SOPHIA

From geometric learning machines to the geometry of AI

Frank Nielsen

@FrnkNlsn

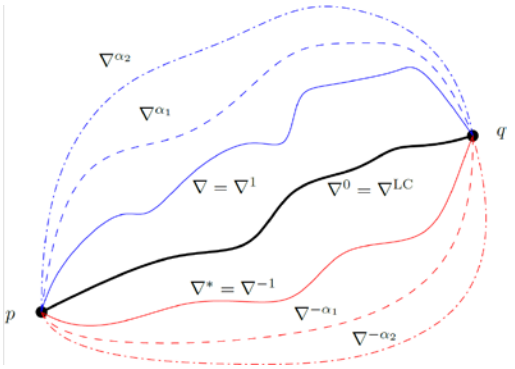
21st November 2019

Sony CSL

Outline of this talk

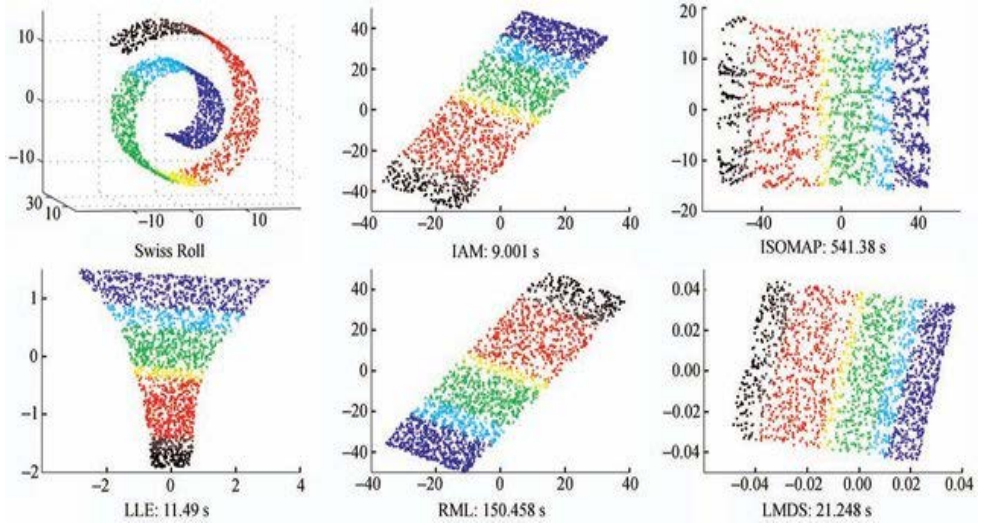
Computational geometry for ML!

- Introduction to some famous **geometric learning machines**
  - **Unsupervised learning**: Riemannian manifold learning
  - **Supervised learning**:
    - Kernel machines and Hilbert geometry (RKHS)
    - Deep learning and trajectories on neuromanifolds
- **Information geometry** and **information projections**:
  - The *dualistic* geometric structures
- Geometry of **interpolation machines**:
  - Double descent learning curves



Non-linear versus linear dimension reduction

- Many **non-linear techniques**:
  - LLE,
  - ISOMAP
  - etc.



- In very high dimensions, **linear dimension projection** (curse of dimensionality). **Johnson-Lindenstrauss' theorem (1984)**

Let X be a set of n points of $\mathbb{R}^d$ and fix a $\epsilon \in (0,1)$. Then there exists a linear transformation $A : \mathbb{R}^d \to \mathbb{R}^k$ with $k = O(\frac{1}{\epsilon^2} \log n)$ such that:

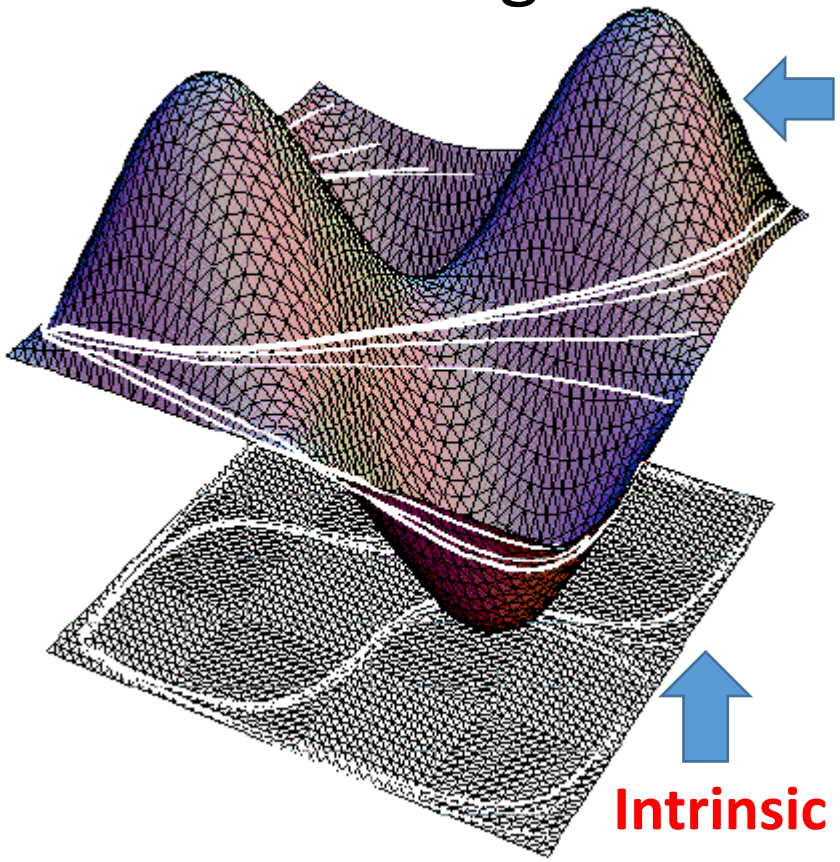
$$\forall x, x' \in X, \quad (1-\epsilon)\|x - x'\|^2 \leq \|xA - x'A\|^2 \leq (1+\epsilon)\|x - x'\|^2$$

Introduction to HPC with MPI for Data Science, Springer, 2016

Riemannian manifolds: Extrinsic view vs intrinsic view

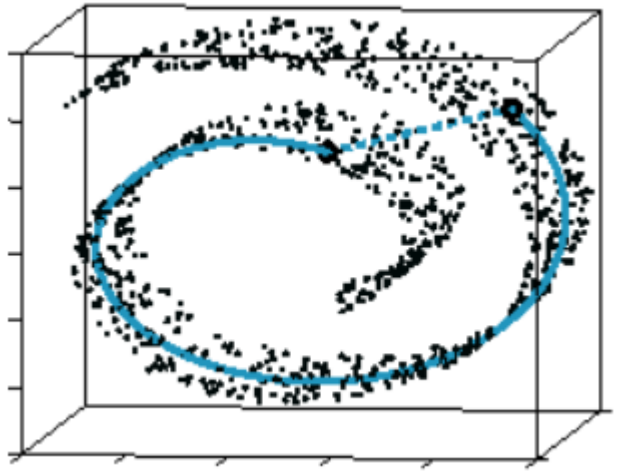
Visualized **extrinsically** as smooth surfaces of the **ambient** Euclidean space: isometric embedding theorem (Nash embedding theorem)

Isometric embedding:



Extrinsic geometry

Intrinsic geometry

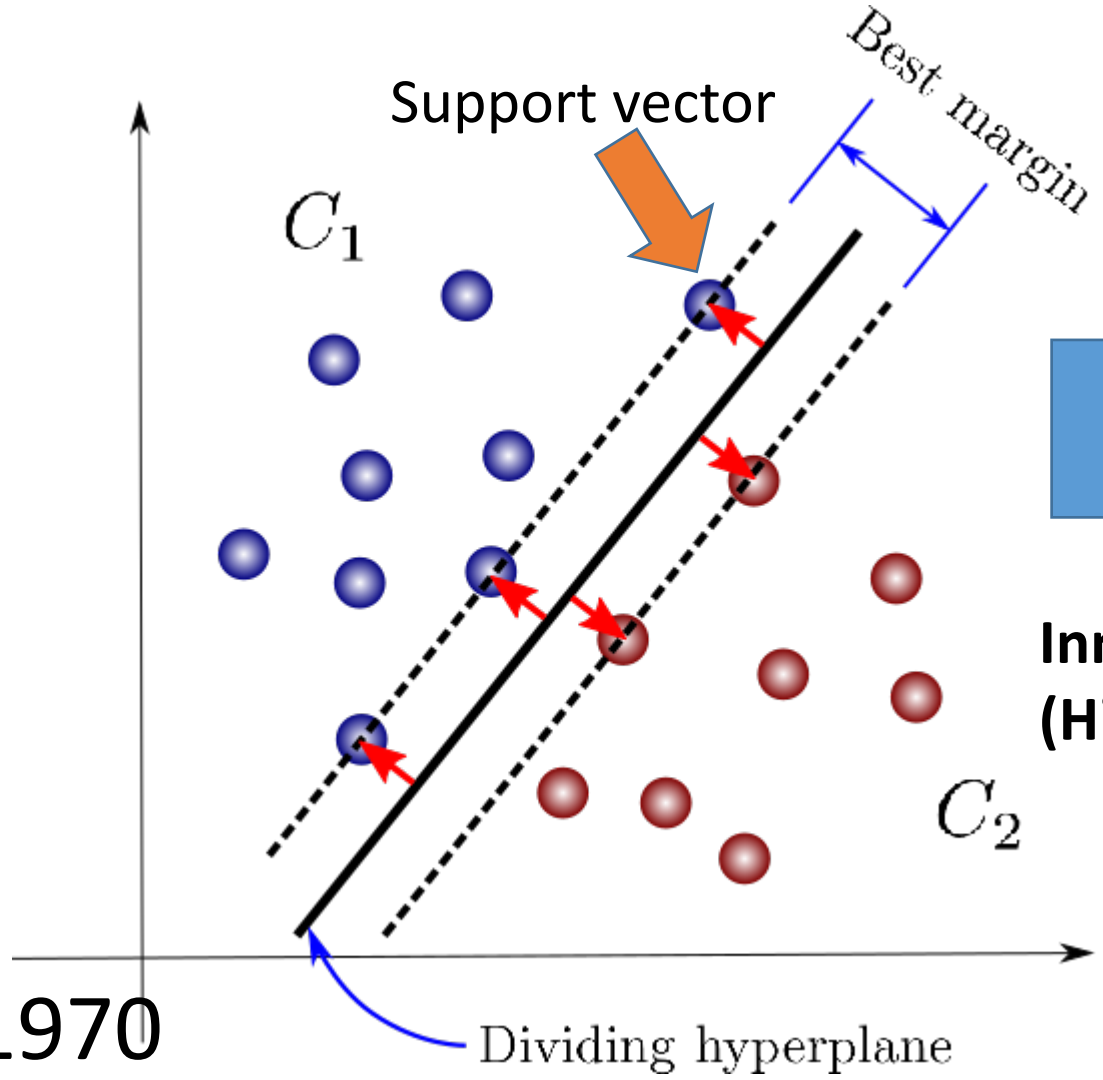


Manifold learning/reconstruction from data points (Swiss roll)

Intrinsic geometry versus **extrinsic** isometric embedding

Kernel Support Vector Machines (SVMs)

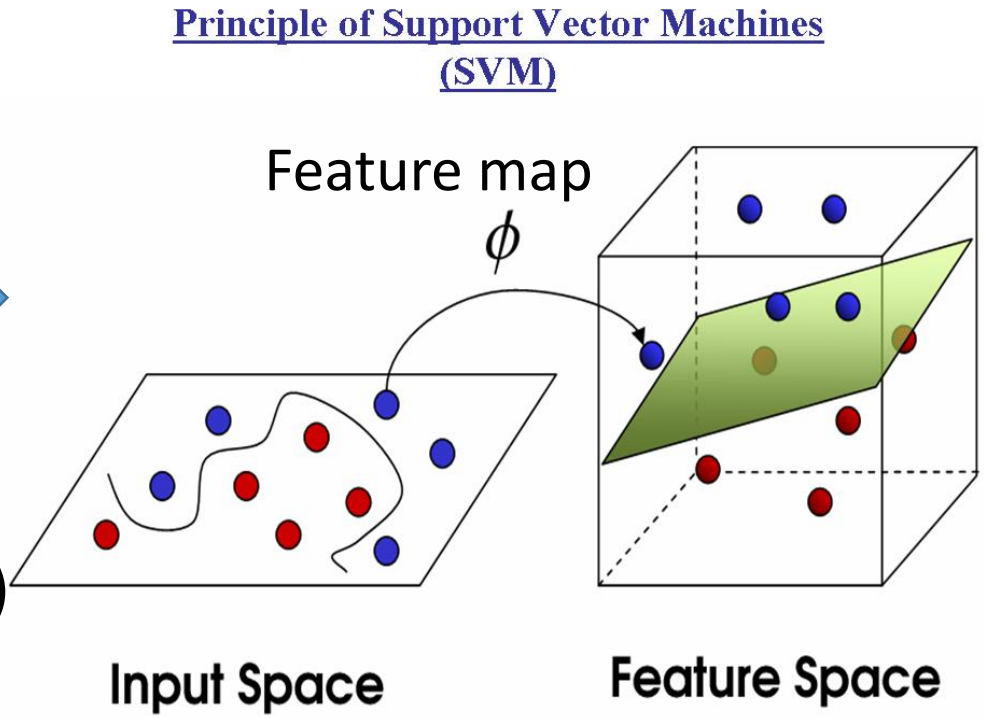
Linear separator



1970

Inner product (Hilbert space)

Non-linear separator

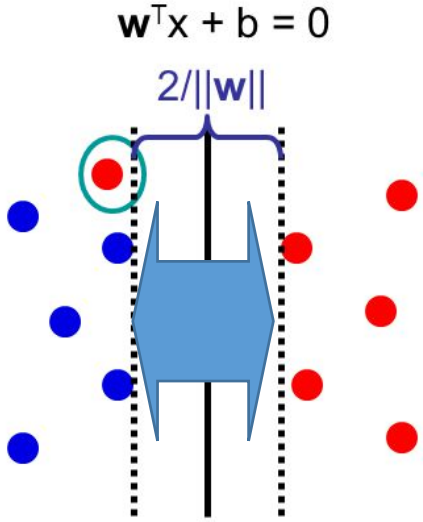


1992

Kernel machines (RKHS geometry)

SVM: Dual quadratic program amount to solve for a smallest enclosing ball (= SEB): Computational geometry !

The SVM Framework



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

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

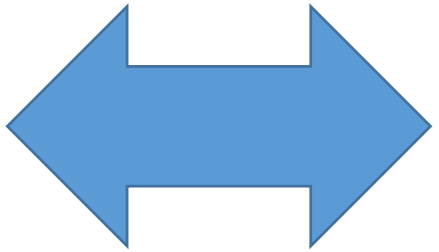
$$\xi_i \geq 0$$

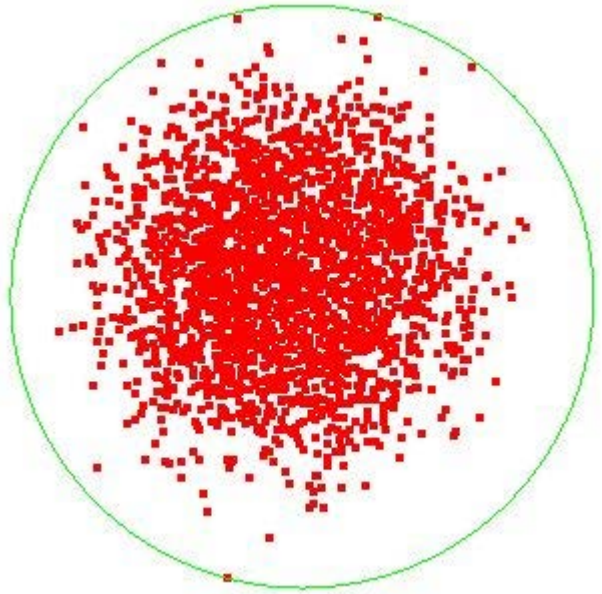
Points $\mathbf{X} = \{\mathbf{x}_i\}$

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

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

Convex Quadratic Program



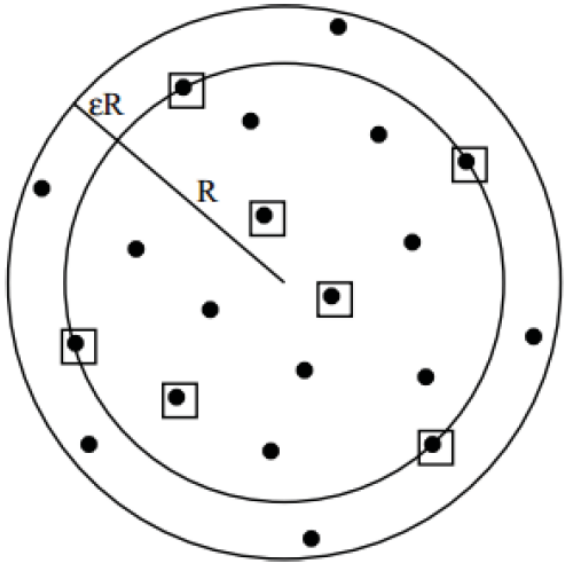


Smallest enclosing ball: Smallest ball with respect to radius or set inclusion

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

Coresets for Smallest Enclosing Balls and Core VMs

- Definition: A **coreset** is a subset $C \subseteq X$ such that [BC 2002]



Selected points inside boxes

$$X \subseteq \mathrm{Ball}(c(C), (1+\epsilon)r(C)).$$

Independent of input size n
Independent of dimension d
Only dependent of epsilon ☺

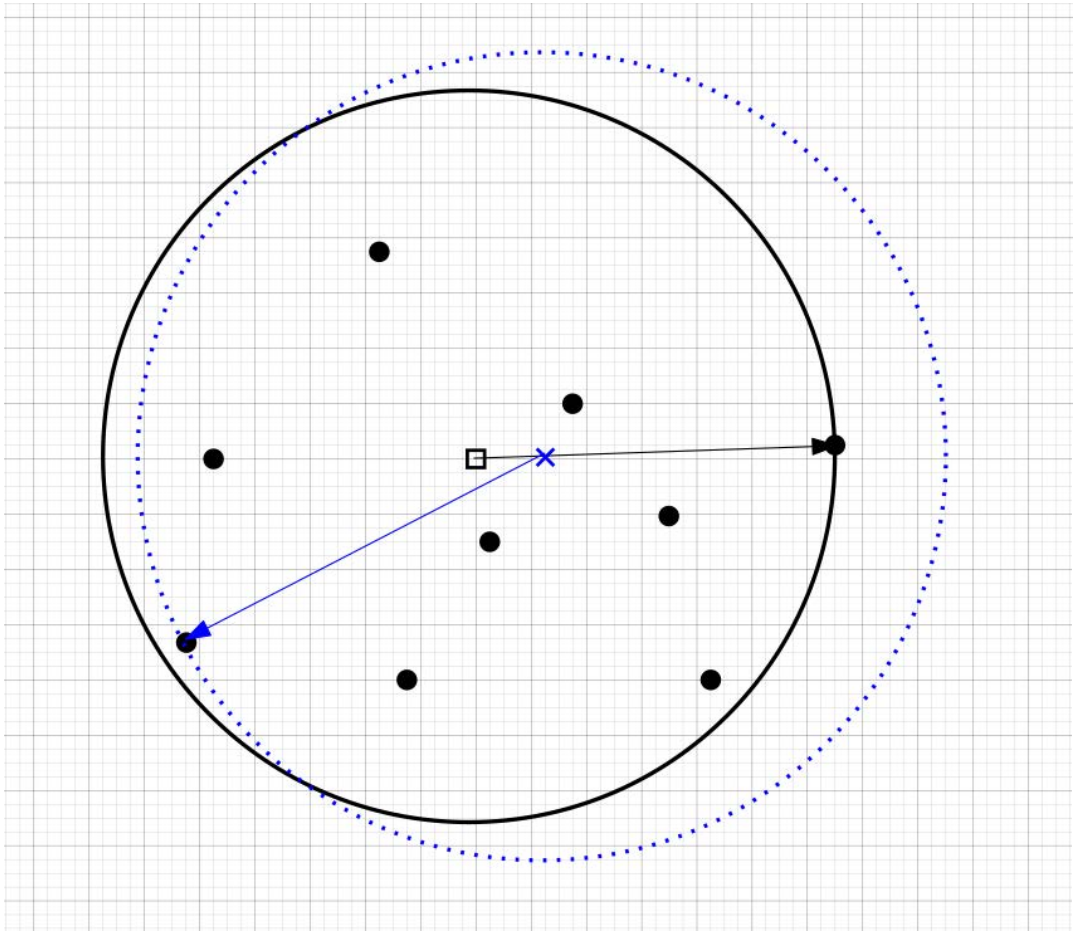
Apply to finite point sets in **infinite-dimensional spaces** too!

-> Kernel machines with Core Vector Machines

A note on kernelizing the smallest enclosing ball for machine learning, 2017

Introduction to HPC with MPI for Data Science, Springer, 2016

Computing a coreset for the Smallest Enclosing Balls



Extremely simple algorithm ☺ !

#iterations: $l = \lceil \frac{1}{\epsilon^2} \rceil$

Running time: $O(\frac{dn}{\epsilon^2})$

BC-ALG:

- Initialize the center $c_1 \in P$, and
- Iteratively update the current center using the rule

$$c_{i+1} \leftarrow c_i + \frac{f_i - c_i}{i+1},$$

where f_i denotes the farthest point of P to c_i.

On approximating the Riemannian 1-center. Comput. Geom. 46(1): 93-104 (2013)

Approximating Covering and Minimum Enclosing Balls in Hyperbolic Geometry. GSI 2015

Approximating the kernelized minimum enclosing ball

Kernel $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ with feature map $\Phi(x) : \mathbb{R}^d \to \mathbb{R}^D$

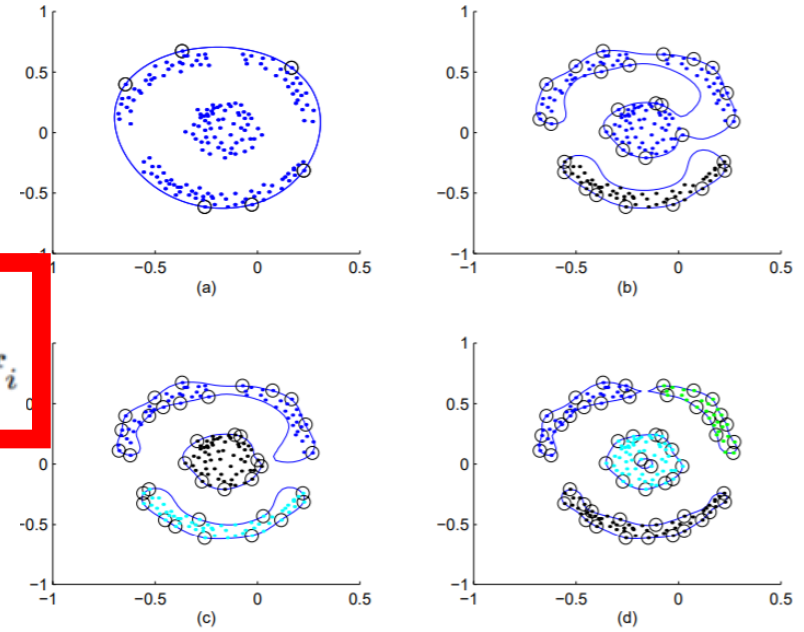
(D may be infinite)

Trick: **Encode implicitly** the circumcenter φ of the enclosing ball as a convex combination of the data points: $\varphi = \sum_i \alpha_i \phi(p_i) = \sum_i \alpha_i \phi_i$

$$\begin{aligned} \|\varphi - \phi(p)\|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \langle \phi_i, \phi_j \rangle - 2\sum_{i=1}^{n} \alpha_i \langle \phi_i, \phi(p) \rangle + \langle \phi(p), \phi(p) \rangle, \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(p_i, p_j) - 2\sum_{i} \alpha_i k(p_i, p) + k(p, p). \end{aligned}$$

Update weights iteratively: $\alpha^{(i+1)} = \frac{i}{i+1}\alpha^{(i)} + \frac{1}{i+1} e_{f_i}$

e_{f_i} Index of the current farthest point



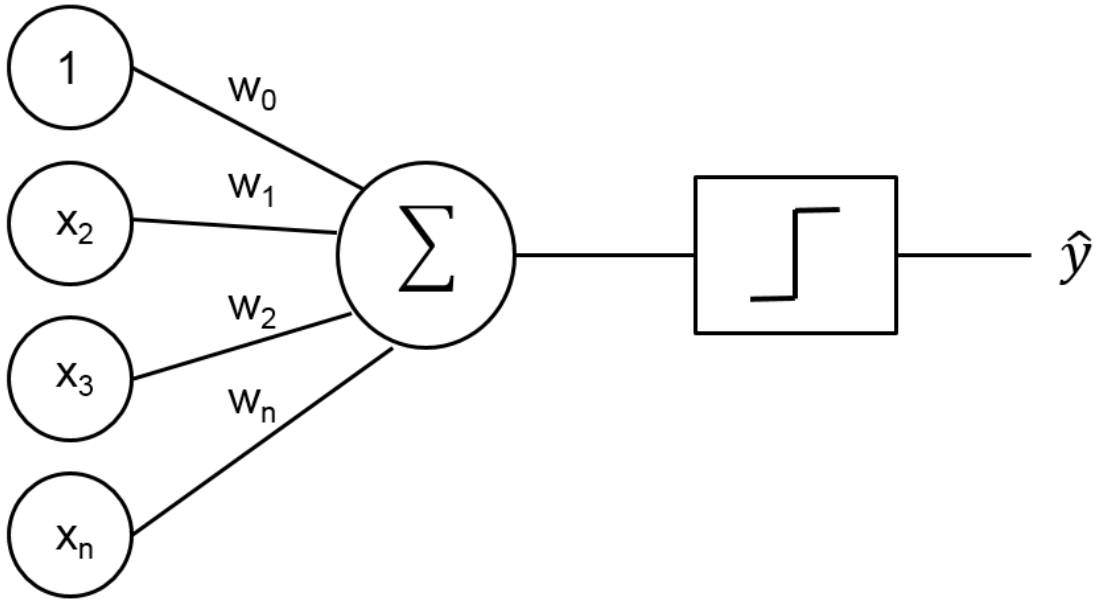
Applications: Support Vector Data Description, Support Vector Data Description

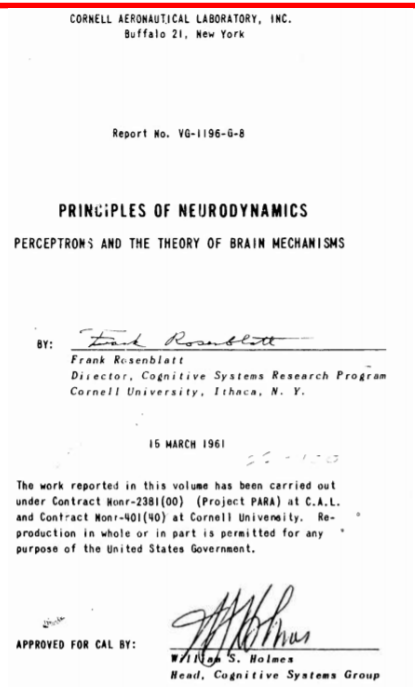
A note on kernelizing the smallest enclosing ball for machine learning, 2017

The 1-layer perceptron: linear separator machine

Frank Rosenblatt:
*Principles of **Neurodynamics**, 1962*

Marvin Minkowski and Seymour Papert:
*Perceptrons: An Introduction to **Computational Geometry**, 1969*





CORNELL AERONAUTICAL LABORATORY, INC.
Buffalo 21, New York

Report No. VG-1196-G-8

PRINCIPLES OF NEURODYNAMICS

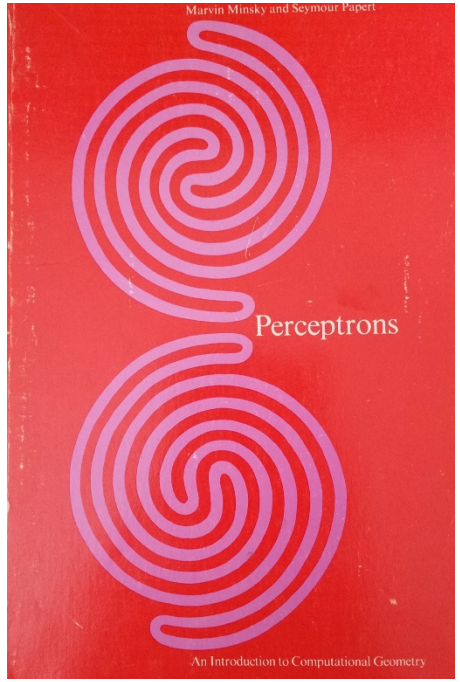
PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS

BY: Frank Rosenblatt
Frank Rosenblatt
Director, Cognitive Systems Research Program
Cornell University, Ithaca, N. Y.

