \xi_i$$

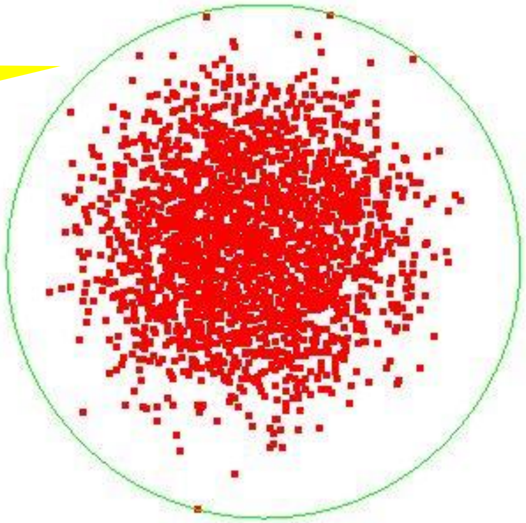
$$y_i\ (\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Convex Quadratic Program

Widest margin hyperplane separator

New insight!

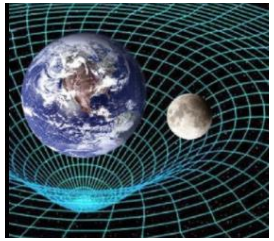


Smallest enclosing ball:

``Smallest'' ball with respect to radius or set inclusion

Approximating Smallest Enclosing Balls with Applications to Machine Learning, IJGA 2009

Information geometry in a nutshell



- Born as a <u>**mathematical curiosity**</u> [Hotelling 1930] [Rao 1945]

  Impacted by the success of Riemannian geometry in *Einstein's general relativity (GR)*

- **Information geometry** studies the ***geometric structures*** and ***statistical invariance*** *(sufficient statistics/Markov kernels)* of a family of probability distributions: the **statistical model**

  \+ demonstrate its use in information sciences: statistics, ML, etc.

- **Geometric method**: **coordinate-free objects** with **computing** operating in (local) **coordinate systems:** free to choose coordinates **to ease** the computations!

- **Dualistic structures** pioneered by Prof. Shun-ichi Amari & statistical invariance pioneered by Chenstov

  [Amari 1985] [Amari & Nagaoka 2000] [Amari 2016] [Chentsov 1982]

The **fabric** of information geometry
and the **untangling** of its geometry, divergence, statistical models

1. Fisher-Rao information geometry

Riemannian geometry

Fisher information matrix (FIM)

- A **parametric** family of distributions $\mathcal{P} = \{p_\theta\}_{\theta \in \Theta}$

- **Fisher information matrix** is positive-semidefinite matrix:

FIM = Covariance of the score: $\boxed{I_X(\theta) = \mathrm{Cov}(s_\theta)}$

$$X = (x_1, \ldots, x_D)^\top \sim p_\theta$$

Positive semi-definite matrix

- Score: $s(\theta) := \nabla_\theta \log p_\theta(x)$

- Under **independence**, Fisher information is **additive**:

$$Y = (Y_1, \ldots, Y_n)_{\sim_{\text{iid}} p_\theta} \quad \Rightarrow \quad I_Y(\theta) = n\, I_X(\theta)$$

Fisher information matrix

$$I_X(\theta) = \mathrm{Cov}(s_\theta)$$

- Under ***regularity conditions I*** = ***FIM type 1*** :

$$I_1(\theta) = E_{p_\theta}\left[(\nabla_\theta \log p_\theta)(\nabla_\theta \log p_\theta)^\top\right]$$

- Under ***regularity conditions II*** = ***FIM type 2*** :

$$I_2(\theta) = -E_{p_\theta}\left[\nabla_\theta^2 \log p_\theta\right]$$

FIM can be **singular** in **hierarchical models** like mixtures & neural networks
FIM can be **infinite** (**irregular models**, e.g., support depend on parameter)
Difficult to estimate FIMs for NNs:
Spectral FIM properties from random matrix theory RMT), relative FIM

Sun & N, Relative Fisher information and natural gradient for learning large modular models, ICML 2017
Soen and Sun, On the Variance of the Fisher Information for Deep Learning, NeurIPS 2021

Fisher information and Cramér-Rao lower bound

- The covariance of any **unbiased estimator** is lower bounded by

$$\mathrm{Cov}[\hat{\theta}] \succeq I_X(\theta)^{-1} \qquad X \sim p_\theta$$

Inverse Fisher Information Matrix (IFIM)

- Since Fisher information is additive:

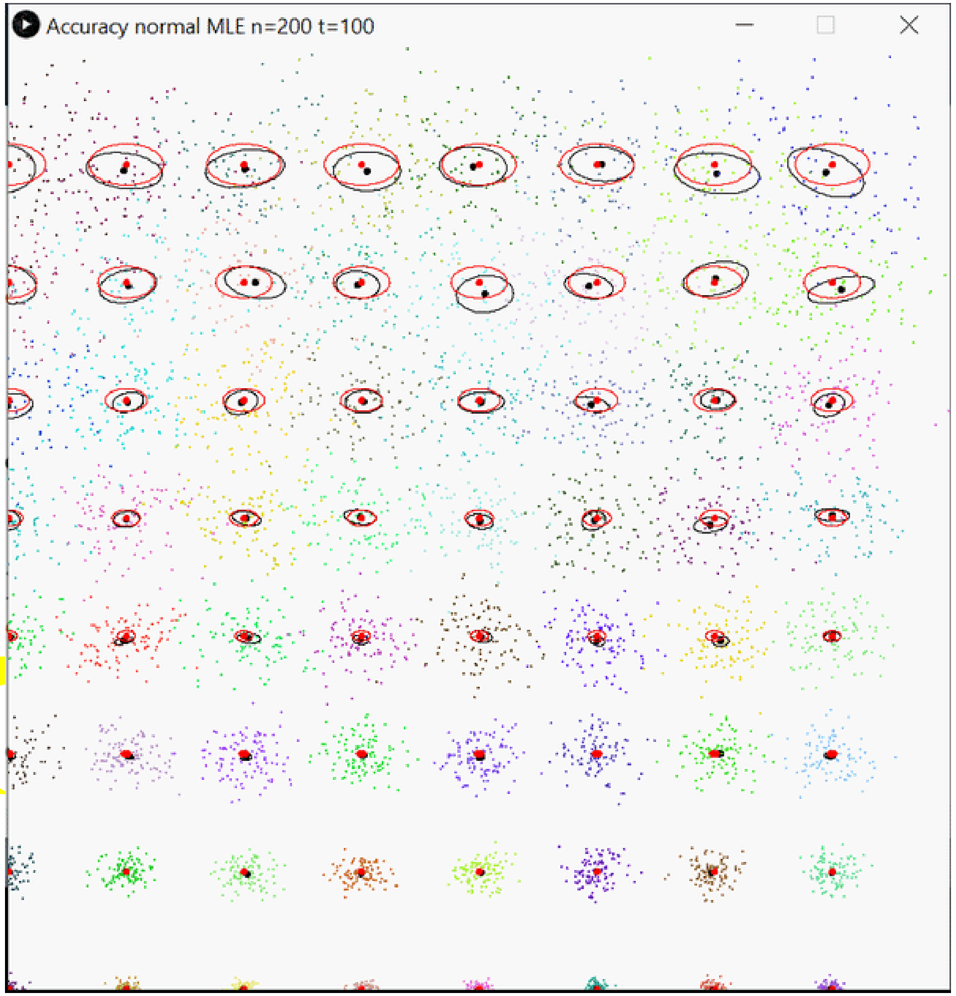
$$\boxed{\mathrm{Cov}[\hat{\theta}_n] \succeq \frac{1}{n} I_X(\theta)^{-1}}$$

Non-asymptotic

$$(X_1, \ldots, X_n) \sim_{\mathrm{iid}} p_\theta$$

$$A \succeq B \Leftrightarrow \forall x, x^\top (A - B)x \geq 0$$

- Accuracy estimators depend on model parameters: **Fisher efficiency**



Empirical estimator covariance matrix

IFIM (Tissot indicatrix)

N., Cramér-Rao lower bound and information geometry, Connected at Infinity II, 2013

Rao's length distance: Riemannian metric distance

(M, g_F): **Riemannian manifold**

Parameter space equipped with the **Fisher information metric g_F**

$$\boxed{\rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2}) = \rho_{g_F}(\theta_1, \theta_2)}$$

$$\rho_g(\theta_1, \theta_2) = \min_{\theta(t)} \int_0^1 \mathrm{d}s_\theta(t)\mathrm{d}t$$

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

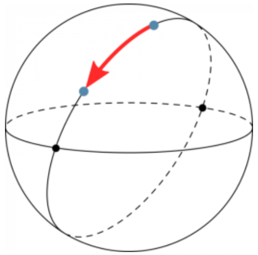
$$\dot{\theta}_k(t) = \frac{d}{\mathrm{d}t}\theta_k(t)$$

C. R. Rao with Sir R. Fisher in 1956

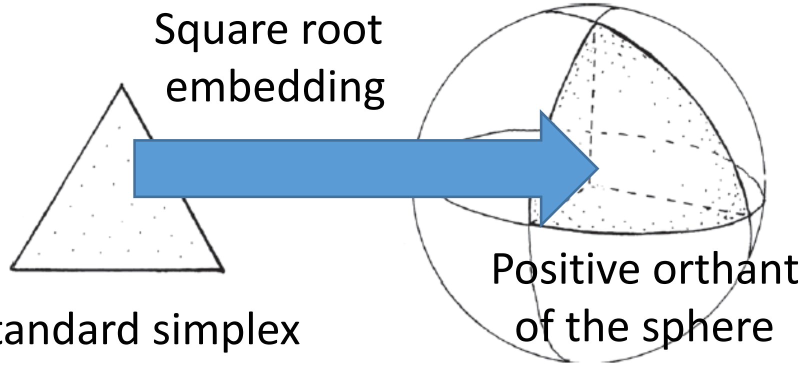
→ need to calculate **Riemannian geodesics** $\theta(t)$:

....characterized as (locally) shortest curves in Riemannian geometry

For example, **Rao distance in the probability simplex**:



$$\rho_{\mathrm{FHR}}(p, q) = 2 \arccos\left(\sum_{i=0}^d \sqrt{\lambda_p^i \lambda_q^i}\right)$$



Reparameterization of the statistical model: Invariance, covariance and contravariance

- Smooth **reparameterization** of the model: $\mathcal{P} = \{p_\theta \ : \ \theta \in \Theta\} = \{p_\eta \ : \ \eta \in H\}$
- **The line element ds is invariant and hence Rao distance is invariant:**

$$\mathrm{d}s_\theta = \mathrm{d}s_\eta \qquad \rho_{\mathrm{Rao}}(p_{\eta_1}, p_{\eta_2}) = \rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})$$

- Fisher information matrix is **covariant**:

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

- Cramér-Rao bound is **contravariant**:

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I_\theta^{-1}(\theta) \xrightarrow{\eta=} \mathrm{Var}[\hat{\eta}_n] \succeq \frac{1}{n} \left[\frac{\partial \eta_i}{\partial \theta_j}\right] I_\theta(\theta(\eta))^{-1} \left[\frac{\partial \eta_i}{\partial \theta_j}\right]^\top$$

- Jacobian calculus: $\mathrm{Jac}_{\eta(\theta)} := \left[\frac{\partial \eta_i}{\partial \theta_j}\right] = (\mathrm{Jac}_{\eta^{-1}(\theta)})^{-1} = (\mathrm{Jac}_{\theta(\eta)})^{-1} := \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^{-1} \qquad \left[\frac{\partial \theta_i}{\partial \eta_j}\right] \times \left[\frac{\partial \eta_i}{\partial \theta_j}\right] = I_{D\times D}$

In practice, calculating Rao's distance can be difficult!

No closed form of Rao's distance between multivariate normals! (MVNs)

Two reasons for intractability:

1. Need to solve the **Ordinary Differential Equation** (ODE) for finding the **geodesic**:

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

But easy to solve when Γ=0: Line segments!

$$\ddot{\theta}=0\Rightarrow\theta(t)=(1-t)\theta_1+t\theta_2$$

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

$\rightarrow$ use the **Levi-Civita connection** derived from the metric tensor g

In general, geodesics depend on choice of the connection via Γ.

2. Need to **integrate** the infinitesimal length elements ds along the geodesics

Natural-gradient descent: Steepest Riemannian descent

Ordinary **gradient descent**:

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

- depends on the choice of the parameterization
- plateau phenomena near singularities

Covariant gradient:
Type mismatch on (M,g)

Contravariant gradient

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

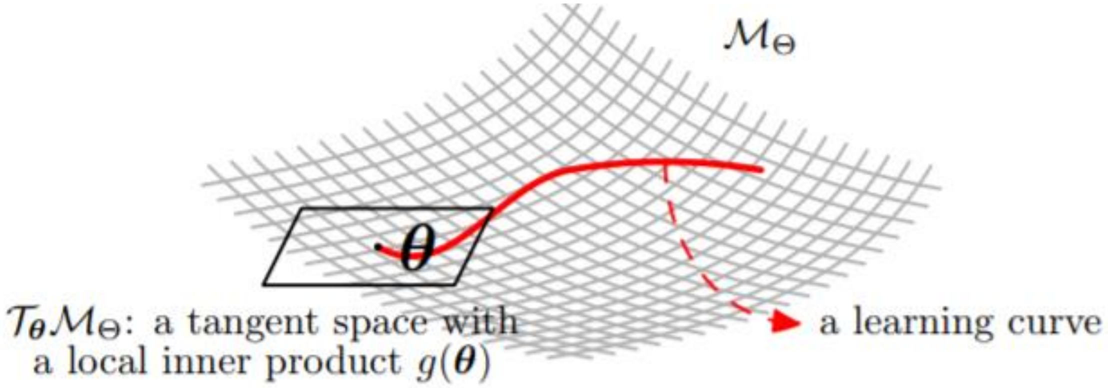
Natural gradient descent with **natural gradient** :

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

- NG **invariant** to reparameterization:

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

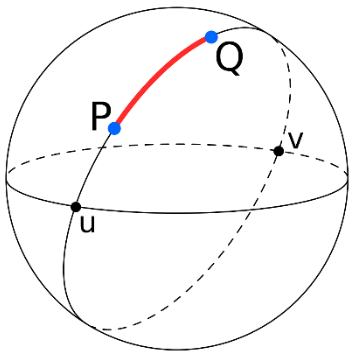
- avoids plateaus

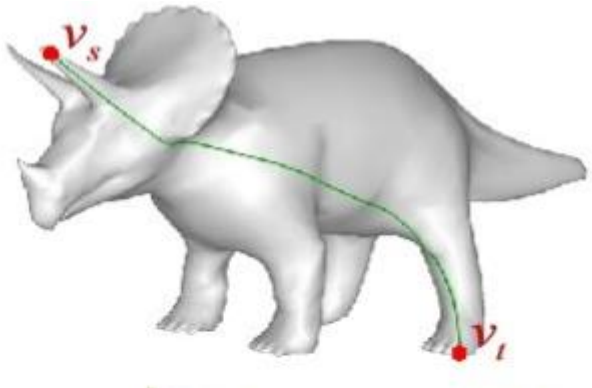


Amari, Natural gradient works efficiently in learning." Neural computation, 1998
Sun & N, Relative Fisher information and natural gradient for learning large modular models, ICML 2017
Li et al., Tractable structured natural gradient descent using local parameterizations, ICML 2021

First-principle of geodesics: Affine connections

- Riemannian geodesics are **locally minimizing length curves**





- *General definition* of geodesics is wrt. to an **affine connection**:

  For Riemannian geodesics, the default connection = **Levi-Civita connection**.

