Computational Information Geometry for Machine Learning

Frank Nielsen

École Polytechnique
Sony Computer Science Laboratories, Inc
e-mail: Frank.Nielsen@acm.org

MLSS 2015

Computational Information Geometry (CIG) : Background

Computational Information Geometry (CIG) relies seamlessly on :

- statistics and probability (STAT & PR),
- information theory (IT),
- differential geometry (DG, including multilinear algebra of tensors),
- computation :
  Yes, we are computer scientists and programmers ! How do we compute friendly ? (make wide & wise use of dualities...)

Many *application fields* : computational statistics, machine learning (ML), information retrievals (IRs), computer vision (CV), medical imaging, radar signal processing, etc.

→ Method of information geometry [2] (2000), prone a framework !


GEOMETRIES
DISTANCES
COORDINATE-FREE
MAXIMUM
COMPUTATIONAL
COMPUTING
ENTROPY
ENTROPIES
MEASURES
LEARNING
LANGUAGE
DUALLY-FLAT
GEOMETRY
META-ALGORITHMS
PRINCIPLE
AXIOMS
POSITIVE
CLASSES
GEOMETRIC
ALGORITHMS
MACHINE
STATISTICS
DISTRIBUTIONS
CROSS-ENTROPIES
RIEMANNIAN
PROBABILITY
FAMILIES
GENERALIZED
EXPONENTIAL

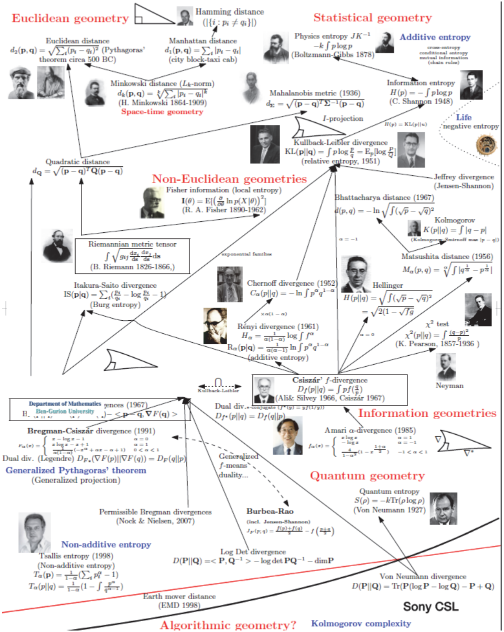
Motivations : Setting goals !

Computational Information Geometry : Main goals

1. understand "distances" and group them axiomatically into classes and build generic meta-algorithms (unifying former algorithms) : Bregman divergences B_F, Csiszár f-divergences I_f, proper scoring rules, etc.
   $\rightarrow$ seek for "properties with exhaustivity",
2. understand relationships between distances and geometries,
3. understand *generalized entropies*, cross-entropies, maximum entropy probability distributions, and their induced geometries (beyond Shannon/Boltzmann/Gibbs).
4. provide (coordinate-free) intrinsic computing using the language/affordances of geometry (for computational statistics, machine learning and predictive analytics)

Goal 1. Dissimilarities (distances) and meta-algorithms

- unify algorithms into meta-algorithms working on classes of distances (metrics, divergences) :
  - parameter estimation (with goodness-of-fit),
  - center-based clustering (with Bregman distances),
  - learning (boosting with surrogate loss functions),
  - forecasting (with proper score functions),
  - etc.
- propose new principled classes of distances :
  total Bregman divergences [17], total Jensen divergences [41], conformal divergences [45], etc.
- understand axiomatically properties and relationships between distances (or multi-entity diversity indexes) and search for their exhaustive characterizations.



http://www.sonycsl.co.jp/person/nielsen/FrankNielsen-distances-figs.pdf

Goal 2. Distances and geometries

Not 1-to-1 (because same geometry can be realized for different distances).
Geometry = meta-model
Embedding (isometrically) a geometry into another geometry : =model interpreted into another larger model.

- Underlying geometries of distances/divergences :
  - Riemannian geometry with metric distances (with the metric Levi-Civita connection),
  - Dually coupled affine differential geometry ($\pm\alpha$-geometry) and non-metric distances (aka. divergences),
  - monotone embeddings into (ρ, τ)-structure (extending I_α-embedding),
  - etc.
- geometries of probability distributions/positive measures and distances :

How to define statistical manifolds?

Goal 3. Entropies, cross-entropies, relative entropies and MaxEnt distributions

- entropies $H(P)$ (Shannon-Boltzmann-Gibbs), cross-entropies $H^\times(P:Q)$ and relative entropies KL. $\mathrm{KL}(P:Q) = H^\times(P:Q) - H(P)$ with $H(P) = H^\times(P:P)$.
- generalized entropies (so called deformed "logarithms"), the concept of escort distributions,
- maximum entropy principle and equilibrium distributions (Boltzmann-Gibbs, Tsallis's heavy tailed distributions, etc.)
- entropies, information (=neg-entropy) and complexity (Kolmogorov, non-computability)

Goal 4. Geometric computing for intrinsic computing

Propose a paradigm for data science : from "datum" (biased) processing to geometric "pointum" (non-biased) coordinate-free computing

- get unbiased processing : coordinate-free !,
- use affordances of the geometric language for building/explaining algorithms :
  points, geodesics, balls, orthogonality, projection, Pythagoras, flat, submanifold, etc.
- analytic and synthetic geometries (closed-form or exact geometric characterization).
  Example : Two pseudo-segments always intersect in a common point... that may not be in closed-form.
- invariance (and statistical invariance) and geometry :
  group of invariance, invariance and sufficiency, statistical invariance, etc.

Geometrizing probability spaces yields statistical manifolds

Part I : Geometry of statistical manifolds

Outline of Part I

1. Fisher information (Cramér-Rao lower bound) & sufficiency (1922)
2. Structures from differential geometry of population spaces (Hotelling, 1930, Rao, 1945, Amari-Centsov 1980's)
3. Maximum entropy principle (exponential families) (1957, Jaynes)
4. Information projections (and Pythagoras' theorem)

I. Statistical Information
Fisher Information
$I(\theta)$

Old days - :) Discrete and Continuous random variables

- Discrete RV : probability mass function (pmf) $X \sim p$, discrete support $\mathcal{X}$.

$$\boxed{\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x) x = \langle X \rangle}$$

Distributions : Bernoulli, binomial, multinomial, Poisson, etc.-∞,

- Continuous RV : probability density function (pdf) $X \sim p$, continuous support $\mathcal{X}$.

$$\boxed{\mathbb{E}[X] = \int_{x \in \mathcal{X}} p(x) x \mathrm{d}x = \langle X \rangle}$$

Distributions : exponential, normal, lognormal, gamma, beta, Dirichlet, Wishart, etc.-∞,

From data sets to empirical (discrete) distributions

Given $X = \{x_1, ..., x_n\}$ observations...
...build the empirical distribution :

$$
\begin{aligned}
p_e(X) &= \frac{1}{n}\sum_{i=1}^{n} \delta(X - X(i))\\
F_e(x) &= \frac{1}{n}\sum_{i=1}^{n} 1_{[x_i \leq x]} \text{ (cdf)}\\
p_e^i &= \frac{1}{n}\#\{x = i\} \text{ (frequency)}
\end{aligned}
$$

Support $\mathcal{X}$ is unknown *a priori* : not a multinomial distribution nor a finite mixture !
Sample mean $\bar{\mu} = \frac{1}{n}\sum_i x_i = \langle X \rangle_{p_e} = \sum_{i\in ???} p_e^i i$.
Estimation $X \sim D(\theta)$ by the method of moments :

$$\boxed{\langle X \rangle_{p_e} = \mathbb{E}[X] = \langle X \rangle}$$

Old days : Discrete and continuous random variables

- Discrete RV. Shannon entropy :

$$\boxed{H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \geq 0}$$

always positive (notion of uncertainty! max uncertainty for uniform distribution : $H(U) = \log n$)

- Continuous RV. Differential entropy :

$$\boxed{H(X) = \int_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \mathrm{d}x}$$

can be negative (physical interpretation!) ...
For example, for multivariate normals (MVNs) $N(\mu, \Sigma)$:

$$H(X) = \frac{1}{2} \log(2\pi e)^d |\Sigma|$$

Mixture sampling : Example of a Gaussian Mixture Model (GMM)

To sample a variate x from a GMM :

- Choose a component l according to the weight distribution $w_1, ..., w_k$,
- Draw a variate x according to $N(\mu_l, \Sigma_l)$.

$\rightarrow$ Sampling is a **doubly stochastic process** :

- throw a biased dice with k faces to choose the component :

$$l \sim \mathrm{Multinomial}(w_1, ..., w_k)$$

  (Multinomial is normalized histogram without void bins)
- then draw at random a variate x from the l-th component

$$x \sim \mathrm{Normal}(\mu_l, \Sigma_l)$$

  $x = \mu + Cz$ with Cholesky : $\Sigma = CC^T$ and $z = [z_1 \ ... \ z_d]^T$ standard normal random variate :$z_i = \sqrt{-2 \log U_1} \cos(2\pi U_2)$

.

Statistical mixtures : discrete, continuous or mixed !

Finite mixture models ($k \in \mathbb{N}$) have pmf/pdf :

$$\boxed{m(x) = \sum_{i=1}^{k} w_i p_i(x)}$$

(not sum of RVs, $M \neq \sum_i w_i X_i$ that have <u>convolutional densities</u>)

- mixtures of Gaussians (universal representation for smooth densities)
- multinomial distribution is a mixture
  (and also an exponential family in information geometry...)

What about the mixture of a standard Gaussian with a binomial distribution ? $\rightarrow$ Neither discrete nor continuous !

Measure theory (axiom system of Kolmogorov, 1933)

- unify discrete and continuous RVs as probability measures (pm) μ, ν, etc.
- can handle RVs that are neither continuous nor discrete (eg., a mixture of Poisson with a Gaussian)
- for probability measures, pmfs/pdfs are Radon-Nikodym derivatives
- expectation notation is unified as :

$$\boxed{\mathbb{E}[X] = \int_{x \in \mathcal{X}} x p(x) \, \mathrm{d}\nu(x)}$$

- Two usual base measures :
  - counting measure : ν_C ($\int \rightarrow \sum$)
  - Lebesgue measure : ν_L

Measure theory : Probability space (recalling terminology)

- $\mathcal{X}$ a set, the sample space
- σ-algebra $\mathcal{F}$ over $\mathcal{X}$: subsets of $\mathcal{X}$ closed under countable many intersections, unions, and complements.
- $(\mathcal{X}, \mathcal{F})$: measurable space
- measure $\mu : \mathcal{F} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ with
  - $\mu(E) \geq 0, \forall E \in \mathcal{F}$, $\mu(\emptyset) = 0$
  - $\mu\left(\cup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \mu(E_i)$ for pairwise disjoint sequence $\{E_i \in \mathcal{F}\}_i$
- $(\mathcal{X}, \mathcal{F}, \mu)$, a (positive) measure space
- $(\mathcal{X}, \mathcal{F}, \mu)$ with $\mu(\mathcal{X}) = 1$, a probability space, $F \in \mathcal{F}$ are events

Measurable functions and random variables

- Measurable function $f : \mathcal{X} \rightarrow \mathcal{Y}$ between two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$:

$$\forall G \in \mathcal{G}, \quad f^{-1}(G) \in \mathcal{F}$$

- Random variable X = measurable function $X : \mathcal{X} \rightarrow \mathbb{R}$. Therefore :

$$\{x \in X \ | \ a < X(x) < b\} \in \mathcal{F}$$

all sample states with X taking values between a and b is an event (CDF)
- continuous RV = measures on Borel σ-algebra

Dominance and Radon-Nikodym derivatives

- measure μ is <u>dominated by</u> mesure ν ($\mu \ll \nu$) iff.

$$\boxed{\nu(E) = 0 \Rightarrow \mu(E) = 0}$$

- $\mu \ll \nu$ <u>σ-finite</u> ($\mathcal{X}$=countable union of measurable sets with finite measure) then μ admits a density f wrt to ν, the Radon-Nikodym derivative :

$$\boxed{f \stackrel{\mathrm{n.}}{=} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}}$$

$$\forall\ \nu - \text{measurable } E, \boxed{\mu(E) \stackrel{\mathrm{n.}}{=} \int_{e \in E} f \mathrm{d}\nu(e)}$$

- $P \ll \nu$, Shannon entropy : $H(P) = -\int p(x) \log p(x) \mathrm{d}\nu(x)$.

Statistical estimation : parametric estimation $\hat{\theta}$

- Given idd. $X = \{x_1, ..., x_n\} \sim p_{\theta_0}(x)$ (hidden by Nature), estimate θ in family $\{p_\theta(x)\}_\theta$?
  $\rightarrow$ from observation sets to random vectors
- Maximum Likelihood Principle (MLE) :

$$\boxed{\hat{\theta}_n = \mathrm{argmax}_\theta \prod_i p_\theta(x_i) = \mathrm{argmax}_\theta l(X;\theta) = \sum_i \log p_\theta(x_i)}$$

- Consistency : $\lim_{n\rightarrow\infty} \hat{\theta}_n = \theta_0$
- score function : $\boxed{s(\theta, x) = \nabla_\theta \log p_\theta(x)}$ with $\nabla_\theta = (\partial_i = \frac{\partial}{\partial\theta^i})_i$. score indicates the *sensitivity of the log-likelihood curve*.
- For strictly concave log-likelihood, unique $\hat{\theta}$ such that $s(\hat{\theta}, x) = 0$ (MVNs, Beta, Poisson, Dirichlet, etc).

Fisher information $I(\theta)$ = Variance of the score

Amount of information that an observable random variable X carries about an unknown parameter θ :

First moment of score : 0, not discriminative !

$$
\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(X;\theta) \mid \theta\right] &= \mathbb{E}\left[\frac{\frac{\partial}{\partial\theta}p(X;\theta)}{p(X;\theta)} \mid \theta\right] = \int \frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)} p(x;\theta)\,\mathrm{d}x \\
&= \int \frac{\partial}{\partial\theta} p(x;\theta)\mathrm{d}x = \frac{\partial}{\partial\theta}\int f(x;\theta)\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}1 = 0.
\end{aligned}
$$

<u>**Second moment of score**</u> : (with $\partial_i l(x;\theta) = \frac{\partial}{\partial\theta_i} l(x;\theta)$)

$$
\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \middle| \theta\right] = \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,\mathrm{d}x > 0
$$

Multi-parameter : $\boxed{I_{i,j}(\theta) = \mathbb{E}_\theta[\partial_i l(x;\theta)\partial_j l(x;\theta)]}$, $\quad I(\theta) \succeq 0$, PS(S)D

Fisher information and Cramér-Rao lower bound

How good is an estimator ? how to measure goodness ?

- Mean Square Error (MSE) : $\mathrm{MSE}(\theta) \stackrel{\mathrm{eq}}{=} \mathbb{E}[\|\hat{\theta} - \theta_0\|^2]$ (consistency : MSE$\rightarrow$ 0)
- Cramér-Rao lower bound : for an *unbiased estimator* $\hat{\theta}$:

$$\boxed{\mathbb{V}[\hat{\theta}] \succeq I^{-1}(\theta_0)}$$

- efficiency : unbiased estimator matching the CR lower bound
- asymptotic normality of $\hat{\theta}$ (on random vectors) :

$$\boxed{\hat{\theta} \sim N\left(\theta_0, \frac{1}{n} I^{-1}(\theta_0)\right)}$$

Fisher Information Matrix (FIM)

$$I(\theta) = [I_{i,j}(\theta)]_{i,j}, \quad I_{i,j}(\theta) = \mathbb{E}_\theta[\partial_i l(x;\theta)\partial_j l(x;\theta)]$$

- For <u>multinomials</u> $(p_1, ..., p_d)$:

$$\boxed{I(\theta) = \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \dots & -p_1 p_k \\ -p_1 p_2 & p_2(1-p_2) & \dots & -p_2 p_k \\ \vdots & & & \vdots \\ -p_1 p_k & -p_2 p_k & \dots & p_k(1-p_k) \end{bmatrix}}$$

- For <u>multivariate normals</u> (MVNs) $N(\mu, \Sigma)$:

$$\boxed{I_{i,j}(\theta) = \frac{\partial \mu^\top}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right)}$$

matrix trace : tr.

Reparameterization of the Fisher information matrix

- Let $\theta = \theta(\eta)$ and η be two 1-to-1 parameterizations
- $J = [J_{i,j}]_{i,j}$: Jacobian matrix $J_{i,j} = \frac{\partial \theta_i}{\partial \eta_j}$.

$$\boxed{I_\eta(\eta) = J^\top \times I_\theta(\theta(\eta)) \times J}$$

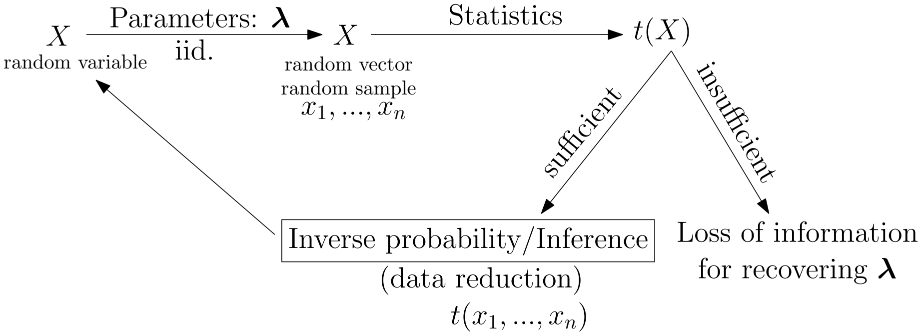
Fisher information matrix depends on the parameterization of the parameter space (covariant)

Statistics : Information and sufficiency

- sufficiency : $\mathbb{P}(x|t,\theta) = \mathbb{P}(x|t)$
  $\Rightarrow$ **all information about θ is contained inside t**
- $I_{s(X)}(\theta) \leq I_X(\theta)$ for a statistic s, with equality iff. s is sufficient
- Fisher-Neyman's factorization criterion : $t(x)$ is sufficient then we have the following canonical factorization :

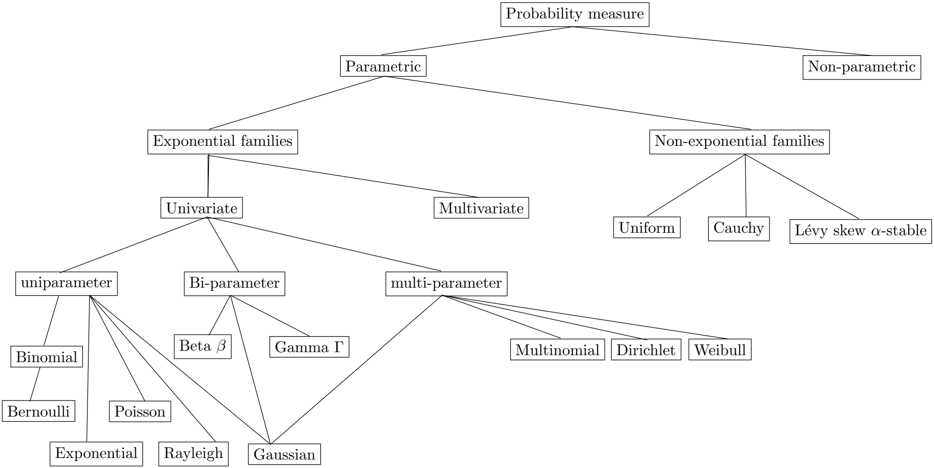
$$p(x;\theta) = g(t(x);\theta)h(x)$$

- Ex. : $t(x) = (\sum_i x_i, \sum_i x_i^2)$ sufficient for univariate normals.
  - All information about θ in two quantities : data reduction without loss of statistical information
  - sample mean $\bar{\mu} = \frac{1}{n}\sum_i x_i$, sample variance $\bar{v} = \frac{1}{n}\sum_i (x_i - \bar{\mu})^2 = \frac{1}{n}\sum_i x_i^2 - \bar{\mu}^2 = \boxed{\frac{1}{n}\sum_i x_i^2 - (\frac{1}{n}\sum_i x_i)^2}$
- not all statistics carry information on θ : ancillary statistics, statistics that does not depend on the parameter θ.



