

# Information Geometry: An *Invitation* for Machine Learning

Frank NIELSEN  
Sony Computer Science Laboratories Inc.  
Tokyo, Japan

<https://franknielsen.github.io/>



Sony CSL

NeurIPS meetup Japan  
December 2021

# Outline of this talk

## Introduction

ML & computational geometry: A perfect matching with long history!

## 1. Fisher-Rao information geometry

Natural-gradient descent

## 2. Bregman information geometry

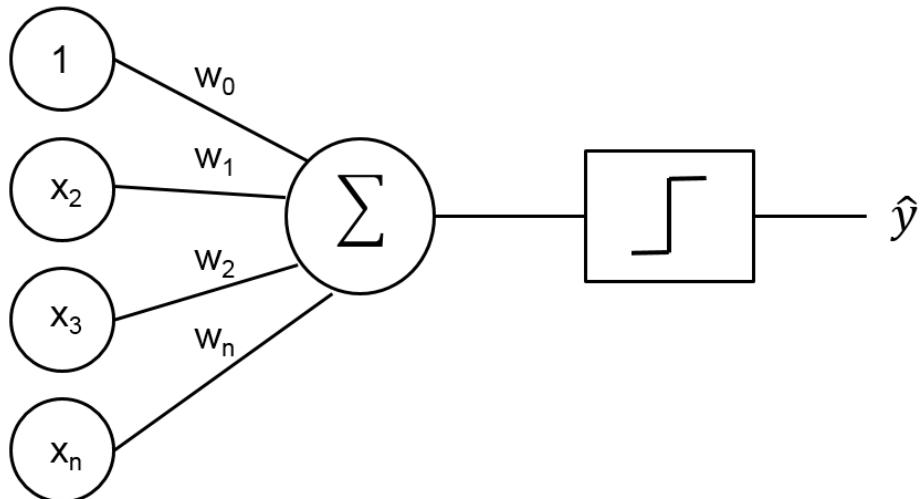
Exponential family manifold

## Perspectives

# Introduction:

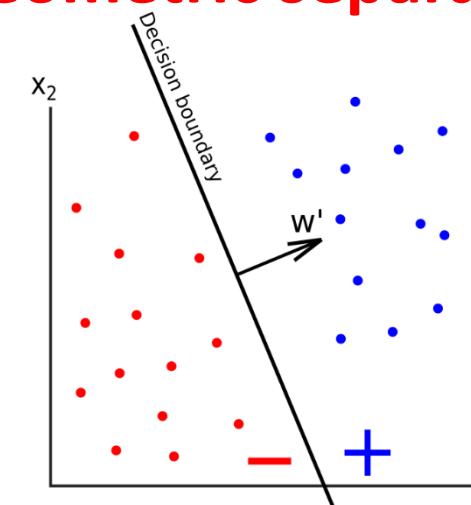
## Machine Learning & computational geometry: A perfect match with a long history!

# Learning machines: Perceptron & geometry (1960's)

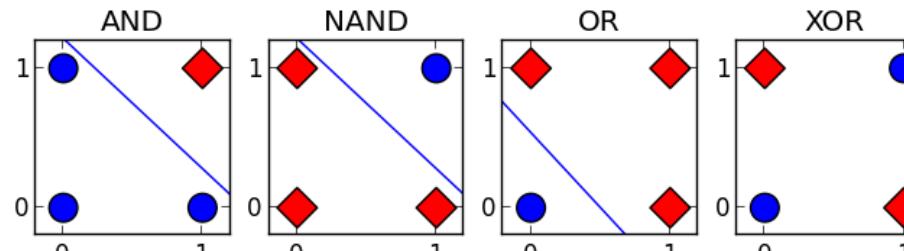


**Connectionism machine**  
≠ von Neumann machine

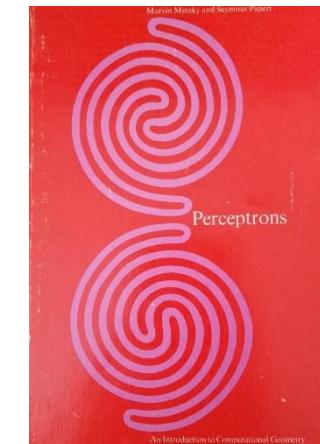
**geometric separator**



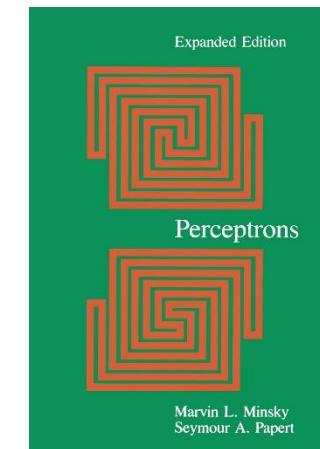
Decision boundary: hyperplanes



XOR cannot be learned... NN winter... MIT Press, 1987  
Marvin Minsky and Seymour Papert:  
*Perceptrons: An Introduction to Computational Geometry*, 1969