  This special Levi-Civita connection is derived from the metric tensor g.

- A geodesic γ(t) with respect to a connection ∇ is an **∇-autoparallel curve**

In physics, "straight" free fall particle

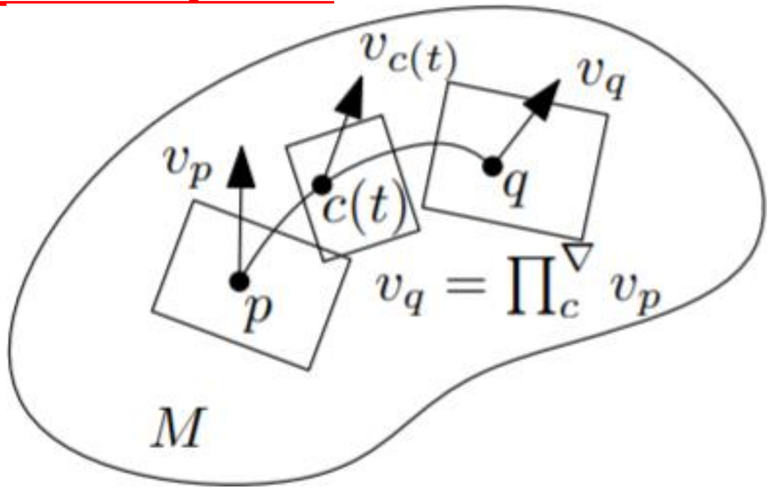
$$\boxed{\nabla_{\dot\gamma}\dot\gamma = 0, \quad \dot\gamma = \frac{d}{dt}\gamma(t)}$$

where $\nabla_X T$ is the **covariant derivative** of a tensor T wrt. a vector field X

[EIG 2020] An elementary introduction to information geometry, Entropy 22.10 (2020)

What makes the Levi-Civita connection so special?

An affine connection ∇ defines how to **∇-parallel transport** a vector from one tangent plane to another tangent plane



- **Fundamental theorem of Riemann geometry**:

Levi-Civita connection is the **unique torsion-free metric connection** induced by the metric tensor g

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

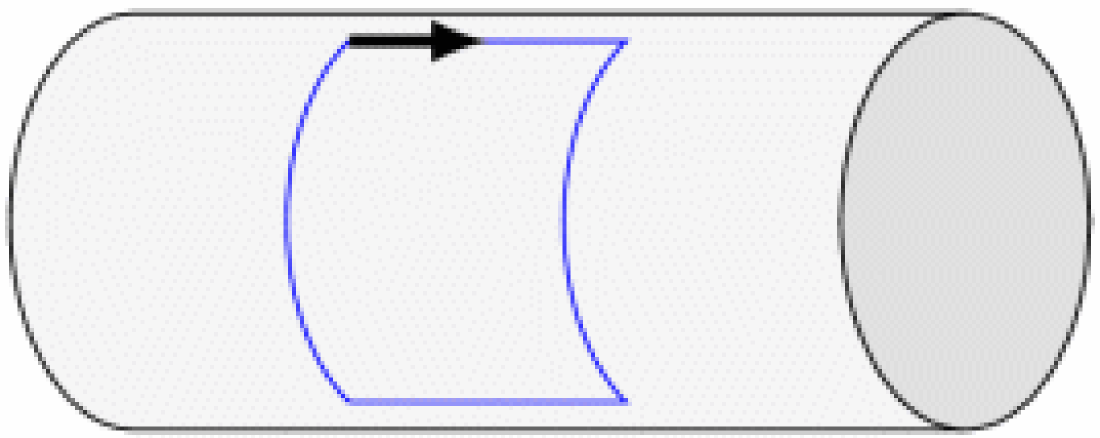
Metric compatible

$$\nabla = {}^g\nabla$$

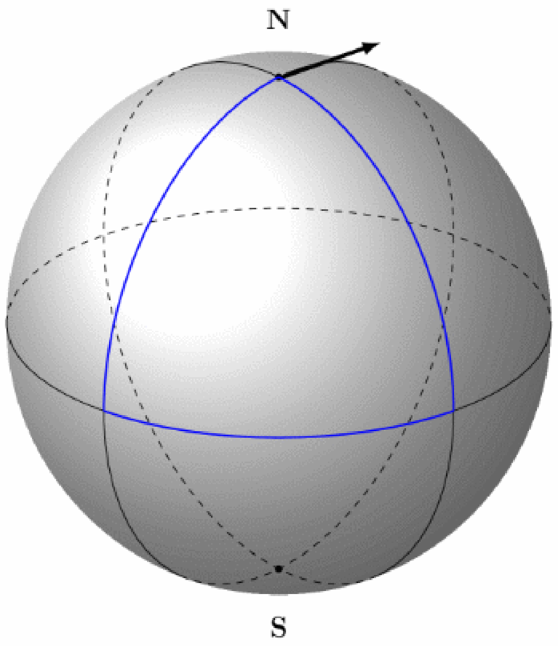
Levi-Civita connection

[EIG 2020]

Affine connection ∇: Curvature & parallel transport on infinitesimal loops



Cylinder is **flat:**
Parallel transport is
independent of path



Sphere has constant curvature:
Parallel transport is path-dependent

Élie Cartan
1869-1951

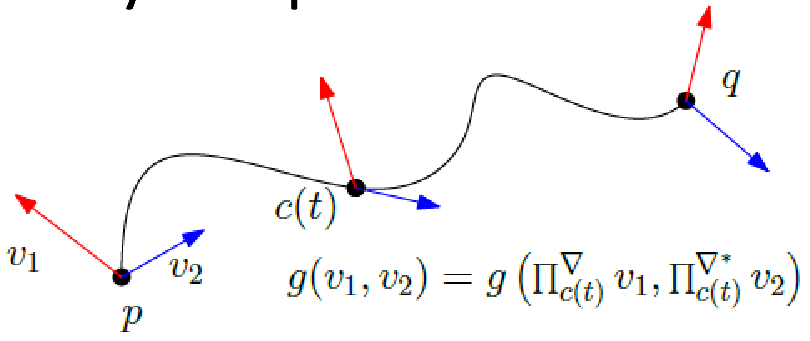
A connection is **flat** is there exists locally a coordinate system θ such that the Christoffel symbols **Γ** are all zero: **$\Gamma(\theta)=0$**

→ Geodesics plotted in that coordinate system are line segments

Dualistic information geometry: (M, g, ∇, ∇^*)

- Given an **affine torsion-free connection ∇ and a metric g**, we can build a **unique** dual affine torsion-free connection: the **dual connection** ∇^* such that the metric (inner product) is preserved by the primal and **metric-compatible dual parallel transports**:

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



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

- This amounts to say that ∇^* is defined uniquely by geometric equation:

$$\boxed{Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z),}$$

Meaning for each point p of M: $X_p g_p(Y_p, Z_p) = g_p((\nabla_X Y)_p, Z_p) + g_p(Y_p, (\nabla^*_X Z)_p).$

- The dual of a dual connection is the primal connection: $(\nabla^*)^* = \nabla.$

[EIG 2020]

Statistical invariance wrt sufficient statistics

- A **statistic** is a function of a random vector (e.g., mean, variance)
- A **sufficient statistic** collect and concentrate from a random sample

all necessary information for estimating the parameters.

Informally, a statistical lossless compression scheme…

- **Definition:** conditional distribution of X given t *does not depend* on θ

$$\boxed{\Pr(x|\theta) = \Pr(x|t)}$$

t=T(X) contains all Information about θ

- **Fisher-Neyman factorization theorem**: Statistic t(x) sufficient iff. the density can be decomposed as: $p(x;\lambda) = a(x) b_\lambda(t(x))$

Example: **Normal distributions have D=2 sufficient statistics**:

$$N(\mu, \sigma) \qquad t_1(X_1, \ldots, X_n) = \sum_i X_i \qquad t_2(X_1, \ldots, X_n) = \sum_i X_i^2$$

Statistical exponential families: A digest with flash cards, arXiv:0911.4863 (2009)

Natural exponential families: Finite sufficient statistics

- Consider a positive measure μ (usually counting or Lebesgue)
- A **natural exponential family** is a parametric family of densities that write as

$$\boxed{p(x;\theta) = \exp(\theta x - F(\theta))}$$

$$p(x;\lambda) = a(x) b_\lambda(t(x))$$

Sufficient statistic T(X)=X

where F is **real-analytic, strictly convex and differentiable**:

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

Natural parameter space $\Theta = \left\{\theta : \int \exp(\theta x) \mathrm{d}\mu(x) < \infty \right\}$

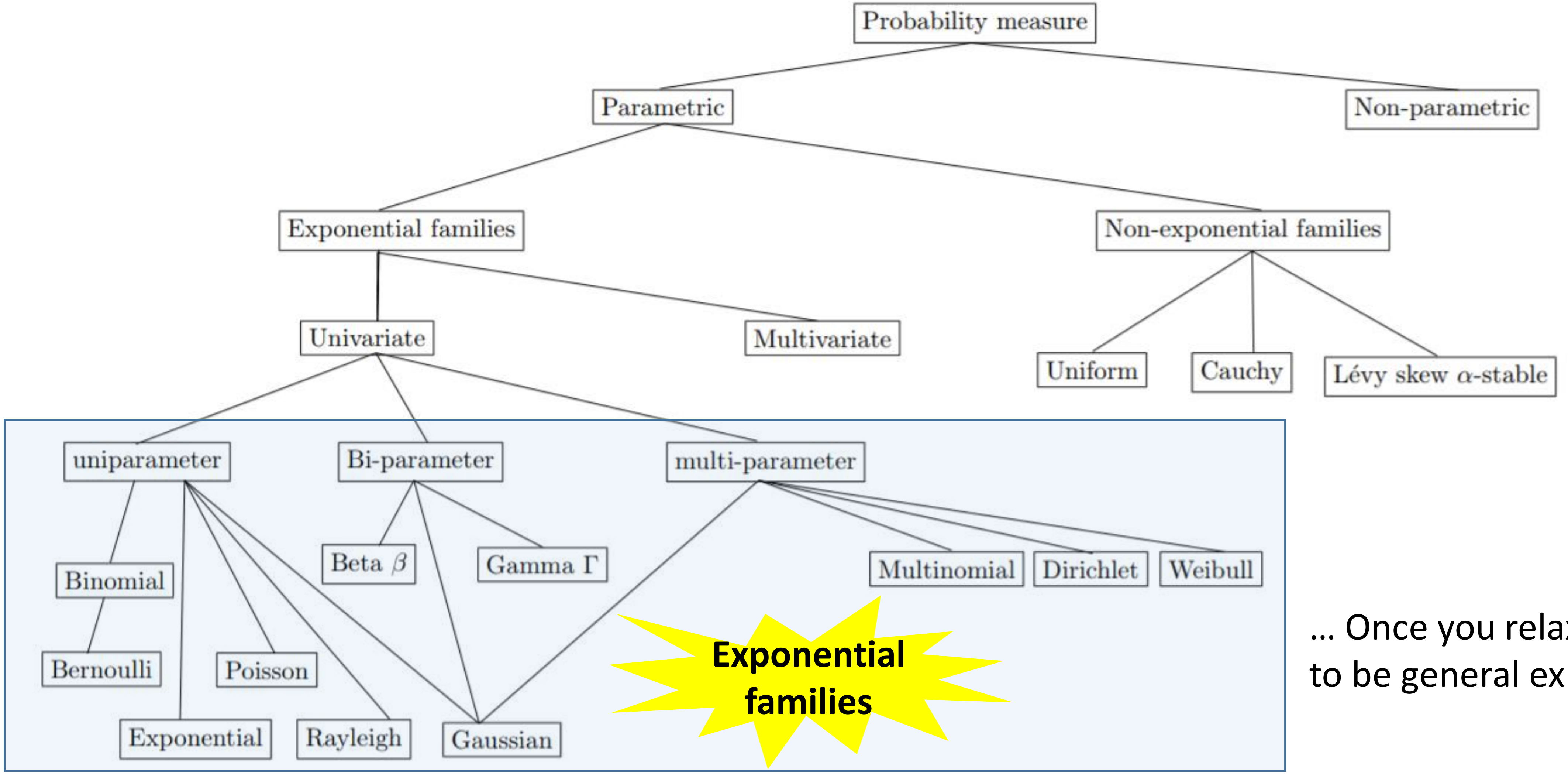
F: **Log-normalizer** (also known as log partition function or cumulant function)

Barndorff-Nielsen, Information and exponential families: in statistical theory. John Wiley & Sons, 2014

Sundberg, Statistical modelling by exponential families. Vol. 12. Cambridge University Press, 2019

N., Garcia, Statistical exponential families: A digest with flash cards." arXiv:0911.4863

Many common distributions are exponential families in disguise…



… Once you relax natural EFs
to be general exponential families

Statistical exponential families: A digest with flash cards, arXiv:0911.4863 (2009)
Tojo and Yoshino, On a method to construct exponential families by representation theory, GSI 2019 (Springer)

Exponential families: From Natural EFs to simply Efs!

