Computational Information Geometry

From Euclidean to dually flat spaces

Frank Nielsen

École Polytechnique, LIX, France

Sony Computer Science Laboratories, FRL, Japan

Acknowledgments

These slides were prepared and used for a 2-hour lecture at the French Computational Geometry days (JGA 2009).

`http://www.lmia.uha.fr/~mage/jga2009/1.htm`

- Collaborators: Jean-Daniel Boissonnat (INRIA) and Richard Nock (UAG)
- ANR-07-BLAN-0328-01 GAIA (Computational Information Geometry and Applications).
- DIGITEO GAS 2008-16D (Geometric Algorithms & Statistics)

```
@misc{jga-compgeom-flatspaces-2009
, title           = "Computational Geometry in Dually Flat Spaces"
, author          = "Frank Nielsen"
, year            = "2009"}
```

Summary of the lecture

High-dimensional noisy datasets encoding heterogeneous feature vectors abound in many real-world applications. On one hand, practitioners face the crucial problem of choosing the *best-suited* distance allowing them to use their algorithmic toolboxes. On the other hand, theoreticians are focusing on *recovering* the intrinsic dimensionality and underlying topology of datasets, and focus on learning the most appropriate distances. In the first part of the lecture, we start by presenting recent advances of the celebrated k-means clustering algorithm (namely, G-means++ that automatically learns the proper number of clusters using a careful seeding). We then present broad generalizations of k-means: convex k-means and *Bregman k-means* (that preserves the centroid relocation scheme). We introduce the notion of Bregman divergences and Bregman information, and explain the fundamental convex duality arising from the Legendre-Fenchel transform. We describe the 1-to-1 map from Bregman divergences to statistical exponential families that allows one to design easy *soft-clustering algorithms* (expectation-maximization). In the second part of the lecture, we present the underlying geometry induced by a convex contrast function: Dually flat spaces generalizing the traditional Euclidean geometry (self-dual). We explain how the standard Euclidean notions of bisectors/projection/orthogonality and geodesics extend to these spaces, and present generalizations of common algorithms and data-structures in computational geometry: smallest enclosing ball, Voronoi and Delaunay/regular triangulations, etc. We then restate these results under the framework of information geometry. Finally, we conclude with other broad classes of distances (eg., Csiszár's f-divergences, f-Jensen semi-distances, α-divergences) with their corresponding non-flat geometries.

High-dimensional datasets abound in applications

Heterogeneous feature **vector** spaces (ie., $\mathcal{F} = \prod_{i=1}^{m} \mathcal{F}_i$).
Noisy datasets (non-Gaussian, spatially-variant anisotropic).

- 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, bioinformatics, etc.

Eg., UCI and KDD machine learning and knowledge discovery repositories.

21^{st} Century data processing challenges

Practitioners.

- Which distance function is most **appropriate** ?
- Does my toolbox (eg., algorithms/DSs) **handles** that distance?

Theoreticians.

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

Lecture plan

Part I. Extending Euclidean (L_2) algorithms to Bregman divergences :

- Modern k-means clustering,
- Bregman k-means,
- Bregman soft clustering
  (as known as *expectation-maximization made easy*),

Part II. Information geometry, flat and curved spaces :

- Bregman ball trees and vantage point trees (nearest neighbours),
- Bregman smallest enclosing balls.
- Bregman Voronoi & dual Bregman triangulations,
- Geometrization of statistics (differential geometry/invariance),
- Bregman divergences as canonical distances of flat spaces,
- Sea of geometries, distances, densities and means.

Clustering with k-means
— Les nuées dynamiques —

Lloyd's iterative k-means refinement (1956, 1957)

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** :

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

- 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 $\mathcal{V}_i$

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

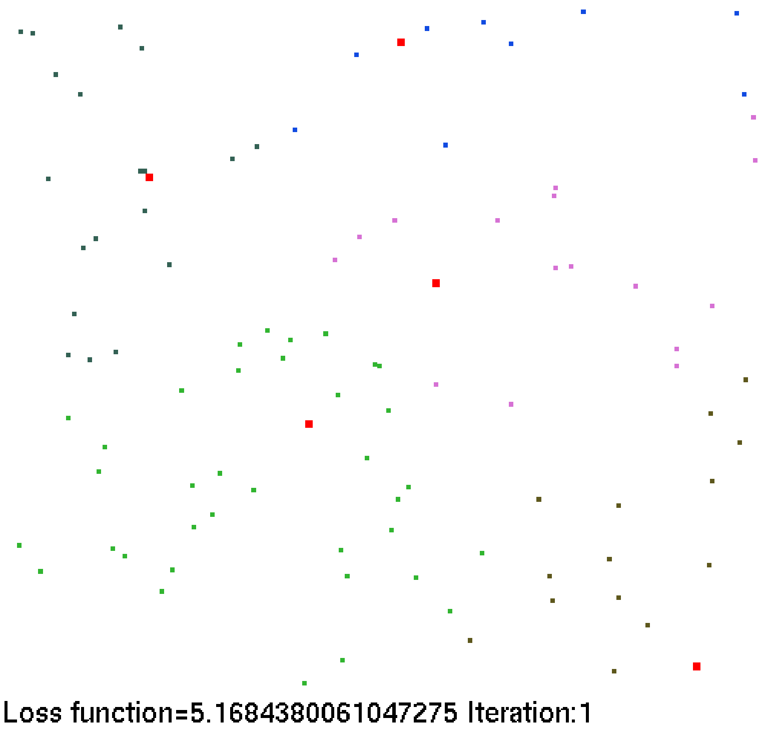
k-means: Potential/loss function

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

L **monotonically** converges. Lloyd' iterations only minimize **locally** L.