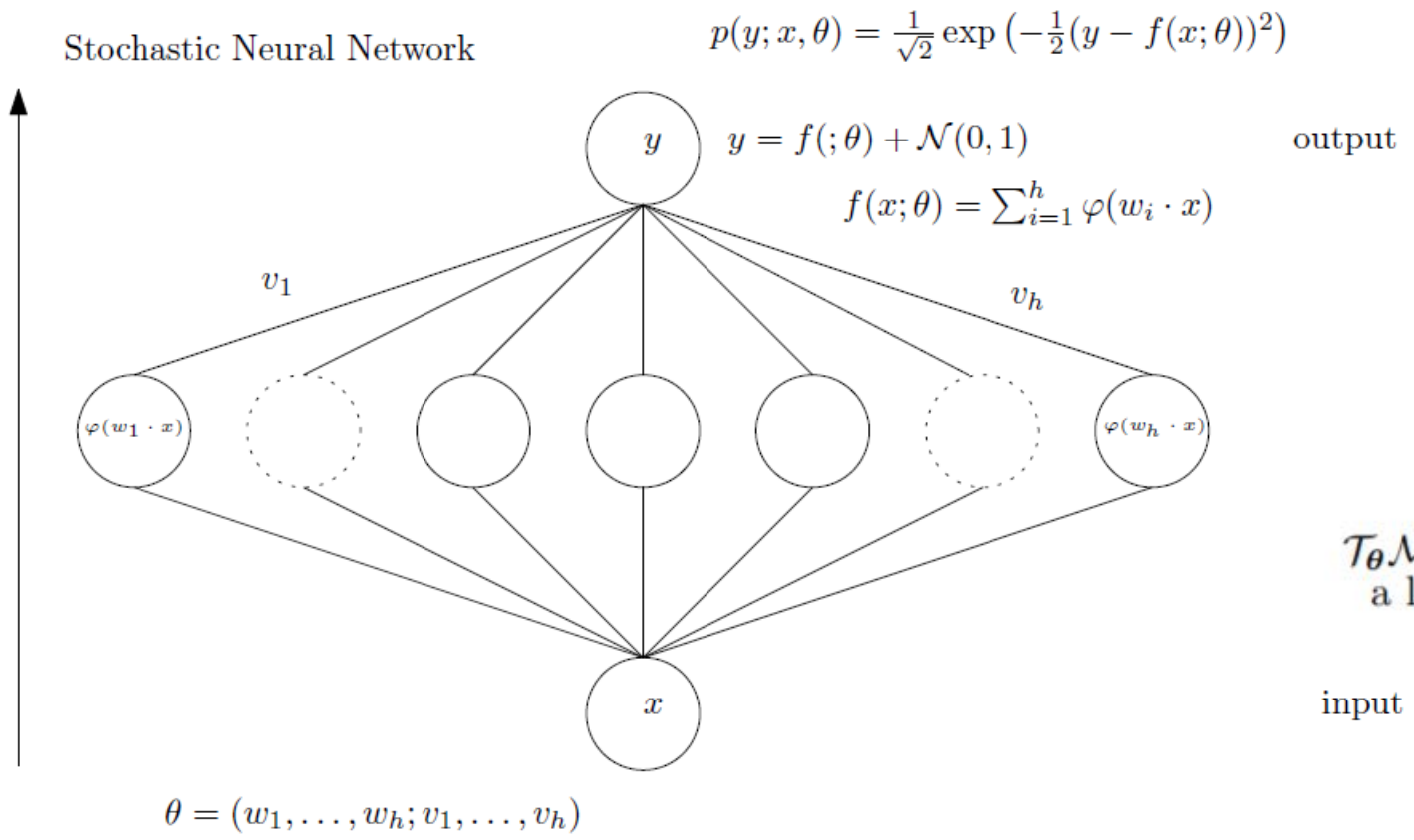
15 MARCH 1961

The work reported in this volume has been carried out under Contract Nonr-2381(00) (Project PARA) at C.A.L. and Contract Nonr-401(40) at Cornell University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

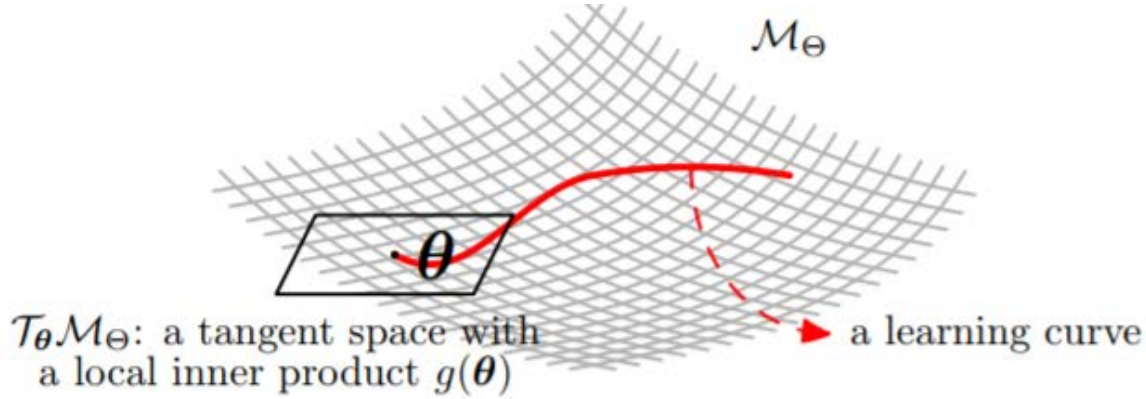
APPROVED FOR CAL BY: Wilbur S. Holmes
Head, Cognitive Systems Group



Stochastic MLPs and neuromanifolds



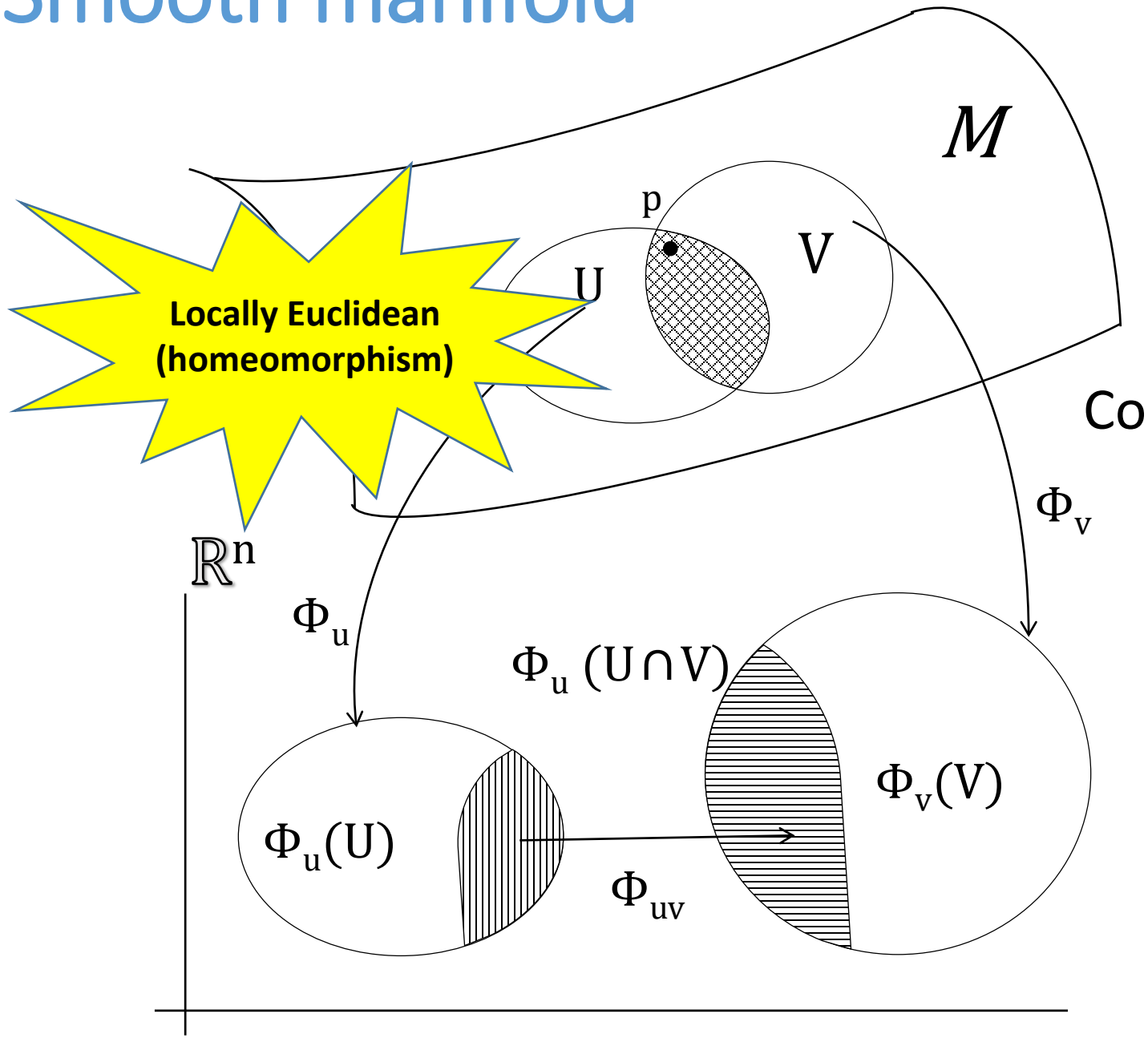
Neurodynamics=
Learning trajectory on the manifold



Parameter space

Hidden layers: universal function approximators (post XOR)
Supervised learning: gradient descent + backpropagation

Smooth manifold



Global geometric objects

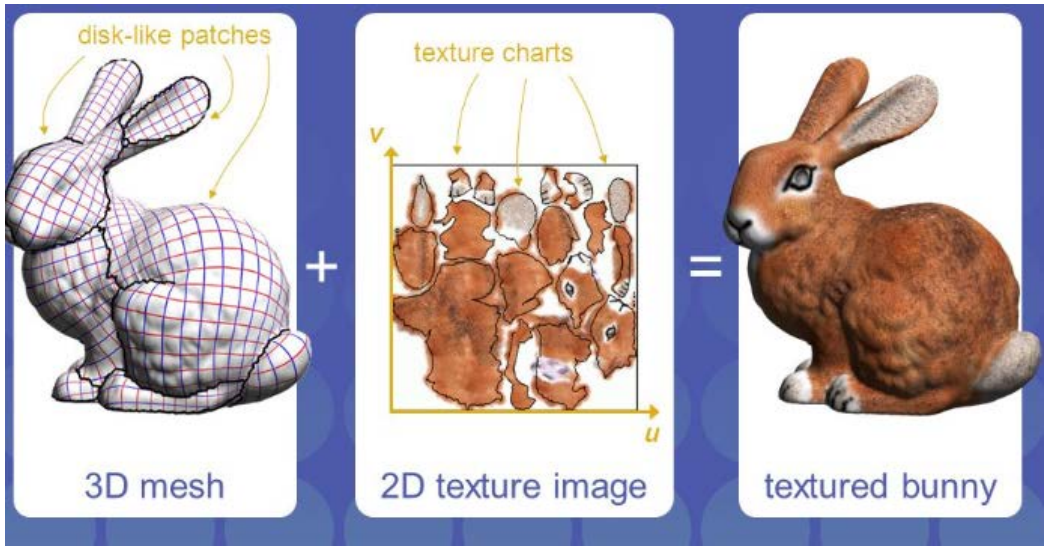
vs

Local descriptions

in local chart coordinates

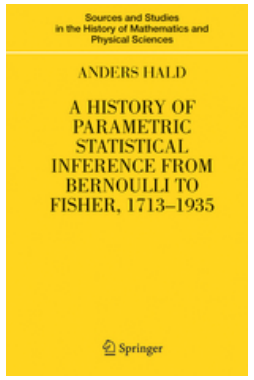
Atlas= set of charts

Coordinates in charts+transition maps



UV mapping in computer graphics

Fisher information matrix (FIM)



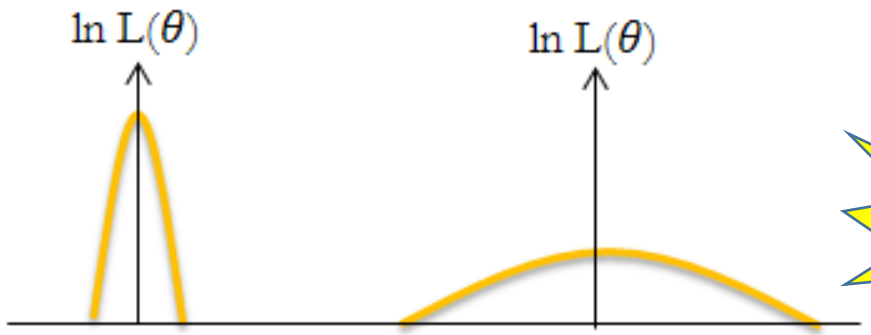
$$g(\xi)=E_\xi[\frac{\partial}{\partial\xi}\log(p_\xi)\,\frac{\partial}{\partial\xi}\log(p_\xi)]$$

$$g_{ij}(\xi)=\int\frac{\partial}{\partial\xi}\log(p_\xi(x))\frac{\partial}{\partial\xi}\log(p_\xi(x))p_\xi(x)dx$$

Sir Ronald Fisher

FIM is positive-semidefinite, **positive-definite** for regular models

$$\text{Curvature} = -\frac{\partial^2}{\partial\theta^2}[\ln L(\theta)]$$



More Sharpness
Less Variance
High Fisher Information

Less Sharpness
More Variance
Low Fisher Information

$$g_{ij}(\theta) = E\left\{\frac{\partial}{\partial\theta_i}\log p(X\mid\theta)\,\frac{\partial}{\partial\theta_j}\log p(X\mid\theta)\mid\theta\right\}$$

1922

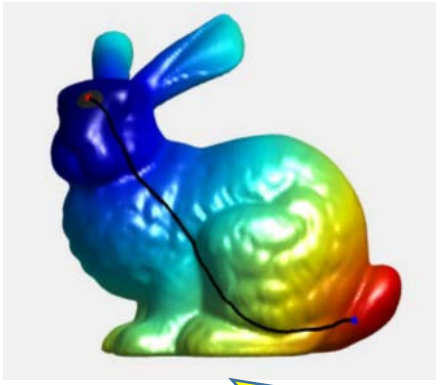
IX. *On the Mathematical Foundations of Theoretical Statistics.*

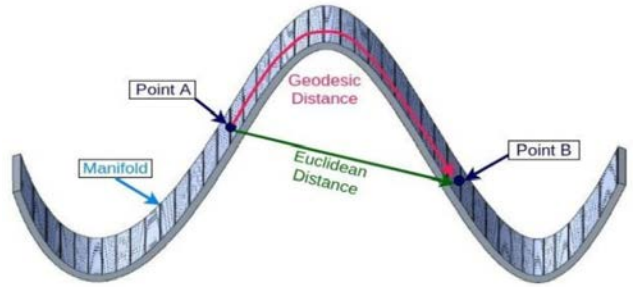
By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Fisher-Rao geometry and geodesic distance

Fisher information metric (FIM)





- Infinitesimal length element :

$$\mathrm{d}s^2 = \sum_{ij} g_{ij}(\theta)\mathrm{d}\theta_i \mathrm{d}\theta_j = \mathrm{d}\theta^T I(\theta)\mathrm{d}\theta$$

independent to reparameterization of θ

- Geodesic and distance are hard to explicitly calculate :

$$\rho(p(x;\theta_1), p(x;\theta_2)) = \min_{\substack{\theta(s) \\ \theta(0)=\theta_1 \\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{\mathrm{d}\theta}{\mathrm{d}s}\right)^T I(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}s}}\mathrm{d}s$$

Riemannian geodesics locally minimize lengths

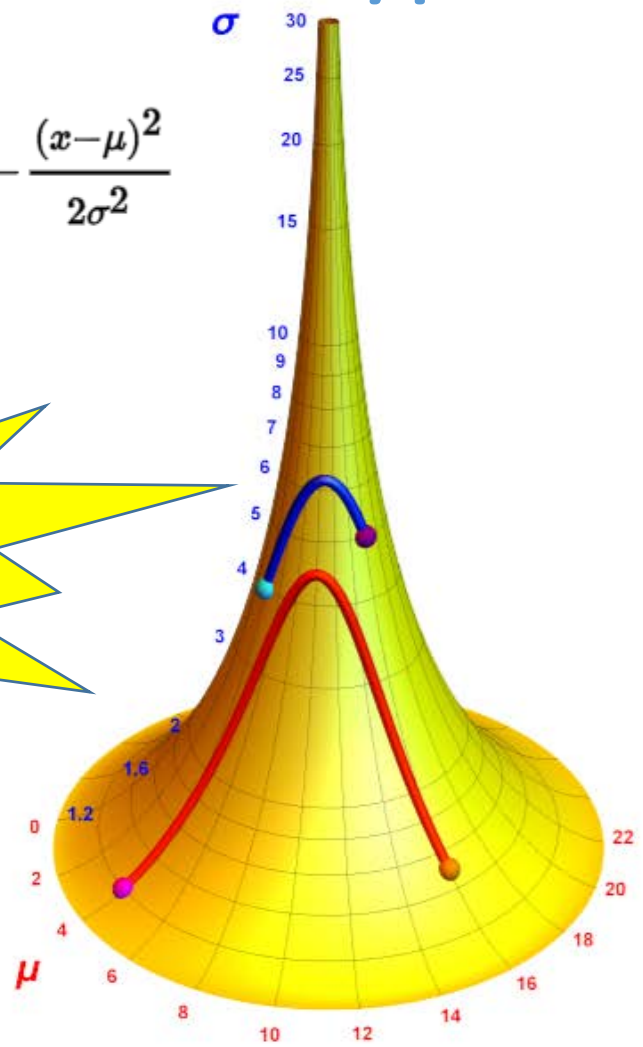
- Metric property of ρ, many tools [1] : Riemannian Log/Exp tangent/manifold mapping

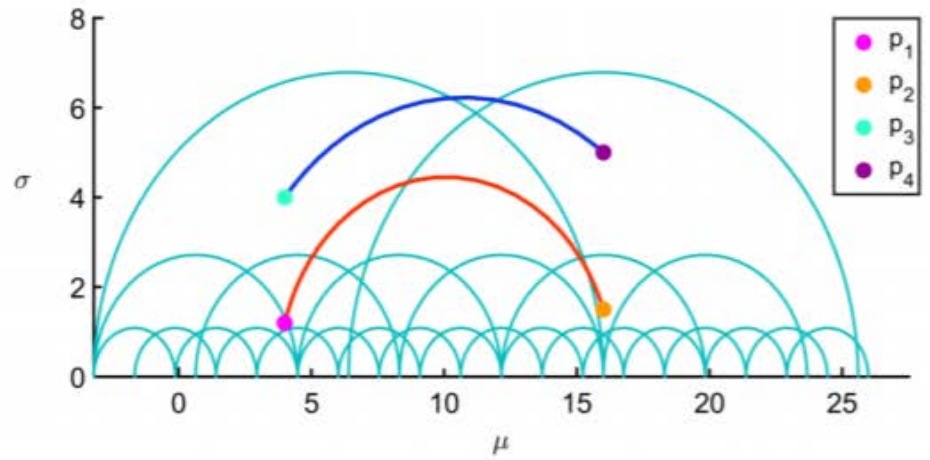
Riemannian geometry of normal distributions equipped with FIM: hyperbolic geometry

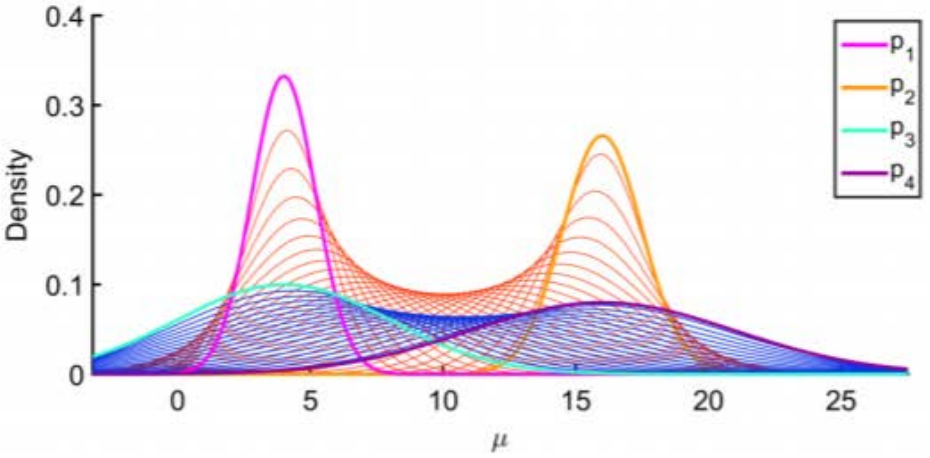
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Metric tensor: Measure vector lengths, angles on tangent space

Pseudo-sphere
(negative curvature -1/2)







Pattern recognition in nuclear fusion data by means of geometric methods in probabilistic spaces, 2017

Cramer-Rao lower bound: Inverse of Fisher information

Löwner partial ordering on positive-semi-definite matrices: $A \succeq B \Leftrightarrow A - B \succeq 0$

CRLB Theorem:

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta_0)^{-1}$$

Accuracy of estimator depends on true parameter

$$\begin{aligned}[I(\theta)]_{ij} &= E_\theta \left[\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right], \\ &= \int \left(\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right) p_\theta(x) \mathrm{d}x.\end{aligned}$$

Under regularity conditions:

$$[I(\theta)]_{ij} = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x) \right]$$

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

Calculating Rao's distance is often untractable

$$d\left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\right) = \min_{\boldsymbol{\theta}(t)} \int_{t_1}^{t_2} \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} g_{ij}(\boldsymbol{\theta}(t)) \frac{d\theta_i(t)}{dt}\frac{d\theta_j(t)}{dt}}\, dt.$$

- Need to solve the **Ordinary Differential Equation** (ODE) for find the geodesic (but trivial in 1D):

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

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

- Need to **integrate** the infinitesimal length elements along the geodesics…

No closed-form Fisher-Rao distance between multivariate normals!

-> geodesic shooting (BVP: boundary value problem, IVP: initial value problem)

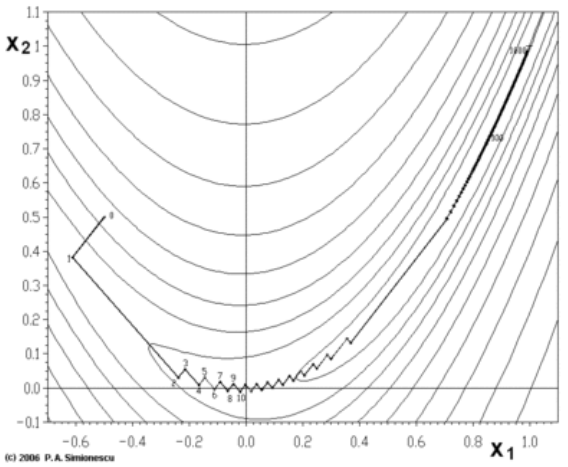
Using the Fisher Information Matrix without geodesics?

Ordinary steepest gradient descent method

- Iterative optimization algorithm
- Start from an initial parameter value θ_0
- **Update iteratively** the current parameter using a learning rate α (step size) and the gradient of the energy function:

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

- First-order optimization method
- Zig-zag local minimum convergence
- Stopping criterion



Similarly, maximization with hill climbing, steepest ascent

Steepest descent in a Riemannian space:

The natural gradient

- The steepest descent direction of E(θ) in a Riemannian space is given by

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

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

Type checking: Contravariantform of the ordinary gradient

Learning rate

Computing the inverse of the Fisher information matrix is tricky!

Amari, Shun-Ichi. "Natural gradient works efficiently in learning." *Neural computation* 10.2 (1998): 251-276.