- Consider a **(sufficient) statistic** t(x), **model order D**, **d-variate densities**
- Consider an **additional carrier measure term** k(x)
- Consider an **inner product** between t(x) and θ

  (usual scalar/dot product)

$$\mathrm{d}\nu(x) = e^{k(x)}\mathrm{d}\mu(x)$$

$$p(x;\lambda) = a(x) b_\lambda(t(x))$$

$$\boxed{p_\theta(x) = \exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))}$$

Sufficient statistic t(X)

Exponential families have <u>finite moments of any order</u>

Properties:

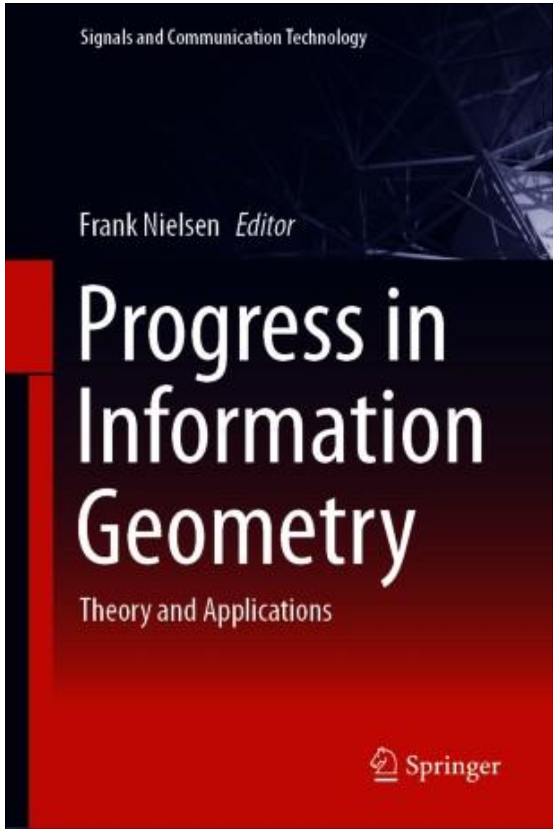
$$E[t(X)] = \nabla F(\theta)$$

$$\boxed{\mathrm{Cov}[t(X)] = \nabla^2 F(\theta) = I(\theta)}$$

Hessian of –log $p_\theta(x)$:
This is FIM of type 2

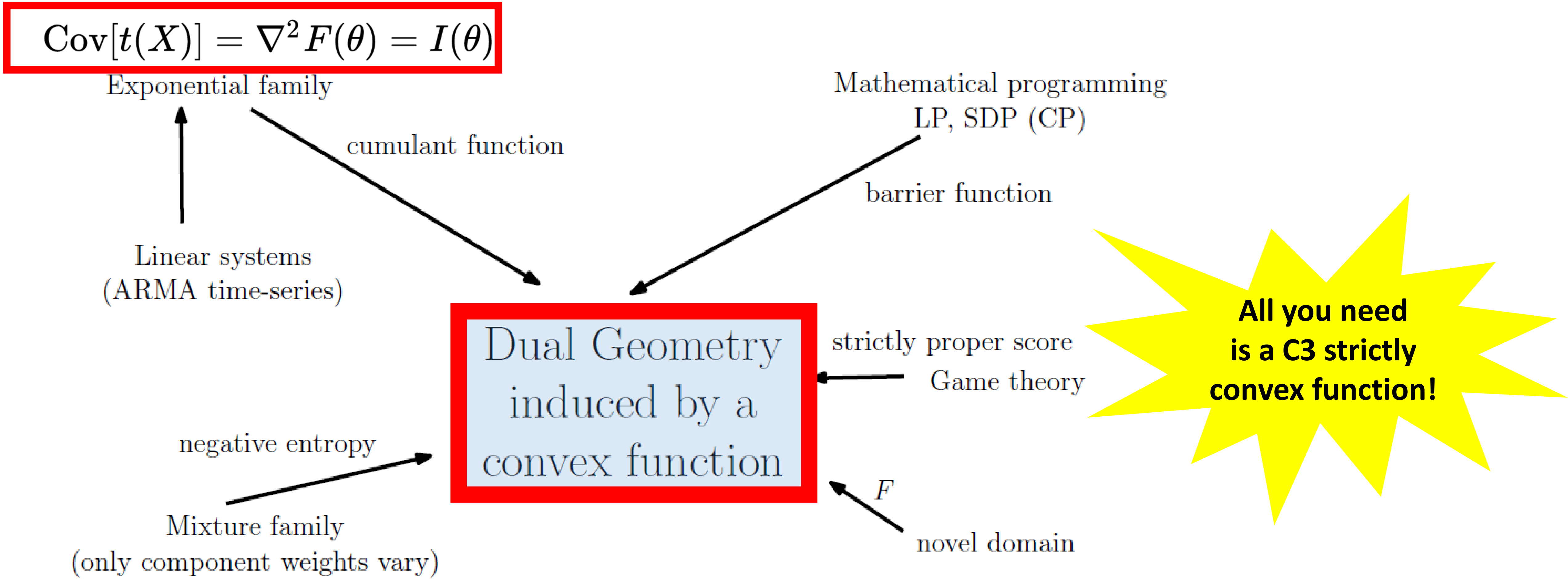
2. Bregman manifolds:

As known as…
…Dually flat spaces in IG



2021

Dually flat geometry from *any* strictly convex function



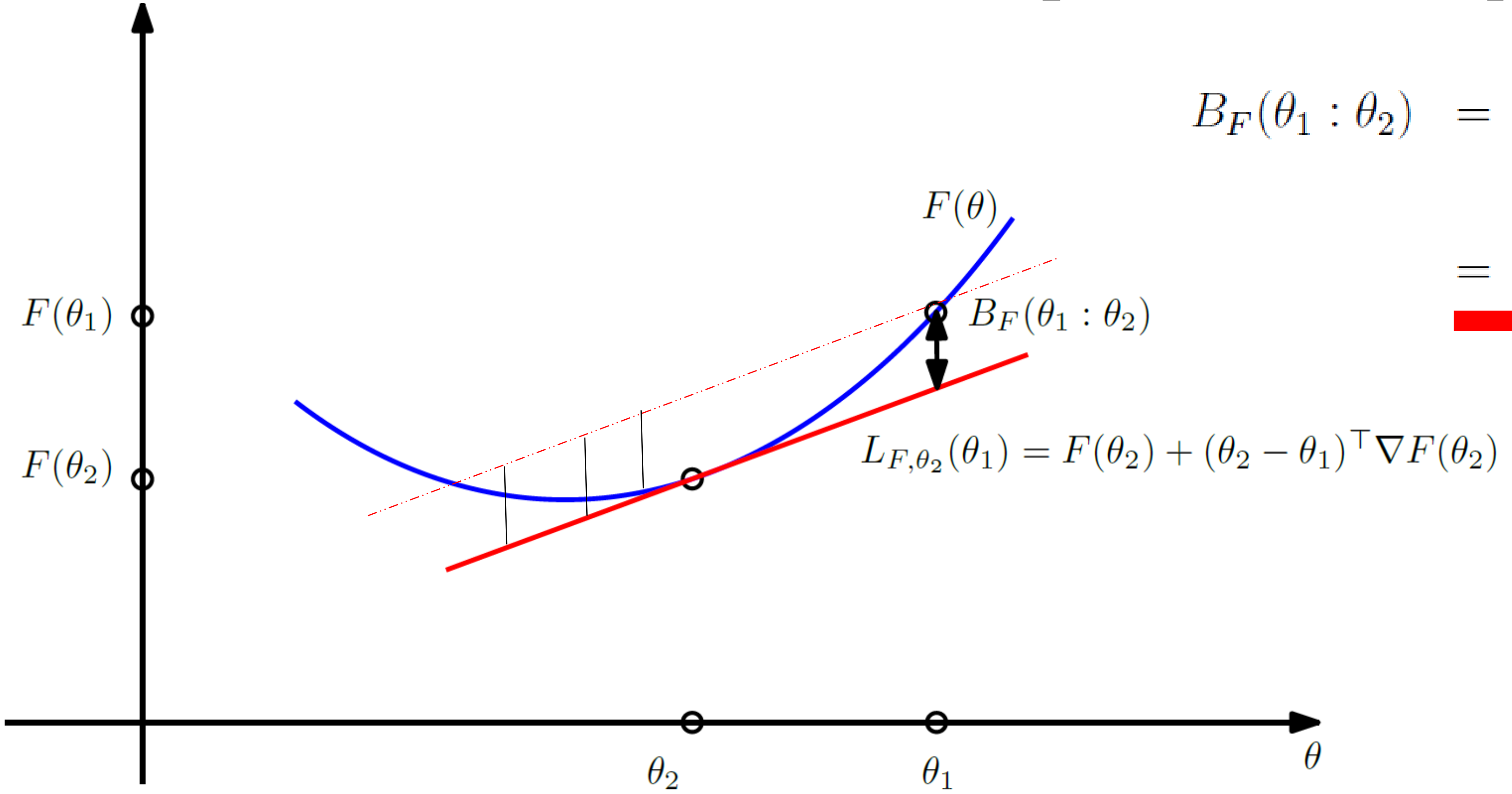
Bregman manifolds are not necessarily related to statistical models,

but can **always be realized by a regular statistical model**

Vân Lê, Hông. "Statistical manifolds are statistical models." *Journal of Geometry* 84.1-2 (2006)

Bregman divergences from strictly convex function

- $F(\theta)$: strictly convex and differentiable convex function on an open convex domain Θ
- Design the **Bregman divergence** as the vertical gap between $F(\theta_1)$ and the linear approximation of $F(\theta)$ at θ_2 evaluated at θ_1 :



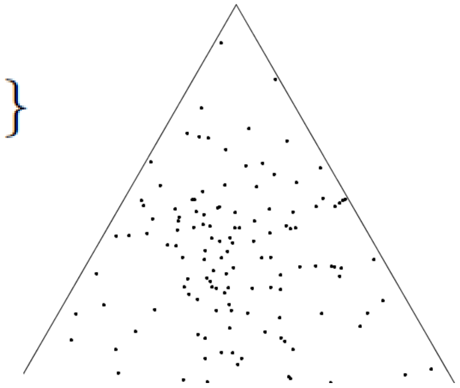
$$\begin{aligned} B_F(\theta_1 : \theta_2) &= F(\theta_1) - \underbrace{\left(F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2)\right)}_{L_{F,\theta_2}(\theta_1)} \\ &= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) \end{aligned}$$

[Bregman 1967]

Discrete Kullback-Leibler divergence: A non-separable Bregman divergence

- The KLD between two **categorical distributions** a.k.a. *multinoulli* amounts to a **non-separable Bregman divergence** on the **natural parameters** of the multinoulli distributions interpreted as an **exponential family**.

$$p_\lambda = (p_\lambda^1, \ldots, p_\lambda^d) \in \Delta_{d-1}^\circ, \quad \sum_{i=1}^d p_\lambda^i = 1 \qquad \theta^i = \log \frac{\lambda^i}{\lambda^D}, i \in \{1, \ldots, D = d-1\}$$



$$\mathcal{D}_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] := \sum_{i=1}^D \lambda_1^i \log \frac{\lambda_1^i}{\lambda_2^i} =: \underline{B_{F_{\mathrm{KL}}}(\theta_1 : \theta_2)}$$

$$F_{\mathrm{KL}}(\theta) = \log(1 + \sum_{i=1}^D \exp(\theta_i)) =: \underline{\mathrm{LogSumExp}_+(\theta_1, \ldots, \theta_D)}$$

LogSumExp is only convex but **$\mathrm{LogSumExp}_+$ is strictly convex** [NH 2019]

[NH 2019] Monte Carlo information-geometric structures, Geometric Structures of Information, 2019.
Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities, Entropy, 18(12), 2016

Legendre-Fenchel transformation: Duality

- 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.

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

<u>Concave programming:</u>

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

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

- Legendre-Fenchel transformation applies to any multivariate function
- 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

Duality regular exponential families/Bregman divergences

$$\boxed{B_F(\theta_1 : \theta_2) \;=\; F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)}$$

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

$$\boxed{\log p_F(x;\theta) = -B_{F^*}(t(x) : \eta) + F^*(t(x)) + k(x)}$$

Exponential Family $p_F(x\|\theta)$	$\Leftrightarrow$ duality	Dual Bregman divergence B_{F^*}
Spherical Gaussian	$\Leftrightarrow$	Squared Euclidean divergence
Multinomial	$\Leftrightarrow$	Kullback-Leibler divergence
Poisson	$\Leftrightarrow$	I-divergence
Geometric	$\Leftrightarrow$	Itakura-Saito divergence
Wishart	$\Leftrightarrow$	log-det/Burg matrix divergence

Maximum Likelihood Estimator = **Bregman centroid** for the dual convex conjugate:

$$\max_{\theta\in\mathbb{N}} \quad \bar{l}(\theta; x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(\langle t(x_i), \theta\rangle - F(\theta) + k(x_i)) \Leftrightarrow \min_{\eta\in\mathbb{M}} \quad \frac{1}{n}\sum_{i=1}^{n} B_{F^*}(t(x_i) : \eta)$$

Inference exponential families wrt F(θ)	Dual Bregman clustering wrt $F^*(\eta)$
Maximum likelihood (MLE)	Bregman centroid
Expectation/Maximization (EM) MEF	Bregman soft clustering
Classification EM = **k-MLE**	Bregman k-means

Banerjee et al., Clustering with Bregman divergences, JMLR 2005
k-MLE: A fast algorithm for learning statistical mixture models, ICASSP, arxiv:2012 1203.5181

Legendre-Fenchel transform: Mixed coordinates and dual Fenchel-Young divergences

- **Dual parameterizations**: $\theta = \nabla F^*(\eta) \Longleftrightarrow \eta = \nabla F(\theta)$
- Convex conjugate expressed as : $F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta))$
- To get in closed form the convex conjugate F^*, we need $\nabla F^*(\eta)$, i.e., **invert** $\nabla F(\theta)$: difficult in general! $\nabla F^* = (\nabla F)^{-1}$
- **Fenchel-Young inequality:** $F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2$

  with equality if and only if
- **Fenchel-Young divergence** use **mixed parameterization** θ/η: $\eta_2 = \nabla F(\theta_1)$

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

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius, Entropy 2021
Blondel et al., Learning with Fenchel-Young losses, JMLR 2020

Dual Bregman & dual Fenchel-Young divergences

- In general, **dual divergence** or **reverse divergence**: $D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$
- **Identity of dual Bregman divergences**: $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$
- **Primal**, **dual** or **mixed** parameterizations of Bregman divergences:

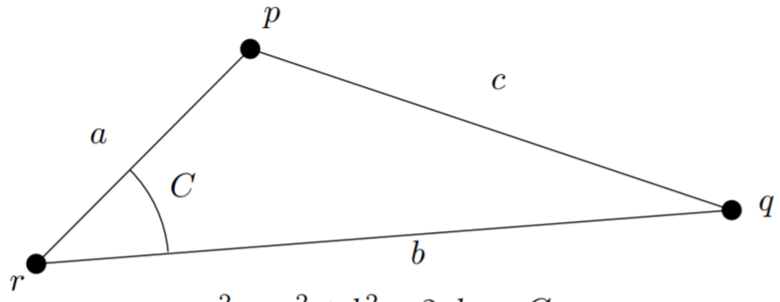
$$\boxed{B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2) = Y_{F^*,F}(\eta_2, \theta_1) = B_{F^*}(\eta_2 : \eta_1)}$$

On a Bregman manifold, 2^n equivalent formula with n terms!