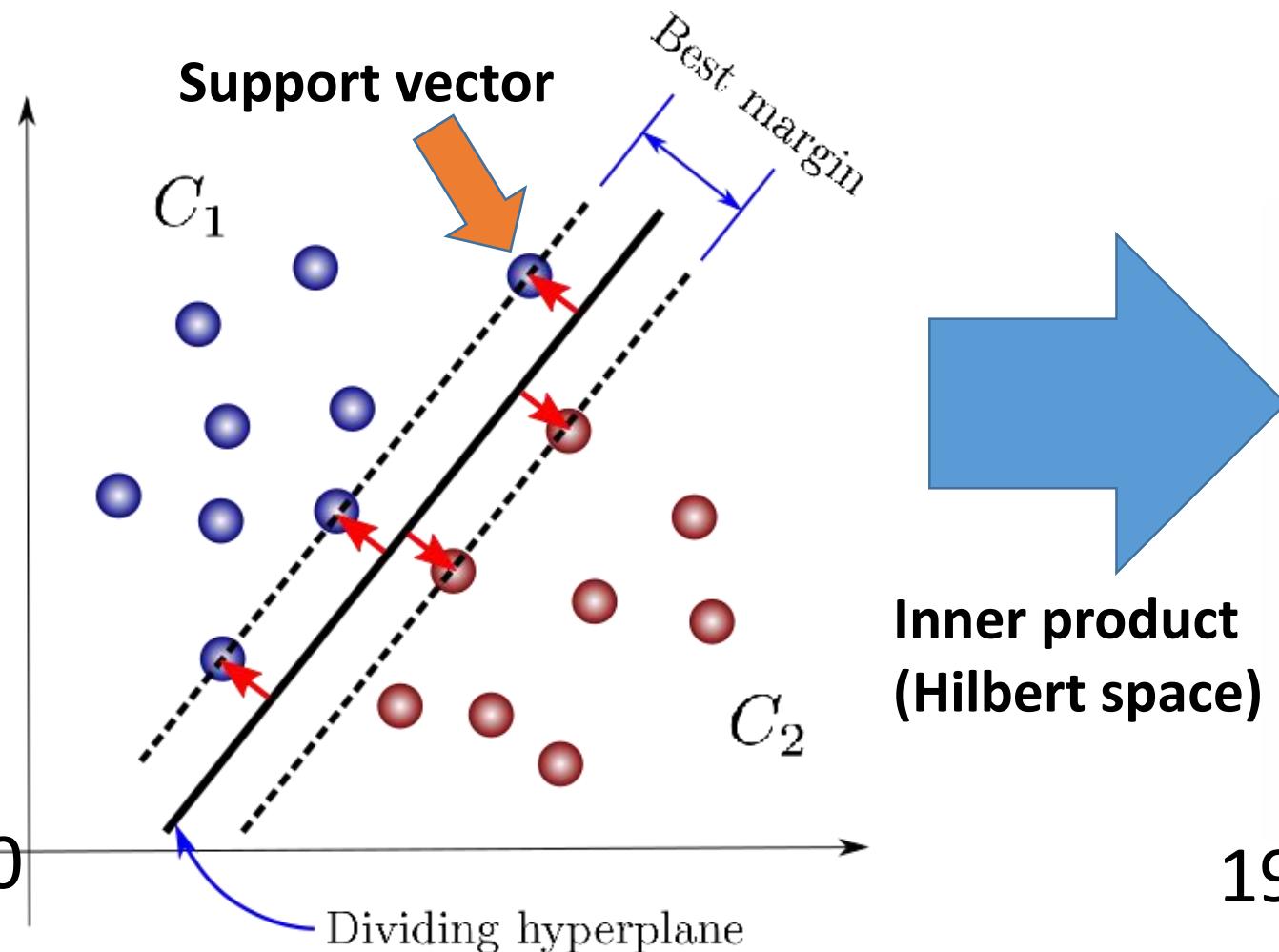
MIT Press, 1969



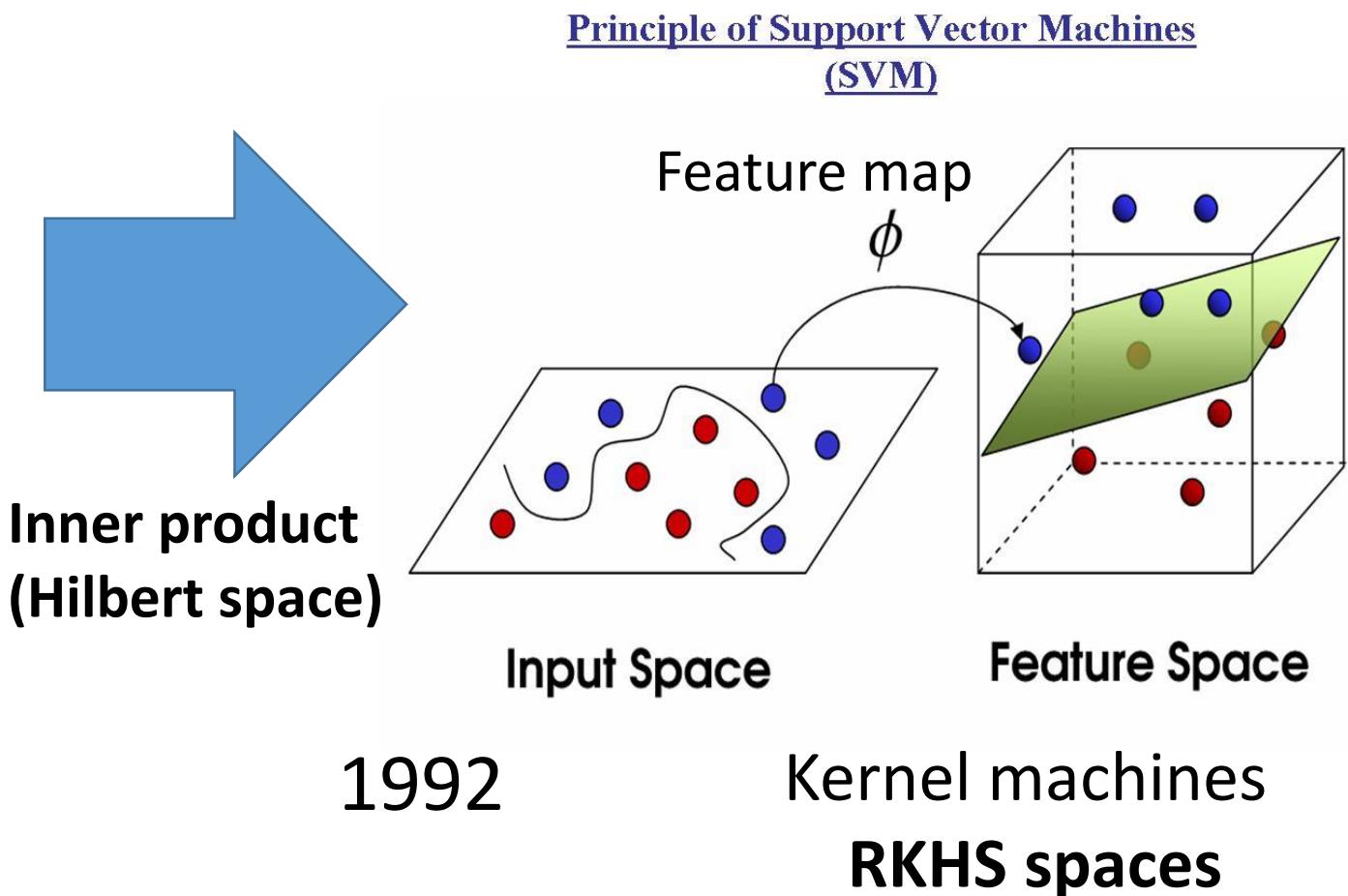
Marvin L. Minsky  
Seymour A. Papert

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

## Linear separator



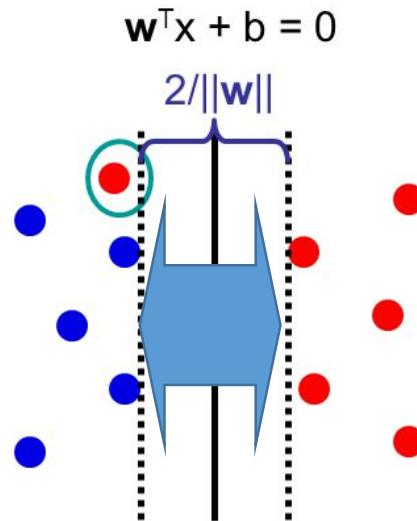
## Non-linear separator



**VC-dimension: expressive power of (geometric) separators**

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

## The SVM Framework



Points  $X = \{x_i\}$

Labels  $y = \{y_i\}$

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

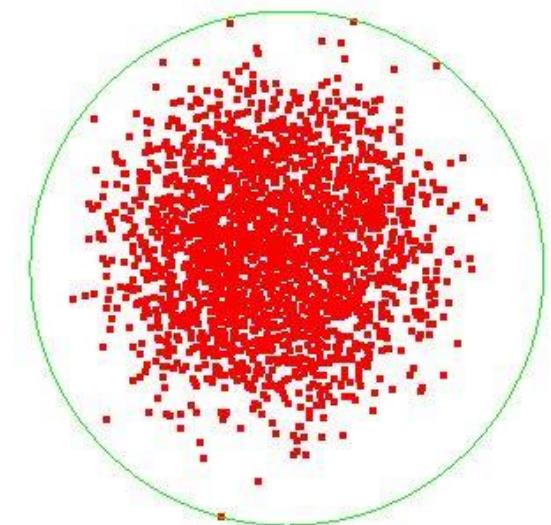
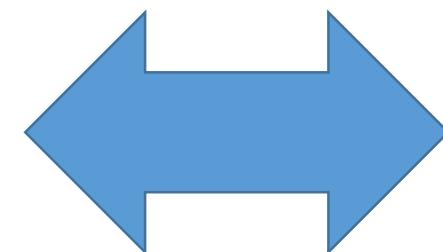
$$\min 1/2 w^T w + C \sum \xi_i$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Convex Quadratic Program

## Widest margin hyperplane separator



**Smallest enclosing ball:**  
“Smallest” ball with respect to  
radius or set inclusion

# 1. Fisher-Rao information geometry

## Riemannian geometry



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

# Information geometry

- Born as a **mathematical curiosity** [Hotelling 1930] [Rao 1945]  
Impacted by the success of Riemannian geometry in Einstein's general relativity
- **Information geometry** studies the **geometric structures** and **statistical invariance principles** (*sufficient statistics, Markov kernels*) of a family of probability distributions (= **statistical model**), and demonstrate their use in information sciences (e.g., statistics & ML).
- 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]

# Fisher information matrix (FIM)



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

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

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

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

Positive  
semi-definite  
matrix

- Score:  $s(\theta) := \nabla_\theta \log p_\theta(x)$
- Under **independence**, Fisher information is **additive**:

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

# Fisher information matrix

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

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

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

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

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

- 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, relative FIM

# Fisher information and Cramér-Rao lower bound

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

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

**Inverse Fisher Information Matrix (IFIM)**

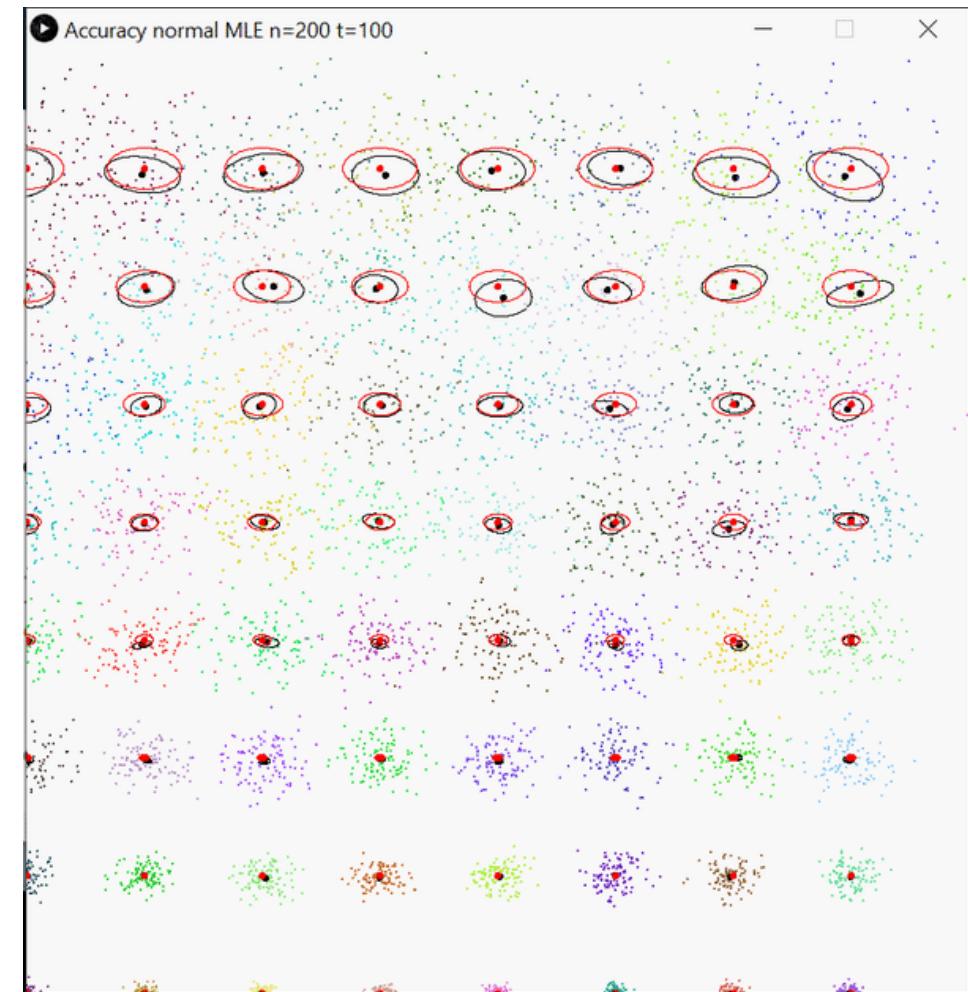
- Since Fisher information is additive:

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

$$(X_1, \dots, X_n) \sim_{\text{iid}} p_\theta$$

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

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



**Empirical estimator covariance matrix  
IFIM (Tissot indicatrix)**

# Rao's length distance: Riemannian metric distance

$(M, g_F)$ : Riemannian manifold

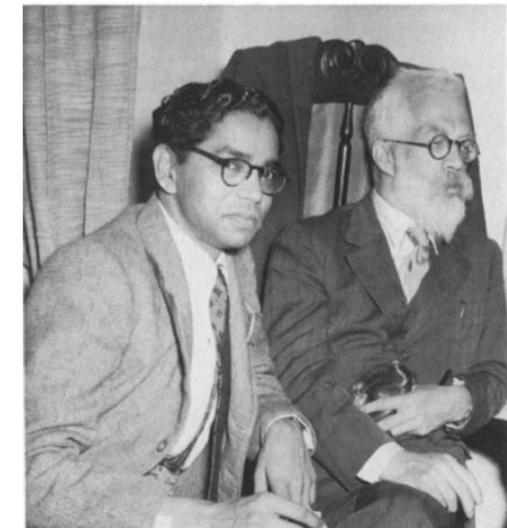
Parameter space equipped with the **Fisher information metric**

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

$$\rho_g(\theta_1, \theta_2) = \min_{\theta(t)} \int_0^1 ds_\theta(t) dt$$

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

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

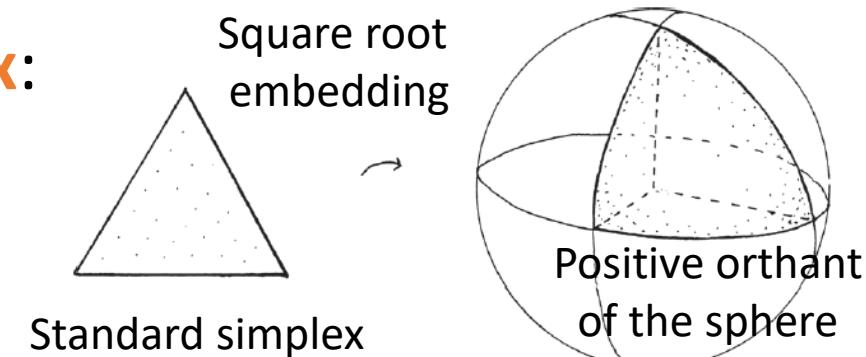


C. R. Rao with  
Sir R. Fisher in 1956

Need to calculate **Riemannian geodesics**  $\theta(t)$ :  
characterized as locally shortest curves in Riemannian geometry

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

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



# 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**:

$$ds_\theta = ds_\eta \quad \rho_{\text{Rao}}(p_{\eta_1}, p_{\eta_2}) = \rho_{\text{Rao}}(p_{\theta_1}, p_{\theta_2})$$

- Fisher information matrix is **covariant**:

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

- Cramer-Rao bound is **contravariant**:

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

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

# In practice, calculating Rao's distance may be difficult!

E.g., no closed form of Rao's distance between multivariate normals (MVs)

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

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

But easy to solve  
when  $\Gamma=0$ :  
Line segments!

