# Pattern learning and recognition on statistical manifolds: An information-geometric review

Frank Nielsen
Frank.Nielsen@acm.org
www.informationgeometry.org

Sony Computer Science Laboratories, Inc.

July 2013, SIMBAD, York, UK.

# Praise for Computational Information Geometry

- What is Information? $=$ *Essence* of data (datum="thing")
  (make it tangible $\rightarrow$ e.g., parameters of generative models)

- Can we do Intrinsic computing?
  (unbiased by any particular "data representation" $\rightarrow$ same results after recoding data)

- Geometry $\stackrel{?!}{\longrightarrow}$ Science of invariance
  (mother of Science, compass & ruler, Descartes analytic=coordinate/Cartesian, imaginaries, ...).
  The open-ended poetic mathematics...

# Rationale for Computational Information Geometry

- Information is *...never void!* → lower bounds
  - Cramér-Rao lower bound and Fisher information (estimation)
  - Bayes error and Chernoff information (classification)
  - Coding and Shannon entropy (communication)
  - Program and Kolmogorov complexity (compression). (Unfortunately not computable!)
- Geometry:
  - Language (point, line, ball, dimension, orthogonal, projection, geodesic, immersion, etc.)
  - Power of characterization (eg., intersection of two pseudo-segments not admitting closed-form expression)
- Computing: Information computing. Seeking for mathematical *convenience* and mathematical *tricks* (eg., kernel).
  Do you know the "space of functions" ?!?
  (Infinity and and 1-to-1 mapping, language vs continuum)

# This talk: Information-geometric Pattern Recognition

- Focus on statistical pattern recognition $\leftrightarrow$ geometric computing.
- Consider probability distribution families (parameter spaces) and statistical manifolds.
  Beyond traditional Riemannian and Hilbert sphere representations $p \to \sqrt{p}$
- Describe dually flat spaces induced by convex functions:
  - Legendre transformation $\to$ dual and mixed coordinate systems
  - Dual similarities/divergences
  - Computing-friendly dual affine geodesics

## Information-geometric Pattern Recognition

*By departing from vector-space representations one is confronted with the challenging problem of dealing with (dis)similarities that do not necessarily possess the Euclidean behavior or not even obey the requirements of a metric. The lack of the Euclidean and/or metric properties undermines the very foundations of traditional pattern recognition theories and algorithms, and poses totally new theoretical/computational questions and challenges.*

Thank you to Prof. Edwin Hancock (U. York, UK) and Prof. Marcello Pelillo (U. Venice, IT)

# Statistical Pattern Recognition

Models data with distributions (generative) or stochastic processes:

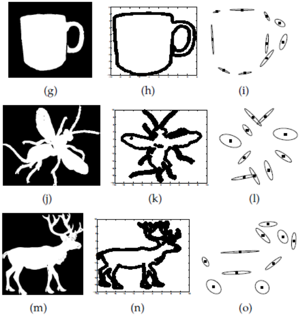
- parametric (Gaussians, histograms) [model size $\sim D$],
- semi-parametric (mixtures) [model size $\sim kD$],
- non-parametric (kernel density estimators [model size $\sim n$], Dirichlet/Gaussian processes [model size $\sim D \log n$],)

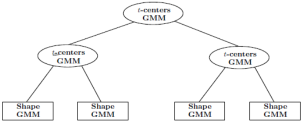
Data = Pattern ($\rightarrow$information) + noise (independent)

Statistical machine learning hot topic: Deep learning (restricted Boltzmann machines)

# Example I: Information-geometric PR (I)

Pattern = Gaussian mixture models (universal class)
Statistical (dis)similirty/distance: total Bregman divergence (tBD, tKL).
Invariance: ..., $p_i(x) \sim N(\mu_i, \Sigma_i)$, $y = A(x) = Lx + t$,
$y_i \sim N(L\mu_i + t, L\Sigma_i L^\top)$, $D(X_1 : X_2) = D(Y_1 : Y_2)$
($L$: any invertible affine transformation, $t$ a translation)




*Shape Retrieval using Hierarchical Total Bregman Soft Clustering* [11], IEEE PAMI, 2012.

# Example II: Information-geometric PR

DTI: diffusion ellipsoids, tensor interpolation.

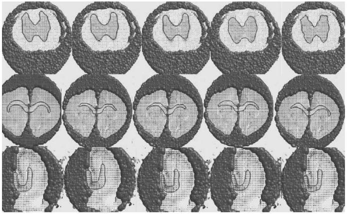
Pattern = zero-centered “Gaussians“

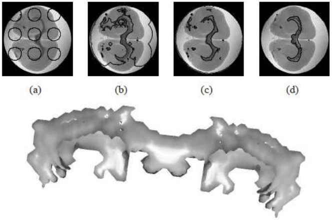
Statistical (dis)similarity/distance: total Bregman divergence (tBD, tKL).

Invariance: ..., $D(A^\top PA : A^\top QA) = D(P : Q)$, $A \in \mathrm{SL}(d)$: orthogonal matrix

(volume/orientation preserving)

total Bregman divergence (tBD).






(3D rat corpus callosum)

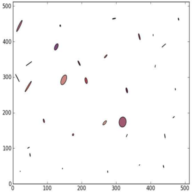
---

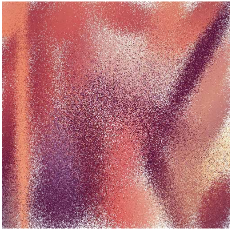
*Total Bregman Divergence and its Applications to DTI Analysis* [34], IEEE TMI, 2011.

# Statistical mixtures: Generative models of data sets

GMM = *feature descriptor* for information retrieval (IR)
→ classification [21], matching, etc.
Increase dimension using color image patches.
Low-frequency information encoded into compact statistical model.

Generative model → statistical image by GMM sampling.



→ A mixture $\sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$ is interpreted as a **weighted point set** in a parameter space: $\{w_i, \theta_i = (\mu_i, \Sigma_i)\}_{i=1}^{k}$.

# Statistical invariance

Riemannian structure $(M, g)$ on $\{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^D\}$

- $\theta$-Invariance under non-singular parameterization:

$$\rho(p(x;\theta), p(x;\theta')) = \rho(p(x;\lambda(\theta)), p(x;\lambda(\theta')))$$

Normal parameterization $(\mu, \sigma)$ or $(\mu, \sigma^2)$ yields same distance

- $x$-Invariance under different $x$-representation:
Sufficient statistics (Fisher, 1922):

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

All information for $\theta$ is contained in $T$.
$\rightarrow$ Lossless information data reduction (exponential families).
Markov kernel = statistical morphism (Chentsov 1972,[6, 7]).
A particular Markov kernel is a deterministic mapping
$T : X \rightarrow Y$ with $y = T(x)$, $\quad p_y = p_x T^{-1}$.

