

A glance at
information-geometric signal processing

Frank Nielsen

Sony Computer Science Laboratories, Inc.

22nd October 2012 (MAHI)

Information geometry in Statistical Signal Processing

Statistical signal processing (SSP) models data with distributions:

- ▶ 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 $\sim D \log n$],)

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

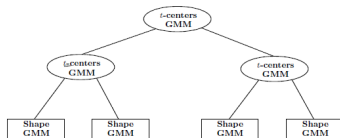
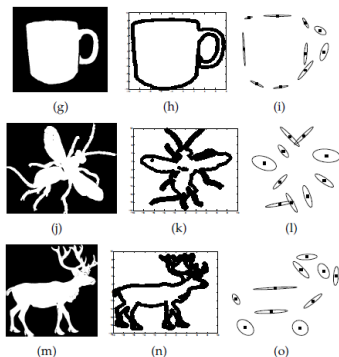
Paradigm of *computational information geometry* provides:

- ▶ Information (entropy), statistical invariance & geometry,
- ▶ Language of geometry for intuitive reasoning,
- ▶ Novel geometric algorithms for signal processing.

→ Intrinsic data analysis

Example of information-geometric SSP (I)

Statistical distance: **total Bregman divergence** (tBD).

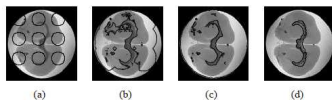
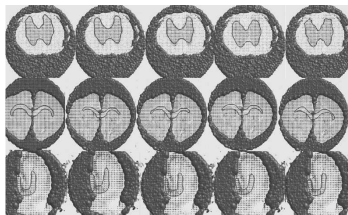


Shape Retrieval using Hierarchical Total Bregman Soft Clustering, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 2012.

Example of information-geometric SSP (II)

DTI: diffusion ellipsoids interpreted as zero-centered Gaussian distributions.

total Bregman divergence (tBD).



(3D rat corpus callosum)

Total Bregman Divergence and its Applications to DTI Analysis, IEEE Trans. Medical Imaging (TMI), 2010.

Statistical mixtures: Generative models of data sets

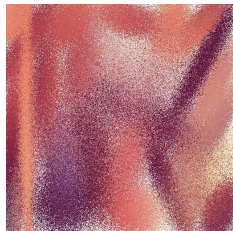
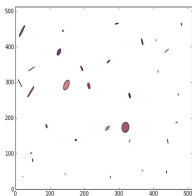
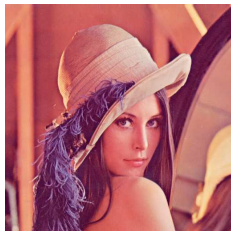
GMM = *feature descriptor* for information retrieval (IR)

→ classification [19], 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: $\{w_i, \theta_i = (\mu_i, \Sigma_i)\}_{i=1}^k$.

Information-geometric hyperspectral imaging

Image with z-axis = spectral bands (radiance or reflectance).
→ characterize spectral variability, similarity and discrimination.

- ▶ Normalize hyperspectral pixel vector (→ **histogram**):

$$p_i = \frac{x_i}{\sum_{i=1}^L x_i}.$$

- ▶ Spectral information divergence (single pixel):

$$SID(x, y) = D(x||y) + D(y||x),$$

$$D(p||q) = \sum_{i=1}^L p_i \log \frac{p_i}{q_i}$$

(aka. Jeffreys symmetrized Kullback-Leibler divergence [23])

C.-I. Chang, *An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis*, IEEE Trans. Information Theory, 2000.

Sided and symmetrized Bregman centroids, IEEE Trans. Information Theory, 2009.

Fisher-Rao Riemannian geometry (1945)

- ▶ D -parametric distribution family: $\{p(x; \theta) \mid \theta \subseteq \mathbb{R}^D\}$.
- ▶ **Fisher Information matrix** (FIM):

$$I(\theta) = [I_{ij}], \quad I_{ij} = E_{\theta} \left[\frac{\partial \log p(x; \theta)}{\partial \theta_i} \frac{\partial \log p(x; \theta)}{\partial \theta_j} \right],$$

$$I(\theta) = \text{Var} \left[\frac{\partial}{\partial \theta} \log p(x; \theta) \right] \succeq 0,$$

always **semi-positive definite**: $\forall x, x^T I(\theta) x \geq 0$.

- ▶ **Cramér-Rao lower bound** (CRLB) for an unbiased estimator $\hat{\theta}$:

$$\boxed{\text{Var}[\hat{\theta}] \succeq I^{-1}(\theta)}$$

Löwner ordering for cone of positive definite matrices:

$$A \succeq B \Leftrightarrow A - B \succeq 0.$$

→ FIM interpreted as **curvature** of the log-likelihood function (derivative of log-l. = **score function**).

→ Estimation efficiency of $\hat{\theta}$ depends on true hidden θ parameter.

Fisher-Rao Riemannian geometry (1945)

Rao chose the FIM for defining a *statistical manifold* (\mathcal{M}, g)

- ▶ Infinitesimal length element:

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

- ▶ Geodesic and distance are hard to explicitly calculate:

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

- ▶ Metric property of ρ , log/exp tangent/manifold mapping

→ FR geometry limited from the viewpoint of computation.

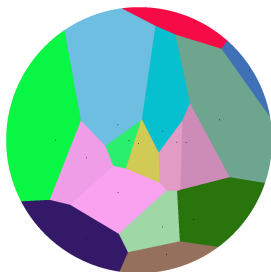
A particular case of Fisher-Rao Riemannian geometry

For **location-scale families** (normal, Cauchy, Laplace, uniform, elliptical): $p(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$.