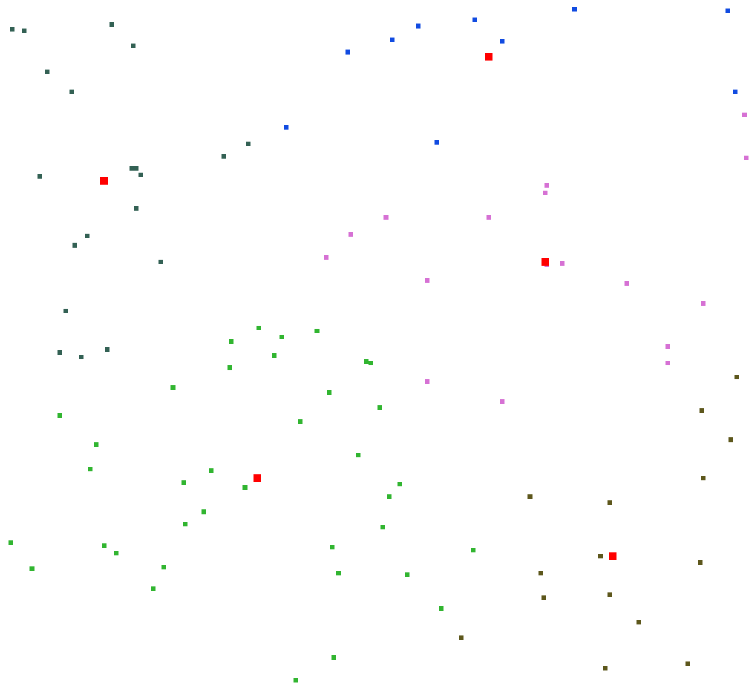
All clusters are always **non-empty**

Never repeat a configuration: upper bound #iter $= O(k^n)$



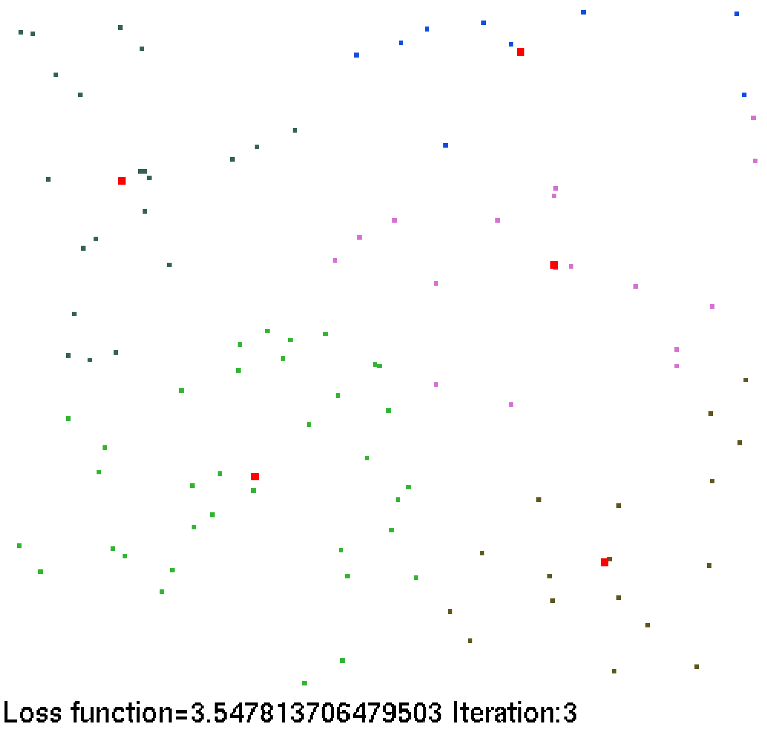
```
Loss function=5.1684380061047275 Iteration:1
```

k-means in action (Iteration 2)



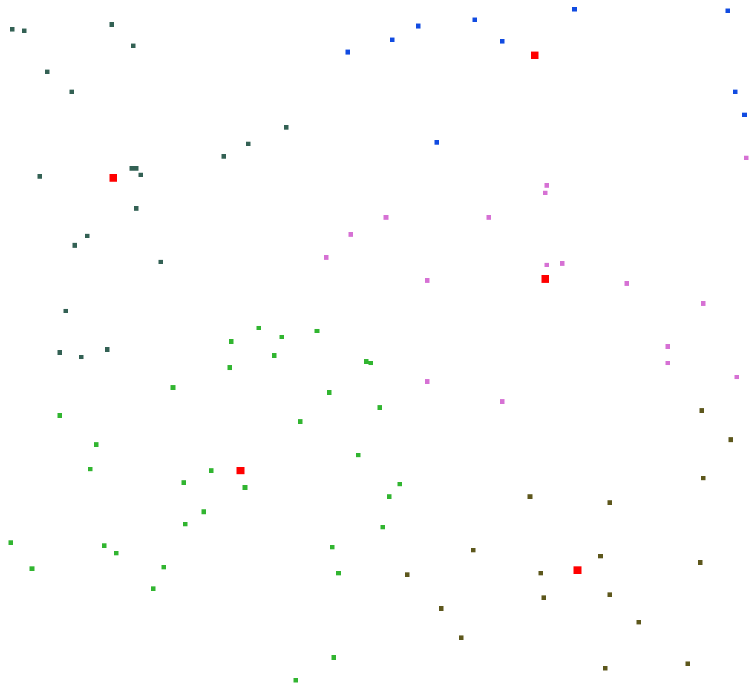
Loss function=3.688985750441276 Iteration:2

k-means in action (Iteration 3)



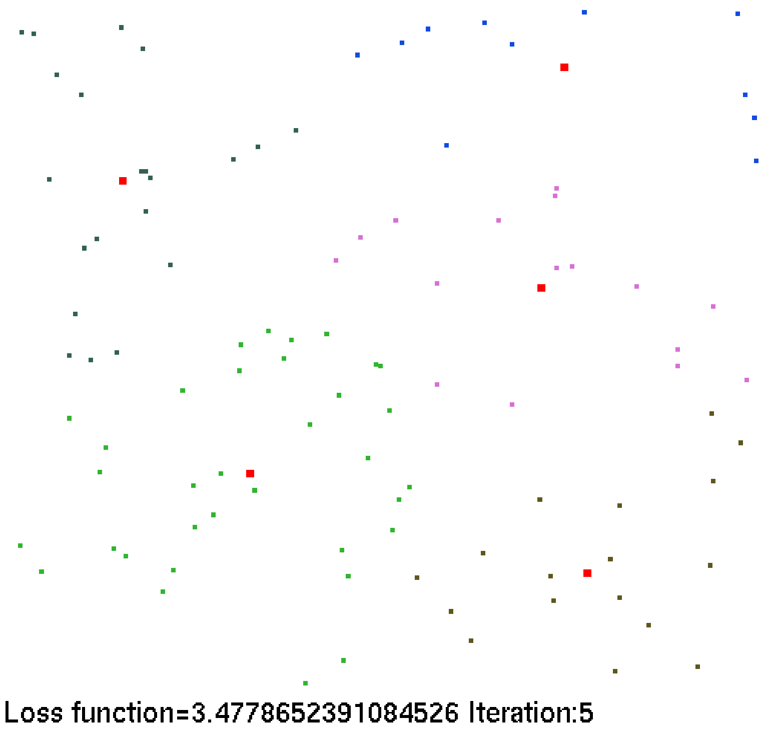
Loss function=3.547813706479503 Iteration:3

k-means in action (Iteration 4)