**Invariance if and only if $g \propto$ Fisher information matrix**

# $f$-divergences (1960's)

A statistical non-metric distance between two probability measures:

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

$f$: continuous convex function with $f(1)=0, f'(1)=0, f''(1)=1$.
$\rightarrow$ asymmetric (not a metric, except TV), modulo affine term.
$\rightarrow$ can always be symmetrized using $s = f + f^*$, with
$f^*(x) = xf(1/x)$.
include many well-known statistical measures: Kullback-Leibler, $\alpha$-divergences, Hellinger, Chi squared, total variation (TV), etc.

$f$-divergences are the only statistical divergences that preserves equivalence wrt. sufficient statistic mapping:

$$I_f(p:q) \geq I_f(p_M:q_M)$$

with equality if and only if $M = T$ (monotonicity property).

# Information geometry: Dually flat spaces, $\alpha$-geometry

Statistical invariance also obtained using $\boxed{(M, g, \nabla, \nabla^*)}$ where $\nabla$ and $\nabla^*$ are dual affine connections.

Riemannian structure $(M, g)$ is particular case for $\nabla = \nabla^* = \nabla^0$, Levi-Civita connection: $(M, g) = (M, g, \nabla^{(0)}, \nabla^{(0)})$

Dually flat space ($\alpha = \pm 1$) are algorithmically-friendly:

- Statistical mixtures of exponential families
- Learning & simplifying mixtures ($k$-MLE)
- Bregman Voronoi diagrams & dually $\perp$ triangulations

Goal: Algorithmics of Gaussians/histograms wrt. Kullback-Leibler divergence.

# Exponential Family Mixture Models (EFMMs)

Generalize Gaussian & Rayleigh MMs to many usual distributions.

$$m(x) = \sum_{i=1}^{k} w_i p_F(x; \lambda_i) \quad \text{with } \forall i \; w_i > 0, \sum_{i=1}^{k} w_i = 1$$

$$p_F(x; \lambda) = e^{\langle t(x), \theta \rangle - F(\theta) + k(x)}$$

$F$: log-Laplace transform (partition, cumulant function):

$$F(\theta) = \log \int_{x \in \mathcal{X}} e^{\langle t(x), \theta \rangle + k(x)} \mathrm{d}x,$$

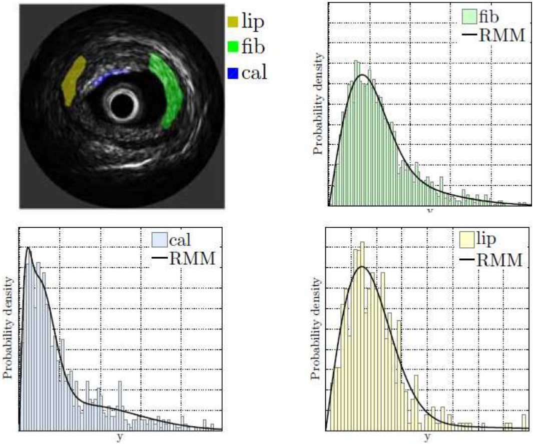
$$\theta \in \Theta = \left\{ \theta \;\middle|\; \int_{x \in \mathcal{X}} e^{\langle t(x), \theta \rangle + k(x)} \mathrm{d}x < \infty \right\}$$

the natural parameter space.

- $d$: Dimension of the support $\mathcal{X}$.
- $D$: order of the family ($= \mathrm{dim}\Theta$). Statistic: $t(x) : \mathbb{R}^d \rightarrow \mathbb{R}^D$.

# Statistical mixtures: Rayleigh MMs [33, 22]

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 $k = 2$)

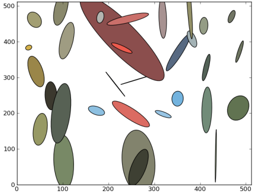
Coronary plaques: fibrotic tissues, calcified tissues, lipidic tissues

**Rayleigh Mixture Models** (**RMMs**):

for *segmentation* and *classification* tasks

# Statistical mixtures: Gaussian MMs [8, 22, 9]

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)
$\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) = (\theta_v, \theta_M)$
$\Theta = \mathbb{R} \times S_{++}^d$
$F(\theta) = \frac{1}{4}\theta_v^T \theta_M^{-1} \theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log\pi$
$t(x) = (x, -xx^T)$
$k(x) = 0$

# Sampling from a Gaussian Mixture Model

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 also an EF, normalized histogram.)

- 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: $z_i = \sqrt{-2 \log U_1} \cos(2\pi U_2)$

# Relative entropy for exponential families

- Distance between features (e.g., GMMs)
- Kullback-Leibler divergence (cross-entropy minus entropy):

$$
\begin{aligned}
\mathrm{KL}(P:Q) &= \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x \geq 0 \\
&= \underbrace{\int p(x)\log\frac{1}{q(x)}\mathrm{d}x}_{H^\times(P:Q)} - \underbrace{\int p(x)\log\frac{1}{p(x)}\mathrm{d}x}_{H(p)=H^\times(P:P)} \\
&= F(\theta_Q) - F(\theta_P) - \langle \theta_Q - \theta_P, \nabla F(\theta_P)\rangle \\
&= B_F(\theta_Q : \theta_P)
\end{aligned}
$$

Bregman divergence $B_F$ defined for a strictly convex and differentiable function up to some affine terms.

- Proof $\mathrm{KL}(P:Q) = B_F(\theta_Q:\theta_P)$ follows from

$$X \sim E_F(\theta) \Longrightarrow \boxed{E[t(X)] = \nabla F(\theta)}$$

# Convex duality: Legendre transformation

- For a strictly convex and differentiable function $F : \mathcal{X} \to \mathbb{R}$:

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

- 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 $F$ ($\nabla^2 F \succ 0$):

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

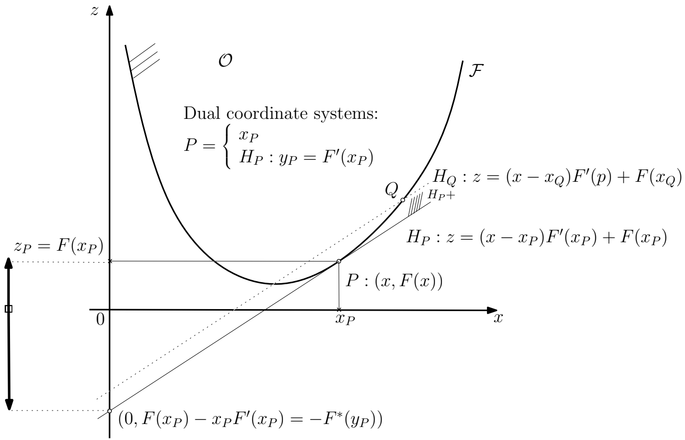
- Convex conjugates:

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

# Legendre duality: Geometric interpretation

Consider the epigraph of $F$ as a convex object:

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




Legendre transform also called *"slope" transform*.