Fisher-Rao geometry amounts to hyperbolic geometry of constant curvature $\kappa = -\frac{1}{(d-1)\beta}$ depending on the **density profile**:

$$\beta = \int \left(x \frac{f'(x)}{f(x)} + 1 \right)^2 f(x) dx.$$

FR statistical Voronoi diagram [25] = Hyperbolic Voronoi diagram on parameter space.



[Klein ball]

Hyperbolic Voronoi diagrams made easy, ICCSA, 2010. [4]

Statistical invariance

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

- ▶ θ -Invariance under non-singular parameterization:

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

Normal parameterization (μ, σ) or (μ, σ^2) yields same distance

- ▶ x -Invariance under different x -representation:

Sufficient statistics (Fisher, 1922):

$$\Pr(X = x \mid t(X) = t, \theta) = \Pr(X = x \mid T(X) = t)$$

All information for θ is contained in T .

→ Lossless information data reduction (exponential families).

Markov kernel = statistical morphism (Chentsov 1972, [7, 8]).

A particular Markov kernel is a deterministic mapping

$$T : X \rightarrow Y \text{ with } y = T(x), \quad p_y = p_x T^{-1}.$$

Invariance if and only if $g =$ Fisher information matrix

f -divergences (1960's)

A **statistical non-metric distance** between two probability measures:

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

f : continuous convex function with $f(1) = 0$, $f'(1) = 0$, $f''(1) = 1$.

→ asymmetric (not a metric, except TV), modulo affine term.

→ can always be symmetrized using $s = f + f^*$, with

$f^*(x) = xf(1/x)$.

include many well-known statistical measures: Kullback-Leibler, α -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**).

Outline: Dually flat spaces

Statistical invariance also obtained using (M, g, ∇, ∇^*) where ∇ and ∇^* 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 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):

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

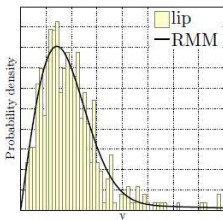
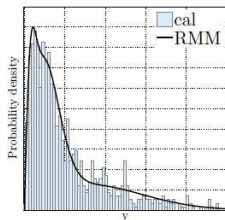
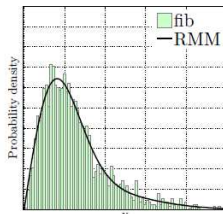
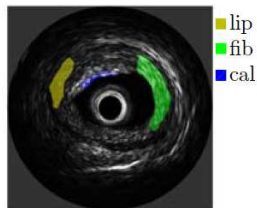
$$\theta \in \Theta = \left\{ \theta \mid \int_{x \in \mathcal{X}} e^{\langle t(x), \theta \rangle + k(x)} dx < \infty \right\}$$

the **natural parameter space**.

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

Statistical mixtures: Rayleigh MMs [35, 20]

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 $k = 2$)

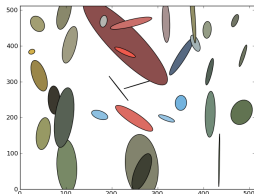
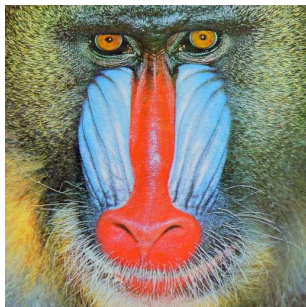
Coronary plaques: fibrotic tissues, calcified tissues, lipidic tissues

Rayleigh Mixture Models (RMMs):

for *segmentation* and *classification* tasks

Statistical mixtures: Gaussian MMs [9, 20, 10]

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)

$$D = \frac{d(d+3)}{2} \text{ (order)}$$

$$\theta = (\Sigma^{-1} \mu, \frac{1}{2} \Sigma^{-1}) = (\theta_v, \theta_M)$$

$$\Theta = \mathbb{R} \times S_{++}^d$$

$$F(\theta) = \frac{1}{4} \theta_v^T \theta_M^{-1} \theta_v - \frac{1}{2} \log |\theta_M| +$$

$$\frac{d}{2} \log \pi$$

$$t(x) = (x, -xx^T)$$

$$k(x) = 0$$

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, \dots, w_k ,
- ▶ Draw a variate x according to $N(\mu_l, \Sigma_l)$.

→ Sampling is a **doubly stochastic process**:

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

$$l \sim \text{Multinomial}(w_1, \dots, w_k)$$

(Multinomial is also an EF, normalized histogram.)

- ▶ then draw at random a variate x from the l -th component

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

$x = \mu + Cz$ with Cholesky: $\Sigma = CC^T$ and $z = [z_1 \dots z_d]^T$
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**):

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

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

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

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

Convex duality: Legendre transformation

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

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

- ▶ 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 ($\nabla^2 F \succ 0$):

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

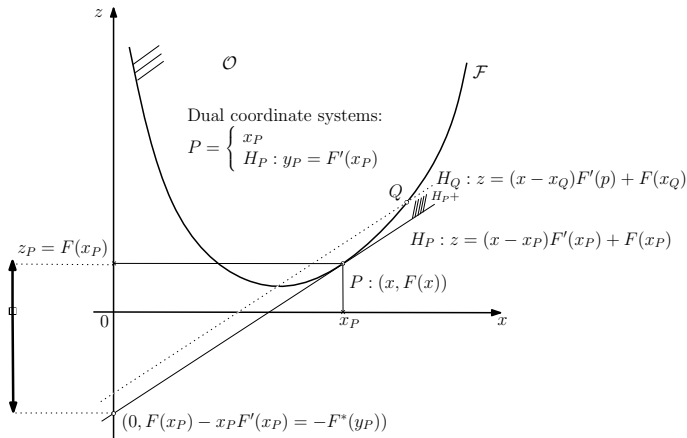
- ▶ Convex conjugates:

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

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

$$\nabla F^{-1} = \nabla F^*$$

∇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 $(\nabla F)^{-1}$:

$$F^*(y) = \langle (\nabla F)^{-1}(y), y \rangle - F((\nabla F)^{-1}(y))$$

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

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

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

Dual Bregman divergences & canonical divergence [24]

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

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

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

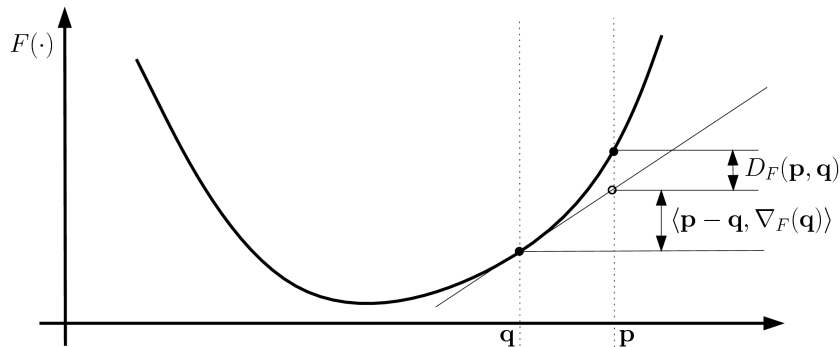
Shannon cross-entropy and entropy of EF [24]:

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

Bregman divergence: Geometric interpretation (I)

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

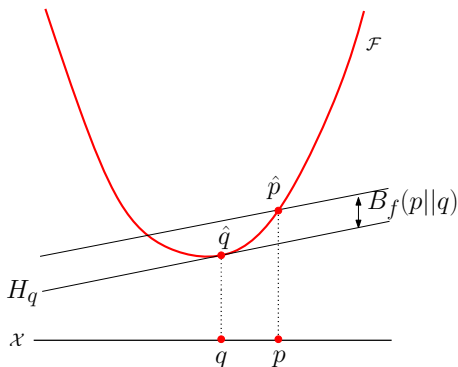
$$D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle$$



Bregman divergence: Geometric interpretation (II)

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

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

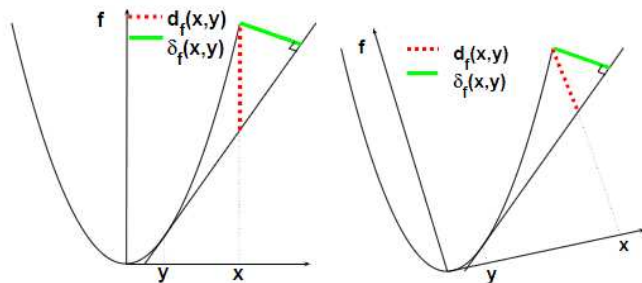


$B_f(\cdot||q)$: vertical distance between the hyperplane H_q tangent to \mathcal{F} at lifted point \hat{q} , and the translated hyperplane at \hat{p} .

total Bregman divergence (tBD)

By analogy to **least squares** and **total least squares**
total Bregman divergence (tBD) [13, 36, 14]

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



Proved **statistical robustness** of tBD.

Bregman sided centroids [23, 19]

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

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

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

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

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

→ Generalized Kolmogorov-Nagumo f -means.

Bregman divergences B_F and ∇F -means

Bijection quasi-arithmetic means (∇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 $\frac{n}{\sum_{j=1}^n \frac{1}{x_j}}$

∇F strictly increasing (like cumulative distribution functions)

Bregman sided centroids [23]

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

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

$$C : \bar{\theta} = \frac{1}{n} \sum_{i=1}^n \theta_i$$

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

$$C' : \bar{\eta} = \frac{1}{n} \sum_{i=1}^n \eta_i$$

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

Bregman information [23]

Bregman information = minimum of loss function

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

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 [5]

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

$$L_F(\mathcal{P} : \mathcal{C}) = \sum_{P \in \mathcal{P}} B_F(P : \mathcal{C})$$
$$B_F(P : \mathcal{C}) = \min_{i \in \{1, \dots, k\}} B_F(P : C_i)$$

→ generalize Lloyd' s quadratic error in Vector Quantization (VQ)

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

$I_F(\mathcal{P})$ → total Bregman information

$I_F(\mathcal{C})$ → between-cluster Bregman information

$L_F(\mathcal{P} : \mathcal{C})$ → within-cluster Bregman information

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

Bregman k -means clustering [5]

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

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: $C_1 = P_1, \dots, C_k = P_k$
- ▶ Repeat until convergence
 - ▶ **Assign** point P_i to its closest centroid:

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

- ▶ **Update** cluster centroids by taking their center of mass:

$$C_i = \frac{1}{|C_i|} \sum_{P \in C_i} P.$$

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 $\{1, \dots, n\}$
- ▶ While $|\mathcal{C}| < k$
 - ▶ Choose $P \in \mathcal{P}$ with probability
$$\frac{B_F(P:\mathcal{C})}{\sum_{i=1}^n B_F(P_i:\mathcal{C})} = \frac{B_F(P:\mathcal{C})}{L_F(\mathcal{P}:\mathcal{C})}$$