We are interested in finite-dimensional sufficient statistics... (statistical lossless data reduction)

Exponential families and finite sufficiency



Beware : Exponential distribution belongs to the exponential families too.

Exponential families : families of parametric distributions

- Canonical decomposition ($t(x)$ sufficient statistics, $k(x)$ auxiliary carrier term) :

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

- log-Laplace transform : $\boxed{F(\theta) = \log \int \exp(\langle t(x), \theta\rangle + k(x))\mathrm{d}x}$
- many distributions $p(x;\lambda)$ (normal, gamma, beta, multinomial, Poisson) are exponential families with $\theta(\lambda)$
- F is *strictly convex* on convex natural parameter space $\Theta = \{\theta \in \mathbb{R}^D \mid F(\theta) < \infty\}$
- Dual parameterizations : $\theta(\lambda)$ or $\eta(\lambda) = \nabla F(\theta(\lambda)) = \mathbb{E}[t(X)]$
- Fisher information matrix : $I(\theta) = \nabla^2 F(\theta) \succ 0$ (Hessian of strictly convex function)
- MLE : $\hat{\eta} = \frac{1}{n}\sum_i t(x_i) = \nabla F(\theta)$ (condition on existence)

Convex duality : Legendre-Fenchel transformation [21, 19]

- For a strictly convex and differentiable function $F : \mathcal{X} \rightarrow \mathbb{R}$, define the <u>convex conjugate</u> :

$$\boxed{F^*(y) = \sup_{x\in\mathcal{X}} \{\underbrace{\langle y, x\rangle - F(x)}_{l_F(y;x);}\}}$$

- Maximum obtained for $y = \nabla F(x)$:

$$\nabla_x l_F(y;x) = y - \nabla F(x) = 0 \Rightarrow y = \nabla F(x)$$

- Maximum *unique* from convexity of F ($\nabla^2 F \succ 0$) :

$$\nabla_x^2 l_F(y;x) = -\nabla^2 F(x) \prec 0$$

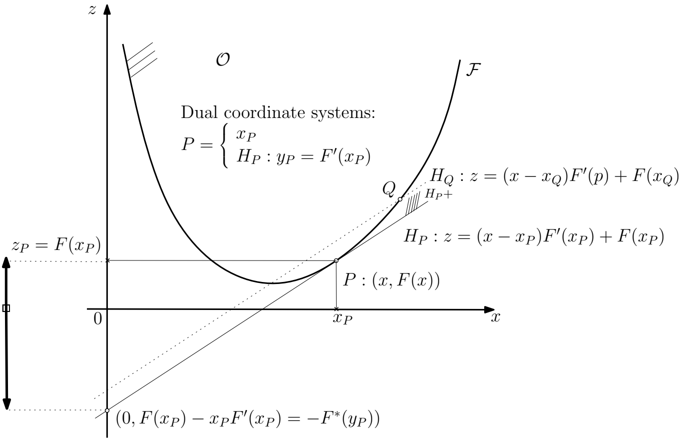
- **Convex conjugates with domains** :

$$(F, \mathcal{X}) \Leftrightarrow (F^*, \mathcal{Y}), \qquad \mathcal{Y} = \{\nabla F(x) \mid x \in \mathcal{X}\}$$

Legendre duality : Geometric interpretation

Consider the epigraph of F as a convex object :

- convex hull (vertex, V-representation), versus
- half-space (halfspace, H-representation).



Legendre transform also called "slope" transform.

Legendre duality & Canonical divergence

- Convex conjugates have *functional inverse* gradients $\nabla F^{-1} = \nabla F^*$
  ∇F^* may require numerical approximation
  (not always available in analytical closed-form)
- Involution : $(F^*)^* = F$ with $\nabla F^* = (\nabla F)^{-1}$.
- Convex conjugate F^* expressed using $(\nabla F)^{-1}$:

$$
\begin{aligned}
F^*(y) &= \langle x, y\rangle - F(x), x = \nabla_y F^*(y) \\
F^*(y) &= \langle (\nabla F)^{-1}(y), y\rangle - F((\nabla F)^{-1}(y))
\end{aligned}
$$

- Fenchel-Young inequality at the heart of the canonical divergence :

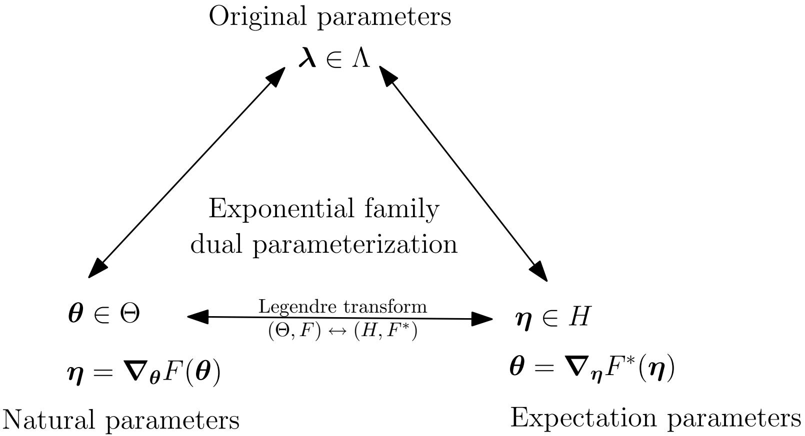
$$F(x) + F^*(y) \geq \langle x, y\rangle$$

$$\boxed{A_F(x:y) = A_{F^*}(y:x) = F(x) + F^*(y) - \langle x, y\rangle \geq 0}$$

Parameters of exponential families

- D : order of the exponential family
- d : uni- ($d = 1$) or multi-variate family

Many parameterizations are possible but only two are canonical : natural parameters and expectation parameters.



Canonical decomposition of exponential families

$\langle \cdot, \cdot \rangle$: inner product on vectors (scalar product), matrices ($\mathrm{Re}\mathrm{Tr}(AB^*)$)
$t(x)$ sufficient statistics, $k(x)$ auxiliary carrier term :

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

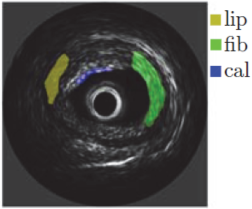
Not unique decomposition because :

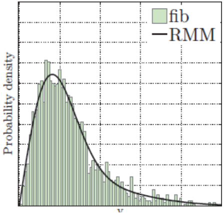
- natural parameter and sufficient statistic : $t'(x) = At(x)$ and $\theta' = A^{-1}\theta$ (for $|A| \neq 0$ affine transformation)
- constant in $F'(\theta) = F(\theta) + c$ and $k'(x) = k(x) - c$

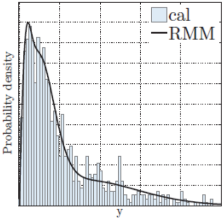
Let us give some decomposition examples...

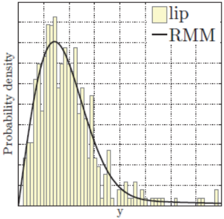
Statistical mixtures : Rayleigh MMs [28]

IntraVascular UltraSound (IVUS) imaging :









Rayleigh distribution :

$p(x;\lambda) = \frac{x}{\lambda^2} e^{-\frac{x^2}{2\lambda^2}}$

$x \in \mathbb{R}^+$

$d = 1$ (univariate)

$D = 1$ (order 1)

$\theta = -\frac{1}{2\lambda^2}$

$\Theta = (-\infty, 0)$

$F(\theta) = -\log(-2\theta)$

$t(x) = x^2$

$k(x) = \log x$

(Weibull $k = 2$)

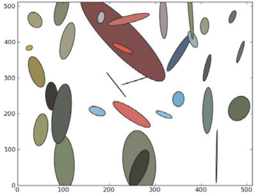
Coronary plaques : fibrotic tissues, calcified tissues, lipidic tissues

Rayleigh Mixture Models (RMMs) :

for *segmentation* and *classification* tasks

Statistical mixtures : Gaussian MMs [12, 28, 13]

Gaussian mixture models (GMMs) : model low frequency.
Color image interpreted as a 5D xyRGB point set.

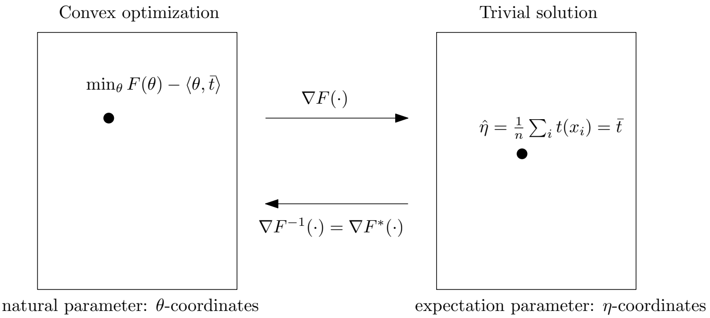


Gaussian distribution $p(x; \mu, \Sigma)$:
$$\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} e^{-\frac{1}{2} D_{\Sigma^{-1}}(x-\mu, x-\mu)}$$

Squared Mahalanobis distance :
$D_Q(x, y) = (x - y)^T Q(x - y)$
$x \in \mathbb{R}^d$
d (multivariate)
$D = \frac{d(d+3)}{2}$ (order)
$\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) = (\theta_v, \theta_M)$
$\Theta = \mathbb{R} \times S^d_{++}$
$F(\theta) = \frac{1}{4}\theta_v^T \theta_M^{-1}\theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log\pi$
$t(x) = (x, -xx^T)$
$k(x) = 0$

MLE of exponential families : Two coordinate systems

$$\eta = \mathbb{E}[t(x)] = \nabla F(\theta), \quad \theta = (\nabla F)^{-1}(\eta) = \nabla F^*(\eta)$$



- Closed-form in expectation parameter coordinate system η : $\hat{\eta} = \frac{1}{n} \sum_i t(x_i)$
- Convex optimization in the natural parameter coordinate system θ. $\max_\theta l(\theta; x_1, ..., x_n) = \frac{1}{n} \sum_i (\langle t(x_i), \theta \rangle - F(\theta)) \equiv \min_\theta F(\theta) - \langle \theta, \bar{t} \rangle$ (that is, $\nabla F(\hat{\theta}) = \bar{t}$)

Exponential families : Universal families !

Universal representations of "smooth" densities :

- mixtures of exponential families approximate any smooth density (mixtures of Gaussians)
- a single exponential family (possibly multimodal) approximates also any smooth density : Similar to approximations of functions by polynomials. We can choose the sufficient statistics in $(1, x, x^2, x^3, \ldots)$ and $(\log x, \log^2 x, \log^3 x, \ldots)$. But then $F(\theta)$ *not* in closed form :

$$\boxed{F(\theta) = \int_x \exp\left(\theta^\top t(x) + k(x)\right) \mathrm{d}\nu(x)}$$

(common problem met in practice not to have closed-form expression of F, Ising and Potts models, etc.)

Boltzmann-Gibbs distribution in statistical physics

Let $E(X;\theta)$ be an energy function.

$$p(X;\theta) = \frac{1}{Z(\theta)} \exp(-E(X;\theta))$$

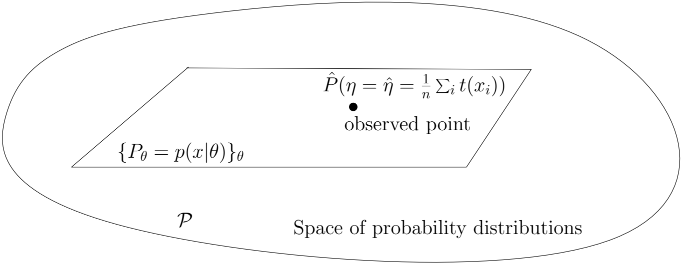
$Z(\theta)$ normalization factor (aka. partition function) :

$$Z(\theta) = \int_x \exp(-E(X;\theta)) \mathrm{d}\nu(x)$$

$$\boxed{F(\theta) = \log Z(\theta)}$$

The observed point $\hat{P}$ in information geometry

- $\{P_\theta\}_\theta$: a parametric (exponential family) model, <u>identifiable</u>
- View P_θ as a point on a manifold (dual coordinates θ and η)
- Observed point $\hat{P}$ with η-coordinate $\overline{t(x)} = \frac{1}{n}\sum_i t(x_i)$ (MLE)



We shall see later that $\hat{P}$ is m-projection of the empirical distribution on the e-flat...

MLE of exponential families [20]

- $\hat{\eta} = \overline{t(x)}$ but we would like $\hat{\theta} = (\nabla F^{-1})(\hat{\eta})$
- value of the maximum likelihood :

$$\boxed{l(\theta; x_1, ..., x_n) = F^*(\hat{\eta}) + \overline{k(x)}}$$

  $\overline{k(x)} = \frac{1}{n}\sum_{i=1}^n k(x_i)$
  F^* is neg-entropy
- When $F(\theta)$ not in closed-form : Contrastive Divergence (MCMC), score matching (Fisher divergence), etc.

II. Geometric structures of probability manifolds :

- (M, g)
- $(M, g, \nabla, \nabla^*) \Leftrightarrow (M, g, T)$

Population space & Parameter space

H. Hotelling [15] (1930), C. R. Rao [47] (1945)

- $\mathcal{P} = \{p(x|\theta) \mid \theta \in \Theta\}$ a *parametric family* of distributions, the population space,
- Θ, the parameter space of dimension D
- immersion $i(\theta) = p(x|\theta)$ from the parameter space to the population space :
  - i : one-to-one (model identifiability)
  - i of rank $\dim(\Theta) = D$:

$$\frac{\partial p(x|\theta)}{\partial \theta_1}, \ldots, \frac{\partial p(x|\theta)}{\partial \theta_D}$$

  ... are *linearly independent*

  - Geometric structures of SPD matrices when we consider the particular space $\{N(0, \Sigma) \mid \Sigma \succ 0\}$

Fisher information matrix (FIM)

- log-likelihood $l(\theta|x) = \log p(x|\theta)$, $\partial_i = \frac{\partial}{\partial\theta_i}$.
- Metric tensor, $D \times D$ matrix : $g = [g_{ij}] = \sum_{i,j} g_{ij} \mathrm{d}x_i \otimes \mathrm{d}x_j$ (tensor product)

$$\boxed{g_{ij} = \mathbb{E}_\theta[\partial_i l(\theta) \partial_j l(\theta)]}$$

- FIM can be rewritten *equivalently* as :

$$\boxed{g_{ij} = 4 \int_x \partial_i \sqrt{p(x|\theta)} \partial_j \sqrt{p(x|\theta)} \mathrm{d}x}$$

- g symmetric positive definite (SPD), non-degenerate when $\{\partial_i p(x|\theta)\}_i$ are linear independent (problem with mixture models where $\exists \theta, I(\theta) = 0$)

Fisher information matrix & Hessian

Negative expectation of the Hessian of the log-likelihood function :

$$
\begin{aligned}
g_{ij} &= \mathbb{E}_\theta[\partial_i l(\theta)\partial_j l(\theta)] \\
g_{ij} &= 4\int_x \partial_i \sqrt{p(x|\theta)}\partial_j \sqrt{p(x|\theta)}\mathrm{d}x \\
g_{ij} &= \boxed{-\mathbb{E}_\theta[\partial_i\partial_j l(\theta)]}
\end{aligned}
$$

For natural exponential families $p(x|\theta) = \exp(\langle\theta, x\rangle - F(\theta))$,

$$
\boxed{I(\theta) = \nabla^2 F(\theta) \succ 0}
$$

Fisher information : invariance and covariance

- Invariant under reparameterization of the sample space : X RV. with $p(x|\theta)$ and $Y = f(X)$ for an invertible transformation $f(\cdot)$ with density $\bar{p}(y|\theta)$.

$$\boxed{g_{ij}(\theta) = \bar{g}_{ij}(\theta)}$$

- Covariant under reparameterization of the parameter space : Let $\eta = \eta(\theta)$ be an invertible transformation with $\bar{p}_\eta(x) = p_{\eta(\theta)}(x)$

$$\boxed{\bar{g}_{ij}(\eta) = g_{kr}\mid_{\eta=\eta(\theta)} \frac{\partial\theta_k}{\partial\eta_i}\frac{\partial\theta_r}{\partial\eta_j}}$$

- sufficient statistics : $p(x|t,\theta) = p(x|t)$, non-deterministic Markov morphism transformations (statistical invariance).

Basics of Riemannian geometry

- (M, g) : Riemannian manifold
- $\langle \cdot, \cdot \rangle$, Riemannian *metric tensor* g : definite positive bilinear form on each tangent space $T_x M$ (depends smoothly on x)
- $\| \cdot \|_x$: $\|u\| = \langle u, u \rangle^{1/2}$: Associated norm in $T_x M$
- $\rho(x, y)$: metric distance between two points on the manifold M (length space)

$$\rho(x, y) = \inf \left\{ \int_0^1 \|\dot{\gamma}(t)\| \,\mathrm{d}t, \quad \gamma \in C^1([0,1], M), \quad \gamma(0) = x, \quad \gamma(1) = y \right\}$$

- Shortest paths (length space)
- but technically parallel transport wrt. Levi-Civita metric connection ∇_{LC}.

Basics of Riemannian geometry : Exponential map

- Local map from the *tangent space* T_xM to the *manifold* defined with geodesics (wrt ∇).

$$\forall x \in M, D(x) \subset T_xM : D(x) = \{v \in T_xM : \gamma_v(1) \text{ is defined}\}$$

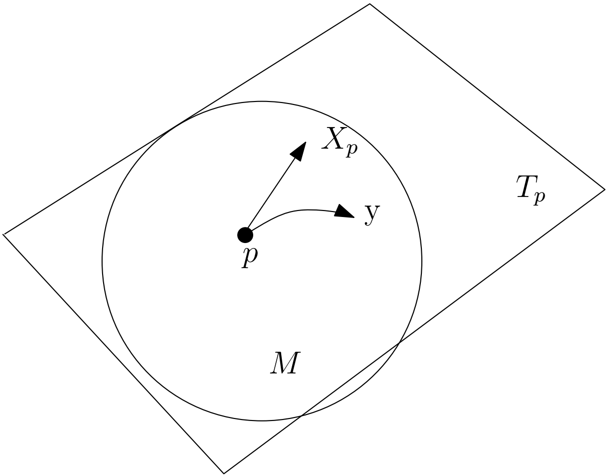
with γ_v maximal (i.e., largest domain) geodesic with $\gamma_v(0) = x$ and $\gamma_v'(0) = v$.