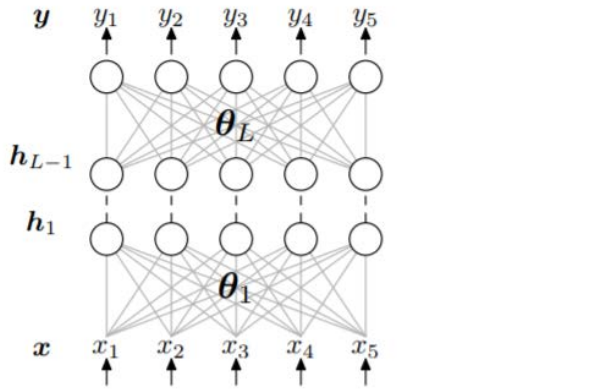
Pros and *cons* of natural gradient

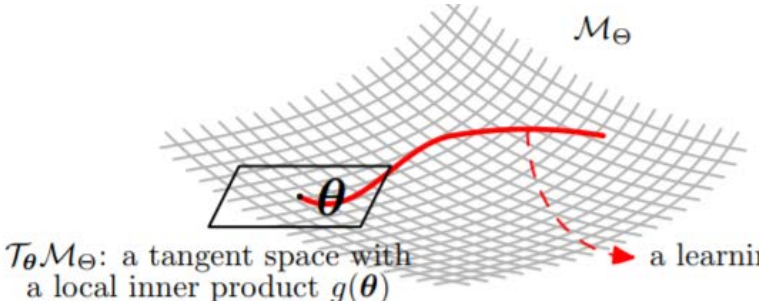
- **Pros:**
  - **Invariant** (intrinsic) gradient (at infinitesimal scale/ODE)
  - **Not trapped in plateaus** (close to degenerate FIM)
  - Achieve **Fisher efficiency** in online learning

- **Cons:**
  - Too *expensive* to compute (no closed-form FIM; need matrix inversion; numerical stability) -> Other Riemannian metrics studied
  - Degenerate for ***irregular models*** (e.g., hierarchical models, Deep learning)
  - Need to adapt the step size

Relative Fisher Information Matrix (RFIM) and Relative Natural Gradient (RNG) for deep learning

$$p(\boldsymbol{y}\,|\,\boldsymbol{x},\Theta) = \sum_{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_{L-1}} p(\boldsymbol{y}\,|\,\boldsymbol{h}_{L-1},\boldsymbol{\theta}_L)\cdots p(\boldsymbol{h}_2\,|\,\boldsymbol{h}_1,\boldsymbol{\theta}_2)p(\boldsymbol{h}_1\,|\,\boldsymbol{x},\boldsymbol{\theta}_1),$$

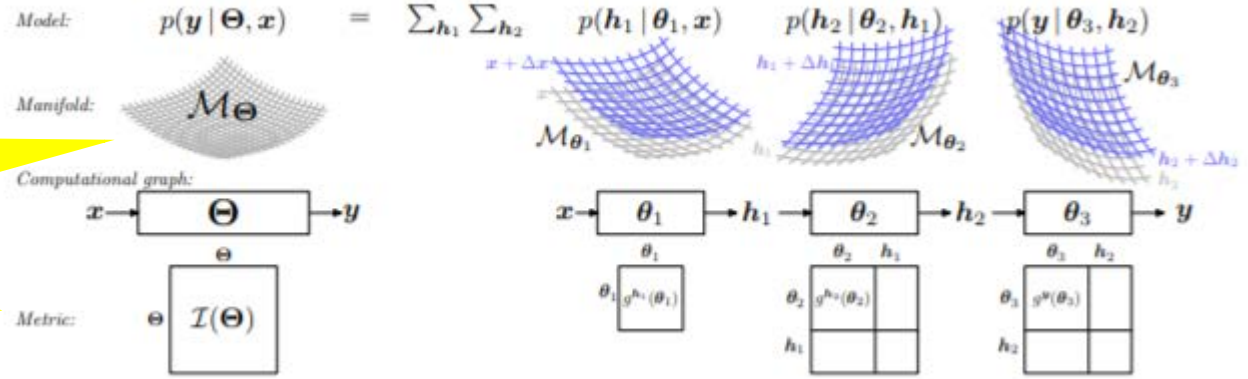




$$g(\Theta) = E_{\boldsymbol{x}\sim\hat{p}(X_n),\,\boldsymbol{y}\sim p(\boldsymbol{y}\,|\,\boldsymbol{x},\Theta)}\left[\frac{\partial l}{\partial\Theta}\frac{\partial l}{\partial\Theta^{\mathsf{T}}}\right] = \frac{1}{n}\sum_{i=1}^{n} E_{p(\boldsymbol{y}\,|\,\boldsymbol{x}_i,\Theta)}\left[\frac{\partial l_i}{\partial\Theta}\frac{\partial l_i}{\partial\Theta^{\mathsf{T}}}\right]$$

Relative Fisher IM: $$g^{\boldsymbol{h}}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_f) = E_{p(\boldsymbol{h}\,|\,\boldsymbol{\theta},\boldsymbol{\theta}_f)}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln p(\boldsymbol{h}\,|\,\boldsymbol{\theta},\boldsymbol{\theta}_f)\frac{\partial}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\ln p(\boldsymbol{h}\,|\,\boldsymbol{\theta},\boldsymbol{\theta}_f)\right]$$

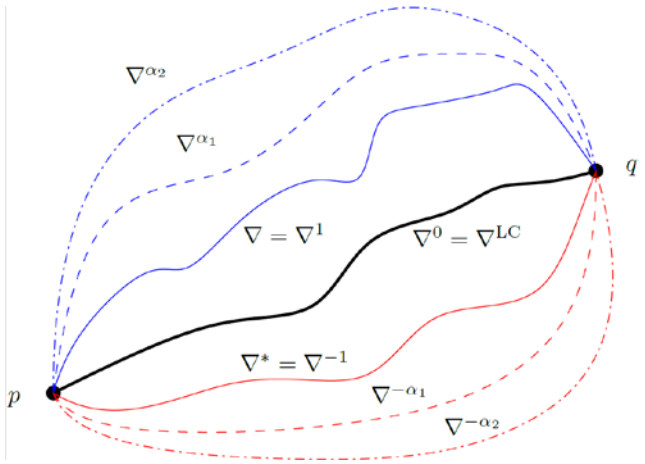
Dynamic geometry!



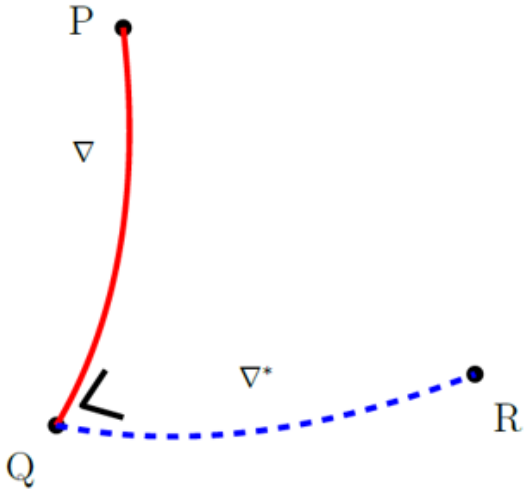
The RFIMs of single neuron models, a linear layer, a non-linear layer, a soft-max layer, two consecutive layers all have simple RFIM closed form solutions

Relative Fisher Information and Natural Gradient for Learning Large Modular Models (ICML'17)

Dualistic structures of Information geometry



$$(M, g, \nabla^e, \nabla^m)$$



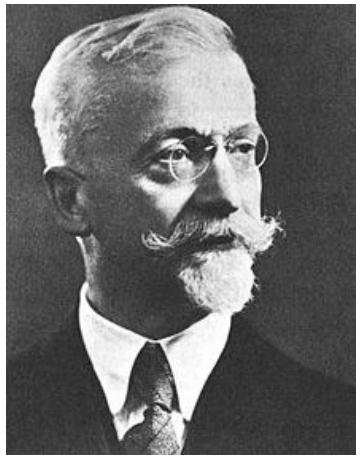
and Information projections

(may seem at first counterintuitive)

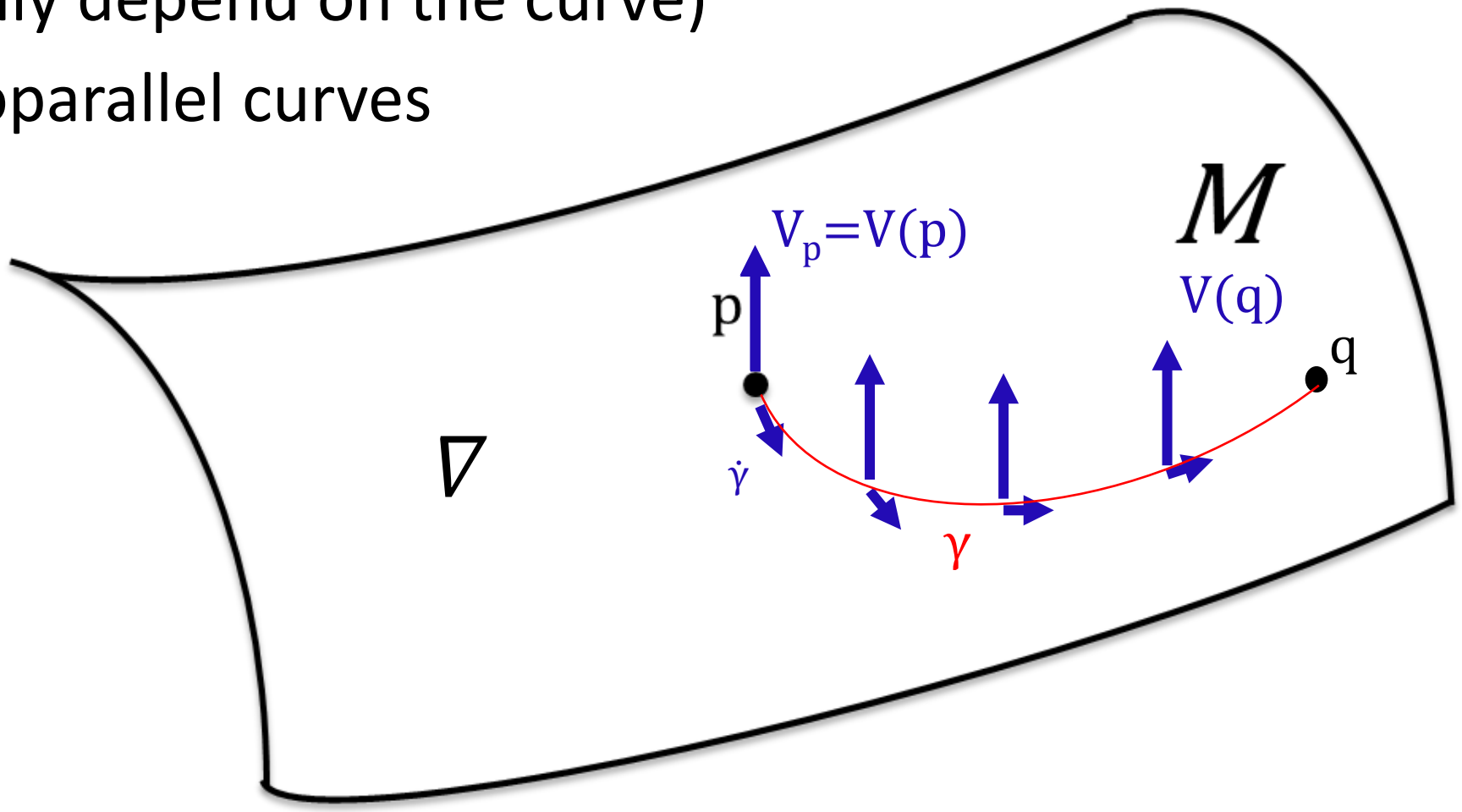
An elementary introduction to information geometry, arXiv:1808.08271

An essential concept: Affine Connection ∇

- Define how to "**parallel transport**" a vector from one tangent plane to another tangent plane by infinitesimally parallel shifting it along a curve (thus generally depend on the curve)
- **∇-geodesics** = autoparallel curves



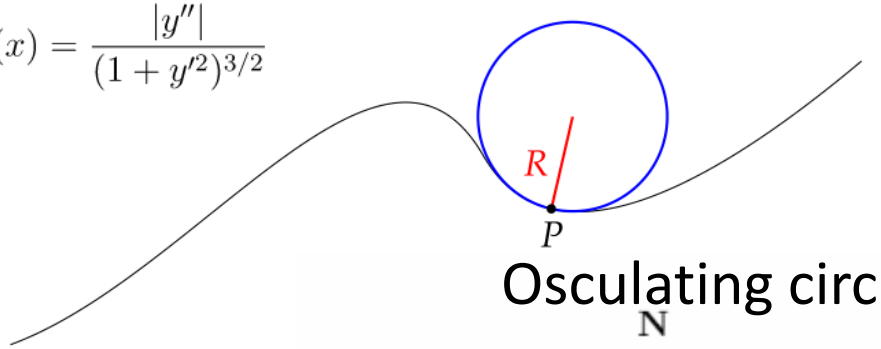
Elie Joseph Cartan
(1869-1951)



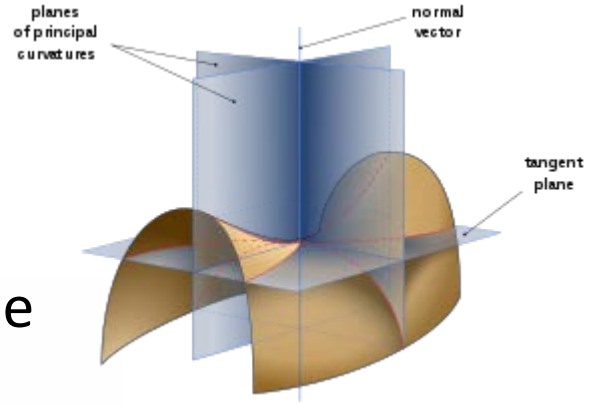
https://arxiv.org/abs/1808.08271

Curvature of a connection ∇

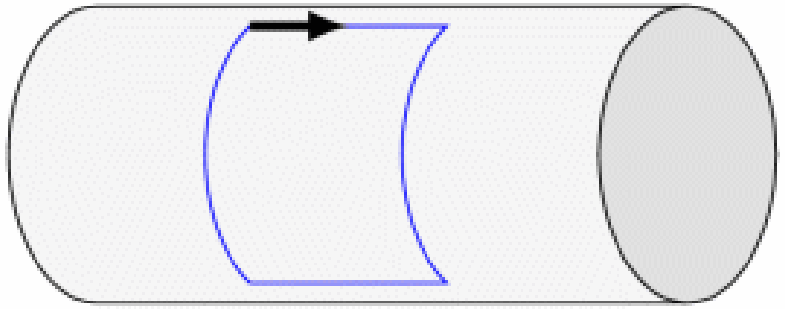
$$\kappa(x) = \frac{|y''|}{(1+y'^2)^{3/2}}$$



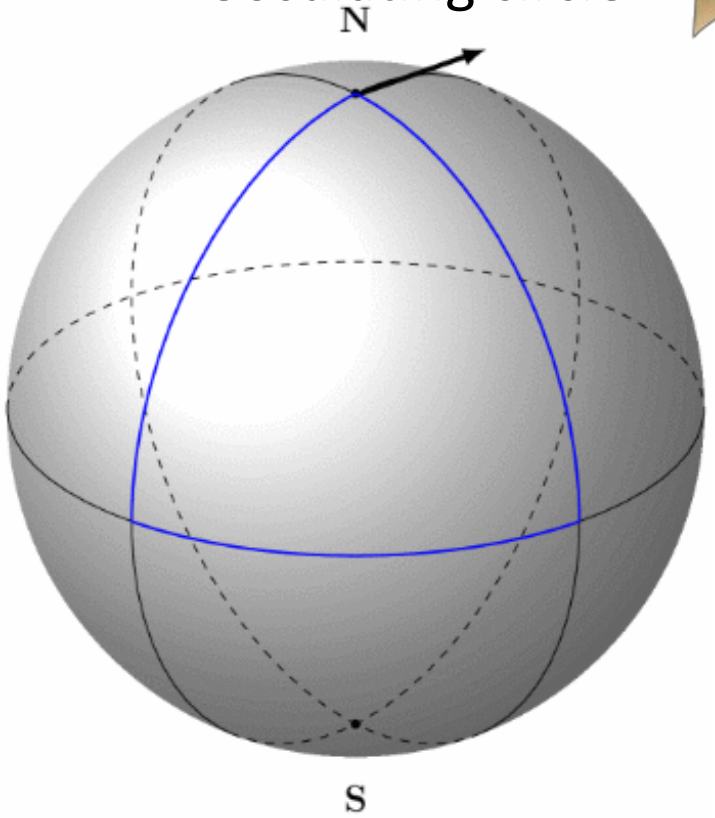
Osculating circle



Sectional curvatures



Cylinder is flat:
Parallel transport is path-independent



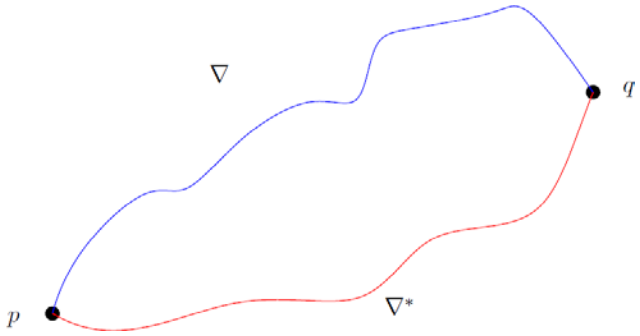
Sphere has constant curved curvature:
Parallel transport is path-dependent

Dual exponential/mixture affine connections

Historically, built the **e-connection** (exponential) and **m-connection** (mixture) for statistical models

$$(M, g, \nabla^e, \nabla^m)$$

Log-likelihood $\ell(p_\xi)(x) = \ln p_\xi(x)$



e-connection ∇^e

$$\Gamma^{(1)}_{ij,k}(\xi) = g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = E_\xi[(\partial_i\partial_j\ell)\,(\partial_k\ell)].$$

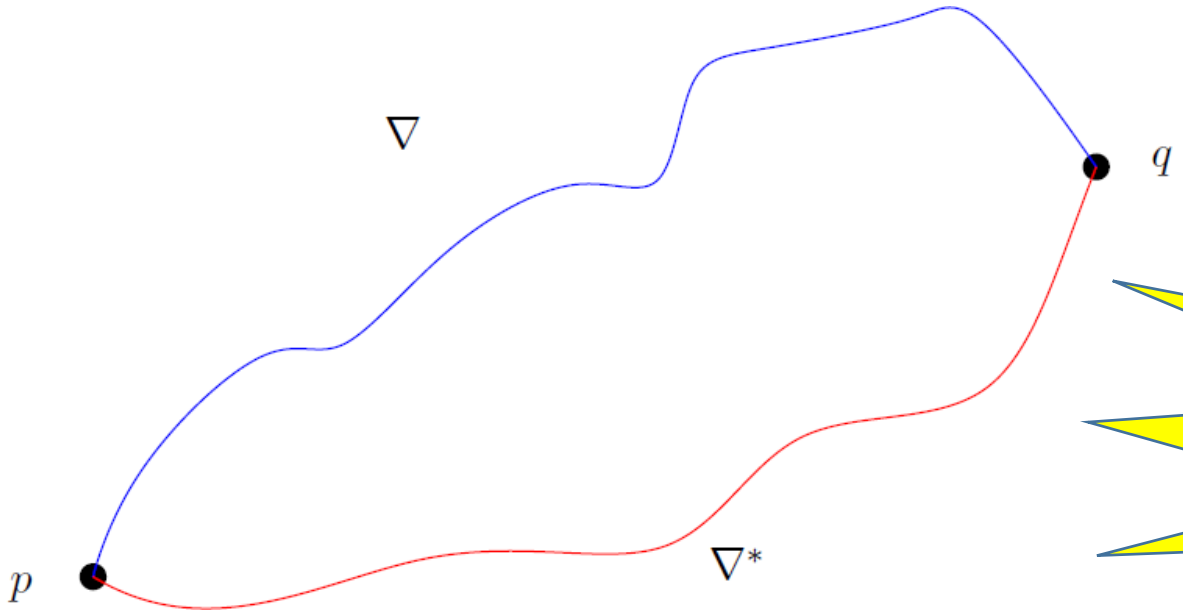
m-connection ∇^m

$$g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(-1)}_{ij,k} = E_\xi[(\partial_i\partial_j\ell + \partial_i\ell\,\partial_j\ell)\,(\partial_k\ell)]$$

May not need distances here

DUAL CONNECTIONS wrt. the Fisher information (Riemannian) metric

Dualistic structure of the Gaussian manifold



∇: e-connection, flat

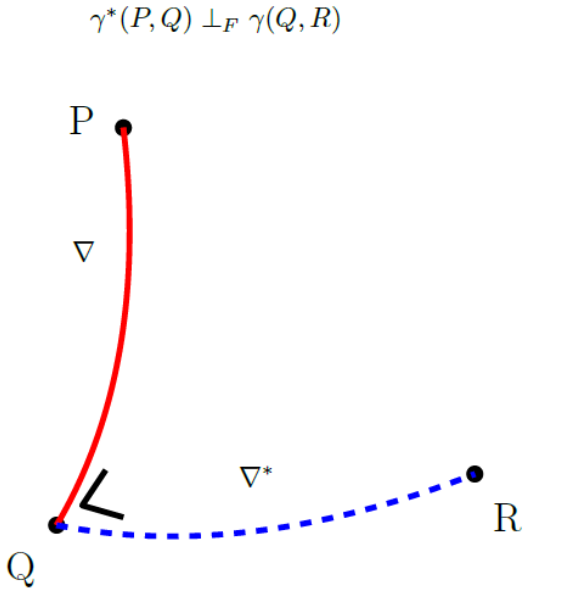
∇*:m-connection, flat

Dually flat space!
(both connections flat)

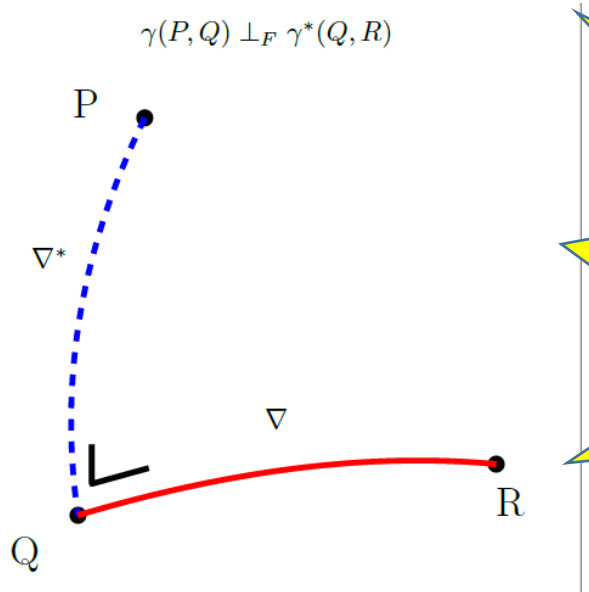
M-geodesic $(p_1 p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ \Sigma^m_\alpha = \bar{\Sigma}_\alpha + (1-\alpha)\mu_1\mu_1^\top - \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha{}^\top \end{cases}$

E-geodesic $(p_1 p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \Sigma^e_\alpha((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2) \\ \Sigma^e_\alpha = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1} \end{cases}$

In a Dually flat space, dual Pythagoras' theorem



$$D(P:R) = D(P:Q) + D(Q:R)$$



$$D^*(P:R) = D^*(P:Q) + D^*(Q:R)$$

Bregman manifold
induced by a convex function

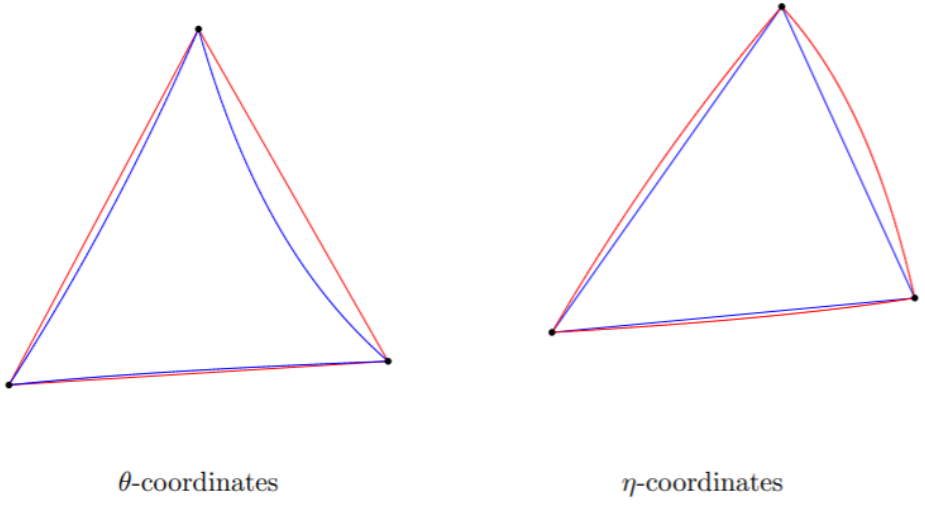
Two (affine) coordinate systems coupled by **Legendre-Fenchel transformation**

Two dually flat connections with respect to the metric tensor

Canonical distance = Bregman divergence induced by convex generator F

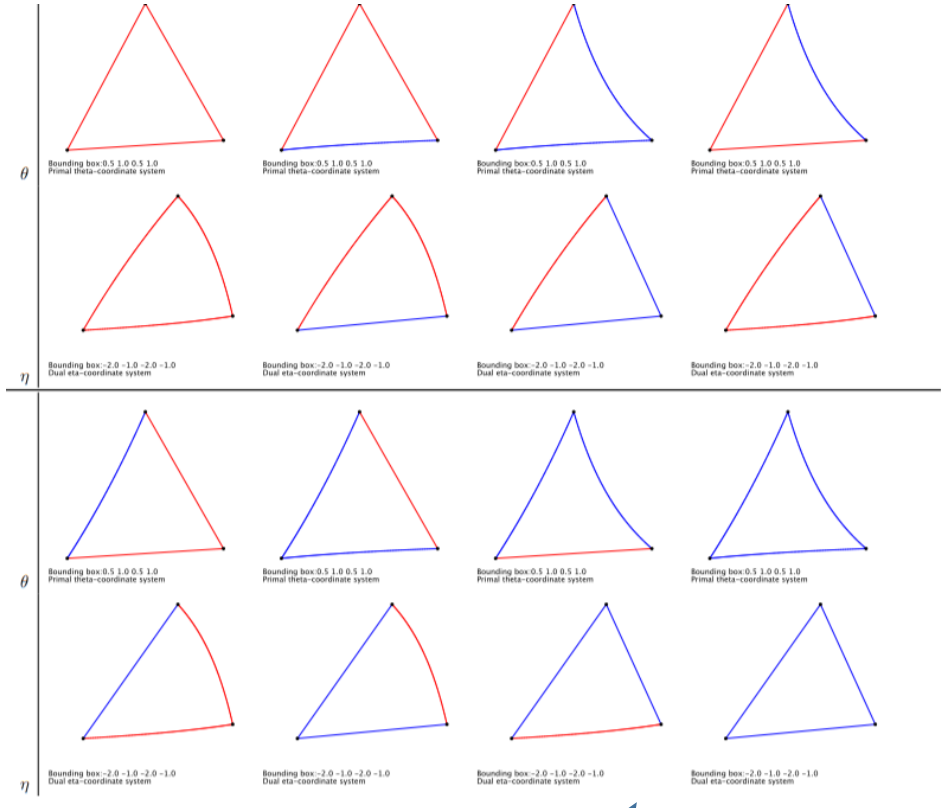
Generalize Euclidean space, ☺ very practical ☺ for computing!

Geodesic triangles in Bregman manifolds

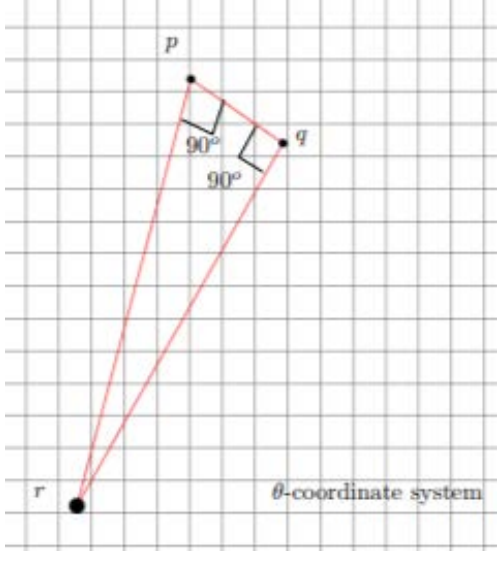


θ-coordinates

η-coordinates



3 vertices define 6 geodesic edges from which
8 geodesic triangles can be built, defining 18 interior angles



Geodesic triangle
with two right angles
(geometry is NOT CONFORMAL)

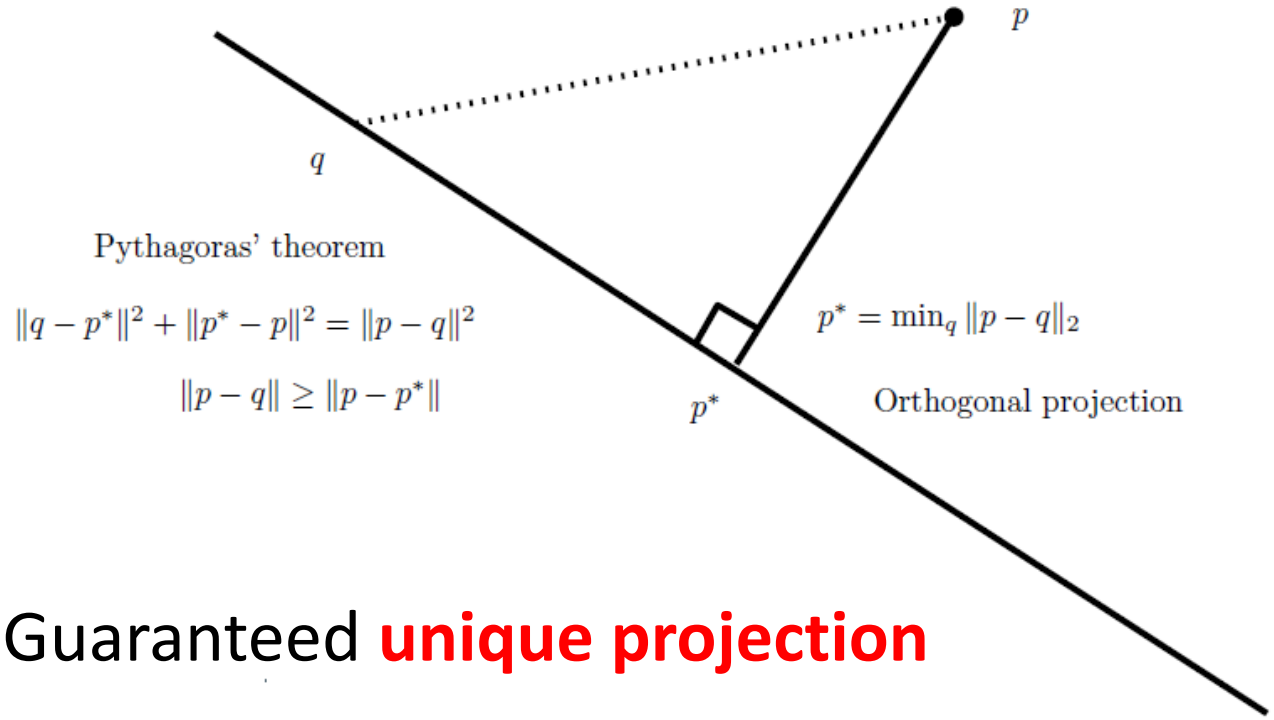
Geometry induced by dual convex potentials

https://arxiv.org/abs/1910.03935

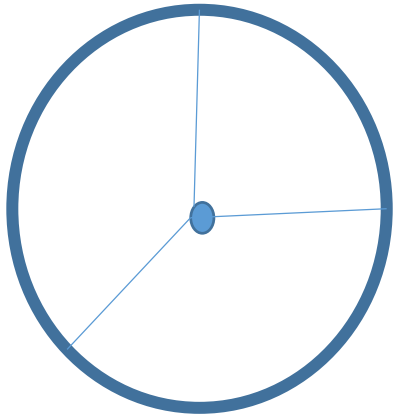
Recalling projections in Euclidean geometry….