# Legendre duality & Canonical divergence

- Convex conjugates have *functional inverse* gradients
  $\nabla F^{-1} = \nabla F^*$
  $\nabla F^*$ 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 $(\nabla F)^{-1}$:

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

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

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

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

# Dual Bregman divergences & canonical divergence [26]

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

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

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

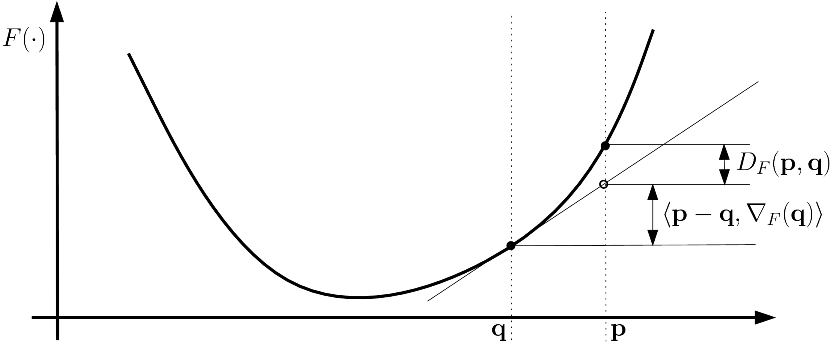
Shannon cross-entropy and entropy of EF [26]:

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

# Bregman divergence: Geometric interpretation (I)

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

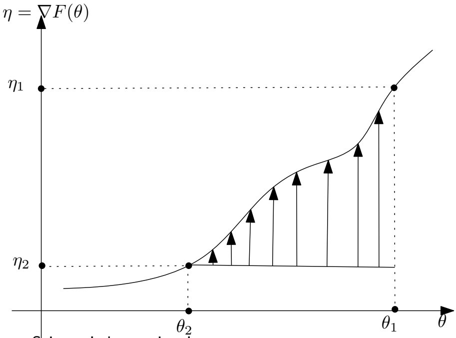
$$\boxed{D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle}$$



# Bregman divergence: Geometric interpretation (III)

Bregman divergence and path integrals

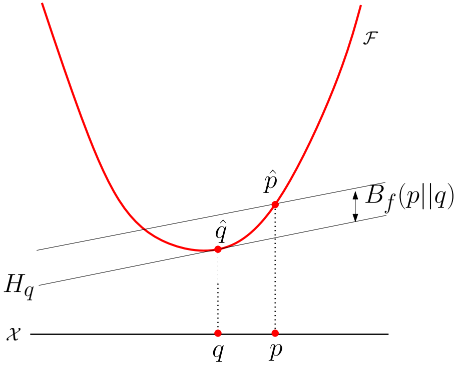
$$\begin{aligned} B(\theta_1 : \theta_2) &= F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle, & (1) \\ &= \int_{\theta_2}^{\theta_1} \langle \nabla F(t) - \nabla F(\theta_2), \mathrm{d}t \rangle, & (2) \\ &= \int_{\eta_1}^{\eta_2} \langle \nabla F^*(t) - \nabla F^*(\eta_1),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



# Bregman divergence: Geometric interpretation (II)

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

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

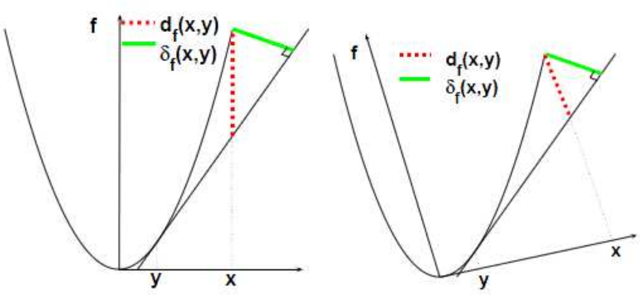



$B_f(.||q)$: vertical distance between the hyperplane $H_q$ tangent to $\mathcal{F}$ at lifted point $\hat{q}$, and the translated hyperplane at $\hat{p}$.

# total Bregman divergence (tBD)

By analogy to least squares and total least squares
total Bregman divergence (tBD) [11, 34, 12]

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




Proved statistical robustness of tBD.

# Bregman sided centroids [25, 21]

Bregman centroids = unique minimizers of average Bregman divergences ($B_F$ convex in right argument)

$$\bar{\theta} = \mathrm{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} B_F(\theta_i : \theta)$$

$$\bar{\theta}' = \mathrm{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} B_F(\theta : \theta_i)$$

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \theta_i, \text{ center of mass, independent of } F$$

$$\bar{\theta}' = (\nabla F)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} (\nabla F)(\theta_i) \right)$$

→ Generalized Kolmogorov-Nagumo $f$-means.

# Bregman divergences $B_F$ and $\nabla F$-means

Bijection quasi-arithmetic means ($\nabla F$) $\Leftrightarrow$ Bregman divergence $B_F$.

| Bregman divergence $B_F$ (entropy/loss function $F$) | $F$ | $\longleftrightarrow$ | $f = F'$ | $f^{-1} = (F')^{-1}$ | $f$-mean (Generalized means) |
|---|---|---|---|---|---|
| Squared Euclidean distance (half squared loss) | $\frac{1}{2}x^2$ | $\longleftrightarrow$ | $x$ | $x$ | Arithmetic mean $\sum_{j=1}^{n} \frac{1}{n} x_j$ |
| Kullback-Leibler divergence (Ext. neg. Shannon entropy) | $x \log x - x$ | $\longleftrightarrow$ | $\log x$ | $\exp x$ | Geometric mean $(\prod_{j=1}^{n} x_j)^{\frac{1}{n}}$ |
| Itakura-Saito divergence (Burg entropy) | $-\log x$ | $\longleftrightarrow$ | $-\frac{1}{x}$ | $-\frac{1}{x}$ | Harmonic mean $\frac{n}{\sum_{j=1}^{n} \frac{1}{x_j}}$ |

