Computational Information Geometry

From Euclidean to flat Pythagorean geometries

Frank Nielsen

École Polytechnique, LIX, France

Sony Computer Science Laboratories, FRL, Japan

Mathematics and Image Analysis 2009 (MIA)

14 December 2009

High-dimensional datasets abound in applications

Heterogeneous feature vector spaces

($\mathcal{F} = \prod_{i=1}^{m} \mathcal{F}_i$).

Noisy datasets

- Image indexing and searching,
  ($\mathcal{F}$: color, texture, shape, location, etc.)
- Sound/speech processing,
  ($\mathcal{F}$: loudness, pitch, timbre, textures, etc.)
- Hypertext documents,
  ($\mathcal{F}$: words, out-links, in-links, etc.)
- XML data objects,
  ($\mathcal{F}$: textual/referential/graphical/numerical/categorical features.)
- Social networks, bio-informatics, etc.

21^{st} Century data processing challenges

Practitioners.

- Which distance function is most **appropriate** ?
- Does my algorithmic toolbox **handles** that distance?

Theoreticians.

- Recover **intrinsic dimensionality** (eg., MST & entropy),
  $\rightarrow$ Degree of freedom of datasets (eg., dof. of face illumination)
- Recover **topology** (eg., theory of zigzag persistence),
- Recover **intrinsic geometry** (eg., distance learning, invariants)

Computational information geometry: Customize geometries to datasets, generic non-Euclidean algorithmic toolboxes.

George E. P. Box (Statistician)

All models are false but some models are useful.

Lecture plan

Part I. **Extending Euclidean algorithms to Bregman divergences** :

- From Lloyd to modern k-means clustering,
- Bregman k-means,
- Bregman soft clustering
  (*expectation-maximization of mixture models made easy*),

Part II. **Information geometry, flat and curved spaces** :

- Nearest neighbor queries: Bregman ball trees and vantage point trees
- Bregman smallest enclosing balls
- Bregman Voronoi and dual Bregman triangulations
- Geometrization of statistics (differential geometry/invariance)

Clustering with k-means
— Les nuées dynamiques —

Lloyd's iterative k-means refinement

Vector quantization (VQ)
(codeword $\in$ codebook for compression/transmission; rate distortion theory)

Hard clustering : Find a **partition** of $\mathcal{V} = \{v_1, ..., v_n\}$ into k **clusters** $\mathcal{V}_1, ..., \mathcal{V}_k$ such as to minimize the **intra-cluster variance** :

$$\underbrace{L(\mathcal{V}) = \sum_{i=1}^{k} \underbrace{\sum_{v_j \in \mathcal{V}_i} ||v_j - c_i||^2}_{\text{cluster}}}_{\text{partition } \mathcal{V} = \uplus_{i=1}^{k} \mathcal{V}_i} \geq 0$$

- Initialization: Seed $\{c_i\}_{i=1}^{k}$ uniformly chosen at random from $\mathcal{V}$ [Forgy].
- Repeat until convergence:

Assignment. Assign vectors to their **nearest cluster**

$$\forall i,\ \mathcal{V}_i = \{v_j \mid ||v_j - c_i|| \leq ||v_j - c_l||\ \forall l \in \{1, ..., k\}\}$$

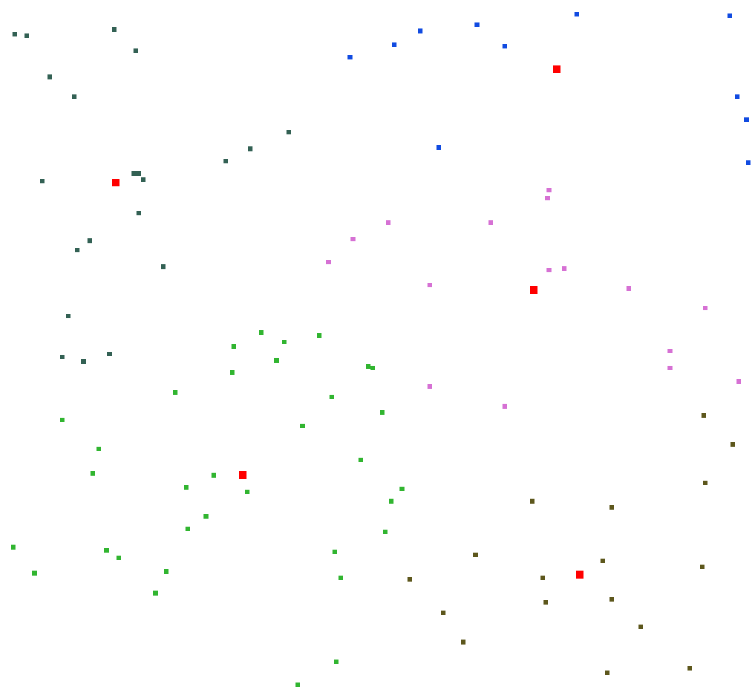
Cluster relocation. Update cluster center c_i as the **centroid** of $\mathcal{V}_i$

$$c_i = \frac{1}{|\mathcal{V}_i|} \sum_{v \in \mathcal{V}_i} v \text{ (=center of mass)}$$

k-means: Potential/loss function

$$L(\mathcal{V}) = \sum_{i=1}^{k} \sum_{v_j \in \mathcal{V}_i} ||v_j - c_i||^2 \geq 0$$

L **monotonically** converges. Lloyd' iterations only minimize **locally** L.
All clusters are always **non-empty**
Never repeat a configuration: upper bound #iter $= O(k^n)$



$\longrightarrow$ repeat until convergence (**local minimum**)

Framework for a generic k-means paradigm

- Initialize cluster centers from randomly choosing k seeds
- Repeat until convergence
  - **Partition assignment** : Allocate points to their nearest cluster center
  - **Center relocation** : Adjust centers of each cluster

Properties of k-means:

- Potential (loss) function **monotonically** decreases (and hence converge): $L_D(\mathcal{V}) = \sum_{j=1}^{k} \sum_{v_i \in C_j} D(v_i, c_j) \geq 0$
- Center relocation of each cluster can be solved as a MINAVG optimization: $c^* = \arg\min_c \sum_{v \in \mathcal{V}} D(v, c)$

Some k-means-like algorithms

For example,

- Euclidean (Lloyd) k-means: $D(p,q) = ||p-q||^2$ ($\rightarrow$ center of mass)
- Spherical k-means: relocate $\frac{\sum_i v_i}{\|\sum_i v_i\|}$ (centers lie on unit sphere)
- Convex k-means
  (assume convexity of $D(\cdot,\cdot)$ in the *second* argument)

Convex k-means [Modha & Spangler, 2003]

A Unified Continuous Optimization Framework for Center-Based Clustering Methods
[Teboulle, JMLR 2007].

k-means with MinAvg optimizer as the centroid

- Initialize k seeds
- Repeat until convergence
  - **Partition assignment** : Allocate points to their nearest center wrt. D
  - **Center relocation** : Move the cluster center to the **cluster centroid**

Problem: Find class of distances D that yield centroid as the MinAvg minimizer

> **Bregman divergences** are the **only** distances such that MinAvg optimizer is the data centroid.

(**Only** the distance change in your k-means code! $\|p-q\|^2 \rightarrow D(p,q)$)

- Bregman k-means [Banerjee et al., JMLR'2005]
- Axiomatization and **exhaustiveness** :
  On the optimality of conditional expectation as a Bregman predictor [IEEE TIT'05]

Bregman divergences B_F

Bregman generator : **Strictly convex** and **differentiable** function F.

For scalars: $B_f(p||q) = f(p) - f(q) - (p-q)f'(q)$ with $f'(x)$ the derivative function.

For vectors: $B_F(p||q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle$ with $\nabla F(x) = [\frac{\partial F(x)}{\partial x_i}]_i^T$ the gradient vector, and $\langle \cdot, \cdot \rangle$ the inner product.

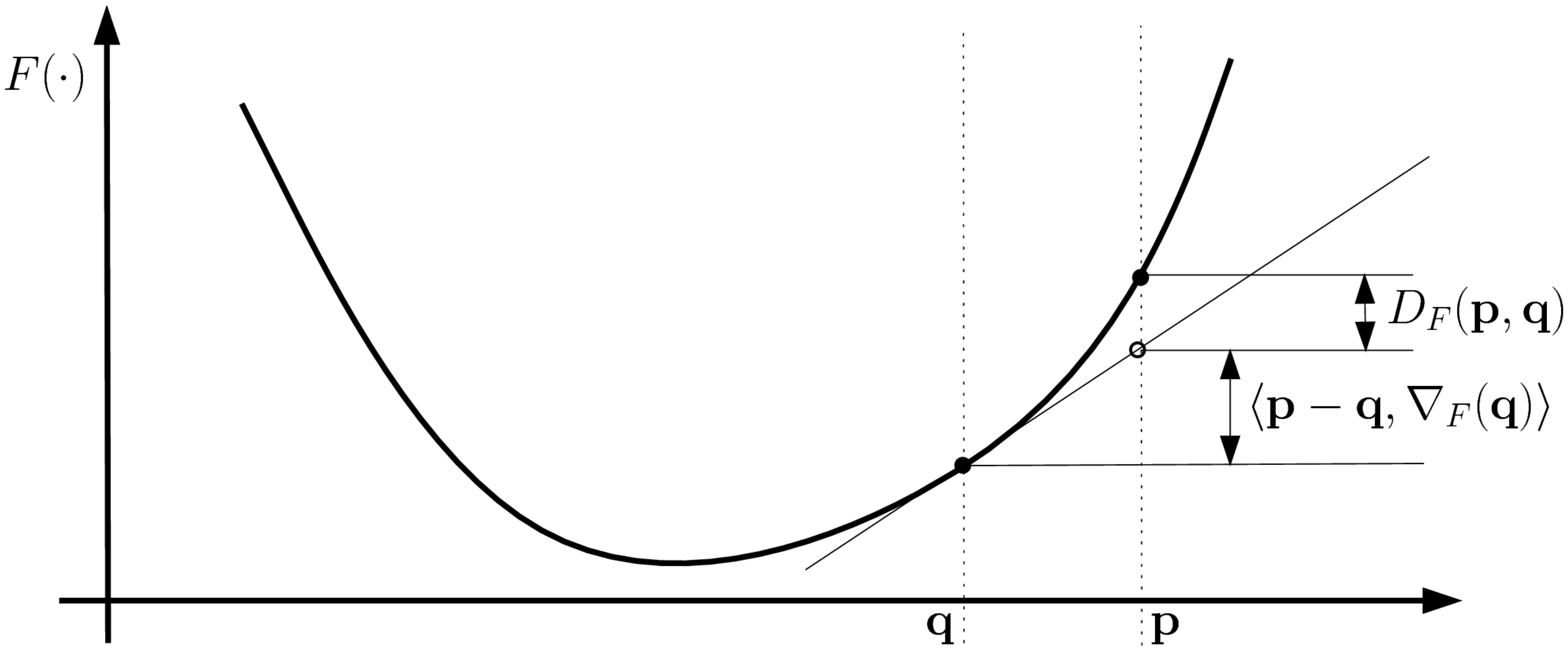
Strictly convex F: Hessian $\nabla^2 F \succ 0$ (psd.) $\longrightarrow$ ∇F is monotonous.

Separable Bregman divergences: $F(x) = \sum_{i=1}^d f_i(x_i)$.

Bregman divergences: A geometric visualization

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

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

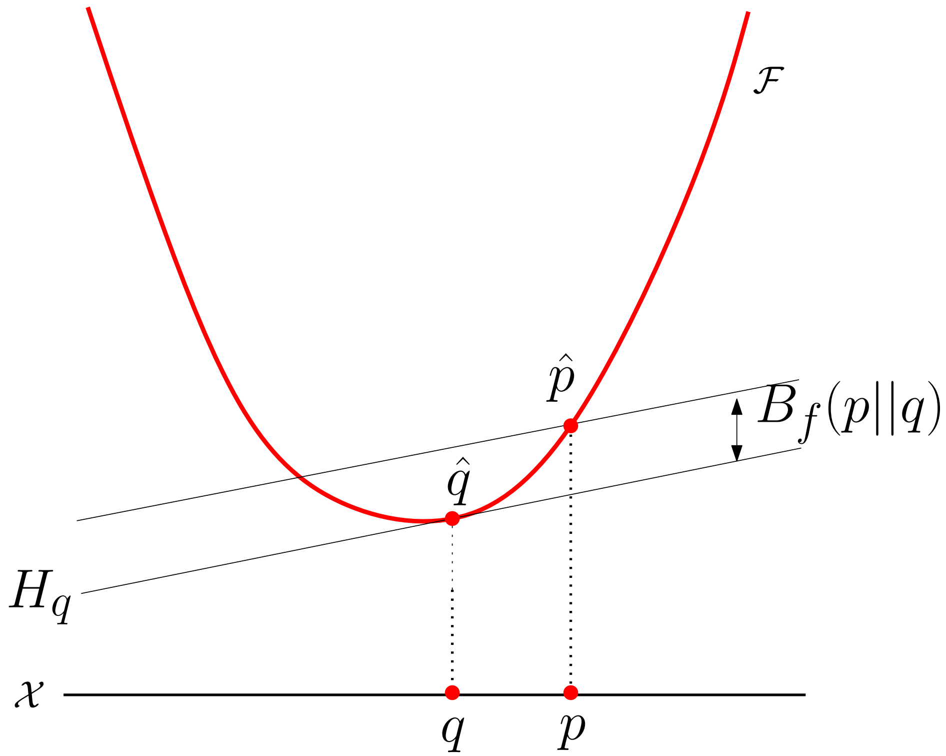


http://www.sonycsl.co.jp/person/nielsen/BregmanDivergence/

Bregman divergences: Another geometric visualization

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

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



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

Squared Euclidean distance (aka. L_2^2)

Take $F(x) = x^T x$. Gradient $\nabla F(x) = 2x$.

$$\begin{aligned} B_F(p,q) &= F(p) - F(q) - (p-q)^T \nabla F(q) \\ &= p^T p + q^T q - 2p^T q \\ &= ||p-q||^2 \end{aligned}$$

Squared Euclidean distance is a Bregman divergence.