Loss function=3.493301240276355 Iteration:4

k-means in action (Iteration 5)



Loss function=3.4778652391084526 Iteration:5

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

Convergence proof ($k = 1$)

The **centroid** minimizes the following MinAvg optimization problem:

$$p^* = \arg\min_p \frac{1}{n}\sum_i ||p_i - p||^2 = \frac{1}{n}\sum_i p_i \overset{\text{equal}}{=} \bar{p}$$

Proof:

$$\begin{aligned}
&= && \frac{1}{n}\left(\sum_{i=1}^n ||p_i||^2 - 2\sum_{i=1}^n \langle p_i, p\rangle + ||p||^2\right) \\
&= && \left(\frac{1}{n}\sum_{i=1}^n ||p_i||^2\right) - 2\langle p_i, \bar{p}\rangle + ||p||^2 \\
&\overset{\text{equiv.}}{\equiv} && ||p||^2 - 2\langle p_i, \bar{p}\rangle \\
&\equiv && ||p||^2 - 2\langle p_i, \bar{p}\rangle + ||\bar{p}||^2 \\
&= && ||p - \bar{p}||^2 \geq 0
\end{aligned}$$

→ minimum obtained for the **center of mass** : $p = \bar{p}$.
Distortion: Cluster **variance** wrt. to the centroid: $\frac{1}{n}\sum_i ||p_i - \bar{p}||^2$
Note: MinAvg wrt. L_2 yields **Fermat-Weber** point (non-closed formula).

Convergence proof of k-means

Loss function to minimize:

$$\begin{aligned} L(\mathcal{P}) &= \sum_{p\in\mathcal{P}} \min_{c\in\mathcal{C}} \|p-c\|^2 \\ &= \sum_{i=1}^{k} \sum_{p\in\mathcal{C}_i} \|p-c\|^2 \overset{\text{equal}}{=} \mathrm{cost}(\mathcal{C}_1, ..., \mathcal{C}_k; c_1, ..., c_k) \end{aligned}$$

Assign each point to its closest center: (assign for each point a "closer" center)

$$\mathrm{cost}(\mathcal{C}_1^{(t+1)}, ..., \mathcal{C}_k^{(t+1)}; c_1^{(t)}, ..., c_k^{(t)}) \leq \mathrm{cost}(\mathcal{C}_1^{(t)}, ..., \mathcal{C}_k^{(t)}; c_1^{(t)}, ..., c_k^{(t)})$$

Relocate cluster centers: (update MinAvg in each cluster)

$$\mathrm{cost}(\mathcal{C}_1^{(t+1)}, ..., \mathcal{C}_k^{(t+1)}; c_1^{(t+1)}, ..., c_k^{(t+1)}) \leq \mathrm{cost}(\mathcal{C}_1^{(t+1)}, ..., \mathcal{C}_k^{(t+1)}; c_1^{(t)}, ..., c_k^{(t)})$$

Thus at each Lloyd iteration:

$$\mathrm{cost}(\mathcal{C}_1^{(t+1)}, ..., \mathcal{C}_k^{(t+1)}; c_1^{(t+1)}, ..., c_k^{(t+1)}) \leq \mathrm{cost}(\mathcal{C}_1^{(t)}, ..., \mathcal{C}_k^{(t)}; c_1^{(t)}, ..., c_k^{(t)})$$

How fast/slow is k-means?

Fast in practice (*de facto* algorithm):

$O(kn^2\Delta^2)$ iterations for point set with spread Δ

(→ bound obtained by Har Peled's 1-point variant=online mode)

→ Seeding is crucial

(Many variants: Forgy, furthest point, pairwise NN, SVD, etc.).

Slow in theory :

There exists point sets requiring **super-polynomial** number of iterations: $2^{\Omega(\sqrt{n})}$.

Fortunately, smooth probabilistic analysis reconciles practice:

$\tilde{O}(n^k)$ (point sets drawn from iid. $N(I, \sigma^2)$)

Trivial upper-bounds: combinatorial $O(k^n)$, geometric $O(n^{kd})$

But finding **global minimum** $L^*(\mathcal{P})$ is **NP-hard** for $k \geq 2$!

Careful seeding of k-means++

Good seeding is required for nice clustering.
(beyond Forgy's classic random seed initialization & several tries)

$\boxed{k\textbf{-means++ } D^2 \textbf{ initialization}}$ [SODA'07]:

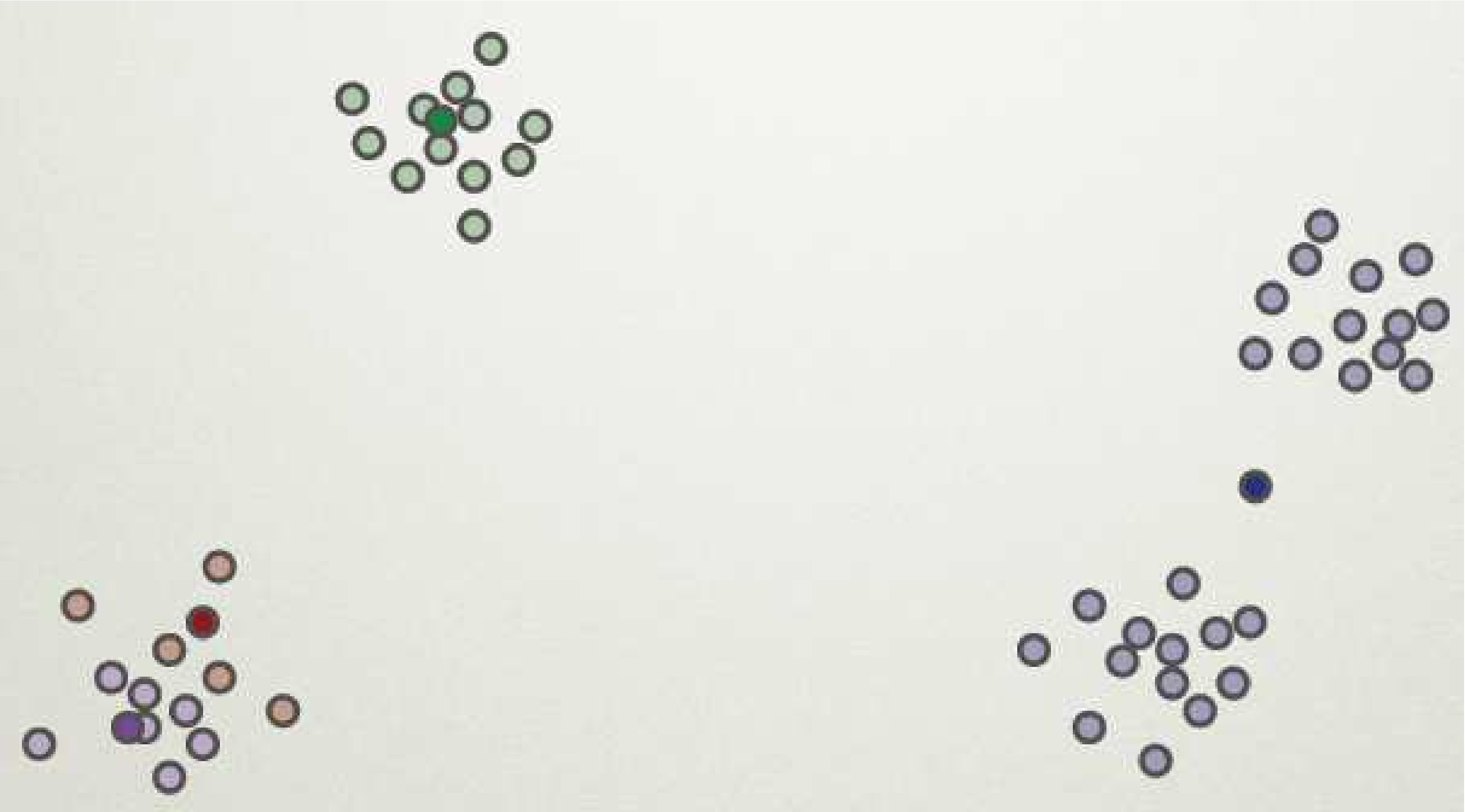
- First, choose $\mathcal{C}_1 = \{c_1\}$ uniformly at random from $\mathcal{V}$,
- Choose $c_2, ..., c_k$ *iteratively* at random, selecting $c_i = v \in \mathcal{V}$ with probability $\Pr(v) = \frac{D^2(v,\mathcal{C}_{i-1})}{\sum_{v\in\mathcal{V}} D^2(v,\mathcal{C}_{i-1})}$, where $D^2(v,\mathcal{V}) = \min_{v_j\in\mathcal{V}} ||v_j - v||^2$ and $\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{c_i\}$.

Squared loss function: $L(\mathcal{V},\mathcal{C}) = \sum_{i=1}^n \min_{j=1}^k ||v_i - c_j||^2$.
Global optimum: $L^*(\mathcal{V}) = \min_{\mathcal{C}} L(\mathcal{V},\mathcal{C})$ (NP-hard even for $k=2$).

$\boxed{k\text{-means++ is } 8(\log k + 2)\text{-competitive: } L(\mathcal{V}) \leq 8(\log k+2)L^*(\mathcal{V}).}$

$(9+\epsilon)$-competitive scheme in $O(n^3\epsilon^{-d})$ [Kanungo, 2002]

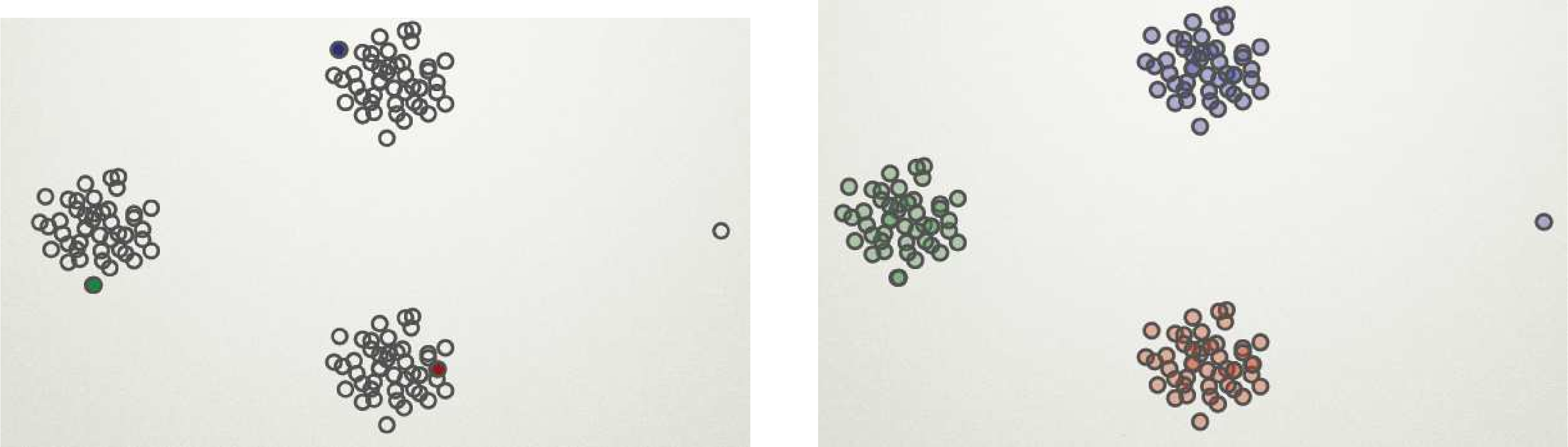
k-means with Forgy initialization ($k = 4$)



$\Longrightarrow$ Not all clusters were captured (bad luck case)

Picture from `http://www.stanford.edu/~darthur/`

k-means++: Probabilistic initialization



$\longrightarrow$ k-means++ initialization is robust to noise and outliers.

Picture from `http://www.stanford.edu/~darthur/`

D^2-initialization generalizes former heuristics. Consider D^α:

- Forgy: $\alpha = 0$ (equiprobability),
- Further point: $\alpha = \infty$ (=2-approximation to k-center heuristic). Sensitive to outliers.
- k-means++: $\alpha = 2$.

Guessing the k in k-means

How many globular clusters is there?

Gaussian means (G-means, NIPS 2003)

G-means :

- Initialization: Let $\mathcal{C} = \{c_1 = \bar{v}\}$
- Repeat until all clusters are statistically checked to be Gaussian
  - $\mathcal{V}_1, ..., \mathcal{V}_{|\mathcal{C}|} \leftarrow k$-means($\mathcal{V}$)
  - For all clusters $\mathcal{V}_i$, **check** whether $\mathcal{V}_i$ is Gaussian or not.
    If not, split c_i into two centers (eg., 2-means++ initialization on $\mathcal{V}_i$).