$\nabla F$ strictly increasing (like cumulative distribution functions)

# Bregman sided centroids [25]

Two sided centroids $\bar{C}$ and $\bar{C}'$ expressed using two $\theta/\eta$ coordinate systems: $=4$ equations.

$$\begin{aligned}\bar{C} &: \boxed{\bar{\theta}}, \bar{\eta}' \\ \bar{C}' &: \bar{\theta}', \boxed{\bar{\eta}}\end{aligned}$$

$$\begin{aligned}C : \bar{\theta} &= \frac{1}{n}\sum_{i=1}^{n} \theta_i \\ \bar{\eta}' &= \nabla F(\bar{\theta}) \\ C' : \bar{\eta} &= \frac{1}{n}\sum_{i=1}^{n} \eta_i \\ \bar{\theta}' &= \nabla F^*(\bar{\eta})\end{aligned}$$

# Bregman centroids and Jacobian fields

Centroid $\theta$ zeroes the left-sided Jacobian vector field:

$$\nabla_\theta \left( \sum_t w_t A(\theta : \eta_t) \right)$$

Sum of tangent vectors of geodesics from centroid to points $= 0$

$$\nabla_\theta A(\theta : \eta') = \eta - \eta'$$

$$\nabla_\theta (\sum_t w_t A(\theta : \eta_t)) = \eta - \bar{\eta}$$

since $\sum_t w_t = 1$, with $\bar{\eta} = \sum_t w_t \eta_t$.

# Bregman information [25]

Bregman information = minimum of loss function

$$
\begin{aligned}
I_F(\mathcal{P}) &= \frac{1}{n}\sum_{i=1}^{n} B_F(\theta_i : \bar{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n} F(\theta_i) - F(\bar{\theta}) - \langle \theta_i - \bar{\theta}, \nabla F(\bar{\theta})\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n} F(\theta_i) - F(\bar{\theta}) - \left\langle \underbrace{\frac{1}{n}\sum_{i=1}^{n} \theta_i - \bar{\theta}}_{=0}, \nabla F(\bar{\theta}) \right\rangle \\
&= J_F(\theta_1, ..., \theta_n)
\end{aligned}
$$

Jensen diversity index (e.g., Jensen-Shannon for $F(x) = x \log x$)

- For squared Euclidean distance, Bregman information = cluster variance,
- For Kullback-Leibler divergence, Bregman information related to mutual information.

# Bregman $k$-means clustering [4]

Bregman $k$-means: Find $k$ centers $\mathcal{C} = \{C_1, ..., C_k\}$ that minimizes the loss function:

$$
\begin{aligned}
L_F(\mathcal{P} : \mathcal{C}) &= \sum_{P \in \mathcal{P}} B_F(P : \mathcal{C}) \\
B_F(P : \mathcal{C}) &= \min_{i \in \{1,...,k\}} B_F(P : C_i)
\end{aligned}
$$

$\rightarrow$ generalize Lloyd' s quadratic error in Vector Quantization (VQ)

$$
\boxed{L_F(\mathcal{P} : \mathcal{C}) = I_F(\mathcal{P}) - I_F(\mathcal{C})}
$$

| | | |
|---|---|---|
| $I_F(\mathcal{P})$ | $\rightarrow$ | total Bregman information |
| $I_F(\mathcal{C})$ | $\rightarrow$ | between-cluster Bregman information |
| $L_F(\mathcal{P} : \mathcal{C})$ | $\rightarrow$ | within-cluster Bregman information |

total Bregman information = within-cluster Bregman information + between-cluster Bregman information

# Bregman $k$-means clustering [4]

$$I_F(\mathcal{P}) = L_F(\mathcal{P} : \mathcal{C}) + I_F(\mathcal{C})$$

Bregman clustering amounts to find the partition $\mathcal{C}^*$that *minimizes the information loss*:

$$L_F{}^* = L_F(\mathcal{P} : \mathcal{C}^*) = \min_{\mathcal{C}}(I_F(\mathcal{P}) - I_F(\mathcal{C}))$$

Bregman $k$-means:

- Initialize distinct seeds: $C_1 = P_1, ..., C_k = P_k$
- Repeat until convergence
  - Assign point $P_i$ to its closest centroid:

$$\mathcal{C}_i = \{P \in \mathcal{P} \mid B_F(P : C_i) \leq B_F(P : C_j) \; \forall j \neq i\}$$

  - Update cluster centroids by taking their center of mass: $C_i = \frac{1}{|\mathcal{C}_i|}\sum_{P \in \mathcal{C}_i} P$.

Loss function monotonically decreases and converges to a *local* optimum. (Extend to weighted point sets using barycenters.)

# Bregman $k$-means++ [1]: Careful seeding (only?!)

(also called Bregman $k$-medians since $\min \sum_i B_F^1(p_i : x)$).
Extend the $D^2$-initialization of $k$-means++

Only seeding stage yields probabilistically guaranteed global approximation factor:

Bregman $k$-means++:

- Choose $\mathcal{C} = \{C_l\}$ for $l$ uniformly random in $\{1, ..., n\}$
- While $|\mathcal{C}| < k$
  - Choose $P \in \mathcal{P}$ with probability $\frac{B_F(P:\mathcal{C})}{\sum_{i=1}^n B_F(P_i:\mathcal{C})} = \frac{B_F(P:\mathcal{C})}{L_F(\mathcal{P}:\mathcal{C})}$

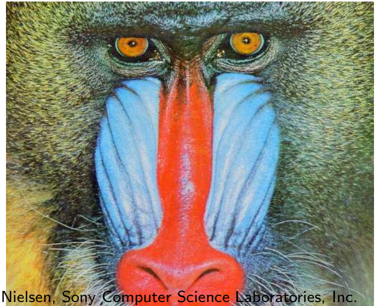
$\rightarrow$ Yields a $O(\log k)$ approximation factor (with high probability).
Constant in $O(\cdot)$ depends on ratio of min/max $\nabla^2 F$.

