The Centroids of Symmetrized Bregman Divergences

Frank Nielsen[1] Richard Nock[2]

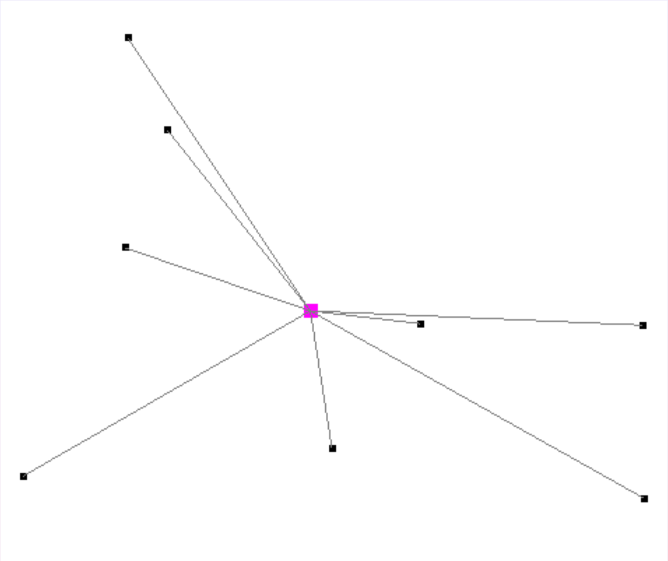
[1]Sony Computer Science Laboratories Inc., FRL
École Polytechnique, LIX
Frank.Nielsen@acm.org

[2]University of Antilles-Guyanne
CEREGMIA
Richard.Nock@martinique.univ-ag.fr

December 2007

The centroid in Euclidean geometry

Given a point set $\mathcal{P} = \{p_1, ..., p_n\}$ of $\mathbb{E}^d$, the centroid $\bar{c}$:



- Is the *center of mass*: $\bar{c} = \frac{1}{n} \sum_{i=1}^{n} p_i$,
- Minimizes $\min_{c \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{n} ||cp_i||^{\boxed{2}}$:
  MINAVG **squared** Euclidean distance optimization,
- Plays a central role in *center-based clustering* methods (k-means of Lloyd'1957)

Centroid and barycenters

Notion of centroid extends to barycenters:

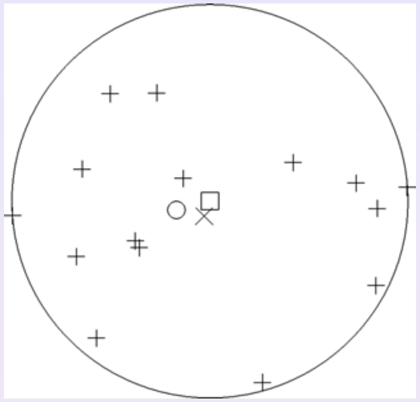
$$\bar{b}(w) = \sum_{i=1}^{n} w_i p_i$$

$\rightarrow$ barycentric coordinates for interpolation, with $||w|| = 1$.
Barycenter $\bar{b}(w)$ minimizes:

$$\min_{c \in \mathbb{R}^d} \sum_{i=1}^{n} w_i ||cp_i||^2$$

$\rightarrow$ weighted MinAvg optimization.
Intracluster min. weighted average $\sum_{i=1}^{n} w_i ||\bar{b}(w) p_i||^2$.

Center points in Euclidean geometry



- **Centroid** $\times$: robust to outliers with simple closed-form sol.
  "Mean" radius is $\frac{1}{n}\sum_{i=1}^{n}||\bar{c}p_i||^2$.
- **Circumcenter** $\square$: minimizes the radius of enclosing ball.
  Combinatorially defined by at most $d+1$ points.
  (MiniBall Welzl'1991)
  MiniMax (non-differentiable) optimization problem:

$$C = \min_c \boxed{\max_i} ||cp_i||^2$$

- MinAvg(L_2) $\rightarrow$ **Fermat-Weber point** $\circ$
  $\rightarrow$ no closed form solution.

Bregman divergences

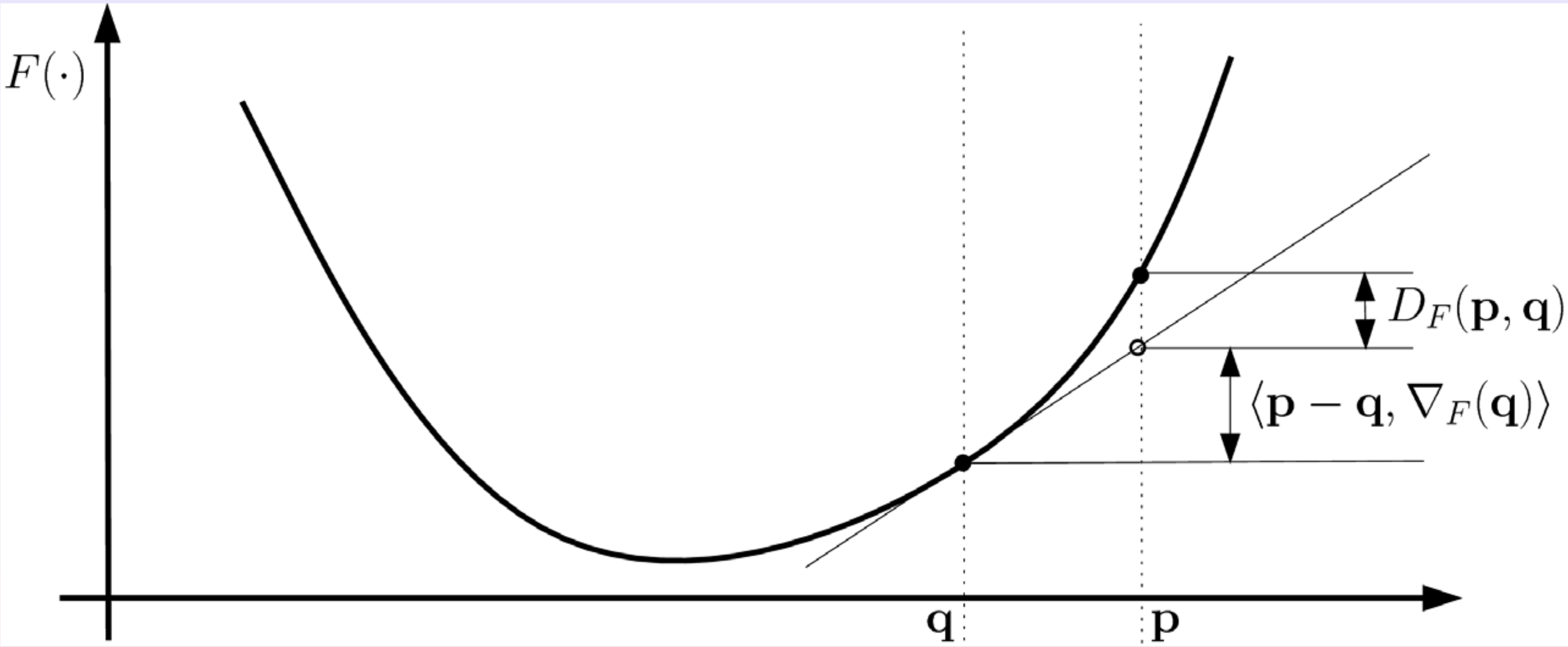
Aim at generalizing Euclidean centroids to *dually flat spaces*.

Bregman divergences D_F:
$F : \mathcal{X} \to \mathbb{R}$ *strictly convex* and *differentiable function* defined over an open *convex domain* $\mathcal{X}$:

$$D_F(p||q) = F(p) - F(q) - \langle p - q, \nabla_F(q) \rangle$$

- not a metric (symmetry and triangle inequality may fail)
- versatile family, popular in Comp. Sci.–Machine learning.

Bregman divergence: Geometric interpretation



$$D_F(p||q) = F(p) - F(q) - \langle p - q, \nabla_F(q) \rangle$$
$$D_F(p||q) \geq 0$$