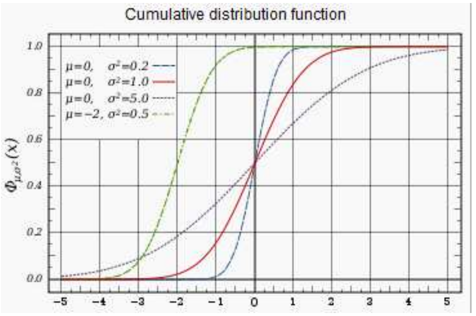
Statisticians: Normality test, goodness of fit, ...
Sea of approaches (100+) !

D'Agostino R. and Stephens M. (1986)
Goodness-of-Fit Techniques, Marcel Dekker, New York.

Normal/Gaussian test: Anderson-Darling test

Are 1D $x_1, ..., x_n$ gaussianly distributed? (normality test)
Anderson-Darling (1952) test for scalars with **confidence level** γ (eg., $\gamma = 0.0001$).
Based on **empirical cumulative distribution function** (ecdf):

- Remap $x_1, ..., x_n$ into **normalized ordered** sequence $x^{(1)}, ..., x^{(n)}$ (with centered mean zero and unit variance)
- Let $z_i = \mathrm{cdf}(x^{(i)})$ where cdf is the Gaussian cdf.:



- Compute statistics:
$$A^2(\mathcal{Z}) = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)(\log z_i + \log(1 - z_{n+1-i})) - n$$
- Readjust statistics (to take into account **sample** mean/variance)
$$A^2_*(\mathcal{Z}) = A^2(\mathcal{Z})(1 + \frac{4}{n} - \frac{25}{n^2})$$

G-means: Statistical multivariate normality test

For d-dimensional vector points $\mathcal{V}$, check whether it is gaussianly distributed or not. (with arbitrary mean and variance-covariance matrix)

Statistical test AndersonDarling$(\mathcal{V}, c, \gamma)$:

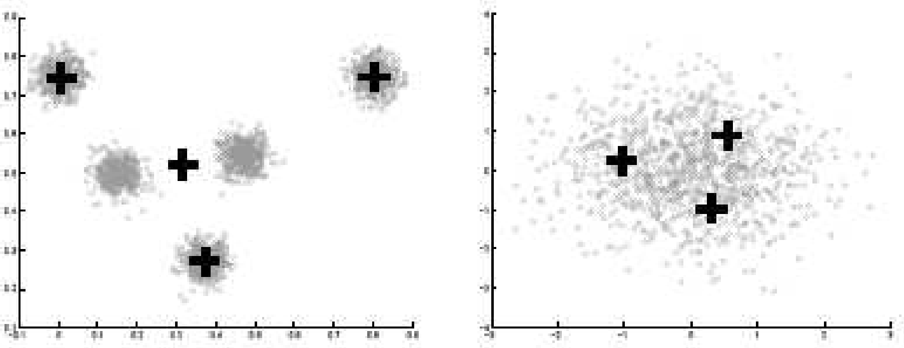
- Run a 2-means++ to get centers c_1 and c_2
- Let $a = c_1 - c_2$ denote the orientation connecting c_1 with c_2
- Projection onto line $(c_1 c_2)$: Consider scalars $v_i' = \frac{<v_i, a>}{||a||^2}$
- Normalize: Adjust $\mathcal{V}'$ so that it has sample mean zero, and unit variance: $\rightarrow$ yields Z (1D ordered sequence).
- Compute $A_*^2(Z)$
- If $A_*^2(Z) \leq t_\gamma$ (**critical value**) then accept Gaussian hypothesis, otherwise reject hypothesis and replace c by c_1 and c_2.

(Eg., for $\gamma = 0.0001$, critical value $t_\gamma = 38.103$)

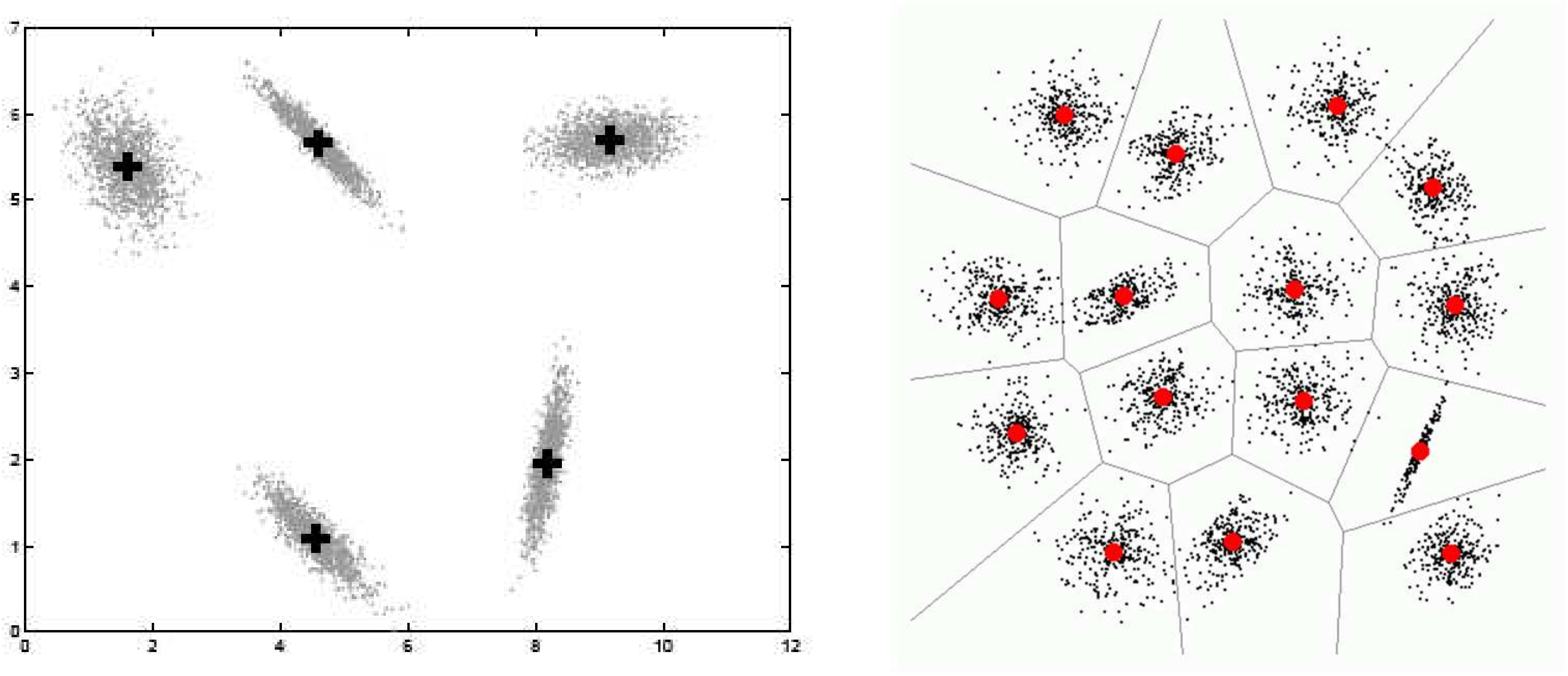
G-means behaves experimentally better than BIC/MDL criteria [NIPS*03].