# Anisotropic Voronoi diagram (for MVN MMs) [10, 14]

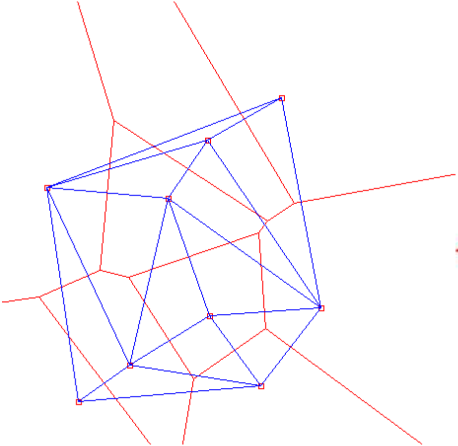
Learning mixtures, Talk of O. Schwander on EM, $k$-MLE
From the source color image (a), we buid a 5D GMM with $k = 32$ components, and color each pixel with the mean color of the anisotropic Voronoi cell it belongs to. ($\sim$ weighted squared Mahalanobis distance per center)

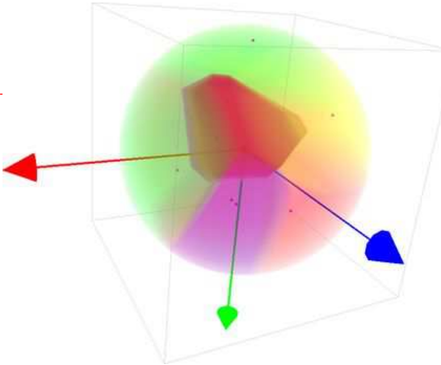
$$\log p_F(x; \theta_i) \propto -B_{F^*}(t(x) : \eta_i) + k(x)$$

Most likely component segmentation $\equiv$ Bregman Voronoi diagram (squared Mahalanobis for Gaussians)



# Voronoi diagrams in dually flat spaces...





Voronoi diagram, dual ⊥ Delaunay triangulation (general position)

# Bregman dual bisectors: Hyperplanes & hypersurfaces [5, 24, 27]

Right-sided bisector: $\rightarrow$ Hyperplane ($\theta$-hyperplane)

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

$H_F$ :

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

Left-sided bisector: $\rightarrow$ Hypersurface ($\eta$-hyperplane)

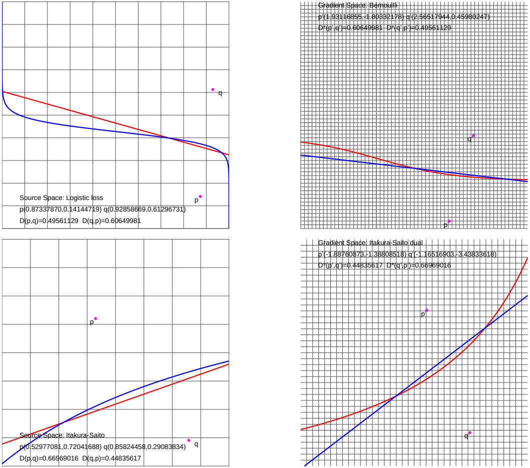
$$H'_F(p,q) = \{x \in \mathcal{X} \mid B_F(p:x) = B_F(q:x)\}$$

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

# Visualizing Bregman bisectors

Primal coordinates $\theta$ — Dual coordinates $\eta$

natural parameters — expectation parameters



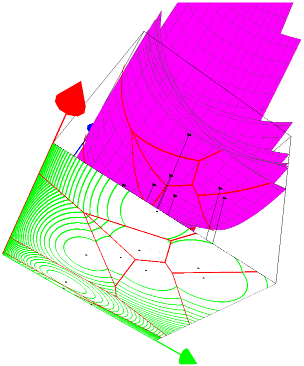
# Bregman Voronoi diagrams as minimization diagrams [5]

A subclass of affine diagrams which have all non-empty cells .

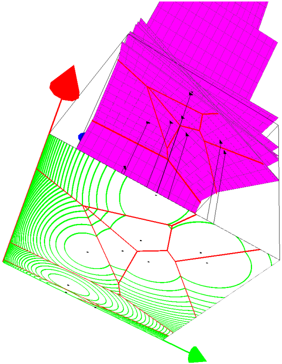
Minimization diagram of the $n$ functions
$D_i(x) = B_F(x : p_i) = F(x) - F(p_i) - \langle x - p_i, \nabla F(p_i)\rangle$.

$\equiv$ minimization of $n$ linear functions:

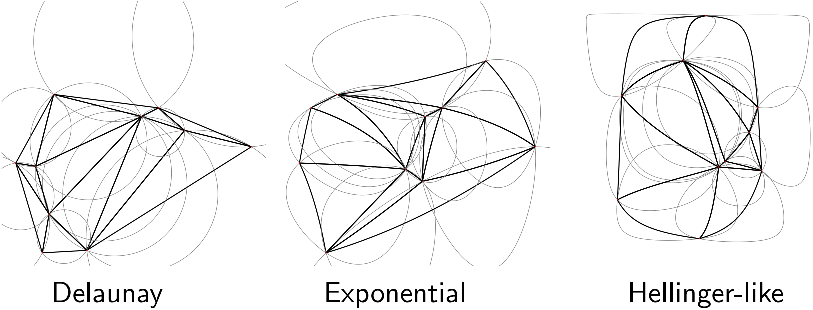
$$\boxed{H_i(x) = (p_i - x)^T \nabla F(q_i) - F(p_i)}$$



$\Longleftrightarrow$



# Bregman dual Delaunay triangulations



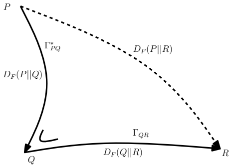

- empty Bregman sphere property,
- geodesic triangles.

BVDs extends Euclidean Voronoi diagrams with similar complexity/algorithms.

# Non-commutative Bregman Orthogonality

3-point property (generalized law of cosines):

$$B_F(p:r) = B_F(p:q) + B_F(q:r) - (p-q)^T(\nabla F(r) - \nabla F(q)))$$




$(pq)_\theta$ Bregman orthogonal to $(qr)_\eta$ iff.

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

(Equivalent to $\langle \theta_p - \theta_q, \eta_r - \eta_q \rangle = 0$)
Extend Pythagoras theorem

$$(pq)_\theta \perp_F (qr)_\eta$$

$\rightarrow \perp_F$ is not commutative...

... except in the squared Euclidean/Mahalanobis case,

# Dually orthogonal Bregman Voronoi & Triangulations

Ordinary Voronoi diagram is perpendicular to Delaunay triangulation.