3-parameter identity of Bregman divergences

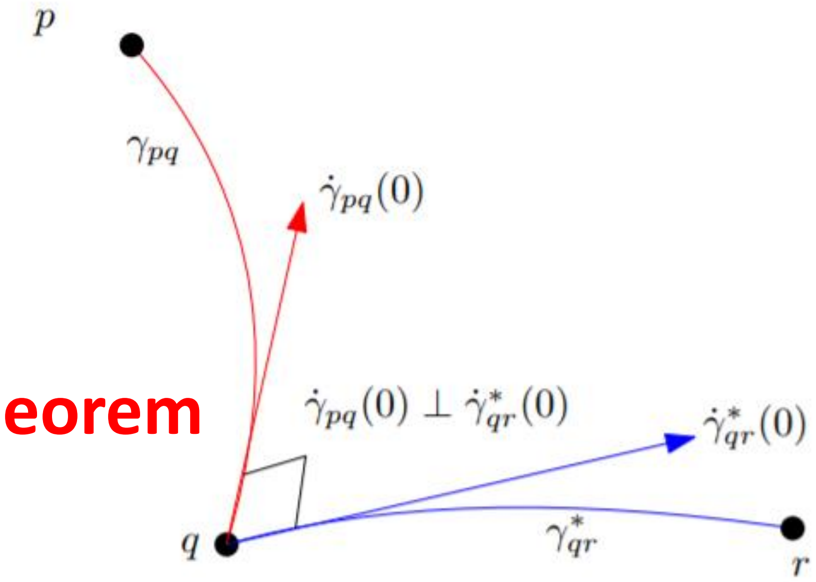
- Generalize the **law of cosines** for the squared Euclidean distance

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



$$c^2=a^2+b^2-2ab\cos C$$

- yields a **generalization of the Pythagorean theorem** when $(\theta_1-\theta_3)^\top(\nabla F(\theta_2)-\nabla F(\theta_3))\ =0$



$$D_F(p:r))=D_F(p:q)+D_F(q:r)$$

$$B_F(\theta(p):\theta(r))=B_F(\theta(p):\theta(q))+B_F(\theta(q):\theta(r))$$

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

On geodesic triangles with right angles in a dually flat space, Progress in Information Geometry: Theory and Applications, 2021

Statistical divergences between parametric models = parameter divergences

Statistical divergences between densities of a **parametric model** $\mathcal{F} = \{f_\theta(x)\}_\theta$ amount equivalently to (parameter) divergences between corresponding parameters:

$$\boxed{\mathcal{D}[f_{\theta_1} : f_{\theta_2}] =: D_{\mathcal{M}}(\theta_1 : \theta_2)}$$

For which statistical models and statistical divergences,

do we obtain $D_M(\theta_1 : \theta_2)$ as a Bregman divergence?

Example 1: Natural exponential family models & KLD*

- Parametric model $\mathcal{E} = \{e_\theta(x)\}_\theta$ with densities $e_\theta(x) = \exp\left(\sum_{i=1}^{D} t_i(x)\theta_i - F(\theta) + k(x)\right)$

- Examples of **natural exponential families**:
  - Exponential distributions (continuous): p.d.f. $\lambda e^{-\lambda x} \quad x \geq 0$
  - Poisson distributions (discrete): p.m.f. $\Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$

- Examples of **exponential families** with density $e_\lambda(x) = \exp\left(\sum_{i=1}^{D} t_i(x)\theta_i(\lambda) - F(\theta) + k(x)\right)$

  Gaussian distributions once reparameterized with natural parameters

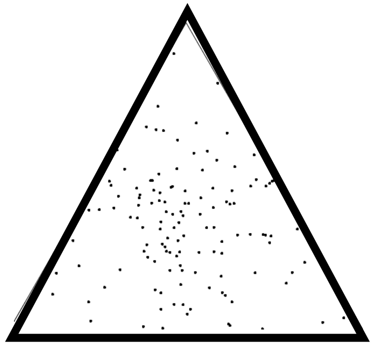
  $\theta(\lambda)=\theta(\mu,\sigma^2)$

- We have $\boxed{\mathcal{D}_{\mathrm{KL}}[e_{\theta_1} : e_{\theta_2}] = \underbrace{B_F{}^*(\theta_1 : \theta_2)}_{D_{\mathcal{E}}(\theta_1:\theta_2)} = B_F(\theta_2 : \theta_1)}$ with Bregman generator:

the **log-normalizer convex real-analytic function**: $F_{\mathcal{E}}(\theta) = \log\left(\int \exp(\sum_{i=1}^{D} t_i(x)\theta_i + k(x))\,\mathrm{d}\mu(x)\right)$

Example 2: Mixture family models & KLD

- Let 1, $p_0(x)$, …,$p_D(x)$ be (D+2) **linearly independent** densities



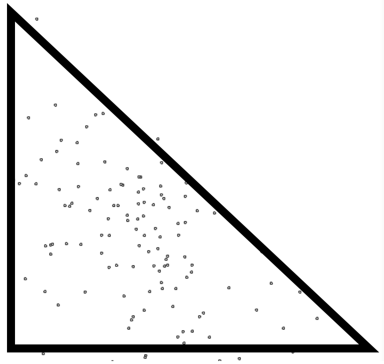
- **Mixture family** $\mathcal{M} = \{m_\theta(x)\}_\theta$ with densities:

$$m_\theta(x) = \sum_{i=1}^{D} w_i p_i(x) + \left(1 - \sum_{i=1}^{D} w_i\right) p_0(x)$$

- We have: $\boxed{\mathcal{D}_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}] = \underbrace{B_{F_{\mathcal{M}}}(\theta_1 : \theta_2)}_{D_{\mathcal{M}}(\theta_1:\theta_2)}}$

Information geometry/reconstruction

$$\theta = (w_1, \ldots, w_D)$$



Natural parameters

- with the Bregman generator = **Shannon negentropy**:

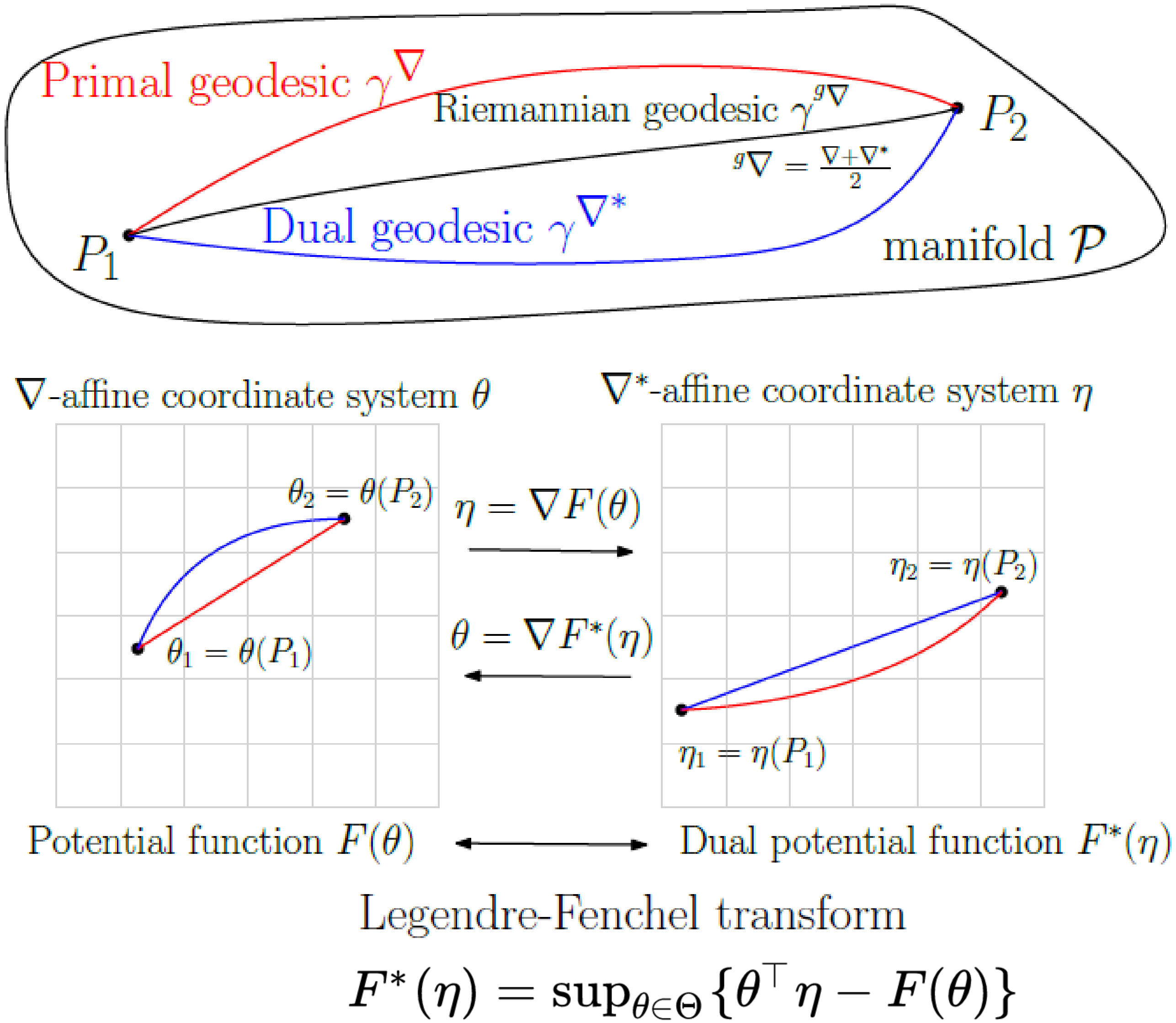
$$F_{\mathcal{M}}(\theta) = \int m_\theta(x) \log m_\theta(x) \mathrm{d}\mu(x)$$

Usually $F_M(\theta)$ not in closed-form…

But 2-mixture family of Cauchy distributions has closed-form!

The dually flat information geometry of the mixture family of two prescribed Cauchy components, arXiv:2104.13801

Bregman information geometry: Bregman manifolds



- Start from a potential function F(θ)

$$ {}^F g = \nabla^2 F(\theta) $$

- Get the dual potential function F*(η)

$$ {}^F g^* = \nabla^2 F^*(\eta) $$

- Define the primal flat connection:

$$ {}^F\Gamma_{ijk}(\theta) = 0 $$

- Define the dual flat connection:

$$ {}^F\Gamma^{*ijk}(\eta) = 0 $$

- Get the dual Bregman divergences

(or dual Fenchel-Young divergences)

In a Bregman manifold, natural gradient = ordinary gradient for the dual parameter!

On a **Bregman manifold**, we have

$$I_\theta(\theta) = \nabla_\theta^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$$

Natural gradient wrt θ

$$\begin{aligned}\tilde{\nabla}_\theta L_\theta(\theta) &:= I_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta) \\ &= (\nabla_\theta \eta)^{-1} \nabla_\theta \eta \nabla_\eta L_\eta(\eta) \\ &= \nabla_\eta L_\eta(\eta)\end{aligned}$$

Ordinary gradient wrt η

Used in variational inference (VI)

Khan & D. Nielsen, Fast yet simple natural-gradient descent for variational inference in complex models, ISITA 2018 arXiv:1807.04489

A note on the natural gradient and its connections with the Riemannian gradient, the mirror descent, and the ordinary gradient

Amari's α-geometry of probability families

$$\boxed{\{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\}_{\alpha\in\mathbb{R}}}$$

Dual structure

- Regular statistical parametric models $\mathcal{P}:=\{p_\theta(x)\}_{\theta\in\Theta}$
  (identifiable and finite positive-definite FIM)

- Amari's **α-connections** ∇^α

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

$$l(\theta;x) := \log L(\theta;x) = \log p_\theta(x)$$

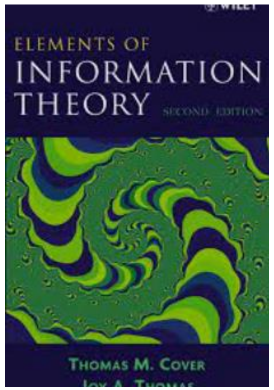
- 0-connection is **Fisher Levi-Civita connection**
- 1-connection is **exponential connection: flat for exponential families!**
- -1 connection is **mixture connection: flat for mixture families!**

NB: **A dually flat is usually not 0-flat**! (eg., normal manifolds)

Amari, Differential geometry of curved exponential families-curvatures and information loss, Annals of Statistics (1982)

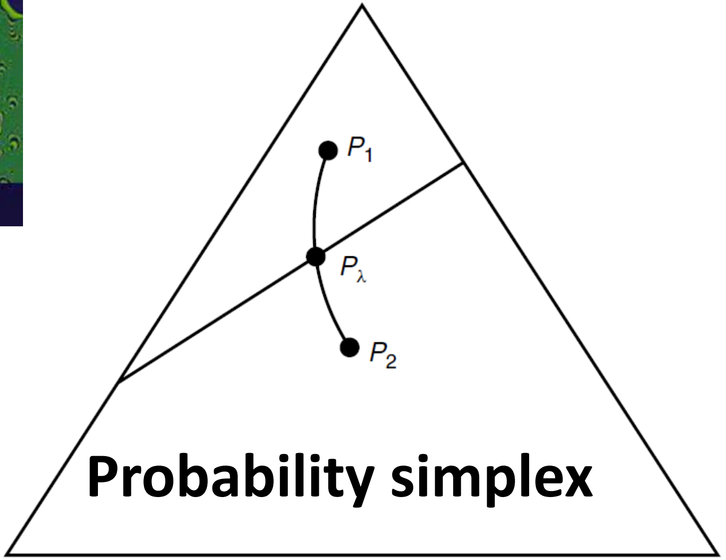
Chernoff information on exponential family manifolds

Probability of error in binary Bayesian hypothesis testing wrt MAP rule $P_e^n = 2^{-nC(P_1,P_2)}$
(equal prior, asymptotic regime)



$$C(P_1, P_2) \stackrel{\Delta}{=} -\min_{0\leq\lambda\leq 1} \log\left(\sum_x P_1^\lambda(x) P_2^{1-\lambda}(x)\right)$$

$$= D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2)$$



Geodesic exponential arc

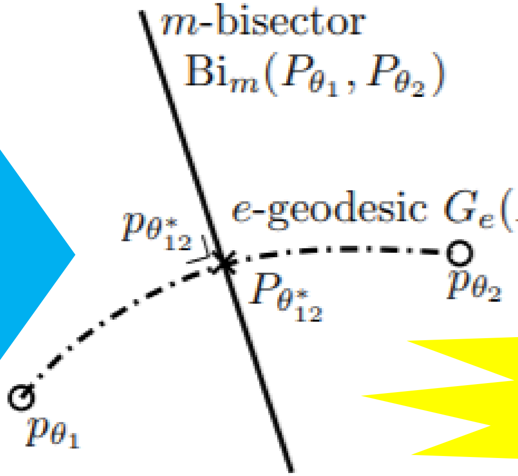
$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}$$

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

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

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

η-coordinate system



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

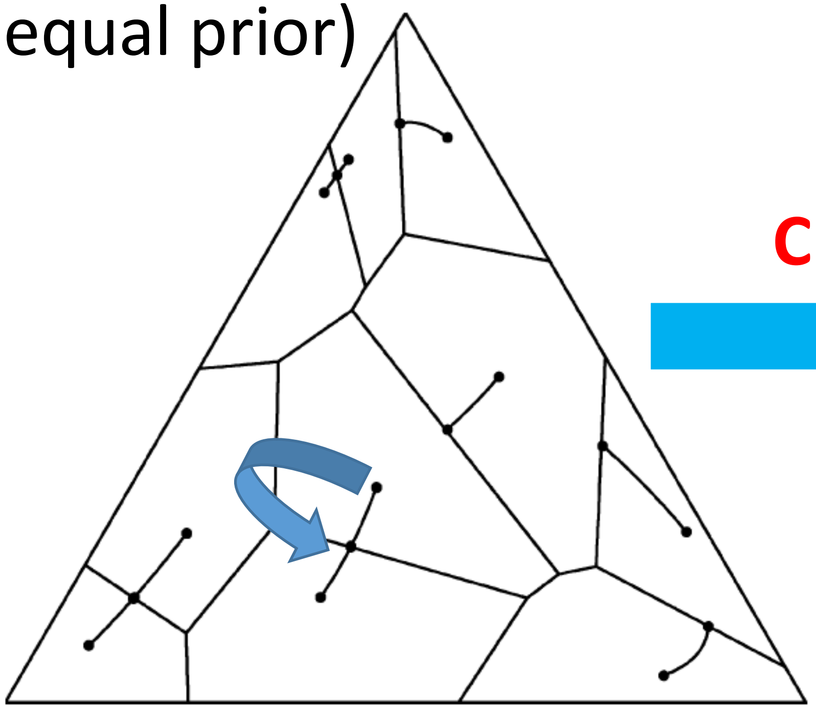
Geometrically exact!