G-means in action

Failure of appropriate k selection in k-means (under/overfitting):

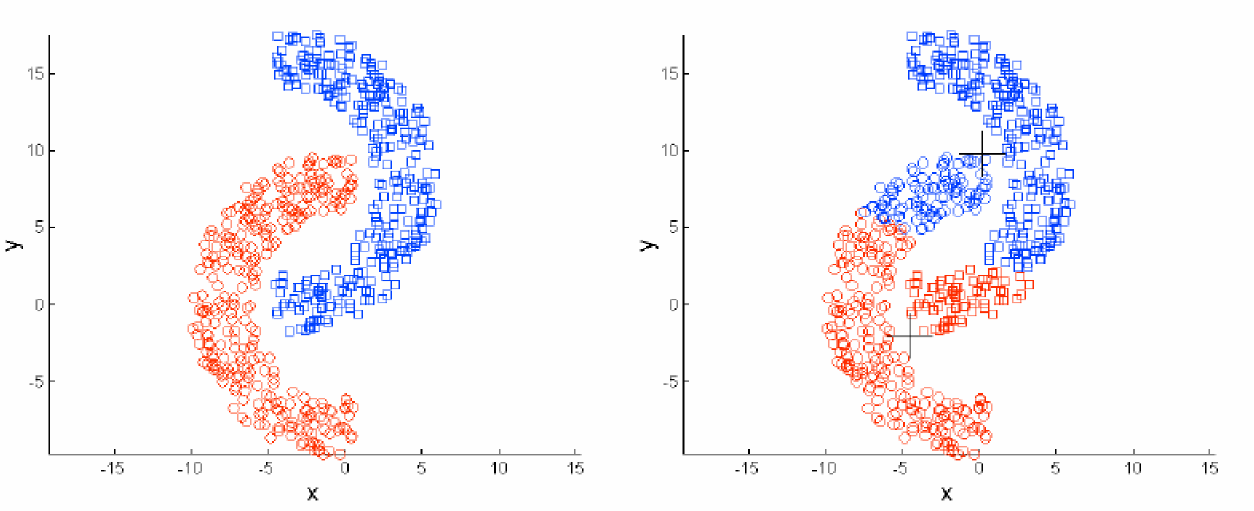


G-means learn the “right” k (assuming clusters are gaussian):

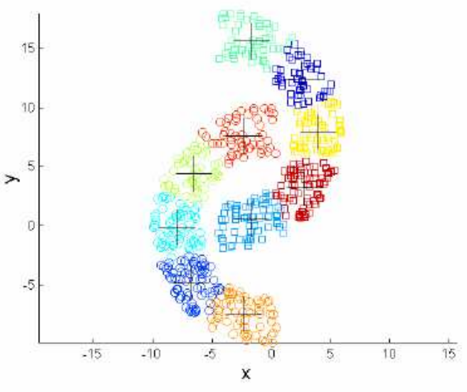


G-means++ yields not ultimate clustering!

Not all data-sets exhibit *globular* patterns...



can be bypassed by oversegmenting:



→ Better **spectral clustering** techniques...
(...that anyway finishes up clustering with k-means on eigenvectors).

Anil K. Jain , Data Clustering: 50 Years Beyond K-means (KS Fu Prize Lecture, ICPR’08)

Bibliographic references for k-means

- H. Steinhaus, Sur la division des corps matériels en parties. Bull. Acad. Polon. Sci., C1. III vol IV:801Ű804, 1956.
- S. Lloyd, Least square quantization in PCM. Bell Telephone Laboratories Paper (1957).
- D. Arthur, S. Vassilvitskii: How Slow is the k-means Method?, SoCG 2006.
- D. Arthur, S. Vassilvitskii: Worst-case and smoothed analysis of the ICP algorithm, with an application to the k-means method, FOCS 2006.
- D. Arthur, S. Vassilvitskii: k-means++ The Advantages of Careful Seeding, SODA 2007.
- R. Ostrovsky, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-Means problem, FOCS 2006.
- T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), 881-892.
- G. Hamerly, C. Elkan: Learning the k in k-means. NIPS 2003.
- G. Frahling, C. Sohler: A fast k-means implementation using coresets, SoCG 2006.

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: $D(p,q) = ||p-q||^2$ (→ 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]

Metrics, semi-distances, and divergences

Distortions: metrics $D(p,q)$, semi-distances $D(p;q)$ and divergences $D(p||q)$.

A **metric** $D(p,q) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ fulfills:

- **Non-negativity** : $D(p,q) \geq 0$.
- **Indiscernibility** : $D(p,q) = 0 \Leftrightarrow p = q$.
- **Symmetry** : $D(p,q) = D(q,p)$
- **Triangle inequality** : $D(p,q) \leq D(p,r) + D(r,q)$

  Qualitative interpretation : If both p,r and r,q are “close” so are p,q.

A **semi-distance** $D(p;q)$ may not satisfy the triangle inequality

A **divergence** $D(p||q)$ may not satisfy the symmetry nor the triangle inequality. It is only **positive definite** and satisfies the law of indiscernibility.

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.