Dual line segment geodesics:

$$\begin{aligned}(pq)_\theta &= \{\theta = \theta_p + (1-\lambda)\theta_q \,|\lambda \in [0,1]\} \\ (pq)_\eta &= \{\eta = \eta_p + (1-\lambda)\eta_q \,|\lambda \in [0,1]\}\end{aligned}$$

Bisectors:

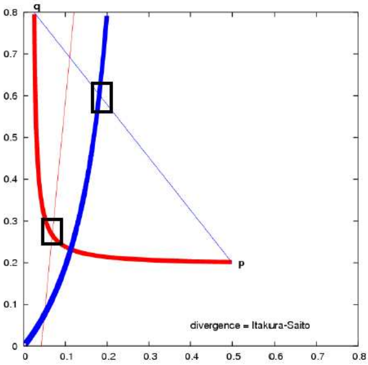
$$\begin{aligned}B_\theta(p,q) &: \langle x, \theta_q - \theta_p\rangle + F(\theta_p) - F(\theta_q) = 0 \\ B_\eta(p,q) &: \langle x, \eta_q - \eta_p\rangle + F^*(\eta_p) - F^*(\eta_q) = 0\end{aligned}$$

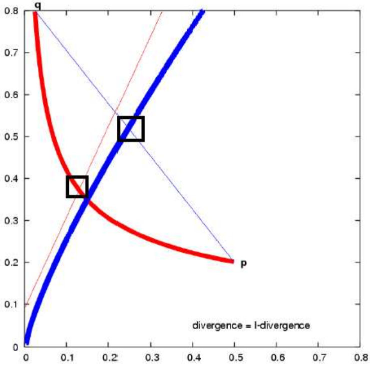
Dual orthogonality:

$$\begin{aligned}B_\eta(p,q) &\perp (pq)_\eta \\ (pq)_\theta &\perp B_\theta(p,q)\end{aligned}$$

# Dually orthogonal Bregman Voronoi & Triangulations

$$B_\eta(p,q) \quad \perp \quad (pq)_\eta$$
$$(pq)_\theta \quad \perp \quad B_\theta(p,q)$$

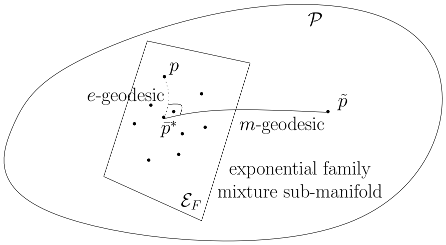


# Simplifying mixture: Kullback-Leibler projection theorem

An exponential family mixture model $\tilde{p} = \sum_{i=1}^{k} w_i p_F(x; \theta_i)$
Right-sided KL barycenter $\bar{p}^*$ of components interpreted as the *projection* of the mixture model $\tilde{p} \in \mathcal{P}$ onto the model exponential family manifold $\mathcal{E}_F$ [32]:

$$\boxed{\bar{p}^* = \arg\min_{p \in \mathcal{E}_F} \mathrm{KL}(\tilde{p} : p)}$$



manifold of probability distribution

Right-sided KL centroid = Left-sided Bregman centroid

# A simple proof

$$\mathrm{argmin}_\theta \mathrm{KL}(m(x):p_\theta(x)) \quad = \quad E_m[\log m] - E_m[\log p_\theta]$$

$$\begin{aligned} &= E_m[\log m] - E_m[k(x)] - E_m[\langle t(x), \theta\rangle - F(\theta)] \\ &\equiv \mathrm{argmax}_\theta E_m[\langle t(x), \theta\rangle] - F(\theta) \\ &= \mathrm{argmax}_\theta \langle E_m[t(x)], \theta\rangle] - F(\theta) \end{aligned}$$

$$E_m[t(x)] = \sum_t w_t E_{p_{\theta_t}}[t(x)] = \sum_t w_t \eta_t = \bar{\eta}$$

$$\nabla F(\theta) = \eta = \bar{\eta}$$

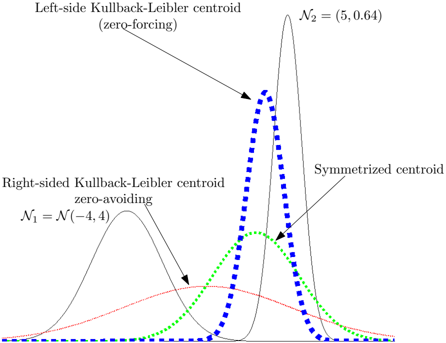
$$\breve{\theta} = \nabla F^*(\bar{\eta})$$

# Left-sided or right-sided Kullback-Leibler centroids?

Left/right Bregman centroids=Right/left entropic centroids (KL of exp. fam.)

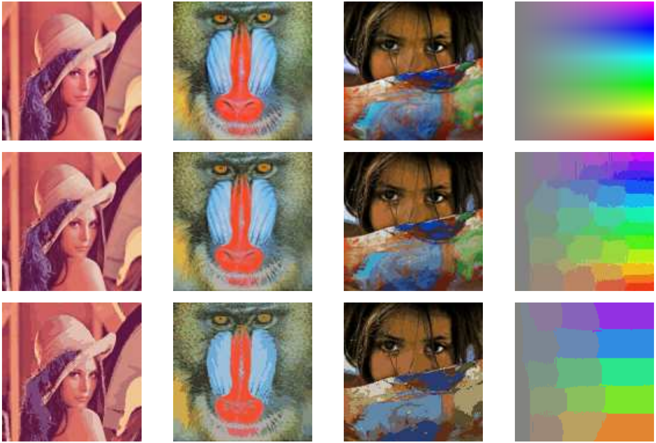
Left-sided/right-sided centroids: *different* (statistical) properties:

- Right-sided entropic centroid: zero-avoiding (cover support of pdfs.)
- Left-sided entropic centroid: zero-forcing (captures highest mode).



# Hierarchical clustering of GMMs (Burbea-Rao)

Hierarchical clustering of GMMs wrt. Bhattacharyya distance.
Simplify the number of components of an initial GMM.