Squared Euclidean distance is not a metric: Triangle inequality **fails** .
E.g., $q = 2p$ and $r = \frac{3}{2}p$.

However Euclidean distance is a metric (with triangular inequality).

$\longrightarrow$ Many square root symmetrized Bregman divergences are metrics iff. $(\log f'')'' \geq 0$ [Chen'08].
For example, the square root of Jensen-Shannon divergence.
Metrics defined by Bregman Divergences, Commun. Math. Sci. 6(4) 4 (2008), 915-926.

Entropy H, uncertainty and information

The **entropy** of a random variable X is its amount of **uncertainty** .
Shannon entropy [1948, communication in noisy/gaussian channels]:

Discrete random variable:

$X \sim \{E_1, ..., E_n\}$ with $\Pr(X = E_i) \overset{\text{equal}}{=} p_i$ (**probability mass function**):

$$H(X) = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^n p_i \log_2 p_i$$ (in bits, or nats for base e)

Continuous random variable:

$X \sim \mathcal{X}$ with $\Pr(X = x) \overset{\text{equal}}{=} p(x)$ (**probability density function**):

$$H(X) = \int_{x\in\mathcal{X}} p(x) \log_2 \frac{1}{p(x)} \mathrm{d}x = -\int p(x) \log p(x) = \mathrm{E}_{\mathcal{X}}[-\log p(X)]$$

Two remarkable facts:

- Maximum uncertainty (entropy) is obtained for the **uniform distribution** :
$$0 \leq H(X) \leq \log n$$
- Maximum entropy for unit variance pdf. is the **Gaussian distribution** .

Cross-entropy $H^\times$

Measures the average number of bits needed to identify an event from a set of possibilities **when** a coding scheme is used based on a given probability distribution $\tilde{P}$, rather than the true (unknown) distribution P.

$$H^\times(P||\tilde{P}) = \mathrm{E}_P[-\log p(\tilde{P})] \geq H(P) \geq 0$$

In modeling probability, P is the *true/target* distribution and $\tilde{P}$ is the **model** .

The closer the cross-entropy is to the entropy, the better the model .
(for $\tilde{P} = P$, $H^\times(P||P) = H(P)$).

- Discrete rv.: $H^\times(P||\tilde{P}) = \sum_i p_i \log_2 \frac{1}{\tilde{p}_i} = -\sum_i p_i \log_2 \tilde{p}_i$
- Continuous rv.:
$H^\times(P||\tilde{P}) = \int p(x) \log_2 \frac{1}{\tilde{p}(x)} \mathrm{d}x = -\int_x p(x) \log_2 \tilde{p}(x) \mathrm{d}x$

Statistical distance: Relative entropy (KL)

The **Kullback-Leibler** measures the **divergence** between two distributions.

$$\boxed{D(P||\tilde{P}) = H^{\times}(P||\tilde{P}) - H(P) \geq 0}$$

KL = cross-entropy of true/model distributions minus the true entropy.
Expected extra message-length per symbol that must be communicated if a code for a given (approximated) distribution $\tilde{P}$ is used instead of optimal P [Covers & Thomas'06].

For probability mass functions:

$$D(P||\tilde{P}) = \sum_i p_i \log_2 \frac{p_i}{\tilde{p}_i} \qquad D(P||\tilde{P}) = \int_{\mathcal{X}} p(x) \log_2 \frac{p(x)}{\tilde{p}(x)} \mathrm{d}x$$

Many synonyms: Information discrimination, relative entropy, etc.

Relative entropy is also a Bregman divergence

Bregman divergence on probability measures p, q:

$$B_f(p||q) = \int (f(p) - f(q) - (p-q)f'(q))\,\mathrm{d}\mu$$

Take $f(x) = x\log x = -x\log\frac{1}{x}$, the **negative (convex) Shannon entropy** (with $f'(x) = 1 + \log x$ and $f''(x) = \frac{1}{x} > 0$ for all $x \in \mathbb{R}_*^+$):

$$\begin{aligned}
B_f(p(x)||q(x)) &= \int (p(x)\log p(x) - q(x)\log q(x) - (p(x) - q(x))(1 + \log q(x)))\,\mathrm{d}\mu \\
&= \int \left(p(x)\log\frac{p(x)}{q(x)}\right)\mathrm{d}\mu - \underbrace{\int p(x)\,\mathrm{d}\mu}_{=1} + \underbrace{\int q(x)\,\mathrm{d}\mu}_{=1} \\
&= \int \left(p(x)\log\frac{p(x)}{q(x)}\right)\mathrm{d}\mu = D(p(x)||q(x))
\end{aligned}$$

(I-divergence is KL divergence for **unnormalized** measures)

Bregman k-means

Bregman divergences **unify** geometric **squared Euclidean distance** with entropic asymmetric **Kullback-Leibler divergence** .

Bregman divergences are always convex in the first argument but may *not* be convex in the second argument (eg., $F(x) = -\log x$, the Burg entropy).

Thus Bregman k-means is not necessarily a convex k-means [Modha & Spangler'03] (actually, it is! -:) using Legendre transformation).

However, the right-side MinAvg optimization problem **surprisingly** always yield the **centroid** (center of mass) as the minimizer.

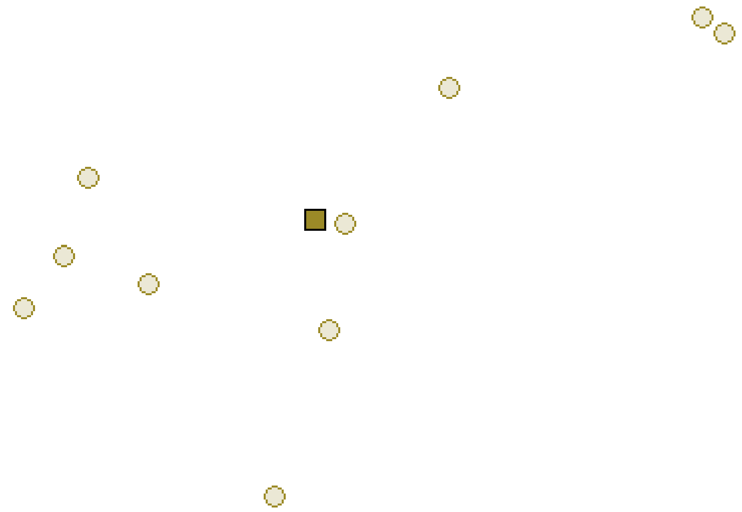
$\longrightarrow$ Bregman divergences allows us to generalize Lloyd k-means.

Bregman representative and Bregman information

Bregman representative : center cluster, (Bregman) centroid

Bregman information : minimum loss function $I_F(\mathcal{P}) = \frac{1}{n}\sum_i B_F(p_i||\bar{p})$, center radius, Bregman/information radius

For squared Euclidean distance, Bregman information = **cluster variance** .



Sample variance $\frac{1}{n}\sum_i (x_i - \bar{x})^2$.

(For Kullback-Leibler divergence, it is *related* to the **mutual information** .)

Quantization: Potential/loss function of k-means

A careful **rewritting** of the loss function yields [Duda et al., 2001]:

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

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

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

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

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

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

...preserve as much as possible Bregman information.

Bregman k-means: Unifying former algorithms

Bregman generator	Bregman divergence	Clustering algorithm
Squared norm	Squared loss	k-means (1956, 1957)
Negative Shannon entropy	Kullback-Leibler divergence	Information-theoretic clustering (2003)
Burg entropy	Itakura-Saito divergence	Linde-Buo-Gray (1980)
...F...	...B_F...	...Bregman k-means...

Bregman k-means yields a **parametric** family of clustering algorithms.

$\longrightarrow$ **Meta-algorithm** .

Key question: How to choose F?

$\longrightarrow$ Many works involve generalized quadratic/Mahalanobis distances.

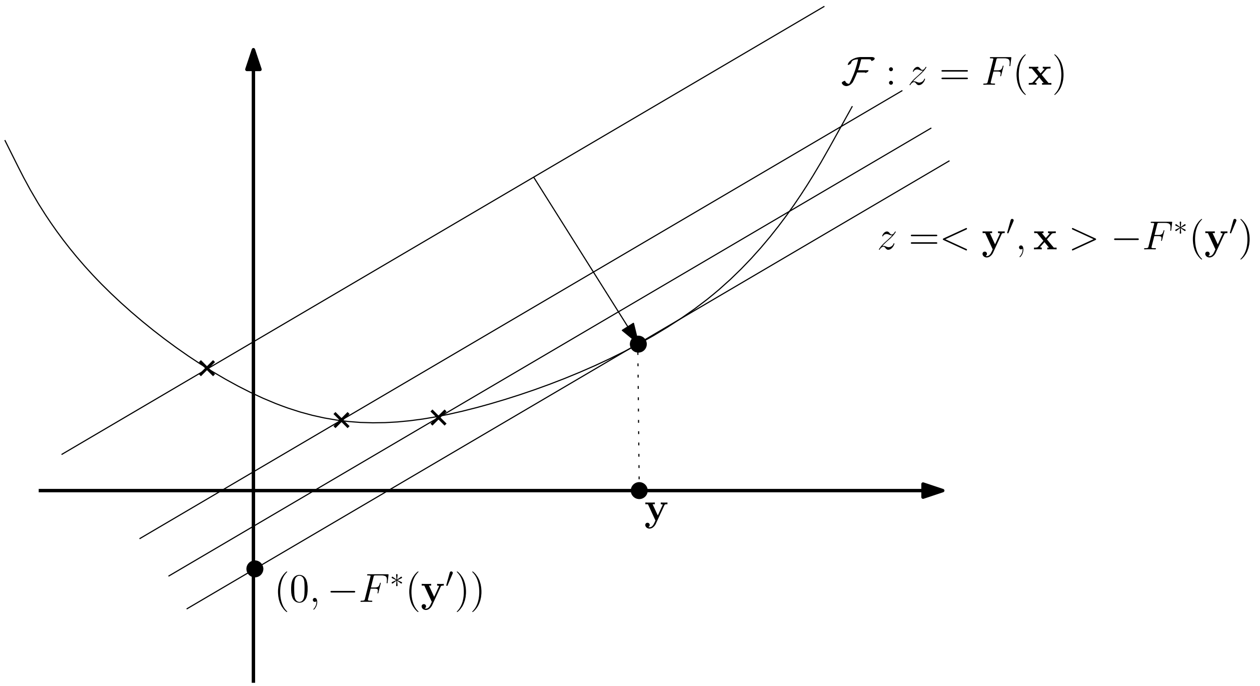
Legendre transformation: Convex conjugates

Let F^* be the **Legendre convex conjugate** of F:

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

The supremum is reached at the *unique* point y where the gradient of $F^*(x) = \langle y, x\rangle - F(x)$ vanishes: $\frac{\partial F^*(x)}{\partial x} = 0 \Longrightarrow y = \nabla F(x)$.

Convex functions come **pairwise** with their domains: $(F, \mathcal{X}) \Leftrightarrow (F^*, \mathcal{X}^*)$



Computing Legendre transformation

Legendre transformation = slope transformation
(dual parameterizations of convex functions: x, $\nabla F(x)$)

In practice:

- Get ∇F from F (easy, fully automatic)
- Compute **reciprocal gradient** : $(\nabla F)^{-1} = \nabla F^*$
  (For non-closed form solutions, perform Householder's root-finding algorithm)
- Compute integral $F^* = \int \nabla F^* = \int (\nabla F)^{-1}$
  (can be tricky too)

$$\boxed{(F^*)^* = F}$$

For example, consider $f(x) = \exp x$, $f'(x) = \exp x$, $f' * (y) = \log y$,
$\Rightarrow f^*(y) = y \log y - y$.