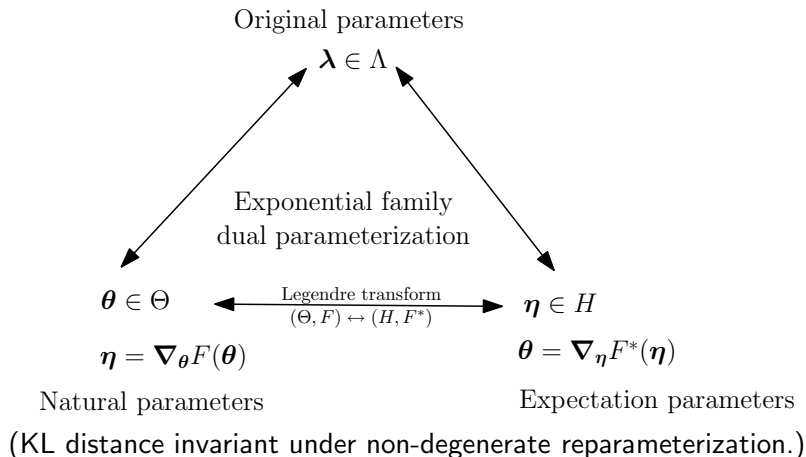
→ Yields a $O(\log k)$ approximation factor (with high probability).
Constant in $O(\cdot)$ depends on ratio of min/max $\nabla^2 F$.

Exponential family mixtures: Dual parameterizations

A finite *weighted point set* $\{(w_i, \theta_i)\}_{i=1}^k$ in a *statistical manifold*.

Many coordinate systems but two natural for computing:

- ▶ usual λ -parameterization or $\text{map} \circ \lambda$,
- ▶ natural θ -parameterization and dual η -parameterization.



Maximum Likelihood Estimator (MLE)

Given n iid. observations x_1, \dots, x_n

Maximum Likelihood Estimator

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_F(x_i; \theta) = \operatorname{argmax}_{\theta \in \Theta} e^{\sum_{i=1}^n \langle t(x_i), \theta \rangle - F(\theta) + k(x_i)}$$

is **unique** maximum since $\nabla^2 F \succ 0$. MLE equation:

$$\nabla F(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n t(x_i)$$

MLE is **consistent**, **efficient** with **asymptotic normal distribution**:

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

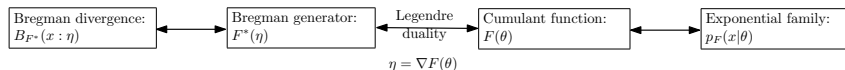
Fisher information matrix for exponential families:

$$I(\theta) = \operatorname{var}[t(X)] = \nabla^2 F(\theta) = (\nabla^2 F^*(\eta))^{-1}$$

MLE may be biased (eg, normal distributions).

→ called **observed point** \hat{P} in information geometry.

Duality Bregman \leftrightarrow Exponential families [5]



An exponential family...

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

has the **log-density** interpreted as a **Bregman divergence**:

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

Exponential families \Leftrightarrow Bregman divergences: Examples

Identify iso-distance contour as iso-probability contour
(Bregman divergences always convex on rhs.)

$F(x)$	$p_F(x \theta)$	\Leftrightarrow	B_{F^*}
Generator	Exponential Family	\Leftrightarrow	Dual Bregman divergence
x^2	Spherical Gaussian	\Leftrightarrow	Squared loss
$x \log x$	Multinomial	\Leftrightarrow	Kullback-Leibler divergence
$x \log x - x$	Poisson	\Leftrightarrow	I -divergence
$-\log x$	Geometric	\Leftrightarrow	Itakura-Saito divergence
$\log X $	Wishart	\Leftrightarrow	log-det/Burg matrix div. [37]

Maximum likelihood estimator revisited

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n p_F(x_i; \theta)$$

$$\max_{\theta} \sum_{i=1}^n (\langle t(x_i), \theta \rangle - F(\theta) + k(x_i))$$

$$\max_{\theta} \sum_{i=1}^n -B_{F^*}(t(x_i) : \eta) + \underbrace{F^*(t(x_i)) + k(x_i)}_{\text{constant}}$$

$$\equiv \min_{\theta} \sum_{i=1}^n B_{F^*}(t(x_i) : \boxed{\eta})$$

Right-sided Bregman centroid = center of mass:

$$\boxed{\hat{\eta} = \frac{1}{n} \sum_{i=1}^n t(x_i)}$$

η -MLE is center of mass of sufficient statistics $\{y_i = t(x_i)\}_{i=1}^n$.

Learning a mixture using the Expectation-Maximization [5]

- ▶ EM increases monotonically the *expected complete likelihood* L (or log-likelihood function l). (Marginalize the hidden variables z_i 's)
- ▶ EM needs an initialization Θ_0 . (Usually by k-means: E.g., for each cluster we fit a Gaussian centered at the cluster centroid with covariance matrix the covariance of the cluster, and weight the relative proportion of points in that cluster.)
- ▶ EM needs a stopping criterion. EM keeps improving the expected log-likelihood. Need to *break* the loop when the difference of log-likelihood between successive iterations $<$ threshold.

Learning a mixture using the Expectation-Maximization [5, 13]

EM for EFMM is equivalent to a Bregman soft clustering.
Bregman EM soft clustering algorithm on $\{x_1, \dots, x_n\}$:

Initialization. Set $\{w_i, \eta_i\}_{i=1}^k$ with $\sum_{i=1}^k w_i = 1$

Loop until improvement $<$ threshold.