Exponential family manifold

An information-geometric characterization of Chernoff information, IEEE Signal Processing Letters, 2013

Chernoff information: Multiple hypothesis testing

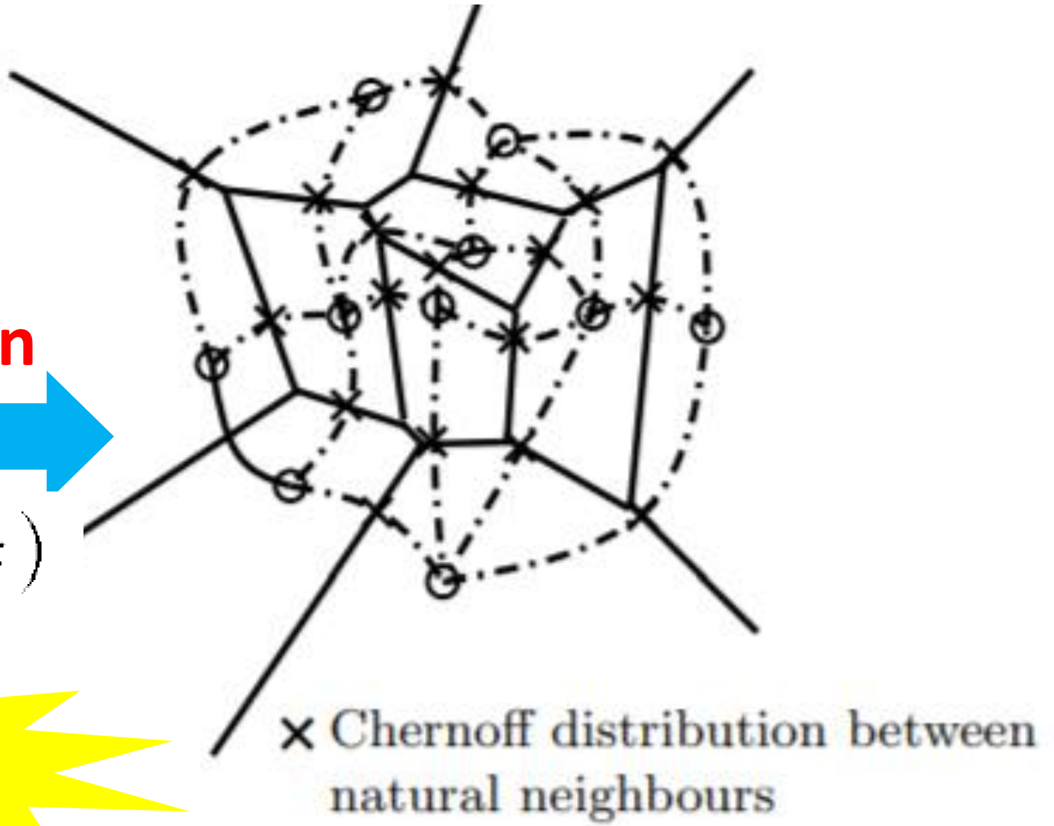
Probability of error: $P_e^n = 2^{-nC(P_i^*, P_j^*)}$
(equal prior)



Closest pair with respect to Chernoff information

$$\arg\min_{i \neq j} C(P_i, P_j)$$

Computational Geometry!

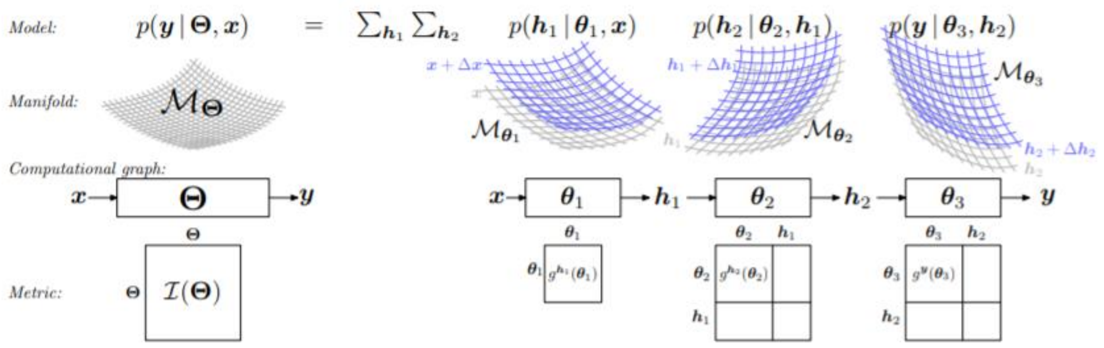


Probability simplex manifold
Kullback-Leibler Voronoi diagram

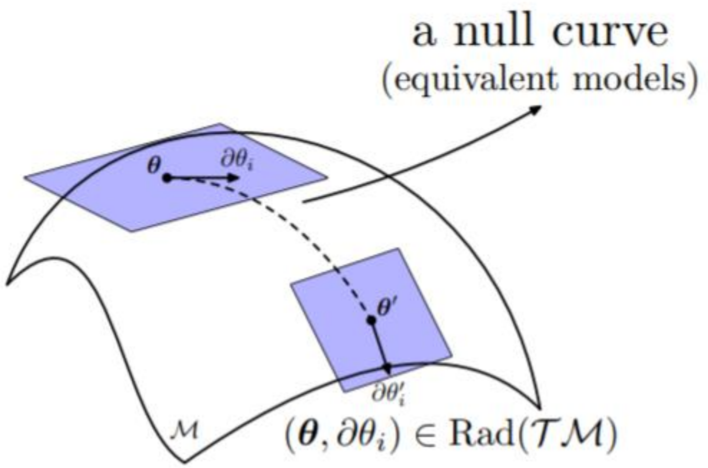
Exponential family manifold
Bregman Voronoi Diagrams

Westover, Asymptotic geometry of multiple hypothesis testing, IEEE Trans. IT, 2008
Hypothesis testing, information divergence and computational geometry, GSI 2013, Springer LNCS
Bregman Voronoi diagrams, Discrete & Computational Geometry, 2010

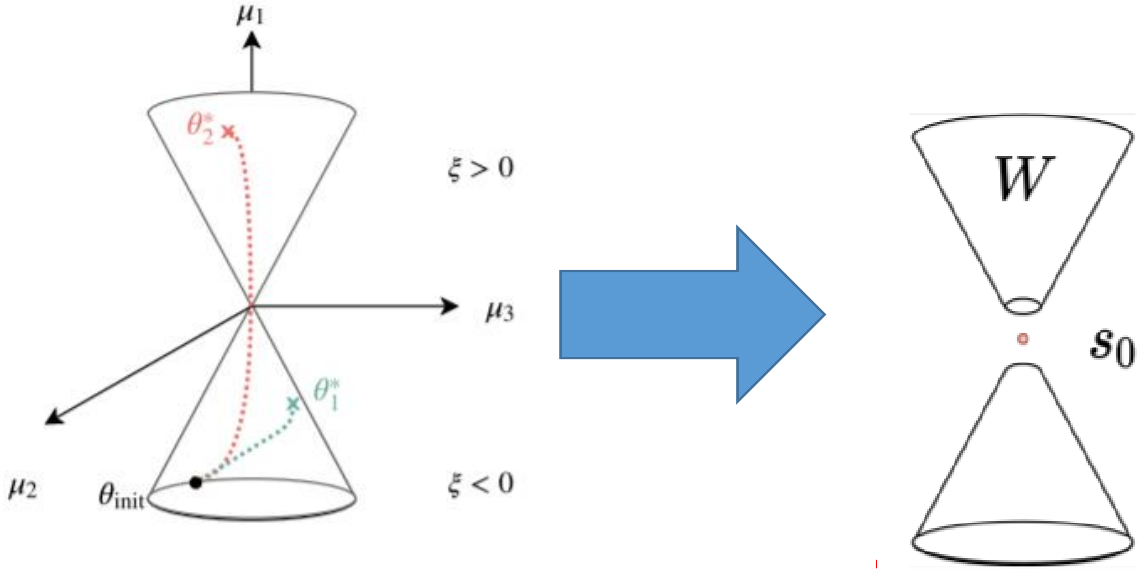
Challenge: IG/NGD for large-size hierarchical singular NN models!



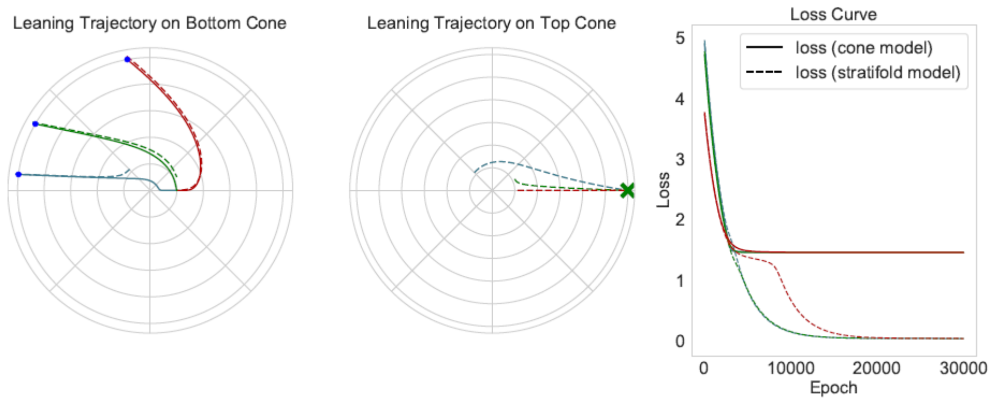
Relative Fisher information matrix, Relative NGD



Semi-Riemannian geometry: **Lightlike manifolds**



Sometimes you need to go through singularities!



Relative Fisher Information and Natural Gradient for Learning Large Modular Models, ICML 2017
Lightlike Neuromanifolds, Occam's Razor and Deep Learning, arXiv:1905.11027
Towards Modeling and Resolving Singular Parameter Spaces using Stratifolds, OPT2021, arXiv:2112.03734

Fisher-Rao Riemannian geometry vs Amari's dual α-geometry

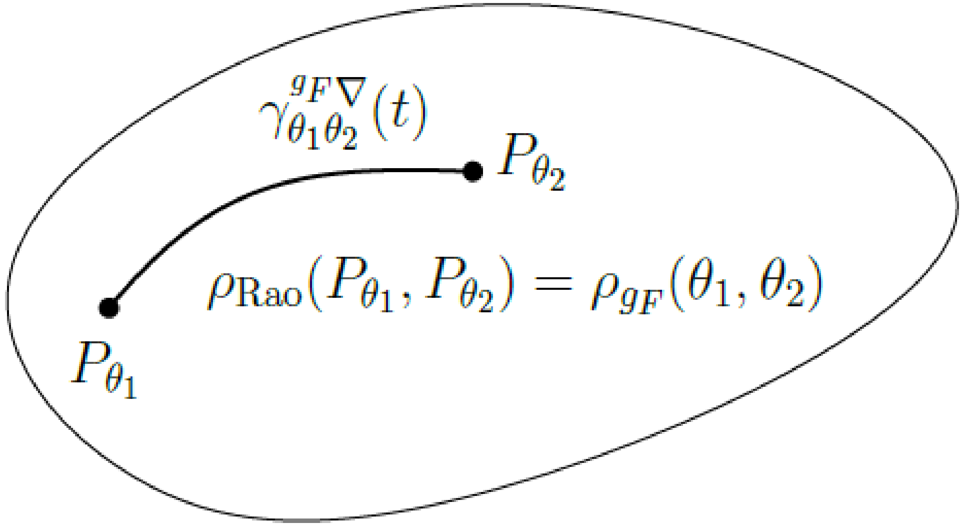
- From the viewpoint of **statistical invariance**, **Fisher information metric** is unique (up to a scaling factor):

Riemannian manifold with **Rao distance**

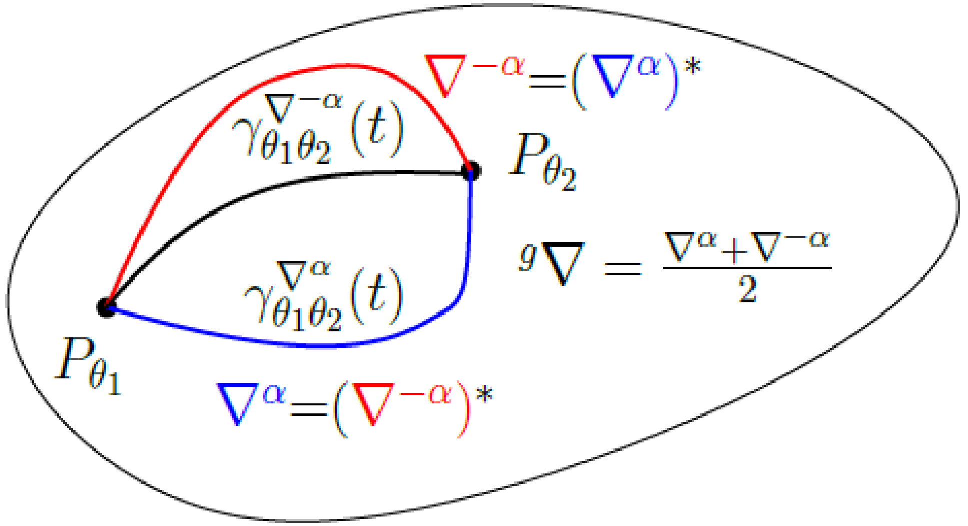
- Given a parametric statistical model, get a **dualistic α-geometry**

- For exponential families and mixture families, **±1-structure** yields **Bregman manifolds** (**dually flat spaces**) with **generalized Pythagoras theorems**

Fisher-Rao geometry
→ Fisher-Rao geodesic distance



versus



Dual α-geometry
→ No default divergence

Thank you very much for your attention.