- Exponential map :

$$\begin{aligned} \exp_x(\cdot) \quad &: \quad D(x) \subseteq T_xM \rightarrow M \\ & \quad\quad \exp_x(v) = \gamma_v(1) \end{aligned}$$

D is *star-shaped*.

Riemannian geometry : Exponential and Logarithmic maps



$$\begin{aligned} \exp &: \quad y \in M \rightarrow X_p \in T_p \\ \log = \exp^{-1} &: \quad X_p \in T_p \rightarrow y \in M \end{aligned}$$

Basics of Riemannian geometry : Geodesics

- *Geodesic* : smooth path which locally minimizes the distance between two points.
- Given a vector $v \in T_x M$ with base point x, there is a unique geodesic started at x with speed v at time 0 : $t \mapsto \exp_x(tv)$ or $t \mapsto \gamma_t(v)$.
- Geodesic on $[a, b]$ is *minimal* if its length is less or equal to others. For *complete* M (i.e., $\exp_x(v)$), taking $x, y \in M$, there exists a *minimal* geodesic from x to y in time 1.
  $\gamma_\cdot(x, y) : [0, 1] \to M$, $t \mapsto \gamma_t(x, y)$ with the conditions $\gamma_0(x, y) = x$ and $\gamma_1(x, y) = y$.
- $U \subseteq M$ is convex if for any $x, y \in U$, there exists a unique minimal geodesic $\gamma_\cdot(x, y)$ in M from x to y. Geodesic *fully lies* in U and depends smoothly on x, y, t.

Basics of Riemannian geometry : Geodesics

- Geodesic $\gamma(x,y)$: locally minimizing curves linking x to y
- Speed vector $\gamma'(t)$ parallel along γ :

$$\boxed{\frac{D\gamma'(t)}{\mathrm{d}t} = \nabla_{\gamma'(t)}\gamma'(t) = 0}$$

- When manifold M embedded in $\mathbb{R}^d$, acceleration is normal to tangent plane :

$$\gamma''(t) \perp T_{\gamma(t)}M$$

- $\|\gamma'(t)\| = c$, a constant (say, unit).

$\Rightarrow$ Parameterization of curves with constant speed (otherwise, you get the trace of the geodesic only...)

Basics of Riemannian geometry : Geodesics and means

Constant speed geodesic $\gamma(t)$ so that $\gamma(0) = x$ and $\gamma(\rho(x,y)) = y$ (constant speed 1, the unit of length).

$$\boxed{x\#_t y = m = \gamma(t) : \rho(x,m) = t \times \rho(x,y)}$$

For example, in the Euclidean space :

$$x\#_t y = (1-t)x + ty = x + t(y-x) = m$$

$$\rho_E(x,m) = \|t(y-x)\| = t\|y-x\| = t \times \rho(x,y), t \in [0,1]$$

$\Rightarrow$ m **interpreted as a** **<u>mean</u>** **(barycenter) between** x **and** y

Basics of Riemannian geometry : Injectivity radius

Diffeomorphism from the tangent space to the manifold

- *Injectivity radius* $\mathrm{inj}(M)$: largest $r > 0$ such that **for all** $x \in M$, the map $\exp_x(\cdot)$ restricted to the open ball in $T_x M$ with radius r is an embedding.
- *Global injectivity radius* : infimum of the injectivity radius over all points of the manifold.

Important for navigating back and forth from $T_x M$ to M (extrinsic/intrinsic computing)...

Riemannian geometry of population spaces

- Consider (M, g) with $g = I(\theta)$, Hotelling (1930), Rao (1945). Fisher information matrix is unique up to a constant (for statistical invariance).
- Geometry of multinomials is spherical (on the orthant)
- For univariate location-scale families, hyperbolic geometry or Euclidean geometry (location only)

$$p(x|\mu,\sigma) = \frac{1}{\sigma} p_0\left(\frac{x-\mu}{\sigma}\right), \quad X = \mu + \sigma X_0$$

(Normal, Cauchy, Laplace, Student t-, etc.)

Tangent planes, tangent bundles, vector fields

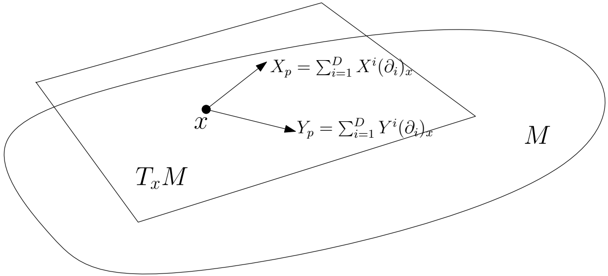
- T_p : tangent plane at p
- TM, tangent bundle
- vector field = global section of the tangent bundle
- Mahalanobis metric distance on tangent planes T_x :

$$\boxed{M_Q(p,q) = \sqrt{(p-q)^\top Q(x)(p-q)}}$$

  axioms of the metric for $Q(x) = g(x) \succ 0$ (SPD).
- Rao's distance between close points amounts to $\rho \simeq \sqrt{2\mathrm{KL}} = \sqrt{\mathrm{SKL}}$. For exponential families, $\rho \simeq \mathrm{Mahalanobis} = \sqrt{\Delta\theta^\top I(\theta)\Delta\theta}$.

Tangent plane : basis vectors

- $(\partial_i)_x = \left(\frac{\partial}{\partial\theta^i}\right)_x$
- $X_x = \sum_{i=1}^{D} X^i (\partial_i)_x$
- Define proper metric tensor : $g_{ij}(x) = g_x(\partial_i, \partial_j) > 0$



α-representations and parameterizations of the tangent planes

$$f_\alpha(u) = \begin{cases} \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}}, & \alpha \neq 1 \\ \log u, & \alpha = 1. \end{cases}$$

- $\alpha = -1 : p(x|\theta) \rightarrow f_{-1}(p(x|\theta)) = p(x|\theta)$: usual parameterization of the tangent plane $T_x^{(-1)} M$ with basis $\partial_i^{(-1)} = \partial_i$.
- $\alpha = 0$: square root representation : $p(x|\theta) \rightarrow f_0(p(x|\theta)) = 2\sqrt{p(x|\theta)}$. $\partial^{(0)}$ perpendicular to θ, identified with the tangent plane $T_x^{(0)} M$.
- $\alpha = 1$: logarithmic representation : $p(x|\theta) \rightarrow f_1(p(x|\theta)) = \log p(x|\theta)$. $\partial^{(1)} = \partial_i f_1(p(x|\theta)) = \frac{1}{p(x|\theta)} \partial_i p(x|\theta)$

Tangent planes are invariant objects : do not depend on the α-representation.

Extrinsic Computational Geometry on tangent planes

- Tensor $g = Q(x) \succ 0$ defines smooth inner product $\langle p, q \rangle_x = (p-q)^\top Q(x)(p-q)$ that induces a normed distance : $d_x(p,q) = \|p-q\|_x = \sqrt{(p-q)^\top Q(x)(p-q)}$
- Mahalanobis metric distance on tangent planes :

$$\boxed{\Delta_\Sigma(X_1, X_2) = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)} = \sqrt{\Delta\mu^\top \Sigma^{-1} \Delta\mu}}$$

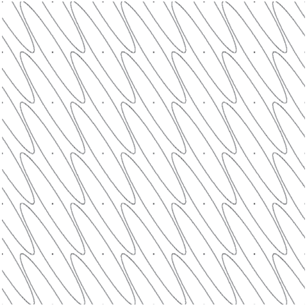
- Cholesky decomposition $\Sigma = LL^\top$, lower triangular matrix L :

$$\boxed{\Delta(X_1, X_2) = D_E(L^{-1}\mu_1, L^{-1}\mu_2)}$$

- **Computing on tangent planes = Euclidean computing on transformed points $x' \leftarrow L^{-1}x$.**
  Extrinsic vs *intrinsic* computations.

Riemannian Mahalanobis metric tensor (Σ^{-1}, PSD)

$$\rho(p_1, p_2) = \sqrt{(p_1 - p_2)^\top \Sigma^{-1} (p_1 - p_2)}, \quad g(p) = \Sigma^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$



non-conformal geometry : $g(p) \neq f(p) I$
(Visualization with Tissot indicatrix)

Normal/Gaussian family and 2D location-scale families

- Fisher Information Matrix (FIM) :

$$\boxed{I(\theta) = \left[I_{i,j}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log p(x|\theta)\frac{\partial}{\partial\theta_j}\log p(x|\theta)\right]\right]} = \mathbb{E}_\theta[\partial_i l \partial_j l]$$

- FIM for univariate normal/multivariate spherical distributions :

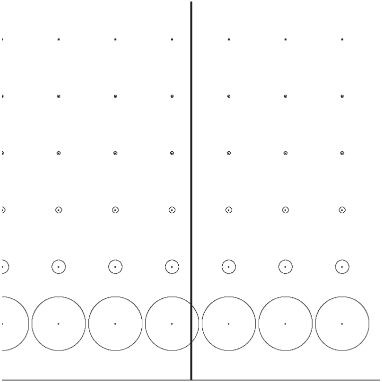
$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

$I(\mu,\sigma) = \mathrm{diag}\left(\frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2}, \frac{2}{\sigma^2}\right)$

- $\rightarrow$ amount to Poincaré metric $\frac{\mathrm{d}x^2+\mathrm{d}y^2}{y^2}$, hyperbolic geometry in upper half plane/space.

Riemannian Poincaré upper plane metric tensor (conformal)

$$\cosh \rho(p_1, p_2) = 1 + \frac{\|p_1 - p_2\|^2}{2 y_1 y_2}, \quad g(p) = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{bmatrix} = \frac{1}{y^2} I$$



<u>conformal</u> : $g(p) = \frac{1}{y^2} I$

Matrix SPD spaces and hyperbolic geometry

Symmetric Positive Definite matrices $M : \forall x \neq 0, x^\top M x > 0$.

- 2D $\mathrm{SPD}(2)$ matrix space has dimension $d = 3$: A positive cone.

$$\mathrm{SPD}(2) \left\{ (a, b, c) \in \mathbb{R}^3 : a > 0, \quad ab - c^2 > 0 \right\}$$

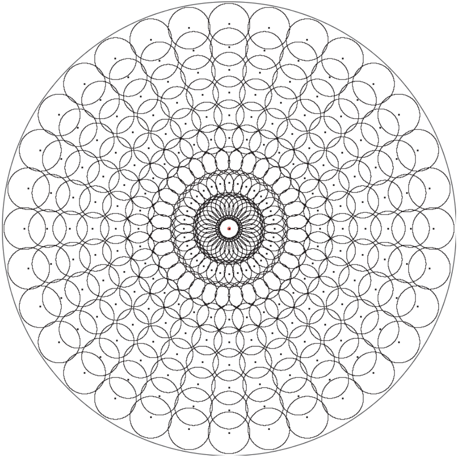
- Can be *peeled into sheets* of dimension 2, each sheet corresponding to a *constant value of the determinant* of the elements

$$\boxed{\mathrm{SPD}(2) = \mathrm{SSPD}(2) \times \mathbb{R}^+}$$

where $\mathrm{SSPD}(2) = \{a, b, c = \sqrt{1 - ab}) : a > 0, ab - c^2 = 1\}$

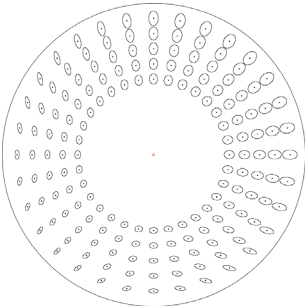
- Mapping $M(a, b, c) \to \mathbb{H}^2$:
  - $\left(x_0 = \frac{a+b}{2} \geq 1, x_1 = \frac{a-b}{2}, x_2 = c\right)$ in hyperboloid model [39]
  - $z = \frac{a-b+2ic}{2+a+b}$ in Poincaré disk [39].

Riemannian Poincaré disk metric tensor (conformal)



→ often used in Human Computer Interfaces, network routing (embedding trees), etc.

Riemannian Klein disk metric tensor (non-conformal)



- recommended for "computing space" since geodesics are straight line segments
- Klein is also **conformal at the origin** (so we can perform translation from and back to the origin via Möbius transform.)
- Geodesics passing through O in the Poincaré disk are straight (so we can perform translation from and back to the origin)

Riemannian geometry : Optimization on the manifold with the natural gradient [1]

Numerical optimization on manifolds :

- defined on a manifold, generalize Euclidean gradient $\nabla_x f(x) = (\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_D} f(x))$.
- natural gradient respects intrinsic geometry of the manifold :

$$\boxed{\tilde{\nabla}_\theta f(\theta) = (I(\theta))^{-1} \times \nabla_\theta f(\theta)}$$

  (Euclidean geometry : $I(\theta) = I$.)
- invariant under changes of the parameterization (natural gradient = contravariant form of the gradient)
- Information-geometric optimization (IGO), black-box optimization

Jeffrey's prior from volume element

- Volume of the manifold :

$$v(M) = \int \sqrt{|g(\theta)|}\mathrm{d}\theta < \infty$$

- Consider the prior distribution :

$$\boxed{q(\theta) = \frac{1}{v(M)}\sqrt{|g(\theta)|}}$$

- invariant under reparameterization
- Bayesian statistics (and other $\pm\alpha$-volume element in IG : $|g(\theta)|^{\frac{1\pm\alpha}{2}}$)

Affine differential geometry : dual connections ∇ and ∇^* coupled with a metric g

Connections $\prod$ and covariant derivatives ∇

- Connections $\prod$ set correspondences between vectors in tangent spaces T_p and T_q. When manifold M is embedded in $\mathbb{R}^d$, there exists a natural correspondence. Otherwise, connections $\prod$ need to be formally defined.
- Covariant derivatives ∇ : differentiation of a vector field Y in the direction of another vector field X, yielding a vector field $Z = \nabla_X Y$.

- Connections and covariant derivatives induce the same geometric structure. Yield notions of geodesics, flatness/curvature, parallelness, torsion.
- Riemannian structure (M, g) has an induced metric connection $\nabla_g = \nabla_{\mathrm{LC}} = \nabla^{(0)}$, called the Levi-Civita connection.

Connections and parallel transport

- $\prod_{p,q}$ a connection from T_p to T_q
$$\prod_{p,q} : T_p \to T_q$$
so that $v \in T_p$ yields $w = \prod_{p,q}(v) \in T_q$
- from linear isomorphism between tangent spaces of neighboring points to tangent points between arbitrary points by integrating along a curve $\gamma_{p,q}$ connecting p with q.
- d^3 coefficients $\Gamma_{ijk}(p)$ required for defining $\prod$.
- Vector field X along γ with $X(t + \mathrm{d}t) = \prod_{\gamma(t),\gamma(t+\mathrm{d}t)} X(t)$. We say vector fields $\{X(t) \mid t\}$ along γ are <u>parallel</u> with respect to the connection $\prod$. Parallel transport.

Covariant derivatives ∇

∇ : differentiation of a vector field Y in the direction of another vector field X, yielding a vector field $Z = \nabla_X Y$.

$$\nabla : V(M) \times V(M) \rightarrow V(M)$$

Properties ∇ should have :

$$\begin{aligned} \nabla_{f_1 X_1 + f_2 X_2} Y &= f_1 \nabla_{X_1} Y + f_2 \nabla_{X_2} Y \\ \nabla_X (Y_1 + Y_2) &= \nabla_X Y_1 + \nabla_X Y_2 \\ \nabla_X (fY) &= f \nabla_X Y + (Xf) Y \end{aligned}$$

Linear combinations of covariant derivatives is a covariant derivative

Vector field parallel to a curve

Vector field $Y \in V(M)$ is <u>∇-parallel to a curve</u> $\gamma(t)$:

$$\forall t, \forall X \in V(M), \quad \nabla_{\dot{\gamma}(t)} Y = 0$$

Geodesics in differential geometry

Curves γ on (M, ∇) such that

$$\boxed{\forall t, \quad \nabla_{\dot\gamma(t)} \dot\gamma(t) = 0}$$

Affine coordinate system and flat connection

In general, specify a connection/covariant ∇ by D^3 coefficients :

$$\boxed{\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \partial_k, \quad \forall i,j,k \in \{1,...,D\}}$$

(M,∇), θ a coordinate system.

θ is an affine coordinate system iff :

- Vector fields $\{\partial_i = \frac{\partial}{\partial\theta_i}\}$ are parallel in M
- Equivalent to $\forall i,j, \quad \nabla_{\partial_i}\partial_j = 0$
- Equivalent to $\forall i,j,k, \quad \Gamma_{ij}^k = 0$ (Christoffel symbols)

When there exists an affine coordinate system for (M,∇), we say that M is flat.

Metric connection : Special case of Levi-Civita connection $\nabla_{LC} = \nabla^{(0)}$

Given (M, g), there exists a unique metric connection, the Levi-Civita connection :

- $\Gamma_{ij}^{k} = \frac{\partial_i g_{jk} + \partial_j g_{kj} - \partial_k g_{ij}}{2}$
- and we have $g(\nabla^{(0)}_{\partial_i} \partial j, \partial_k) = \Gamma_{ij}^{k}$.
- Parallel transport of tangent vectors preserves the inner product.
- Therefore angles are kept, henceforth "parallel transport"

Autoparallel submanifold

$N \subset M$ of (M, N) is autoparallel :

- Property on the tangent bundle TN

$$\boxed{\forall X, Y \in TN, \quad \nabla_X Y \in TN}$$

- Parallel (∇)-transport of tangent vectors for N are tangent vectors of N.
- Notion of "hyperplanes" in differential geometry
- **For an affine connection with coordinate system θ, equivalent to an affine subspace of $\theta \in \mathbb{R}^D$.**

Differential-geometric structures : Summary

Manifold M

g

$\prod, \nabla$

Riemannian manifold
metric tensor g (inner product)
(angle, orthogonality)
(M, g)

connection $\prod$
covariant derivatives ∇
$\prod \Leftrightarrow \nabla$
parallel transport
(flatness, autoparallel)
(M, ∇)

Differential structure (M, g, ∇)

Levi-Civita connection
$\nabla_{\mathrm{LC}} = \nabla(g)$ (coefficients Γ_{ij}^k)
geodesics preserves $\langle \cdot, \cdot \rangle$
$\rho(P, Q)$ metric distance
(shortest paths)

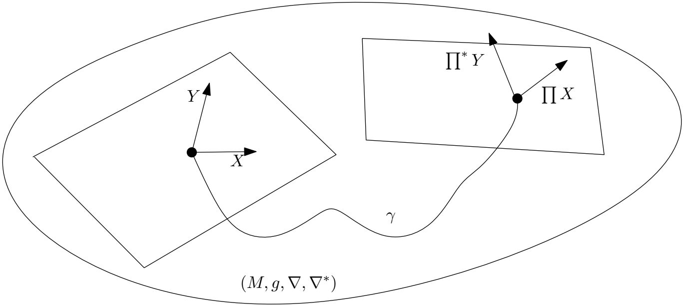
Dual connections (M, g, ∇, ∇^*)

Dually affine connections

- Two affine connections $\prod$ and $\prod^*$ (and covariant derivatives ∇ and ∇^*)
- Property of inner product :

$$\boxed{\langle X, Y\rangle_g = \langle \prod X, \prod^* Y\rangle_g}$$

- Riemannian geometry : $\prod = \prod^*$



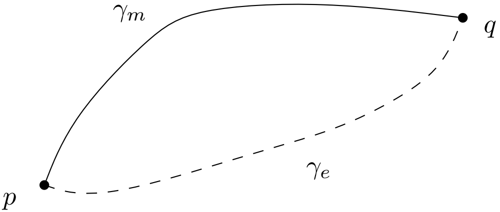
$$\langle X, Y\rangle_g = \langle \prod X, \prod^* Y\rangle_g$$

Dually affine connections : e-connection and m-connection

Exponential e-geodesics and mixture m-geodesics for probability densities :

$$
\begin{aligned}
\gamma_m(p,q,\alpha) &: \quad r(x,\alpha) = \alpha p(x) + (1-\alpha) q(x) \\
\gamma_e(p,q,\alpha) &: \quad \log r(x,\alpha) = \alpha p(x) + (1-\alpha) q(x) - F(t)
\end{aligned}
$$

$$
\boxed{\nabla^{(e)}_{\dot\gamma_e} \dot\gamma_e(t) = 0, \quad \nabla^{(m)}_{\dot\gamma_m} \dot\gamma_m(t) = 0}
$$



Flat but not Riemannian flat : e-flat and m-flat.