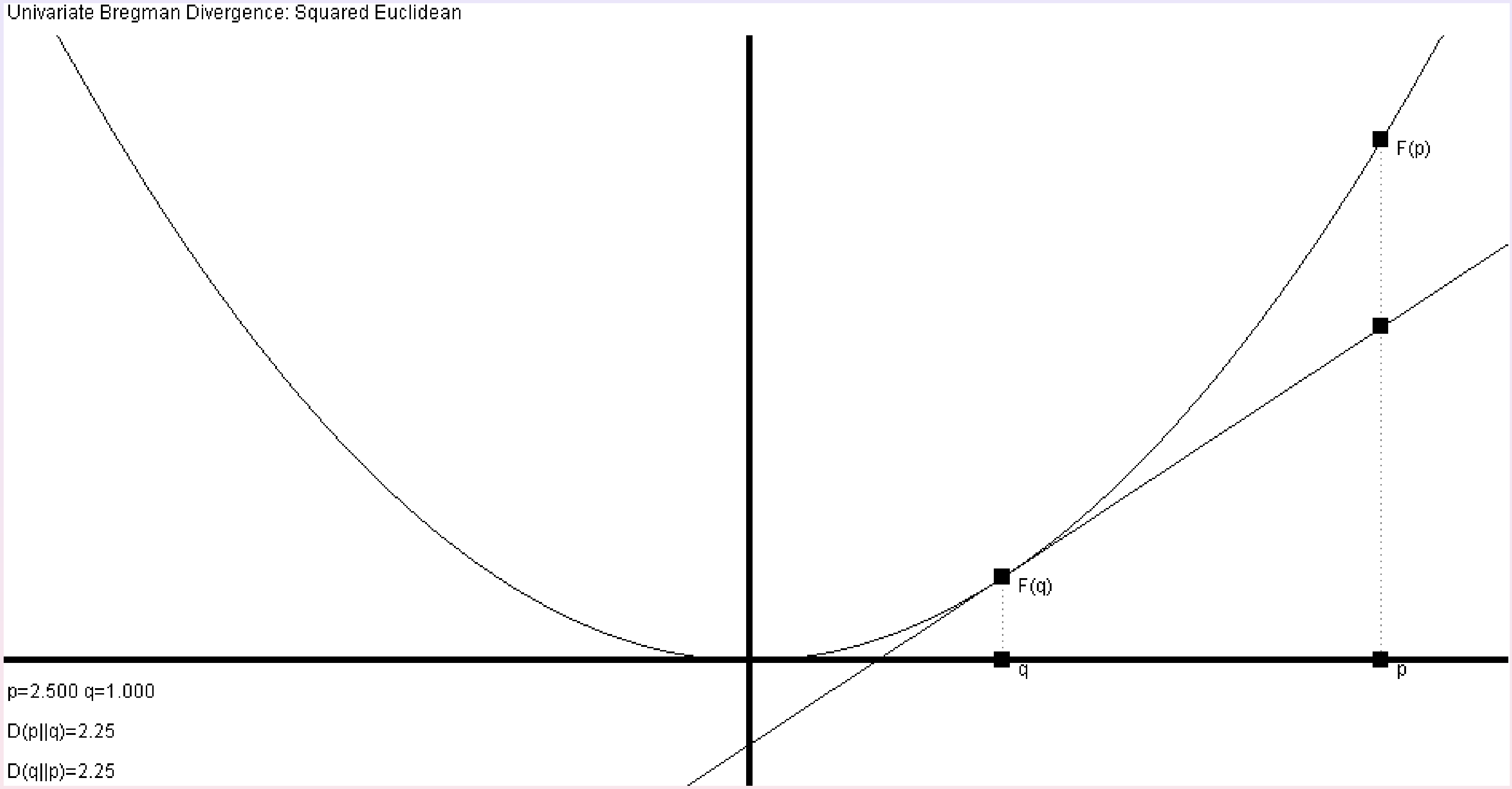
(with equality iff. $p = q$)
F: generator, contrast function or potential function (inf. geom.).

Example 1: The squared Euclidean distance

- $F(x) = x^2$: strictly convex and differentiable over $\mathbb{R}^d$ (Multivariate $F(x) = \sum_{i=1}^{d} x_i^2$, obtained coordinatewise)

$$
\begin{aligned}
D_F(p||q) &= F(p) - F(q) - \langle p - q, \nabla_F(q) \rangle \\
&= p^2 - q^2 - \langle p - q, 2q \rangle \\
&- p^2 - q^2 - 2\langle p, q \rangle + 2q^2 \\
&= \|p - q\|^2
\end{aligned}
$$

Example 1: The squared Euclidean distance



Java applet: http://www.sonycsl.co.jp/person/nielsen/BregmanDivergence/

Example 2: The relative entropy (Kullback-Leibler divergence)

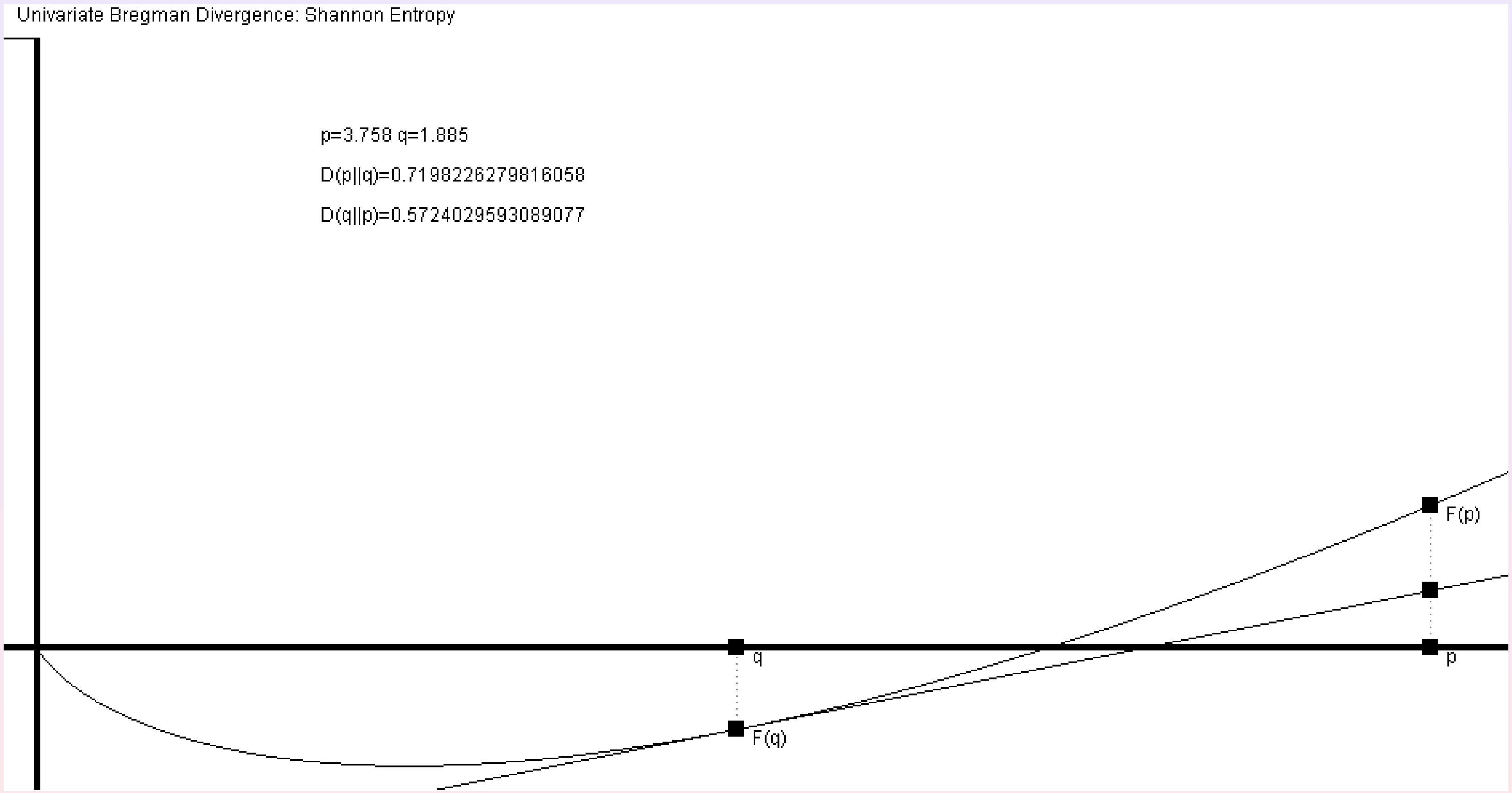
- $F(p) = \int p(x) \log p(x) \,\mathrm{d}x$ (negative Shannon entropy)
  (Discrete distributions $F(p) = \sum_x p(x) \log p(x) \,\mathrm{d}x$)

$$
\begin{aligned}
D_F(p||q) &= \int (p(x) \log p(x) - q(x) \log q(x) \\
&\quad - \langle p(x) - q(x), \log q(x) + 1 \rangle) \,\mathrm{d}x \\
&= \int p(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}x \qquad \text{(KL divergence)}
\end{aligned}
$$

Kullback-Leiber divergence also known as:
relative entropy, discrimination information or I-divergence.

(Defined either on the probability simplex or *extended* on the full positive quadrant – unnormalized pdf.)

Example 2: The relative entropy (Kullback-Leibler divergence)



Java applet: http://www.sonycsl.co.jp/person/nielsen/BregmanDivergence/

Bregman divergences: A versatile family of measures

Bregman divergences are *versatile*, suited to *mixed-type* data.

(Build mixed-type multivariate divergences *dimensionwise* using elementary uni-type divergences.)

Fact (Linearity)

Bregman divergence is a linear operator:
$\forall F_1 \in \mathcal{C}\ \forall F_2 \in \mathcal{C} \quad D_{F_1+\lambda F_2}(p||q) = D_{F_1}(p||q) + \lambda D_{F_2}(p||q)$ *for any* $\lambda \geq 0$.

Fact (Equivalence classes)

Let $G(x) = F(x) + \langle a, x\rangle + b$ *be another strictly convex and differentiable function, with* $a \in \mathbb{R}^d$ *and* $b \in \mathbb{R}$. *Then* $D_F(p||q) = D_G(p||q)$.

(Simplify Bregman generators by removing affine terms.)