Orthogonality and uniqueness of projection

Proof using the Pythagoras' theorem, min. distance



Guaranteed **unique projection**



Non-unique projection

On uniqueness of projections in dually flat spaces

Projection to a submanifold with respect to a connection

$$\mathrm{proj}_S^{\nabla}(p) = \{q \in S \ : \quad \gamma_{pq} \perp_q S\}.$$

In dually flat spaces, uniqueness of projections when

Theorem (Uniqueness of projections) *The* ∇-projection P_S *of* P *on* S *is* unique *if* S *is* ∇^*-flat *and minimizes the divergence* $D(\theta(P) : \theta(Q))$*:*

$$\nabla\text{-projection:} \quad P_S = \arg\min_{Q \in S} D(\theta(P) : \theta(Q)).$$

The dual ∇^**-projection* P_S^* *is* unique *if* $M \subseteq S$ *is* ∇-flat *and minimizes the divergence* $D(\theta(Q) : \theta(P))$*:*

$$\nabla^*\text{-projection:} \quad P_S^* = \arg\min_{Q \in S} D(\theta(Q) : \theta(P)).$$

Maximum likelihood estimator for an exponential family as an information m-projection

$$\max_{\theta\in\Theta}\prod_{i=1}^{n} p_\theta(x_i),$$

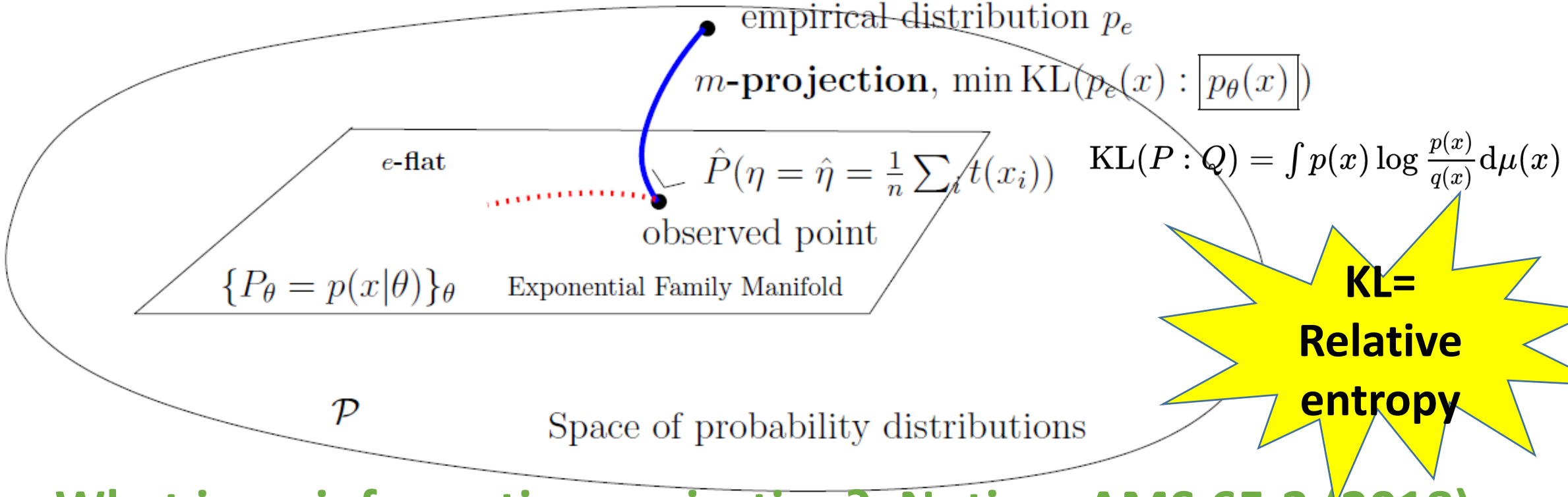
$$\max_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)\log p_\theta(x),$$

$$\equiv \min_{\theta\in\Theta} -E_{p_e}[\log p_\theta(x)] - H(p_e),$$

$$\equiv \min_{\theta\in\Theta} \mathrm{KL}(p_e : p_\theta),$$

Exponential Family Manifold (EFM) is **e-flat**

Observed point



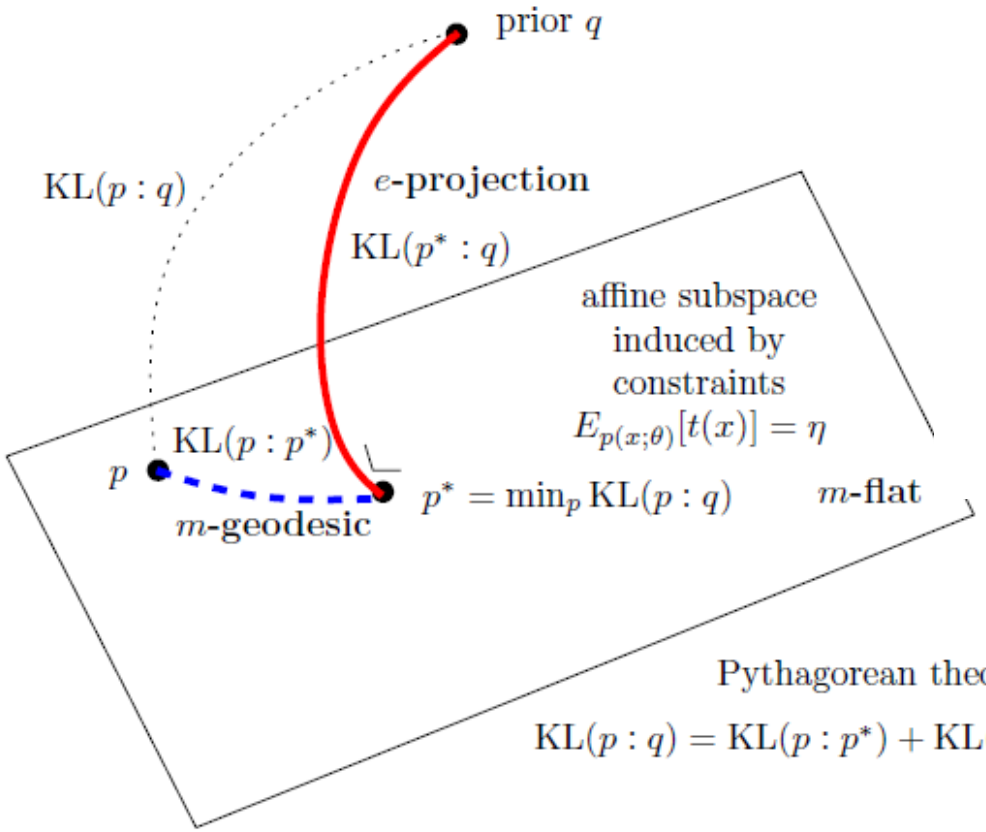
$$\mathrm{KL}(P : Q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$$

KL= Relative entropy

What is an information projection?, Notices AMS 65.3 (2018)

MaxEnt as an information e-projection

- MaxEnt linear constraints define a **m-flat**



$$\max_p H(p) = \min_p \mathrm{KL}(p:u),$$

$$\int p(x) t_i(x) \mathrm{d}\mu(x) = m_i, \forall i \in \{1, \ldots, D\},$$

$$\int p(x) \mathrm{d}\mu(x) = 1,$$

Pythagoras' theorem (Fisher orthogonality) $\gamma_m\left(p, p^*\right) \perp_{\mathrm{FIM}} \gamma_e\left(p^*, q\right)$

What is an information projection?, Notices AMS 65.3 (2018)

Geometric framework yields interpretations of MLE/MaxEnt of KL divergence minimizations as information projections (and uniqueness proofs)

MaxEnt (with prior q)

e-projection on m-flat

$$\min_p \mathrm{KL}(p : q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\sum_x p(x) t_i(x) = m_i, \quad \forall i \in \{1, \ldots, D\}$$

$$p(x) \geq 0, \quad \forall x \in \{1, \ldots, n\}$$

$$\sum_x p(x) = 1$$

Maximum Likelihood Estimate

m-projection on e-flat

$$\begin{aligned}
\min \quad & \mathrm{KL}(p_e(x) : p_\theta(x)) \\
&= \int p_e(x) \log p_e(x) \mathrm{d}x - \int p_e(x) \log p_\theta(x) \mathrm{d}x \\
&= \min -H(p_e) - \underbrace{E_{p_e}[\log p_\theta(x)]} \\
&\equiv \max \frac{1}{n} \sum \delta(x - x_i) \log p_\theta(x) \\
&= \max \frac{1}{n} \sum_i \log p_\theta(x_i) = \mathrm{MLE}
\end{aligned}$$

Divergences: Statistical (oriented) distances or smooth parametric distances

- In information theory, **relative entropy** called Kullback-Leibler divergence

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- KLD can be extended to **f-divergences**

$$D_f(P \parallel Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x).$$

- Plug f(u)=-log(u) and the f-divergence is the KLD

Classes of distances: Csiszar's f-divergence

- Function f convex, **strictly convex at 1**, with f(1)=0

$$I_f(p:q) = \int pf\left(\frac{q}{p}\right) d\mu \geq f(1)$$

Name of the f-divergence	Formula $I_f(P:Q)$	Generator $f(u)$ with $f(1)=0$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int (\sqrt{p(x)}-\sqrt{q(x)})^2 \mathrm{d}\nu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-\lambda p(x))^k}{p^{k-1}(x)} \mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-\lambda p(x)\rvert^k}{p^{k-1}(x)} \mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)} \mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)} \mathrm{d}\nu(x)$	$u\log u$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x)q^{1+\alpha}(x)\mathrm{d}\nu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int (p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

On the chi square and higher-order chi distances for approximating f-divergences, IEEE SPL 2013

The Fisher information matrix and f-divergences

- *Fisher Information Metric* (FIM): Fisher Riemannian metric

$$g_{jk}(\theta) = \int_X \frac{\partial \log p(x,\theta)}{\partial \theta_j} \frac{\partial \log p(x,\theta)}{\partial \theta_k} p(x,\theta)\, dx.$$

- Infinitesimally, the Kullback-Leibler or f-div related to the FIM

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

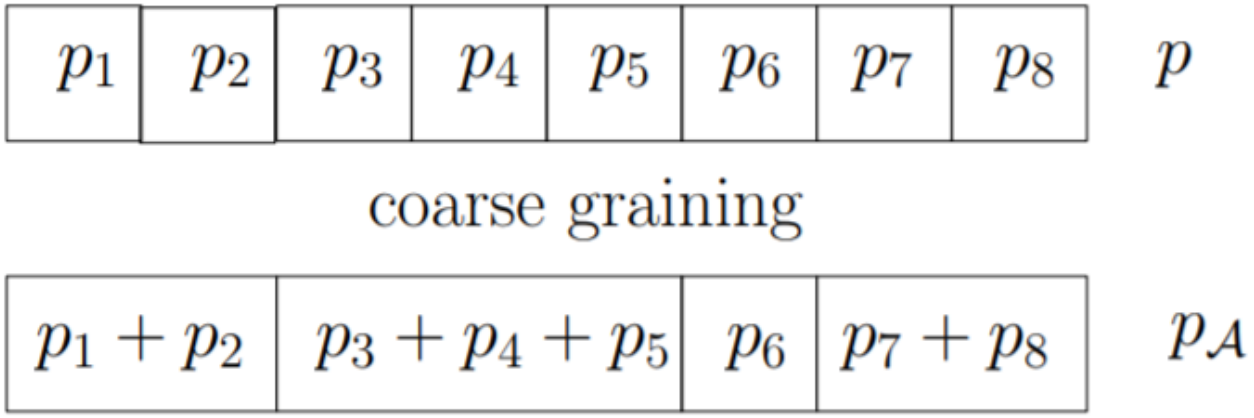
$$D_{\mathrm{KL}}[P(\theta_0) \| P(\theta)] = \frac{1}{2} \sum_{jk} \Delta\theta^j \Delta\theta^k g_{jk}(\theta_0) + \mathrm{O}(\Delta\theta^3).$$

Invariant divergence = f-divergences

- Lump or coarse-bin a separable distance, and ask for

information monotonicity

$$D(\theta_{\bar{\mathcal{A}}} : \theta'_{\bar{\mathcal{A}}}) \leq D(\theta : \theta')$$



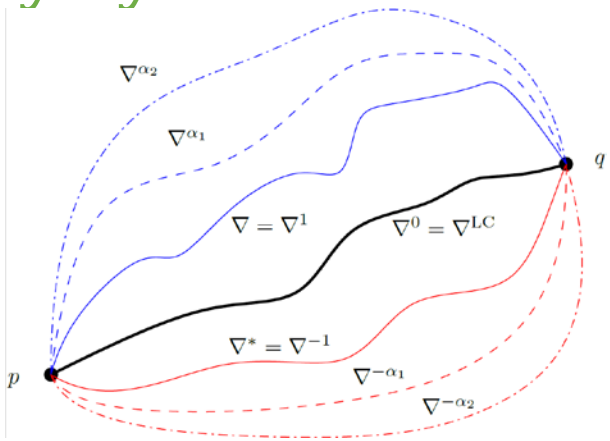
Theorem: The only monotone *separable* divergences are f-divergences (except for the curious case of binary alphabets),
f-divergences are invariant by diffeormorphisms of the sample space

$$\begin{aligned} D_f(q_i, q_j) &= \int_{\mathcal{Y}} q_j(y) f\left(\frac{q_i(y)}{q_j(y)}\right) dy \\ &= \int_{\mathcal{X}} p_j(x)|\mathcal{J}(x)|^{-1} f\left(\frac{p_i(x)|\mathcal{J}(x)|^{-1}}{p_j(x)|\mathcal{J}(x)|^{-1}}\right) |\mathcal{J}(x)|dx \\ &= \int_{\mathcal{X}} p_j(x) f\left(\frac{p_i(x)}{p_j(x)}\right) dx = D_f(p_i, p_j). \end{aligned}$$

Dual connections from *any* divergence! $(M, {}_D g, {}_D\nabla, {}_D\nabla^*)$

Dual connections from any smooth parametric distance,
called a (parameter) divergence D: D is not necessarily symmetric

- a tensor metric g: $g_{ij}(p_\xi)=\frac{\partial}{\partial \xi^i_1}\frac{\partial}{\partial \xi^j_2} D(p_{\xi_1},p_{\xi_2})|_{\xi_1=\xi_2=\xi}$



- a torsion-less affine connection ∇:

$$\Gamma_{ijk}(p_\xi)= -\frac{\partial}{\partial \xi^i_1}\frac{\partial}{\partial \xi^j_2}\frac{\partial}{\partial \xi^k_2} D(p_{\xi_1},p_{\xi_2})|_{\xi_1=\xi_2=\xi}$$

Dual divergences
and dual connections

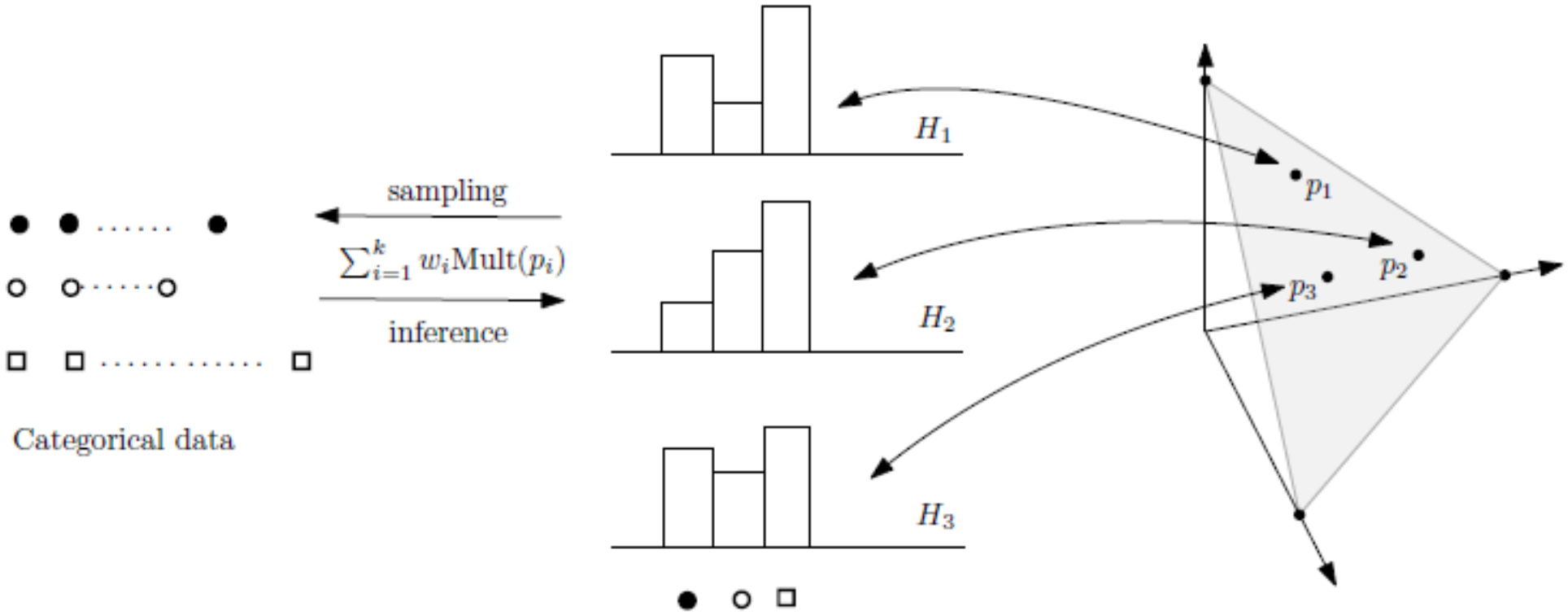
$$D^*(p_{\xi_1},p_{\xi_2})= D(p_{\xi_2},p_{\xi_1})$$

Symmetric divergences yields the same connection:

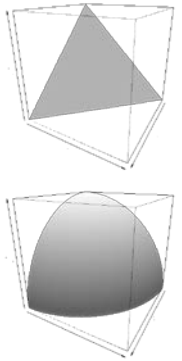
The Levi-Civita connection

Which geometry is best suited for clustering normalized histograms?

Bag of words

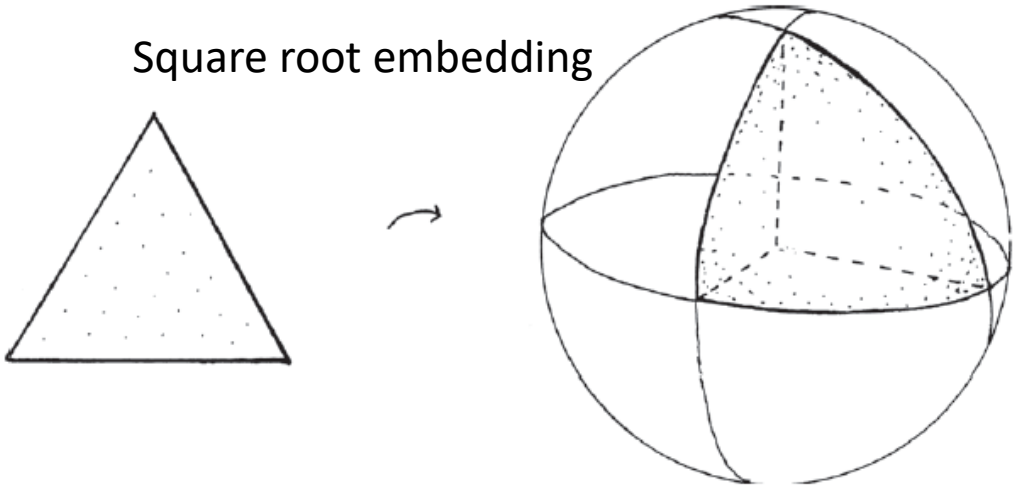


Fisher-Rao geometry of the categorical distribution (standard simplex)



- Trinomial (trinoulli)

Square root embedding



Embedding to the sphere positive orthant

Fisher information metric:

$$g_{ij}(p) = \frac{\delta_{ij}}{\lambda_p^i} + \frac{1}{\lambda_p^0}.$$

(Hotelling)-Fisher-Rao distance:

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

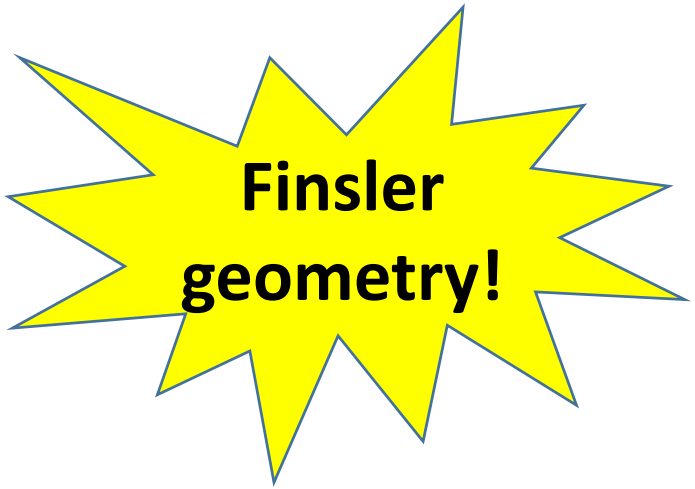
Pattern Learning and Recognition on Statistical Manifolds: An Information-Geometric Review, SIMBAD 2013

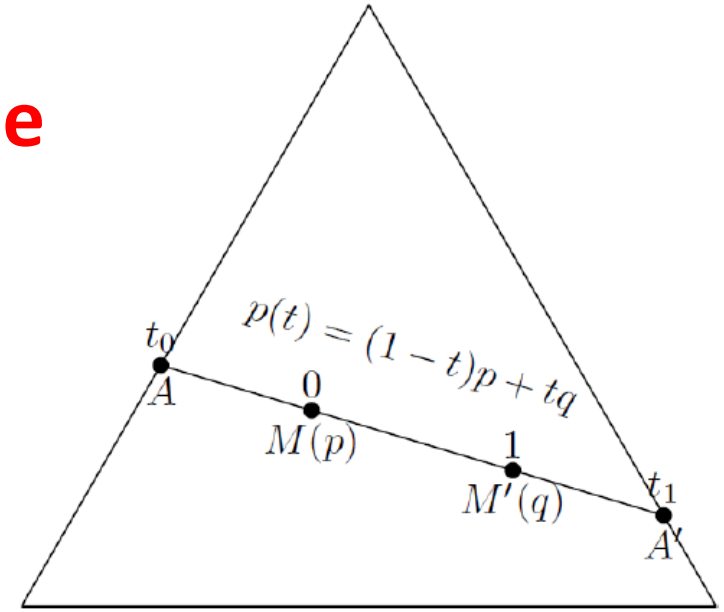
Clustering in Hilbert simplex geometry, arXiv:1704.00454

Hilbert log cross-ratio metric

$$\rho_{\mathrm{HG}}(M,M') = \begin{cases} \left|\log \frac{|A'M||AM'|}{|A'M'||AM|}\right|, & M \neq M', \\ 0 & M = M'. \end{cases}$$

Geodesics are straight lines but not unique





Clustering in Hilbert simplex geometry. CoRR abs/1704.00454 (2017)

Hilbert log cross-ratio metric for the standard simplex

Isometry of Hilbert simplex geometry with a normed vector space $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$

- $V^d = \{v \in \mathbb{R}^{d+1} \ : \ \sum_i v^i = 0\} \subset \mathbb{R}^{d+1}$
- Map $p = (\lambda^0, \ldots, \lambda^d) \in \Delta^d$ to $v(x) = (v^0, \ldots, v^d) \in V^d$:

$$v^i = \frac{1}{d+1}\left(d \log \lambda^i - \sum_{j \neq i} \log \lambda^j\right) = \log \lambda^i - \frac{1}{d+1}\sum_j \log \lambda^j.$$

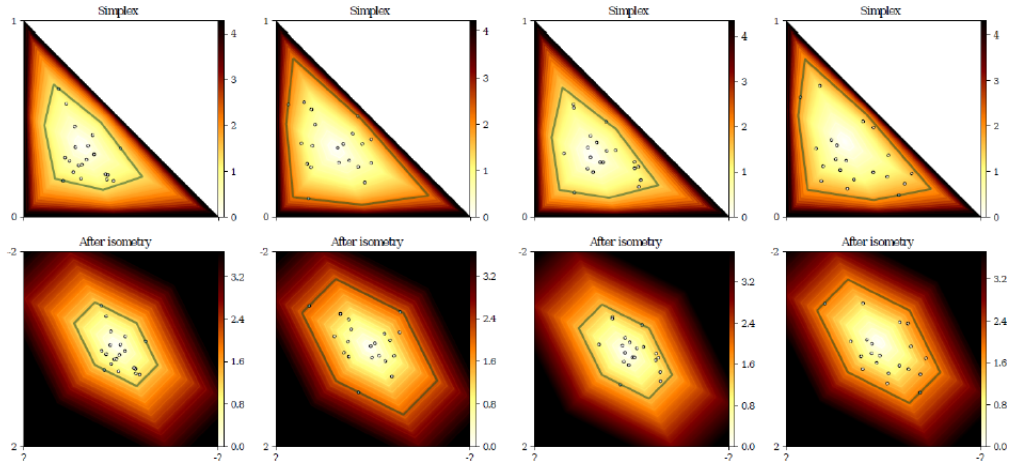
$$\lambda^i = \frac{\exp(v^i)}{\sum_j \exp(v^j)}.$$

- Norm $\|\cdot\|_{\mathrm{NH}}$ in V^d defined by the shape of its unit ball $B_V = \{v \in V^d \ : \ |v^i - v^j| \leq 1, \forall i \neq j\}$.
- Polytopal norm-induced distance:

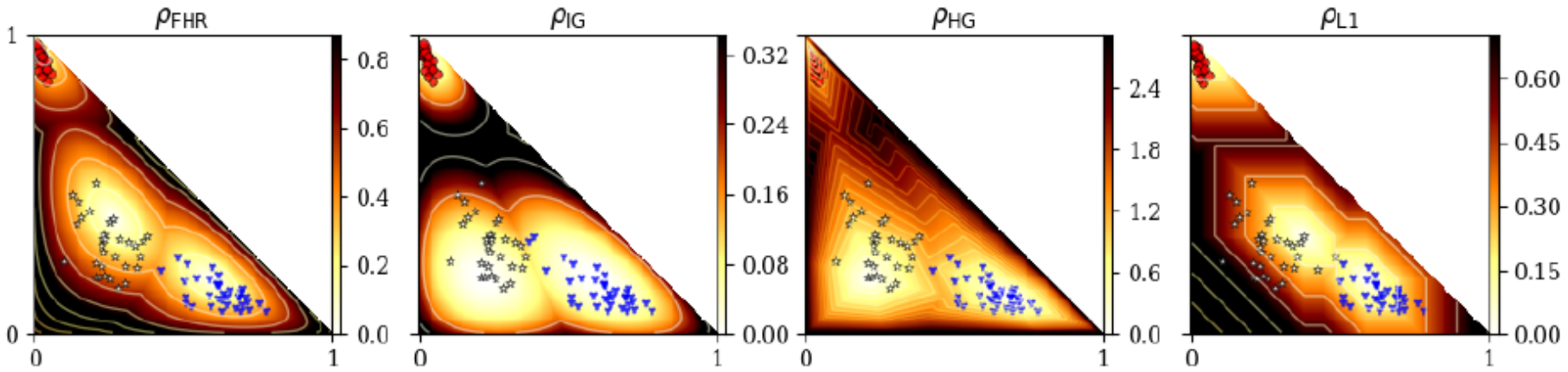
$$\rho_V(v, v') = \|v - v'\|_{\mathrm{NH}} = \inf\left\{\tau \ : \ v' \in \tau(B_V \oplus \{v\})\right\},$$

- Norm does not satisfy parallelogram law (no inner product)

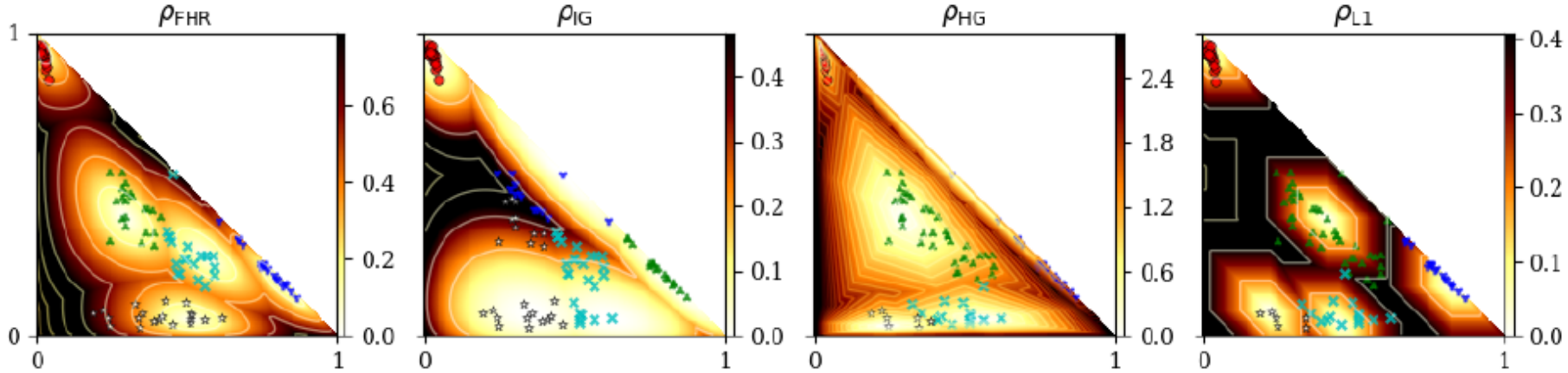
Visualizing the isometry: $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$



Experiments
With
K-means

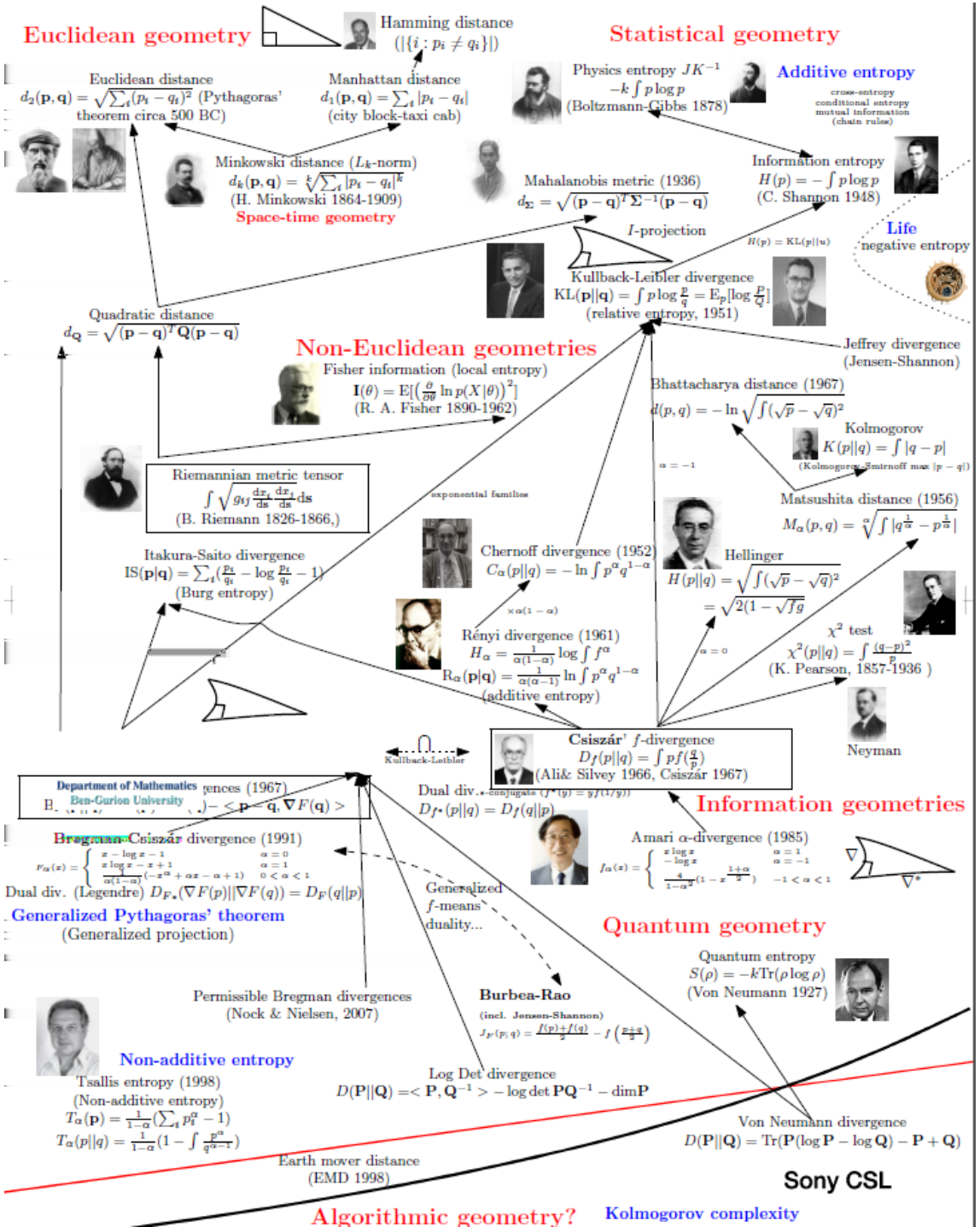


$k = 3$ clusters



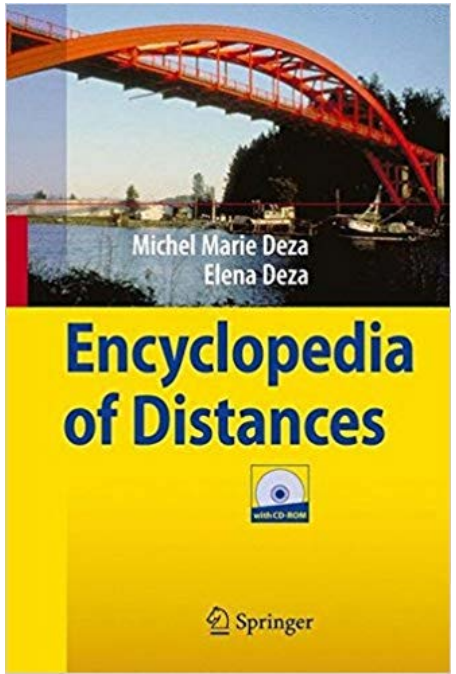
$k = 5$ clusters

Clustering in Hilbert simplex geometry. CoRR abs/1704.00454 (2017)

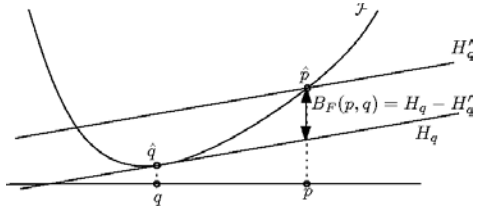


Sailing on a sea of distances:

- Which distance is suitable?
- Which loss function to minimize?
- Which "metric" to evaluate?
- …
- Are there **first principles**?



Classes of distances: Bregman divergence

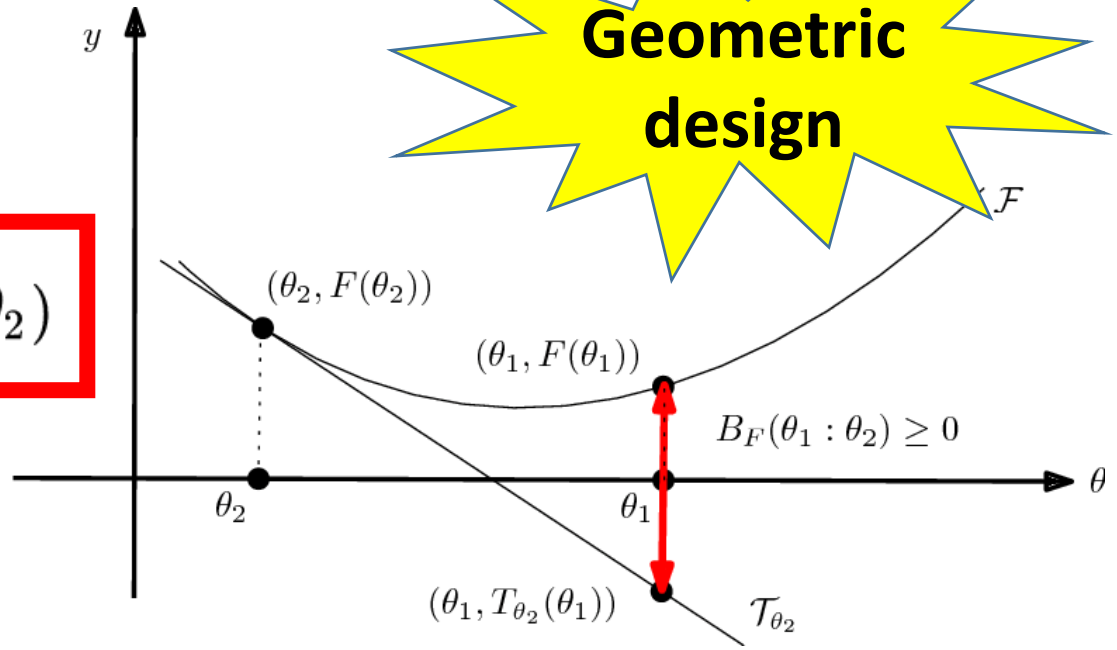


- **Bregman divergence** between parameters for a strictly convex and differentiable convex function F

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)$$

Unify squared Euclidean geometry and geometry of information theory

Geometric design



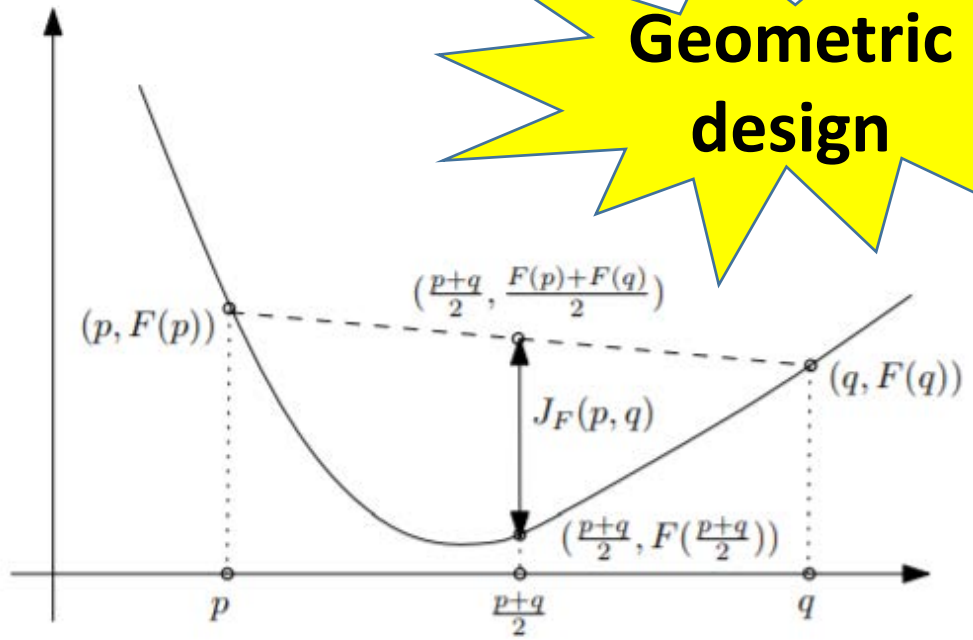
- The canonical divergence of dually flat spaces

Mining matrix data with Bregman matrix divergences for portfolio selection."*Matrix Information Geometry*. Springer, Berlin, Heidelberg, 2013. 373-402.

Jensen difference/Jensen divergence (Burbea-Rao)

- Introduced by Burbea and Rao
- Vertical gap induced by Jensen inequality

$$J_F(\theta_1,\theta_2) = \frac{F(\theta_1)+F(\theta_2)}{2} - F\left(\frac{\theta_1+\theta_2}{2}\right) \geq 0$$



Asymptotic scaled Jensen divergence amount to a Bregman or reverse Bregman divergence

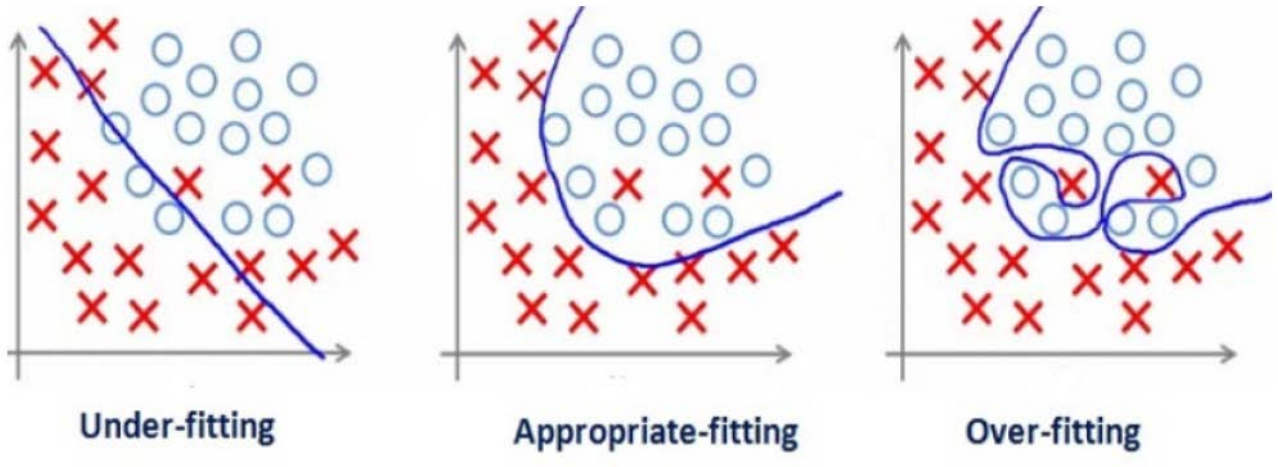
$$J_\alpha^F(\theta_1 : \theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J'^F(\theta_1 : \theta_2) & \alpha \neq \{0,1\} \\ B_F(\theta_1 : \theta_2) & \alpha = 1 \\ B_F(\theta_2 : \theta_1) & \alpha = 0 \end{cases}$$

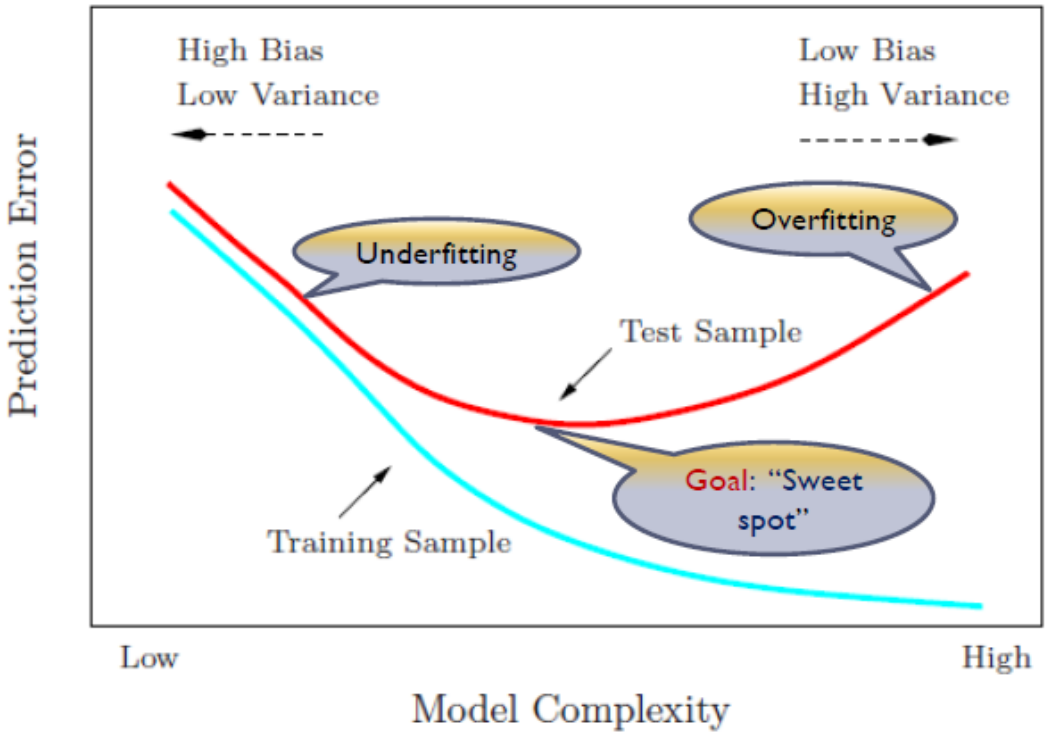
The Burbea-Rao and Bhattacharyya centroids." *IEEE Transactions on Information Theory* 57.8 (2011): 5455-5466.

Bregman chord divergence: https://arxiv.org/abs/1810.09113

A family of statistical symmetric divergences based on Jensen's inequality, arXiv:1009.4004

Classical wisdom of machine learning: The bias-variance tradeoff…



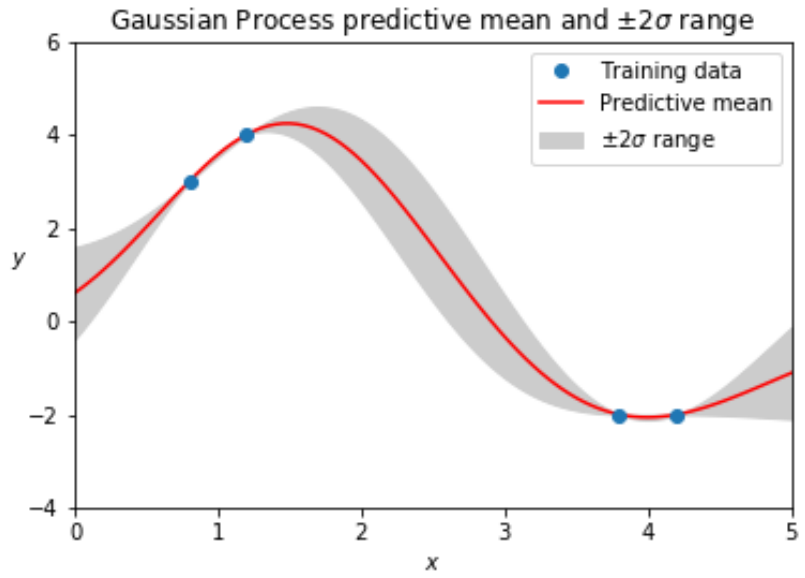


We used to think: Do not overfit for better generalization!

Test sample for evaluating generalization error

Modern view of machine learning: The age of Interpolation machines!

- Ok, let us do zero-training error! (Gaussian processes or Neural Networks)



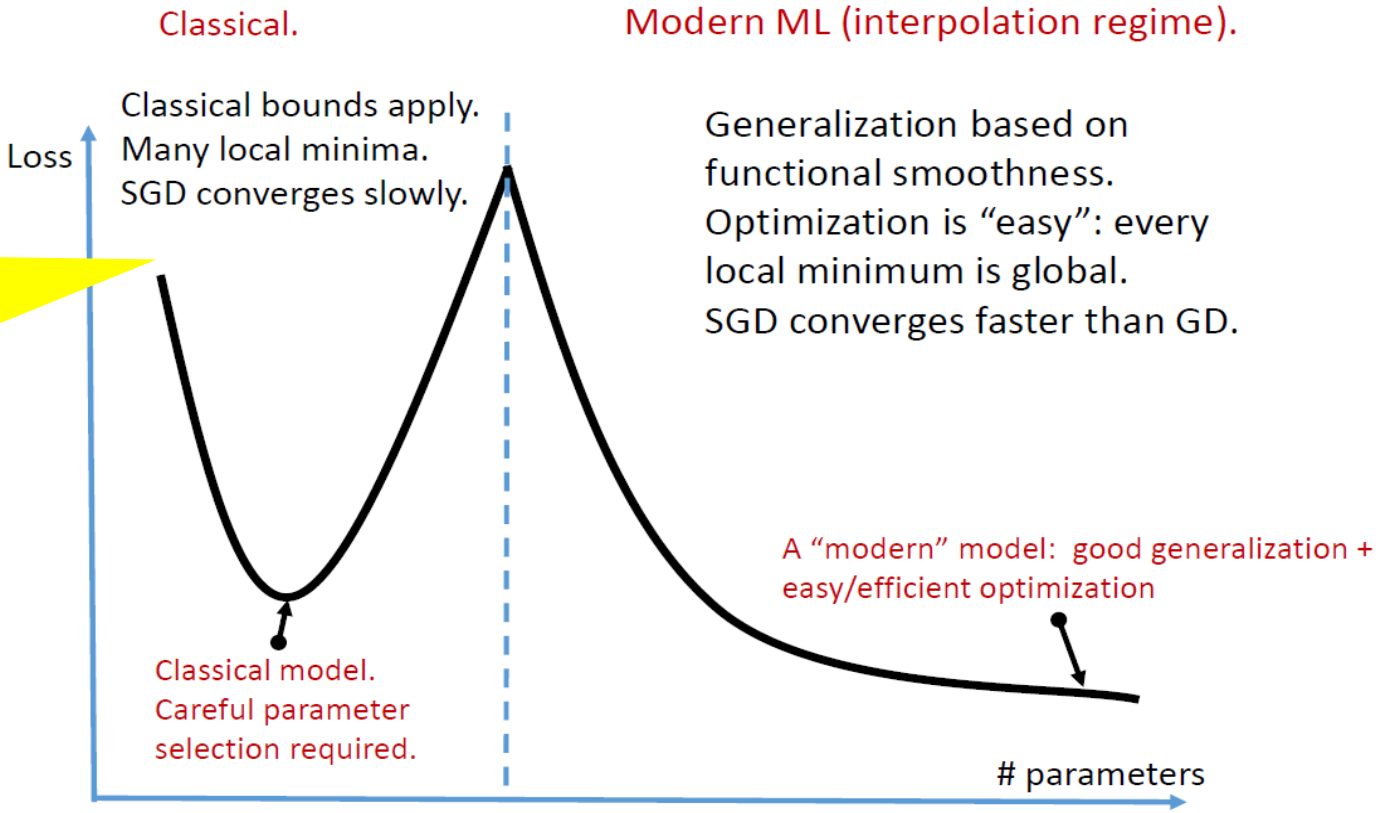
Neural networks perform well even when overparameterizated

- But to have good models, let us set Occam's razor principle to choose the **smoothest interpolating function**

Deep Neural Networks as Gaussian Processes, https://arxiv.org/abs/1711.00165

Modern view of machine learning:
The double descent view/regime of models

Semi-Riemannian geometry of neuromanifold and learning trajectories



Reconc ing modern machine learning practice and the bias–variance trade–off, https://arxiv.org/abs/1812.11118

Lightlike Neuromanifolds, Occam's Razor and Deep Learning, 1905.11027

Concluding remarks

Computational information geometry for ML+AI!

- From the very beginning, **computational geometry** played a major role in machine learning!

- Geometry: Design guiding principle promoting insightful intuition and science of **invariance.** Meaning of **distances.**

- Dualistic structure of information geometry + **information projection**: Role of **Fisher information matrix/metric** in ML (Fisher kernel, etc.)