Dually α-affine connections

$$\alpha \in \mathbb{R}, \quad \boxed{\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*}$$

- $\nabla = \nabla^e$ or ∇^m
- Dually-coupled affine connections : $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$
- $\alpha = 0$: $\nabla^{(0)} = \frac{\nabla+\nabla^*}{2} = \nabla_{\mathrm{LC}}$, Levi-Civita metric connection (self-dual $\nabla^{(0)} = \nabla^{(0)^*}$)
- 0-geometry is Riemannian geometry (often curved but not for isotropic Gaussians)

Dually flat orthogonal coordinate systems

- θ- and η-coordinate systems
- partial derivatives : $\partial_i = \frac{\partial}{\partial \theta_i}$, $\partial^i = \frac{\partial}{\partial \eta^i}$
- $\langle \partial_i, \partial^j \rangle = \delta_{ij}$ (biorthogonal coordinate systems)
- metric-coupled connection :

$$\boxed{X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle}$$

- $\Gamma_{ijk}(\theta) = \Gamma^*_{ijk}(\eta) = 0$

This is key advantage over the Riemannian (∇_{LC}) structure : Geodesics are known in closed form with the affine coordinate systems. Line segments in either the θ- or η-coordinate systems.

Dually flat manifolds from a convex function F

Canonical geometry induced by strictly convex and differentiable convex function F.

- Potential functions : F and Legendre convex conjugate $G = F^*$
- Dual coordinate systems : $\theta = \nabla F^*(\eta)$ and $\eta = \nabla F(\theta)$.
- Metric tensor g : written equivalently using the two coordinate systems :

$$\boxed{g_{ij}(\theta) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} F(\theta), \quad g^{ij}(\eta) = \frac{\partial^2}{\partial\eta_i \partial\eta_j} G(\eta)}$$

- Divergence from Young's inequality of convex conjugates :

$$\boxed{D(P : Q) = F(\theta(P)) + F^*(\eta(Q)) - \langle \theta(P), \eta(Q) \rangle}$$

  This is a Bregman divergence in disguise - :) ...
- exponential family : $p(x|\theta) = \exp(\langle \theta, x \rangle - F(\theta))$
  Terminology : F=cumulant function, G=negative entropy

Geometry induced from a potential function

F a strictly convex potential function

$$g_{ij} = \frac{\partial^2 F}{\partial_i \partial_j}$$

$$\Gamma_{ijk}^{(\alpha)} = \frac{1-\alpha}{2} \frac{\partial^3 F}{\partial_i \partial_j \partial_k}$$

Dually coupled $\pm\alpha$-connections (affine torsion-free, Kurose [16], 1994) :

$$\boxed{\forall X, Y, Z \in V(M), \quad Xg(Y,Z) = g(\nabla_X^{(\alpha)} Y, Z) + g(Y, \nabla_X^{(\alpha)} Z)}$$

Curvature : $\kappa = \frac{1-\alpha^2}{4}$ (and hence $\alpha = \pm 1 \Leftrightarrow \kappa = 0$, flat)

Bregman divergences : An old friend from the optimization community

Bregman divergences

$$\boxed{D_F(p:q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle}$$

includes...

- squared Euclidean distance : $F(x) = \langle x, x \rangle$, and squared Mahalanobis $F(x) = x^\top Q x$ (only symmetric divergences)
- (extended) Kullback-Leibler divergence : $F(x) = \sum_i x_i \log x_i - x_i$ (Shannon information),

$$\boxed{\mathrm{eKL}(p:q) = \sum_i \left(p_i \log \frac{p_i}{q_i} + q_i - p_i \right)}$$

- $F(x) = -\sum_i \log x_i$ (Burg information), Itakura-Saito divergence :

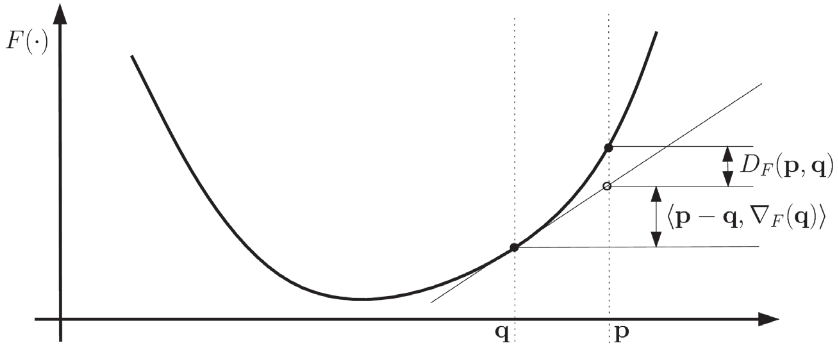
$$\mathrm{IS}(p:q) = \sum_i \left(\frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1 \right)$$

- and many others !

Bregman divergence : Geometric interpretation (I)

Potential function F, graph plot $\mathcal{F} : (x, F(x))$.

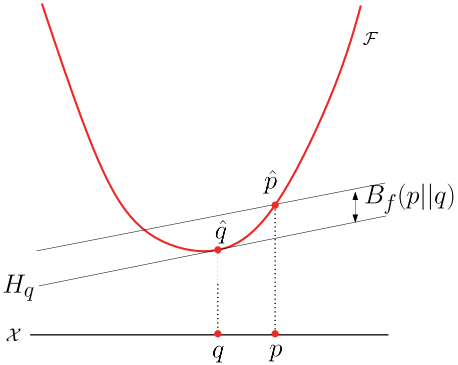
$$\boxed{D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle}$$



Bregman divergence : Geometric interpretation (II)

Potential function f, graph plot $\mathcal{F}: (x, f(x))$.

$$B_f(p||q) = f(p) - f(q) - (p-q)f'(q)$$

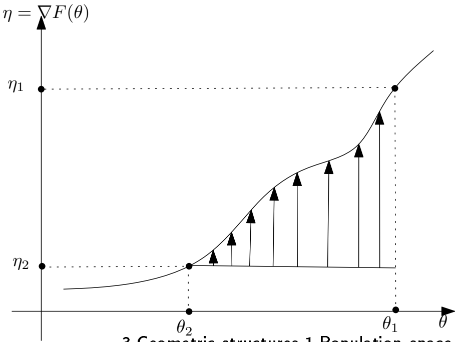


$B_f(.||q)$: vertical distance between the hyperplane H_q tangent to $\mathcal{F}$ at lifted point $\hat{q}$, and the translated hyperplane at $\hat{p}$.

Bregman divergence : Geometric interpretation (III)

Bregman divergence and path integrals

$$
\begin{aligned}
B(\theta_1 : \theta_2) &= F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle, & (1)\\
&= \int_{\theta_2}^{\theta_1} \langle \nabla F(t) - \nabla F(\theta_2), \mathrm{d}t \rangle, & (2)\\
&= \int_{\eta_1}^{\eta_2} \langle \nabla F^*(t) - \nabla F^*(\eta_1), \mathrm{d}t \rangle, & (3)\\
&= B^*(\eta_2 : \eta_1) & (4)
\end{aligned}
$$



Dual Bregman divergences & canonical divergence [34]

For P and Q belonging to the <u>same</u> exponential families]

$$
\begin{aligned}
\mathrm{KL}(P:Q) &= E_P\left[\log\frac{p(x)}{q(x)}\right] \geq 0 \\
&= B_F(\theta_Q:\theta_P) = B_{F^*}(\eta_P:\eta_Q) \\
&= F(\theta_Q) + F^*(\eta_P) - \langle\theta_Q,\eta_P\rangle \\
&= A_F(\theta_Q:\eta_P) = A_{F^*}(\eta_P:\theta_Q)
\end{aligned}
$$

with θ_Q (natural parameterization) and $\eta_P = E_P[t(X)] = \nabla F(\theta_P)$ (moment parameterization).

$$
\mathrm{KL}(P:Q) = \underbrace{\int p(x)\log\frac{1}{q(x)}\mathrm{d}x}_{H^\times(P:Q)} - \underbrace{\int p(x)\log\frac{1}{p(x)}\mathrm{d}x}_{H(p)=H^\times(P:P)}
$$

Shannon cross-entropy and entropy of EF [34] :

$$
\begin{aligned}
H^\times(P:Q) &= F(\theta_Q) - \langle\theta_Q,\nabla F(\theta_P)\rangle - E_P[k(x)] \\
H(P) &= F(\theta_P) - \langle\theta_P,\nabla F(\theta_P)\rangle - E_P[k(x)] \\
H(P) &= -F^*(\eta_P) - E_P[k(x)]
\end{aligned}
$$

III. Principle of Maximum Entropy (MaxEnt)

Maximum entropy (MaxEnt)

Underconstrained optimization problem (Jaynes's principle for maximum ignorance) :

$$\max_p H(p) = \sum_x p(x) \log \frac{1}{p(x)}$$

$$\begin{aligned}
\sum_x p(x) t_i(x) &= m_i, \quad \forall i \in \{1, ..., D\} \\
p(x) &\geq 0, \quad \forall x \in \{1, ..., n\} \\
\sum_x p(x) &= 1
\end{aligned}$$

- Maximizing a concave function (H) subject to linear constraints
- Convex optimization problem.

A more general setting for MaxEnt

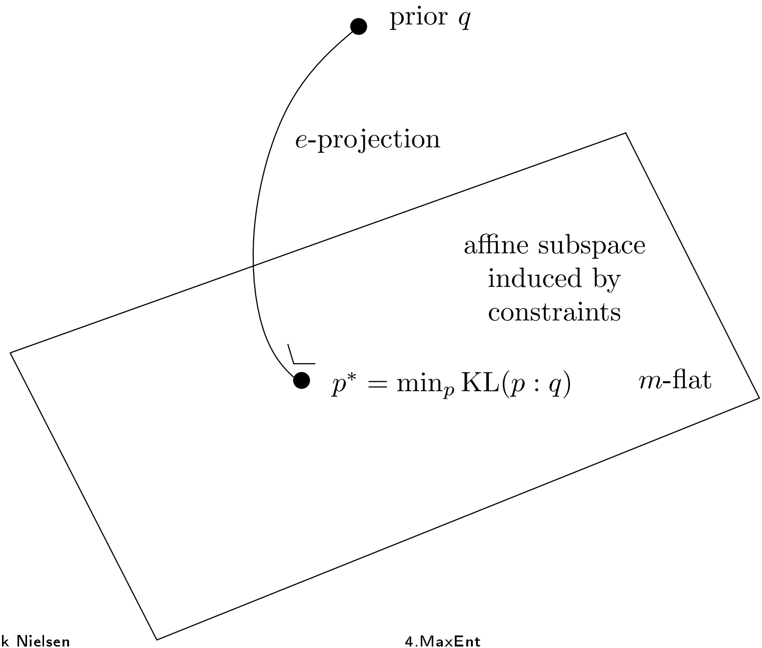
Given a prior q, find the closest distribution which satifies the linear constraints :

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

$$\begin{aligned} \sum_x p(x) t_i(x) &= m_i, \quad \forall i \in \{1, ..., D\} \\ p(x) &\geq 0, \quad \forall x \in \{1, ..., n\} \\ \sum_x p(x) &= 1 \end{aligned}$$

$\rightarrow$ **Maximum entropy when $q = \frac{1}{n}$, the uniform prior**

An illustration...



Analytic solution : exponential families !

Using Lagrange multipliers θ with $t(x) = (t_1(x), ..., t_D(x))$:

$$\boxed{p(x) = \frac{1}{Z(\theta)} \exp\left(\langle \theta, t(x) \rangle\right) q(x)}$$

... but Lagrange multipliers usually not in explicit form.

- Canonical exponential families : $\exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))$
- Prior q gives the carrier measure $q(x) = e^{k(x)}$
- $Z(\theta)$ is the normalizer
- called Gibbs distribution, Maxwell-Boltzmann distribution in statistical mechanics

A toy example for MaxEnt

- A distribution p with support $\mathbb{R}$ has $\mathbb{E}[X] = 3$ and $\mathbb{E}[X^2] = 25$. Which distribution should we choose for p?
- $t(x) = (x, x^2)$ defines the univariate Gaussian family of distributions.
- So we choose $p \sim N(\mu = 3, \sigma = 5)$

in general not so easy if we are given $E[X^k]$ for $k > 2$... uniqueness but no closed form...

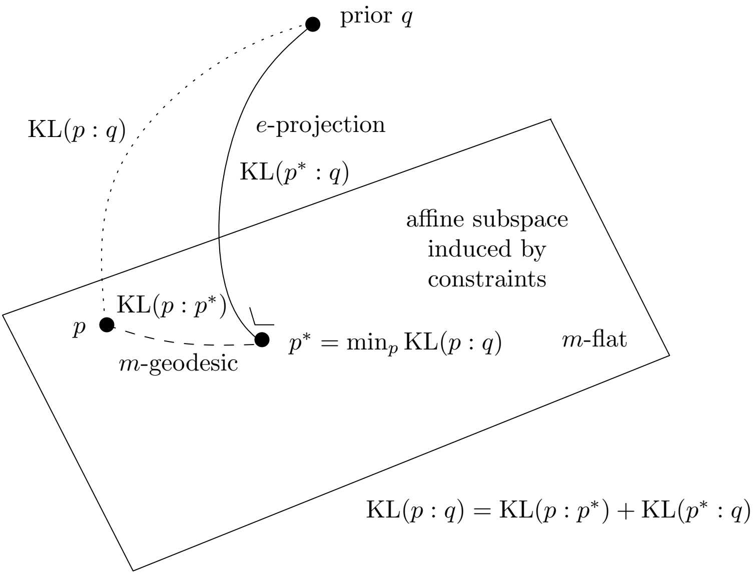
Another insightful proof

Any other distribution $p \neq p^*$ satisfying the constraints is such that $\mathrm{KL}(p:q) > \mathrm{KL}(p^*:q)$.
Consider the difference $\mathrm{KL}(p:q) - \mathrm{KL}(p^*:q)$:

$$\begin{aligned}
&= \sum_x p(x)\log\frac{p(x)}{q(x)} - \sum_x p^*(x)\log\frac{p^*(x)}{q(x)}\\
&\ldots\\
&= \sum_x p(x)\log\frac{p(x)}{q(x)} - \sum_x p(x)\log\frac{p^*(x)}{q(x)}\\
&= \sum_x p(x)\log\frac{p(x)}{p^*(x)} = \mathrm{KL}(p:p^*) > 0
\end{aligned}$$

Pythagorean relation : $\boxed{\mathrm{KL}(p:q) = \mathrm{KL}(p:p^*) + \mathrm{KL}(p^*:q)}$

An illustration of MaxEnt with prior $q(x)$...

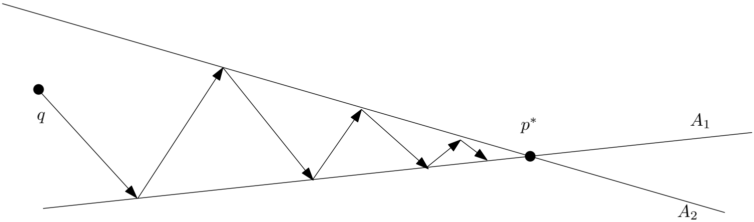


Pythagoras' theorem...

Computing information projections easily

- Project the prior q onto $A = \{p \mid \mathbb{E}_p[t_i(x)] = m_i, \ \forall i \in \{1, ..., D\}\}$. Let $A_i = \{p \mid \mathbb{E}_p[t_i(x)] = m_i\}$
- Let $t = 0$ and $p_0 = q$
- Repeat until convergence (within a threshold) :
$$p_{t+1} = \text{I-projection of } p_t \text{ onto } L_{t \mod D}$$
- 1D projection easy : Find θ_i such that $F_{\neq i}(\theta_i) = m_i$ (for example, using line search)

Cyclic (line search) 1D information projections



IV. Information projection

Projections : e-projection and m-projection

$$\boxed{\nabla^{(e)} = \nabla^{(1)}, \quad \nabla^{(m)} = \nabla^{(-1)}}$$

- e-projection q is **unique** if $M \subseteq S$ is m-flat and minimizes the m-divergence $\mathrm{KL}(\boxed{q} : p)$.
- m-projection q is **unique** if $M \subseteq S$ is e-flat and minimizes the e-divergence $\mathrm{KL}(p : \boxed{q})$.

KL and reverse KL are α-divergences for $\alpha = \pm 1$...

MLE as min KL : Information projection

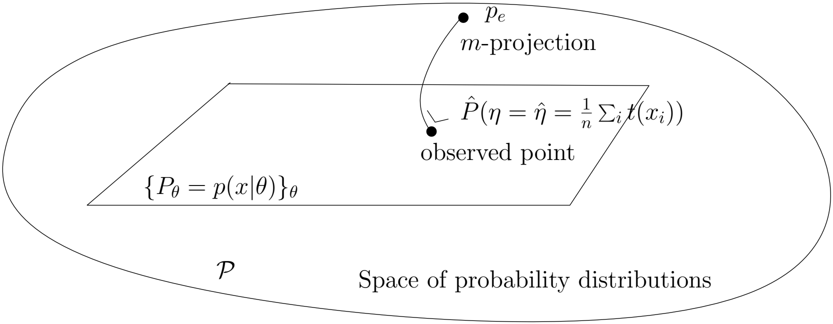
- Empirical distribution : $p_e(x) = \frac{1}{n}\sum_i \delta(x - x_i)$.
- p_e is absolutely continuous with respect to $p_\theta(x)$

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

Log-likelihood function

$$l(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta) = \langle \log p(x|\theta) \rangle_{p_e}$$

Empirical distribution : $p_e(X) = \frac{1}{n}\sum_{i=1}^{n} \delta(X - X(i))$
MLE = m-projection from p_e to the model submanifold



Nested and curved exponential families

$\mathcal{P}(\theta)$ an exponential family

- <u>nested EFs</u> : Fix some parameters $\theta = (\theta_{\mathrm{fixed}}, \theta_{\mathrm{variable}})$. Then $\mathcal{P}_{\theta_{\mathrm{fixed}}}(\theta_{\mathrm{variable}})$ is a nested exponential family. Get stratified EFs with uni-order EF easy to handle algorithmically (Legendre)
- <u>curved EFs</u> : $\mathcal{C}(\gamma) \subseteq \mathcal{P}(\theta)$ embedded in $\mathcal{P}(\theta)$. Example : $\{N(\mu, \mu^2) \mid \mu \in \mathbb{R}\}$ is embedded into $\{N(\mu, \sigma^2)\}$.