Sided and symmetrized centroids from MinAvg

Right-sided and left-sided centroids:

$$c_R^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(p_i||\boxed{\mathrm{c}})$$

$$c_L^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(\boxed{\mathrm{c}}||p_i)$$

Symmetrized centroid:

$$c^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{D_F(p_i||\boxed{\mathrm{c}}) + D_F(\boxed{\mathrm{c}}||p_i)}{2}}_{S_F}$$

Symmetrized asymmetric Bregman divergence S_F is not a Bregman divergence (because of domain convexity)

Sided right-type centroid

Theorem

The right-type sided Bregman centroid c_R^F *of a set* $\mathcal{P}$ *of* n *points* $p_1, ...p_n$*, defined as the minimizer for the average right divergence* $c_R^F = \arg\min_c \sum_{i=1}^n \frac{1}{n} D_F(p_i||c) = \arg\min_c \mathrm{AVG}_F(\mathcal{P}||c)$*, is unique,* **independent** *of the selected divergence* D_F*, and coincides with the* **center of mass** $c_R^F = c_R = \bar{p} = \frac{1}{n}\sum_{i=1}^n p_i$*.*

Proof:
Start with the function to minimize:

$$\mathrm{AVG}_F(\mathcal{P}||q) = \sum_{i=1}^n \frac{1}{n} D_F(p_i||q)$$

$$\mathrm{AVG}_F(\mathcal{P}||q) = \sum_{i=1}^{n} \frac{1}{n} (F(p_i) - F(q) - < p_i - q, \nabla F(q) >)$$

$$\begin{aligned}
\mathrm{AVG}_F(\mathcal{P}, q) &= \left(\sum_{i=1}^{n} \frac{1}{n} F(p_i) - F(\bar{p}) \right) + \left(F(\bar{p}) - F(q) - \sum_{i=1}^{n} \frac{1}{n} < p_i - q, \nabla F(q) > \right), \\
&= \left(\sum_{i=1}^{n} \frac{1}{n} F(p_i) - F(\bar{p}) \right) + \left(F(\bar{p}) - F(q) - \left\langle \sum_{i=1}^{n} \frac{1}{n} (p_i - q), \nabla F(q) \right\rangle \right), \\
&= \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} F(p_i) - F(\bar{p}) \right)}_{\text{Independent of } q} + D_F(\bar{p}||q).
\end{aligned}$$

- Minimized for $q = \bar{p}$ since $D_F(\bar{p}||q) \geq 0$ with equality iff $q = \bar{p}$
- Information radius: $\mathrm{JS}_F(\mathcal{P}) = \frac{1}{n} \sum_{i=1}^{n} F(p_i) - F(\bar{p}) \geq 0$
  $\rightarrow$ Jensen F remainder aka. **Burbea-Rao divergences**
  $\leftarrow$ Generalize Jensen-Shannon divergences (Lin'1991)

Dual Bregman divergence

Legendre-Fenchel (slope) transformation $F \rightarrow G = \mathcal{L}F$:

$$G(y) = \sup_{x \in \mathcal{X}} \{ < y, x > - F(x) \}$$

$G = F^*$ (with $G^* = F^{**} = F$), minimized for $y = x' \stackrel{\text{def}}{=} \nabla F(x)$.

$$F^*(x') = < x, x' > - F(x)$$

Dual Bregman divergence

$$\begin{aligned} D_F(p||q) &= F(p) + F^*(\nabla F(q)) - < p, \nabla F(q) > \\ &= F(p) + F^*(q') - < p, q' > \\ &= D_{F^*}(q'||p') \end{aligned}$$

More details in "Bregman Voronoi diagrams", arXiv:0709.2196

Left-sided Bregman centroid

Theorem

The left-type sided Bregman centroid c_L^F, defined as the minimizer for the average left divergence $c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_L^F(c||\mathcal{P})$, is the unique point $c_L^F \in \mathcal{X}$ such that $c_L^F = (\nabla F)^{-1}(\bar{p}') = (\nabla F)^{-1}(\sum_{i=1}^n \nabla F(p_i))$, where $\bar{p}' = c_R^{F^*}(\mathcal{P}_F{}')$ is the center of mass for the gradient point set $\mathcal{P}_F{}' = \{p_i' = \nabla F(p_i) \mid p_i \in \mathcal{P}\}$.

$$c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(\boxed{\mathrm{c}}||\mathcal{P}) \Leftrightarrow \arg\min_{c'\in\mathcal{X}}^{\prime} \mathrm{AVG}_{F^*}(\mathcal{P}_F{}'||\boxed{\mathrm{c'}}) = c_R^{F^*}(\mathcal{P}_F')$$

The information radii of sided Bregman centroids are equal:

$$\mathrm{AVG}_F(\mathcal{P}||c_R^F) = \mathrm{AVG}_F(c_L^F||\mathcal{P}) = \mathrm{JS}_F(\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n F(p_i) - F(\bar{p}) > 0$$

is the F-Jensen-Shannon divergence for the **uniform weight**

Generalized means

A sequence $\mathcal{V}$ of n real numbers $V = \{v_1, ..., v_n\}$

$$M(\mathcal{V}; f) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(v_i)\right)$$

For example, Pythagoras' means:

- Arithmetic: $f(x) = x$ (average)
- Geometric: $f(x) = \log x$ (central tendency)
- Harmonic: $f(x) = \frac{1}{x}$ (average of rates)

$$\min_i x_i \leq M(\mathcal{V}; f) \leq \max_i x_i$$

min and max: **power means** ($f(x) = x^p$) for $p \rightarrow \pm\infty$

Bijection: Bregman divergences and means

Bijection: Bregman divergence $D_F \leftrightarrow \nabla F$-means

$$M(\mathcal{S}; f) = M(\mathcal{S}; af + b)\ \forall a \in \mathbb{R}_*^+ \text{ and } \forall b \in \mathbb{R}$$

Recall one property of Bregman divergences:

Fact (Equivalence classes)

Let $G(\mathbf{x}) = F(\mathbf{x}) + \langle \mathbf{a}, \mathbf{x} \rangle + b$ be another strictly convex and differentiable function, with $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then $D_F(\mathbf{p}||\mathbf{q}) = D_G(\mathbf{p}||\mathbf{q})$.

Left- and right-sided Bregman barycenters

Left- and right-sided Bregman centroids extend to barycenters.

Theorem

Bregman divergences are in bijection with generalized means. The right-type barycenter $b_R^F(w)$ *is independent of* F *and computed as the weighted arithmetic mean on the point set, a generalized mean for the identity function:* $b_R^F(\mathcal{P};w) = b_R(\mathcal{P};w) = M(\mathcal{P};x;w)$ *with* $M(\mathcal{P};f;w) \stackrel{\mathrm{def}}{=} f^{-1}(\sum_{i=1}^n w_i f(v_i))$. *The left-type Bregman barycenter* b_L^F *is computed as a generalized mean on the point set for the gradient function:* $b_L^F(\mathcal{P}) = M(\mathcal{P};\nabla F;w)$.
The information radius of sided barycenters is: $\mathrm{JS}_F(\mathcal{P};w) = \sum_{i=1}^d w_i F(p_i) - F(\sum_{i=1}^d w_i p_i)$.

Symmetrized Bregman centroid

Symmetrized centroid defined from the minimum average optimization problem:

$$c^F = \arg\min_{c \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} \frac{D_F(c||p_i) + D_F(p_i||c)}{2} = \arg\min_{c \in \mathcal{X}} \mathrm{AVG}(\mathcal{P}; c)$$

Lemma

The symmetrized Bregman centroid c^F is **unique** *and obtained by minimizing $\min_{q \in \mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$:*

$$c^F = \arg\min_{q \in \mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F).$$

$\longrightarrow$ Minimization problem depends only on sided centroids.

Proof

$$
\begin{aligned}
\mathrm{AVG}_F(\mathcal{P}||q) &= \left(\sum_{i=1}^{n} \frac{1}{n} F(p_i) - F(\bar{p})\right) + D_F(\bar{p}||q) \\
\mathrm{AVG}_F(q||\mathcal{P}) &= \mathrm{AVG}_{F^*}({\mathcal{P}_F}'||q') \\
&= \left(\sum_{i=1}^{n} \frac{1}{n} F^*(p_i') - F^*(\bar{p'})\right) + D_{F^*}(\bar{p'_F}||q'_F)
\end{aligned}
$$

But $D_{F^*}(\bar{p'_F}||q'_F) = D_{F^{**}}(\nabla F^* \circ \nabla F(q)||\nabla F^*(\sum_{i=1}^n \nabla F(p_i))) = D_F(q||c_L^F)$ since $F^{**} = F$, $\nabla F^* = \nabla F^{-1}$ and $\nabla F^* \circ \nabla F(q) = q$.

$$
\arg\min_{c\in\mathcal{X}} \frac{1}{2}\left(\mathrm{AVG}_F(\mathcal{P}||q) + \mathrm{AVG}_F(q||\mathcal{P})\right) \Longleftrightarrow \arg\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)
$$

(removing all terms independent of q)