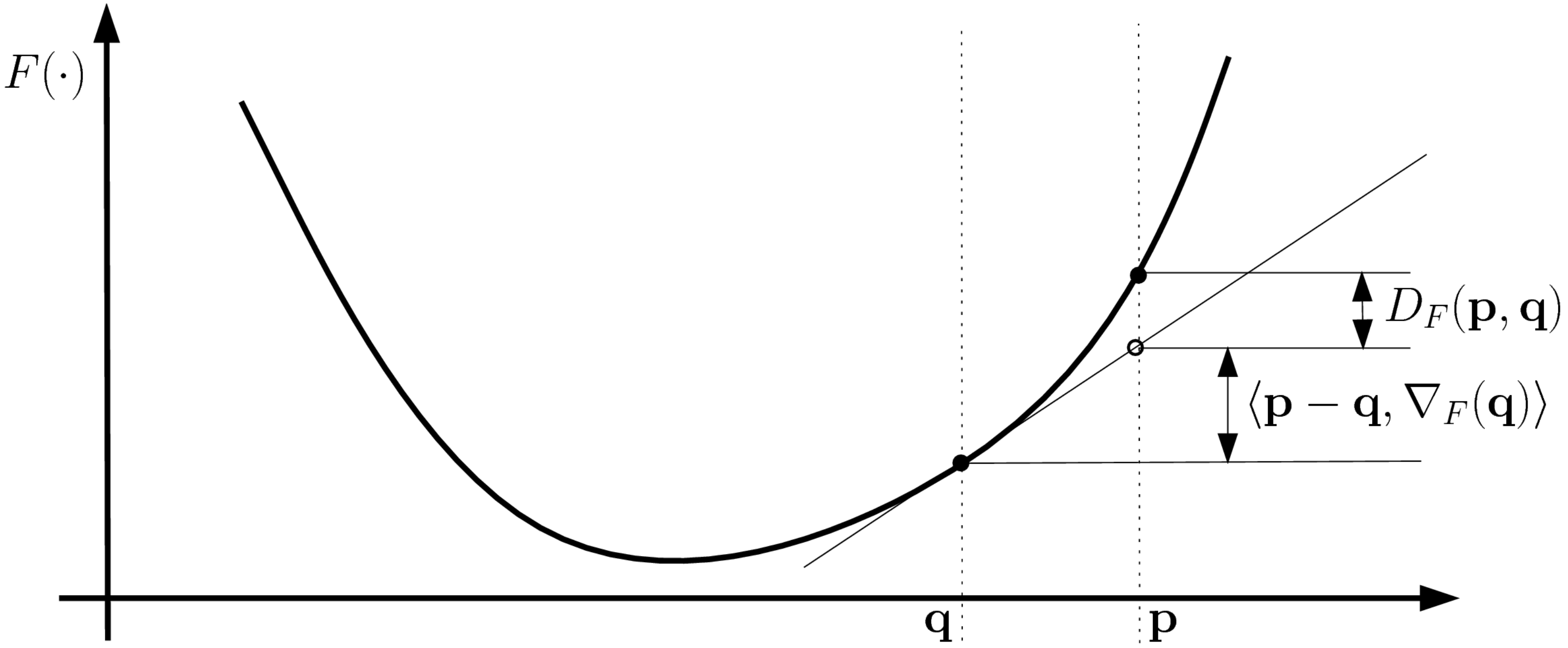
Strictly convex F: Hessian $\nabla^2 F \succ 0$ (psd.) $\longrightarrow$ ∇F is monotonous.