Expectation. (compute posterior probabilities)

For all observations x

For all model components i :

$$\Pr(i|x) = \frac{w_i e^{-B_{F^*}(x; \eta_i)}}{\sum_{j=1}^k w_j e^{-B_{F^*}(x; \eta_j)}}$$

Maximization. For all model components i

$$w_i = \frac{1}{n} \sum_{j=1}^n \Pr(i|x_j)$$
$$\eta_i = \frac{\sum_{j=1}^n \Pr(i|x_j) x_j}{\sum_{j=1}^n \Pr(i|x_j)} \rightarrow \text{barycenter}$$

Monotonous convergence of the expected complete likelihood.

!!! But sampling variates is a doubly stochastic process... !!!

k -MLE for EFMM = Bregman Hard Clustering [18]

Bijection exponential families (distributions) \leftrightarrow Bregman distances

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

k -MLE (F) = Bregman hard k -means for F^* + cross-entropy minimization for weights

Complete log-likelihood:

$$\max_{\Theta} \sum_{i=1}^n \sum_{j=1}^k \delta_j(z_i) (\log p_F(x_i | \theta_j) + \log w_j)$$

$$\min_H \sum_{i=1}^n \sum_{j=1}^k \delta_j(z_i) \left((B_{F^*}(t(x_i) : \eta_j) - \log w_j) - \underbrace{k(x_i) - F^*(t(x_i))}_{\text{constant}} \right)$$

$$\equiv \min_H \sum_{i=1}^n \min_{j=1}^k B_{F^*}(t(x_i) : \eta_j) - \log w_j$$

\rightarrow guarantees the (local) convergence of the *complete likelihood* of k -MLE. (Assign a sample to a unique cluster: Hard clustering).

k -MLE for EFMMs [18]

- ▶ **0. Initialization:** $\forall i \in \{1, \dots, k\}$, let $w_i = \frac{1}{k}$ and $\eta_i = t(x_i)$ (initialization is discussed later on).

- ▶ **1. Assignment:**

$$\forall i \in \{1, \dots, n\}, z_i = \operatorname{argmin}_{j=1}^k B_{F^*}(t(x_i) : \eta_j).$$

Let $\mathcal{C}_i = \{x_j | z_j = i\}$, $\forall i \in \{1, \dots, k\}$ be the cluster partition

- ▶ **2. Update the η -parameters:**

$$\forall i \in \{1, \dots, k\}, \eta_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} t(x).$$

Goto step 1 unless local convergence of the complete likelihood is reached.

- ▶ **3. Update the weights:** $\forall i \in \{1, \dots, k\}$, $w_i = \frac{1}{n} |\mathcal{C}_i|$.

Goto step 1 unless local convergence of the complete likelihood is reached.

→ Steps 2 and 3 iterated until convergence: k -MLE = Hard EM

→ Can use **other k -means heuristics** (like Hartigan greedy swap)

Further generalization of k -MLE

Each mixture component can have its own exponential family

Infinitely many families of exponential families:

- ▶ Weibull (incl. Rayleigh or exponential),
- ▶ generalized Gaussians (incl. normal, Laplace, uniform).

$$p(x; \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\frac{|x - \mu|^\beta}{\alpha}\right)$$

with $\alpha > 0$ (scale parameter) and $\beta > 0$ (shape parameter).

Apply k -MLE by adding at each round a component family selection (eg., select the best β for each component).

k-MLE for mixtures of generalized Gaussians, ICPR, 2012. [34]

k-MLE++ [18]

- ▶ k -MLE++ = Bregman F^* k -means++ initialization
Guaranteed approximation on the best complete average log-likelihood.
→ Single step mixture learning (fast and good)
- ▶ **Indivisibility**: Robustness when identifying statistical mixture models? Which k ?

$$\forall k \in \mathbb{N}, N(\mu, \sigma^2) = \sum_{i=1}^k N\left(\frac{\mu}{k}, \frac{\sigma^2}{k}\right)$$

(add small perturbations → we should cluster MMs to get compact high quality equivalent MMs)

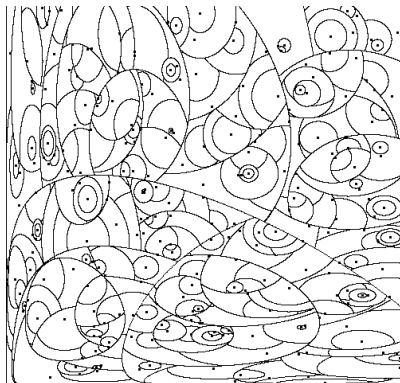
- ▶ → Choose large k (like $k = n$ for Kernel Density Estimators), and simplify MMs [9, 33]

Speeding-up k -MLE... Fast assignment

- ▶ Proximity data-structures for Bregman k -means:

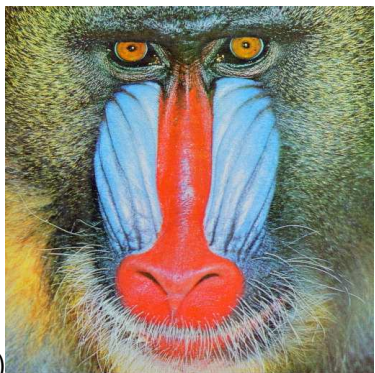
$$C_i = \{P \in \mathcal{P} \mid B_F(P : C_i) \leq B_F(P : C_j) \forall j \neq i\}$$

- ▶ Bregman Voronoi diagrams [6]
- ▶ Bregman Nearest Neighbors: ball trees [30] or vantage point trees [29].