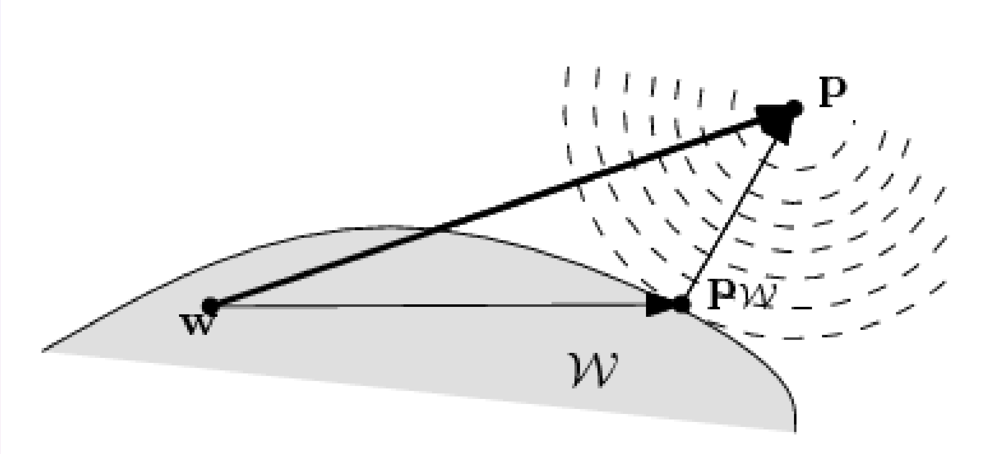
Geometric characterization

Theorem

The symmetrized Bregman centroid c^F is uniquely defined as the minimizer of $D_F(c_R^F||q) + D_F(q||c_L^F)$. It is defined **geometrically** *as $c^F = \Gamma_F(c_R^F, c_L^F) \cap M_F(c_R^F, c_L^F)$, where $\Gamma_F(c_R^F, c_L^F) = \{(\nabla F)^{-1}((1-\lambda)\nabla F(c_R^F) + \lambda \nabla F(c_L^F)) \mid \lambda \in [0,1]\}$ is the geodesic linking c_R^F to c_L^F, and $M_F(c_R^F, c_L^F)$ is the mixed-type Bregman bisector: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid D_F(c_R^F||x) = D_F(x||c_L^F)\}$.*

Proof by contradiction using Bregman Pythagoras' theorem.

Generalized Bregman Pythagoras' theorem

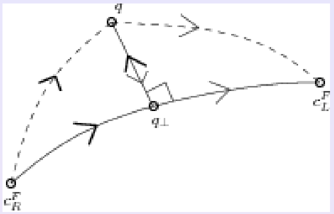


Projection $\mathbf{p}_{\mathcal{W}}$ of point $\mathbf{p}$ to a *convex subset* $\mathcal{W} \subseteq \mathcal{X}$.

$$D_F(\mathbf{w}||\mathbf{p}) \geq D_F(\mathbf{w}||\mathbf{p}_{\mathcal{W}}) + D_F(\mathbf{p}_{\mathcal{W}}||\mathbf{p})$$

with equality for and only for *affine sets* $\mathcal{W}$

Proof by contradiction



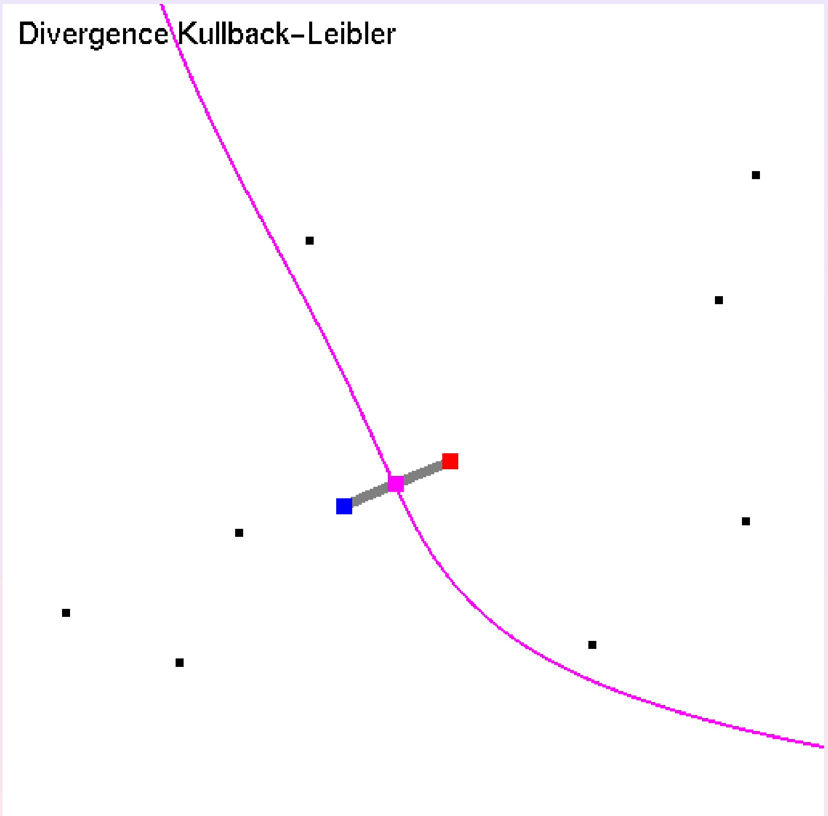
Bregman projection: $q_\perp = \arg\min_{t \in \Gamma(c_R^F, c_L^F)} D_F(t||q)$

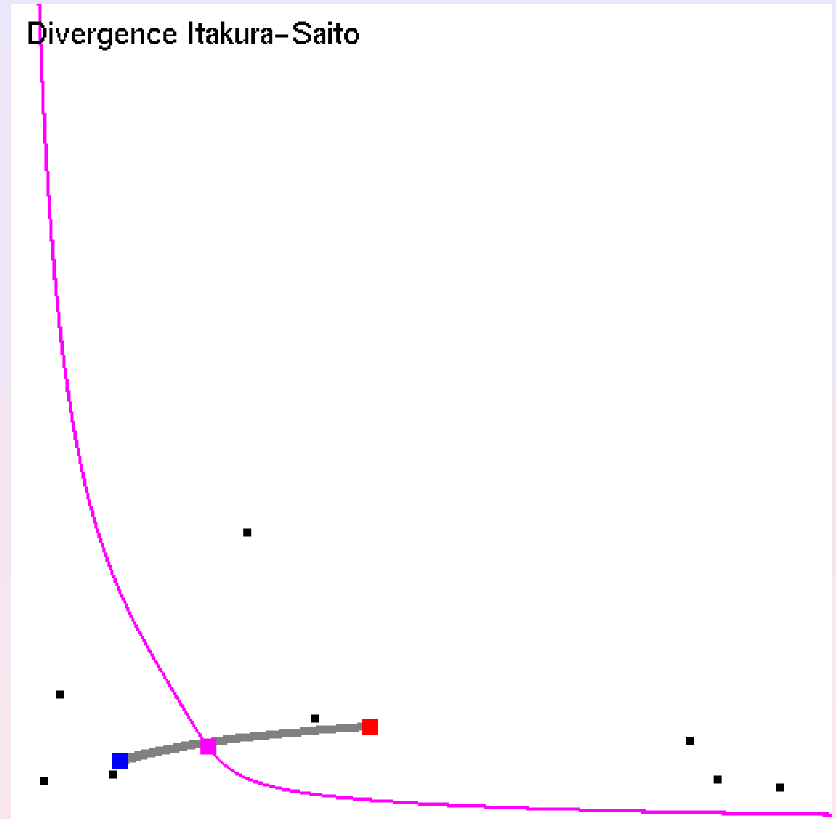
$$\begin{aligned} D_F(p||q) &\geq D(p||P_\Omega(q)) + D_F(P_\Omega(q)||q) \\ P_\Omega(q) &= \arg\min_{\omega \in \Omega} D_F(\omega||q) \\ D_F(c_R^F||q) &\geq D_F(c_R||q_\perp) + D_F(q_\perp||q) \\ D_F(q||c_L^F) &\geq D_F(q||q_\perp) + D_F(q_\perp||C_L^F) \end{aligned}$$

$D_F(c_R^F||q) + D_F(q||c_L^F) \geq D_F(c_R^F||q_\perp) + D_F(q_\perp||c_L^F) + (D_F(q_\perp||q) + D_F(q||q_\perp))$