(a) source

(b) $k = 48$

(c) $k = 16$

# Two symmetrizations of Bregman divergences

- **Jeffreys-Bregman divergences.**

$$
\begin{aligned}
S_F(p;q) &= \frac{B_F(p,q)+B_F(q,p)}{2} \\
&= \frac{1}{2}\langle p-q, \nabla F(p) - \nabla F(q)\rangle,
\end{aligned}
$$

- **Jensen-Bregman divergences (diversity index).**

$$
\begin{aligned}
J_F(p;q) &= \frac{B_F(p,\frac{p+q}{2}) + B_F(q,\frac{p+q}{2})}{2} \\
&= \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) = \mathrm{BR}_F(p,q)
\end{aligned}
$$

Skew Jensen divergence [21, 29]
$J_F^{(\alpha)}(p;q) = \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) = \mathrm{BR}_F^{(\alpha)}(p;q)$
(Jeffreys and Jensen-Shannon symmetrization of Kullback-Leibler)

# Burbea-Rao centroids ($\alpha$-skewed Jensen centroids)

Minimum average divergence:

$$\mathrm{OPT} : c = \arg\min_x \sum_{i=1}^{n} w_i J_F^{(\alpha)}(x, p_i) = \arg\min_x L(x)$$

Equivalent to minimize:

$$\boxed{E(c) = (\sum_{i=1}^{n} w_i \alpha) F(c) - \sum_{i=1}^{n} w_i F(\alpha c + (1-\alpha) p_i)}$$

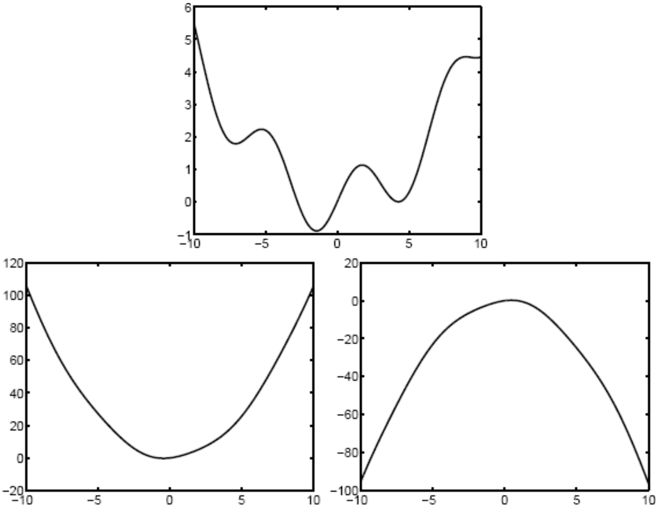
Sum $E = F + G$ of convex $F$ + concave $G$ function $\Rightarrow$ Convex-ConCave Procedure (CCCP)
Start from arbitrary $c_0$, and iteratively update as:

$$\nabla F(c_{t+1}) = -\nabla G(c_t)$$

$\Rightarrow$ *guaranteed convergence* to a local minimum.

# ConCave Convex Procedure (CCCP)

$$\min_x E(x) = F(x) + G(x)$$
$$\nabla F(c_{t+1}) = -\nabla G(c_t)$$



# Iterative algorithm for Burbea-Rao centroids

Apply CCCP scheme

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

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

Get arbitrarily fine approximations of the (skew) Burbea-Rao centroids and barycenters.

Unique GLOBAL minimum when divergence is separable [21].

Unique GLOBAL minimum for matrix mean [23] for the logDet divergence.

# Information-geometric computing on statistical manifolds

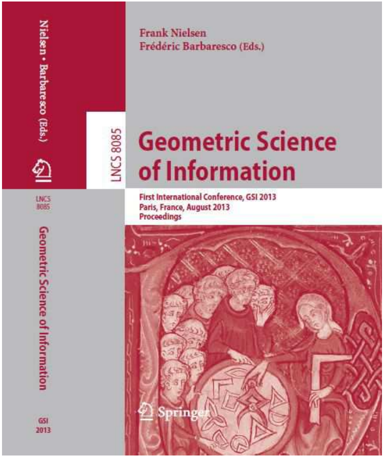
- Cramér-Rao Lower Bound and Information Geometry [13] (overview article)
- Chernoff information on statistical exponential family manifold [19]
- Hypothesis testing and Bregman Voronoi diagrams [18]
- Jeffreys centroid [20]
- Learning mixtures with $k$-MLE [17]
- Closed-form divergences for statistical mixtures [16]

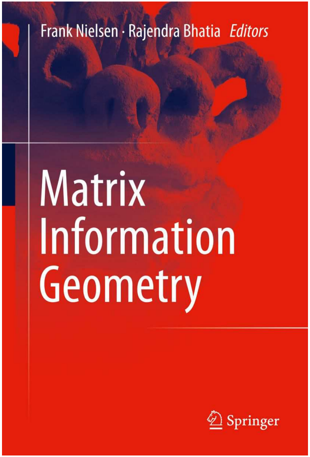
# Summary

**Information-geometric pattern recognition:**

- Statistical manifold $(M, g)$: Rao's distance and Fisher-Rao curved Riemannian geometry.
- Statistical manifold $(M, g, \nabla, \nabla^*)$: dually flat spaces, Bregman divergences, geodesics are straight lines in either $\theta/\eta$ parameter space. $\pm 1$-geometry, special case of $\alpha$-geometry
- Clustering in dually flat spaces
- Software library: jMEF [8] (Java), pyMEF [31] (Python)
- ... but also many other geometries to explore: Hilbertian, Finsler [3], Kähler, Wasserstein, Contact, Symplectic, etc. (it is easy to require non-Euclidean geometry but then space is wild open!)

# Edited book (MIG) and proceedings (GSI)



# Thank you.

# Exponential families & statistical distances

Universal density estimators [2] generalizing Gaussians/histograms (single EF density approximates any smooth density)

Explicit formula for

- Shannon entropy, cross-entropy, and Kullback-Leibler divergence [26]:
- Rényi/Tsallis entropy and divergence [28]
- Sharma-Mittal entropy and divergence [30]. A 2-parameter family extending extensive Rényi (for $\beta \to 1$) and non-extensive Tsallis entropies (for $\beta \to \alpha$)

$$H_{\alpha,\beta}(p) = \frac{1}{1-\beta}\left(\left(\int p(x)^\alpha \mathrm{d}x\right)^{\frac{1-\beta}{1-\alpha}} - 1\right),$$

  with $\alpha > 0, \alpha \neq 1, \beta \neq 1$.
- Skew Jensen and Burbea-Rao divergence [21]
- Chernoff information and divergence [15]
- Mixtures: total Least square, Jensen-Rényi, Cauchy-Schwarz divergence [16].

# Statistical invariance: Markov kernel

Probability family: $p(x;\theta)$.

$(X,\sigma)$ and $(X',\sigma')$ two measurable spaces.
$\sigma$: A $\sigma$-algebra on $X$
(non-empty, closed under complementation and countable union).