MLE for curved exponential families

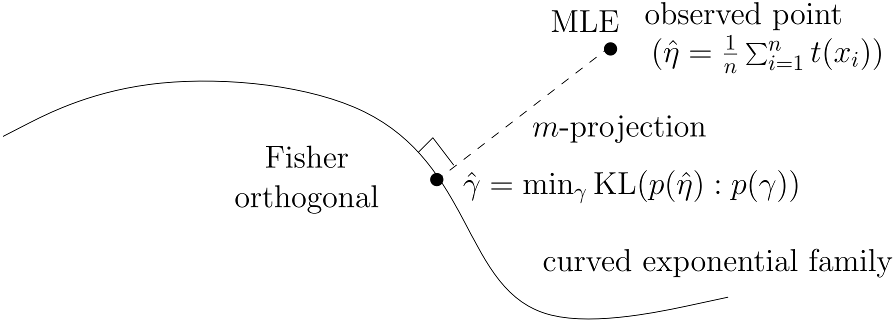
Entropy $H(\theta) = -E_\theta[\log p(x|\theta)] = F(\theta) - \langle \theta, \nabla F(\theta) \rangle = -F^*(\eta)$ (when $k(x) = 0$, otherwise add $-E[k(x)]$).

$$D(p(\hat{\eta}) : p(\gamma)) = -H(\hat{\eta}) - \frac{1}{n} \log L(\gamma)$$

$$\boxed{\max_\gamma L(\gamma) \equiv \min_\gamma D(p(\hat{\eta}) : p(\gamma))}$$

$\hat{\gamma}$ is the m-projection of the observation point (with η-coordinate $\hat{\eta}$)

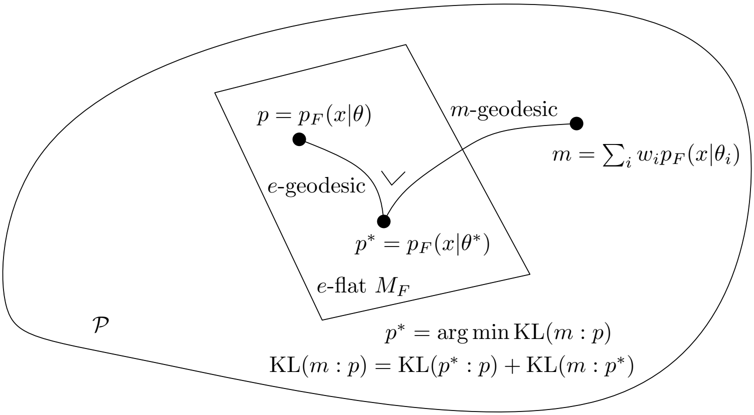
Illustration : MLE for curved exponential families



information loss, statistical curvature.

Simplifying a mixture model into a single component [48]

m-projection of the mixture model m onto the e-flat (exponential family manifold) : Best single distribution that approximates an exponential family mixture is found by taking the center of mass of the moment parameters : $\bar{\eta} = \sum_i w_i \eta_i$.



Kullback-Leibler divergence and Fisher information

$$\boxed{\mathrm{KL}(\theta + \Delta\theta : \theta) \approx \frac{1}{2}\theta^\top I(\theta)\theta}$$

... square Mahalanobis induced locally by half squared Mahalanobis distance for the Fisher information matrix.

$$g_{ij}(\theta_0) = \left.\frac{\partial^2}{\partial\theta^i \partial\theta^j}\right|_{\theta=\theta_0} \mathrm{KL}(P(\theta)\|P(\theta_0))$$

This holds for <u>f-divergences</u> $\int p(x) f(\frac{q(x)}{p(x)})\mathrm{d}\nu(x)$ (that includes Kullback-Leibler divergence) : divergence inducing a metric proportional to Fisher information (Part II).

Additive Shannon/Rényi versus non-additive Tsallis entropies

- additive (Shannon-Rényi)

$$\boxed{H(P \times Q) = H(P) + H(Q)}$$

- non-additive (Tsallis) $T_q(X) = \frac{1}{q-1}(1 - \sum_i p_i^q)$

$$\boxed{T_q(X \times Y) = T_q(X) + T_q(Y) + (1-q) T_q(X) T_q(Y)}$$

- Both can be unified with Sharma-Mittal [37] 2-parameter family of entropies
- Sharma-Mittal entropies, cross-entropies and relative entropies are known in closed-form for exponential families.

Part I : Summary

- Fisher information (Cramér-Rao lower bound) & sufficiency (1922)
- Differential geometry of population spaces :
  - Fisher-Rao geometry (Hotelling, 1930) : $g(\theta) = I(\theta)$
  - Dually-coupled connection geometry (1970's-1980's, Cencov, Amari, Kurose) : $(M, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$, or (M, g, T)
  - Dually-flat manifold from a potential function F and canonical divergence (=Bregman divergence).
- Exhaustivity : Bregman divergences=canonical divergences in dually flat spaces
- Maximum entropy principle (Shannon entropy & exponential families)
- Information-geometric projections : MLE from empirical distribution, MLE in curved exponential families, and in mixture simplification.

Part II : Algorithms & Space of spheres

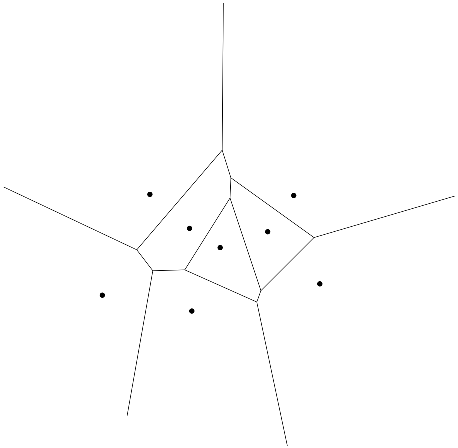
Brief historical review of Computational Geometry (CG)

- Three research periods :
  1. **Geometric algorithms** :
     Voronoi/Delaunay, minimum spanning trees, data-structures for proximity queries
  2. **Geometric computing** :
     robustness, algebraic degree of predicates, programs that work/scale !
  3. **Computational topology** (global geometry) :
     simplicial complexes, filtrations, input=distance matrix
     $\rightarrow$ paradigm of Topological Data Analysis (TDA)
- Showcasing libraries for CG software :
  - CGAL `http://www.cgal.org/`
    Geometry Factory `http://geometryfactory.com/`
  - Gudhi `https://project.inria.fr/gudhi/`
    Ayasdi `http://www.ayasdi.com/`

Basics of Euclidean Computational Geometry : Voronoi diagrams and dual Delaunay complexes

Euclidean (ordinary) Voronoi diagrams

$\mathcal{P} = \{P_1, ..., P_n\}$: n distinct **point generators** in Euclidean space $\mathbb{E}^d$



$$\boxed{V(P_i) = \{X : D_E(P_i, X) \leq D_E(P_j, X),\ \forall j \neq i\}}$$

Voronoi diagram = <u>cell complex</u> $V(P_i)$'s with their faces

Voronoi diagrams from bisectors and $\cap$ halfspaces

Bisectors

$$\boxed{\mathrm{Bi}(P,Q) = \{X : D_E(P,X) = D_E(Q,X)\}}$$

$\rightarrow$ are **hyperplanes** in Euclidean geometry
Voronoi cells as halfspace intersections :

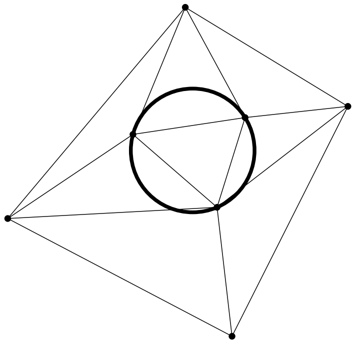
$$\boxed{V(P_i) = \{X : D_E(P_i,X) \leq D_E(P_j,X),\ \forall j \neq i\} = \cap_{i=1}^{n} \mathrm{Bi}^{+}(P_i,P_j)}$$

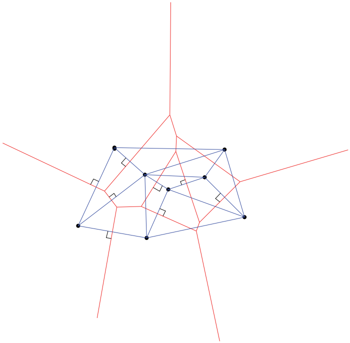
$$D_E(P,Q) = \|\theta(P) - \theta(Q)\|_2 = \sqrt{\sum_{i=1}^{d} (\theta_i(P) - \theta_i(Q))^2}$$

$\theta(P) = p$: Cartesian coordinate system with $\theta_j(P_i) = p_i^{(j)}$.

$\Rightarrow$ Many applications of Voronoï diagrams : crystal growth, codebook/quantization, molecule interfaces/docking, motion planning, etc.

Voronoi diagrams and **dual** Delaunay simplicial complex





- **Empty sphere** property, **max min angle** triangulation, etc
- Voronoi & dual **Delaunay triangulation**
  $\rightarrow$ non-degenerate point set = no $(d+2)$ points co-spherical
- Duality : $\boxed{\text{Voronoi } k\text{-face} \Leftrightarrow \text{Delaunay } (d-k)\text{-simplex}}$
- Bisector $\mathrm{Bi}(P, Q)$ **perpendicular** $\perp$ to segment $[PQ]$

Voronoi & Delaunay : Complexity and algorithms

- Combinatorial complexity : $\Theta(n^{\lceil \frac{d}{2} \rceil})$ ($\rightarrow$ quadratic in 3D)
  matched for points on the moment curve : $t \mapsto (t, t^2, .., t^d)$
- Construction : $\Theta(n \log n + n^{\lceil \frac{d}{2} \rceil})$, optimal
- some output-sensitive algorithms but...
- $\Omega(n \log n + f)$, **not yet optimal output-sensitive algorithms**.

Population spaces : Hotelling (1930) [15] & Rao (1945) [47]

Birth of **differential-geometric methods in statistics**.

- Fisher information matrix (non-degenerate positive definite) can be used as a (smooth) *Riemannian metric tensor* g.
- Distance between two populations indexed by θ_1 and θ_2 : Riemannian distance (metric length)

First applications in statistics :

- Fisher-Hotelling-Rao (FHR) geodesic distance used in **classification** : Find the closest population to a given set of populations
- Used in **tests of significance** (null versus alternative hypothesis), power of a test : $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is false})$
  $\rightarrow$ define surfaces in population spaces

Rao's distance (1945, introduced by Hotelling 1930 [15])

- Infinitesimal squared length element :

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

- Geodesic and distance are **hard to explicitly calculate** :

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

Rao's distance not known in closed-form for multivariate normals

- Advantages : Metric property of ρ + many tools of differential geometry [3] : Riemannian Log/Exp tangent/manifold mapping

Extrinsic Computational Geometry on tangent planes

- Tensor $g = Q(x) \succ 0$ defines smooth inner product $\langle p, q\rangle_x = (p-q)^\top Q(x)(p-q)$ that induces a normed distance : $d_x(p,q) = \|p-q\|_x = \sqrt{(p-q)^\top Q(x)(p-q)}$
- Mahalanobis metric distance on tangent planes :

$$\boxed{\Delta_\Sigma(X_1, X_2) = \sqrt{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)} = \sqrt{\Delta\mu^\top \Sigma^{-1} \Delta\mu}}$$

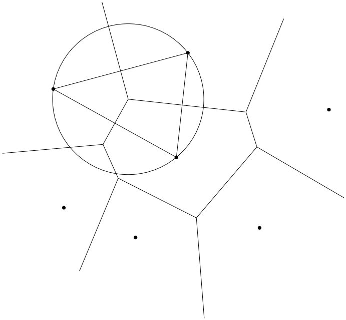
- Cholesky decomposition $\Sigma = LL^\top$

$$\boxed{\Delta(X_1, X_2) = D_E(L^{-1}\mu_1, L^{-1}\mu_2)}$$

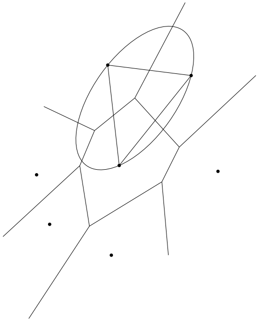
- CG on tangent planes = **ordinary CG** on transformed points $x' \leftarrow L^{-1}x$. *Extrinsic* vs *intrinsic* means [11]

Mahalanobis Voronoi diagrams on tangent planes (extrinsic)

In statistics, covariance matrix Σ account for both **correlation** and dimension (feature) **scaling**



$\Leftrightarrow$



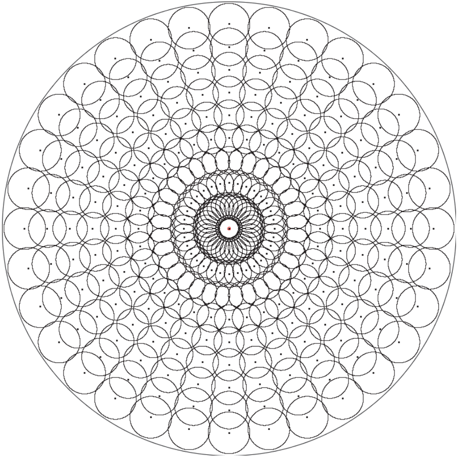
Dual structure $\equiv$ anisotropic Delaunay triangulation
$\Rightarrow$ "empty circumellipse" property (Cholesky decomposition)

Riemannian manifolds : Choice of equivalent models ?

Many **equivalent** models of hyperbolic geometry :

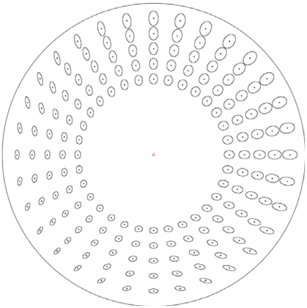
- **Conformal** (good for visualization since we can measure angles) versus **non-conformal** (computationally-friendly for geodesics) models.
- Convert *equivalently* to other models of hyperbolic geometry : Poincaré disk, upper half space, hyperboloid, **Beltrami** hemisphere, etc.

Riemannian Poincaré disk metric tensor (conformal)



$\rightarrow$ often used in Human Computer Interfaces, network routing (embedding trees), etc.

Riemannian Klein disk metric tensor (non-conformal)



- recommended for "computing space" since geodesics are straight line segments
- Klein is also **conformal at the origin** (so we can perform translation from and back to the origin)
- Geodesics passing through O in the Poincaré disk are straight (so we can perform translation from and back to the origin)

Hyperbolic Voronoi diagrams [35, 40]

In arbitrary dimension, $\mathbb{H}^d$

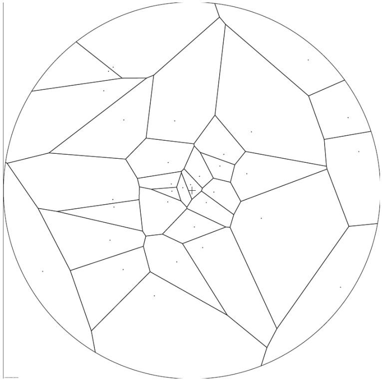
- In Klein disk, the hyperbolic Voronoi diagram amounts to a **clipped affine Voronoi diagram**, or a **clipped power diagram** with efficient clipping algorithm [6].
- then convert to other models of hyperbolic geometry : Poincaré disk, upper half space, hyperboloid, **Beltrami** hemisphere, etc.
- **Conformal** (good for visualization) versus **non-conformal** (good for computing) models.

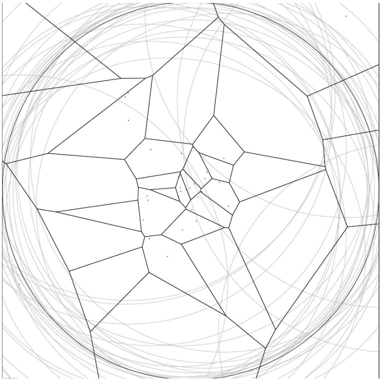
Hyperbolic Voronoi diagrams [35, 40]

Hyperbolic Voronoi diagram in Klein disk = clipped power diagram.
Power distance :

$$\boxed{\|x-p\|^2 - w_p}$$

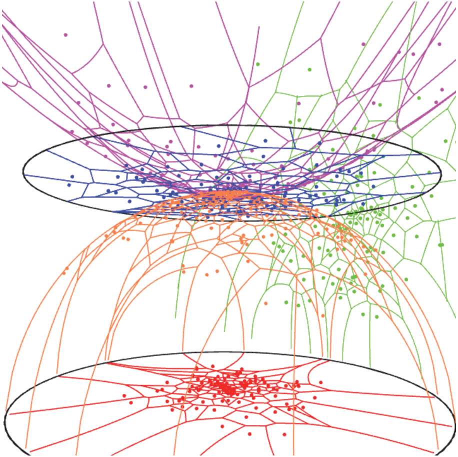
$\rightarrow$ additively weighted ordinary Voronoi = ordinary CG





Hyperbolic Voronoi diagrams [35, 40]

5 common models of the abstract hyperbolic geometry



`https://www.youtube.com/watch?v=i9IUzNxeH4o` (5 min. video)
ACM Symposium on Computational Geometry (SoCG'14)

Voronoi diagrams in dually affine information geometry

Dually flat space construction from convex functions F

- Convex and strictly differentiable function $F(\theta)$ admits a Legendre-Fenchel convex conjugate $F^*(\eta)$:

$$\boxed{F^*(\eta) = \sup_\theta (\theta^\top \eta - F(\theta)), \quad \nabla F(\theta) = \eta = (\nabla F^*)^{-1}(\theta)}$$

- Young's inequality gives rise to **canonical divergence** [19] :

$$F(\theta) + F^*(\eta') \geq \theta^\top \eta' \Rightarrow \boxed{A_{F,F^*}(\theta, \eta') = F(\theta) + F^*(\eta') - \theta^\top \eta'}$$

- Writing using **single coordinate system**, get dual **Bregman divergences** :

$$\begin{aligned} B_F(\theta_p : \theta_q) &= F(\theta_p) - F(\theta_q) - (\theta_p - \theta_q)^\top \nabla F(\theta_q) \\ &= B_{F^*}(\eta_q : \eta_p) = A_{F,F^*}(\theta_p, \eta_q) = A_{F^*,F}(\eta_q : \theta_p) \end{aligned}$$

- dual affine coordinate systems with geodesics "straight" : $\eta = \nabla F(\theta) \Leftrightarrow \theta = \nabla F^*(\eta)$. Tensor $\boxed{g(\theta) = g^*(\eta)}$

Dual divergence/Bregman dual bisectors [7, 32, 36]

Bregman sided (reference) bisectors related by convex duality :

$$\begin{aligned}\mathrm{Bi}_F(\theta_1,\theta_2) &= \{\theta\in\Theta\ | B_F(\theta:\theta_1)=B_F(\theta:\theta_1)\}\\ \mathrm{Bi}_{F^*}(\eta_1,\eta_2) &= \{\eta\in H\ | B_{F^*}(\eta:\eta_1)=B_{F^*}(\eta:\eta_1)\}\end{aligned}$$

Right-sided bisector : $\rightarrow$ θ-hyperplane, η-hypersurface

$$H_F(p,q)=\{x\in\mathcal{X}\mid B_F(x:\boxed{p})=B_F(x:\boxed{q})\}.$$

$$\boxed{H_F:\langle\nabla F(p)-\nabla F(q),x\rangle+(F(p)-F(q)+\langle q,\nabla F(q)\rangle-\langle p,\nabla F(p)\rangle)=0}$$