Dual Bregman divergences

Follows from the Legendre transformation:

$$B_F(p||q) = F(p) + F^*(\nabla F(q)) - \langle p, \nabla F(q) \rangle = B_{F^*}(\nabla F(q)||\nabla F(p))$$

Dual divergences :

$$B_F(p||q) = B_{F^*}(\nabla F(q)||\nabla F(p)) \; \forall (p,q) \in \mathcal{X} \times \mathcal{X}$$

$$B_{F*}(r||s) = B_F(\nabla F^*(s)||\nabla F^*(r)) \; \forall (r,s) \in \mathcal{X}^* \times \mathcal{X}^*$$

(See information geometric interpretation and canonical divergences)

Generalized means: f-means

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

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

Pythagoras' means :

- Arithmetic: $f(x) = x$
- Geometric: $f(x) = \log x$
- Harmonic: $f(x) = \frac{1}{x}$

Property:

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

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

Left-sided and right-sided Bregman barycenters

The *right-sided barycenter* $b_F(w)$ is **independent** of F and computed as the **weighted arithmetic mean** on the point set, a generalized mean for the identity function: $b_F(\mathcal{P}; w) = b(\mathcal{P}; w) = M(\mathcal{P}; x; w)$ with $M(\mathcal{P}; f; w) = f^{-1}(\sum_{i=1}^n w_i f(v_i))$.

The *left-sided Bregman barycenter* b_F^* is computed as a **generalized mean** on the point set for the gradient function ∇F: $b_F^*(\mathcal{P}) = M(\mathcal{P}; \nabla F; w)$.

The **Bregman information** (information radius) of sided barycenters is a **F-Jensen remainder** (also known as Burbea-Rao divergences):

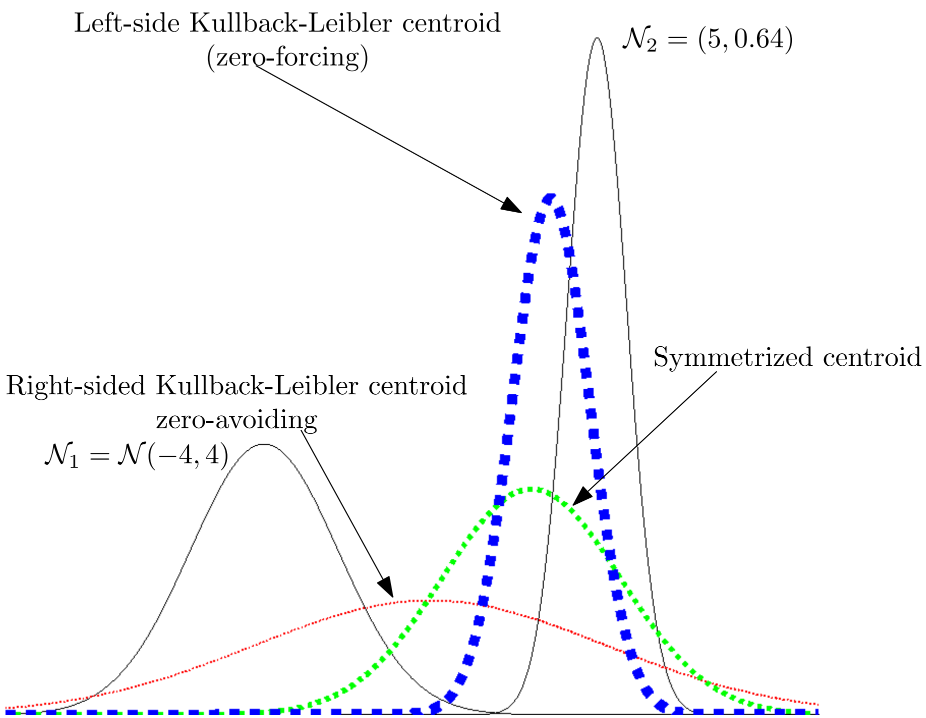
$$\mathrm{JS}_F(\mathcal{P}; w) = \sum_{i=1}^{d} w_i F(p_i) - F\left(\sum_{i=1}^{d} w_i p_i\right) \geq 0$$

(Jensen's inequality)

Left-sided or right-sided centroids (k-means) ?

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).

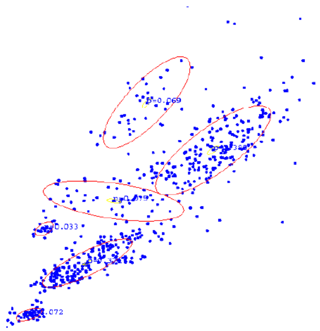


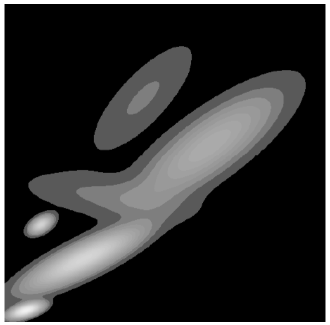
Soft clustering & EM algorithm

Soft clustering: each point belongs to all clusters according to a weight distribution (=density). $\rightarrow$ statistical modeling

Gaussian mixture models (GMMs, MoGs: mixture of Gaussians):
Probabilistic modeling of data:
$\Pr(X=x)=\sum_{i=1}^{k} w_i \Pr(X=x|\mu_i,\Sigma_i)$ (with $\sum_i w_i = 1$ and all $w_i \geq 0$).

$$\Pr(X=x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Sigma}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}.$$





Similar to k-means, soft clustering wrt. to **log-likelihood** is minimized by the expectation-maximization (EM) algorithm [Dempster'77]

Exponential families in statistics

$\longrightarrow$ Workhorses of probabilistic modeling.

Canonical decomposition of the probability measures:

$$\boxed{p_F(x|\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right)}$$

- F: **log-normalizer**, **strictly convex function** characterizing the family: Gaussian, Multinomial, Poisson, Beta, Gamma, Rayleigh, Weibull, Wishart, von Mises, etc. (∞ many)
- θ: **natural parameters** (fix a family member)
- $t(x)$: **sufficient statistics** (for recovering parameters from observations)
- $k(x)$: **carrier measure** (usually Lebesgue or counting)

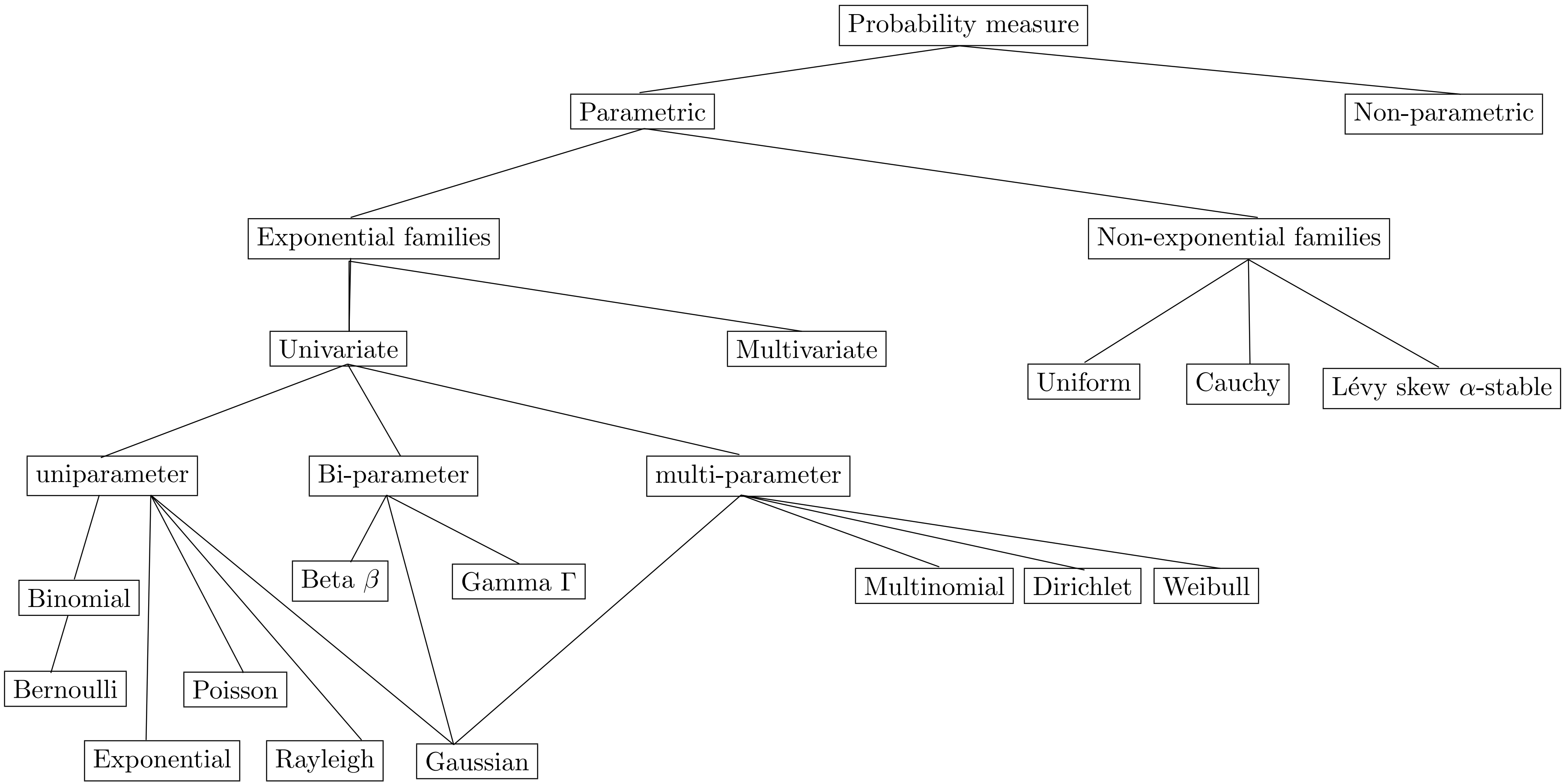
Log normalizer : From $\int_{x\in\mathcal{X}} p_F(x|\theta)\mathrm{d}x = 1$

$$\Rightarrow F(\theta) = \log \int e^{\langle t(x),\theta\rangle + k(x)}\mathrm{d}x$$

Exponential family=**log-linear model** .

(Convex conjugate F^*: negative entropy)

A taxonomy of probability measures



Expectation and variance of exponential families

$$X \sim p_F(\theta)$$

- Expectation:

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

- Variance:

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

(for natural sufficient statistics $t(X)$s)

Exponential families have always **finite moments** . F is C^∞
(incl. expectations & variances.)
($\rightarrow$ Cauchy distributions have not finite moments, $\rightarrow$ do not belong to the exponential families).

Example of exponential families: Gaussian distributions

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

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

Σ: Variance/covariance matrix (=dispersion matrix)
Source parameter is a **mixed-type** of *vector* $\mu \in \mathbb{R}^d$ and *matrix* $\Sigma \succ 0$:

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

Order of the parametric distribution:

$$D = d + \underbrace{\frac{d+1}{2}}_{\Sigma \succ 0 \text{ is symmetric psd.}} = \frac{d(d+3)}{2} > d.$$

(cone of positive semidefinite matrices)

Multivariate normals: Canonical decomposition

Multivariate normal distribution belongs to the exponential families:

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

- Sufficient statistics: $\tilde{x} = (x, -\frac{1}{2}xx^T)$ ($\rightarrow$ mean & sample covariance)
- 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:

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

Multivariate normals: Dual Legendre log normalizers

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

Converting parameters: $\tilde{H} \leftrightarrow \tilde{\Theta} \leftrightarrow \Lambda$

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

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

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

Maximum likelihood estimator of exponential families

d: dimensionality of the observations

D: order (=#parameters) of the exponential family (ie., $\frac{d(d+3)}{2}$ for normals)

The **maximum likelihood estimator** (MLE) is:

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

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

(called **observed point** in information geometry)

(For non-closed form solutions, use Newton or Householder **root findings** to get *arbitrary* fine approximations of ∇F^{-1}.)