Theory of communication between data/(sub)models and models

(but may seem at first counterintuitive)

Geometric Science of Information (GSI)

Co-organized every 2 years since 2013

180 participants, August, Toulouse, France, 2019

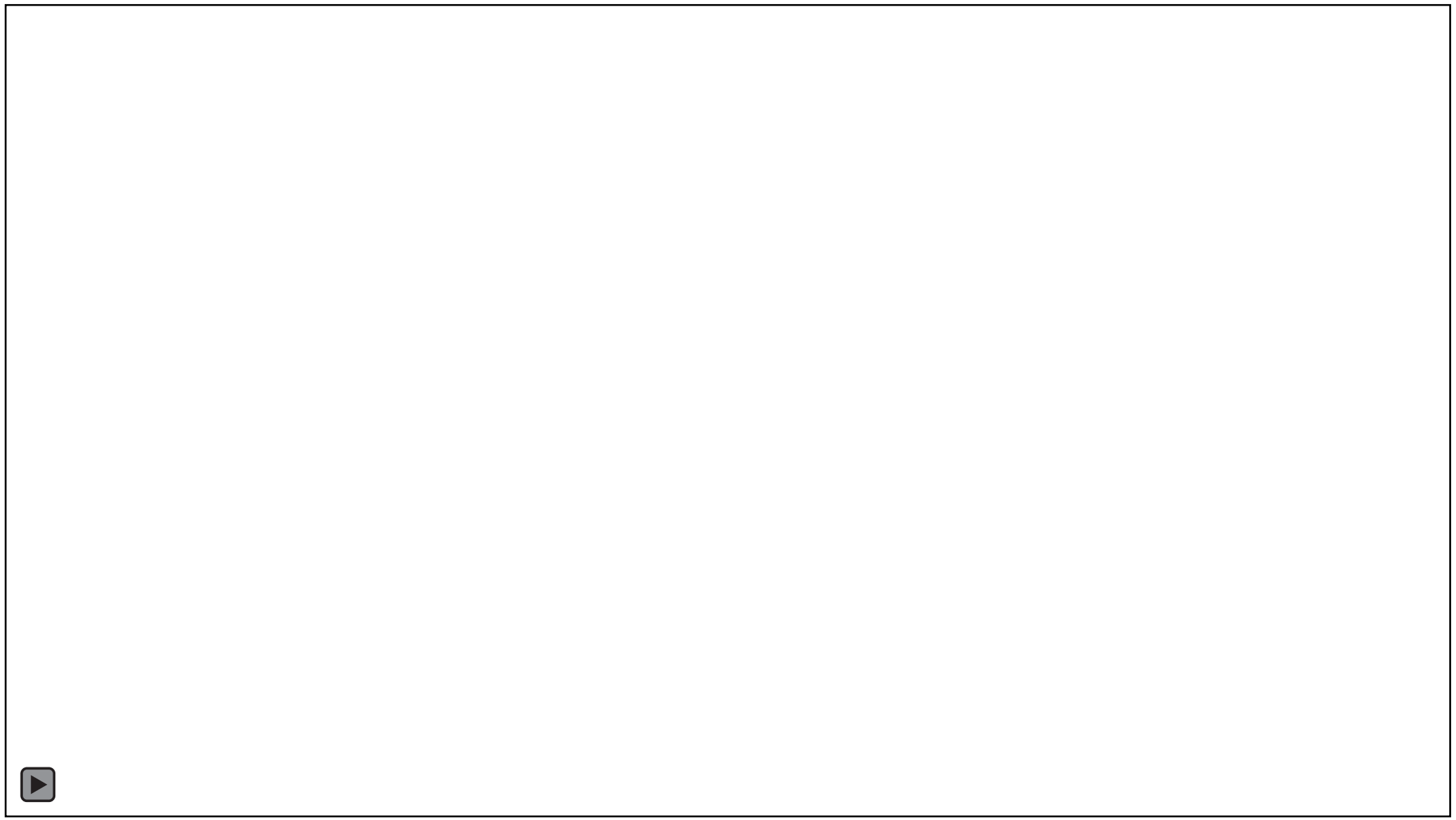
Joint Structures and Common Foundations of Statistical Physics, Information Geometry and Inference for Learning

26th July to 31st July 2020

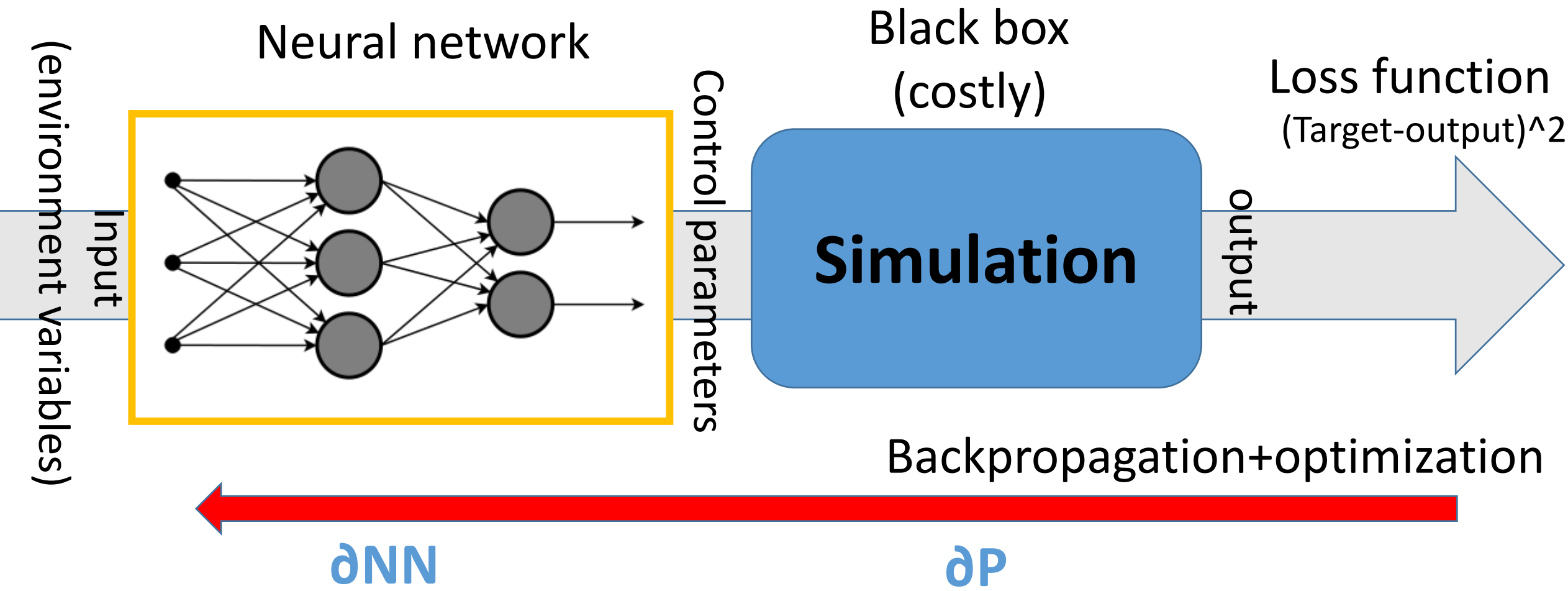
Ecole de Physique des Houches

Thank you!

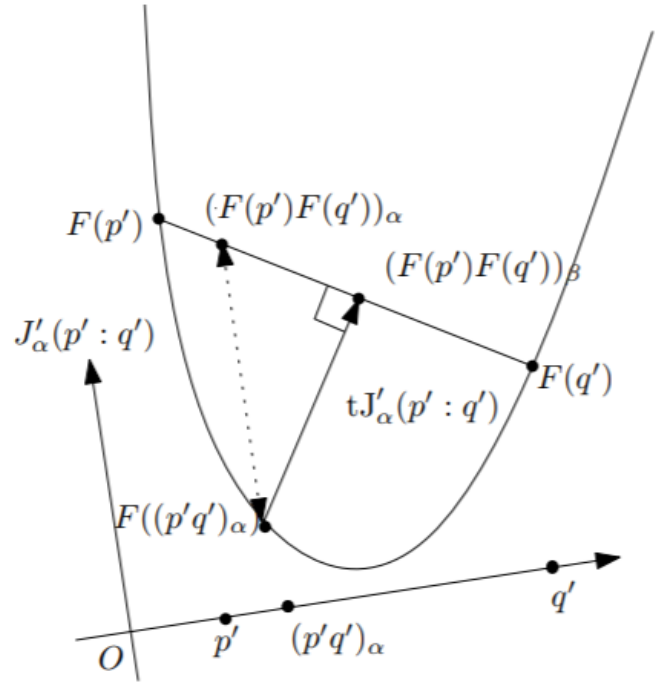
video



Machine learning changing the scientific computing: Neural Network (NN)+Differentiable programming (∂P)



Total Jensen divergence



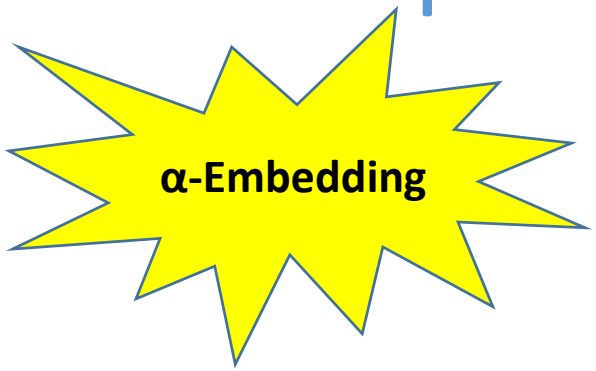
Invariant to axis rotation

$$\mathrm{tB}(p:q) = \rho_B(q)B(p:q), \quad \rho_B(q) = \sqrt{\frac{1}{1+\langle \nabla F(q), \nabla F(q)\rangle}}$$

$$\mathrm{tJ}_\alpha(p:q) = \rho_J(p,q)J_\alpha(p:q), \quad \rho_J(p,q) = \sqrt{\frac{1}{1+\frac{(F(p)-F(q))^2}{\langle p-q,p-q\rangle}}}$$

Total Jensen divergences: Definition, properties and clustering. ICASSP 2015

α-representations of the FIM



$$I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)}(x;\theta)\partial_j l^{(-\alpha)}(x;\theta)d\nu(x)$$

- 0-representation (square root) : $I'_{ij}(\theta) := 4\int \partial_i\sqrt{p(x;\theta)}\partial_j\sqrt{p(x;\theta)}\mathrm{d}\nu(x)$
- 1-representation (log): $I_{ij}(\theta) := E_{p(x;\theta)}[\partial_i l(x;\theta)\partial_j l(x;\theta)]$
- Under mild regularity conditions:

$$I_{ij}^{(\alpha)}(\theta) = -\frac{2}{1+\alpha}\int p(x;\theta)^{\frac{1+\alpha}{2}}\partial_i\partial_j l^{(\alpha)}(x;\theta)d\nu(x)$$

- Coefficients of the connection: $\Gamma_{ij,k}^{(\alpha)} = \int \partial_i\partial_j l^{(\alpha)}\partial_k l^{(-\alpha)}d\nu(x)$

The α-representations of the Fisher Information Matrix, 2017

(ρ, τ)-representations of the FIM

Smooth convex function and convex conjugates: $f^*(t) = t(f')^{-1}(t) - f\left((f')^{-1}(t)\right)$

τ -representation $\tau(p) = f'(\rho(p)) = \left((f^*)'\right)^{-1}(\rho(p))$

ρ-representation $\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p))$

(ρ, τ)-FIM $g_{ij}(\theta) = E_\mu \left\{ f'' \left(\rho(p(\zeta|\theta)) \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \right\} \right.$

(ρ, τ)-α-connections $\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} f'''(\rho(p(\zeta|\theta)))A_{ijk} + f''(\rho(p(\zeta|\theta)))B_{ijk} \right\}$

$$A_{ijk}(\zeta,\theta) = \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k}, \quad B_{ijk}(\zeta,\theta) = \frac{\partial^2 \rho(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k}$$

Zhang, Jun. "On monotone embedding in information geometry." *Entropy* 17.7 (2015

Standard invariant f-divergences

- f strictly convex at 1 (for ensuring the law of the indiscernibles)
- Choose f(1)=0 (for lower bound of f-divergence being 0)
- Choose f'(1)=0 to fix lambda in equivalent class of generators:

$$f_\lambda(u) = f(u) + \lambda(u-1)$$

- Expansion of

$$I_f(p:p+dp) = f''(1)\frac{1}{2}dp^\top g(p)dp$$

- Choose f''(1)=1 to get **standard f-divergence** with infinitesimal distance expressed using the Fisher information matrix tensor
- **The α-connection for any standard f-divergence corresponds to the expected α-connections for**

$$\alpha = 2f'''(1) + 3$$

M-statistical mixture

$$(pq)_\alpha^M(x) := \frac{M_\alpha(p(x), q(x))}{Z_\alpha^M(p:q)}$$

Need to normalize M-mixtures

$$Z_\alpha^M(p:q) = \int_{t\in\mathcal{X}} M_\alpha(p(t), q(t))\mathrm{d}\mu(t)$$

$$(p_1 \ldots p_k)_\alpha^M := \frac{p_1(x)^{\alpha_1} \times \ldots \times p_k(x)^{\alpha_k}}{Z_\alpha(p_1, \ldots, p_k)}$$

$$\mathrm{JS}_D^{M_\alpha}(p:q) := (1-\alpha)D\left(p:(pq)_\alpha^M\right) + \alpha D\left(q:(pq)_\alpha^M\right)$$

$$\mathrm{JS}^{M_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p:(pq)_\alpha^M\right) + \alpha\mathrm{KL}\left(q:(pq)_\alpha^M\right)$$

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means, Entropy 2019

Case study of multivariate Gaussians

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) = \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta_\mu^\top \Sigma_2^{-1}\Delta_\mu + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\right\}$$

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}) + \alpha\mathrm{KL}(p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}),\\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2),\\
&= \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \right.\\
&\quad \left.(1-\alpha)(\mu_\alpha-\mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_1) + \alpha(\mu_\alpha-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_2) - d\right)\\
\mathrm{JS}^{G_\alpha}_*(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}),\\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha),\\
&= J_F(\theta_1 : \theta_2),\\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right)
\end{aligned}$$

$$\Sigma_\alpha = (\Sigma_1\Sigma_2)^\Sigma_\alpha = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1} \qquad \mu_\alpha = (\mu_1\mu_2)^\mu_\alpha = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right)$$

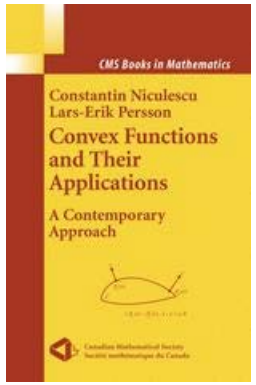
New Bregman divergences from abstract means

A function is **(M,N)-convex** (comparative convexity) if and only if

$$F(M(p,q)) \leq N(F(p),F(q)), \quad \forall p,q \in \mathcal{X}$$

A mean is **regular** if it is:

1. homogeneous
2. symmetric,
3. continuous
4. increasing in each variable.



Skewed (M,N)-Jensen-divergence for regular means:

$$J_F^{M,N}(p,q) = N(F(p),F(q))) - F(M(p,q))$$

$$J_{F,\alpha}^{M,N}(p:q) \geq 0$$

Example of non-regular means: Lehmer mean (also Bajraktarevic mean) $L_\delta(x_1,\ldots,x_n;w_1,\ldots,w_n) = \frac{\sum_{i=1}^n w_i x_i^{\delta+1}}{\sum_{i=1}^n w_i x_i^\delta}$

Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity, IEEE SPL 2017

(M,N)-Bregman divergences from comparative convexity

(M,N) Bregman divergences obtained in the scaled limit case of Jensen divergence:

$$B_F^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} \left(N_\alpha(F(p),F(q))) - F(M_\alpha(p,q))\right)$$

Quasi-arithmetic Bregman divergences obtained

$$M_f(p,q) = f^{-1}\left(\frac{f(p)+f(q)}{2}\right)$$

$$B_F^{\rho,\tau}(p:q) = \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p)-\rho(q)}{\rho'(q)} F'(q).$$

$$B_F^{\rho,\tau}(p:q) = \kappa_\tau(F(q):F(p)) - \kappa_\rho(q:p)F'(q)$$

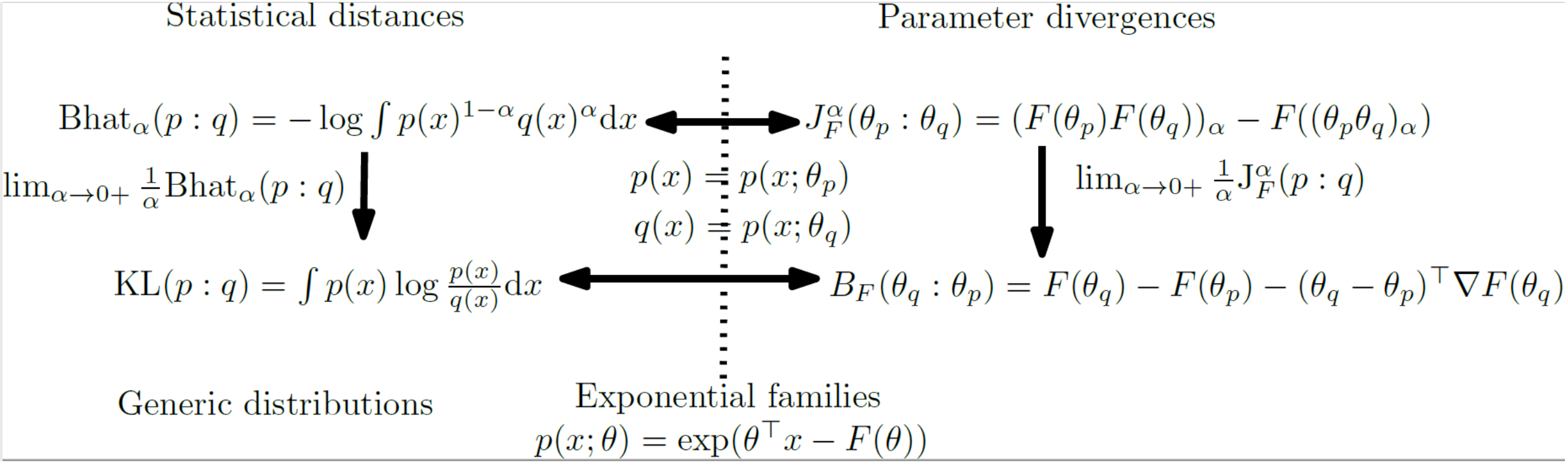
Type	γ	$\kappa_\gamma(x:y) = \frac{\gamma(y)-\gamma(x)}{\gamma'(x)}$
A	$\gamma(x) = x$	$y - x$
G	$\gamma(x) = \log x$	$x \log \frac{y}{x}$
H	$\gamma(x) = \frac{1}{x}$	$x^2\left(\frac{1}{y} - \frac{1}{x}\right)$
$P_\delta, \delta \neq 0$	$\gamma_\delta(x) = x^\delta$	$\frac{y^\delta - x^\delta}{\delta x^{\delta-1}}$

For example, the **power mean Bregman divergences**:

$$B_F^{\delta_1,\delta_2}(p:q) = \frac{F^{\delta_2}(p) - F^{\delta_2}(q)}{\delta_2 F^{\delta_2-1}(q)} - \frac{p^{\delta_1} - q^{\delta_1}}{\delta_1 q^{\delta_1-1}} F'(q)$$

Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity, IEEE SPL 2017

Statistical divergences amount to parameter divergences for exponential families:



The Burbea-Bao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57(8), 2011

Symmetrizing the KL divergence

Jeffreys divergence:

$$J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p-q) \log \frac{p}{q} \mathrm{d}\mu = J(q;p).$$

Resistor average divergence:

$$\begin{aligned} \frac{1}{R(p;q)} &= \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right), \\ R(p;q) &= \frac{2\left(\mathrm{KL}(p:q) + \mathrm{KL}(q:p)\right)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)} = \frac{2J(p;q)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)}. \end{aligned}$$

Historically, Amari's expected α-geometry

- Given a parametric family of distributions, consider the Fisher information matrix and a family of connections: **α connections**
- **Exponential e-mixture connection and m-mixture connection**

$$g_{p_\xi}(\nabla^\alpha{}_{\partial_i}\partial_j(p_\xi),\partial_k(p_\xi)) = \Gamma^\alpha{}_{ijk}(p_\xi) = E_\xi[(\frac{\partial}{\partial\xi^i}\frac{\partial}{\partial\xi^j}\log(p_\xi) + \frac{1-\alpha}{2}\frac{\partial}{\partial\xi^i}\log(p_\xi)\frac{\partial}{\partial\xi^j}\log(p_\xi))\frac{\partial}{\partial\xi^k}\log(p_\xi)]$$

- No associated distance in the alpha-expected geometry

$$\text{Levi-Civita connection} : \nabla^0 = \nabla^{LC}$$

https://arxiv.org/abs/1808.08271

Expected α-geometry for a parametric model

$$\mathcal{P}:=\{p_\theta(x)\}_{\theta\in\Theta} \quad \Rightarrow \quad \{(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})\}_{\alpha\in\mathbb{R}}$$

- Use Fisher information metric (FIM)
- Define the expected α-connections:
- **Amari-Chentsov cubic tensor**

$$C_{ijk}:=E_\theta\left[\partial_i l \partial_j l \partial_k l\right]$$

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

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

$$=E_\theta\left[\left(\partial_i\partial_j l+\frac{1-\alpha}{2}\partial_i l\partial_j l\right)(\partial_k l)\right].$$

Key concept: 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 recovering/estimating the parameters.

  Informally, a statistical lossless compression scheme…

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

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

- Fisher-Neyman factorization theorem: Statistic t(x) sufficient iff. the density can be decomposed as:

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

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

Natural exponential families (NEF)

- Consider a positive measure μ
- An exponential family is a parametric family of densities that write as

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

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

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

Log-Laplace transform

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

F: Log-normalizer (also known as partition function, cumulant function, etc.)

Exponential families (from Natural EFs to EFs)

- Consider a **(sufficient) statistic** t(x)
- Consider an **additional carrier measure term** k(x)
- Consider an **inner product** between t(x) and θ

  (usual scalar/dot product)

Rk: Canonical decomposition not unique

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

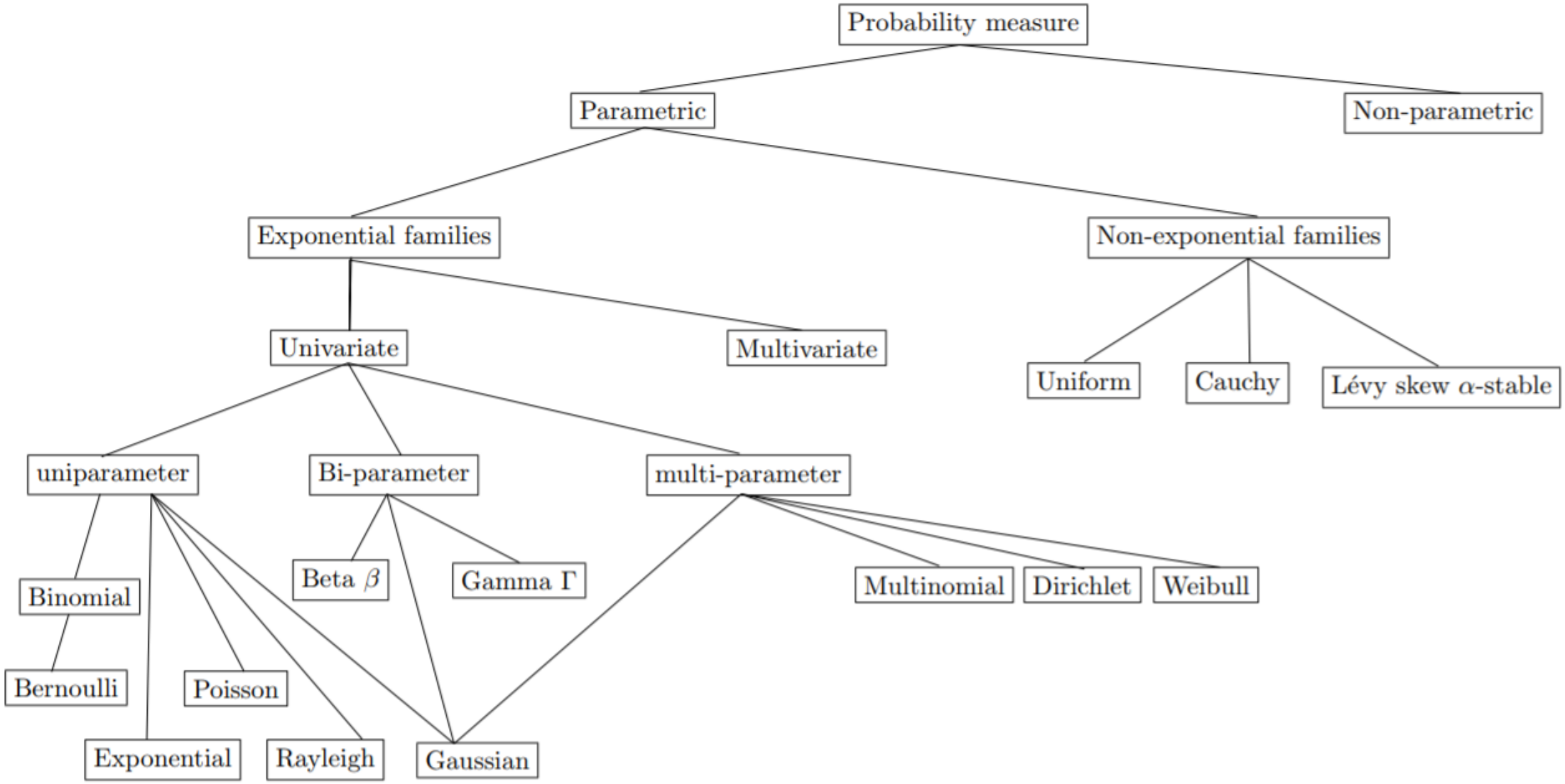
Properties:

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

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

Exponential families have finite moments of any order

Many common distributions are exponential families in disguise



In a dually flat space, natural gradient is ordinary gradient for the dual coordinates

In a dually flat space (Hessian manifold), we have

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

Natural gradient

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

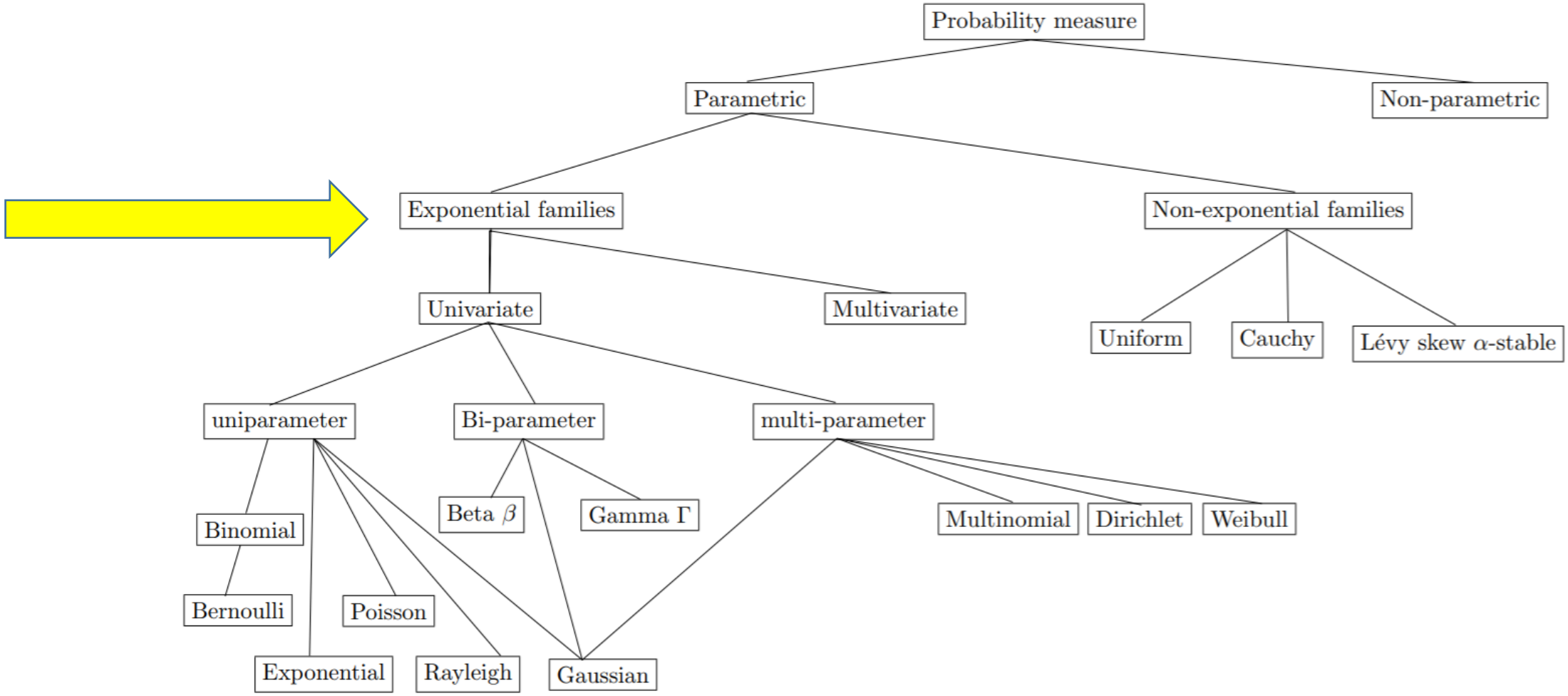
Ordinary gradient

Used in variational inference (VI)

Zhang, Guodong, et al. "Noisy natural gradient as variational inference." ***arXiv:1712.02390*** **(2017).**

Exponential families have finite dimensional sufficiency

- You can compress all n observations into **a constant dimensional** vector t(X) without loosing any statistical information for likelihood inference



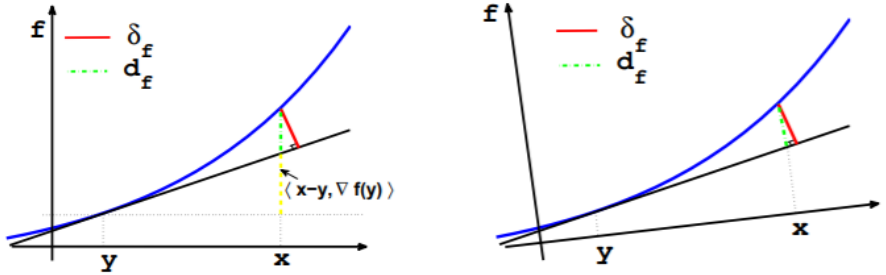
Shape Retrieval Using Hierarchical Total Bregman Soft Clustering

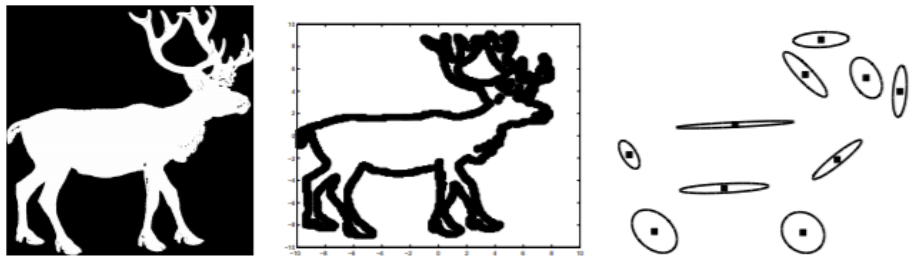
Definition *The total Bregman divergence δ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,*

$$\delta_f(x,y) = \frac{f(x) - f(y) - \langle x-y, \nabla f(y)\rangle}{\sqrt{1+\|\nabla f(y)\|^2}},$$

$\langle\cdot,\cdot\rangle$ is inner product and $\|\nabla f(y)\|^2 = \langle \nabla f(y), \nabla f(y)\rangle$ generally.