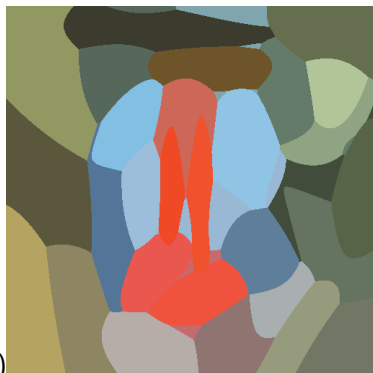


Anisotropic Voronoi diagram (for MVN MMs) [12, 15]

From the source color image (a), we build 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)

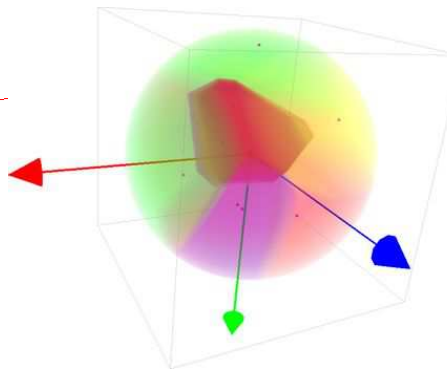
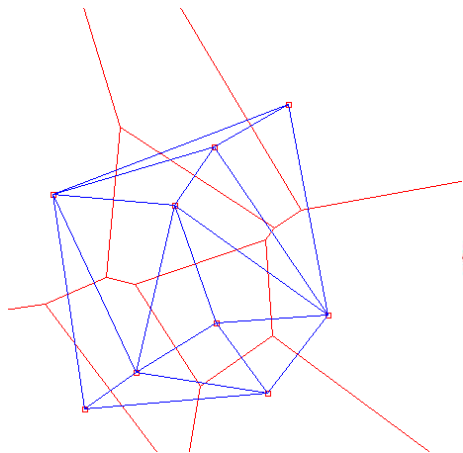


(a)



(b)

Voronoi diagrams



Voronoi diagram, dual \perp Delaunay triangulation (general position)

Bregman dual bisectors: Hyperplanes & hypersurfaces [6, 22, 25]

Right-sided bisector: \rightarrow Hyperplane (θ -hyperplane)

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

H_F :

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

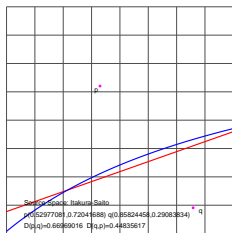
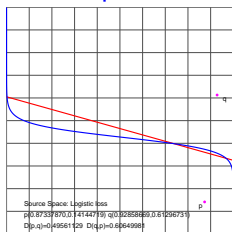
Left-sided bisector: \rightarrow Hypersurface (η -hyperplane)

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

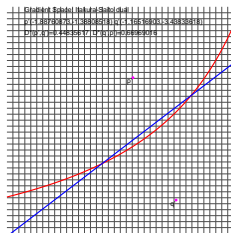
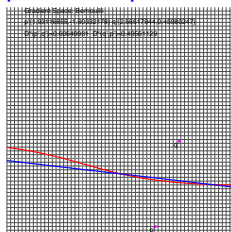
$$H'_F : \langle \nabla F(x), q - p \rangle + F(p) - F(q) = 0$$

Visualizing Bregman bisectors

Primal coordinates θ
natural parameters



Dual coordinates η
expectation parameters



Bregman Voronoi diagrams as minimization diagrams [6]

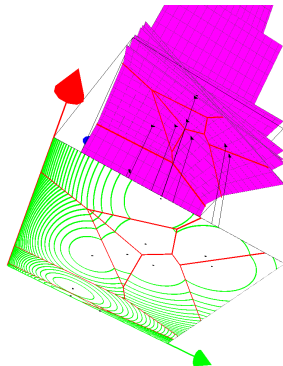
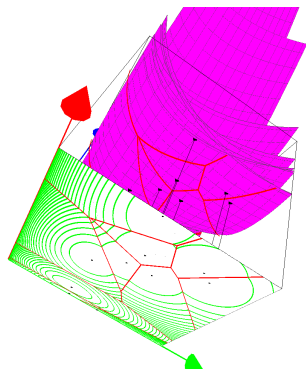
A subclass of affine diagrams which have all non-empty cells .

Minimization diagram of the n functions

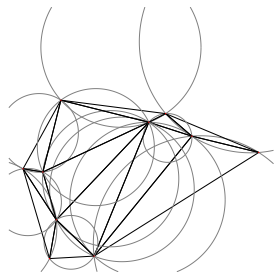
$$D_i(x) = B_F(x : p_i) = F(x) - F(p_i) - \langle x - p_i, \nabla F(p_i) \rangle.$$

\equiv minimization of n linear functions:

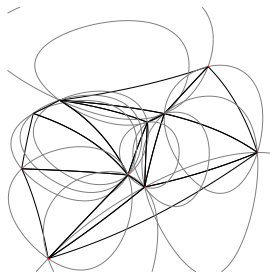
$$H_i(x) = (p_i - x)^T \nabla F(q_i) - F(p_i)$$



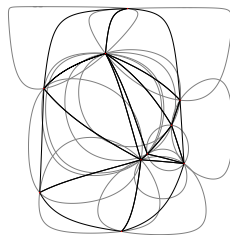
Bregman dual Delaunay triangulations



Delaunay



Exponential



Hellinger-like

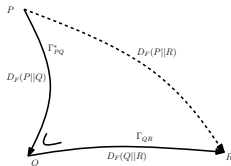
- ▶ empty Bregman sphere property,
- ▶ geodesic triangles.