Kullback-Leibler & Bregman divergences

The **KL divergence** of distributions of the **same** exponential family is the Bregman divergence induced by the log-normalizer on the corresponding natural parameters (with parameter order swapping):
Furthermore, generic formula for entropy and relative entropy:

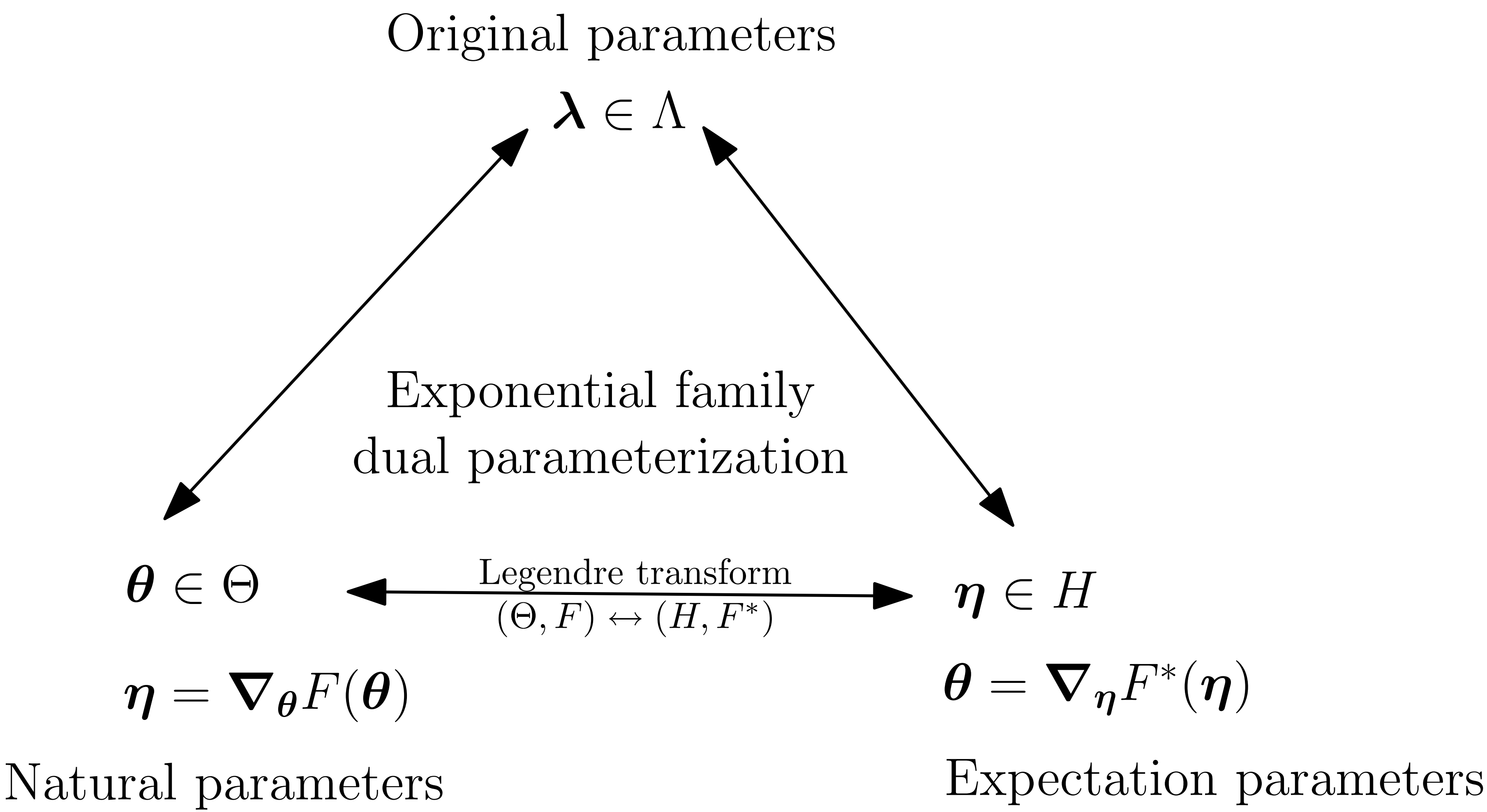
$$
\begin{aligned}
\mathrm{KL}(p_F(x;\theta_1)||p_F(x;\theta_2)) &= B_F(\theta_2||\theta_1) \\
&= H^\times(p_F(x;\theta_1)||p_F(x;\theta_2)) - H(p_F(x;\theta_1))
\end{aligned}
$$

$$
\begin{aligned}
H(p) &= H_F(\theta_p) = F(\theta_p) - \langle \theta_p, \nabla F(\theta_p) \rangle + b \\
H_F^\times(\theta_p||\theta_q) &= F(\theta_q) - \langle \theta_q, \nabla F(\theta_p) \rangle + b
\end{aligned}
$$

$b = -\int k(x) p_F(x;\theta) \mathrm{d}x$ (0 for some exponential families like Gaussians).

Statistical exponential families: A digest with flash cards, arXiV, http://arxiv.org/abs/0911.4863, 2009.

Parameterization of exponential families



(θ/η are the two dual affine biorthogonal coordinate systems.)

Bijection: Exponential families $\Leftrightarrow$ Bregman divergences

Regular exponential family $\Longleftrightarrow$ regular Bregman divergence:

$$\log p_F(x;\theta) = -B_{F^*}(x||\nabla F(\theta)) + \log c_{F^*}(x)$$

$\mu \stackrel{\text{equal}}{=} \nabla F(\theta)$ is the expectation of the distribution.
F^*: generalized entropy functional.

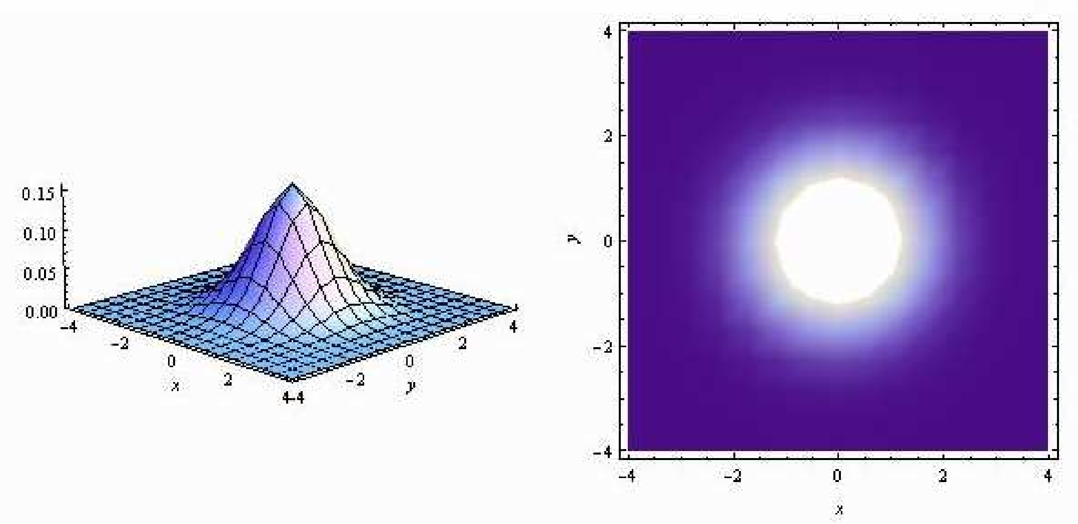
Note: For some generators F, the Legendre dual F^* may not have closed-form solutions. Use Householder formula to approximate the $\nabla F(\theta)$ for a given θ (root finding).

[JMLR'05] Clustering with Bregman divergences.

Visualizing the density/distance bijection

Consider a probability measure of an exponential family $p_F(x;\theta)$.

The **expectation** is $\int p_F(x;\theta)\mathrm{d}x = \nabla F(\theta) \stackrel{\text{equal}}{=} \mu$.
The **iso-density** is $p_F(x;\theta) = \lambda \iff c_{F^*}(x)\exp - B_{F^*}(x||\nabla F(\theta)) = \lambda$.
Bregman divergences are **always convex** in first argument.
$\mu = \nabla F(\theta)$ is located as the **second argument**



Convex contours centered at the mean
(=convex distance centered at a point = iso-density).

Exponential families have always finite means
(Therefore Cauchy distributions does not belong to exp. fam.)

Bregman divergences $\Leftrightarrow$ Exponential families

Bregman divergence: $B_{F^*}(x||\mu)$ $\longleftrightarrow$ Bregman generator: $F^*(\mu)$ $\xleftrightarrow{\text{Legendre duality}}$ Cumulant function: $F(\theta)$ $\longleftrightarrow$ Exponential family: $p_F(x|\theta)$

$\mu = \nabla F(\theta)$

$$p_F(x;\theta) = \lambda \Longleftrightarrow c_{F^*}(x) \exp -B_{F^*}(x||\nabla F(\theta)) = \lambda$$

Dual coordinate systems $(\theta, \eta = \nabla F(\theta) = \mu)$

Examples: Exponential families $\Leftrightarrow$ Bregman divergences

Generator	Distribution	Loss/energy	
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
...$F(x)$...	...$p_F(x\|\theta)$...	$\Leftrightarrow$	...B_{F^*}... (∞ many)

Soft Bregman clustering

Model data with mixture of the **same** exponential family.
Expectation-maximization (EM) for exponential families:

$$X \sim \sum_{i=1}^{k} w_i p_F(x|\theta_i)$$

with $w_i > 0$ and $\sum_i w_i = 1$.

From $\log p_F(x|\theta) \propto -B_{F*}(x||\mu)$, with $\mu = \nabla F(\theta)$:

Maximum log-likelihood (max. $\log p_F$) $\Leftrightarrow$ Minimum Bregman divergence(B_{F^*})

$\longrightarrow$ yields **very efficient** soft clustering.
Soft clustering (EM) extends hard (k-means) clustering

Bregman soft clustering made easy

Bregman EM clustering algorithm on $\{x_1, ..., x_n\}$:

Initialization. Set $\{w_i, c_i\}_{i=1}^k$ with $\sum_i w_i = 1$
(eg., Bregman k-means++ using:
the centroids of sufficient statistics per cluster)

Loop until convergence.

Expectation. (compute the **posterior** probability)
For all observations x
For all model component i:

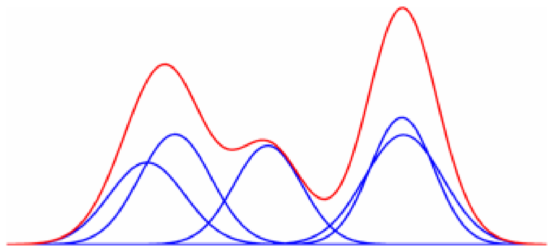
$$\Pr(i|x) = \frac{w_i \exp -B_{F^*}(x||c_i)}{\sum_{j=1}^k w_j \exp -B_{F^*}(x||c_j)}$$

Maximization. For all model components i

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

$$c_i = \frac{\sum_{j=1}^n \Pr(i|x_j)x_j}{\sum_{j=1}^n \Pr(i|x_j)}$$

jMEF: Mixture of exponential families



- A Java library to create, process and manage mixtures of exponential families (MEF):
  - Estimation of a MEF using the Bregman soft clustering.
  - Simplification of a MEF using the Bregman hard clustering.
  - Hierarchical representation of a MEF using the Bregman hierarchical clustering.
  - Learn the *optimal* MEF using the Bregman hierarchical clustering.
- Open-source:

  `http://www.lix.polytechnique.fr/~nielsen/MEF/`
- Cross platform

Bregman dual bisectors: Hyperplane/hypersurface

Right-sided bisector: $\rightarrow$ Hyperplane

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

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

Left-sided bisector: $\rightarrow$ Hypersurface

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

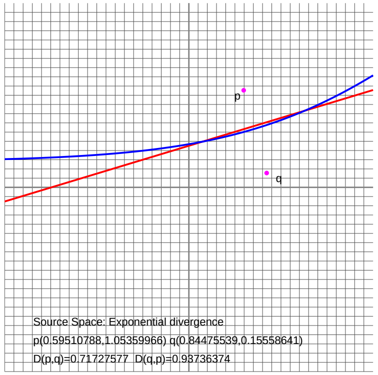
$$H'_F : \langle \nabla F(x), q - p\rangle + F(p) - F(q) = 0$$

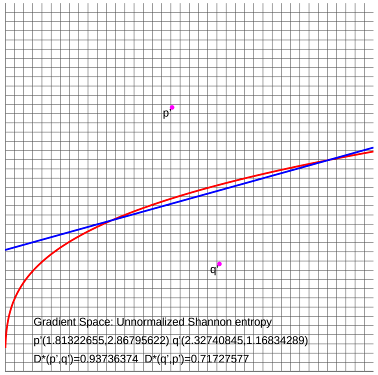
(hyperplane in the “gradient space” $\nabla\mathcal{X}$ = dual coordinate system)

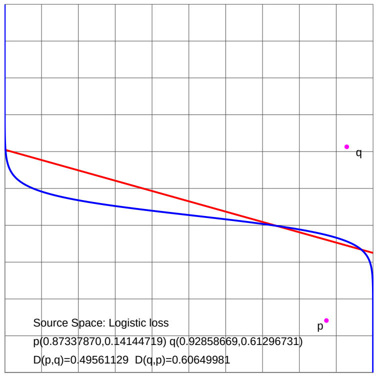
Visualizing Bregman bisectors

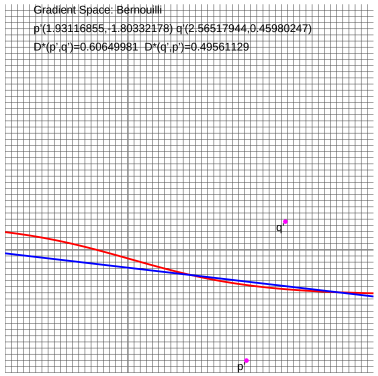
Primal coordinates θ — natural parameters

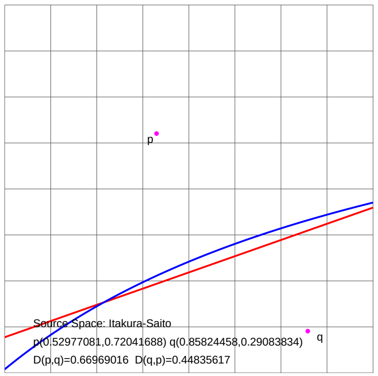
Dual coordinates η — expectation parameters