X	$f(x)$	$\delta_f(x,y)$	t-center	ℓ_1-norm BD center	Remark
$\mathbb{R}$	x^2	$\frac{(x-y)^2}{\sqrt{1+4y^2}}$	$\sum_i w_i x_i$	$\sum_i x_i$	total square loss (tSL)
$\mathbb{R}-\mathbb{R}_-$	$x\log x$	$\frac{x\log\frac{x}{y}+\bar{x}\log\frac{\bar{x}}{\bar{y}}}{\sqrt{1+y(1+\log y)^2+\bar{y}(1+\log\bar{y})^2}}$	$\prod_i (x_i)^{w_i}$	$\sum_i x_i$	
$[0,1]$	$-\log x$	$\frac{\frac{x}{y}-\log\frac{x}{y}-1}{\sqrt{1+y^{-2}}}$	$\frac{\sum_i (x_i/(1-x_i))^{w_i}}{1+\sum_i (x_i/(1-x_i))^{w_i}}$	$\sum_i x_i$	total logistic loss
$\mathbb{R}_+$	$-\log x$	$\frac{\frac{x}{y}-\log\frac{x}{y}-1}{\sqrt{1+y^{-2}}}$	$\frac{1}{\sum_i w_i/x_i}$	$\sum_i x_i$	total Itakura-Saito distance
$\mathbb{R}$	e^x	$\frac{e^x-e^y-(x-y)e^y}{\sqrt{1+e^{2y}}}$	$\sum_i w_i x_i$	$\sum_i x_i$	
$\mathbb{R}^d$	$\|x\|^2$	$\frac{\|x-y\|^2}{\sqrt{1+4\|y\|^2}}$	$\sum_i w_i x_i$	$\sum_i x_i$	total squared Euclidean
$\mathbb{R}^d$	$x^t A x$	$\frac{(x-y)^t A(x-y)}{\sqrt{1+4\|Ay\|^2}}$	$\sum_i w_i x_i$	$\sum_i x_i$	total Mahalanobis distance
Δ^d	$\sum_{j=1}^d x_j \log x_j$	$\frac{\sum_{j=1}^d x_j \log\frac{x_j}{y_j}}{\sqrt{1+\sum_{j=1}^d y_j(1+\log y_j)^2}}$	$c\prod_i (x_i)^{w_i}$	$\sum_i x_i$	total KL divergence (tKL)
$\mathbb{C}^{m\times n}$	$\|x\|_F^2$	$\frac{\|x-y\|_F^2}{\sqrt{1+4\|y\|_F^2}}$	$\frac{\|x-y\|_F^2}{\sqrt{1+4\|y\|_F^2}}$	$\sum_i x_i$	total squared Frobenius

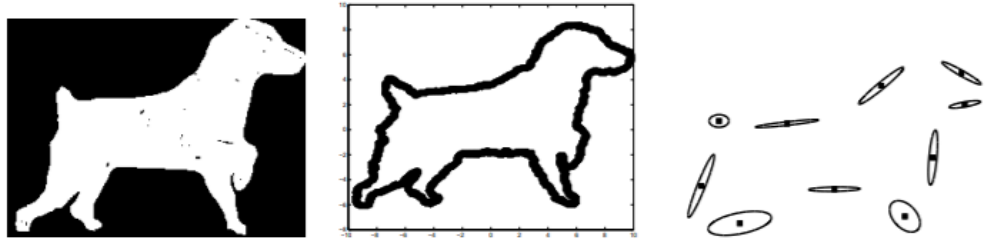




(m) (n) (o)

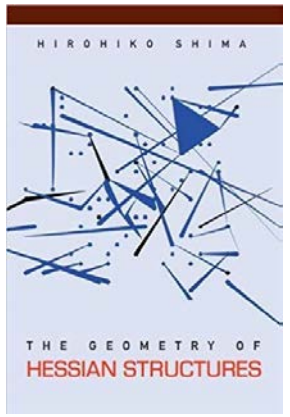
t-center: $\bar{x} = \arg\min_x \delta_f^1(x,E) = \arg\min_x \sum_{i=1}^n \delta_f(x,x_i)$

Robust to noise/outliers

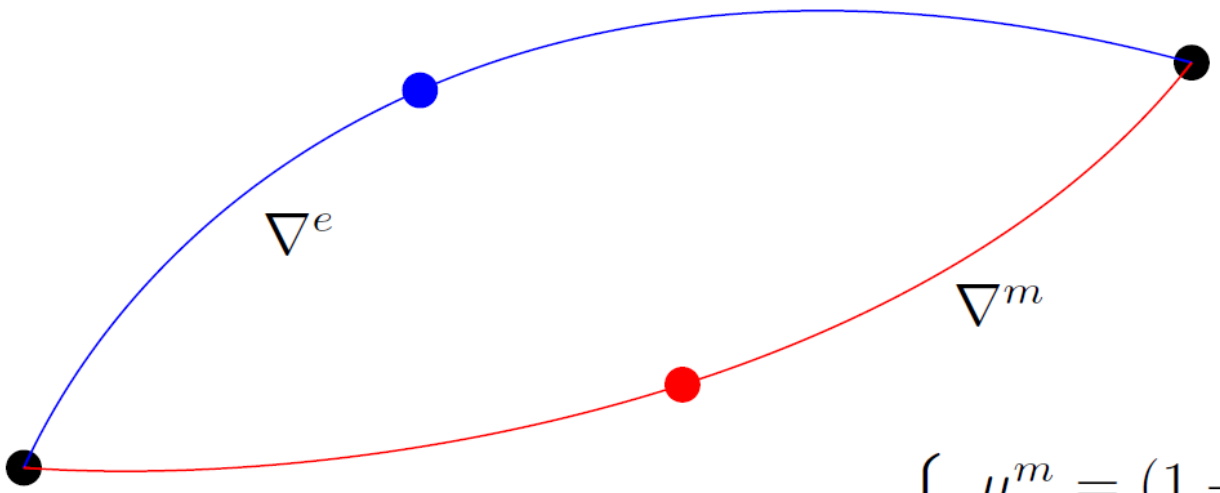


IEEE TPAMI 34, 2012

Example of dual e-/m-connections for the univariate Gaussian 2D manifold



$$(p_1p_2)^e_\alpha = \begin{cases} \mu^e_\alpha = \frac{(1-\alpha)\mu_1 v_2 + \alpha\mu_2 v_1}{(1-\alpha)v_2 + \alpha v_1} \\ v^e_\alpha = \frac{v_1 v_2}{(1-\alpha)v_2 + \alpha v_1} \end{cases}$$



$p_2 = (\mu_2, v_2 = \sigma_2^2)$

$p_1 = (\mu_1, v_1 = \sigma_1^2)$

$$(p_1p_2)^m_\alpha = \begin{cases} \mu^m_\alpha = (1-\alpha)\mu_1 + \alpha\mu_2 \\ v^m_\alpha = (1-\alpha)v_1 + \alpha v_2 - \alpha(1-\alpha)(\mu_1-\mu_2)^2 \end{cases}$$

Misconception: The m-geodesic between two Gaussians of a Gaussian manifold is a Gaussian (and not a mixture of Gaussian!)

The Gaussian is obtained from linear interpolation on the moment parameters

Converting similarities S ⟷ distances D

D: Distance measure S: Similarity measure

$$S(p,q) = \exp(-D(p,q)) \in (0,1]$$
$$D(p,q) = -\log S(p,q) \in [0,\infty)$$

Additive triangular inequality of metric distances:

$$D(p,q) + D(q,r) \geq D(p,r)$$

Multiplicative triangular inequality of similarities:

$$S(p,q) \times S(q,r) \leq S(p,r)$$

Metric distances and metric spaces (X,D)

A **metric** D is a (distance) function that satisfies the following axioms:

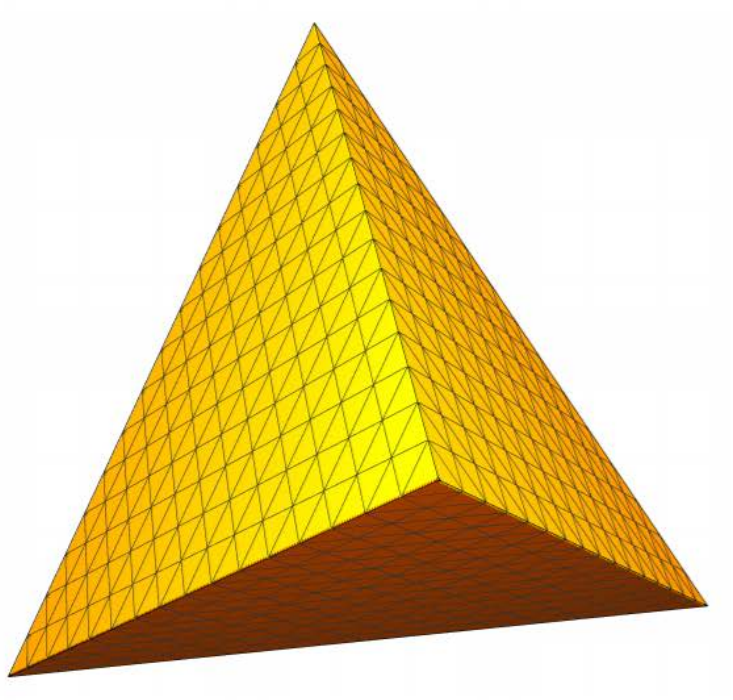
- M1. (Non-negativity) $D(p_1, p_2) \geq 0$
- M2. (Identity of the indiscernibles) $D(p_1, p_2) = 0 \Leftrightarrow p_1 = p_2$
- M3. (Symmetry) $D(p_1, p_2) = D(p_2, p_1)$
- M4. (Triangle inequality/subadditivity)

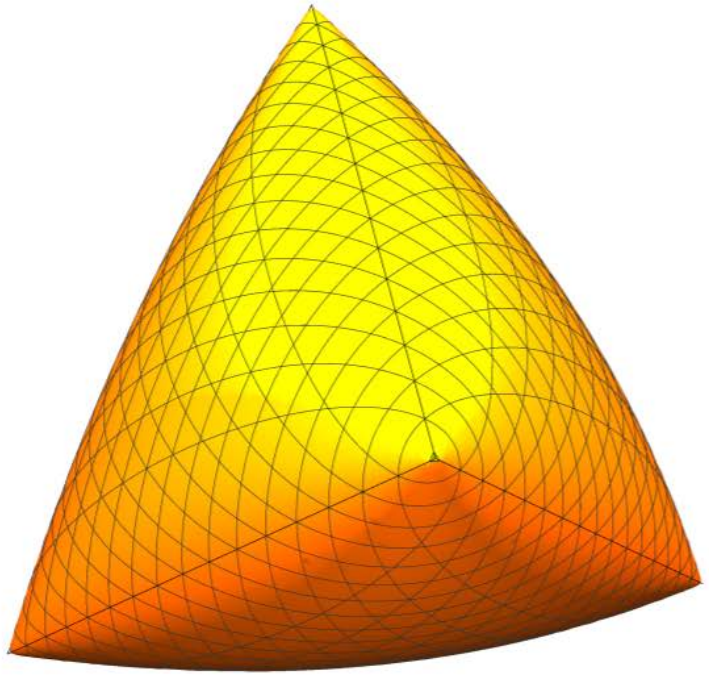
$$D(p_1, p_2) + D(p_2, p_3) \geq D(p_1, p_3)$$

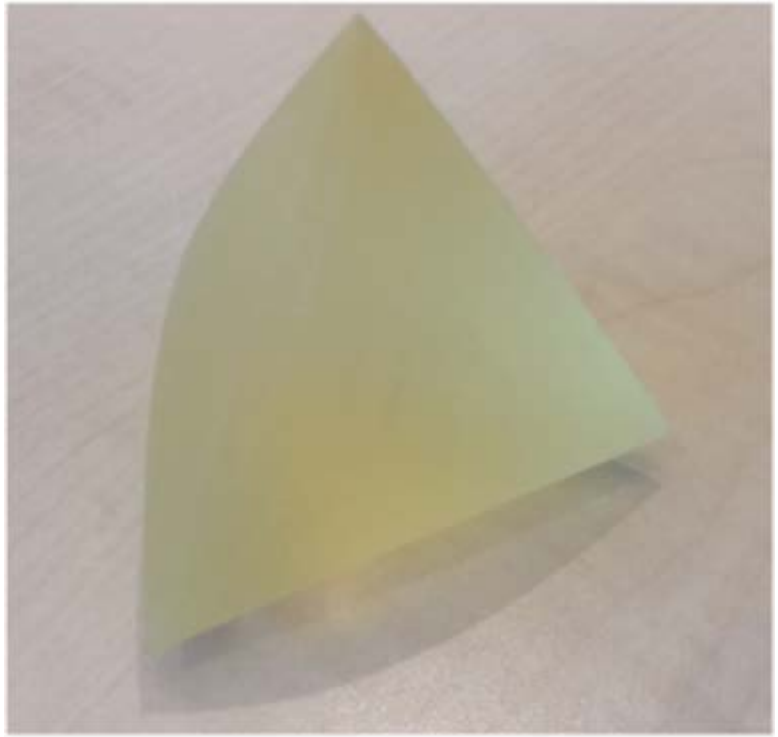
Clustering correlation matrices (elliptope)

Covariance matrices with unit diagonal, correlation coefficients

$$\mathcal{C}^d = \{C_{d\times d} \,:\, C \succ 0; C_{ii} = 1, \forall i\}$$



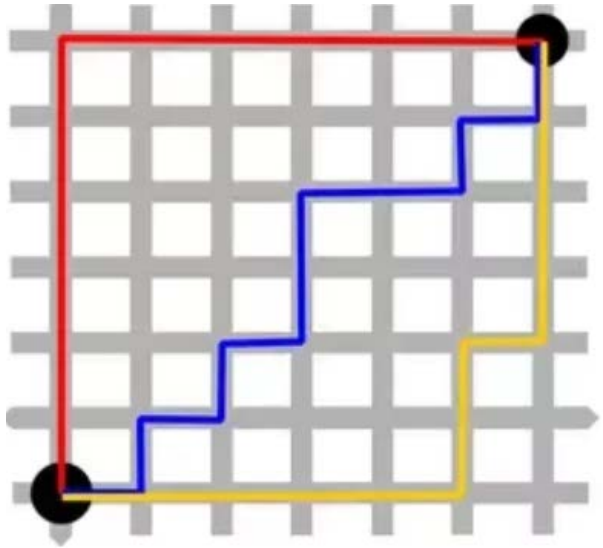




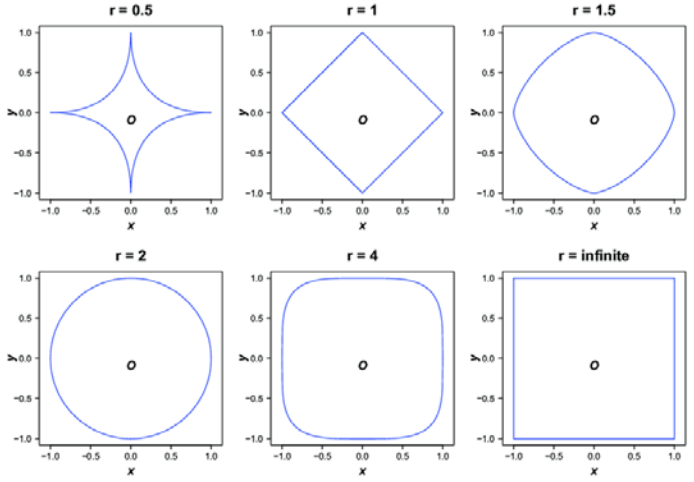
Examples of metric spaces

- Euclidean distance $D_E(p,q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$
- Manhattan/Taxi cab distance $M_1(p,q) = \sum_{i=1}^{d}|p_i - q_i|$
- Minkowski metric distances

$$M_\alpha(p,q) = \left(\sum_{i=1}^{d}|p_i - q_i|^\alpha\right)^{\frac{1}{\alpha}}, \quad \alpha \geq 1$$



L1 is not geodesic



Non-metric (not convex) and metric balls (convex)

Inner product, induced norms and induced distance

- **Inner product** $\langle x, y \rangle_G$
- Induced **norm** $\|x\|_G = \sqrt{\langle x, x \rangle_G}$
- Induced **metric distance** $D_G(p, q) = \|p - q\|_G$
- Example with Euclidean distance an its dot/scalar product

$$\langle x, y \rangle_E = \sum_{i=1}^{d} x_i y_i \quad \Rightarrow \quad D_E(p, q) = \|p - q\|_E = \|p - q\|_2$$

- Example with Minkowski norms

$$\|x\|_\alpha = \left(\sum_i |x_i|^\alpha\right)^{\frac{1}{\alpha}} \quad \Rightarrow \quad M_\alpha(p, q) = \|p - q\|_\alpha$$

Jensen-Bregman divergence as a Jensen divergence

$$
\begin{aligned}
\mathrm{JB}_F(\theta:\theta') &:= \frac{1}{2}\left(B_F\left(\theta:\frac{\theta+\theta'}{2}\right)+B_F\left(\theta':\frac{\theta+\theta'}{2}\right)\right),\\
&= \frac{F(\theta)+F(\theta')}{2}-F\left(\frac{\theta+\theta'}{2}\right)=:J_F(\theta:\theta'),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{JB}_F^{\alpha}(\theta:\theta') &:= (1-\alpha)B_F\left(\theta:(\theta\theta')_\alpha\right)+\alpha B_F\left(\theta':(\theta\theta')_\alpha\right),\\
&= (F(\theta)F(\theta'))_\alpha-F\left((\theta\theta')_\alpha\right)=:J_F^{\alpha}(\theta:\theta'),
\end{aligned}
$$

Skew Jensen-Bregman Voronoi diagrams, 2011

Generalizing Jensen-Shannon divergences

$$\begin{aligned}\mathrm{JS}(p;q) &:= \frac{1}{2}\left(\mathrm{KL}\left(p:\frac{p+q}{2}\right)+\mathrm{KL}\left(q:\frac{p+q}{2}\right)\right),\\ &= \frac{1}{2}\int\left(p\log\frac{2p}{p+q}+q\log\frac{2q}{p+q}\right)\mathrm{d}\mu.\end{aligned}$$

$$\mathrm{JS}(p;q)=h\left(\frac{p+q}{2}\right)-\frac{h(p)+h(q)}{2}.$$

Jensen-Shannon divergence is the total divergence to the average divergence
Always bounded by log 2, and the square root of JSD is a metric

On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means, Entropy 2019

https://www.mdpi.com/1099-4300/21/5/485

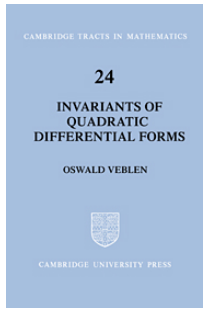
Fisher-Riemannian geometry (1930/1945)

Spaces Of Statistical Parameters.

By Harold Hotelling , Stanford University.

For a space of n dimensions representing the parameters $p_1, \ldots, p_n$ of a frequenny distribution, a statistically significant metric is defined by means of the variances and

Oswald Veblen,
Advisor of Hotelling



1930

Use Fisher information for the Riemannian metric tensor

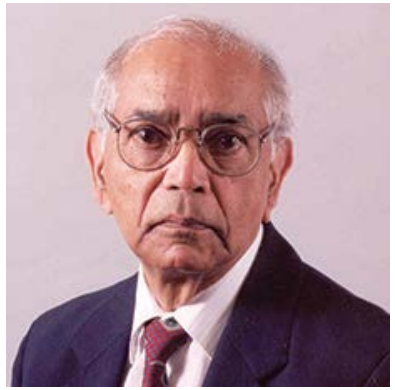
Harold Hotelling
Econometrician

Information and the Accuracy Attainable in the Estimation of Statistical Parameters

C. Radhakrishna Rao

1. Cramer-Rao lower bound CRLB
2. Rao-Blackwellization
3. Fisher-Rao distance

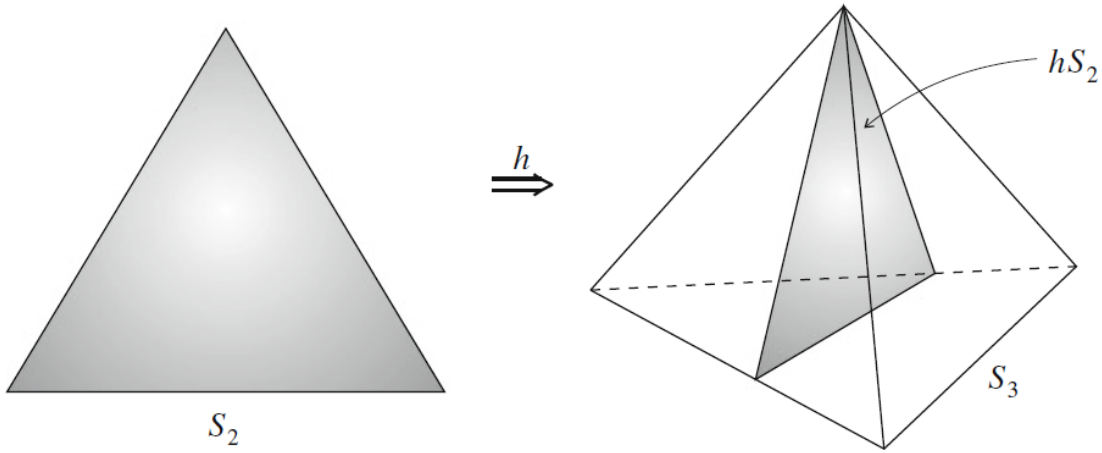
1945



C. R. Rao
Statistician

Cramér-Rao Lower Bound and Information Geometry, 2013
https://arxiv.org/abs/1301.3578

Statistical invariance



- **Fisher-Rao distance** is independent of parameterization (but FIM is covariant!)

  Same Fisher-Rao distance for parameterizations {N(μ, σ)} or {N(μ, σ2)}

- Fisher information metric is the **only invariant metric tensor** (up to a scale factor)

- Metric tensor induced by any **standard f-divergence** coincides with the Fisher information metric

- Dual connections induced by any f-divergence yield **expected alpha-connections**

Lightlike neuromanifolds, Occam's Razor and Deep Learning

Question: Why do DNNs generalize well with huge number of free parameters?

Problem: Generalization error of DNNs is experimentally
not U-shaped but a double descent risk curve (arxiv 1812.11118)

Occam's razor for Deep Neural Networks (DNNs):

(uniform width M, L layers, N #observations, d: dimension of screen distributions in lightlike neuromanifold)

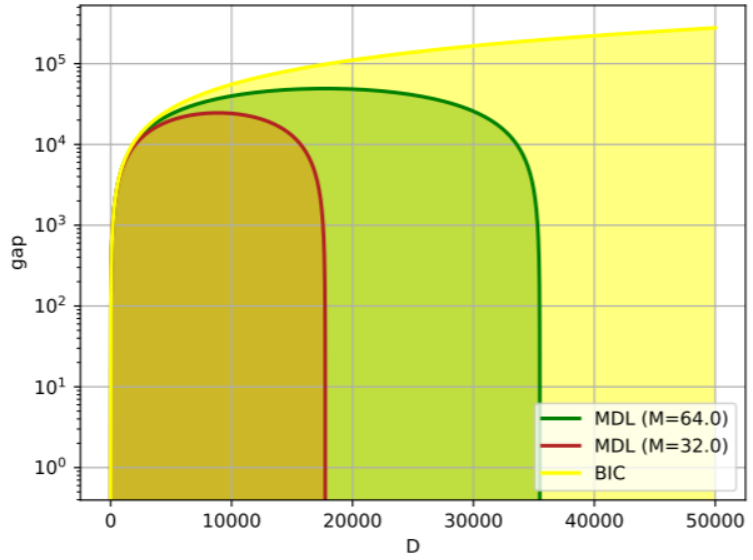
Θ: parameters of the DNN, $\hat{\Theta}$: estimated parameters

$$\mathcal{O} = -\log P(X \mid \hat{\Theta}) + \frac{d}{2}\log N + \frac{d}{2}\int_0^{\infty} \rho_{\mathcal{I}}(\lambda)\log\lambda d\lambda$$

$$\mathcal{O} \approx -\log P(X \mid \hat{\Theta}) + \frac{d}{2}\log N - \frac{d}{2}\gamma LM$$

$\rho_{\mathcal{I}}$ Spectrum density of the Fisher Information Matrix (FIM)

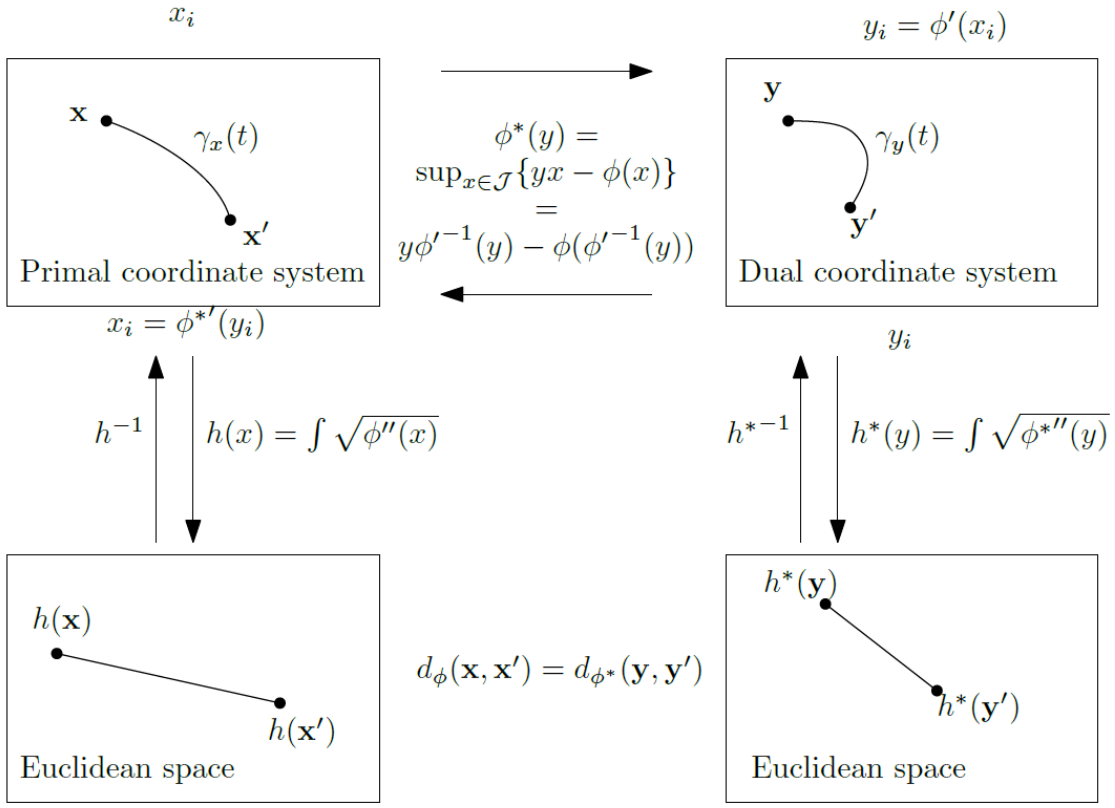
$$\mathcal{I}(\Theta) = E_p\left(\frac{\partial \log p(X \mid \Theta)}{\partial \Theta}\frac{\partial \log p(X \mid \Theta)}{\partial \Theta^{\intercal}}\right)$$