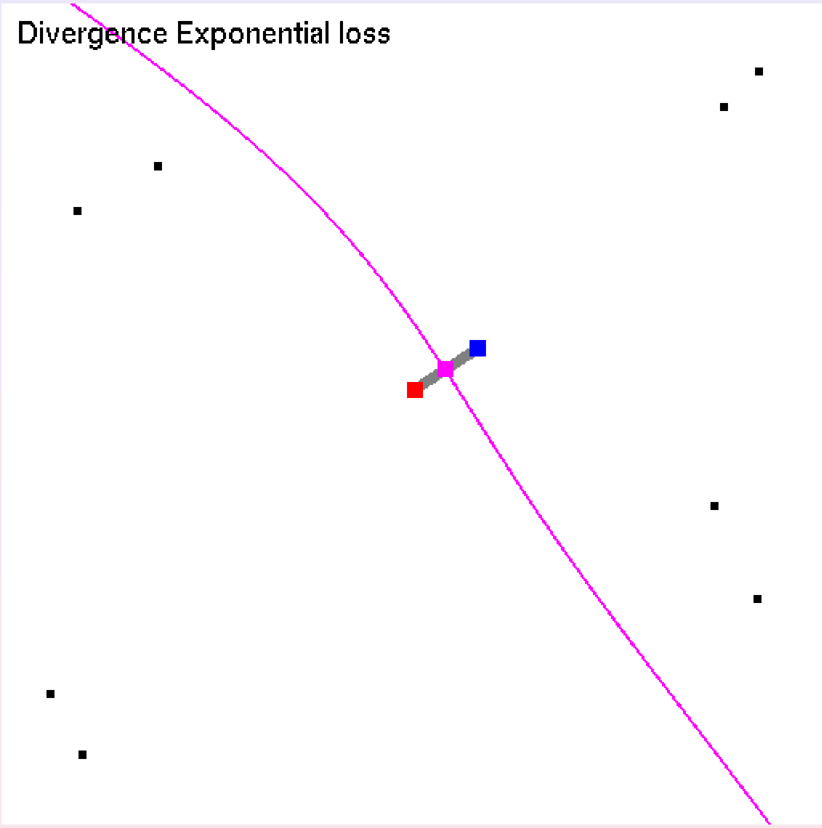
But $D_F(q_\perp||q) + D_F(q||q_\perp) > 0$ yields contradiction.
Proved mixed-type bisector by moving q along the geodesic while reducing $|D_F(c_R^F||q) - D_F(q||c_L^F)|$.

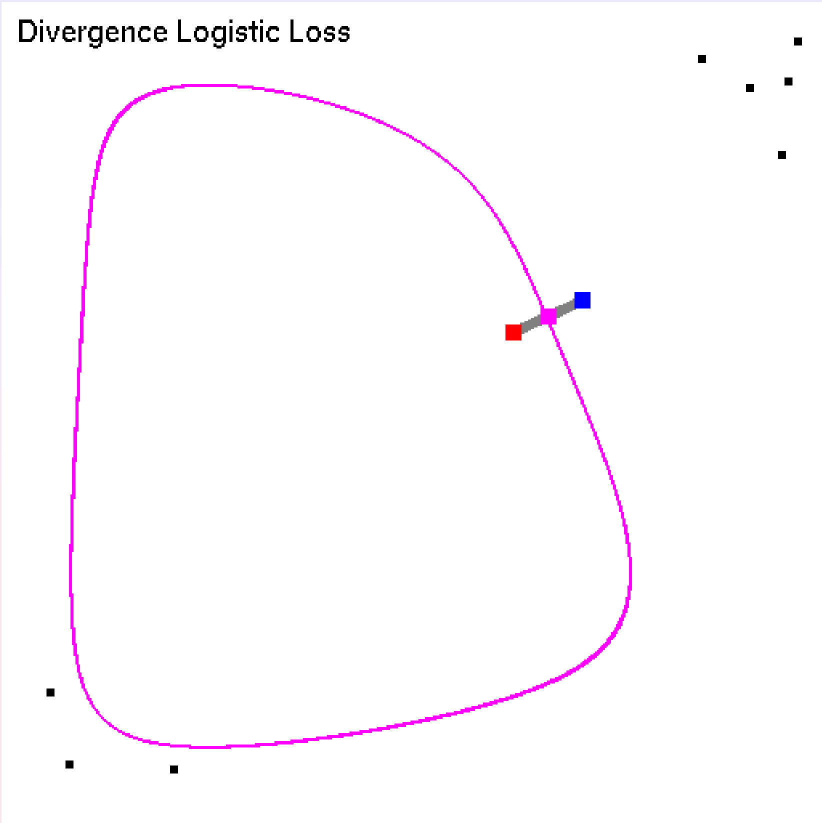
A few examples: Kullback-Leibler & Itakura-Saito





A few examples: Exponential & Logistic losses





Non-linear mixed-type bisector

$$M_F(p,q) = \{ \quad x \in \mathcal{X} \mid F(p) - F(q) - 2\boxed{F(x)} - \\ < p, x' > + < x, x' > + < x, q' > - < q, q' > = 0\}$$

$\rightarrow$ non-closed form solution.

Corollary

The symmetrized Bregman divergence minimization problem is both lower and upper bounded as follows:

$$\mathrm{JS}_F(\mathcal{P}) \leq \mathrm{AVG}_F(\mathcal{P}; c^F) \leq D_F(c_R^F || c_L^F)$$

Geodesic-walk dichotomic approximation algorithm

ALGORITHM:
Geodesic is parameterized by $\lambda \in [0, 1]$.
Start with $\lambda_m = 0$ and $\lambda_M = 1$.

Geodesic walk. Compute interval midpoint $\lambda_h = \frac{\lambda_m + \lambda_M}{2}$ and corresponding geodesic point

$$q_{\lambda_h} = (\nabla F)^{-1}((1 - \lambda_h)\nabla F(c_R^F) + \lambda_h \nabla F(c_L^F)),$$

Mixed-type bisector side. Evaluate the sign of $D_F(c_R^F || q_h) - D_F(q_h || c_L^R)$, and

Dichotomy. Branch on $[\lambda_h, \lambda_M]$ if the sign is negative, or on $[\lambda_m, \lambda_h]$ otherwise.

Precision and number of iterations as a function of $h_F = \max_{x \in \Gamma(c_R^F, c_L^F)} ||H_F(x)||^2$. (Bregman balls, ECML'05)

Applications

Applications of the dichotomic geodesic walk algorithm

- Symmetrized non-parametric Kullback-Leibler (SKL),
- Symmetrized Kullback-Leibler of multivariate normals, (parametric distributions)
- Symmetrized Bregman-Csiszár centroids,
  $\longrightarrow$ include J-divergence and COSH distance (Itakura-Saito symmetrized divergence).

The symmetrized Kullback-Leibler divergence

For two discrete probability mass functions p and q:

$$\mathrm{KL}(p||q) = \sum_{i=1}^{d} p^{(i)} \log \frac{p^{(i)}}{q^{(i)}}$$

For continuous distributions

$$\mathrm{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)}.$$

Symmetric Kullback-Leibler is called J-divergence.

$\boxed{\text{Observe that finite discrete distributions are parametric.}}$

$\rightarrow$ Degree of freedom is cardinal of sample space minus one.

Exponential families and divergences

Exponential families in statistics have the following pdf.:

$$\exp(< \theta, t(x) > -F(\theta) + C(x))$$

- θ: natural parameter (dimension: **order** of the exp. fam.)
- $t(x)$: sufficient statistics ($\leftarrow$ Fisher-Neyman factorization),
- F: log normalizer (cumulant) characterizes the family,
- $C(x)$: carrier measure (usually Lebesgue or counting).

Kullback-Leibler of exp. fam. is a Bregman divergence

$$\mathrm{KL}(p(\theta_p|F)||q(\theta_q|F)) = D_F(\theta_q||\theta_p)$$

Discrete distributions are multinomials

Multinomials extend Bernoulli distributions.

Conversion from source q to natural θ parameters:

$$\theta^{(i)} = \log \frac{q^{(i)}}{q^{(d)}} = \log \frac{q^{(i)}}{1 - \sum_{j=1}^{d-1} q^{(j)}}$$

with $q^{(d)} = 1 - \sum_{j=1}^{d-1} q^{(j)}$
... and to convert back from natural to source parameters:

$$q^{(i)} = \frac{\exp \theta^{(i)}}{1 + \sum_{j=1}^{d-1}(1 + \exp \theta^{(j)})}$$

with $q^{(d)} = \frac{1}{1+\sum_{j=1}^{d-1}(1+\exp \theta^{(j)})}$

Dual log normalizers

Log normalizer of multinomials is

$$F(\theta) = \log\left(1 + \sum_{i=1}^{d-1} \exp \theta^{(i)}\right)$$

Logistic entropy on open space $\Theta = \mathbb{R}^{d-1}$.

Dual Legendre function $F^* = \mathcal{L}F$ is d**-ary entropy**:

$$F^*(\eta) = \left(\sum_{i=1}^{d-1} \eta^{(i)} \log \eta^{(i)}\right) + \left(1 - \sum_{i=1}^{d-1} \eta^{(i)}\right) \log\left(1 - \sum_{i=1}^{d-1} \eta^{(i)}\right)$$

Geodesic of multinomials

Both ∇F and ∇F^* are necessary for:

- computing the left-sided centroid (∇F-means),
- walking on the geodesic ($(\nabla F)^{-1} = \nabla F^*$):

$$q_\lambda = (\nabla F)^{-1}\left((1-\lambda)\nabla F(c_R^F) + \lambda \nabla F(c_L^F)\right)$$