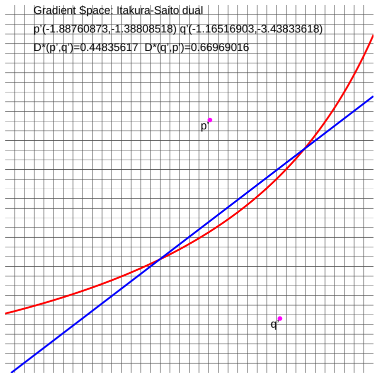












Bregman MINIBALL (infsup/minimax) algorithm

Problem: Given a point set $\mathcal{P} = \{p_1, ..., p_n\}$, finds the smallest enclosing ball with respect to a Bregman divergence B_F:

$$c^* = \arg\min_{c\in\mathcal{X}} \max_{i=1}^{n} B_F(c||p_i)$$

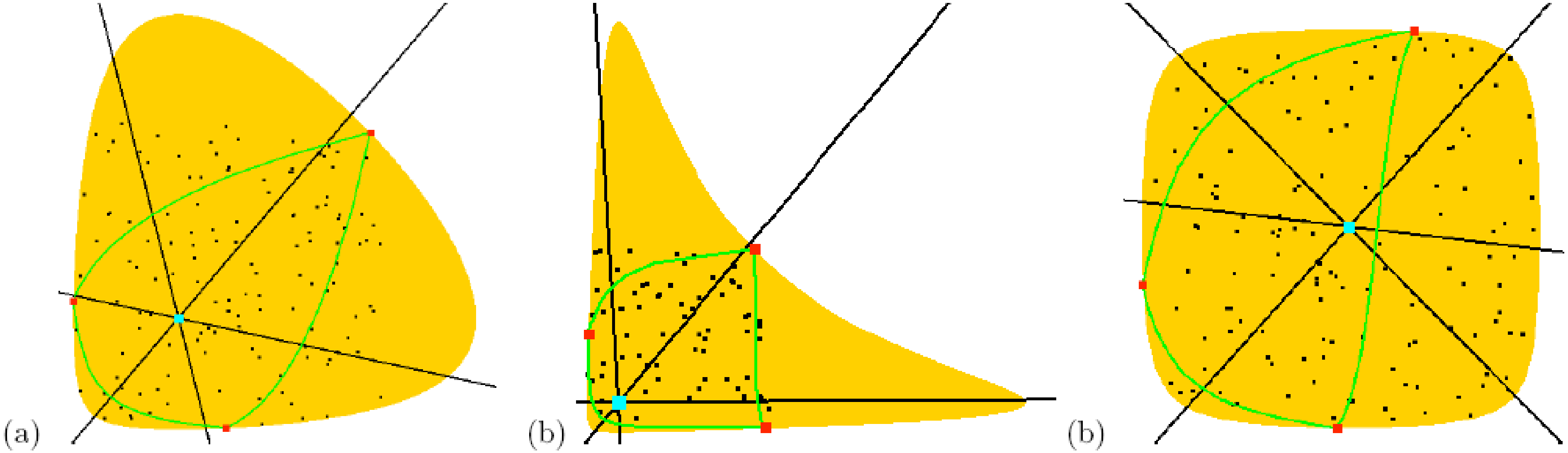
$\rightarrow$ unique ball/circumcenter, and
$\rightarrow$ unique radius $r^* = \min_{c\in\mathcal{X}} \max_{i=1}^{n} B_F(c||p_i)$.

Fit the **LP-type** framework

[IPL'08] On the smallest enclosing information disk. (2008)

Bregman MiniBall: Demo

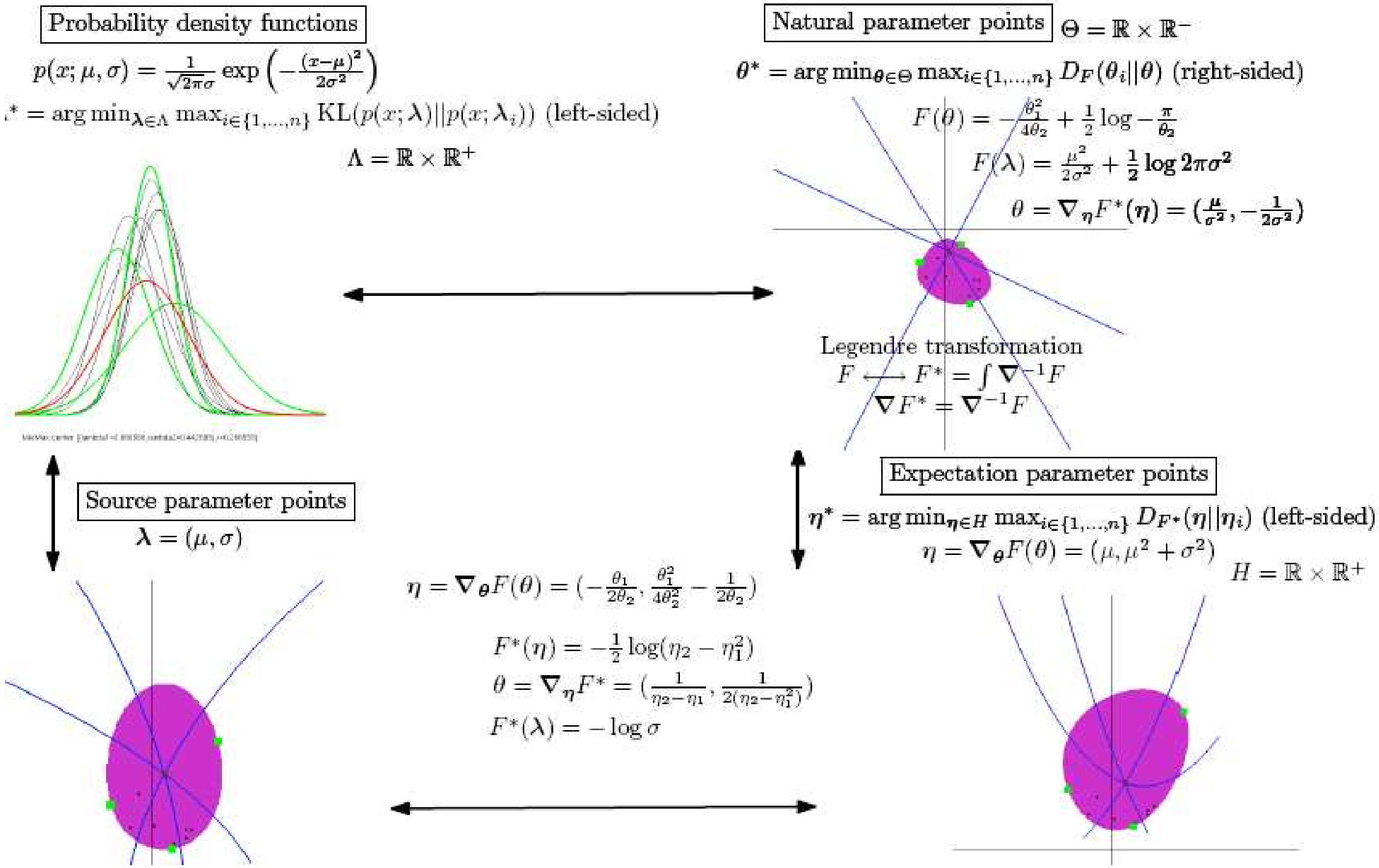
[DEMO]
`http://www.sonycsl.co.jp/person/nielsen/BregmanBall/MINIBALL/`



[IPL'08] On the smallest enclosing information disk.(2008)

Bregman MINIBALL: **Entropic centers**

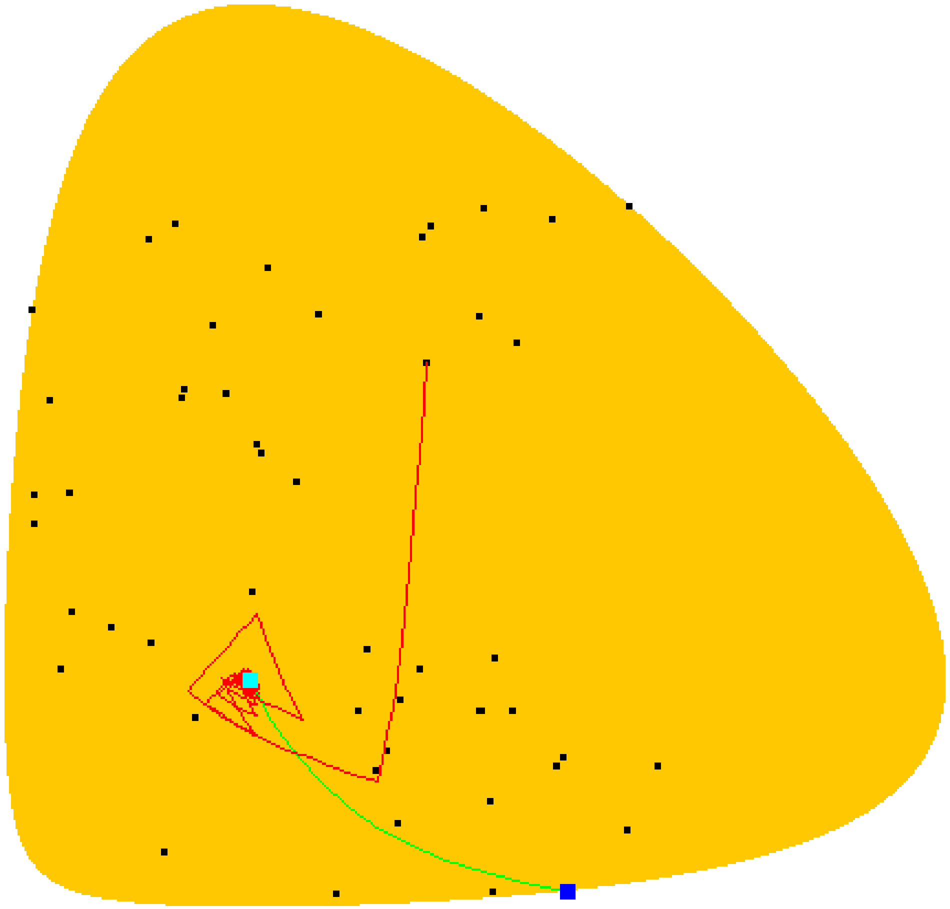
Given a set of n normal distributions $\mathcal{N}_1, ..., \mathcal{N}_n$, find the unique distribution $\mathcal{N}^*$ that **minimizes** the maximum Kullback-Leibler divergence to the others. (KL of exp. fam.= Bregman divergences with parameters swap)



[EuroCG'08] The Entropic Centers of Multivariate Normal Distributions.

Bregman core-sets: Demo

In high dimensions:



`http://www.sonycsl.co.jp/person/nielsen/BregmanBall/MINIBALL/`

[ECML'05] Fitting the smallest enclosing Bregman ball.

Space of Bregman spheres

Right-centered and **left-centered** Bregman balls (with bounding spheres):

$$\mathrm{Ball}^r_F(c,r) = \{x \in \mathcal{X} \mid B_F(x||c) \leq r\} \quad \text{and} \quad \mathrm{Ball}^l_F(c,r) = \{x \in \mathcal{X} \mid B_F(c||x) \leq r\}$$

From Legendre duality, $\mathrm{Ball}^l_F(c,r) = (\nabla F)^{-1}(\mathrm{Ball}^r_{F^*}(\nabla F(c), r))$.

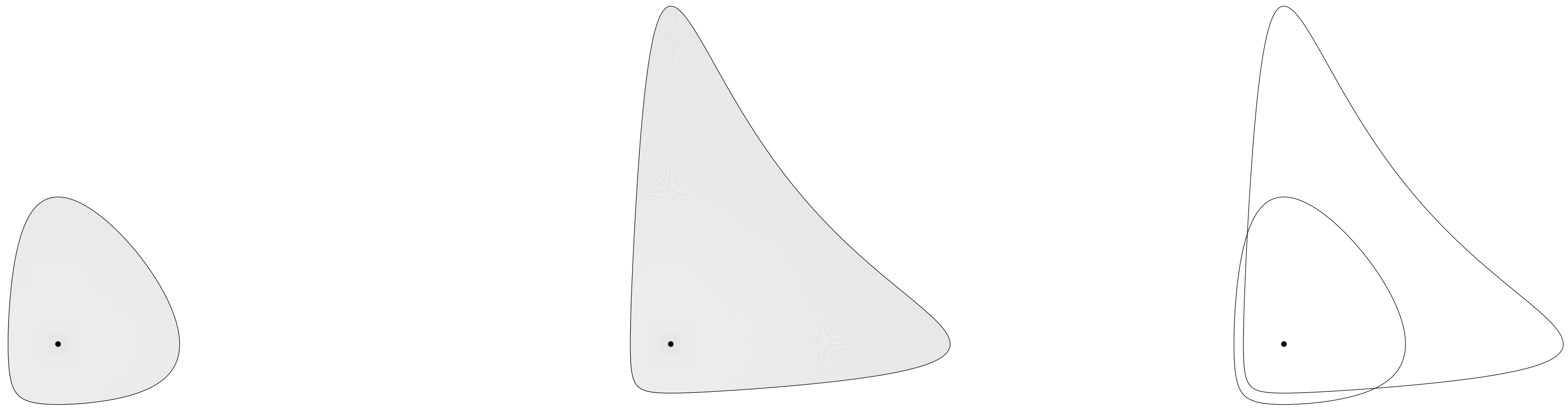
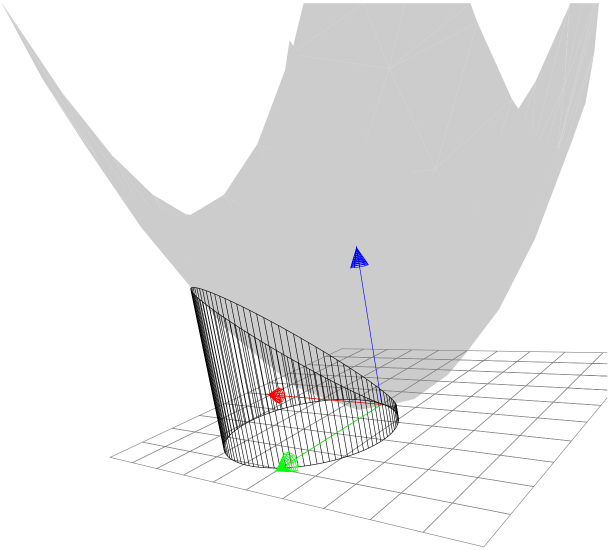
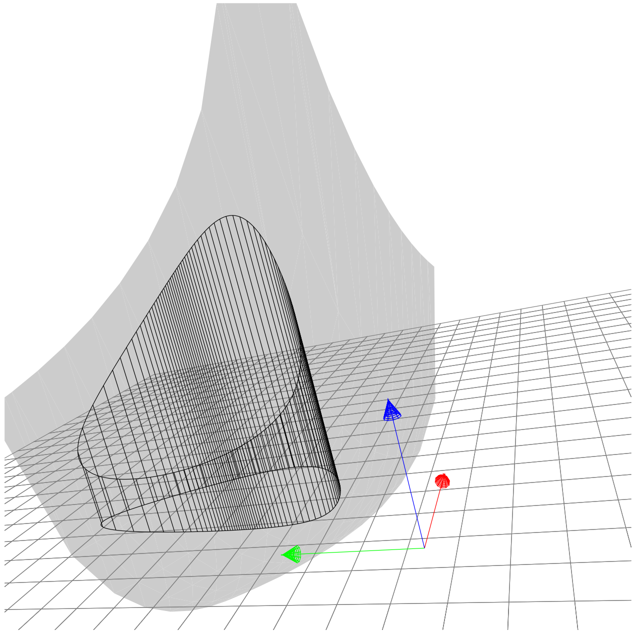