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

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

**Geodesic depends on choice of the connection via  $\Gamma$ .**

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

# 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  $\epsilon$  ( $10^{-5}$ ), error =  $10^6$ .

**while**  $\text{error} \geq \epsilon$  **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 =  $\|W(1)\|_{P_1} = \sqrt{W_\mu(1)^T \Sigma_1^{-1} W_\mu(1) + \frac{1}{2} \text{tr}((\Sigma_1^{-1} W_\Sigma(1))^2)}$ ;

Numerically integrate the parallel transport equations (18) and (19) for given trajectory  $(\mu(t), \Sigma(t))$  and final velocities  $W(1)$ , backward in time from  $t = 1$  to  $t = 0$ ;

Numerically calculate Jacobi field  $J(1)$  from (22),

$$J(1) = \frac{\exp_{P_0}(V(0) + \alpha W(0)) - \exp_{P_0}(V(0))}{\alpha}, \text{ where } \alpha \text{ is sufficiently small value and we use } \frac{\epsilon}{\|W(0)\|_{P_0}}$$

Determine proper update size  $s$ :

$$s_1 = \frac{\langle W(1), J(1) \rangle_{P(1)}}{\|J(1)\|_{P(1)}^2}$$

**if**  $\|W(1)\|_{P(1)} > 0.05$  **then**

$$s = 0.05 / \|W(1)\|_{P(1)} s_1;$$

**else**

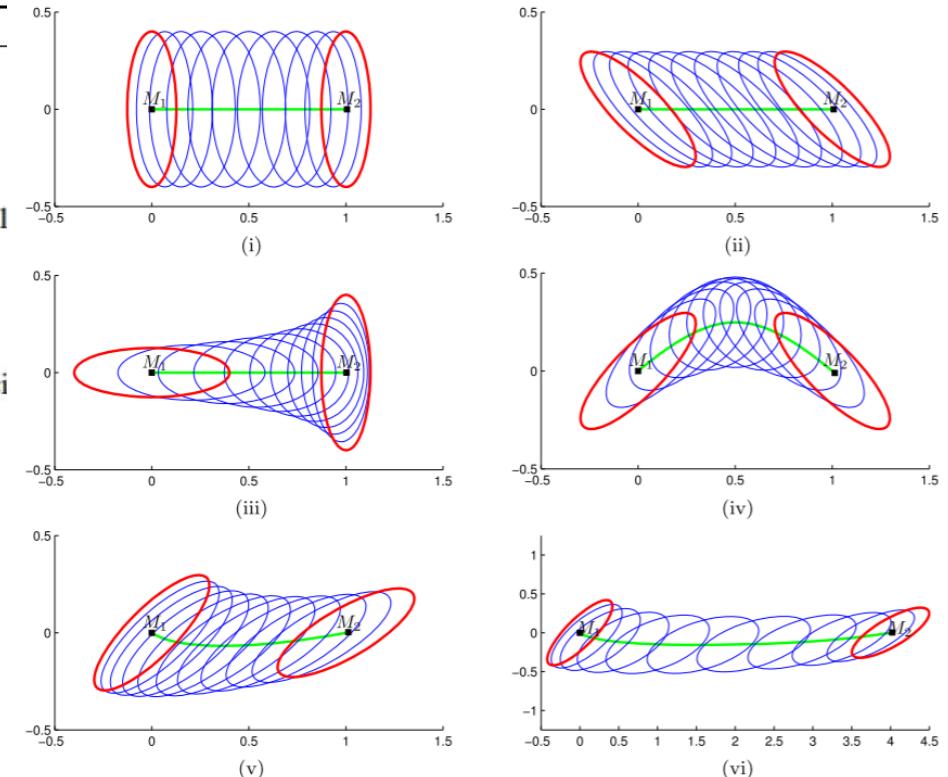
$$s = s_1;$$

**end if**

$$V(0) \leftarrow V(0) + s W(0);$$

**end while**

## Bivariate normals interpolation



## ODE with boundary value conditions

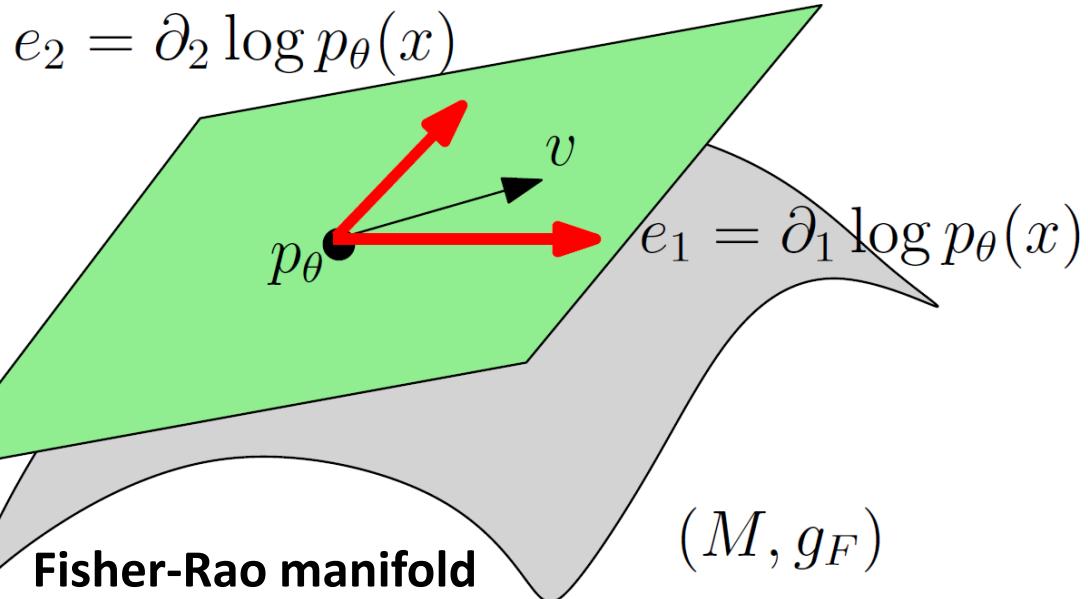
# Fisher-Rao manifolds: Interpreting the inner product

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

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

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

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



Vector expressed using **score functions**:

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

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

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

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

Other basis:  **$\alpha$ -representations** with inner product expressed as  **$\alpha$ -expectations**

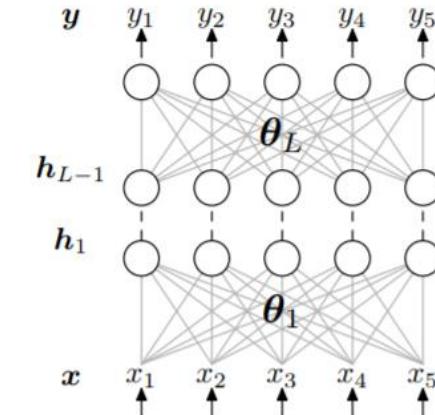
# Natural-gradient descent

Ordinary **gradient descent**:

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

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

$$p(\mathbf{y} | \mathbf{x}, \Theta) = \sum_{\mathbf{h}_1, \dots, \mathbf{h}_{L-1}} p(\mathbf{y} | \mathbf{h}_{L-1}, \Theta_L) \cdots p(\mathbf{h}_2 | \mathbf{h}_1, \Theta_2) p(\mathbf{h}_1 | \mathbf{x}, \Theta_1),$$

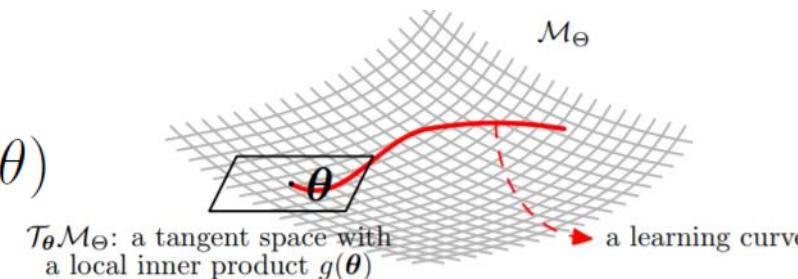


**Natural gradient descent** with **natural gradient** :

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

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

- steepest descent on Riemannian manifold  $(M, g)$
- NG invariant to reparameterization:  $\tilde{\nabla} E_\eta(\eta) = \tilde{\nabla} E_\theta(\theta)$
- avoids plateaus



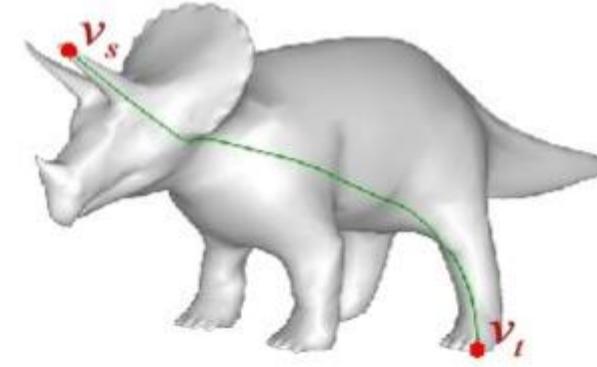
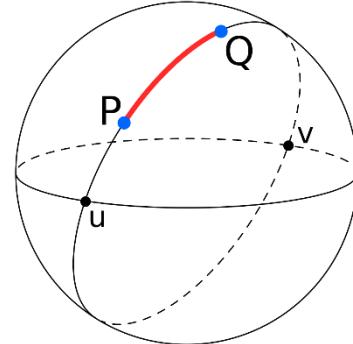
Amari, "Natural gradient works efficiently in learning." Neural computation, 1998

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

# Another look at Riemannian geodesics: Connections

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



- The *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  $\gamma(t)$  with respect to a connection  $\nabla$  is an  **$\nabla$ -autoparallel curve**  
(straight free fall particle in physics):

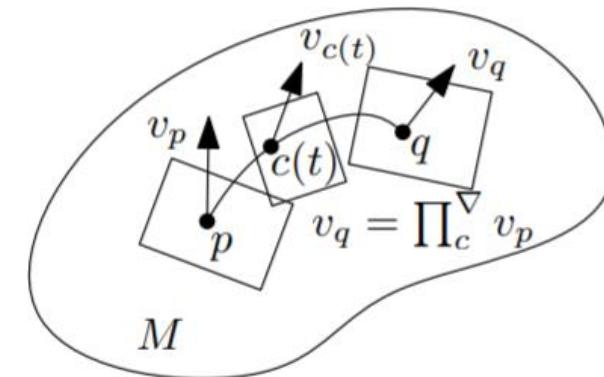
$$\nabla_{\dot{\gamma}} \dot{\gamma} = 0, \quad \dot{\gamma} = \frac{d}{dt} \gamma(t)$$

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

# What makes the Levi-Civita connection so special?

- A connection is described by **Christoffel symbols** (functions  $\Gamma$ ), and the geodesics is described by this ODE:  $\ddot{\gamma}(t) + \Gamma_{ij}^k \dot{\gamma}(t) \dot{\gamma}(t) = 0, \quad \gamma^I(t) = x^I \circ \gamma(t),$

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



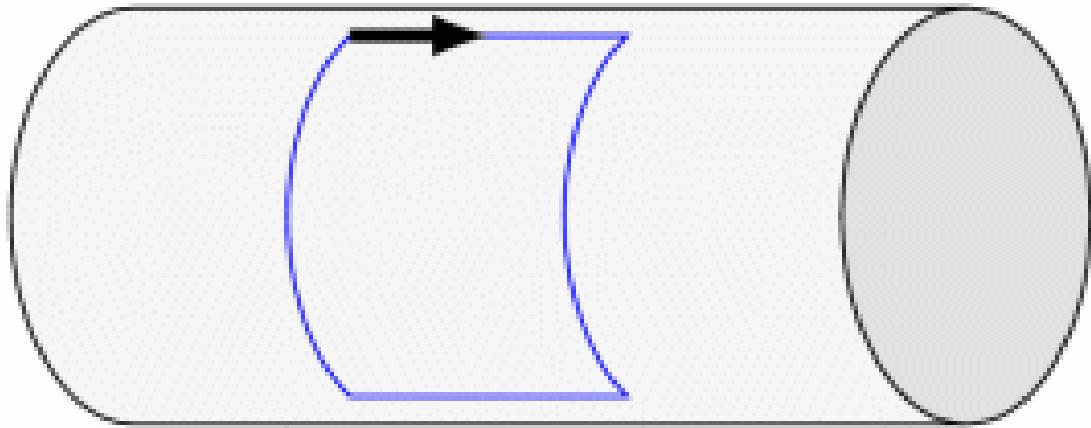
- **Fundamental theorem of Riemann geometry:**

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

$$\text{LC}\Gamma_{ij}^k \stackrel{\Sigma}{=} \frac{1}{2} g^{kl} (\partial_i g_{il} + \partial_j g_{il} - \partial_l g_{ij})$$