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 probability distributions, called the statistical model, as a Riemannian manifold equipped with the Fisher metric 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

$$I_X(\theta) := E_{p_\theta}\left[s_\theta(x)s_\theta(x)^\top\right],$$

AMS Notices feature article, January 2022
8 pages + 1 historical poster

THE GRADUATE STUDENT SECTION

WHAT IS...

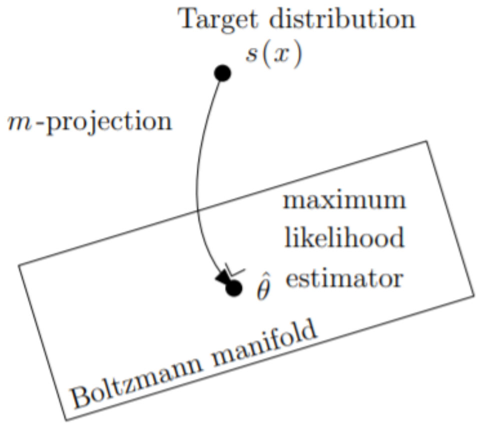
an Information Projection?

Frank Nielsen

Communicated by Cesar E. Silva

Orthogonal Projections as Distance Minimizers

In Euclidean geometry, the orthogonal projection p_S of a vector p onto a subset S as in Figure 1 can be defined as the point(s) q of S minimizing the distance $D(p,q)$ from p to q. In general, the projection may not be unique: for example, projecting the center of a unit ball onto its boundary sphere yields the full boundary sphere. However, the projection p_S is always guaranteed to be unique when S is an affine subspace.



we use the notation $D(p:q)$ to highlight the asymmetric property of information distances and call $D(p:q)$ a *divergence*, assumed to be infinitely differentiable.

Here the word "divergence" is not to be confused with the divergence operator from calculus. Similar to the Euclidean case, an information projection of $p \in M$ onto $S \subset M$ can be defined by minimizing the divergence $D(q:p)$ for $q \in S$. Since the divergence is asymmetric, we define a dual divergence $D^*(p:q) = D(q:p)$.

AMS Notices, March 2018
3 pages


Applied Mathematical Sciences
Shun-ichi Amari
Information Geometry and Its Applications
Springer
Information Geometry and Its Applications
Springer
Translations of MATHEMATICAL MONOGRAPHS
Volume 53
Statistical Decision Rules and Optimal Inference
N. N. Čencov
American Mathematical Society
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
Springer Complexity
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
Cambridge Monographs on Applied and Computational Mathematics
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
CAMBRIDGE
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
Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge
A Series of Modern Surveys in Mathematics 64
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

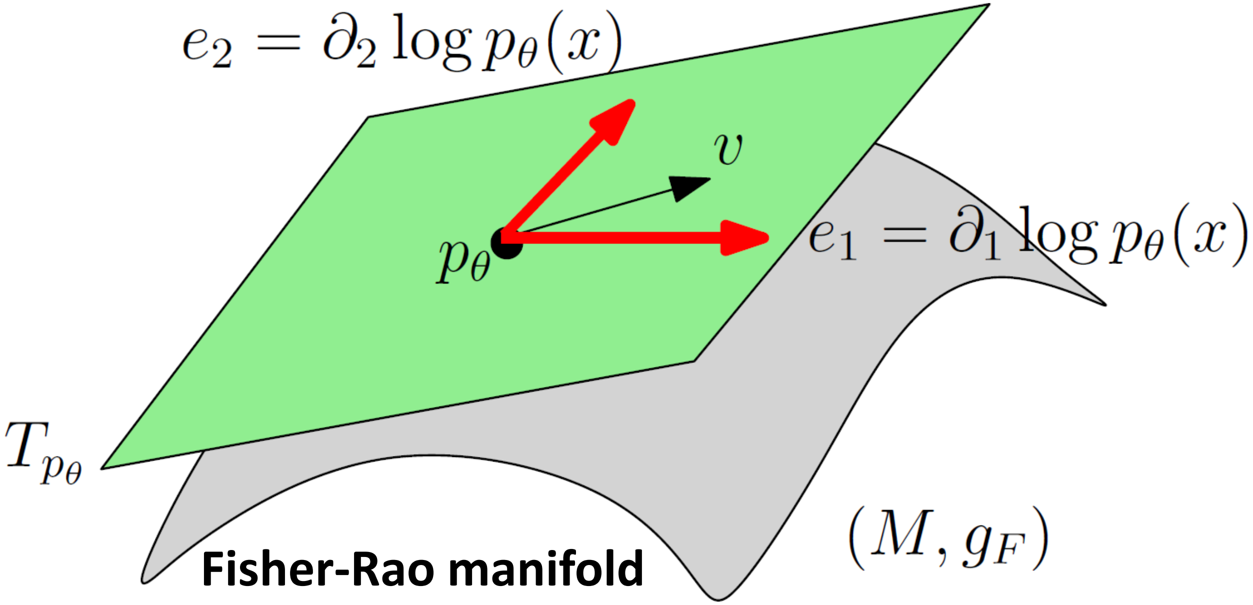
Fisher-Rao manifolds: Interpreting the inner product

- (M, g_F): Riemannian manifold equipped with the **Fisher information metric**
- **Inner product at tangent plane T_p expressed using the metric tensor g:**

$$\langle v_1, v_2 \rangle_p = [v_1]_{\mathcal{B}}^\top [g_{ij}(p)]_{\mathcal{B}} [v_2]_{\mathcal{B}} \quad \text{using basis} \quad \mathcal{B} = \{e_1, \ldots, e_D\}$$

$$\langle v_1, v_2 \rangle_p = [v_1]_{\mathcal{B}'}^\top [g_{ij}(p)]_{\mathcal{B}'} [v_2]_{\mathcal{B}'} \quad \text{using basis} \quad \mathcal{B}' = \{e'_1, \ldots, e'_D\}$$

- **Interpret back tangent planes** and **inner product** from statistical viewpoint:



Vector expressed using **score functions**:

$$v = \sum_{i=1}^{D} v^i \partial_i l_\theta(x) \qquad l_\theta(x) = \log p_\theta(x)$$

Basis wrt 1-resp: $\mathcal{B}_1 = \{\partial_1 l_\theta(x), \ldots, \partial_D l_\theta(x)\}$

Fisher-Rao inner product as **expectation**:

$$\boxed{\langle v_1, v_2 \rangle_{g_F(p_\theta)} = E_{p_\theta}\left[[v_1]_{\mathcal{B}_1}^\top [v_2]_{\mathcal{B}_1}\right]}$$

Other basis: **α-representations** with inner product expressed as **α-expectations**

Other information metrics

- Energetic information metric
- Wasserstein information metrics [LZ 2019]
- φ-entropy metrics (e.g., entropy metric of order α) [AR 2008]

Adrian & Rangarajan, Information geometry for landmark shape analysis: Unifying shape representation and deformation, IEEE TPAMI 2008
Lightlike Neuromanifolds, Occam’s Razor and Deep Learning,arXiv:1905.11027
Li & Zhao, Wasserstein information matrix." arXiv preprint arXiv:1910.11248, 2019

Bhattacharyya arc: Likelihood Ratio Exponential Family

- **Bhattacharyya arc** or **Hellinger arc** induced by two mutually absolutely continuous **arbitrary distributions** p and q (same support $\mathcal{X}$):

$$\mathcal{E}(p,q) := \left\{ p_\lambda(x) := \frac{p^{1-\lambda}(x) q^\lambda(x)}{Z_\lambda^G(p,q)}, \quad \lambda \in (0,1) \right\} \qquad Z_\lambda^G(p,q) := \int_{\mathcal{X}} p^{1-\lambda}(x) q^\lambda(x) \mathrm{d}\mu(x)$$

- Strictly convex log-normalizer F(λ) (i.e., Z is strictly log-convex)
- Bhattacharyya arc (geometric mixtures) = **1D exponential family:**

$$\begin{aligned}
p_\lambda(x) &= \frac{p_0^{1-\lambda}(x) p_1^\lambda(x)}{Z_\lambda^G(p,q)} \\
&= p_0(x) \exp\left(\lambda \log\left(\frac{p_1(x)}{p_0(x)} \right) - \log Z_\lambda^G(p,q) \right) \\
&= \exp\left(\lambda t(x) - F(\lambda) + k(x) \right)
\end{aligned}$$

$$\boxed{\begin{aligned}
F(\lambda) &:= \log(Z_\lambda^G(p,q)) = \log\left(\int_{\mathcal{X}} p^{1-\lambda}(x) q^\lambda(x) \mathrm{d}\mu(x) \right) \\
&=: -D_\lambda^{\mathrm{Bhat}}[p:q]
\end{aligned}}$$

Log-likelihood sufficient statistics:

$$t(x) := \log\left(\frac{p_1(x)}{p_0(x)} \right)$$

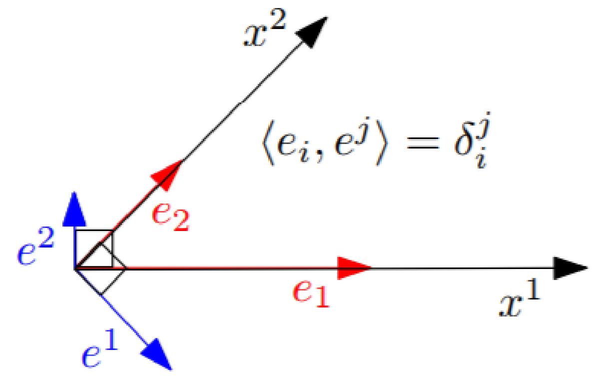
Base measure is p_0 $\quad k(x) := \log p_0(x)$

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

Generalizing the Geometric Annealing Path using Power Means, UAI 2021

Likelihood Ratio Exponential Families, NeurIPS Workshop on Deep Learning through Information Geometry 2020

Metric tensor using covariant/contravariant notations



Reciprocal basis

2-covariant metric tensor in local coordinates:

$$g_{ij}(\theta) = \nabla^2 F(\theta)$$

Dual metric tensor in local coordinates:

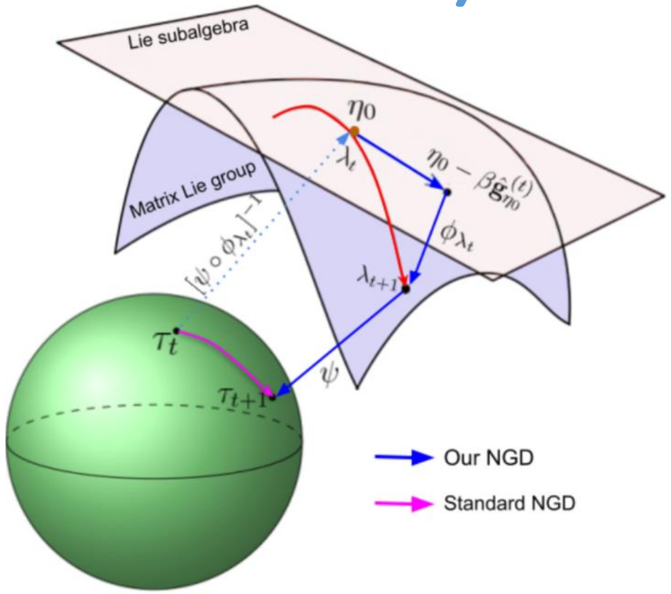
$$g^{ij}(\eta) = g^{*ij}(\eta) = \nabla^2 F^*(\eta)$$

Crouzeix's identity: x of Hessians of convex conjugates= Id:

$$\nabla^2 F(\theta) \nabla^2 F^*(\eta) = I$$

An elementary introduction to information geometry." Entropy 22.10 (2020)

Structured natural-gradient descent (Struct-NGD)



- Consider the **general optimization problem**:

$$\min_{\tau \in \Omega_\tau} \mathcal{L}(\tau) := \mathrm{E}_{q(\mathrm{w}|\tau)}\left[\ell(\mathbf{w})\right] + \gamma \mathrm{E}_{q(\mathrm{w}|\tau)}\left[\log q(w|\tau)\right]$$

- Standard natural-gradient descent (without structure):

$$\tau_{t+1} \leftarrow \tau_t - \beta \left[\mathbf{F}_\tau(\tau_t)\right]^{-1} \nabla_{\tau_t} \mathcal{L}(\tau)$$

- Natural-gradient descent **preserving structure** using **local parameterization**:

$$\boxed{\begin{aligned} \lambda_{t+1} &\leftarrow \phi_{\lambda_t}\left(\eta_0 - \beta \hat{\mathbf{g}}_{\eta_0}^{(t)}\right) \\ \tau_{t+1} &\leftarrow \psi(\lambda_{t+1}) \end{aligned} \quad \text{with} \quad \hat{\mathbf{g}}_{\eta_0}^{(t)} = \mathbf{F}_\eta(\eta_0)^{-1} \left[\nabla_{\eta_0}\left[\psi \circ \phi_{\lambda_t}(\eta)\right] \nabla_{\tau_t} \mathcal{L}(\tau)\right]}$$

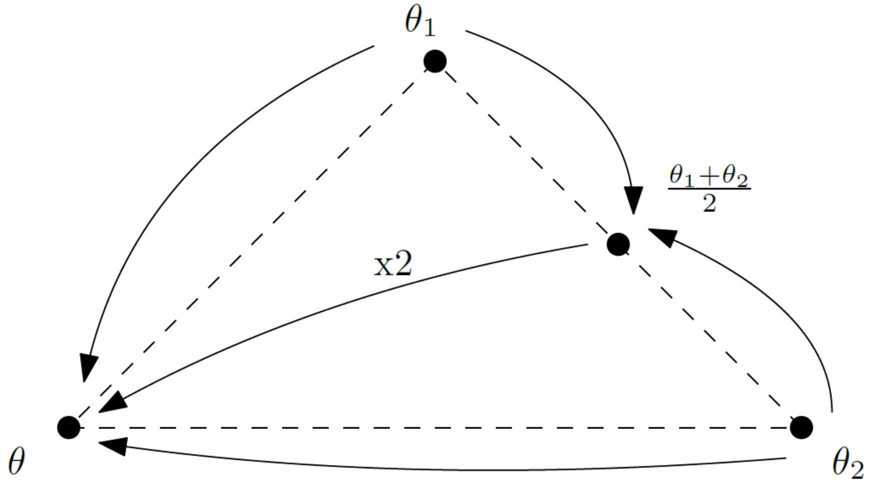
- worked examples on matrix Lie groups and applications: generalizes NGD & xNES evolutionary strategy, recovers Newton-like algorithms, obtained new structured second-order algorithms, etc.