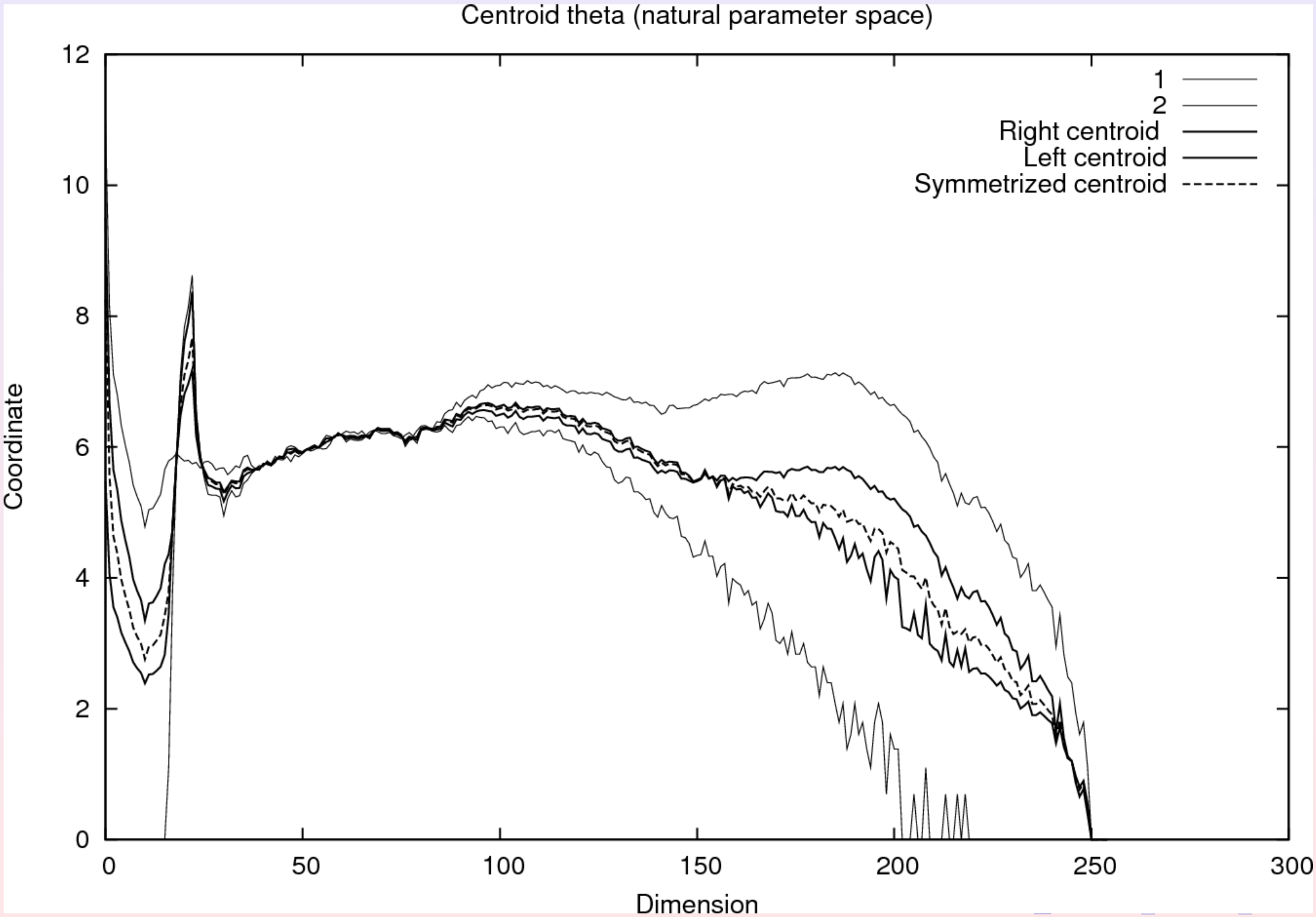
$$\nabla F(\theta) = \left(\frac{\exp \theta^{(i)}}{1 + \sum_{j=1}^{d-1} \exp \theta^{(j)}}\right)_i \stackrel{\text{def}}{=} \eta$$

It follows from Legendre transformation that $\nabla F^* = \nabla F^{-1}$

$$\nabla^{-1} F(\eta) = \left(\log \frac{\eta^{(i)}}{1 - \sum_{j=1}^{d-1} \eta^{(j)}}\right)_i \stackrel{\text{def}}{=} \theta$$

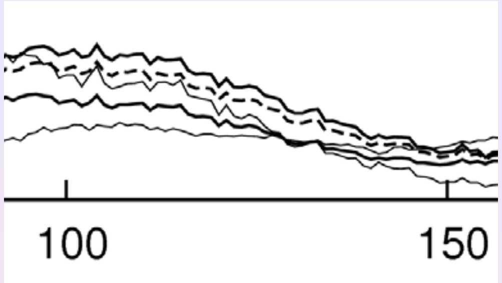
Example: Centroids of histograms

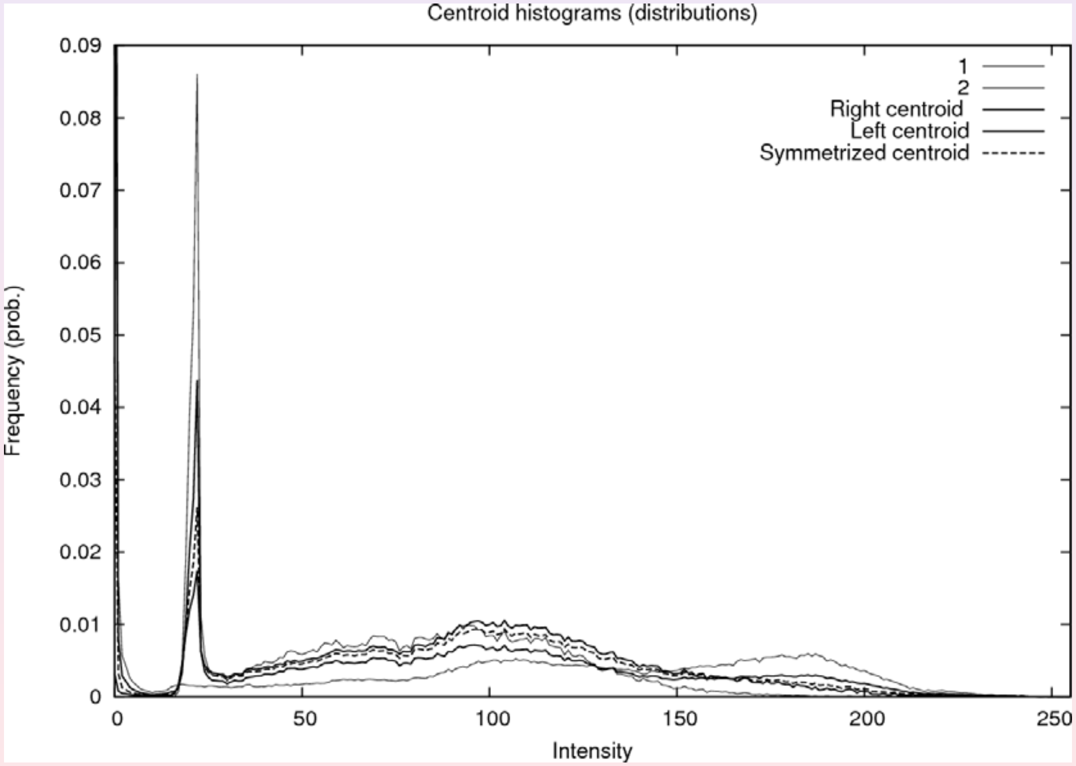
Centroids are generalized means:
$\longrightarrow$ coordinates are always inside the extrema...



Example: Centroids of histograms

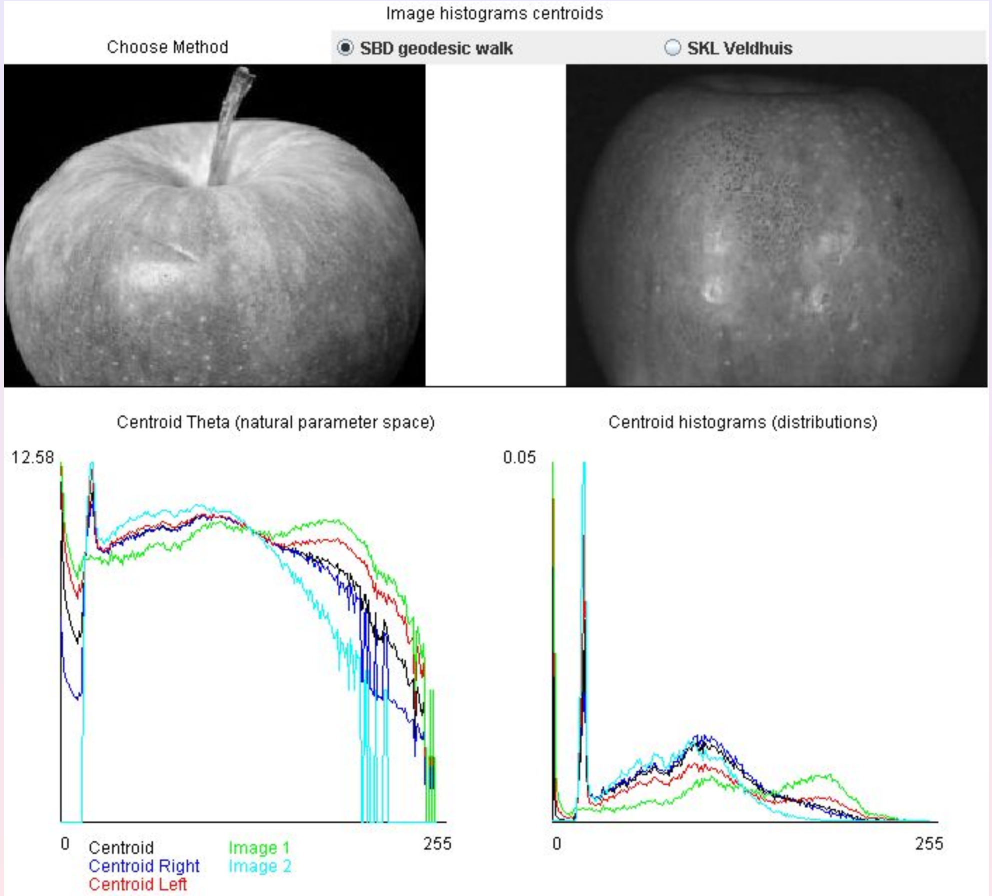
... but not necessarily true for the corresponding histograms!