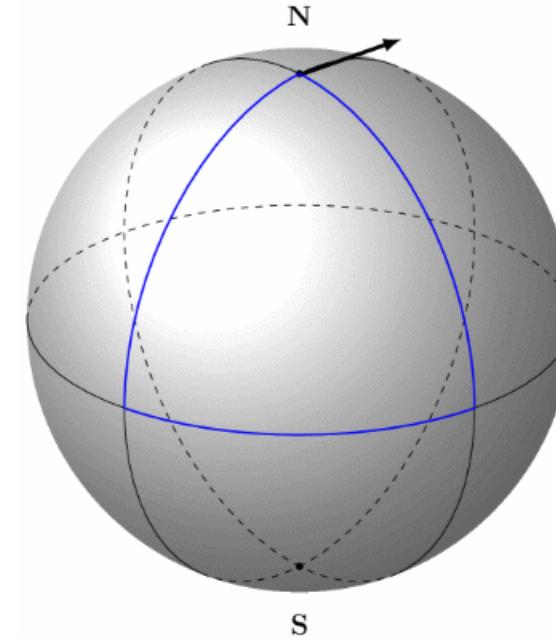
$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \rightarrow c(t)}^\nabla u, \prod_{c(0) \rightarrow c(t)}^\nabla v \right\rangle_{c(t)} \quad \forall t.$$

# Affine connection $\nabla$ : Curvature and parallel transport on infinitesimal loops



Cylinder is **flat**:

Parallel transport is  
**independent of path**



Sphere has constant curvature:  
**Parallel transport is path-dependent**

A connection is flat if there exists locally a coordinate system such that the Christoffel symbols are all zero: Geodesics plotted in that coordinate system are line segments

# Statistical invariance: 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  $\theta$

$$\Pr(x|\theta) = \Pr(x|t)$$

$t=T(X)$  contains all  
Information about  $\theta$

- **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:**

$$N(\mu, \sigma)$$

$$t_1(X_1, \dots, X_n) = \sum_i X_i$$

$$t_2(X_1, \dots, X_n) = \sum_i X_i^2$$

# Natural exponential families (NEF)

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

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

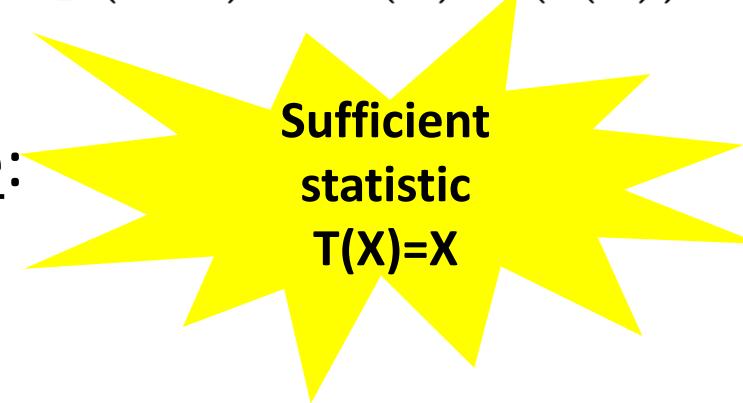
$$p(x; \lambda) = a(x)b_\lambda(t(x))$$

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

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

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

$F$ : **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

# Exponential families (from Natural EFs to 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  $\theta$   
(usual scalar/dot product)

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

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

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

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

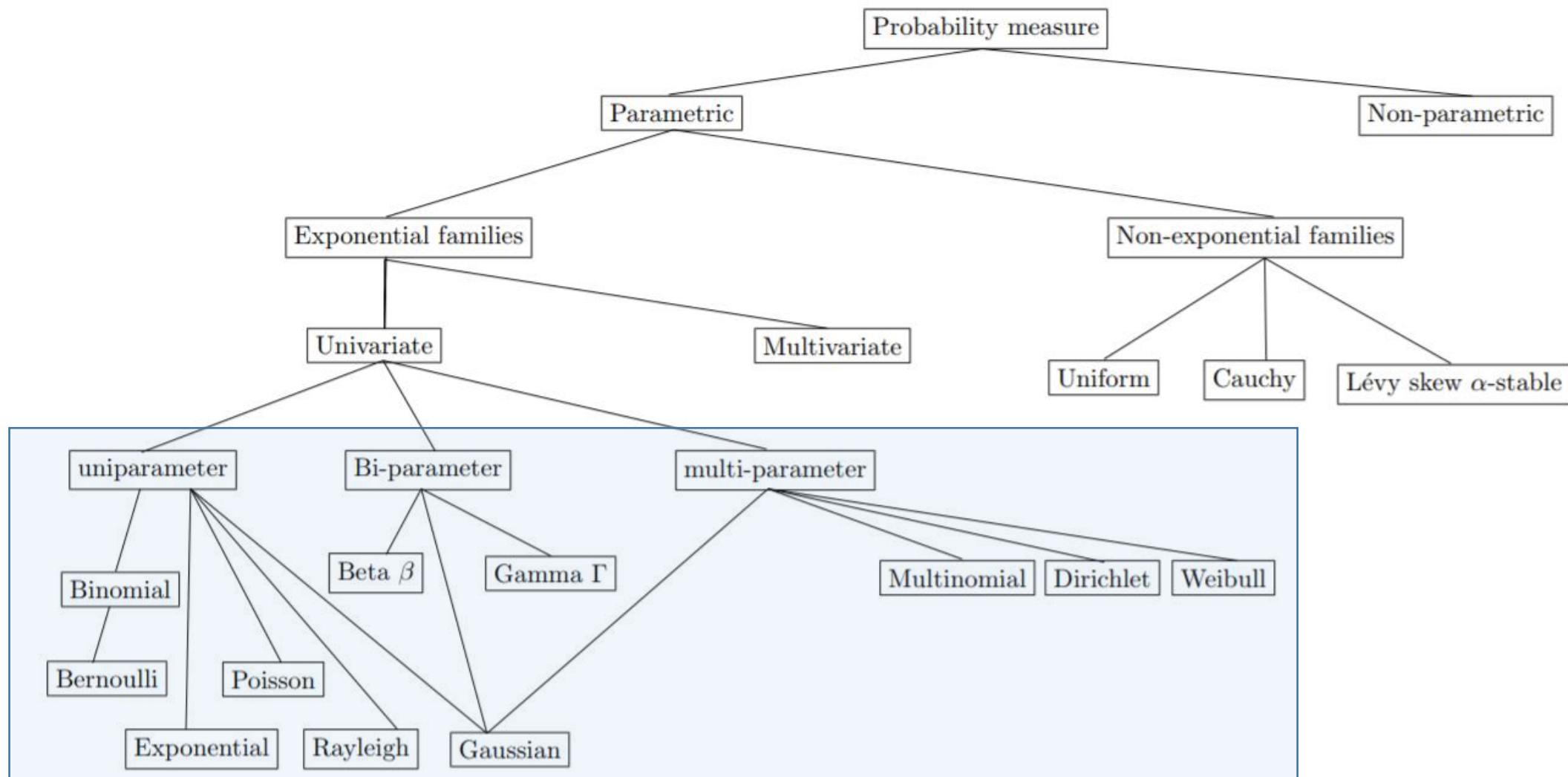
$$\text{Cov}[t(X)] = \nabla^2 F(\theta) = I(\theta)$$



(Hessian of  $-\log p_\theta(x)$ )  
(FIM of type 2)

**Exponential families have finite moments of any order**

# Many common distributions are exponential families in disguise



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)

# Bhattacharyya arc: Likelihood Ratio Exponential Family

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

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

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

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

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

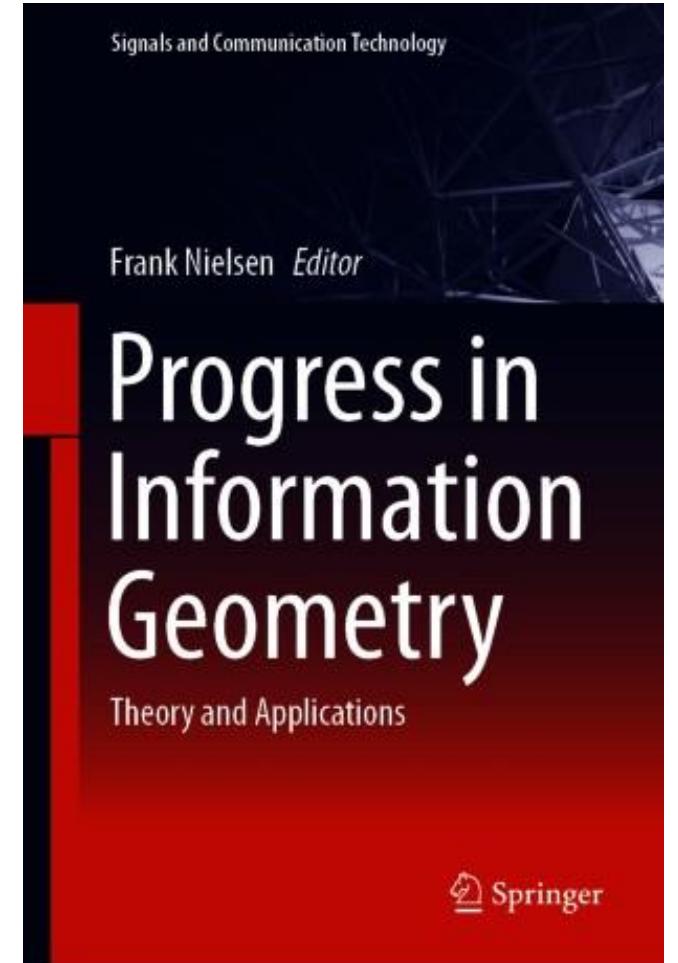
Log-likelihood sufficient statistics:

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

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

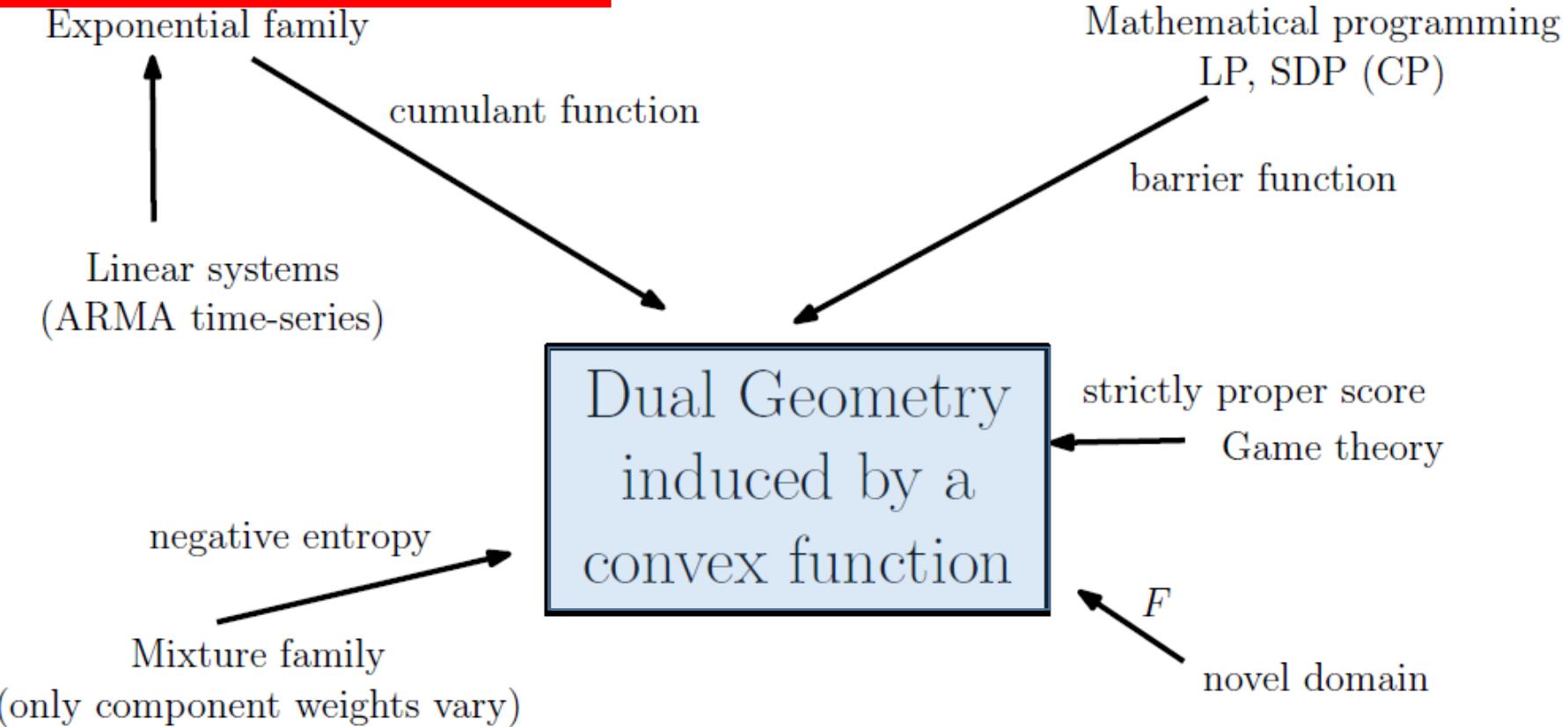
$$D_\alpha^{\text{Bhat}}[p : q] := -\log \left( \int_{\mathcal{X}} p^{1-\alpha}(x)q^\alpha(x)d\mu(x) \right)$$