Illustration for Itakura-Saito divergence, $F(x) = -\log x$

Space of Bregman spheres: Lifting map

$\mathcal{F} : x \mapsto \hat{x} = (x, F(x))$, 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)
$\hat{\sigma} = \mathcal{F} \cap H_\sigma$.
Conversely, the intersection of any hyperplane H with $\mathcal{F}$ projects onto $\mathcal{X}$ as a Bregman sphere:
$H : z = \langle x, a \rangle + b \rightarrow \sigma : \mathrm{Ball}_F(c = (\nabla F)^{-1}(a), r = \langle a, c \rangle - F(c) + b)$





InSphere predicates wrt. Bregman divergences

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

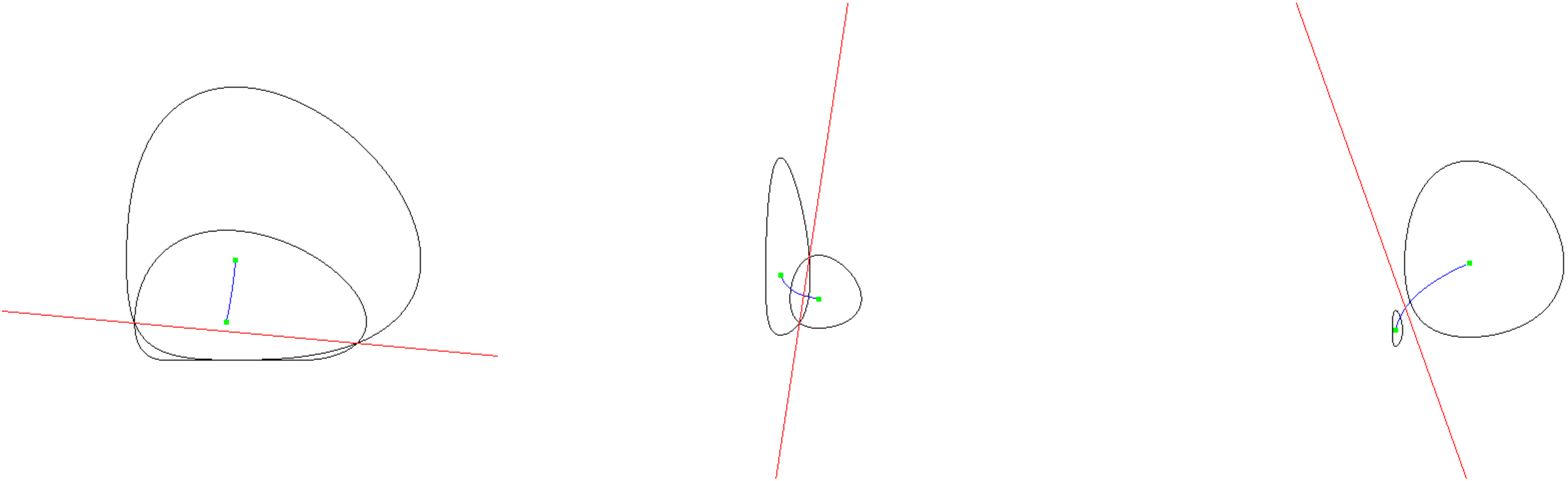
$\texttt{InSphere}(\mathbf{x}; \mathbf{p}_0, ..., \mathbf{p}_d)$ is negative, null or positive depending on whether x lies inside, on, or outside σ.

Space of spheres allows us for practical algorithms for computing the union/intersection of Bregman spheres

[BVD'07] Bregman Voronoi Diagrams: Properties, Algorithms and Applications, arXiv:0709.2196. (SODA'07)

Detecting Bregman ball intersections

$\longrightarrow$ performs a bisection search wrt. the radical axis.



Power to Bregman balls: $H_{12}: \; B_1(x) - B_2(x) = 0$, where
$B_1(x) : B_F(x||p) - r_p = 0$ and $B_2(x) : B_F(x||q) - r_q = 0$

Radical hyperplane

$$H_{12} : F(q) - F(p) + r_2 - r_1 + \langle x, \nabla_F(q) - \nabla_F(p)\rangle + \langle p, \nabla_F(p)\rangle - \langle q, \nabla_F(q)\rangle = 0$$

(EuroCG'09) Tailored Bregman Ball Trees for Effective Nearest Neighbors

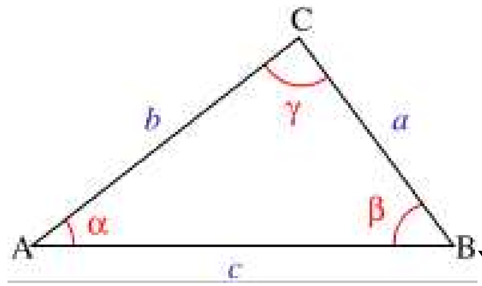
(ICME'09) Bregman vantage point trees for efficient nearest Neighbor Queries

Bregman: Three-point property and Bregman projection

Three-point property:

For any p, q and r of points of $\mathcal{X}$:

$$B_F(p||q) + B_F(q||r) = B_F(p||r) + \underbrace{\langle p - q, \nabla F(r) - \nabla F(q)\rangle}_{\geq 0}$$

(generalizes the law of cosines $c^2 = a^2 + b^2 - 2ab\cos\gamma$ 

)

Bregman projection:

For any p, there exists a **unique point** $x \in \mathcal{W}$ that minimizes $B_F(x||p)$: the Bregman projection of p onto $\mathcal{W}$ ($x^* = p_{\mathcal{W}}$)

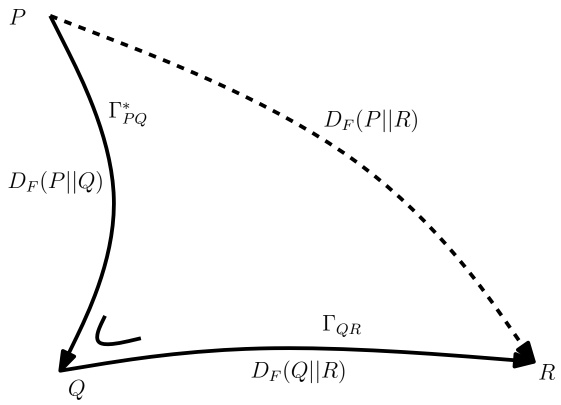
$$p_{\mathcal{W}} = x^* = \arg\min_{x \in \mathcal{W}} B_F(x||p)$$

Note that $p_{\mathcal{W}} = p, \ \forall p \in \mathcal{W}$.

Orthogonality & Generalized Pythagoras' theorem

pq **Bregman orthogonal** to qr iff $B_F(p||q) + B_F(q||r) = B_F(p||r)$.
(Equivalent to $\langle p-q, \nabla F(r) - \nabla F(q)\rangle = 0$) [3-point property])

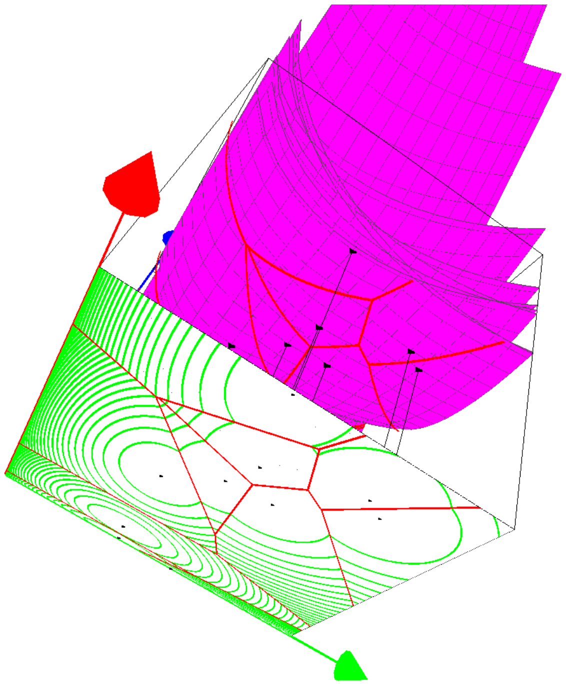
Bregman Pythagoras' inequality:
For convex $\mathcal{W} \subset \mathcal{X}$ and $p \in \mathcal{X}$. We have
$B_F(w||p) \geq B_F(w||p_{\mathcal{W}}) + B_F(p_{\mathcal{W}}||p)$,
with equality for and only for **affine sets** $\mathcal{W}$.



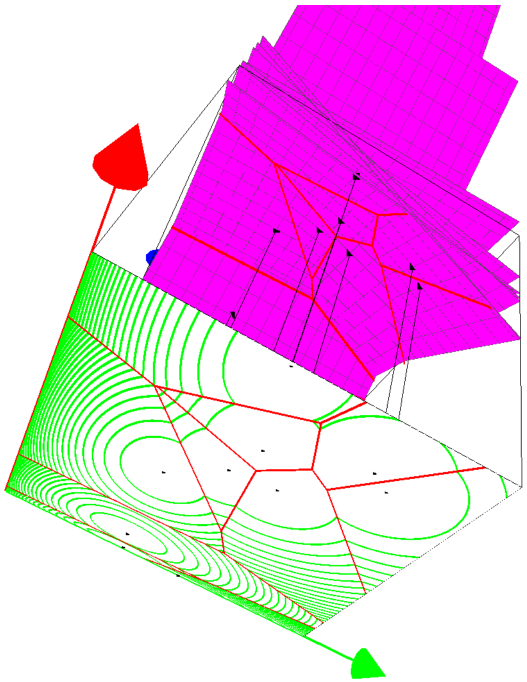
[Bregman'66] The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming.

Bregman Voronoi diagrams as minimization diagrams

A subclass of **affine diagrams** which have all cells non-empty.
Extend Euclidean Voronoi to Voronoi diagrams in dually flat spaces.
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: $H_i(x) = (p_i - x)^T \nabla F(q_i) - F(p_i)$.



$\Longleftrightarrow$



The sided Bregman Voronoi diagrams of n d-dimensional points have complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$
and can be computed in optimal time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.
(SoCG'07) Visualizing Bregman Voronoi diagrams

Bregman Voronoi from Power diagrams

Any affine diagram can be built from a **power diagram** .
(power diagrams are defined in full space $\mathbb{R}^d$, and not only open convex $\mathcal{X}$)

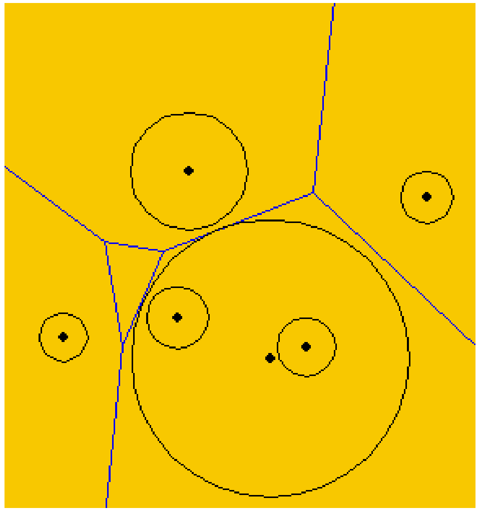
Power distance of x to $\mathrm{Ball}(p, r)$: $||p - x||^2 - r^2$.