Left-sided bisector : $\rightarrow$ θ-hypersurface, η-hyperplane

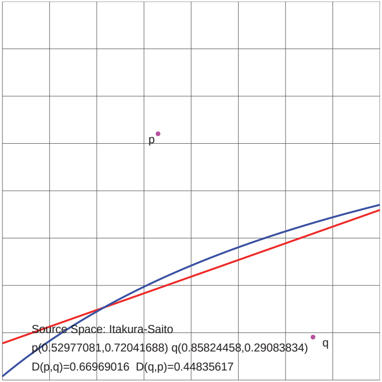
$$H'_F(p,q)=\{x\in\mathcal{X}\mid B_F(\boxed{p}:x)=B_F(\boxed{q}:x)\}$$

$$\boxed{H'_F:\langle\nabla F(x),q-p\rangle+F(p)-F(q)=0}$$

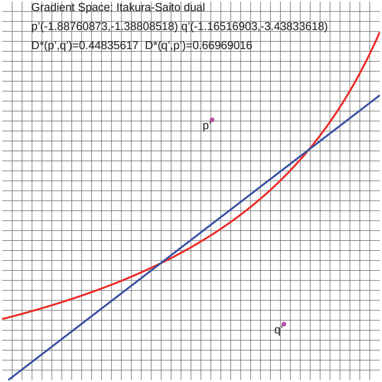
hyperplane = autoparallel submanifold of dimension $d-1$

Visualizing Bregman bisectors in θ- and η-coordinate systems

Primal coordinates θ
natural parameters



Dual coordinates η
expectation parameters



$\mathrm{Bi}(P, Q)$ and $\mathrm{Bi}^*(P, Q)$ can be expressed in either θ/η coordinate systems

Spaces of spheres : 1-to-1 mapping between d-spheres and $(d+1)$-hyperplanes using potential functions

Space of Bregman spheres and Bregman balls [7]

Dual sided Bregman balls (bounding Bregman spheres) :

$$\begin{aligned}\mathrm{Ball}_F^r(c,r) &= \{x \in \mathcal{X} \mid B_F(x:c) \leq r\}\\ \mathrm{Ball}_F^l(c,r) &= \{x \in \mathcal{X} \mid B_F(c:x) \leq r\}\end{aligned}$$

Legendre duality :

$$\boxed{\mathrm{Ball}_F^l(c,r) = (\nabla F)^{-1}(\mathrm{Ball}_{F^*}^r(\nabla F(c),r))}$$

Illustration for Itakura-Saito divergence, $F(x) = -\log x$

Generalized law of cosines and generalized Pythagoras' theorem

- Generalized law of cosines : θ=angle made at Q by the ∇-geodesic γ_{PQ} with the ∇^*-geodesic γ^*_{QR}

$$\boxed{D(P:R) = D(P:Q) + D(Q:R) - \underbrace{\|\dot{\gamma}_{PQ}\| \|\dot{\gamma^*}_{QR}\| \cos(\theta)}_{\langle \theta_P - \theta_Q, \eta_R - \eta_Q \rangle}}$$

- Euclidean law of cosines when $D = B_F$ for $F = \frac{1}{2}x^\top x$:

$$\|\overrightarrow{PR}\|^2 = \|\overrightarrow{PQ}\|^2 + \|\overrightarrow{QR}\|^2 - 2\|\overrightarrow{PQ}\| \|\overrightarrow{QR}\| \cos\theta$$

- Generalized Pythagoras' theorem when $\theta = \frac{\pi}{2}$:

$$\boxed{D(P:R) = D(P:Q) + D(Q:R)}$$

amount to check that $\cos\theta = 0$, that is $\langle \theta_P - \theta_Q, \eta_R - \eta_Q \rangle = 0$

Space of Bregman spheres : Lifting map [7]

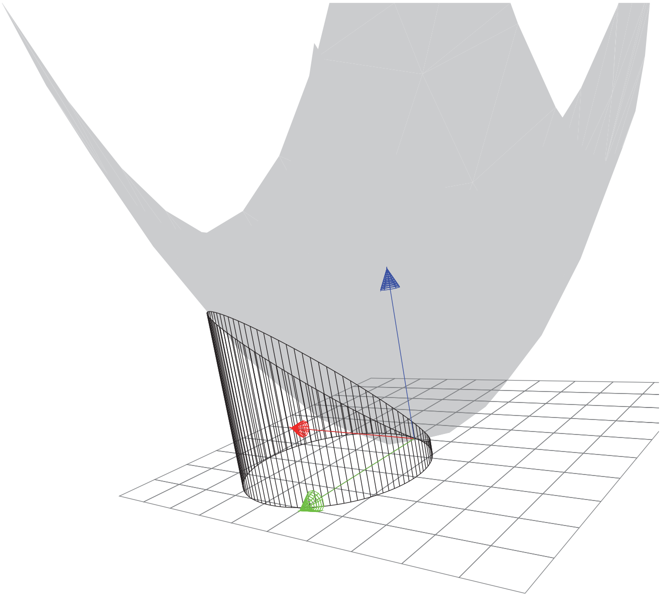
$\mathcal{F} : x \mapsto \hat{x} = (x, F(x))$, hypersurface in $\mathbb{R}^{d+1}$, potential function

H_p : Tangent hyperplane at $\hat{p}$, $z = H_p(x) = \langle x - p, \nabla F(p)\rangle + F(p)$

- Bregman sphere $\sigma \longrightarrow \hat{\sigma}$ with supporting hyperplane $H_\sigma : z = \langle x - c, \nabla F(c)\rangle + F(c) + r$.
  (// to H_c and shifted vertically by r)
  $\hat{\sigma} = \mathcal{F} \cap H_\sigma$.
- intersection of any hyperplane H with $\mathcal{F}$ projects onto $\mathcal{X}$ as a Bregman sphere :

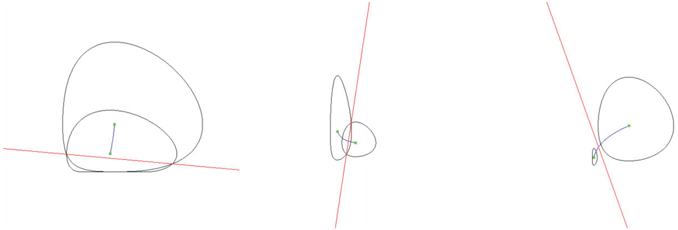
$$H : z = \langle x, a\rangle + b \rightarrow \sigma : \mathrm{Ball}_F(c = (\nabla F)^{-1}(a), r = \langle a, c\rangle - F(c) + b)$$

Lifting/Polarity : Potential function graph $\mathcal{F}$



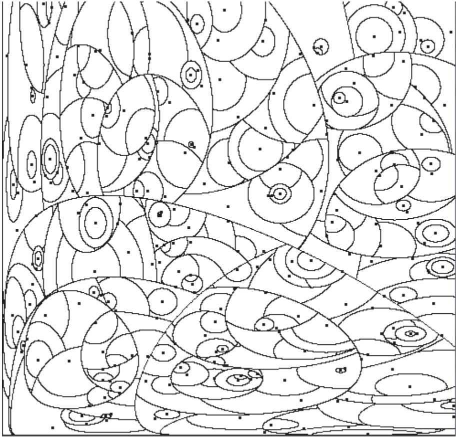
Space of Bregman spheres : Algorithmic applications [7]

- Vapnik-Chervonenkis dimension (VC-dim) is $d+1$ for the class of Bregman balls.
- Union/intersection of Bregman d-spheres from representational $(d+1)$-polytope [7]
- **Radical axis** of two Bregman balls is an **hyperplane** : Applications to Nearest Neighbor search trees like Bregman ball trees or Bregman vantage point trees [43].



Bregman proximity data structures [43]

Vantage point trees : partition space according to Bregman balls



Partitionning space with intersection of Kullback-Leibler balls
$\rightarrow$ efficient nearest neighbour queries in information spaces

Application : **Minimum Enclosing Ball** [30, 44]

To a hyperplane $H_\sigma = H(a,b) : z = \langle a, x\rangle + b$ in $\mathbb{R}^{d+1}$, corresponds a ball $\sigma = \mathrm{Ball}(c,r)$ in $\mathbb{R}^d$ with center $c = \nabla F^*(a)$ and radius :

$$r = \langle a, c\rangle - F(c) + b = \langle a, \nabla F^*(a)\rangle - F(\nabla F^*(a)) + b = \boxed{F^*(a) + b}$$

since $F(\nabla F^*(a)) = \langle \nabla F^*(a), a\rangle - F^*(a)$ (Young equality)

<u>SEB</u> : Find halfspace $H(a,b)^- : z \leq \langle a, x\rangle + b$ that contains all lifted points :

$$\begin{aligned} & \min_{a,b} r = F^*(a) + b, \\ \forall i \in \{1, ..., n\}, \quad & \langle a, x_i\rangle + b - F(x_i) \geq 0 \end{aligned}$$

→ **Convex Program (CP) with linear inequality constraints**
$F(\theta) = F^*(\eta) = \frac{1}{2}x^\top x$: CP → **Quadratic Programming** (QP) [14] used in SVM. **Smallest enclosing ball used as a primitive in SVM [49]**

Smallest Bregman enclosing balls [44, 29]

Algorithm 1: $\mathtt{BBCA}(\mathcal{P}, l)$.

$c_1 \leftarrow$ choose randomly a point in $\mathcal{P}$;

for $i = 2$ **to** $l - 1$ **do**

`// farthest point from` c_i `wrt.` B_F

$s_i \leftarrow \mathrm{argmax}_{j=1}^{n} B_F(c_i : p_j)$;

`// update the center: walk on the` η`-segment` $[c_i, p_{s_i}]_\eta$

$\boxed{c_{i+1} \leftarrow \nabla F^{-1}(\nabla F(c_i) \#_{\frac{1}{i+1}} \nabla F(p_{s_i}))}$;

end

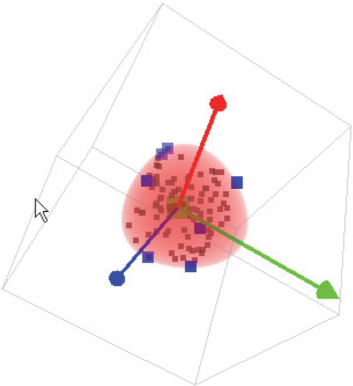
`// Return the SEBB approximation`

return $\mathrm{Ball}(c_l, r_l = B_F(c_l : X))$;

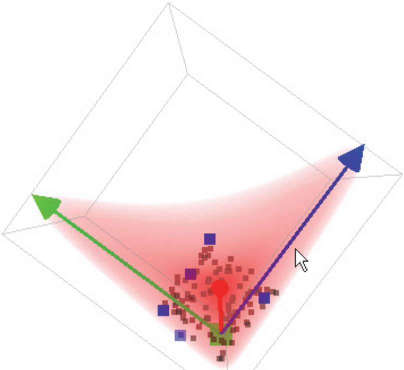
θ-, η-geodesic segments in dually flat geometry.

Smallest enclosing balls : Core-sets [44]

Core-set $\mathcal{C} \subseteq \mathcal{S}$: $\boxed{\mathrm{SOL}(\mathcal{S}) \leq \mathrm{SOL}(\mathcal{C}) \leq (1+\epsilon)\mathrm{SOL}(\mathcal{S})}$



extended Kullback-Leibler



Itakura-Saito

InSphere predicates wrt Bregman divergences [7]

Implicit representation of Bregman spheres/balls : consider $d+1$ support points on the boundary

- **Is x inside the Bregman ball defined by $d+1$ support points ?**

$$\mathrm{InSphere}(x; p_0, ..., p_d) = \begin{vmatrix} 1 & ... & 1 & 1 \\ p_0 & ... & p_d & x \\ F(p_0) & ... & F(p_d) & F(x) \end{vmatrix}$$

- sign of a $(d+2) \times (d+2)$ matrix determinant
- $\mathrm{InSphere}(x; p_0, ..., p_d)$ is negative, null or positive depending on whether x lies inside, on, or outside σ.

Smallest enclosing ball in Riemannian manifolds [3]

$c = a\#_t^M b$: point $\gamma(t)$ on the geodesic line segment $[ab]$ wrt M such that $\rho_M(a, c) = t \times \rho_M(a, b)$ (with ρ_M the metric distance on manifold M)

Algorithm 2: `GeoA`

$c_1 \leftarrow$ choose randomly a point in $\mathcal{P}$;

for $i = 2$ **to** l **do**

  `// farthest point from` c_i

  $s_i \leftarrow \mathrm{argmax}_{j=1}^n \rho(c_i, p_j)$;

  `// update the center: walk on the geodesic line segment` $[c_i, p_{s_i}]$

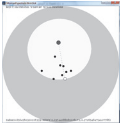
  $c_{i+1} \leftarrow c_i \#^M_{\frac{1}{i+1}} p_{s_i}$;

end

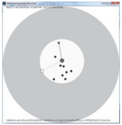
`// Return the SEB approximation`

return Ball$(c_l, r_l = \rho(c_l, \mathcal{P}))$;

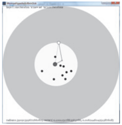
Approximating the smallest enclosing ball in hyperbolic space



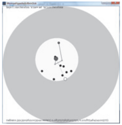
Initialization



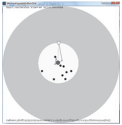
First iteration



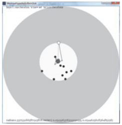
Second iteration



Third iteration



Fourth iteration

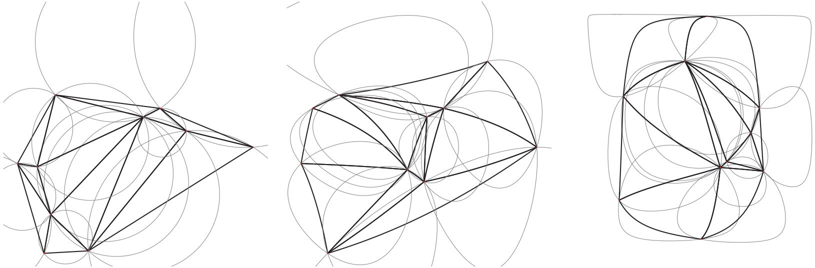


after 104 iterations

http://www.sonycsl.co.jp/person/nielsen/infogeo/RiemannMinimax/

Bregman dual regular/Delaunay triangulations

Embedded geodesic Delaunay triangulations+empty Bregman balls



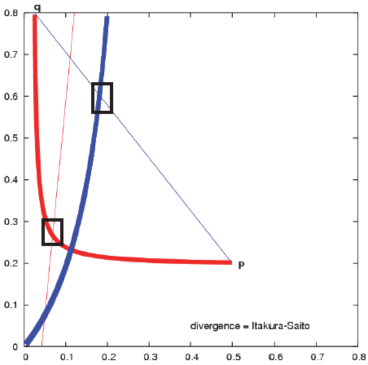
Delaunay Exponential Del. Hellinger-like Del.

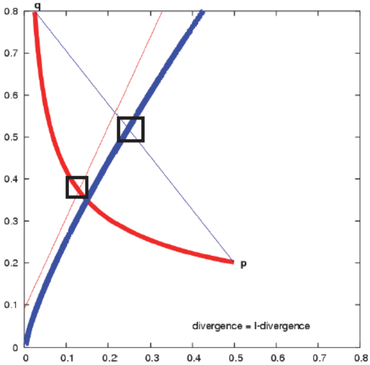
- empty Bregman sphere property,
- geodesic triangles : embedded Delaunay.

Dually orthogonal Bregman Voronoi & Triangulations

Ordinary Voronoi diagram is perpendicular to Delaunay triangulation :
Voronoi k-face $\perp$ Delaunay $d-k$-face

$$\begin{aligned} \mathrm{Bi}(P,Q) &\perp \gamma^*(P,Q) \\ \gamma(P,Q) &\perp \mathrm{Bi}^*(P,Q) \end{aligned}$$





Synthetic geometry : Exact characterization of the Bayesian error exponent but no closed-form known

Bayesian hypothesis testing, MAP rule and probability of error P_e

- Mixture $p(x) = \sum_i w_i p_i(x)$. **Task = Classify** x Which component ?
- Prior probabilities : $w_i = \mathbb{P}(X \sim P_i) > 0$ (with $\sum_{i=1}^n w_i = 1$)
- Conditional probabilities : $\mathbb{P}(X = x | X \sim P_i)$.

$$\mathbb{P}(X = x) = \sum_{i=1}^{n} \mathbb{P}(X \sim P_i)\mathbb{P}(X = x | X \sim P_i) = \sum_{i=1}^{n} w_i \mathbb{P}(X | P_i)$$

- Best rule = Maximum *A Posteriori* probability (MAP) rule :

$$\boxed{\mathrm{map}(x) = \mathrm{argmax}_{i \in \{1,\dots,n\}} \; w_i p_i(x)}$$

where $p_i(x) = \mathbb{P}(X = x | X \sim P_i)$ are the conditional probabilities.

- For $w_1 = w_2 = \frac{1}{2}$, probability of error
  $P_e = \frac{1}{2} \int \min(p_1(x), p_2(x)) \mathrm{d}x \leq \frac{1}{2} \int p_1(x)^\alpha p_2(x)^{1-\alpha} \mathrm{d}x$, for $\alpha \in (0, 1)$.
  Best exponent α^*

Error exponent for exponential families : duality EF⇔BD

- **Exponential families** have finite dimensional sufficient statistics : $\rightarrow$ Reduce n data to D statistics.

$$\boxed{\forall x \in \mathcal{X},\ \mathbb{P}(x|\theta) = \exp(\theta^\top t(x) - F(\theta) + k(x))}$$

  $F(\cdot)$: log-normalizer/cumulant/partition function, $k(x)$: auxiliary term for carrier measure.
- Maximum likelihood estimator (MLE) : $\nabla F(\hat\theta) = \frac{1}{n}\sum_i t(X_i) = \hat\eta$
- **Bijection between exponential families and Bregman divergences** :

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

  Exponential families are log-concave

Geometry of the best error exponent

On the exponential family manifold, **Chernoff α-coefficient** [8] :

$$\boxed{c_\alpha(P_{\theta_1} : P_{\theta_2}) = \int p_{\theta_1}^{\alpha}(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2))}$$

Skew Jensen divergence [26] on the natural parameters :

$$\mathbf{J}_{\mathbf{F}}^{(\alpha)}(\theta_{\mathbf{1}} : \theta_{\mathbf{2}}) = \alpha \mathbf{F}(\theta_{\mathbf{1}}) + (\mathbf{1} - \alpha)\mathbf{F}(\theta_{\mathbf{2}}) - \mathbf{F}(\theta_{\mathbf{12}}^{(\alpha)})$$

Chernoff information = **Bregman divergence for exponential families** :

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

Finding best error exponent α^* ?