# 2. Bregman manifolds



# Dually flat geometry from any strictly convex function

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

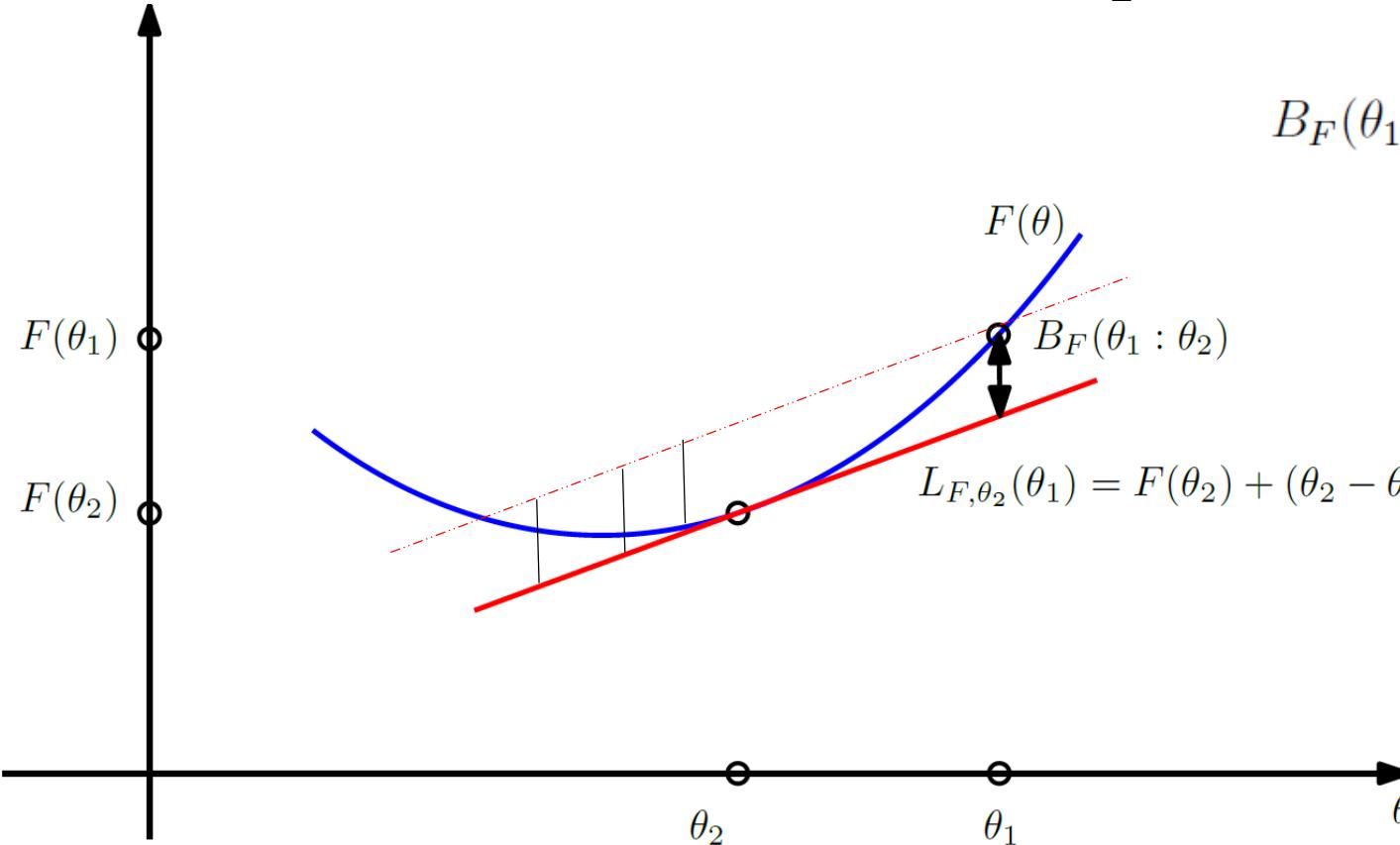


Bregman manifolds are not necessarily related to statistical models,

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

# Bregman divergences from strictly convex function

- $F(\theta)$ : strictly convex and differentiable convex function on an open convex domain  $\Theta$
- Design the **Bregman divergence** as the vertical gap between  $F(\theta_1)$  and the linear approximation of  $F(\theta)$  at  $\theta_2$  evaluated at  $\theta_1$ :



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

[Bregman 1967]

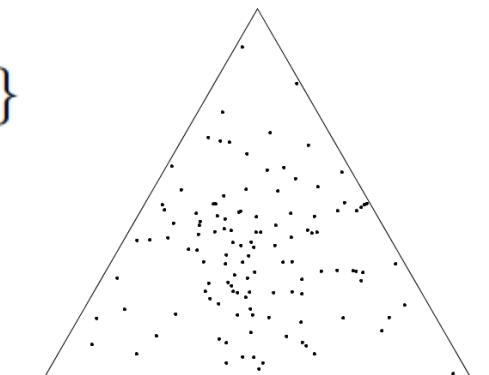
# Discrete Kullback-Leibler divergence: A non-separable Bregman divergence

- 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**.

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

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

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



LogSumExp is only convex but **LogSumExp<sub>+</sub> is strictly convex** [NH 2019]

# Legendre-Fenchel transformation

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

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

Concave programming:

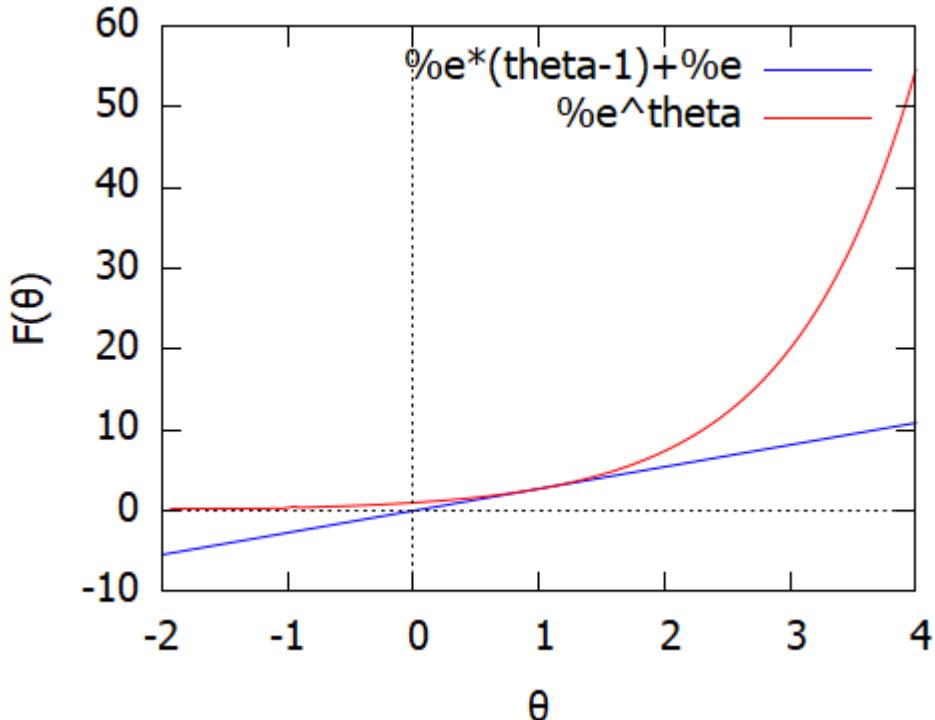
$$F^*(\eta) = \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\} = \sup_{\theta \in \Theta} \{E(\theta)\}$$

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

- Legendre-Fenchel transformation applies to any multivariate function
- Fenchel-Moreau's **biconjugation theorem** for  $F$  of Legendre-type:  $F = (F^*)^*$

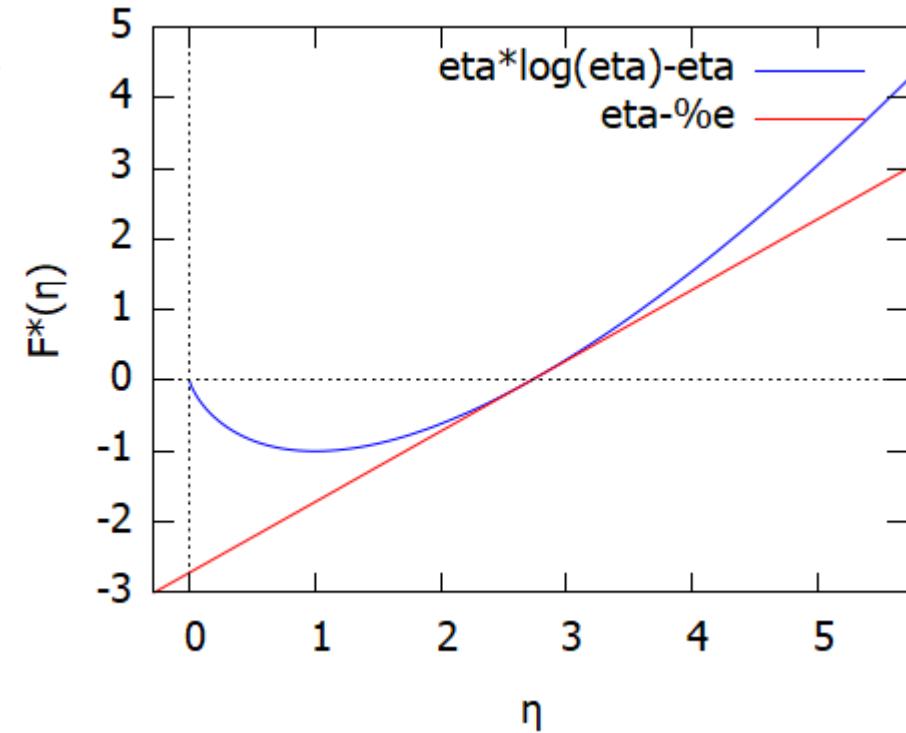
# Reading the Legendre-Fenchel transformation

- Legendre-Fenchel transformation also called the slope transform



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

Points/Lines



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

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

# Legendre-Fenchel transform: Mixed coordinates and Fenchel-Young divergence

- **Dual parameterizations** of epigraph:  $\theta = \nabla F^*(\eta)$  and  $\eta = \nabla F(\theta)$
- 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  $\nabla F(\theta)$
- **Fenchel-Young inequality:**  $F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2$   
with equality if and only if  $\eta_2 = \nabla F(\theta_1)$
- **Fenchel-Young divergence** use mixed parameterization  $\theta/\eta$ :