Power or Laguerre diagram : minimization diagram of $D_i(x) = ||p_i - x||^2 - r_i^2$

Power bisector of $\mathrm{Ball}(p_i, r_i)$ and $\mathrm{Ball}(p_j, r_j)$= **radical hyperplane** :

$$2\langle x, p_j - p_i\rangle + ||p_i||^2 - ||p_j||^2 + r_j^2 - r_i^2 = 0.$$

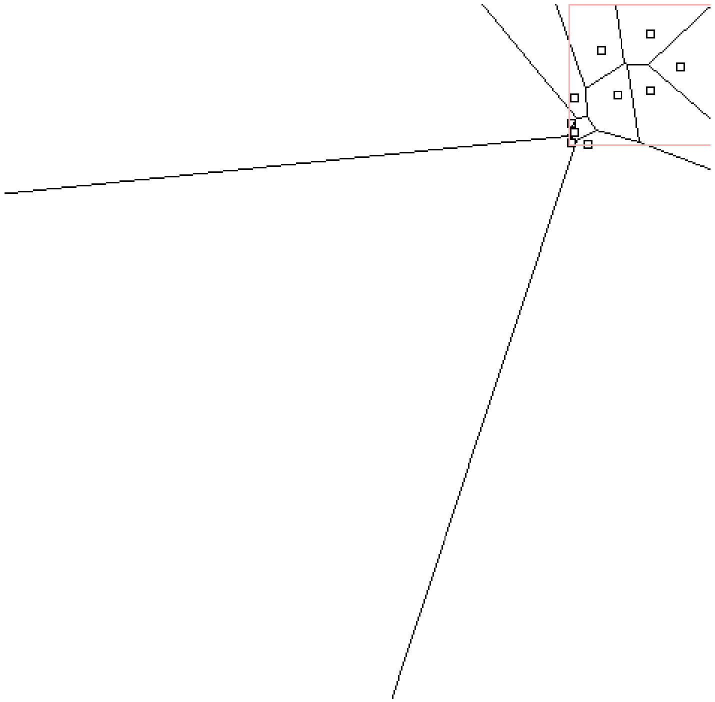
$$\boxed{\text{Affine Bregman Voronoi diagram} \Leftarrow \text{Power diagram}}$$

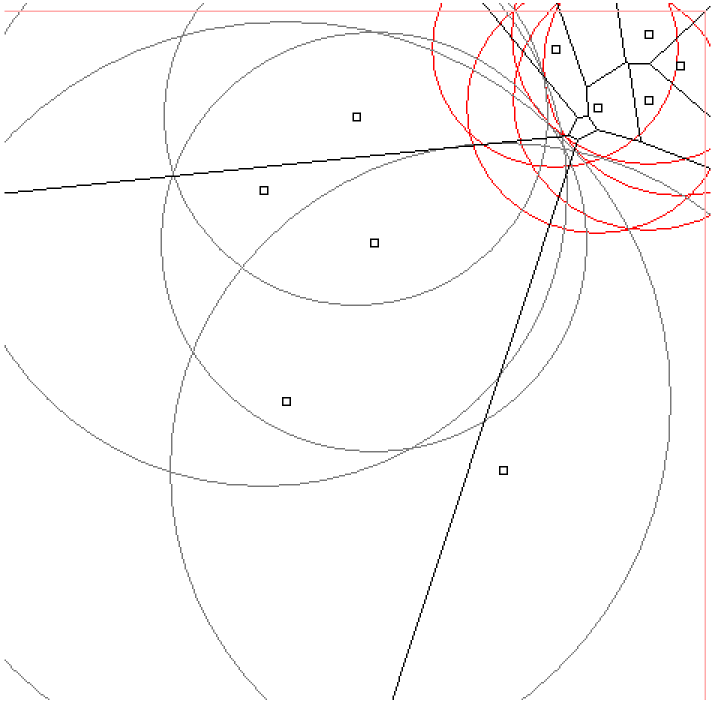


[PVD'87] Franz Aurenhammer: Power Diagrams: Properties, Algorithms and Applications.

Affine Bregman Voronoi diagrams as power diagrams

Equivalence: $B(\nabla F(p_i), r_i)$ with
$r_i^2 = \langle \nabla F(p_i), \nabla F(p_i)\rangle + 2(F(p_i) - \langle p_i, \nabla F(p_i)\rangle$
(**imaginary radii** shown in red)



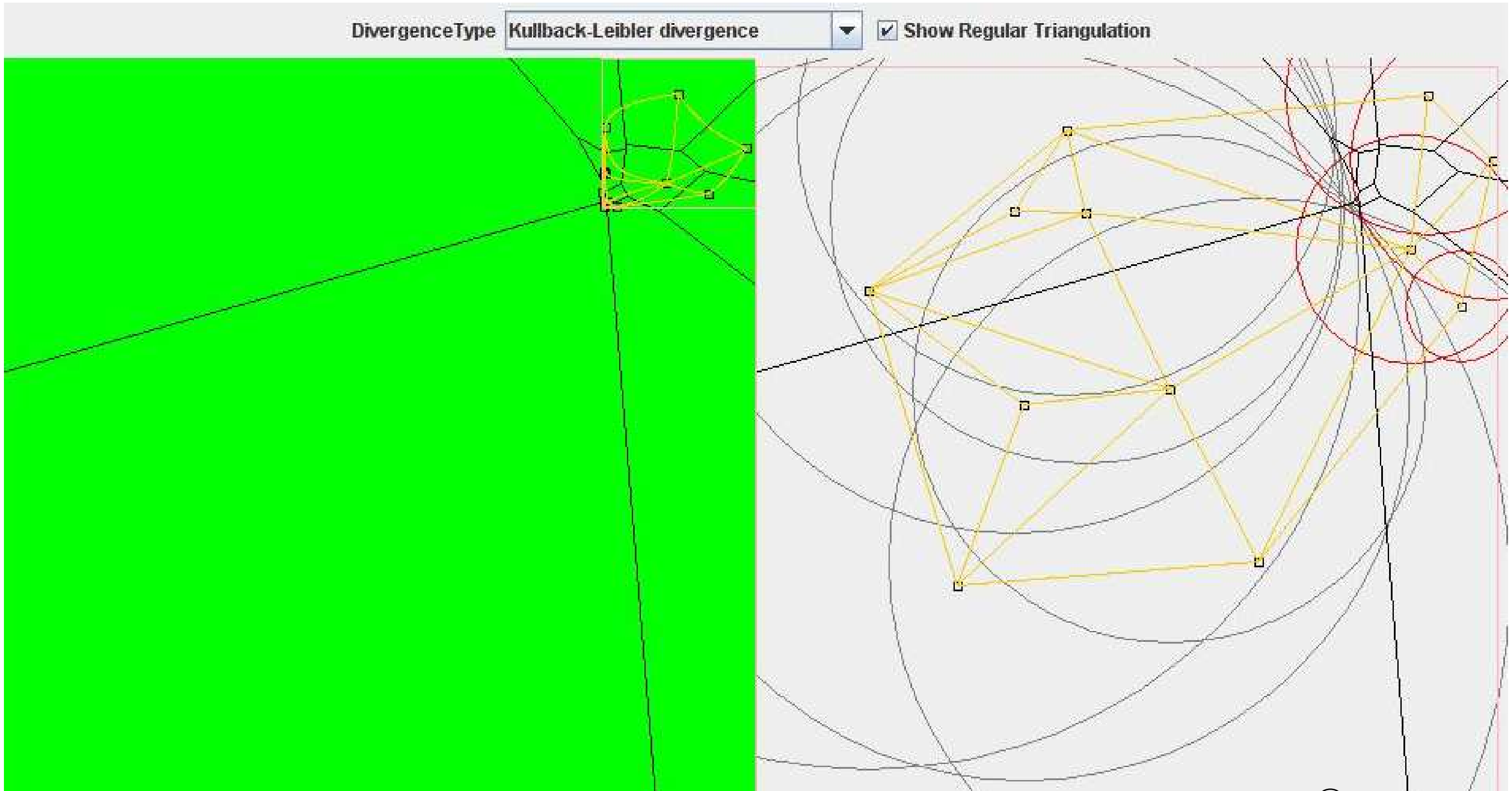


(Some cells may be empty in the Laguerre diagram but not in the Bregman diagram)
Curved Voronoi diagram as dual affine Voronoi diagram
(requires only to compute $\nabla F^* = \nabla F^{-1}$ at source points.)

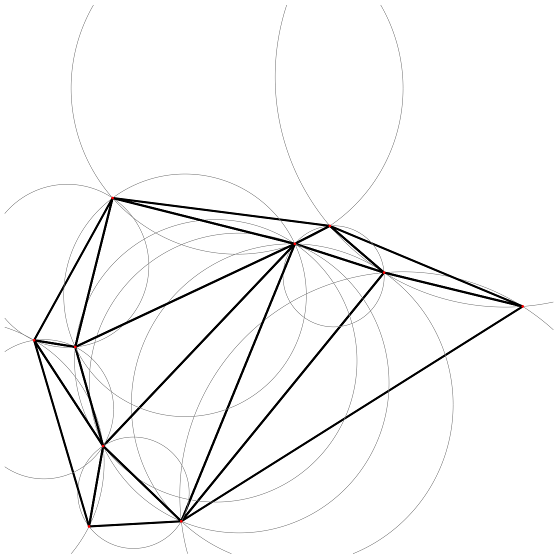
`http://www.csl.sony.co.jp/person/nielsen/BVDapplet/`

Bregman Delaunay/geodesic triangulations

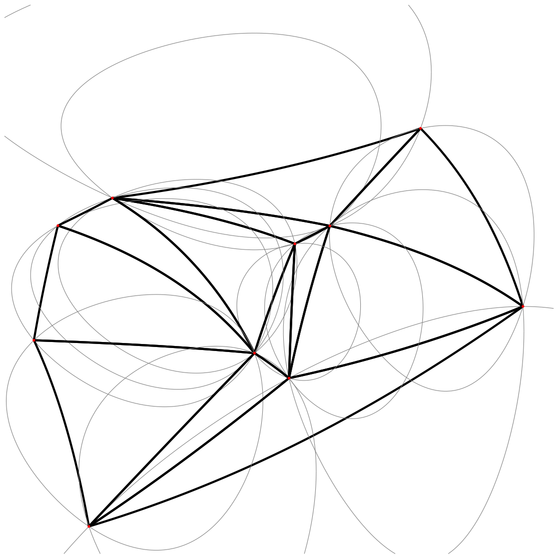
- **Empty-sphere property** : The Bregman sphere circumscribing any simplex of $\mathrm{BT}(\mathcal{P})$ is empty.
- **Optimality** : $\mathrm{BT}(\mathcal{P}) = \min_T \max \tau \in Tr(\tau)$
  ($r(\tau)$: radius of the smallest Bregman ball containing τ)



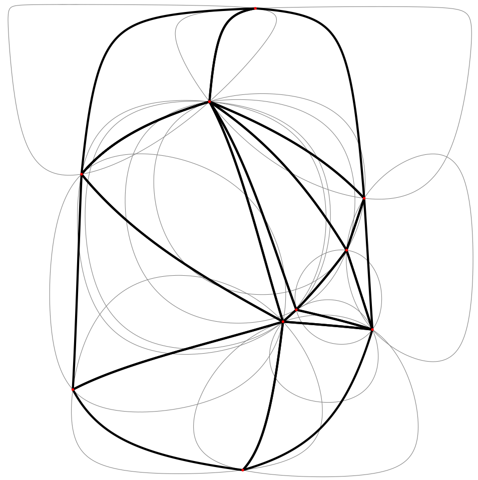
Bregman Delaunay triangulations



Ordinary Delaunay



Exponential loss



Hellinger-like divergence

- empty Bregman sphere property,
- geodesic triangles.

Bregman Voronoi/regular triangulations

Primal space		Dual gradient space
θ		η
Affine BVD($\mathcal{P}$)	=	Laguerre/power diagram on $\nabla F(\mathcal{P})$
$\updownarrow$		$\updownarrow$
Geodesic $\mathrm{BT}(\mathcal{P})$	$\leftrightarrow$	Regular triangulation on $\nabla F(\mathcal{P})$

Bregman Voronoi diagrams extend to **weighted points** :
$W_F(p_i||p_j) = B_F(p_i||p_j) + w_i - w_j$.