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

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

→ \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:

$$(pq)_\theta = \{\theta = \theta_p + (1 - \lambda)\theta_q \mid \lambda \in [0, 1]\}$$

$$(pq)_\eta = \{\eta = \eta_p + (1 - \lambda)\eta_q \mid \lambda \in [0, 1]\}$$

Bisectors:

$$B_\theta(p, q) : \langle x, \theta_q - \theta_p \rangle + F(\theta_p) - F(\theta_q) = 0$$

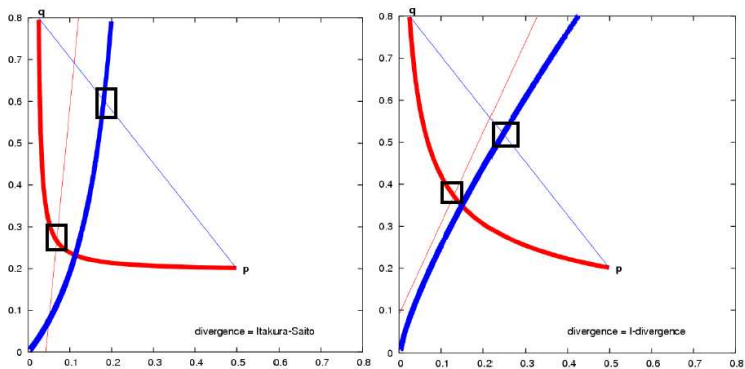
$$B_\eta(p, q) : \langle x, \eta_q - \eta_p \rangle + F^*(\eta_p) - F^*(\eta_q) = 0$$

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

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

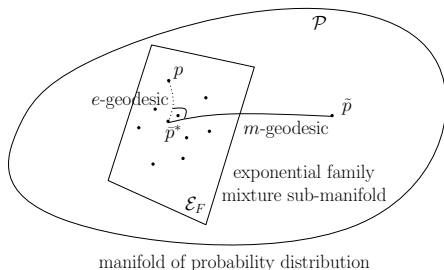


Simplifying mixture: Kullback-Leibler projection theorem

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]:

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



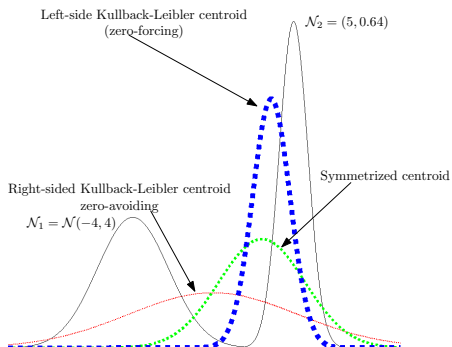
Right-sided KL centroid = Left-sided Bregman centroid

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:

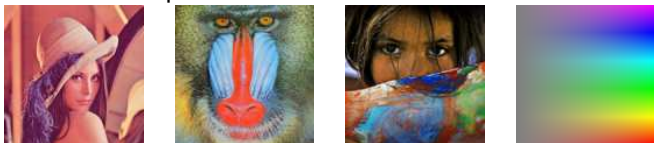
- ▶ Right-sided entropic centroid: **zero-avoiding** (cover support of pdfs.)
- ▶ Left-sided entropic centroid: **zero-forcing** (captures highest mode).



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).**

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

Skew Jensen divergence [19, 27]

$$J_F^{(\alpha)}(p; q) = \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) = \text{BR}_F^{(\alpha)}(p; q)$$

(Jeffreys and Jensen-Shannon symmetrization of Kullback-Leibler)

(Burbea-Rao centroids (α -skewed Jensen centroids))

Minimum average divergence

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

Equivalent to minimize:

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

Sum $E = F + G$ of convex $F +$ concave G function \Rightarrow
Convex-ConCave Procedure (CCCP)

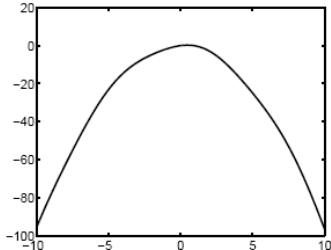
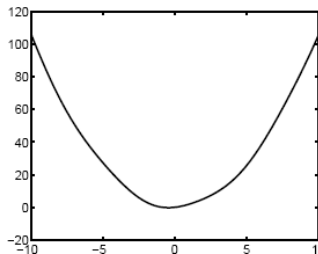
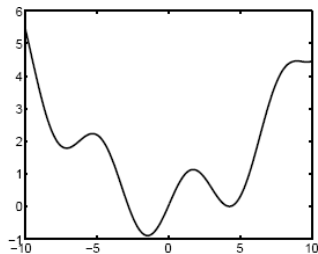
Start from arbitrary c_0 , and iteratively update as:

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

\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

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

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

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

Unique GLOBAL minimum when divergence is separable [19].

Unique GLOBAL minimum for matrix mean [21] for the logDet divergence.

Statistical divergences (Recap.)

- ▶ Kullback-Leibler is a f -divergence (\rightarrow statistical invariance, information monotonicity, curved geometry)
- ▶ Kullback-Leibler of exponential families = Bregman divergences on parameters (dually flat geometry)
- ▶ Skew Jensen-divergence (Burbea-Rao, $\alpha = \frac{1}{2}$) include Bregman divergences in limit cases [19]
- ▶ No known closed form for Kullback-Leibler of mixtures. But closed-form for EFMMs with the **Cauchy-Schwarz** divergence [17]:

$$\text{CS}(P : Q) = -\log \frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2 dx \int q(x)^2 dx}},$$

Closed-Form Information-Theoretic Divergences for Statistical Mixtures, ICPR, 2012.