Geometry of the best error exponent : binary hypothesis [23]

Chernoff distribution P^* :

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

e-geodesic :

$$G_e(P_1, P_2) = \left\{ E_{12}^{(\lambda)} \mid \theta(E_{12}^{(\lambda)}) = (1-\lambda)\theta_1 + \lambda\theta_2, \lambda \in [0,1] \right\},$$

m-bisector :

$$\mathrm{Bi}_m(P_1, P_2) : \left\{ P \mid F(\theta_1) - F(\theta_2) + \eta(P)^\top \Delta\theta = 0 \right\},$$

Optimal natural parameter of P^* :

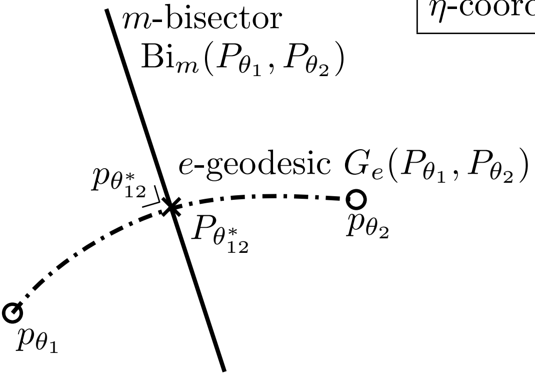
$$\theta^* = \theta_{12}^{(\alpha^*)} = \mathrm{argmin}_{\theta\in\Theta} B(\theta_1 : \theta) = \mathrm{argmin}_{\theta\in\Theta} B(\theta_2 : \theta).$$

$\rightarrow$ closed-form for order-1 family, or efficient bisection search.

Geometry of the best error exponent : binary hypothesis

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

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



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

Binary Hypothesis Testing : P_e bounded using Bregman divergence between Chernoff distribution and class-conditional distributions.

Clustering and Learning finite statistical mixtures

The distortion class of α-divergences

For $\alpha \in \mathbb{R} \neq \pm 1$, α-divergences [9] on positive arrays [51] :

- $\boxed{D_\alpha(p:q) \stackrel{\mathrm{eq}}{=} \sum_{i=1}^{d} \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}p^i + \frac{1+\alpha}{2}q^i - (p^i)^{\frac{1-\alpha}{2}}(q^i)^{\frac{1+\alpha}{2}}\right)}$ with $D_\alpha(p:q) = D_{-\alpha}(q:p)$ and in the limit cases $D_{-1}(p:q) = \mathrm{KL}(p:q)$ and $D_1(p:q) = \mathrm{KL}(q:p)$, where KL is the extended Kullback–Leibler divergence $\mathrm{KL}(p:q) \stackrel{\mathrm{eq}}{=} \sum_{i=1}^{d} p^i \log \frac{p^i}{q^i} + q^i - p^i$
- α-divergences belong to the class of Csiszár f-divergences $I_f(p:q) \stackrel{\mathrm{eq}}{=} \sum_{i=1}^{d} q^i f\left(\frac{p^i}{q^i}\right)$ with the following generator :

$$\boxed{f(t) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - t^{(1+\alpha)/2}\right), & \text{if } \alpha \neq \pm 1, \\ t \ln t, & \text{if } \alpha = 1, \\ -\ln t, & \text{if } \alpha = -1 \end{cases}}$$

Information monotonicity

Pythagoras' theorem for α-divergences [16]

Use $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ dually coupled connections with respect to g.

$$\boxed{Xg(Y,Z) = g(\nabla_X^{(\alpha)}, Z) + g(Y, \nabla_X^{(-\alpha)} Z)}$$

$$\gamma_{PQ}^{(\alpha)} \perp \gamma_{QR}^{(-\alpha)}$$

$$\boxed{D_\alpha(P:Q) = D_\alpha(P:Q) + D_\alpha(Q:R) - \kappa D_\alpha(P:Q) D_\alpha(Q:R)}$$

Curvature $\kappa = \frac{\alpha^2-1}{4}$.

Mixed divergences [42]

Defined on three parameters p, q and r :

$$\boxed{M_\lambda(p:q:r) \stackrel{\mathrm{eq}}{=} \lambda D(p:q) + (1-\lambda)D(q:r)}$$

for $\lambda \in [0,1]$.

Mixed divergences include :

- the **sided divergences** for $\lambda \in \{0,1\}$,
- the **symmetrized** (arithmetic mean) divergence for $\lambda = \frac{1}{2}$, or skew symmetrized for $\lambda \neq \frac{1}{2}$.

Symmetrizing α-divergences

$$\begin{aligned} S_\alpha(p,q) &= \frac{1}{2}\left(D_\alpha(p:q) + D_\alpha(q:p)\right) = S_{-\alpha}(p,q), \\ &= M_{\frac{1}{2}}(p:q:p), \end{aligned}$$

For $\alpha = \pm 1$, we get half of Jeffreys divergence :

$$\boxed{S_{\pm 1}(p,q) = \frac{1}{2}\sum_{i=1}^{d}(p^i - q^i)\log\frac{p^i}{q^i}}$$

- Centroids for symmetrized α-divergence usually not in closed form.
- How to perform center-based clustering without closed form centroids ?

Jeffreys positive centroid [22]

- Jeffreys divergence is symmetrized $\alpha = \pm 1$ divergences.
- The Jeffreys positive centroid $c = (c^1, ..., c^d)$ of a set $\{h_1, ..., h_n\}$ of n weighted positive histograms with d bins can be calculated component-wise exactly using the Lambert W analytic function :

$$\boxed{c^i = \frac{a^i}{W\left(\frac{a^i}{g^i}e\right)}}$$

  where $a^i = \sum_{j=1}^n \pi_j h_j^i$ denotes the coordinate-wise arithmetic weighted means and $g^i = \prod_{j=1}^n (h_j^i)^{\pi_j}$ the coordinate-wise geometric weighted means.
- The Lambert analytic function W [5] (positive branch) is defined by $W(x)e^{W(x)} = x$ for $x \geq 0$.
- $\rightarrow$ Jeffreys k-means clustering . But for $\alpha \neq 1$, how to cluster ?

Mixed α-divergences/α-Jeffreys symmetrized divergence

- Mixed α-divergence between a histogram x to **two** histograms p and q :

$$\begin{aligned} M_{\lambda,\alpha}(p : x : q) &= \lambda D_\alpha(p : x) + (1-\lambda) D_\alpha(x : q), \\ &= \lambda D_{-\alpha}(x : p) + (1-\lambda) D_{-\alpha}(q : x), \\ &= M_{1-\lambda,-\alpha}(q : x : p), \end{aligned}$$

- α-Jeffreys symmetrized divergence is obtained for $\lambda = \frac{1}{2}$:

$$\boxed{S_\alpha(p, q) = M_{\frac{1}{2},\alpha}(q : p : q) = M_{\frac{1}{2},\alpha}(p : q : p)}$$

- skew symmetrized α-divergence is defined by :

$$\boxed{S_{\lambda,\alpha}(p : q) = \lambda D_\alpha(p : q) + (1-\lambda) D_\alpha(q : p)}$$

Mixed divergence-based k-means clustering

k distinct seeds from the dataset with $l_i = r_i$.

Input: Weighted histogram set $\mathcal{H}$, divergence $D(\cdot, \cdot)$, integer $k > 0$, real $\lambda \in [0, 1]$;

Initialize left-sided/right-sided seeds $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k$;

repeat

  //Assignment

  for $i = 1, 2, ..., k$ **do**

    $\mathcal{C}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_\lambda(l_j : h : r_j)\}$;

  end

  // Dual-sided centroid relocation

  for $i = 1, 2, ..., k$ **do**

    $r_i \leftarrow \arg\min_x D(\mathcal{C}_i : x) = \sum_{h \in \mathcal{C}_i} w_j D(h : x)$;

    $l_i \leftarrow \arg\min_x D(x : \mathcal{C}_i) = \sum_{h \in \mathcal{C}_i} w_j D(x : h)$;

  end

until *convergence*;

Mixed α-hard clustering : $\mathrm{MAhC}(\mathcal{H}, k, \lambda, \alpha)$

Input: Weighted histogram set $\mathcal{H}$, integer $k > 0$, real $\lambda \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

  //Assignment

  for $i = 1, 2, ..., k$ **do**

    $\mathcal{A}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

  end

  // Centroid relocation

  for $i = 1, 2, ..., k$ **do**

    $r_i \leftarrow \left(\sum_{h\in\mathcal{A}_i} w_i h^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

    $l_i \leftarrow \left(\sum_{h\in\mathcal{A}_i} w_i h^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

  end

until *convergence*;

Coupled k-Means++ α-Seeding

Algorithm 3: Mixed α-seeding ; $\mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$

Input: Weighted histogram set $\mathcal{H}$, integer $k \geq 1$, real $\lambda \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} \leftarrow h_j$ with uniform probability ;

for $i = 2, 3, \ldots, k$ **do**

Pick at random histogram $h \in \mathcal{H}$ with probability :

$$\pi_{\mathcal{H}}(h) \stackrel{\text{eq}}{=} \frac{w_h M_{\lambda,\alpha}(c_h : h : c_h)}{\sum_{y \in \mathcal{H}} w_y M_{\lambda,\alpha}(c_y : y : c_y)}, \quad (5)$$

//where $(c_h, c_h) \stackrel{\text{eq}}{=} \arg\min_{(z,z) \in \mathcal{C}} M_{\lambda,\alpha}(z : h : z)$;

$\mathcal{C} \leftarrow \mathcal{C} \cup \{(h, h)\}$;

end

Output: Set of initial cluster centers $\mathcal{C}$;

$\rightarrow$ Guaranteed probabilistic bound. Just need to initialize ! No centroid computations

Learning MMs : A geometric hard clustering viewpoint

Learn the parameters of a mixture $m(x) = \sum_{i=1}^k w_i p(x|\theta_i)$

Maximize the **complete data likelihood**=clustering objective function

$$
\begin{aligned}
\max_{W,\Lambda} l_c(W,\Lambda) &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{i,j}\log(w_j p(x_i|\theta_j)) \\
&= \max_{\Lambda} \sum_{i=1}^{n} \max_{j=1}^{k} \log(w_j p(x_i|\theta_j)) \\
&\equiv \boxed{\min_{W,\Lambda} \sum_{i=1}^{n} \min_{j=1}^{k} D_j(x_i)},
\end{aligned}
$$

where $c_j = (w_j, \theta_j)$ (**cluster prototype**) and $D_j(x_i) = -\log p(x_i|\theta_j) - \log w_j$ are **potential distance-like functions**.

further attach to each cluster a different family of probability distributions.

Generalized k-MLE for learning statistical mixtures

Model-based clustering : Assignment of points to clusters :

$$\boxed{D_{w_j,\theta_j,F_j}(x) = -\log p_{F_j}(x;\theta_j) - \log w_j}$$

<u>k-GMLE</u> :

1. Initialize weight $W \in \Delta_k$ and family type $(F_1, ..., F_k)$ for each cluster
2. Solve $\min_\Lambda \sum_i \min_j D_j(x_i)$ (**center-based clustering** for W fixed) with potential functions : $D_j(x_i) = -\log p_{F_j}(x_i|\theta_j) - \log w_j$
3. **Solve family types** maximizing the MLE in each cluster $\mathcal{C}_j$ by choosing the parametric family of distributions $F_j = F(\gamma_j)$ that yields the best likelihood : $\min_{F_1=F(\gamma_1),...,F_k=F(\gamma_k)\in F(\gamma)} \sum_i \min_j D_{w_j,\theta_j,F_j}(x_i)$.
   $\forall l, \gamma_l = \max_j F_j^*(\hat{\eta}_l = \frac{1}{n_l}\sum_{x\in\mathcal{C}_l} t_j(x)) + \frac{1}{n_l}\sum_{x\in\mathcal{C}_l} k(x)$.
4. **Update weight** W as the cluster point proportion
5. Test for convergence and go to step 2) otherwise.

Drawback = biased, non-consistent estimator due to Voronoi support truncation.

Computing f-divergences for generic f : Beyond stochastic Monte-Carlo numerical integration

Ali-Silvey-Csiszár f-divergences

$$\boxed{I_f(X_1 : X_2) = \int x_1(x) f\left(\frac{x_2(x)}{x_1(x)}\right) \mathrm{d}\nu(x) \geq 0}$$

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	$\frac{1}{2}\int(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

Information monotonicity of f-divergences

Do coarse binning : from d bins to $k < d$ bins :

$$\mathcal{X} = \uplus_{i=1}^{k} A_i$$

Let $p^A = (p_i)_A$ with $p_i = \sum_{j \in A_i} p_j$.
<u>Information monotonicity</u> :

$$\boxed{D(p : q) \geq D(p^A : q^A)}$$

$\Rightarrow$ f-divergences are the *only* divergences preserving the information monotonicity.

f-divergences and higher-order Vajda χ^k divergences

$$\boxed{I_f(X_1 : X_2) = \sum_{k=0}^{\infty} \frac{f^{(k)}(1)}{k!} \chi_P^k(X_1 : X_2)}$$

$$\begin{aligned}
\chi_P^k(X_1 : X_2) &= \int \frac{(x_2(x) - x_1(x))^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \\
|\chi|_P^k(X_1 : X_2) &= \int \frac{|x_2(x) - x_1(x)|^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x),
\end{aligned}$$

are f-divergences for the generators $(u-1)^k$ and $|u-1|^k$.

- When $k=1$, $\chi_P^1(X_1 : X_2) = \int (x_1(x) - x_2(x)) \mathrm{d}\nu(x) = 0$ (never discriminative), and $|\chi_P^1|(X_1, X_2)$ is twice the total variation distance.
- χ_P^k is a signed distance

Affine exponential families

Canonical decomposition of the probability measure :

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

consider natural parameter space Θ affine (like multinomials).

$$\begin{aligned}
\mathrm{Poi}(\lambda) &: \quad p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \lambda > 0, x \in \{0, 1, ...\} \\
\mathrm{Nor}_I(\mu) &: \quad p(x|\mu) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}(x-\mu)^\top (x-\mu)}, \mu \in \mathbb{R}^d, x \in \mathbb{R}^d
\end{aligned}$$

Family	θ	Θ	$F(\theta)$	$k(x)$	$t(x)$	ν
Poisson	$\log\lambda$	$\mathbb{R}$	e^θ	$-\log x!$	x	ν_c
Iso.Gaussian	μ	$\mathbb{R}^d$	$\frac{1}{2}\theta^\top\theta$	$\frac{d}{2}\log 2\pi - \frac{1}{2}x^\top x$	x	ν_L

Higher-order Vajda χ^k divergences

The (signed) χ^k_P distance between members $X_1 \sim \mathcal{E}_F(\theta_1)$ and $X_2 \sim \mathcal{E}_F(\theta_2)$ of the same affine exponential family is ($k \in \mathbb{N}$) always bounded and equal to :

$$\boxed{\chi^k_P(X_1 : X_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \frac{e^{F((1-j)\theta_1 + j\theta_2)}}{e^{(1-j)F(\theta_1) + jF(\theta_2)}}}$$

For Poisson/Normal distributions, we get closed-form formula :

$$\begin{aligned} \chi^k_P(\lambda_1 : \lambda_2) &= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\lambda_1^{1-j}\lambda_2^{j} - ((1-j)\lambda_1 + j\lambda_2)}, \\ \chi^k_P(\mu_1 : \mu_2) &= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\frac{1}{2}j(j-1)(\mu_1-\mu_2)^\top(\mu_1-\mu_2)}. \end{aligned}$$

f-divergences : Analytic formula [18]

- $\lambda = 1 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$, f-divergence (Theorem 1 of [4]) :

$$\left| I_f(X_1 : X_2) - \sum_{k=0}^{s} \frac{f^{(k)}(1)}{k!} \chi_P^k(X_1 : X_2) \right| \leq \frac{1}{(s+1)!} \|f^{(s+1)}\|_\infty (M-m)^s,$$

where $\|f^{(s+1)}\|_\infty = \sup_{t\in[m,M]} |f^{(s+1)}(t)|$ and $m \leq \frac{p}{q} \leq M$.

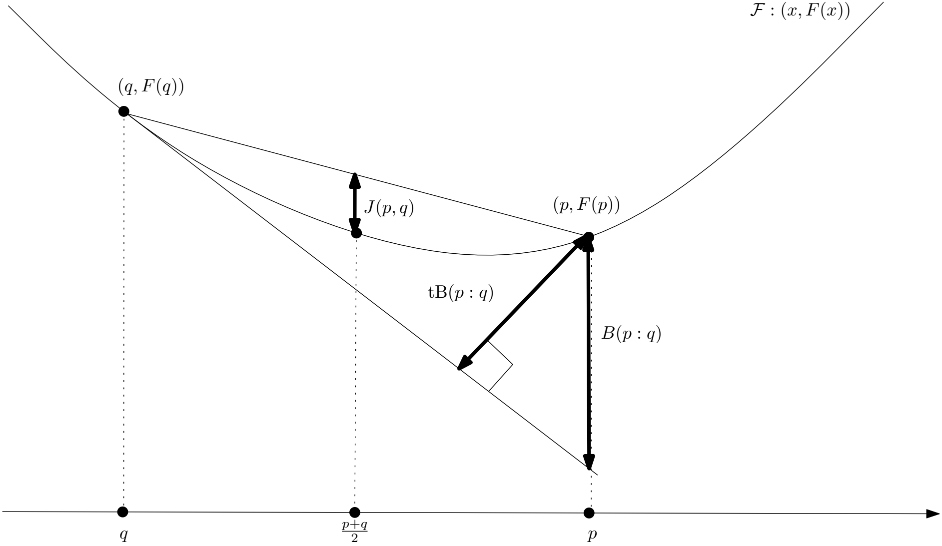
- $\lambda = 0$ (whenever $0 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$) and affine exponential families, simpler expression :

$$\begin{aligned} I_f(X_1 : X_2) &= \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} I_{1-i,i}(\theta_1 : \theta_2), \\ I_{1-i,i}(\theta_1 : \theta_2) &= \frac{e^{F(i\theta_2 + (1-i)\theta_1)}}{e^{iF(\theta_2)+(1-i)F(\theta_1)}}. \end{aligned}$$

Designing conformal divergences : Finding graphical gaps !

Geometrically designed divergences

Plot of the convex generator F.



Divergences : skew Jensen & Bregman divergences

F a smooth convex function, the generator.

- Skew Jensen divergences :

$$
\begin{aligned}
J'_\alpha(p:q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q), \\
&= (F(p)F(q))_\alpha - F((pq)_\alpha),
\end{aligned}
$$

where $(pq)_\gamma = \gamma p + (1-\gamma)q = q + \gamma(p-q)$ and $(F(p)F(q))_\gamma = \gamma F(p) + (1-\gamma)F(q) = F(q) + \gamma(F(p) - F(q))$.

- Bregman divergences :

$$
B(p:q) = F(p) - F(q) - \langle p-q, \nabla F(q) \rangle,
$$

$$
\boxed{\lim_{\alpha \to 0} J_\alpha(p:q) = B(p:q), \quad \lim_{\alpha \to 1} J_\alpha(p:q) = B(q:p)}
$$

- Statistical skewed Bhattacharrya divergence :

$$
\mathrm{Bhat}(p_1:p_2) = -\log \int p_1(x)^\alpha p_2(x)^{1-\alpha} \mathrm{d}\nu(x) = J'_\alpha(\theta_1:\theta_2)
$$

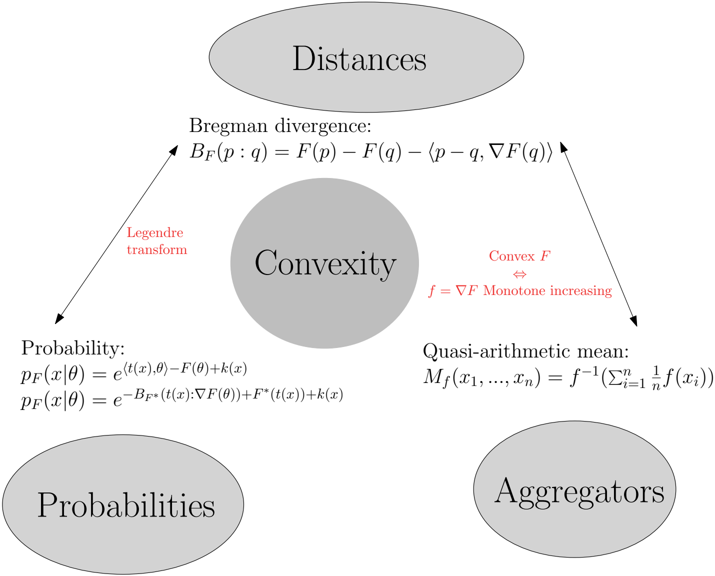
for exponential families [27].