Centroids for symmetrized Bregman divergences

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

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

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

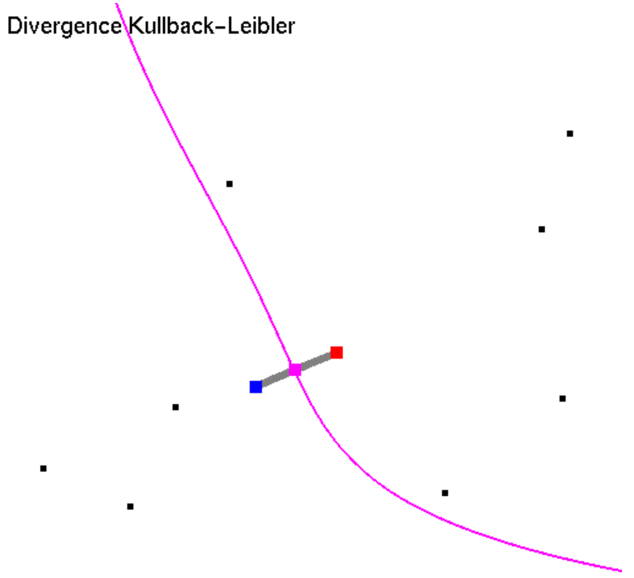
$$\begin{aligned}
\mathrm{AVG}_F(\mathcal{P}||q) &= \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + B_F(\bar{p}||q) \\
\mathrm{AVG}_F(q||\mathcal{P}) &= \mathrm{AVG}_{F^*}({\mathcal{P}_F}'||q') \\
&= (\sum_{i=1}^n \frac{1}{n} F^*(p_i') - F^*(\bar{p'})) + B_{F^*}(\bar{p'_F}||q'_F)
\end{aligned}$$

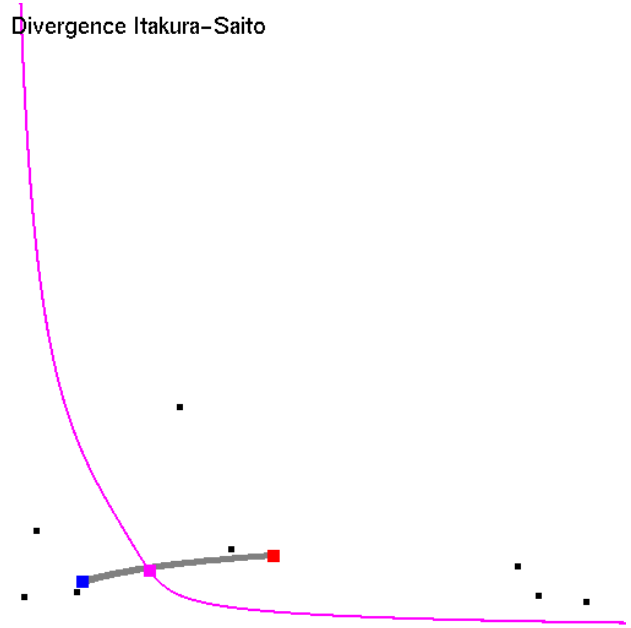
But $B_{F^*}(\bar{p'_F}||q'_F) = B_{F^{**}}(\nabla F^* \circ \nabla F(q)||\nabla F^*(\sum_{i=1}^n \nabla F(p_i))) = B_F(q||c_L^F)$ since $F^{**} = F$, $\nabla F^* = \nabla F^{-1}$ and $\nabla F^* \circ \nabla F(q) = q$.

$\arg\min_{c\in\mathcal{X}} \frac{1}{2}\left(\mathrm{AVG}_F(\mathcal{P}||q) + \mathrm{AVG}_F(q||\mathcal{P})\right) \iff$
$\arg\min_{q\in\mathcal{X}} B_F(c_R^F||q) + B_F(q||c_L^F)$ (removing all terms independent of q)

Symmetrized Bregman centroid

The symmetrized Bregman centroid c^F is **uniquely** defined as the minimizer of $B_F(c_R^F||q) + B_F(q||c_L^F)$. 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: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid B_F(c_R^F||x) = B_F(x||c_L^F)\}$.





[ICPR'08] Bregman Sided and Symmetrized Centroids. arXiv 0711.3242

Meaning of duality/invariance in information geometry

On the manifold of probability measures $\{p_F(x|\theta) \mid \theta \in \Theta\}$:

- **Reparameterization** : $p_F(x|\lambda)$ same as $p_F(x|\theta)$ for a bijective mapping $\lambda \leftrightarrow \theta$.
- **Reference duality** :
  Choice of the reference vs comparison points:
  $B_F(p||q) = B_{F^*}(\nabla F(q)||\nabla F(p))$.
- **Representational duality** :
  Choice of a monotonic scaling (density or positive measures).

Canonical divergence :

$$
\begin{aligned}
A_F(\theta||\eta) &= B_F(\theta||\nabla F^{-1}(\eta)) \\
&= F(\theta) + F^*(\eta) - \langle \theta, \eta \rangle \geq 0 \text{ (Legendre inequality)} \\
&= A_{F*}(\eta||\theta)
\end{aligned}
$$

Given a divergence B_F, we can derive a **Riemannian metric** and a pair of **conjugate affine connections** [Eguchi'83].

Bregman divergence as exact Taylor remainder

For a given fixed point p, we can view the geometry as a Riemannian geometry.
Bregman divergence:

$$
\begin{aligned}
B_F(p||q) &= F(p) - F(q) - (p-q)^T \nabla F(q) \\
& \frac{1}{2}(p-q)^T \nabla^2 F(\varepsilon)(p-q),
\end{aligned}
$$

with $\varepsilon \in [pq]$.
$\nabla^2 F$: Hessian of F, positive-definite matrix (psd.): $\nabla^2 F \succ 0$.

Example for I-divergence $I(p||q) = \sum_{i=1}^d p_i \log \frac{p_i}{q_i} + q_i - p_i$:
$\nabla^2 F(\varepsilon) = \mathrm{diag}(\frac{1}{x_1}, ..., \frac{1}{x_i}, ..., \frac{1}{x_d}) \succ 0$ for $\varepsilon \in \mathbb{R}_+^d$, $\varepsilon_i = \frac{(p_i - q_i)^2}{2p_i \log \frac{p_i}{q_i} + q_i - p_i}$ with $\varepsilon \in [pq]$
Numerical example (1D):

```
p=0.4200869374923376, q=0.5899178549202998, I=0.0272023222342305 8,
vareps=0.5301484973611689, vareps belongs to [p,q]
```

Representational Bregman divergences

- Bregman generator

$$U(\mathbf{x}) = \sum_{i=1}^{d} U(x_i) = \sum_{i=1}^{d} U(k(s_i)) = F(\mathbf{s})$$

  with $F = U \circ k$.

- Dual 1D generator $U^*(x^*) = \max_x \{xx^* - U(x)\}$ induces dual coordinate system $x_i^* = U'(x_i)$, where U' denotes the derivative of U. $\nabla U(\mathbf{x}) = [U'(x_1) \ldots U'(x_d)]^T$.

Canonical **separable representational Bregman divergence**:

$$B_{U,k}(\mathbf{p}||\mathbf{q}) = U(k(\mathbf{p})) + U^*(k^*(\mathbf{q}^*)) - \langle k(\mathbf{p}), k^*(\mathbf{q}^*)\rangle,$$

with $k^*(\mathbf{x}^*) = U'(k(\mathbf{x}))$.

Often, a Bregman by setting $F = U \circ k$. But although U is a strictly convex and differentiable function and k a strictly monotonous function, $F = U \circ k$ may not be strictly convex.

[ISVD'09] The dual Voronoi diagrams with respect to representational Bregman divergences

Amari's α-divergences

α-divergences on positive arrays (unnormalized discrete probabilities), $\alpha \in \mathbb{R}$:

$$D_\alpha(\mathbf{p}||\mathbf{q}) = \begin{cases} \sum_{i=1}^d \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}p_i + \frac{1+\alpha}{2}q_i - p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\right) & \alpha \neq \pm 1 \\ \sum_{i=1}^d p_i \log \frac{p_i}{q_i} + q_i - p_i = \mathrm{KL}(\mathbf{p}||\mathbf{q}) & \alpha = -1 \\ \sum_{i=1}^d q_i \log \frac{q_i}{p_i} + p_i - q_i = \mathrm{KL}(\mathbf{q}||\mathbf{p}) & \alpha = 1 \end{cases}$$

Duality

$$\boxed{D_\alpha(\mathbf{p}||\mathbf{q}) = D_{-\alpha}(\mathbf{q}||\mathbf{p}).}$$

Representational Bregman divergences of α-/β-divergences

Divergence	Convex conjugate functions	Representation functions
Bregman divergences B_F, B_{F^*}	U $U' = (U^{*\prime})^{-1}$ U^*	$k(x) = x$ $k^*(x) = U'(k(x))$
α-divergences$(\alpha \neq \pm 1)$ $F_\alpha(x) = \frac{2}{1+\alpha} x$ $F^*_\alpha(x) = \frac{2}{1-\alpha} x$	$U_\alpha(x) = \frac{2}{1+\alpha}(\frac{1-\alpha}{2} x)^{\frac{2}{1-\alpha}}$ $U'_\alpha(x) = \frac{2}{1+\alpha}(\frac{1-\alpha}{2} x)^{\frac{1+\alpha}{1-\alpha}}$ $U^*_\alpha(x) = \frac{2}{1-\alpha}(\frac{1+\alpha}{2} x)^{\frac{2}{1+\alpha}} = U_{-\alpha}(x)$	$k_\alpha(x) = \frac{2}{1-\alpha} x^{\frac{1-\alpha}{2}}$ $k^*_\alpha(x) = \frac{2}{1+\alpha} x^{\frac{1+\alpha}{2}} = k_{-\alpha}(x)$
β-divergences$(\beta > 0)$ $F_\beta(x) = \frac{1}{\beta+1} x^{\beta+1}$ $F^*_\beta(x) = \frac{x^{\beta+1}-x}{\beta(\beta+1)}$	$U_\beta(x) = \frac{1}{\beta+1}(1+\beta x)^{\frac{1+\beta}{\beta}}$ $U'_\beta(x) = (1+\beta x)^{\frac{1}{\beta}} \qquad U^{*\prime}_\beta(x) = \frac{x^\beta - 1}{\beta}$ $U^*_\beta(x) = \frac{x^{\beta+1}-x}{\beta(\beta+1)}$	$k_\beta(x) = \frac{x^\beta - 1}{\beta}$ $k^*_\beta(x) = x$

α- and β-divergences are representational Bregman divergences in disguise.

Riemannian statistical manifolds

Fisher information and induced Riemannian metric:

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)|\theta\right] = g_{ij}(\theta)$$

For exponential families:

$$I(\theta) = \nabla^2 F(\theta)$$

$\boxed{\text{Distance is geodesic length (Rao, 1945)}}$

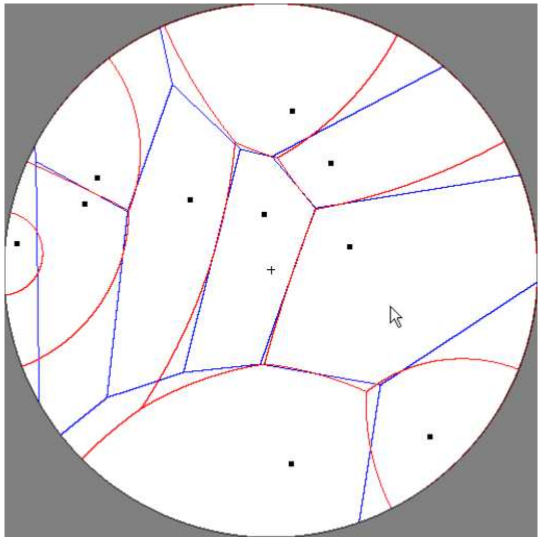
$$D(P,Q) = \int_{t=0}^{t=1}\sqrt{g_{ij}(t(\theta))}\mathrm{d}t, \qquad t(\theta_0) = \theta(P), t(\theta_1) = \theta(Q)$$

Fisher-Rao Riemannian geometries:
Multinomial distributions $\Rightarrow$ Spherical geometries (constant curvature 1).
Normal distributions $\Rightarrow$ Hyperbolic geometries (negative curvature).

Voronoi diagram in embedded geometries

Imaginary geometry can be realized in **many different** ways.
For example, hyperbolic geometry:

- Conformal Poincaré upper half-space,
- Conformal Poincaré disk (in red),
- Non-conformal Klein disk (in blue),
- Pseudo-sphere in Euclidean geometry, etc.



Hyperbolic Voronoi diagrams made easy, arXiv:0903.3287, 2009.

Distance between two corresponding points in *any* isometric embedding is the same.

Summary

- Bregman divergences unifies squared Euclidean distance with Kullback-Leibler divergence.
- Bregman divergences = canonical divergences of flat spaces with ± 1-connections.
- Bregman geometries = flat geometries with Bregman projection/generalized Pythagoras theorems.
- ± 1-connections are compatible with the Fisher metric.

Statistical manifolds with invariance [Chentsov'72]: Fisher metric and α-connections only. $\alpha = \pm 1 \Rightarrow$ Dually flat spaces.
Perspectives:

- α-geometries and its applications
- Choice of embeddings for relevant computations

Thank you

Collaborators: Shun-ichi Amari, Michel Barlaud, Jean-Daniel Boissonnat, Sylvain Boltz, Meizhu Liu, Richard Nock, Paolo Piro, Olivier Schwander, Baba Vemuri.

- ANR-07-BLAN-0328-01 GAIA (Computational Information Geometry and Applications).
- DIGITEO GAS 2008-16D (Geometric Algorithms & Statistics)

- `http://www.informationgeometry.org/`
- `http://blog.informationgeometry.org/`
- ETVC'08: Emerging Trends in Visual Computing: `http://www.lix.polytechnique.fr/Labo/Frank.Nielsen/ETVC08/`

All geometries are false but some geometries are useful.

Remember that all geometries are wrong; the practical question is how wrong do they have to be to not be useful.