Summary

Computational information-geometric signal processing:

- ▶ Statistical manifold (M, g) : Rao's distance and Fisher-Rao curved riemannian geometry.
- ▶ Statistical manifold (M, g, ∇, ∇^*) : dually flat spaces, Bregman divergences, geodesics are straight lines in either θ/η parameter space.
- ▶ Clustering & learning statistical mixtures (EM=soft Bregman clustering, k -MLE, KDE simplification, hierarchical mixtures [11])
- ▶ Software library: JMEF [9] (Java), PYMEF [31] (Python)
- ▶ ... but also many other geometry to explore: Hilbertian, Finsler [3], Kähler, Wasserstein, etc. (it is easy to require non-Euclidean geometry but then [space is wild open!](#))

André Ferrari, Cédric Févotte, Cédric Richard
Caroline Daire & MAHI Organizers.
University of Nice Sophia-Antipolis, Lagrange lab,
Albert Bijaoui, OCA,
CNRS.

Collaborators: Shun-ichi Amari, Marc Arnaudon, Michel Barlaud,
Sylvain Boltz, Jean-Daniel Boissonnat, Eric Debreuve, Vincent
Garcia, Meizhu Liu Richard Nock, Paolo Piro, Olivier Schwander,
Baba C. Vemuri.

Sony Computer Science Laboratories Inc: Professors Hiroaki
Kitano and Mario Tokoro.

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 [24]:
- ▶ Rényi/Tsallis entropy and divergence [26]
- ▶ Sharma-Mittal entropy and divergence [28]. A 2-parameter family extending extensive Rényi (for $\beta \rightarrow 1$) and non-extensive Tsallis entropies (for $\beta \rightarrow \alpha$)

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

with $\alpha > 0, \alpha \neq 1, \beta \neq 1$.

- ▶ Skew Jensen and Burbea-Rao divergence [19]
- ▶ Chernoff information and divergence [16]
- ▶ Mixtures: total Least square, Jensen-Rényi, Cauchy-Schwarz divergence [17].

Statistical invariance: Markov kernel

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

(X, σ) and (X', σ') two measurable spaces.

σ : A σ -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', σ') .

p a pm. on (X, σ) induces Kp a pm., with

$$Kp(E') = \int_X K(x, E')p(dx), \forall E' \in \sigma'$$

Space of Bregman spheres and Bregman balls [6]

Dual Bregman balls (bounding Bregman spheres):

$$\text{Ball}'_F(c, r) = \{x \in \mathcal{X} \mid B_F(x : c) \leq r\}$$

and

$$\text{Ball}'_F(c, r) = \{x \in \mathcal{X} \mid B_F(c : x) \leq r\}$$

Legendre duality:

$$\text{Ball}'_F(c, r) = (\nabla F)^{-1}(\text{Ball}'_{F^*}(\nabla F(c), r))$$



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

Space of Bregman spheres: Lifting map [6]

$\mathcal{F} : x \mapsto \hat{x} = (x, F(x))$, 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 $\sigma \rightarrow \hat{\sigma}$ with supporting hyperplane

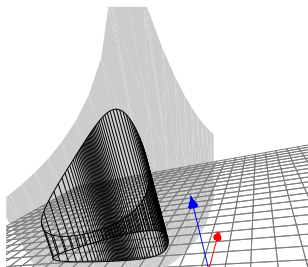
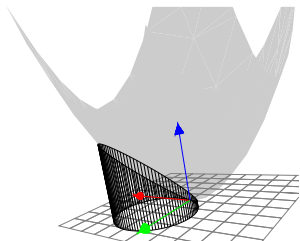
$$H_\sigma : z = \langle x - c, \nabla F(c) \rangle + F(c) + r.$$

(// 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:

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



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.



Marc Arnaudon and Frank Nielsen.

On approximating the Riemannian 1-center.

Computational Geometry, 46(1):93 – 104, 2013.



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 and Computational Geometry, 44(2):281–307, April 2010.

Bibliographic references II



Nikolai Nikolaevich Chentsov.

Statistical Decision Rules and Optimal Inferences.

Transactions of Mathematics Monograph, numero 53, 1982.

Published in russian in 1972.



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.



Vincent Garcia, Frank Nielsen, and Richard Nock.

Hierarchical Gaussian mixture model.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4070–4073, 2010.

Bibliographic references III



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, 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.



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), 2012.

Bibliographic references IV



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.



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.



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.

Bibliographic references V



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.



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.

Bibliographic references VI



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.



Paolo Piro, Frank Nielsen, and Michel Barlaud.

Tailored Bregman ball trees for effective nearest neighbors.

In *European Workshop on Computational Geometry (EuroCG)*, LORIA, Nancy, France, March 2009. IEEE.



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.



Olivier Schwander and Frank Nielsen.

Model centroids for the simplification of kernel density estimators.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 737–740, 2012.



Olivier Schwander, Frank Nielsen, Aurélien Schutz, and Yannick Berthoumieu.

k-MLE for mixtures of generalized Gaussians.

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

Bibliographic references VII



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.

10.1109/TMI.2010.2086464.



Shijun Wang and Rong Jin.

An information geometry approach for distance metric learning.

Journal of Machine Learning Research, 5:591–598, 2009.