Divergences and centroids [33, 27]

Population minimizers : $\arg\min_c \sum_{i=1}^n w_i D(p_i : c)$

- useful for center-based clustering algorithms (k-means)
- For Bregman divergences : $c^R = \sum_i w_i p_i$ (invariant, center of mass). $c^L = (\nabla F)^{-1}(\sum_i w_i \nabla F(p_i))$ a f-mean also called quasi-arithmetic mean : $f^{-1}(\sum_i w_i f(x_i))$ that generalizes arithmetic $f(x) = x$, harmonic $f(x) = \frac{1}{x}$ and geometric means $f(x) = \log x$.
- Bregman information $\sum_{i=1}^n w_i D(p_i : c^R) = F(\sum_i w_i p_i) - \sum_i w_i F(p_i)$, a Jensen diversity index.
- For Jensen divergences, use Concave-Convex Procedure from $c_0 = \sum_i w_i p_i$ to solve $\sum_i w_i J'_\alpha(c : p_i)$:

$$\boxed{c_{t+1} = (\nabla F)^{-1}\left(\sum_i w_i \nabla F(\alpha c_t + (1-\alpha)p_i)\right)}$$


Distances
Bregman divergence:
$B_F(p:q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle$
Legendre transform
Convexity
Convex F
$\Leftrightarrow$
$f = \nabla F$ Monotone increasing
Probability:
$p_F(x|\theta) = e^{\langle t(x),\theta\rangle - F(\theta) + k(x)}$
$p_F(x|\theta) = e^{-B_{F^*}(t(x):\nabla F(\theta)) + F^*(t(x)) + k(x)}$
Quasi-arithmetic mean:
$M_f(x_1, ..., x_n) = f^{-1}(\sum_{i=1}^n \frac{1}{n} f(x_i))$
Probabilities
Aggregators

Total Bregman divergences [17]

Conformal divergence, conformal factor ρ :

$$\boxed{D'(p:q) = \rho(p,q)D(p:q)}$$

plays the rôle of "regularizer" [50]

Invariance by rotation of the axes of the design space

$$\begin{aligned}
\mathrm{tB}(p:q) &= \frac{B(p:q)}{\sqrt{1+\langle \nabla F(q), \nabla F(q)\rangle}} = \rho_B(q)B(p:q),\\
\rho_B(q) &= \frac{1}{\sqrt{1+\langle \nabla F(q), \nabla F(q)\rangle}}.
\end{aligned}$$

For example, total squared Euclidean divergence :

$$tE(p,q) = \frac{1}{2}\frac{\langle p-q, p-q\rangle}{\sqrt{1+\langle q,q\rangle}}.$$

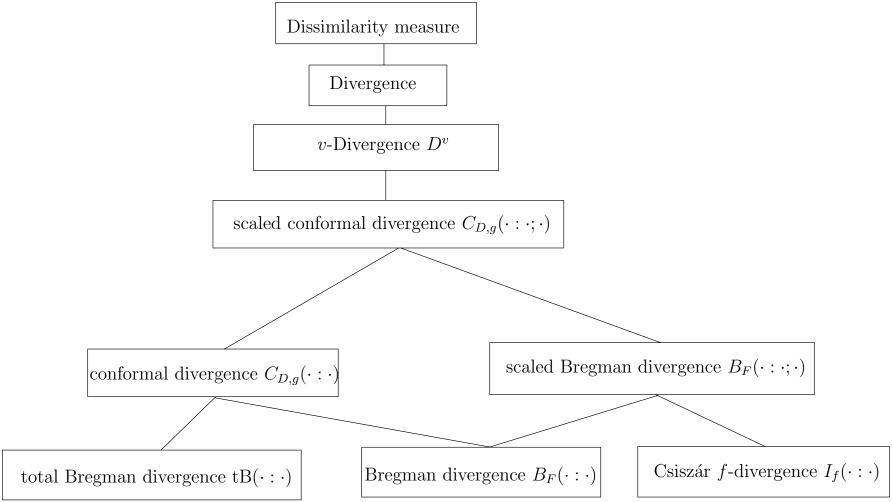
Total skew Jensen divergences [38]

$$
\begin{aligned}
\mathrm{tB}(p:q) &= \rho_B(q) B(p:q), \quad \rho_B(q) = \sqrt{\frac{1}{1+\langle \nabla F(q), \nabla F(q)\rangle}} \\
\mathrm{tJ}_\alpha(p:q) &= \rho_J(p,q) J_\alpha(p:q), \quad \rho_J(p,q) = \sqrt{\frac{1}{1+\frac{(F(p)-F(q))^2}{\langle p-q, p-q\rangle}}}
\end{aligned}
$$

Jensen-Shannon divergence, square root is a metric :

$$
\mathrm{JS}(p,q) = \frac{1}{2}\sum_{i=1}^{d} p_i \log \frac{2p_i}{p_i+q_i} + \frac{1}{2}\sum_{i=1}^{d} q_i \log \frac{2q_i}{p_i+q_i}
$$

But the square root of the total Jensen-Shannon divergence is **not** a metric.



$D^v(P:Q) = D(v(P):v(Q))$

$I_f(P:Q) = \int p(x) f\left(\left(\frac{q(x)}{p(x)}\right)\mathrm{d}\nu(x)\right.$

$B_F(P:Q) = F(P) - F(Q) - \langle P - Q, \nabla F(Q)\rangle$

$\mathrm{tB}_F(P:Q) = \frac{B_F(P:Q)}{\sqrt{1+\|\nabla F(Q)\|^2}}$

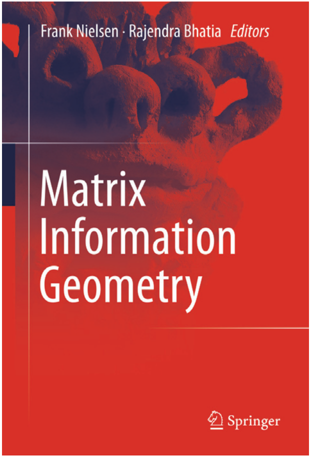
$C_{D,g}(P:Q) = g(Q)D(P:Q)$

$B_{F,g}(P:Q;W) = W B_F\left(\frac{P}{Q}:\frac{Q}{W}\right)$

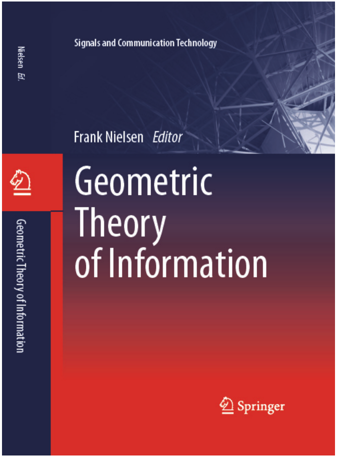
Summary : Part II. Geometric Computing in Information Spaces

- Location-scale families, spherical normal, symmetric positive definite matrices $\rightarrow$ hyperbolic geometry.
- Hyperbolic geometry : CG affine constructions in Klein disk
- Space of spheres in dually affine connection geometry
- Synthetic geometry for characterizing the best error exponent in Bayes error
- Conformal divergences : total Bregman/total Jensen divergences
- Clustering using pair of centroids for clusters using mixed divergences for symmetrized alpha divergences
- Learning statisical mixtures maximizing the complete likelihood as a sequence of geometric clustering problems : k-GLME
- In search of closed-form solutions : Jeffreys centroid using Lambert W function, f-divergence approximation for affine exponential families.

Computational Information Geometry (Edited books)



[25]



[24]

http://www.springer.com/engineering/signals/book/978-3-642-30231-2
http://www.sonycsl.co.jp/person/nielsen/infogeo/MIG/MIGBOOKWEB/

http://www.springer.com/engineering/signals/book/978-3-319-05316-5
http://www.sonycsl.co.jp/person/nielsen/infogeo/GTI/GeometricTheoryOfInformation.html

Geometric Sciences of Information (GSI) 2015

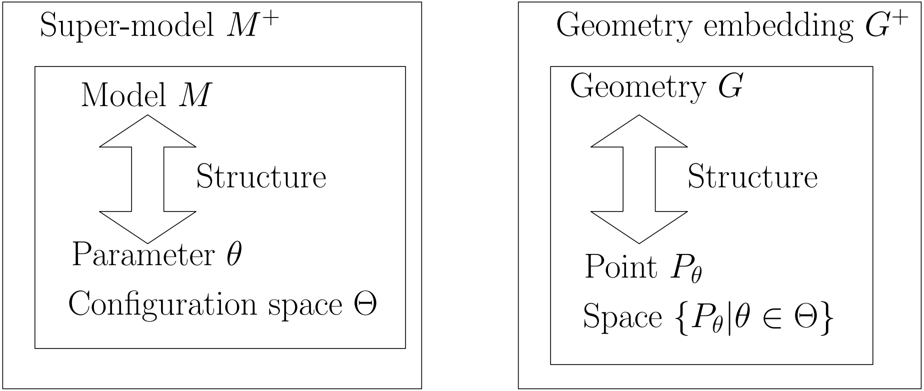
October 28-30th 2015. Deadline **1st March 2015**

http://www.gsi2015.org/

Summary : Computational Information Geometry

- Originally, IG studied the space of (parametric) probability distributions, but now geometry of "parameter spaces" in general (matrices, dynamic systems, etc.)
- Fisher-Rao Riemannian geometry has often geodesics *not* in closed form
- Dual connections coupled with metric has dual geodesics straight in biorthogonal affine coordinate systems
- Bregman divergences are canonical divergences in dually flat spaces
- Csiszár f-divergences preserve information monotonicity and induce locally Fisher tensor metric geometry.
- Algorithm designs are often based on information projections.

Closing philosophical view...



coordinate-based (biased)

coordinate-free!

The next big wave...

Quantum Information Geometry (and QIT)

- Quantum states : density matrices = Hermitian positive semi-definite matrices of unit trace (John von Neumann, 1927)
- A generalization of probability theory (classical probability=diagonal matrices=commutative matrices)
- Several Quantum Fisher Information metrics [46]
- Quantum random walks to define distance between graphs (simulated on classical computers [10])
- Quantum Voronoi diagrams [31]
- etc.

Thank you !

Next time, why not consider CIG for your ML problems - :) ?

Bibliography I

Shun-Ichi Amari.
Natural gradient works efficiently in learning.
Neural computation, 10(2) :251–276, 1998.

Shun-ichi Amari and Hiroshi Nagaoka.
Methods of Information Geometry.
Oxford University Press, 2000.

Marc Arnaudon and Frank Nielsen.
On approximating the Riemannian 1-center.
Computational Geometry, 46(1) :93 – 104, 2013.

N.S. Barnett, P. Cerone, S.S. Dragomir, and A. Sofo.
Approximating Csiszár f-divergence by the use of Taylor's formula with integral remainder.
Mathematical Inequalities & Applications, 5(3) :417–434, 2002.

D. A. Barry, P. J. Culligan-Hensley, and S. J. Barry.
Real values of the W-function.
ACM Trans. Math. Softw., 21(2) :161–171, June 1995.

Jean-Daniel Boissonnat and Christophe Delage.
Convex hull and Voronoi diagram of additively weighted points.
In Gerth StÃÿlting Brodal and Stefano Leonardi, editors, *ESA*, volume 3669 of *Lecture Notes in Computer Science*, pages 367–378. Springer, 2005.

Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock.
Bregman Voronoi diagrams.
Discrete and Computational Geometry, 44(2) :281–307, April 2010.

Bibliography II

Herman Chernoff.
A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations.
Annals of Mathematical Statistics, 23 :493–507, 1952.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari.
Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization.
Entropy, 13(1) :134–170, 2011.

David Emms, Edwin R Hancock, and Richard C Wilson.
Graph similarity using interfering quantum walks.
In *Proceedings of the 12th international conference on Computer analysis of images and patterns*, pages 823–831. Springer-Verlag, 2007.

P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang C. Joshi.
Principal geodesic analysis for the study of nonlinear statistics of shape.
IEEE Trans. Med. Imaging, 23(8) :995–1005, 2004.

Vincent Garcia and Frank Nielsen.
Simplification and hierarchical representations of mixtures of exponential families.
Signal Processing (Elsevier), 90(12) :3197–3212, 2010.

Vincent Garcia, Frank Nielsen, and Richard Nock.
Levels of details for Gaussian mixture models.
In *Asian Conference on Computer Vision (ACCV)*, volume 2, pages 514–525, 2009.

Bernd Gärtner and Sven Schönherr.
An efficient, exact, and generic quadratic programming solver for geometric optimization.
In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 110–118. ACM, 2000.

Bibliography III

Harold Hotelling.
Spaces of statistical parameters.
Bulletin of American Mathematical Society, 36(3) :191, 1930.

Takashi Kurose.
On the divergences of 1-conformally flat statistical manifolds.
Tohoku Mathematical Journal, Second Series, 46(3) :427–433, 1994.

Meizhu Liu, Baba C. Vemuri, Shun-ichi Amari, and Frank Nielsen.
Shape retrieval using hierarchical total Bregman soft clustering.
Transactions on Pattern Analysis and Machine Intelligence, 34(12) :2407–2419, 2012.

F. Nielsen and R. Nock.
On the chi square and higher-order chi distances for approximating f -divergences.
Signal Processing Letters, IEEE, 21(1) :10–13, 2014.

Frank Nielsen.
Legendre transformation and information geometry.
Technical Report CIG-MEMO2, September 2010.

Frank Nielsen.
k-MLE : A fast algorithm for learning statistical mixture models.
In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 869–872. IEEE, 2012.

Frank Nielsen.
Cramér-rao lower bound and information geometry.
arXiv preprint arXiv :1301.3578, 2013.

Bibliography IV

Frank Nielsen.

Jeffreys centroids : A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms.
Signal Processing Letters, IEEE, PP(99) :1–1, 2013.

Frank Nielsen.

Generalized bhattacharyya and chernoff upper bounds on bayes error using quasi-arithmetic means.
Pattern Recognition Letters, 42 :25–34, 2014.

Frank Nielsen.

Geometric Theory of Information.
Springer, 2014.

Frank Nielsen and Rajendra Bhatia, editors.

Matrix Information Geometry (Revised Invited Papers). Springer, 2012.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8) :5455–5466, 2011.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8) :5455–5466, August 2011.

Frank Nielsen and Vincent Garcia.

Statistical exponential families : A digest with flash cards, 2009.
arXiv.org :0911.4863.

Bibliography V

Frank Nielsen and Richard Nock.
On approximating the smallest enclosing Bregman balls.
In *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*, SCG '06, pages 485–486, New York, NY, USA, 2006. ACM.

Frank Nielsen and Richard Nock.
On the smallest enclosing information disk.
Information Processing Letters (IPL), 105(3) :93–97, 2008.

Frank Nielsen and Richard Nock.
Quantum Voronoi diagrams and Holevo channel capacity for 1-qubit quantum states.
In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 96–100. IEEE, 2008.

Frank Nielsen and Richard Nock.
The dual Voronoi diagrams with respect to representational Bregman divergences.
In *International Symposium on Voronoi Diagrams (ISVD)*, pages 71–78, 2009.

Frank Nielsen and Richard Nock.
Sided and symmetrized Bregman centroids.
IEEE Transactions on Information Theory, 55(6) :2882–2904, 2009.

Frank Nielsen and Richard Nock.
Entropies and cross-entropies of exponential families.
In *International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

Frank Nielsen and Richard Nock.
Hyperbolic Voronoi diagrams made easy.
In *2013 13th International Conference on Computational Science and Its Applications*, pages 74–80. IEEE, 2010.

Bibliography VI

Frank Nielsen and Richard Nock.
Hyperbolic Voronoi diagrams made easy.
In *International Conference on Computational Science and its Applications (ICCSA)*, volume 1, pages 74–80, Los Alamitos, CA, USA, march 2010. IEEE Computer Society.

Frank Nielsen and Richard Nock.
A closed-form expression for the Sharma-Mittal entropy of exponential families.
Journal of Physics A : Mathematical and Theoretical, 45(3), 2012.

Frank Nielsen and Richard Nock.
Total jensen divergences : Definition, properties and k-means++ clustering.
CoRR, abs/1309.7109, 2013.

Frank Nielsen and Richard Nock.
Visualizing hyperbolic Voronoi diagrams.
In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, SOCG'14, pages 90 :90–90 :91, New York, NY, USA, 2014. ACM.

Frank Nielsen and Richard Nock.
Visualizing hyperbolic Voronoi diagrams.
In *Symposium on Computational Geometry*, page 90, 2014.

Frank Nielsen and Richard Nock.
Total Jensen divergences : Definition, properties and clustering.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Frank Nielsen, Richard Nock, and Shun-ichi Amari.
On clustering histograms with k-means by using mixed α-divergences.
Entropy, 16(6) :3273–3301, 2014.

Bibliography VII

Frank Nielsen, Paolo Piro, and Michel Barlaud.
Bregman vantage point trees for efficient nearest neighbor queries.
In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME)*, pages 878–881, 2009.

Richard Nock and Frank Nielsen.
Fitting the smallest enclosing Bregman ball.
In *Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*, pages 649–656. Springer Berlin Heidelberg, 2005.

Richard Nock, Frank Nielsen, and Shun-ichi Amari.
On conformal divergences and their population minimizers.
CoRR, abs/1311.5125, 2013.

D. Petz and C. Ghinea.
Introduction to quantum Fisher information, volume 27.
2010.

Calyampudi Radhakrishna Rao.
Information and the accuracy attainable in the estimation of statistical parameters.
Bulletin of the Calcutta Mathematical Society, 37 :81–89, 1945.

Olivier Schwander and Frank Nielsen.
Learning mixtures by simplifying kernel density estimators.
In *Matrix Information Geometry*, pages 403–426. Springer, 2013.

Ivor W. Tsang, Andras Kocsor, and James T. Kwok.
Simpler core vector machines with enclosing balls.
In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 911–918, New York, NY, USA, 2007. ACM.

Bibliography VIII

Baba Vemuri, Meizhu Liu, Shun-ichi Amari, and Frank Nielsen.
Total Bregman divergence and its applications to DTI analysis.
IEEE Transactions on Medical Imaging, pages 475–483, 2011.

Huaiyu Zhu and Richard Rohwer.
Measurements of generalisation based on information geometry.
In StephenW. Ellacott, JohnC. Mason, and IainJ. Anderson, editors, *Mathematics of Neural Networks*, volume 8 of *Operations Research/Computer Science Interfaces Series*, pages 394–398. Springer US, 1997.