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

# Dual Bregman and dual Fenchel-Young divergences

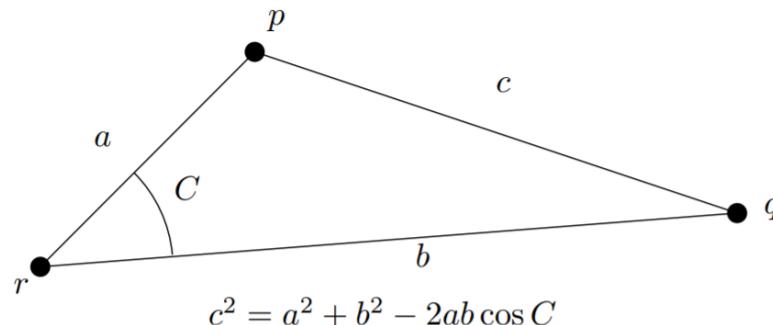
- **Identity of dual Bregman divergences:**  $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$
- In general, dual or **reverse divergence**:  $D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$
- Primal, dual or mixed parameterizations of Bregman divergences:

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

# 3-parameter identity of Bregman divergences

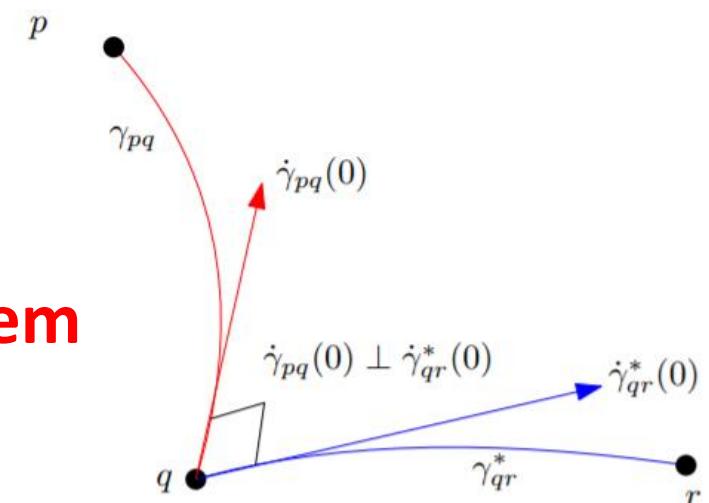
- Generalize the **law of cosines** for the squared Euclidean distance

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



- yields a **generalization of the Pythagorean theorem**

when  $(\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3)) = 0$

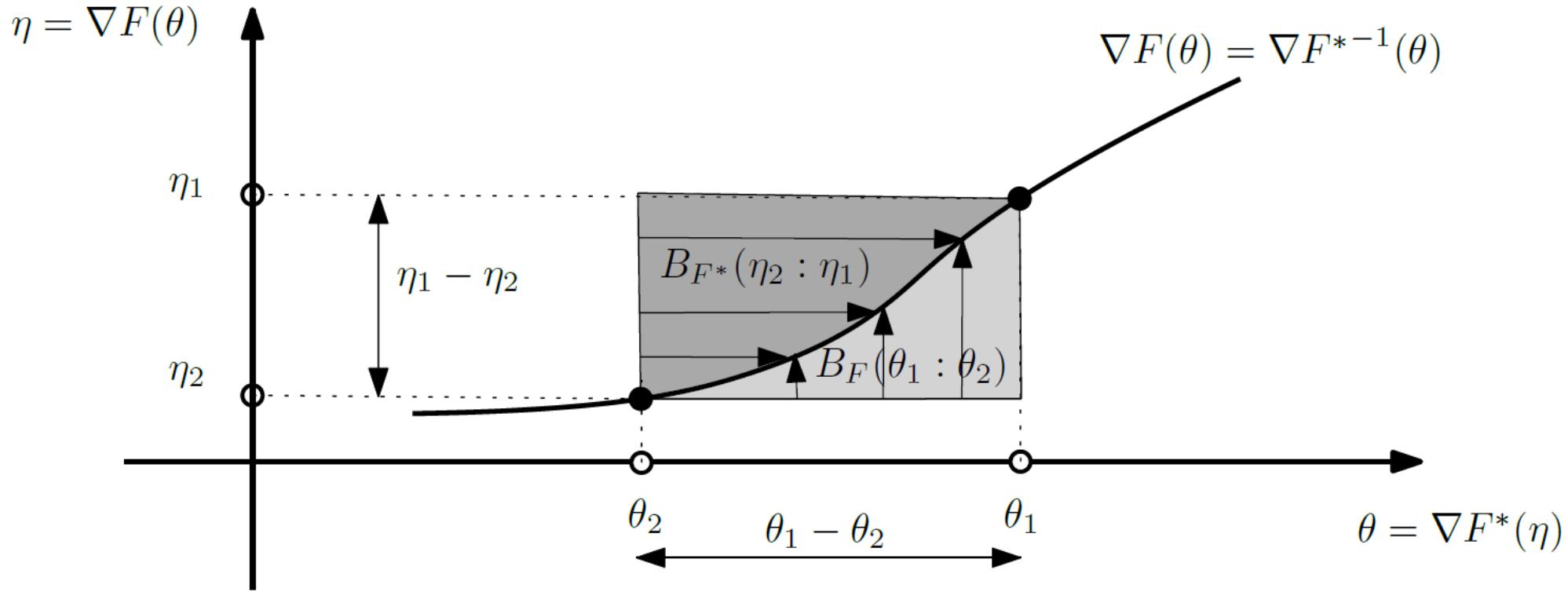


$$D_F(p : r) = D_F(p : q) + D_F(q : r)$$

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

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

# Symmetrized Bregman divergence: Geometric reading



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

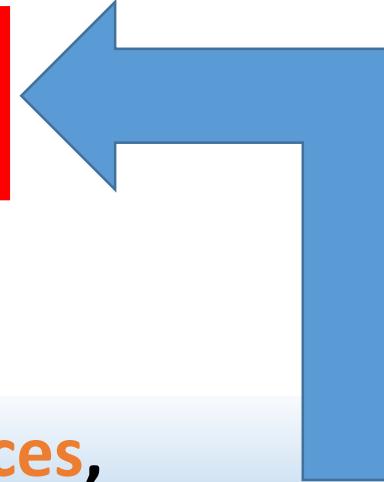
$$B_{F^*}(\eta_2 : \theta_1) = \int_{\eta_1}^{\eta_2} (F^{*\prime}(\eta) - F^{*\prime}(\eta_1)) d\eta$$

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

# 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:

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



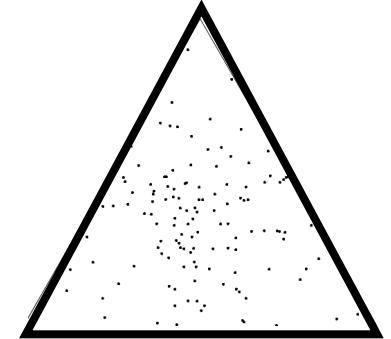
For which **statistical models** and **statistical divergences**,  
do we obtain  $D_{\mathcal{M}}(\theta_1 : \theta_2)$  as a **Bregman divergence**?

# Example 1: Natural exponential family models

- 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  $\mathcal{D}_{\text{KL}}[e_{\theta_1} : e_{\theta_2}] = \underbrace{B_F^*(\theta_1 : \theta_2)}_{D_{\mathcal{E}}(\theta_1 : \theta_2)} = B_F(\theta_2 : \theta_1)$  with Bregman generator:  
the **log-normalizer convex real-analytic function**:  $F_{\mathcal{E}}(\theta) = \log\left(\int \exp\left(\sum_{i=1}^D t_i(x)\theta_i + k(x)\right) d\mu(x)\right)$

## Example 2: Mixture family models

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



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

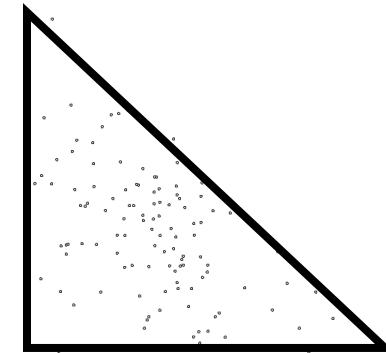
$$m_\theta(x) = \sum_{i=1}^D w_i p_i(x) + \left(1 - \sum_{i=1}^D w_i\right) p_0(x)$$

- We have:  $\mathcal{D}_{\text{KL}}[m_{\theta_1} : m_{\theta_2}] = \underbrace{B_{F_M}(\theta_1 : \theta_2)}_{D_M(\theta_1:\theta_2)}$   $\theta = (w_1, \dots, w_D)$

Information geometry/reconstruction

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

$$F_M(\theta) = \int m_\theta(x) \log m_\theta(x) d\mu(x)$$



Natural parameters

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

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

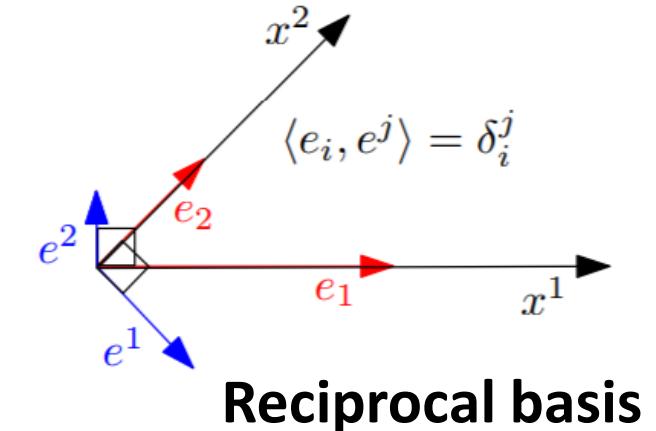
# Metric tensor using covariant/contravariant notations

**2-covariant metric tensor** in local coordinates:

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

**Dual metric tensor** in local coordinates:

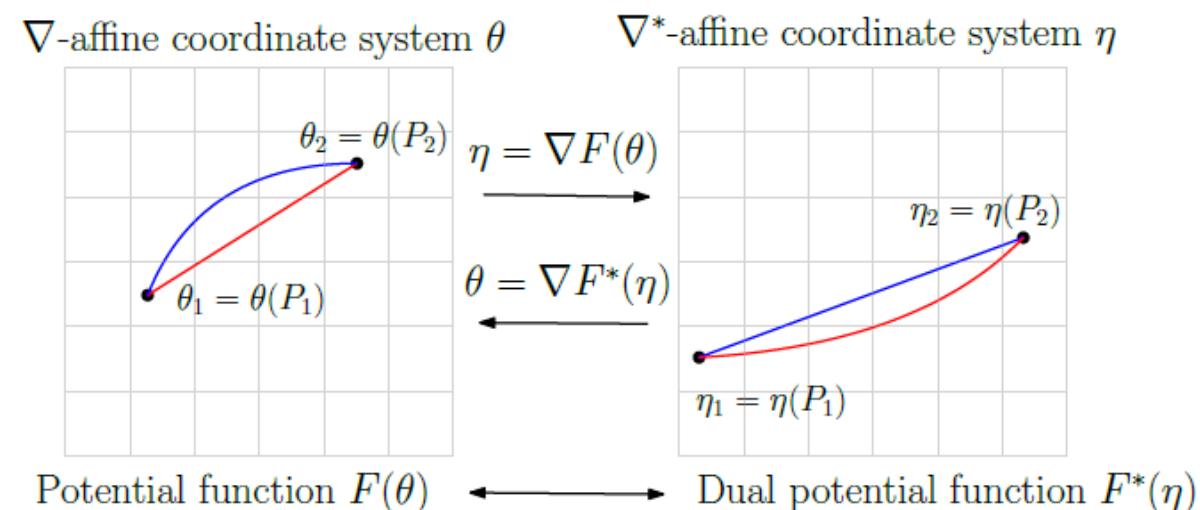
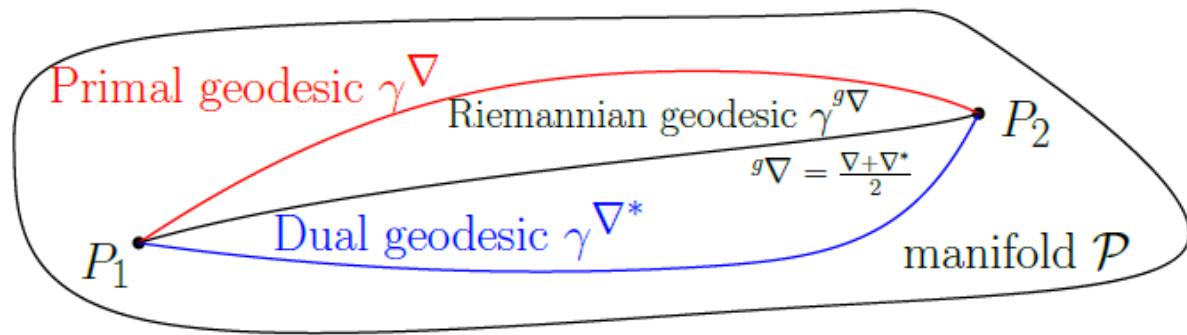
$$g^{ij}(\eta) = g^{*ij}(\eta) = \nabla^2 F^*(\eta)$$