Li et al., Tractable structured natural gradient descent using local parameterizations, ICML 2021

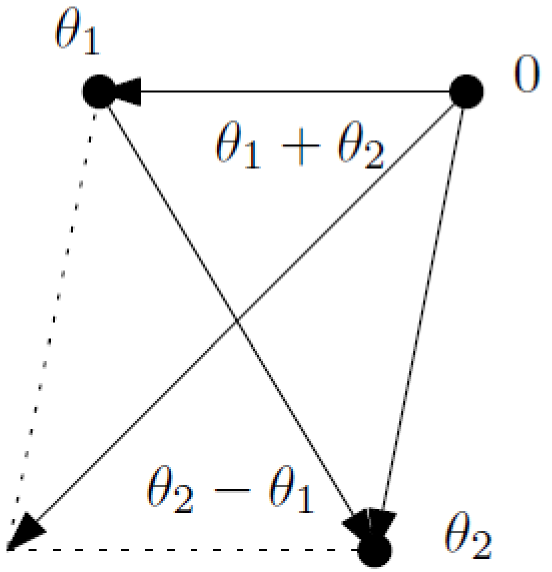
4-parameter identity of Bregman divergences

- Parallelogram identity

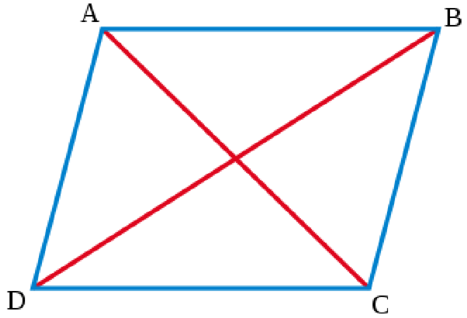
$$B_F(\theta_1 : \theta) + B_F(\theta_2 : \theta) = B_F\left(\theta_1 : \frac{\theta_1+\theta_2}{2}\right) + B_F\left(\theta_2 : \frac{\theta_1+\theta_2}{2}\right) + 2B_F\left(\frac{\theta_1+\theta_2}{2} : \theta\right)$$



$$B_F(\theta_1 : \theta) + B_F(\theta_2 : \theta) = B_F\left(\theta_1 : \frac{\theta_1+\theta_2}{2}\right) + B_F\left(\theta_2 : \frac{\theta_1+\theta_2}{2}\right) + 2\,B_F\left(\frac{\theta_1+\theta_2}{2} : \theta\right)$$



- In Euclidean geometry:



$$2AB^2 + 2BC^2 = AC^2 + BD^2$$

$$2\,\|\theta_1\|^2 + 2\,\|\theta_2\|^2 = \|\theta_1 - \theta_2\|^2 + \|\theta_1 + \theta_2\|^2$$

On geodesic triangles with right angles in a dually flat space, Progress in Information Geometry: Theory and Applications, 2021

Class of Bregman generators modulo affine terms
& KLD between exponential family densities expressed as log-ratio

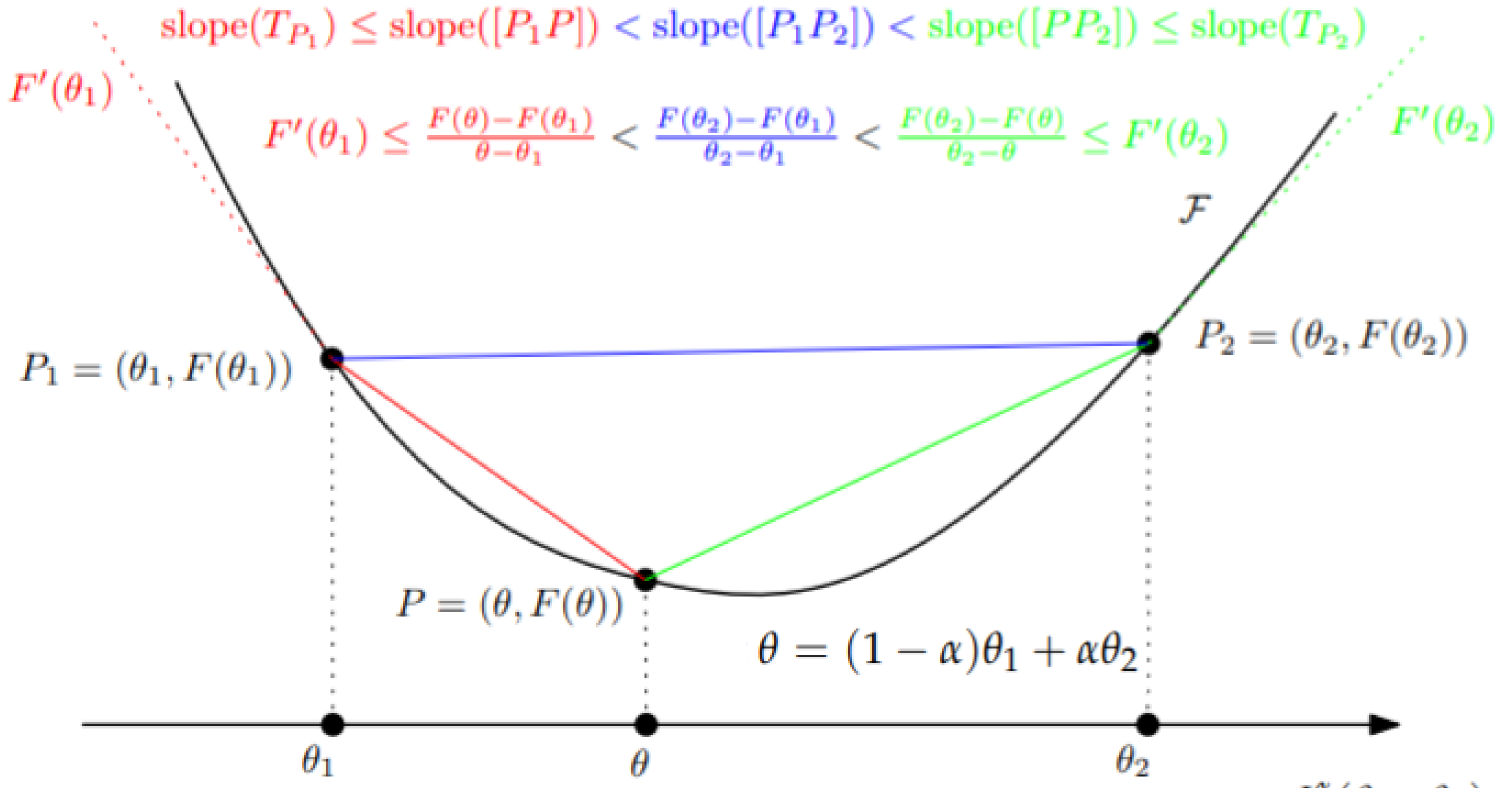
- Bregman generators are strictly convex and differentiable convex functions defined modulo affine terms: **$B_F=B_G$ iff. $F(\theta)=G(\theta)+A\theta+b$**
- Choose for **any** ω in the support of the exponential family the Bregman generator:

$$F_\omega(\theta) := -\log(p_\theta(\omega)) = F(\theta) - \underbrace{(\theta^\top t(\omega) + k(\omega))}_{\text{affine term in } \theta}$$

- We get: $D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \log\left(\frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)}\right) + (\theta(\lambda_2) - \theta(\lambda_1))^\top (t(\omega) - \nabla F(\theta(\lambda_1))), \quad \forall \omega \in \mathcal{X}$
- By choosing s points: $\boxed{D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s}\sum_{i=1}^{s} \log\left(\frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)}\right) \text{ such that } \frac{1}{s}\sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)]}$

Computing Statistical Divergences with Sigma Points. GSI 2021
Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family, arXiv:2003.02469

Chordal slope lemma & Jensen/Bregman divergences



Jensen Divergence (JD)

$$\frac{F(\theta)-F(\theta_1)}{\alpha(\theta_2-\theta_1)} < \frac{F(\theta_2)-F(\theta_1)}{(\theta_2-\theta_1)},$$

$$\Downarrow$$

$$F(\theta)-F(\theta_1) < \alpha(F(\theta_2)-F(\theta_1)),$$

$$\Downarrow$$

$$\alpha(F(\theta_2)-F(\theta_1)) - F(\theta) + F(\theta_1) > 0,$$

$$\Downarrow$$

$$J_F^{\alpha}(\theta_1:\theta_2) := (1-\alpha)F(\theta_1) + \alpha F(\theta_2) - F((1-\alpha)\theta_1 + \alpha\theta_2) > 0.$$

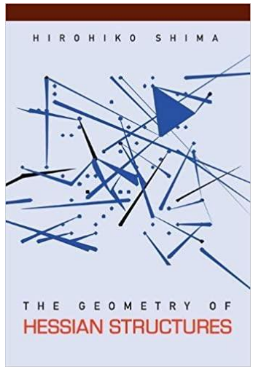
Bregman Divergences (BDs):

$$F'(\theta_1) \leq \frac{F(\theta_2)-F(\theta_1)}{\theta_2-\theta_1} \leq F'(\theta_2) \Rightarrow \begin{array}{rcl} F(\theta_2)-F(\theta_1)-(\theta_2-\theta_1)F'(\theta_1) & \geq & 0, \\ F(\theta_2)-F(\theta_1)-(\theta_2-\theta_1)F'(\theta_2) & \leq & 0. \end{array} \Rightarrow \begin{array}{rcl} B_F(\theta_2:\theta_1) & \geq & 0, \\ B_F(\theta_1:\theta_2) & \geq & 0. \end{array}$$

BD as a limit of a scaled JD: $B_F(\theta_1:\theta_2) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1:\theta_2)$

[EIG, Entropy 2020]

Bregman manifolds vs Hessian manifolds



- **Hessian metric** wrt. a **flat connection** ∇. function is 0-form on M:

  Riemannian Hessian metric when $g = \nabla^2 F_M$

- **Hessian operator:** $\left(\nabla^2 F_M\right)(X, Y) := \left(\nabla_X d\right)\left(F_M(Y)\right) = X(dF_M(Y)) - dF_M\left(\nabla_X Y\right)$

$$\nabla^2 F_M\left(\partial_{x^i}, \partial_{x^j}\right) = \frac{\partial^2 F_M}{\partial x^i \partial x^j} - \Gamma^k_{ij} \frac{\partial F_M}{\partial x^k} \quad \xrightarrow{\nabla \text{ flat}} \quad \nabla^2 F_M\left(\partial_{x^i}, \partial_{x^j}\right) = \frac{\partial^2 F_M}{\partial x^i \partial x^j}$$

- **Bregman manifold**: geometry on an open convex domain:

  Here, ∇ = gradient — Here, ∇, ∇^* = affine flat connections

$$g(\theta) = \nabla^2 F(\theta) \Longrightarrow \nabla : \Gamma_{ijk}(\theta) = 0$$

$$g^*(\eta) = \nabla^2 F^*(\eta) \Longrightarrow \nabla^* : \Gamma^{*ijk}(\eta) = 0$$

N., On geodesic triangles with right angles in a dually flat space, Progress in Information Geometry: Theory and Applications, Springer, 2021

Rao's distance between 1D normal distributions

Fisher information metric becomes the Poincare upper plane metric after scale change of variable

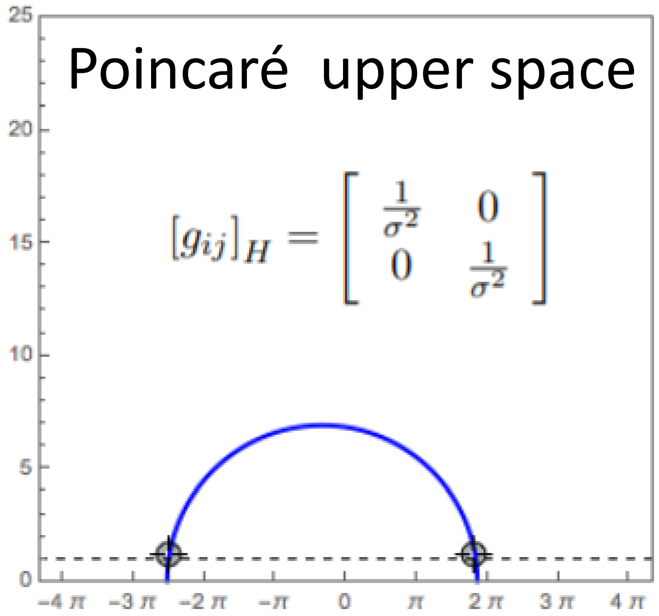
FIM of normal distributions

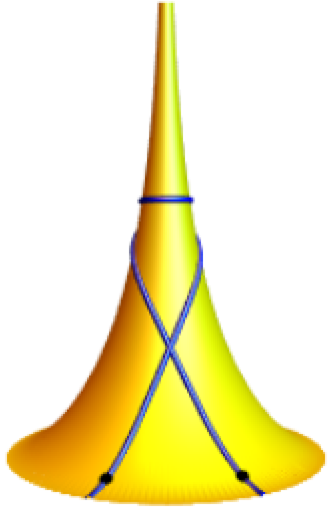
$$[g_{ij}(\mu,\sigma)]_F = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

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

$$d_F((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{2}d_H\left(\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right),\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right)$$

$$\mathrm{dist}(\langle x_1,y_1\rangle,\langle x_2,y_2\rangle) = \mathrm{arcosh}\left(1+\frac{(x_2-x_1)^2+(y_2-y_1)^2}{2y_1y_2}\right)$$