Centroids of histograms: Java Applet

Generalize ad-hoc convex programming method of (Veldhuis'02).



Try with your own images online using Java applet at:

http://www.sonycsl.co.jp/person/nielsen/SBDj/

Side-by-side comparison

INPUT:
n discrete distributions $q_1, ..., q_n$ of $\mathcal{S}^d$ with
$\forall\ i \in \{1, ..., n\}\ q_i = (q_i^{(1)}, ..., q_i^{(d)})$.

INITIALIZATION

Arithmetic mean:
$\forall k\ \bar{q}^{(k)} = \frac{1}{n} \sum_{i=1}^{n} q_i^{(k)}$
Geometric normalized mean:
$\forall k\ \breve{q}^{(k)} = \frac{\tilde{q}^{(k)}}{\sum_{i=1}^{d} \tilde{q}_i}$ with $\forall k\ \tilde{q}^{(k)} = \left(\prod_{i=1}^{n} q_i^{(k)}\right)^{\frac{1}{n}}$
$\alpha = -1$

MAIN LOOP:
For 1 to 10
$\forall k\ y^{(k)} = \frac{\bar{q}^{(k)}}{\breve{q}^{(k)} \exp \alpha}$
$\forall k\ x^{(k)} = 1$

INNER LOOP 1:
For 1 to 5
$\forall k\ x^{(k)} \leftarrow x^{(k)} - \frac{x^{(k)} \log x^{(k)} - y^{(k)}}{\log x^{(k)} + 1}$

INNER LOOP 2:
For 1 to 5
$\alpha \leftarrow \alpha - \frac{(\sum_{i=1}^{d} x^{(k)} \breve{q}^{(k)} \exp \alpha) - 1}{\sum_{i=1}^{d} x^{(k)} \breve{q}^{(k)} \exp \alpha}$

CENTROID:
$\forall k\ c^{(k)} = x^{(k)} \breve{q}^{(k)} \exp \alpha$

(Vedlhuis'02)

INPUT:
n discrete distributions $q_1, ..., q_n$ of $\mathcal{S}^d$ with
$\forall i \in \{1, ..., n\}\ q_i = (q_i^{(1)}, ..., q_i^{(d)})$

CONVERSION:
Probability mass function $\rightarrow$ multinomial
$\forall i\ \forall k\ \theta_i^{(k)} = \log \frac{q_i^{(k)}}{1 - \sum_{i=1}^{d-1} q_i^{(j)}}$
$F(\theta) = \log(1 + \sum_{j=1}^{d-1} \exp \theta^{(j)})$
$\nabla F(\theta) = \left(\frac{\exp \theta^{(i)}}{1 + \sum_{j=1}^{d-1} \exp \theta^{(j)}}\right)_{i \in \{1,...,d-1\}}$
$(\nabla F)^{-1(\eta)} = \left(\log \frac{\eta^{(i)}}{1 - \sum_{i=1}^{d-1} \eta^{(j)}}\right)_{i \in \{1,...,d-1\}}$

INITIALIZATION:
Arithmetic mean: $\theta_R^F = \frac{1}{n} \sum_{i=1}^{n} \theta_i$
∇F-mean: $\theta_L^F = \nabla F^{-1}(\frac{1}{n} \sum_{i=1}^{n} \nabla F(\theta_i))$
$\lambda_m = 0, \lambda_M = 1$

GEODESIC DICHOTOMIC WALK:
While $\lambda_M - \lambda_m >$ precision do
$\lambda = \frac{\lambda_m + \lambda_M}{2}$
$\theta = (\nabla F)^{-1}((1 - \lambda) \nabla F(c_R^F) + \lambda \nabla F(c_L^F))$
if $D_F(c_R^F || \theta) > D_F(\theta || c_L^F)$ then
$\lambda_M = \lambda$
else
$\lambda_m = \lambda$

CONVERSION:
Multinomial $\rightarrow$ Probability mass function
$\forall i\ q_i^{(d)} = \frac{1}{1 + \sum_{j=1}^{d-1} (1 + \exp \theta_i^{(j)})}$
$\forall i\ \forall k\ q_i^{(k)} = \frac{\exp \theta_i^{(k)}}{1 + \sum_{j=1}^{d-1} (1 + \exp \theta_i^{(j)})}$

Geodesic dichotomic walk

Entropic means of multivariate normal distributions

Multivariate normal distributions of $\mathbb{R}^d$ has following pdf.:

$$\begin{aligned} \Pr(X = x) &= \\ & p(x; \mu, \Sigma) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma}} \exp\left(- \frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right) \end{aligned}$$

Source parameter is a **mixed-type** of vector and matrix:

$$\tilde{\Lambda} = (\mu, \Sigma)$$

Order of the parametric distribution family is $D = \frac{d(d+3)}{2} > d$.

Exponential family decomposition

Multivariate normal distribution belongs to the exponential families:

$$\exp(<\theta, t(x)> - F(\theta) + C(x))$$

- Sufficient statistics: $\tilde{x} = (x, -\frac{1}{2}xx^T)$
- Natural parameters: $\tilde{\Theta} = (\theta, \Theta) = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$
- Log normalizer

$$F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log\pi$$

Mixed-type separable inner product:

$$<\tilde{\Theta}_p, \tilde{\Theta}_q> = <\Theta_p, \Theta_q> + <\theta_p, \theta_q>$$

with matrix inner product defined as the following trace:

$$<\Theta_p, \Theta_q> = \mathrm{Tr}(\Theta_p \Theta_q^T)$$

Dual Legendre functions

$$
\begin{aligned}
F(\tilde{\Theta}) &= \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log \det\Theta + \frac{d}{2}\log\pi \\
F^*(\tilde{H}) &= -\frac{1}{2}\log(1+\eta^T H^{-1}\eta) - \frac{1}{2}\log\det(-H) - \frac{d}{2}\log(2\pi e)
\end{aligned}
$$

Parameters transformations $\tilde{H} \leftrightarrow \tilde{\Theta} \leftrightarrow \Lambda$

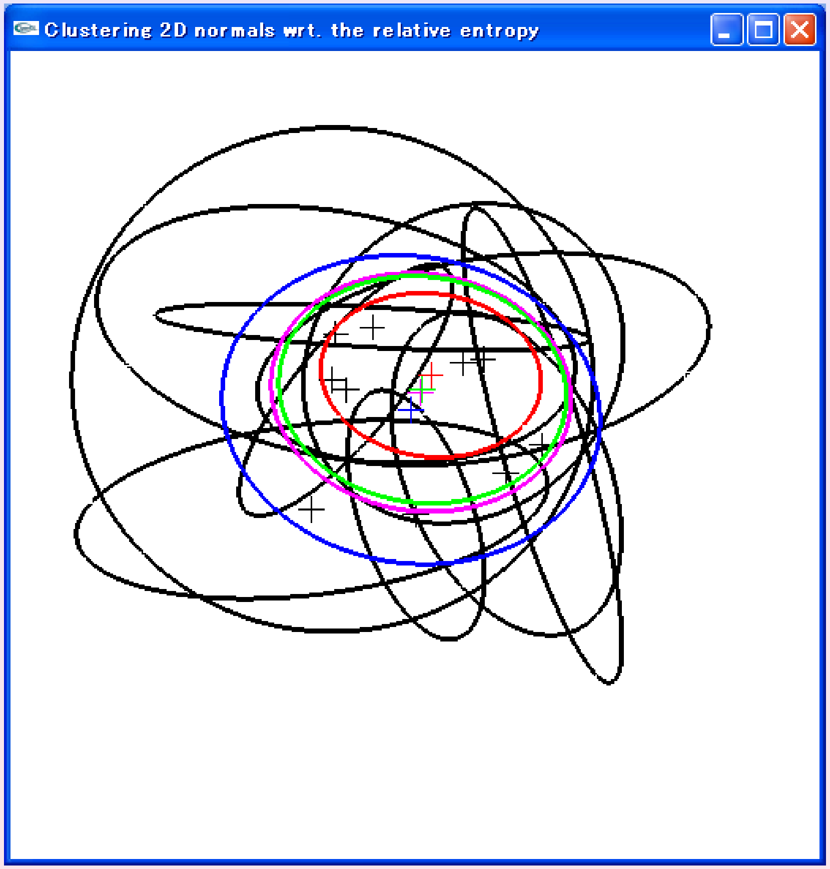
$$
\tilde{H} = \left(\begin{array}{c} \eta = \mu \\ H = -(\Sigma + \mu\mu^T) \end{array}\right) \Longleftrightarrow \tilde{\Lambda} = \left(\begin{array}{c} \lambda = \mu \\ \Lambda = \Sigma \end{array}\right) \Longleftrightarrow \tilde{\Theta} = \left(\begin{array}{c} \theta = \Sigma^{-1}\mu \\ \Theta = \frac{1}{2}\Sigma^{-1} \end{array}\right)
$$