Markov kernel = transition probability kernel
$K : X \times \sigma' \rightarrow [0,1]$:

- $\forall E' \in \sigma', K(\cdot, E')$ measurable map,
- $\forall x \in X, K(x,\cdot)$ is a probability measure on $(X',\sigma')$.

$p$ a pm. on $(X,\sigma)$ induces $Kp$ a pm., with

$$Kp(E') = \int_X K(x,E')p(\mathrm{d}x), \forall E' \subset \sigma'$$

# Space of Bregman spheres and Bregman balls [5]

Dual Bregman balls (bounding Bregman spheres):

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

Legendre duality:

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

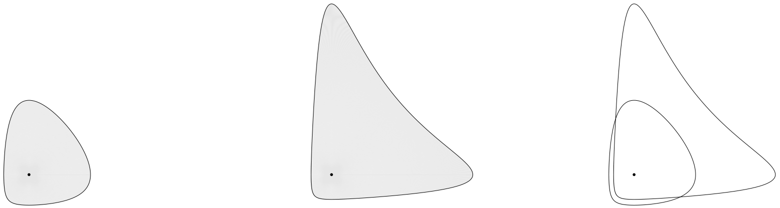


Illustration for Itakura-Saito divergence, $F(x) = -\log x$

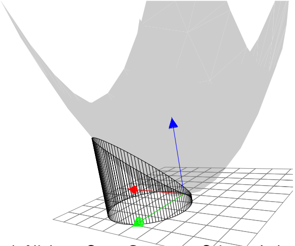
# Space of Bregman spheres: Lifting map [5]

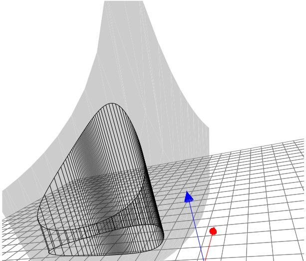
$\mathcal{F} : x \mapsto \hat{x} = (x, F(x))$, hypersurface in $\mathbb{R}^{d+1}$.

$H_p$: 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:

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





# Bibliographic references I


Marcel R. Ackermann and Johannes Blömer.

Bregman clustering for separable instances.

In *Scandinavian Workshop on Algorithm Theory (SWAT)*, pages 212–223, 2010.

Yasemin Altun, Alexander J. Smola, and Thomas Hofmann.

Exponential families for conditional random fields.

In *Uncertainty in Artificial Intelligence (UAI)*, pages 2–9, 2004.

Marc Arnaudon and Frank Nielsen.

Medians and means in Finsler geometry.

*LMS Journal of Computation and Mathematics*, 15, 2012.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh.

Clustering with Bregman divergences.

*Journal of Machine Learning Research*, 6:1705–1749, 2005.

Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock.

Bregman Voronoi diagrams.

*Discrete & Computational Geometry*, 44(2):281–307, 2010.

Nikolai Nikolaevich Chentsov.

*Statistical Decision Rules and Optimal Inferences*.

Transactions of Mathematics Monograph, numero 53, 1982.

Published in russian in 1972.

# Bibliographic references II


José Manuel Corcuera and Federica Giummolé.

A characterization of monotone and regular divergences.

*Annals of the Institute of Statistical Mathematics*, 50(3):433–450, 1998.

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.

François Labelle and Jonathan Richard Shewchuk.

Anisotropic Voronoi diagrams and guaranteed-quality anisotropic mesh generation.

In *Proceedings of the nineteenth annual symposium on Computational geometry*, SCG '03, pages 191–200, New York, NY, USA, 2003. ACM.

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.

Meizhu Liu, Baba C. Vemuri, Shun ichi Amari, and Frank Nielsen.

Total Bregman divergence and its applications to shape retrieval.

In *International Conference on Computer Vision (CVPR)*, pages 3463–3468, 2010.

# Bibliographic references III


Frank Nielsen.

Cramér-rao lower bound and information geometry.

In Rajendra Bhatia, C. S. Rajan, and Ajit Iqbal Singh, editors, *Connected at Infinity II: A selection of mathematics by Indians*. Hindustan Book Agency (Texts and Readings in Mathematics, TRIM).

arxiv 1301.3578.

Frank Nielsen.

*Visual Computing: Geometry, Graphics, and Vision*.

Charles River Media / Thomson Delmar Learning, 2005.

Frank Nielsen.

Chernoff information of exponential families.

*arXiv*, abs/1102.2684, 2011.

Frank Nielsen.

Closed-form information-theoretic divergences for statistical mixtures.

In *International Conference on Pattern Recognition (ICPR)*, pages 1723–1726, 2012.

Frank Nielsen.

$k$-MLE: A fast algorithm for learning statistical mixture models.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012.

preliminary, technical report on arXiv.

# Bibliographic references IV


Frank Nielsen.

Hypothesis testing, information divergence and computational geometry.

In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information (GSI)*, volume 8085 of *LNCS*, pages 241–248. Springer, 2013.

Frank Nielsen.

An information-geometric characterization of Chernoff information.

*IEEE Signal Processing Letters*, 20(3):269–272, 2013.

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 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.

# Bibliographic references V


Frank Nielsen, Meizhu Liu, Xiaojing Ye, and Baba C. Vemuri.

Jensen divergence based SPD matrix means and applications.

In *International Conference on Pattern Recognition (ICPR)*, 2012.

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):2048–2059, June 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 *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.

On rényi and tsallis entropies and divergences for exponential families.

*arXiv*, abs/1105.3259, 2011.

# Bibliographic references VI


Frank Nielsen and Richard Nock.

Skew Jensen-Bregman Voronoi diagrams.

*Transactions on Computational Science*, 14:102–128, 2011.

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.

Olivier Schwander and Frank Nielsen.

PyMEF - A framework for exponential families in Python.

In *IEEE/SP Workshop on Statistical Signal Processing (SSP)*, 2011.

Olivier Schwander and Frank Nielsen.

Learning mixtures by simplifying kernel density estimators.

In Frank Nielsen and Rajendra Bhatia, editors, *Matrix Information Geometry*, pages 403–426, 2012.

Jose Seabra, Francesco Ciompi, Oriol Pujol, Josepa Mauri, Petia Radeva, and Joao Sanchez.

Rayleigh mixture model for plaque characterization in intravascular ultrasound.

*IEEE Transaction on Biomedical Engineering*, 58(5):1314–1324, 2011.

Baba Vemuri, Meizhu Liu, Shun ichi Amari, and Frank Nielsen.

Total Bregman divergence and its applications to DTI analysis.

*IEEE Transactions on Medical Imaging*, 2011.