**Crouzeix's identity**:  $\times$  of Hessians of convex conjugates= Id:

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

# Bregman information geometry: Bregman manifolds



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

- Start from a potential function  $F(\theta)$

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

- Get the dual potential function  $F^*(\eta)$

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

- 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

In a **Bregman manifold**, we have

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

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

$$\begin{aligned} &= (\nabla_\theta \eta)^{-1} \nabla_\theta \eta \nabla_\eta L_\eta(\eta) \\ &= \nabla_\eta L_\eta(\eta) \quad \text{Ordinary gradient} \end{aligned}$$

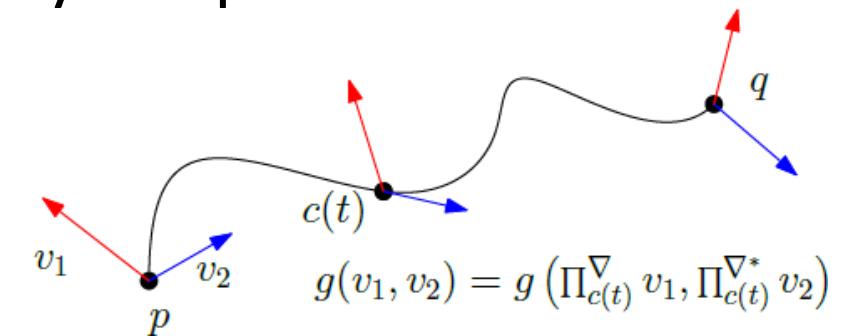


Used in variational inference (VI)

# Dualistic information geometry: $(M, g, \nabla, \nabla^*)$

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


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



- This amounts to say that  $\nabla^*$  is defined uniquely by

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X^* Z),$$

meaning  $X_p g_p(Y_p, Z_p) = g_p((\nabla_X Y)_p, Z_p) + g_p(Y_p, (\nabla_X^* Z)_p)$ .

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

# Amari's $\alpha$ -structures of probability families

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

- Regular statistical parametric models (identifiable and finite positive-definite FIM)  $\mathcal{P} := \{p_\theta(x)\}_{\theta \in \Theta}$
- Amari's  **$\alpha$ -connections**  ${}_{\mathcal{P}}\Gamma^\alpha_{ij,k}(\theta) := E_\theta \left[ \left( \partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right].$   
$$l(\theta; x) := \log L(\theta; x) = \log p_\theta(x)$$
- 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**

# Perspective.

## Information geometry: An invitation to ML

- Given a parametric statistical model, get a **dualistic  $\alpha$ -structure**
- For exponential families and mixture families, get **Bregman manifolds (dually flat spaces,  $\pm 1$ -structure)** with **generalized Pythagoras theorems**
- From the viewpoint of statistical invariance, **Fisher information metric** is unique (up to scaling)
- Challenge: How to deal with **large-size hierarchical singular models?**
- Other **information metrics**: energetic information metric, Wasserstein information metrics [LZ 2019],  $\varphi$ -entropy metrics (e.g., entropy metric of order  $\alpha$ ) [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

Towards Modeling and Resolving Singular Parameter Spaces using Stratifolds, arXiv:2112.03734

# 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  $\mu$ , usually chosen as the Lebesgue measure  $\mu_L$  or the counting measure  $\mu_c$ , and consider a parametric family  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  of probability distributions, all dominated by  $\mu$ . 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  $\theta$ -coordinate system is the Fisher information matrix (FIM) of the random variable  $X$ :  $[g_F]_\theta := I_X(\theta)$  with

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



Review

## An Elementary Introduction to Information Geometry

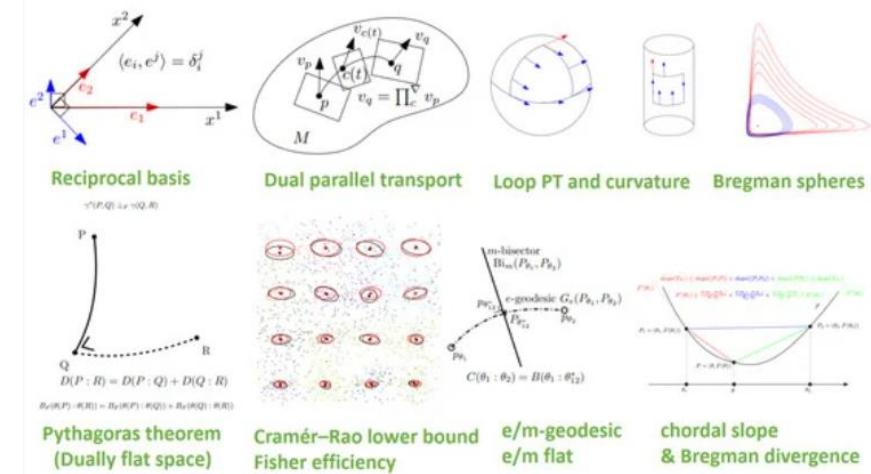
Frank Nielsen

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

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



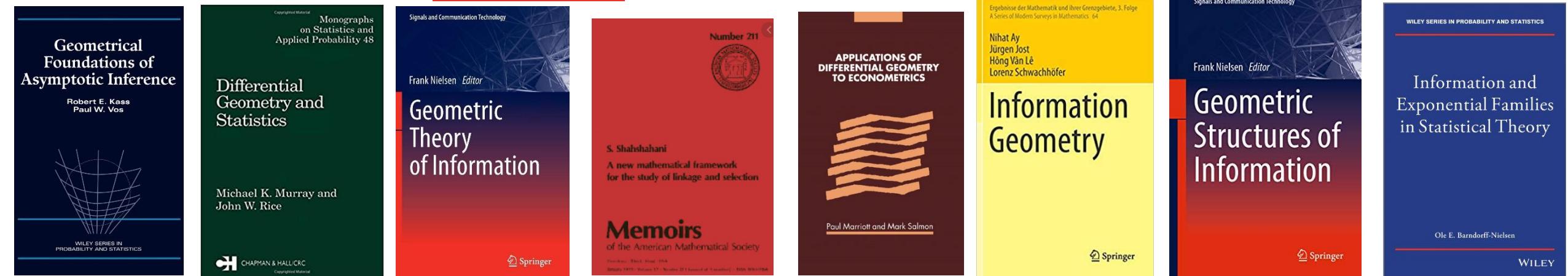
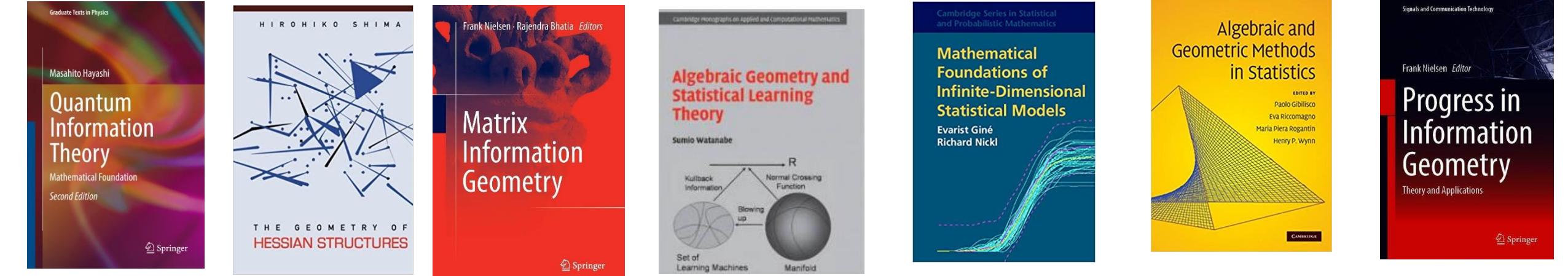
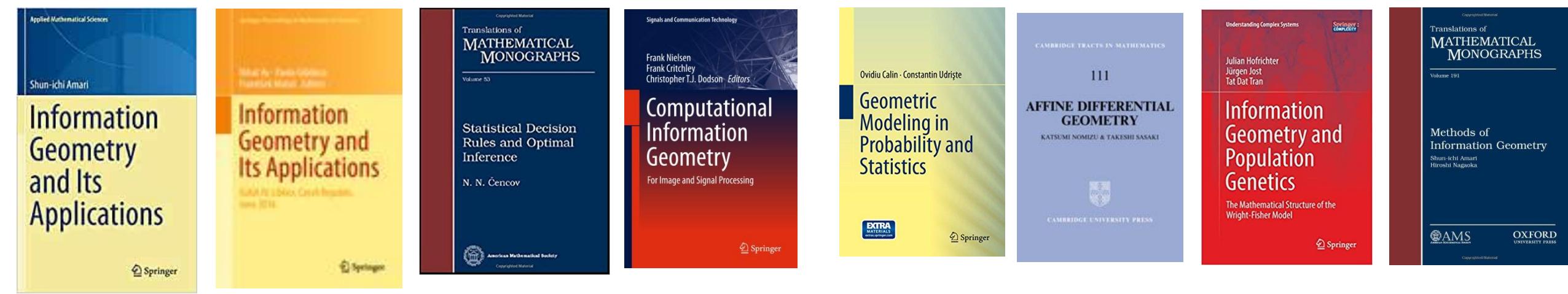
**Abstract:** In this survey, we describe the fundamental differential-geometric structures of information manifolds, state the fundamental theorem of information geometry, and illustrate some use cases of these information manifolds in information sciences. The exposition is self-contained by concisely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



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

Entropy 2020  
61 pages





# Structured natural-gradient descent (Struct-NGD)

- Consider the **general optimization problem**:

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

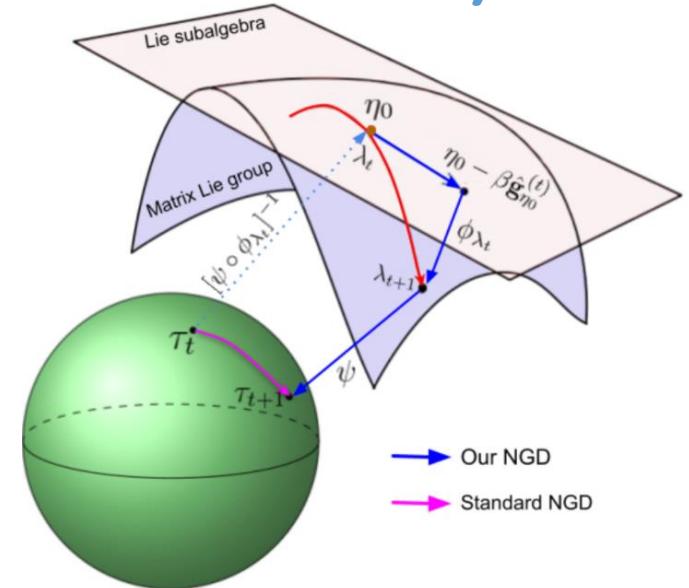
- Standard natural-gradient descent (without structure):