$$
\tilde{H} = \nabla_{\tilde{\Theta}} F(\tilde{\Theta}) = \left(\begin{array}{c} \nabla_{\tilde{\Theta}} F(\theta) \\ \nabla_{\tilde{\Theta}} F(\Theta) \end{array}\right) = \left(\begin{array}{c} \frac{1}{2}\Theta^{-1}\theta \\ -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T \end{array}\right) = \left(\begin{array}{c} \mu \\ -(\Sigma + \mu\mu^T) \end{array}\right)
$$

$$
\tilde{\Theta} = \nabla_{\tilde{H}} F^*(\tilde{H}) = \left(\begin{array}{c} \nabla_{\tilde{H}} F^*(\eta) \\ \nabla_{\tilde{H}} F^*(H) \end{array}\right) = \left(\begin{array}{c} -(H + \eta\eta^T)^{-1}\eta \\ -\frac{1}{2}(H + \eta\eta^T)^{-1} \end{array}\right) = \left(\begin{array}{c} \Sigma^{-1}\mu \\ \frac{1}{2}\Sigma^{-1} \end{array}\right)
$$

Example of entropic multivariate normal means

Simplify and extend (Myrvoll & Soong'03)

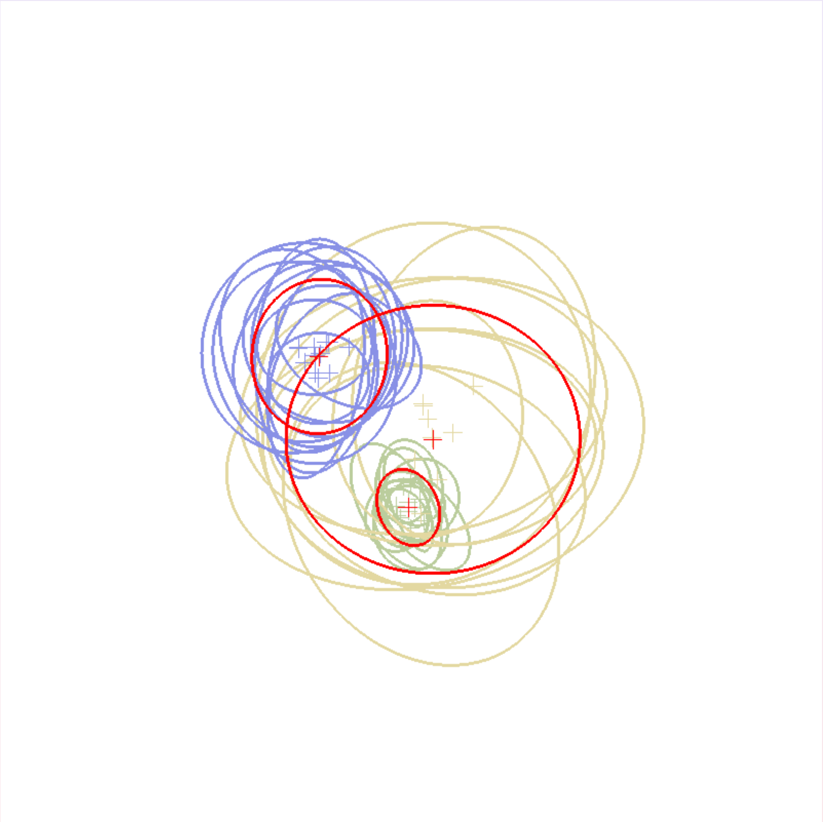


Right in red
Left in blue
Symmetrized in green
Geodesic half point $\lambda = \frac{1}{2}$
in purple

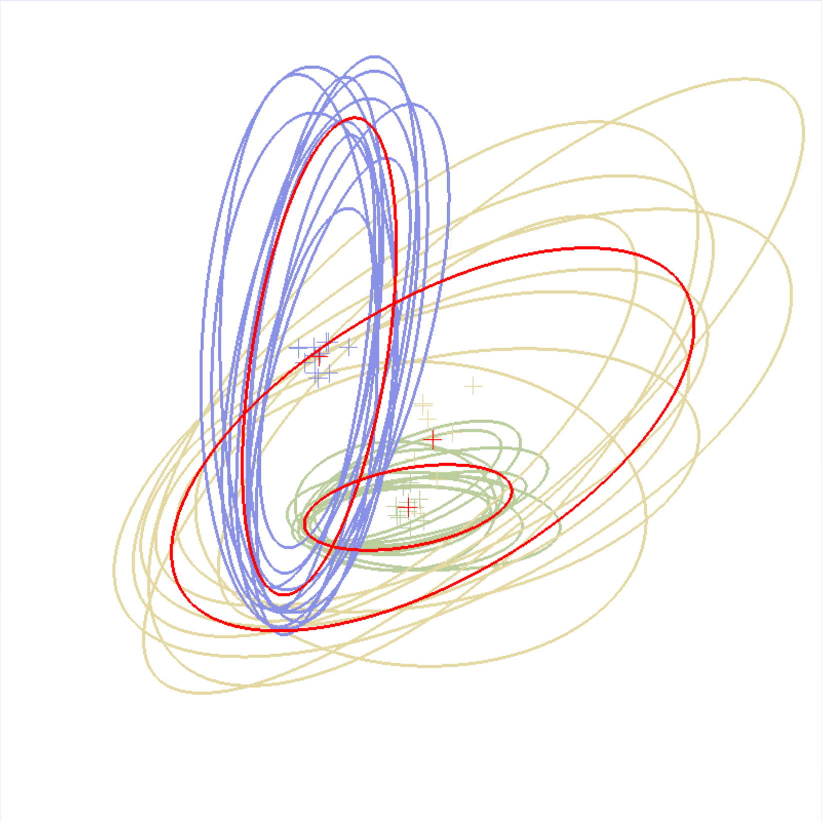
Left D_F centroid is a **right** Kullback-Leibler centroid.
$\rightarrow$ Generalize the approach of (Davis & Dhillon, NIPS'06).

Clustering multivariate normal means

Center-based clustering (Banerjee et al., JMLR'05), (Teboulle'07)



Primal (natural $\tilde{\Theta}$)



Dual (expectation $\tilde{H}$)

COSH distance: Symmetrized Itakura-Saito

Burg entropy $F(x) = -\sum_{i=1}^{d} \log x^{(i)}$ yields Itakura-Saito div.

$$\mathrm{IS}(p||q) = \sum_{i=1}^{d} \left(\frac{p_i}{q_i} + \log \frac{p_i}{q_i} - 1 \right) = D_F(p||q).$$

Symmetrized Itakura-Saito divergence is called COSH distance

$$\mathrm{COSH}(p;q) = \frac{\mathrm{IS}(p||q) + \mathrm{IS}(q||p)}{2}$$

(Wei & Gibson'00) COSH performs "best" for sound processing.

Bregman-Csiszár centroids

(Csiszár Ann. Stat.'91) and (Lafferty COLT'99) used the following *continuous* family of generators:

$$F_\alpha = \begin{cases} x - \log x - 1 & \alpha = 0 \\ \frac{1}{\alpha(1-\alpha)}(-x^\alpha + \alpha x - \alpha + 1) & \alpha \in (0,1) \\ x \log x - x + 1 & \alpha = 1 \end{cases}$$

Continuum of Bregman divergences:

$$\begin{aligned} \mathrm{BC}_0(p||q) &= \log \frac{q}{p} + \frac{p}{q} - 1 \quad \text{Itakura-Saito} \\ \mathrm{BC}_\alpha(p||q) &= \frac{1}{\alpha(1-\alpha)} \left(q^\alpha(x) - p^\alpha(x) + \alpha q(x)^{\alpha-1}(p(x) - q(x)) \right) \\ \mathrm{BC}_1(p||q) &= p \log \frac{p}{q} - p + q \quad \text{Ext. Kullback-Leibler.} \end{aligned}$$

$\rightarrow$ Smooth spectrum of symmetric entropic centroids.

Summary of presentation and contributions

Bregman centroids (Kullback-Leibler exp. fam. centroids)

- Left-sided and right-sided Bregman barycenters are generalized means,
- Symmetrized Bregman centroid c^F is exactly geometrically characterized,
- Simple dichotomic geodesic walk to approximate c^F (2-mean on sided centroids)

Generalize *ad-hoc* solutions with geometric interpretation:

- SKL centroid for non-parameter histograms (Veldhuis'02)
- SKL centroid for parametric multivariate normals (Myrvoll & Soong'03)
- Right KL for multivariate normal (Davis & Dhillon'06)

Thank you!

References:
arXiv.org – cs – cs.CG (Computational Geometry)
arXiv:0711.3242

http://www.sonycsl.co.jp/person/nielsen/BregmanCentroids/

Acknowledgments:
Sony CSL, École Polytechnique (LIX), CEREGMIA, ANR GAIA.