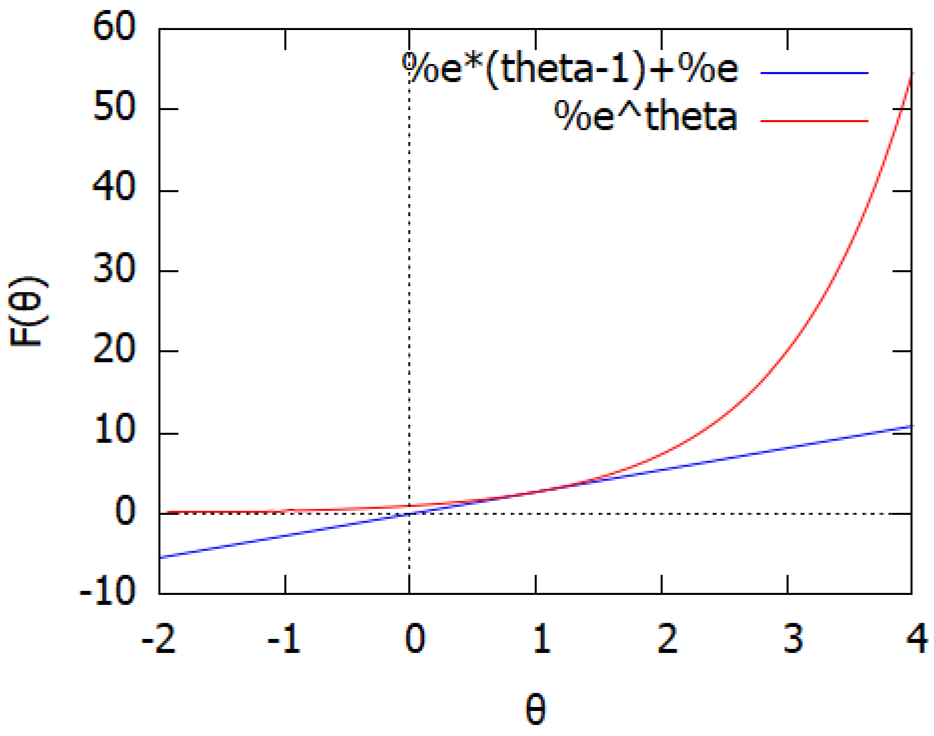


Pseudo-sphere *partial* embedding in R^3

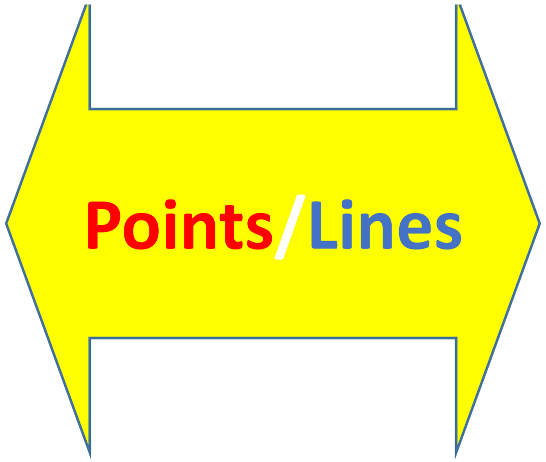
On Voronoi diagrams on the information-geometric Cauchy manifolds, Entropy 22.7 (2020)

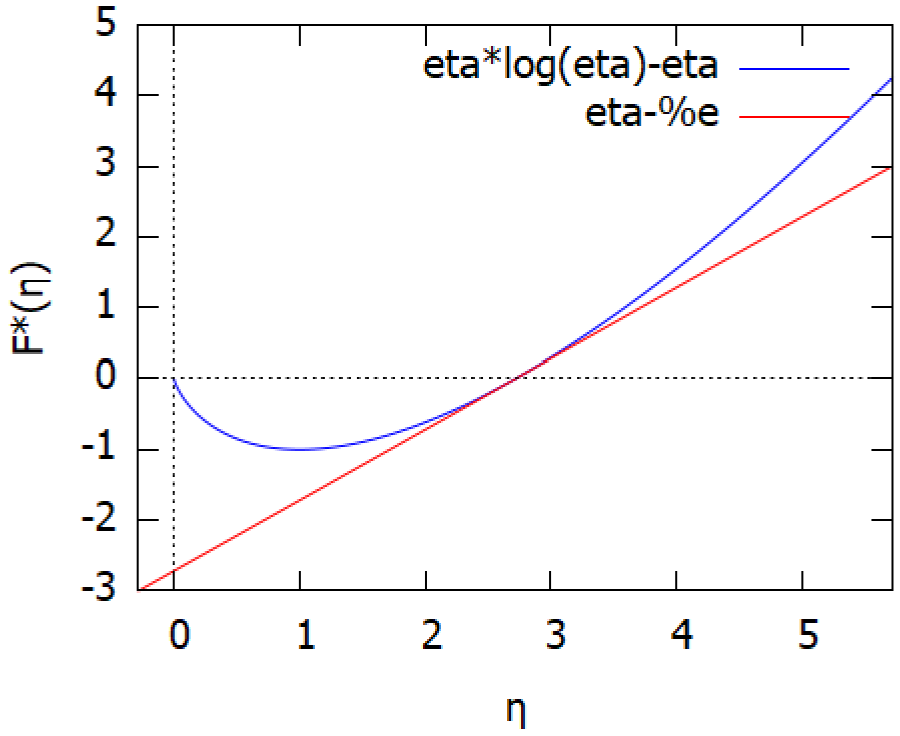
Illustrating the Legendre-Fenchel transformation

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



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





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

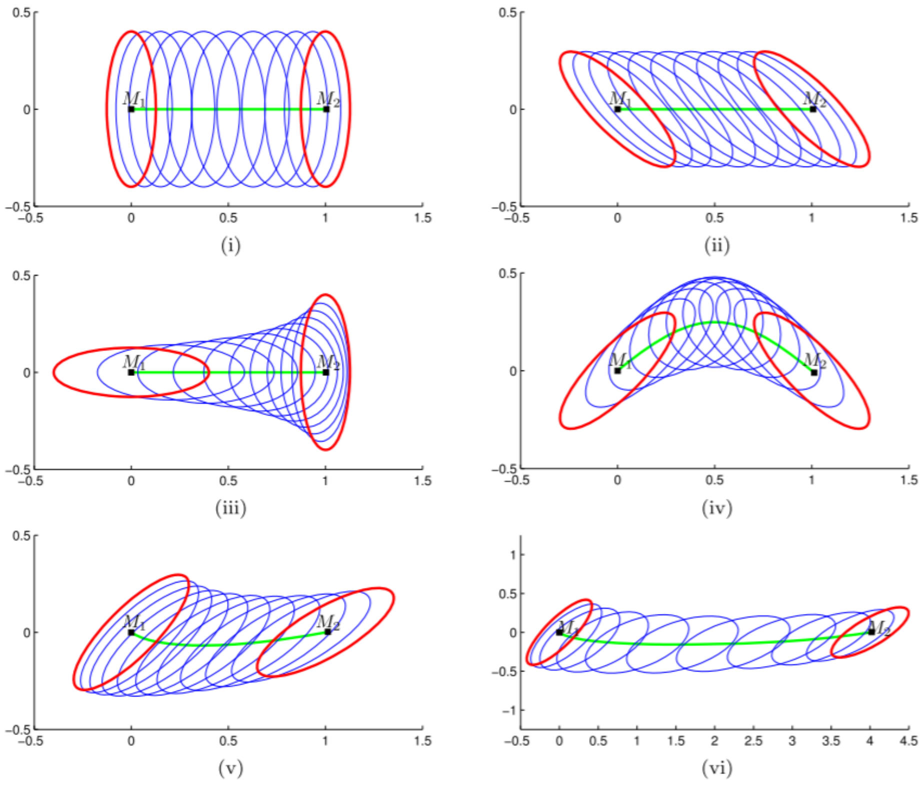
(Here, F was chosen as the cumulant function of the Poisson distributions)

Approximating geodesics for MVNs: geodesic shooting

Algorithm 1 Shooting method for minimal geodesics on $\mathcal{N}(n)$

Given: Initial point $P_0 = (\mu_0, \Sigma_0)$, final point $P_1 = (\mu_1, \Sigma_1)$.
Output: Minimal geodesic $P(t) = (\mu(t), \Sigma(t))$, $t \in [0, 1]$, such that $P(1) = (\mu_1, \Sigma_1)$.
Initialization: Choose initial velocities $V(0) = (\dot{\mu}(0), \dot{\Sigma}(0))$ (e.g., zeroes), initial values for ϵ (10^-5), error $= 10^6$.
while error $\geq \epsilon$ **do**
  Numerically integrate the geodesic equations (13), (14) for given initial conditions $(\mu_0, \Sigma_0, \dot{\mu}_0, \dot{\Sigma}_0)$ from $t = 0$ to $t = 1$
  Denote the solution by $(\mu(t), \Sigma(t))$;
  Set $W(1) = (W_\mu(1), W_\Sigma(1)) = (\mu_1 - \mu(1), \Sigma_1 - \Sigma(1))$;
  Calculate error $= \|W(1)\|_{P_1} = \sqrt{W_\mu(1)^T \Sigma_1^{-1} W_\mu(1) + \frac{1}{2}\mathrm{tr}((\Sigma_1^{-1} W_\Sigma(1))^2)}$;
  Numerically integrate the parallel transport equations (18) and (19) for given trajectory $(\mu(t), \Sigma(t))$ and final veloci
  ties $W(1)$, backward in time from $t = 1$ to $t = 0$;
  Numerically calculate Jacobi field $J(1)$ from (22),
  $J(1) = \frac{\exp_{P_0}(V(0)+\alpha W(0)) - \exp_{P_0}(V(0))}{\alpha}$, where α is sufficiently small value and we use $\frac{\epsilon}{\|W(0)\|_{P_0}}$
  Determine proper update size s:
  $s_1 = \frac{\langle W(1), J(1)\rangle_{P(1)}}{\|J(1)\|^2_{P(1)}}$
  if $\|W(1)\|_{P(1)} > 0.05$ **then**
    $s = 0.05/\|W(1)\|_{P(1)} s_1$;
  else
    $s = s_1$;
  end if
  $V(0) \leftarrow V(0) + sW(0)$;
end while

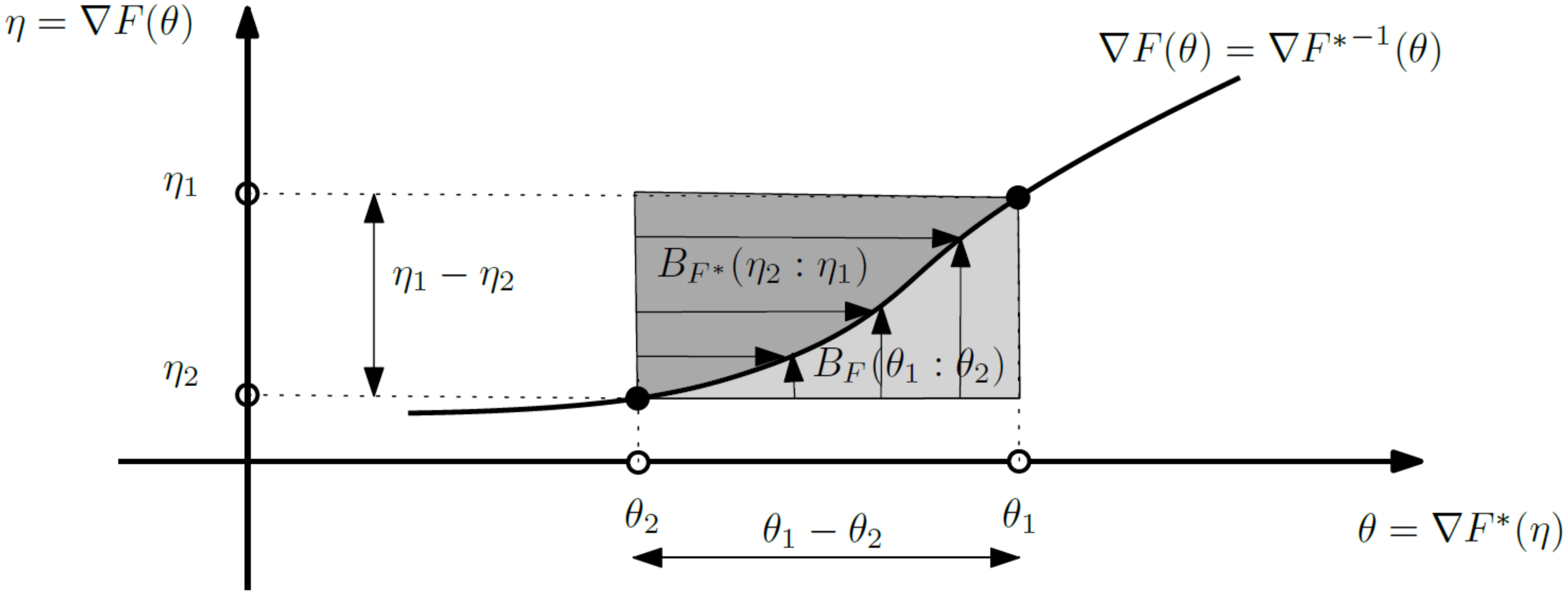
Bivariate normals interpolation



ODE with boundary value conditions

Minyeon Han · F.C. Park, DTI Segmentation and Fiber Tracking Using Metrics on Multivariate Normal Distributions, 2014
Calvo, Miquel, and Josep Maria Oller. "An explicit solution of information geodesic equations for the multivariate normal model." *Statistics & Risk Modeling* 9.1-2 (1991): 119-138.

Symmetrized Bregman divergence: Geometric reading



$$B_F(\theta_1 : \theta_2) = \int_{\theta_2}^{\theta_1} (F'(\theta) - F'(\theta_2))\mathrm{d}\theta$$

$$B_{F^*}(\eta_2 : \theta_1) = \int_{\eta_1}^{\eta_2} (F^{*\prime}(\eta) - F^{*\prime}(\eta_1))\mathrm{d}\eta$$

$$\boxed{\begin{aligned} S_F(\theta_1, \theta_2) &= B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) \\ &= B_F(\theta_1 : \theta_2) + B_{F^*}(\eta_1 : \eta_2) \\ &= (\theta_1 - \theta_2)^\top (\eta_1 - \eta_2) \end{aligned}}$$

[arXiv:2107.05901]

entropy

MDPI

Review

An Elementary Introduction to Information Geometry

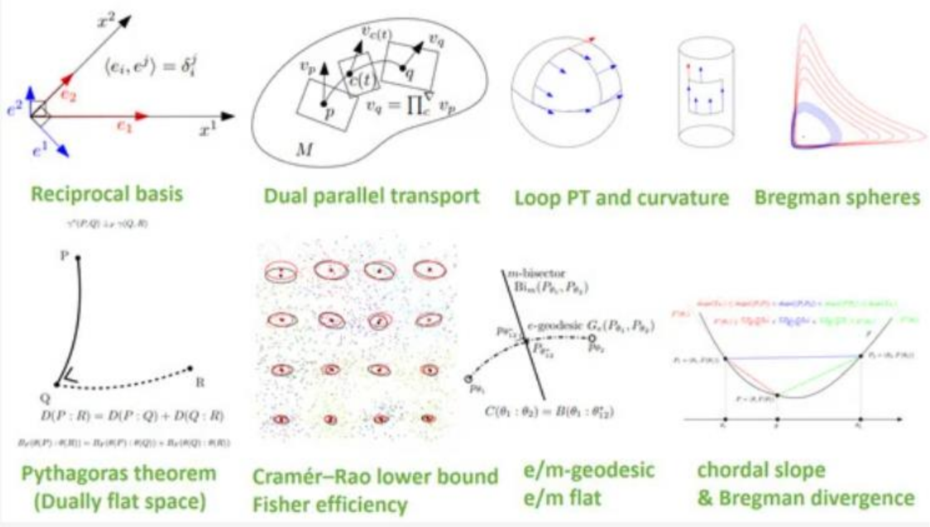
Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; Frank.Nielsen@acm.org

Received: 6 September 2020; Accepted: 25 September 2020; Published: 29 September 2020

check for updates

Abstract: In this survey, we describe the fundamental differential-geometric structures of information manifolds, state the fundamental theorem of information geometry, and illustrate some use cases of these information manifolds in information sciences. The exposition is self-contained by concisely introducing the necessary concepts of differential geometry. Proofs are omitted for brevity.



Entropy 2020
61 pages