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

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

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

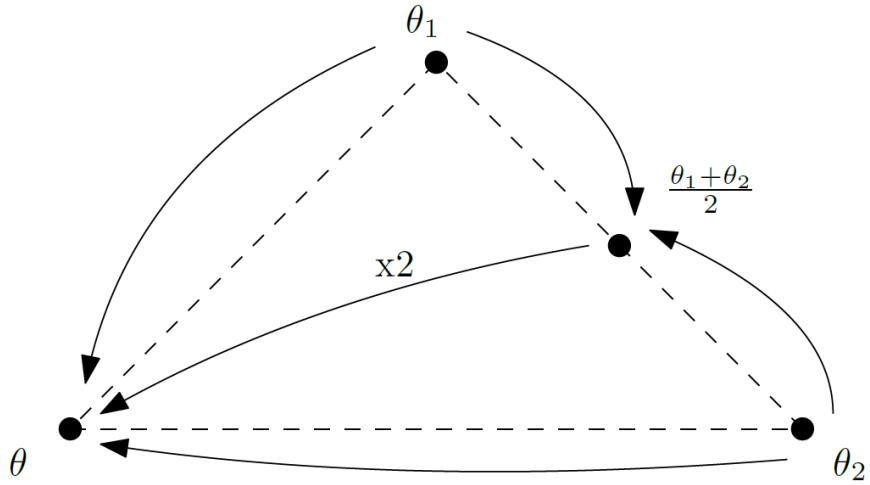
- worked examples on matrix Lie groups and applications: generalizes NGD & xNES evolutionary strategy, recovers Newton-like algorithms, obtained new structured second-order algorithms, etc.



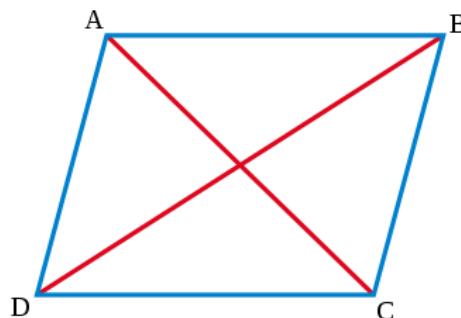
# 4-parameter identity of Bregman divergences

- Parallelogram identity

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

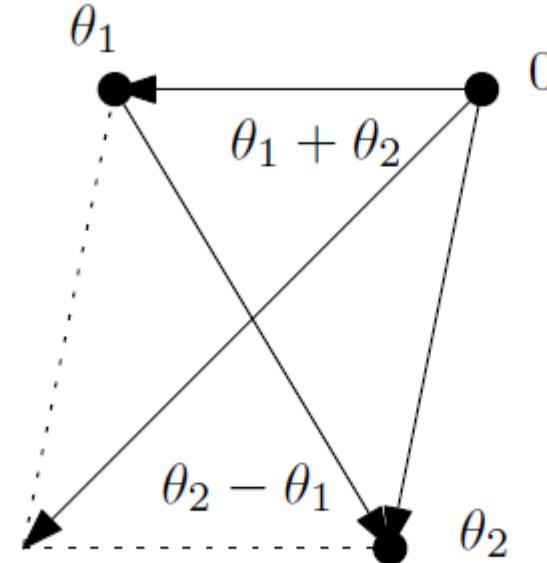


$$\underline{B_F(\theta_1 : \theta) + B_F(\theta_2 : \theta) = B_F\left(\theta_1 : \frac{\theta_1 + \theta_2}{2}\right) + B_F\left(\theta_2 : \frac{\theta_1 + \theta_2}{2}\right) + 2B_F\left(\frac{\theta_1 + \theta_2}{2} : \theta\right)}$$



- In Euclidean geometry:

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



$$\underline{2\|\theta_1\|^2 + 2\|\theta_2\|^2 = \|\theta_1 - \theta_2\|^2 + \|\theta_1 + \theta_2\|^2}$$

# 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**  $\omega$  in the support of the exponential family the Bregman generator:

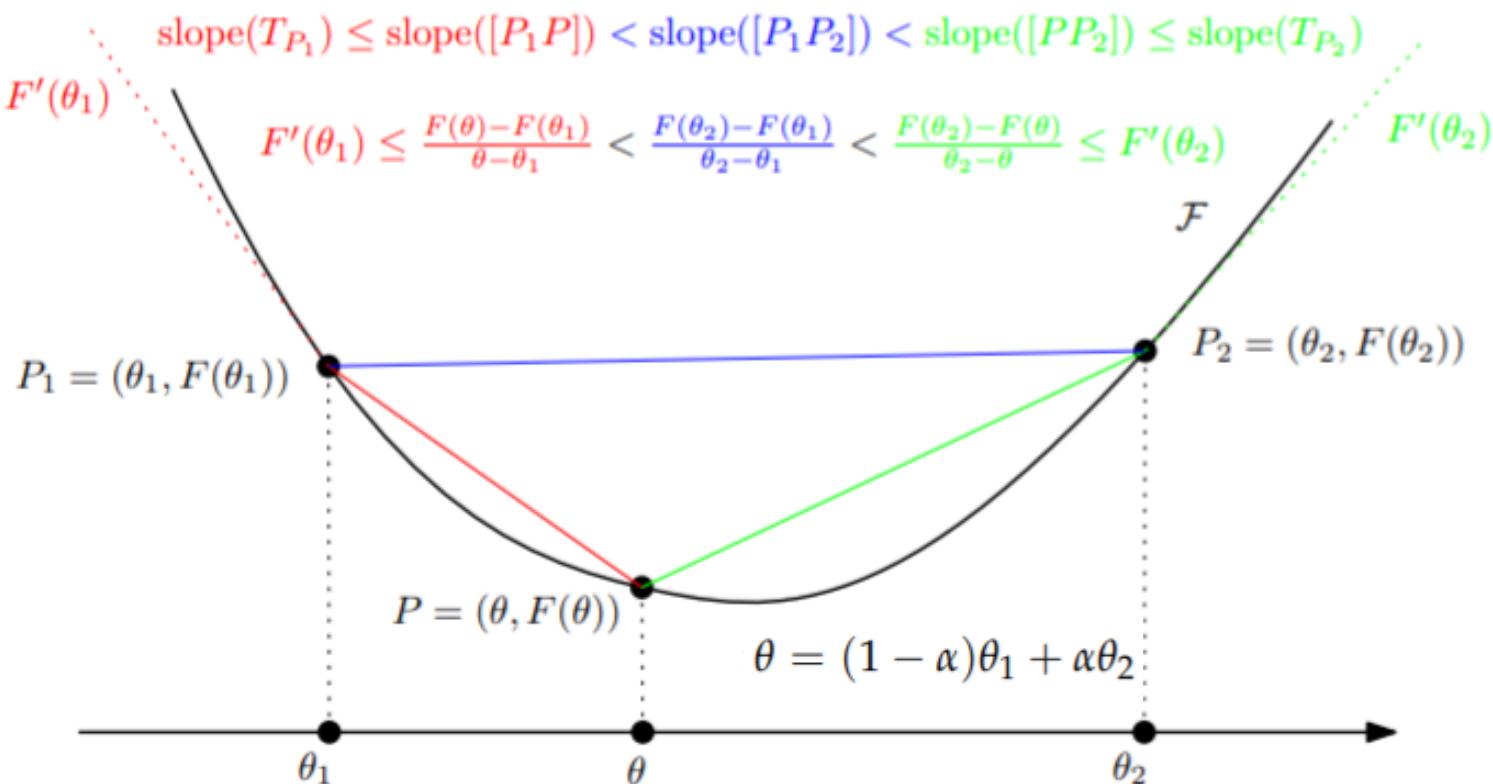
$$F_\omega(\theta) := \underline{-\log(p_\theta(\omega))} = F(\theta) - \underbrace{(\theta^\top t(\omega) + k(\omega))}_{\text{affine term in } \theta}$$

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

# Statistical invariance: Information & Divergence

- Fisher information is invariant under sufficient statistics:
- A statistical divergence is a smooth dissimilarity distance, not necessarily symmetric nor satisfying the triangle inequality of distance metric.

# Chordal slope lemma & Jensen/Bregman divergences



**Jensen  
Divergence (JD)**

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

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

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

---


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

**Bregman  
Divergences  
(BDs):**

$$F'(\theta_1) \leq \frac{F(\theta_2) - F(\theta_1)}{\theta_2 - \theta_1} \leq F'(\theta_2)$$



$$\begin{aligned} F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)F'(\theta_1) &\geq 0, \\ F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)F'(\theta_2) &\leq 0. \end{aligned}$$



$$\begin{aligned} \frac{B_F(\theta_2 : \theta_1)}{B_F(\theta_1 : \theta_2)} &\geq 0, \\ \frac{B_F(\theta_1 : \theta_2)}{B_F(\theta_2 : \theta_1)} &\geq 0. \end{aligned}$$

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

[EIG, Entropy 2020]

# Bregman manifolds vs Hessian manifolds



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

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

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

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

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

Here,  $\nabla$  = gradient

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

Here,  $\nabla, \nabla^*$  = affine flat connections

$$\nabla : \Gamma_{ijk}(\theta) = 0$$

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

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

# Rao's distance between 1D normal distributions

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

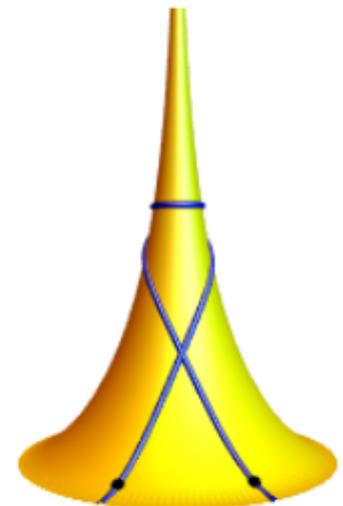
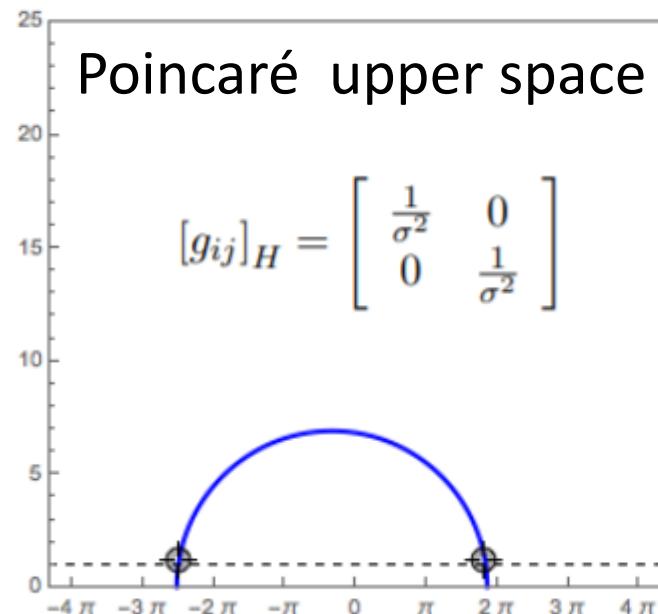
FIM of normal distributions

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

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

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

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



Pseudo-sphere  
*partial* embedding  
in  $\mathbb{R}^3$