Estimated generalisation gap (in log scale) against the number of free parameters.

https://arxiv.org/abs/1905.11027

Dual Riemann geodesic distances induced by a separable Bregman divergence



$$\rho_\phi(x,x') = \rho_{\phi^*}(y,y') = \rho_{\phi^*}(\nabla\Phi(x),\nabla\Phi(x'))$$

Legendre conjugate: $\phi^*(y) = y\phi'^{-1}(y) - \phi(\phi'^{-1}(y))$

Bregman divergence:

$$B_\Phi(x,x') := \Phi(x) - \Phi(x') - (x-x')^\top \nabla\Phi(x')$$

Separable Bregman generator:

$$\Phi(x) := \sum_{j=1}^K \phi(x_j) \text{ with } \phi : \mathcal{J} \to \mathbb{R}$$

Riemannian metric tensor:

$$g_{ij}(x) = \phi''(x_i)\delta_{ij}$$

Geodesics:

$$\gamma_i(t) = h^{-1}\Big((1-t)h(x_i) + th(x_i')\Big), \quad t \in [0,1].$$

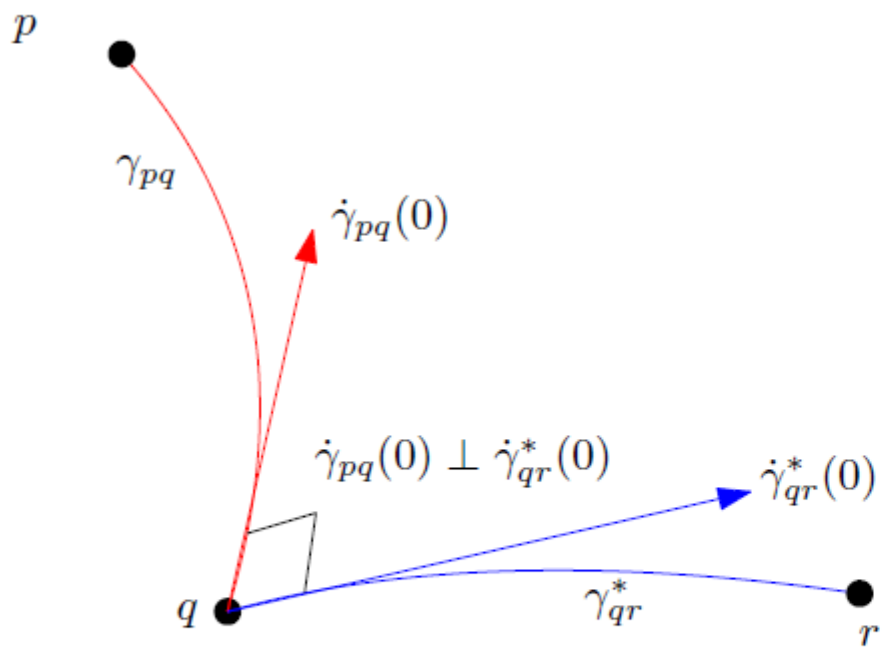
Riemannian distance (metric):

$$\rho_\phi(x,x') = \sqrt{\sum_{j=1}^K \big(h(x_j) - h(x_j')\big)^2}$$

where $h(x) := \int \sqrt{\phi''(x)}$

Geometry and clustering with metrics derived from separable Bregman divergences, arXiv:1810.10770

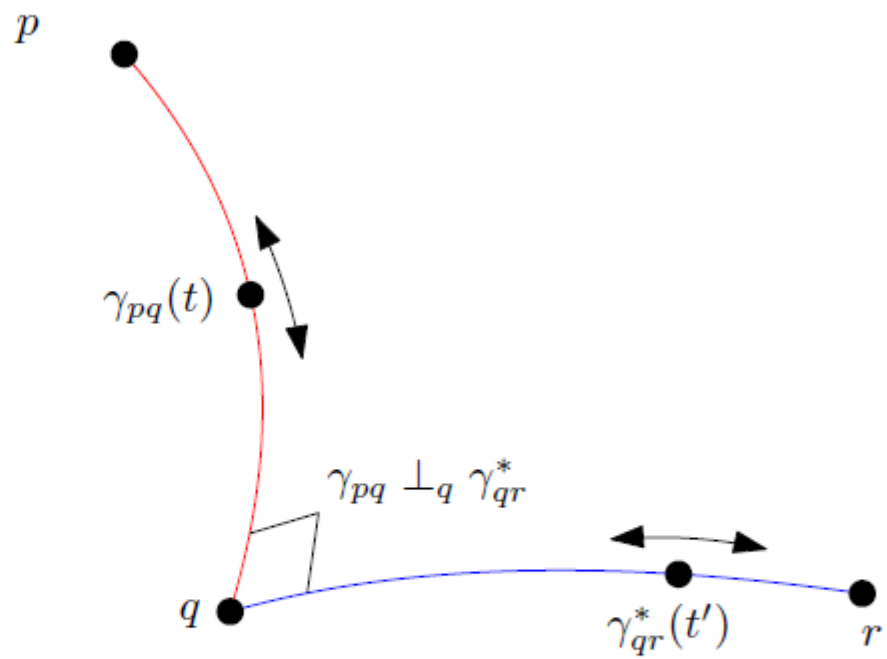
Bregman manifold: Generalized Pythagorean theorem



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

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

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



$$D_F(\gamma_{pq}(t):\gamma_{qr}(t')) = D_F(\gamma_{pq}(t):q) + D_F(q:\gamma^*_{qr}(t')),$$

https://arxiv.org/abs/1910.03935

Bregman 3-parameter/3-point identity

$$B_F(\theta_p : \theta_r) = B_F(\theta_p : \theta_q) + B_F(\theta_q : \theta_r) - (\theta_p - \theta_q)^\top (\nabla F(\theta_r) - \nabla F(\theta_q))$$

Dual parameterization $\eta_x = \nabla F(\theta_x)$

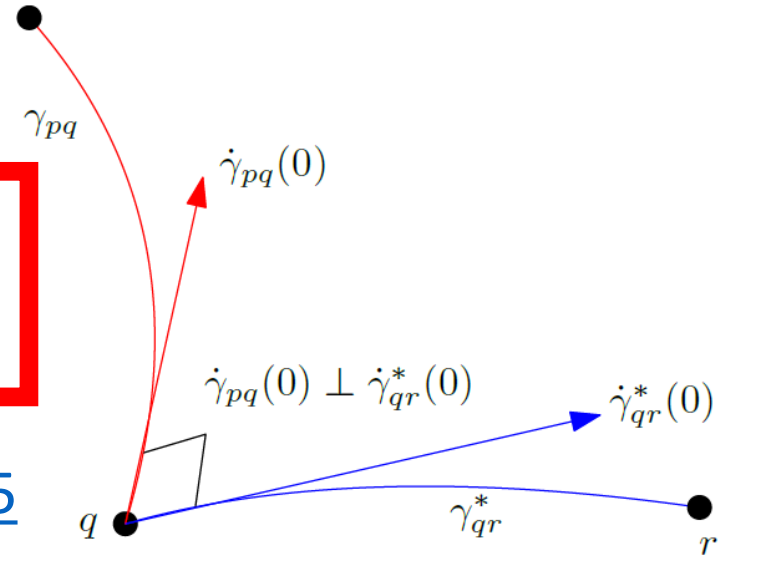
Divergence between points: $D_F(x : y) = B_F(\theta_x : \theta_y)$

Contravariant components of tangent vector to primal geodesic $\dot{\gamma}_{qp}(0)$ at q: $\theta_p - \theta_q$

Covariant components of tangent vector to dual geodesic $\dot{\gamma}^*_{qr}(0)$ at q: $\eta_r - \eta_q$

$$(\theta_p - \theta_q)^\top (\eta_r - \eta_q) = 0 \Leftrightarrow \dot{\gamma}_{qp}(0) \perp_q \dot{\gamma}^*_{qr}(0)$$

$$\begin{aligned} D(p : r) &= D(p : q) + D(q : r) - g_q(\dot{\gamma}_{qp}(0), \dot{\gamma}^*_{qr}(0)), \\ &= D(p : q) + D(q : r) - \|\dot{\gamma}_{qp}(0)\|_q \, \|\dot{\gamma}^*_{qr}(0)\|_q \cos(\alpha_q(\dot{\gamma}_{qp}, \dot{\gamma}^*_{qr})). \end{aligned}$$



https://arxiv.org/abs/1808.08271 https://arxiv.org/abs/1910.03935

Statistical manifolds from Bregman divergences

Bregman divergence (1967, on Operations research):

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$$

$$(M, F) \equiv (M, {}^{B_F} g, {}^{B_F} \nabla, {}^{B_F} \nabla^* = {}^{B_{F^*}} \nabla)$$

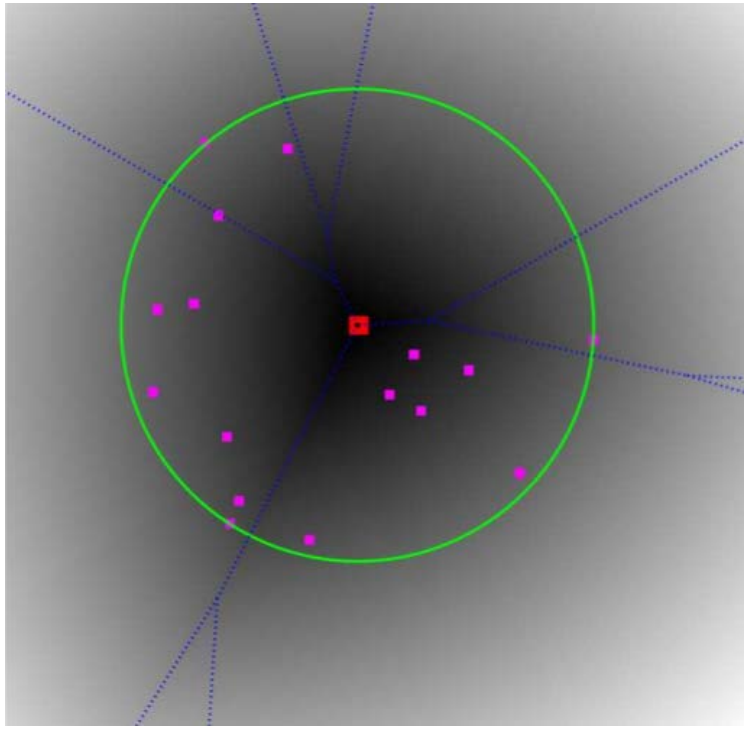
Dually flat space!

Dual Bregman divergence and Legendre-Fenchel transformation F*

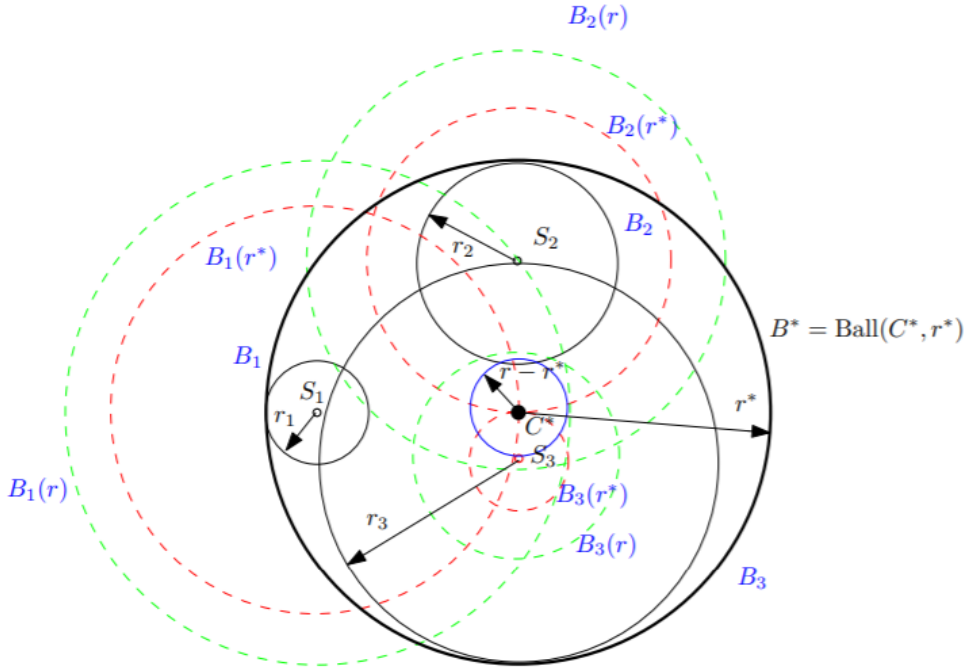
$$B_F^*(\theta : \theta') = B_F(\theta' : \theta) = B_{F^*}(\eta' : \eta)$$

$$\eta = \nabla F(\theta), \theta = \nabla F(\theta)$$

SVMs: SEBs and computational geometry



Lemma 2. *For $r \geq r^*$, there exists a ball B of radius $r(B) = r - r^*$ centered at $C(B) = C^*$ fully contained inside the intersection $\cap\mathcal{B}(r)$.*



SEB is **non-differentiable** at further Voronoi diagram boundaries

Decision problem: Covering a point set amount to pierce a corresponding point of balls

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

Cramer-Rao lower bound: Inverse of Fisher information

Löwner partial ordering on positive-semi-definite matrices: $A \succeq B \Leftrightarrow A - B \succeq 0$

CRLB Theorem:

$$\mathrm{Var}[\hat{\theta}_n] \succeq \frac{1}{n} I(\theta_0)^{-1}$$

Accuracy of estimator depends on true parameter

$$[I(\theta)]_{ij} = E_\theta \left[\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right],$$
$$= \int \left(\frac{\partial}{\partial \theta_i} \log p_\theta(x) \frac{\partial}{\partial \theta_j} \log p_\theta(x) \right) p_\theta(x) \mathrm{d}x.$$

Under regularity conditions:

$$[I(\theta)]_{ij} = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x) \right]$$

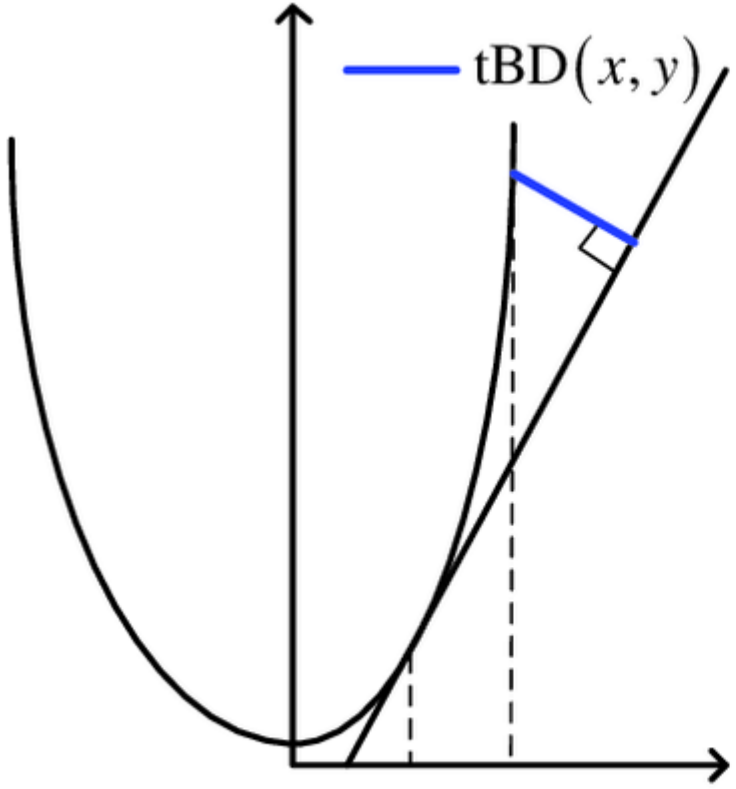
Fisher vector and kernel in ML

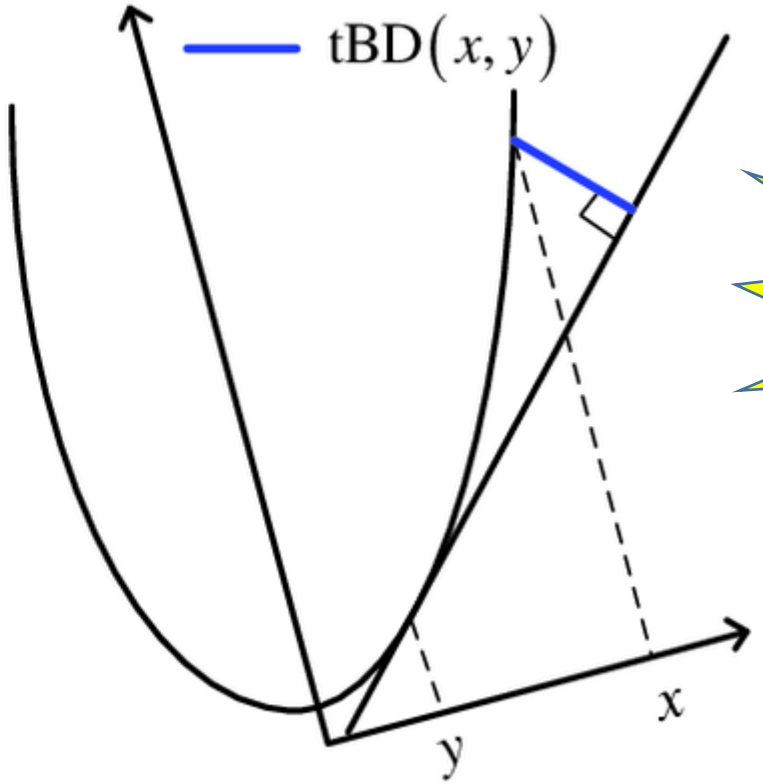
$$\phi_{Fisher}(z) = \nabla_\theta \log p_\theta(z)$$

$$k_{Fisher}(z, z') = \phi_{Fisher}(z)^T I^{-1} \phi_{Fisher}(z')$$

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

Total Bregman divergence: Further invariance!





Geometric design

$$\mathrm{TBD}(p:q)=\frac{\varphi(p)-\varphi(q)-\nabla\varphi(q)\cdot(p-q)}{\sqrt{1+|\nabla\varphi(q)|^2}}$$

Invariant to axis rotation

Total Jensen divergences: Definition, properties and clustering. ICASSP 2015

Total Bregman divergence and its applications to DTI analysis

IEEE Transactions on medical imaging, *30*(2), 475-483, 2010.

Definition *The total Bregman divergence (TBD) δ_f associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,*

$$\delta_f(x,y) = \frac{f(x) - f(y) - \langle x-y, \nabla f(y)\rangle}{\sqrt{1+\|\nabla f(y)\|^2}}, \qquad (2)$$

$\langle\cdot,\cdot\rangle$ is inner product as in definition II.1, and $\|\nabla f(y)\|^2 = \langle \nabla f(y), \nabla f(y)\rangle$ generally.

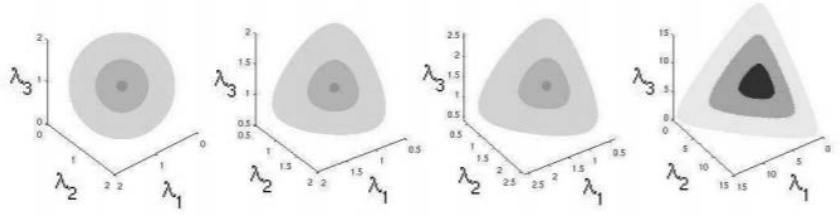
$$\iota KL(P,Q) = \frac{\int p \log \frac{p}{q} dx}{\sqrt{1+\int (1+\log q)^2 q dx}}$$

$$= \frac{\log(\det(P^{-1}Q)) + tr(Q^{-1}P) - n}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{n(1+\log 2\pi)}{2}\log(\det Q)}}$$

$$tKL(P,Q) = tKL(A'PA, A'QA), \quad \forall A \in SL(n),$$

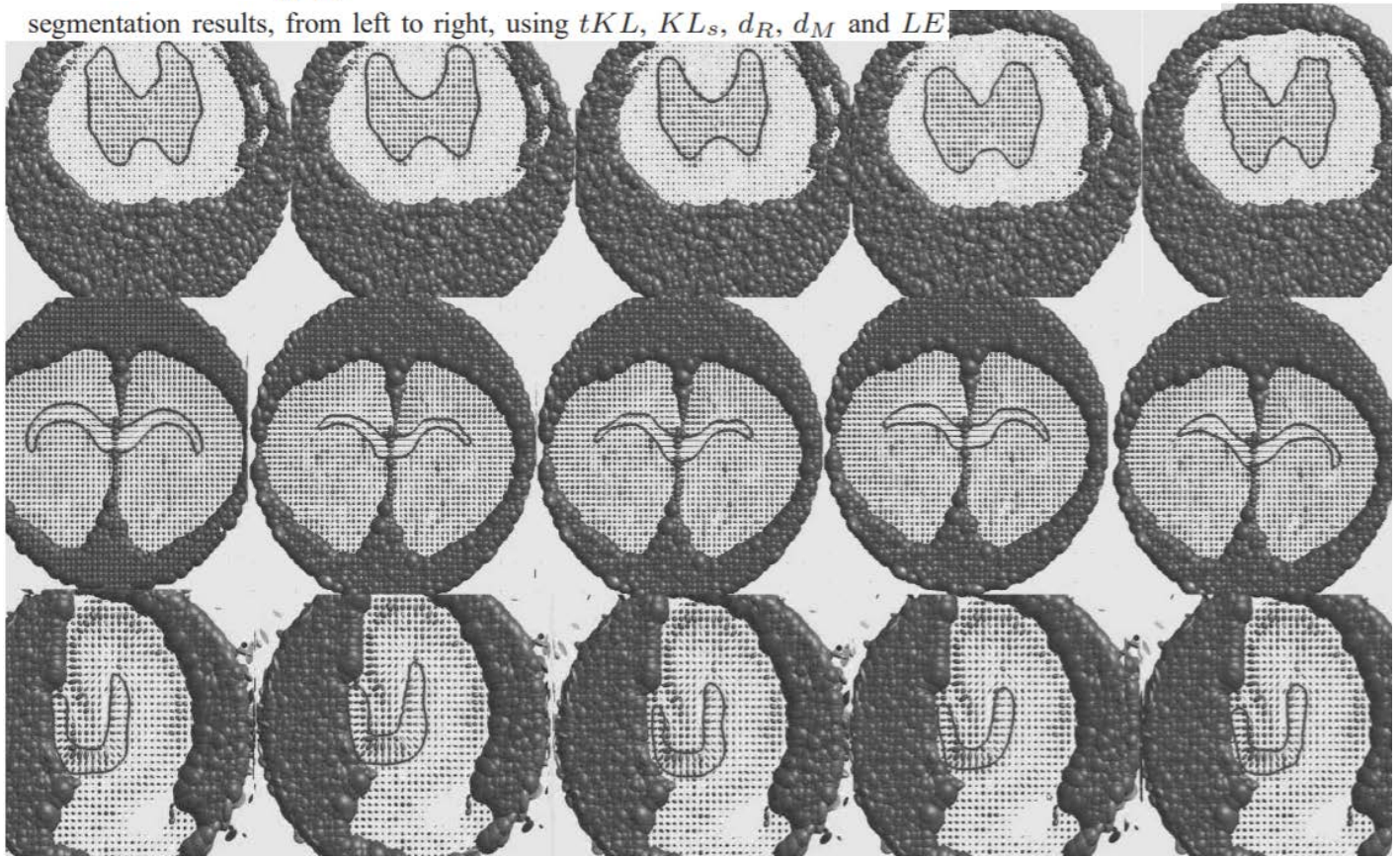
$$tSL(P,Q) = \frac{\int (p-q)^2 dx}{\sqrt{1+\int (2q)^2 q dx}} =$$

$$\frac{1/\sqrt{\det(2P)} + 1/\sqrt{\det(2Q)} - 2/\sqrt{\det(P+Q)}}{(2\pi)^n + 4\sqrt{(2\pi)^n}/\sqrt{\det(3Q)}}$$



The isosurfaces of $d_F(P,I) = r$, $d_R(P,I) = r$, $KL_s(P,I) = r$ and $tKL(P,I) = r$ shown from left to right. The three axes are eigenvalues of P.

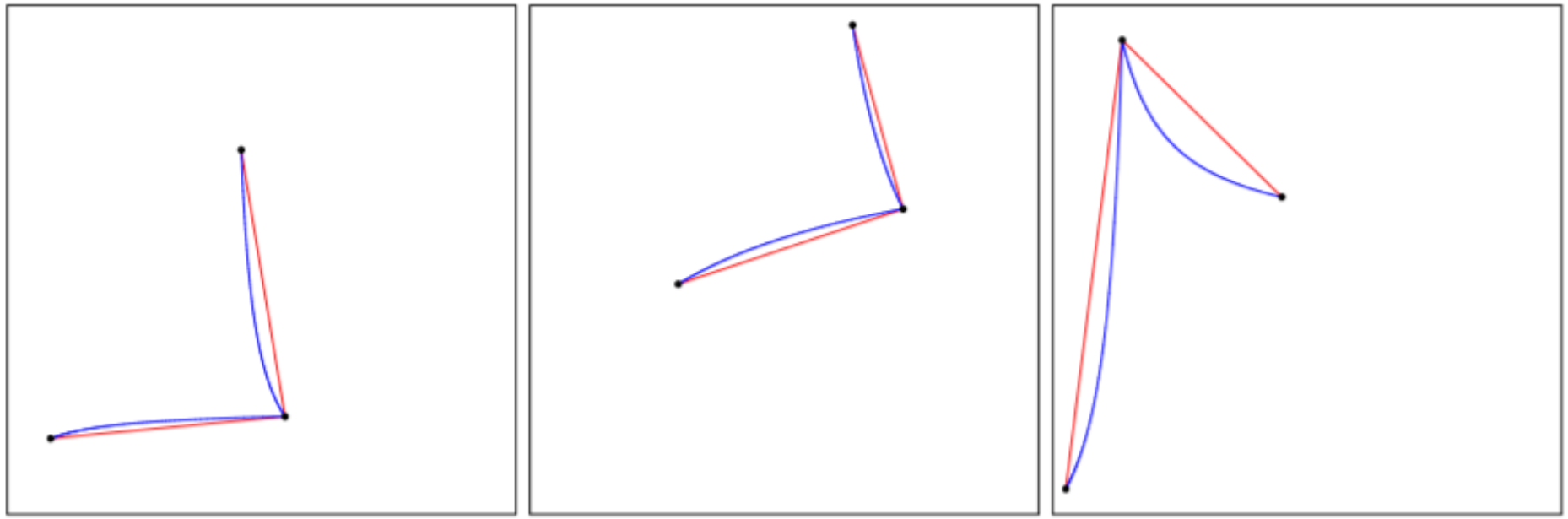
segmentation results, from left to right, using tKL, KL_s, d_R, d_M and LE



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

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

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



Itakura-Saito Manifold
(solve quadratic system)

Two blue-red geodesic pairs orthogonal at q

https://arxiv.org/abs/1910.03935