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

Bregman divergences: A geometric visualization

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

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

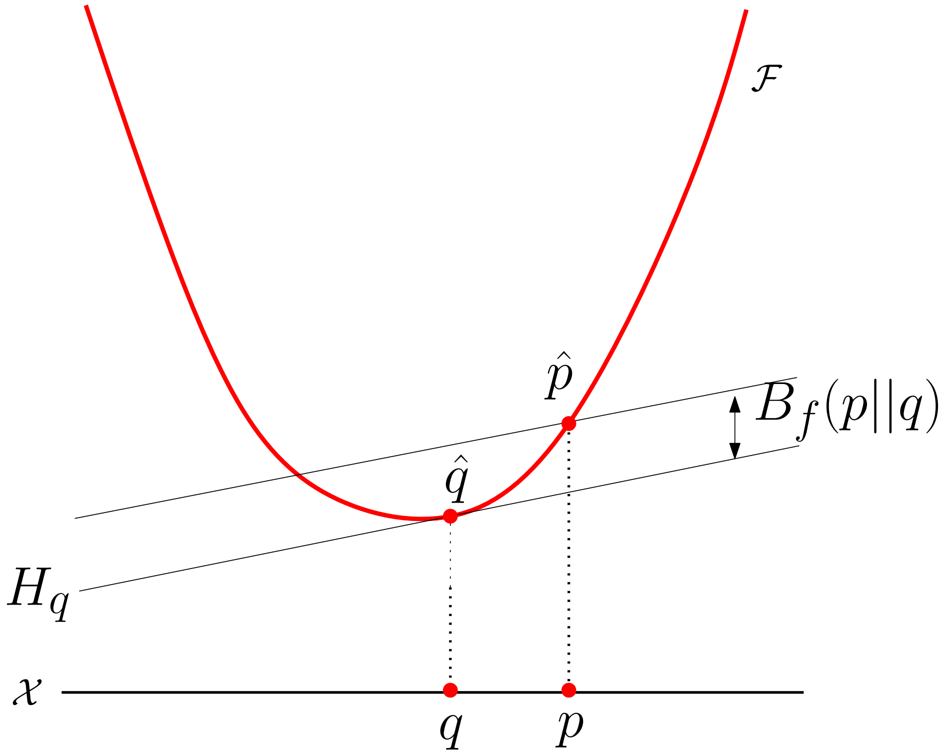

$F(\cdot)$
$D_F(\mathbf{p},\mathbf{q})$
$\langle \mathbf{p}-\mathbf{q}, \nabla_F(\mathbf{q})\rangle$
$\mathbf{q}$
$\mathbf{p}$

```
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].
Metrics defined by Bregman Divergences, Commun. Math. Sci. Volume 6, Number 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.: $H^\times(P||Q) = \sum_i p_i \log_2 \frac{1}{\tilde{p}_i} = -\sum_i p_i \log_2 \tilde{p}_i$
- Continuous rv.:
  $H^\times(P||Q) = \int p(x) \log_2 \frac{1}{\tilde{p}(x)} \mathrm{d}x = -\int_x p(x) \log_2 \tilde{p}(x) \mathrm{d}x$

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:

$$D(P||\tilde{P}) = \sum_i p_i \log_2 \frac{p_i}{\tilde{p}_i}$$

For probability density functions:

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

Many synonyms: Information discrimination, relative entropy, etc.

Kullback-Leibler divergence for probability measures

Use more general random variables than discrete/continuous rv.
Modern probability theory of **measures** : [Kolmogorov, 1933]
probability measures rather than probability mass/density distributions.
(E.g., some mixture models *are neither* continuous nor discrete.)
For a **probability measure** with **Radon-Nikodym** derivatives:

$$D(P||Q) = -\int \log \frac{\mathrm{d}Q}{\mathrm{d}P}\mathrm{d}P = \int \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}Q$$

Let μ be the **Lebesgue/counting measure** , $p \overset{\text{equal}}{=} \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q \overset{\text{equal}}{=} \frac{\mathrm{d}Q}{\mathrm{d}\mu}$:

$$D(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} \underbrace{\mathrm{d}\mu}_{\text{our former } \mathrm{d}x \text{ in pdf.}}$$

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.

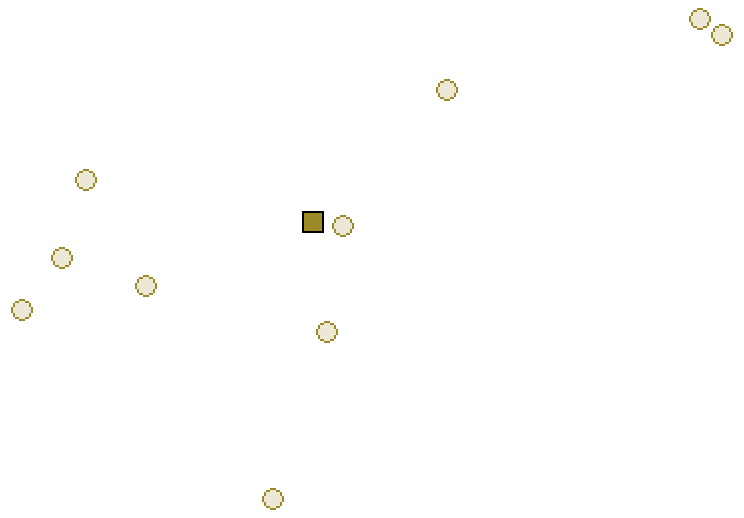
$\longrightarrow$ Bregman divergences allows us to generalize Lloyd k-means.

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** .)

Bregman information wrt. statistical random variables

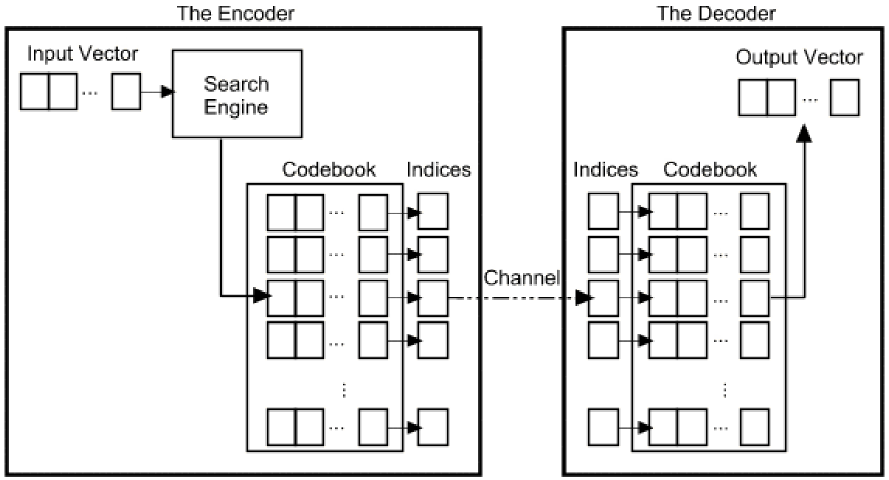
Consider $F(x) = x^T x$, $\mathcal{X} = \mathbb{R}^d$. Bregman information = **Variance**

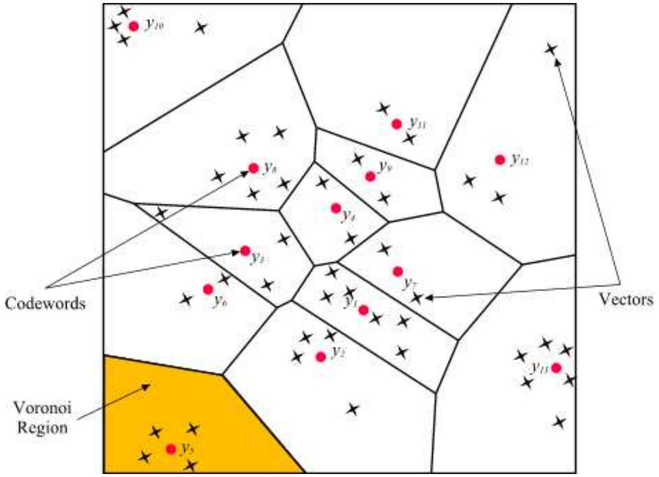
$$\begin{aligned} I_F(X) &= E_{\mathcal{X}}[B_F(X||E[X]]] \\ &= E[||X - E[X]||^2] \end{aligned}$$

(Relation to rate distortion & information bottleneck theory in [JMLR'05])

$\rightarrow$ Measure allows us to weight each sample
(distribution of weights, $\sum_i w_i = 1$ with all $w_i \geq 0$).

MinAvg cluster minimizer becomes the **barycenter** : $\bar{b} = \sum_i w_i p_i$, $\bar{b} = \int p(x) \mathrm{d}x$.





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.

→ **Meta-algorithm** .

Key question: How to choose F?

→ Many works involve generalized quadratic/Mahalanobis distances.
[ICML'07] Information-Theoretic Metric Learning
`http://videolectures.net/icml07_kulis_itml/`

Bibliographic references for k-means generalizations

- A. Banerjee, X. Guo, H. Wang: On the optimality of conditional expectation as a Bregman predictor. IEEE Transactions on Information Theory 51(7): 2664-2669 (2005)
- A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, Clustering with Bregman Divergences. 6(Oct):1705-1749, 2005.
- D. S. Modha, W. S. Spangler, Feature weighting in k-means clustering, Machine Learning, 52(3):217-237, 2003.
- Y. Linde, A. Buzo, R. Gray, An Algorithm for Vector Quantizer Design, IEEE Transactions on Communications, vol. 28, pp. 84-94, 1980.
- M. Teboulle: A Unified Continuous Optimization Framework for Center-Based Clustering Methods. Journal of Machine Learning Research 8: 65-102 (2007)

But also kernel k-means, k-means and spectral clustering, etc.:

- Weighted Graph Cuts without Eigenvectors: A Multilevel Approach, TPAMI, vol. 29:11, pp. 1944-1957, 2007.
- Kernel kmeans, Spectral Clustering and Normalized Cuts, KDD 2004.
- A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts, UTCS Technical Report TR-04-25, 2005. `http://www.cs.utexas.edu/users/dml/Software/graclus.html`

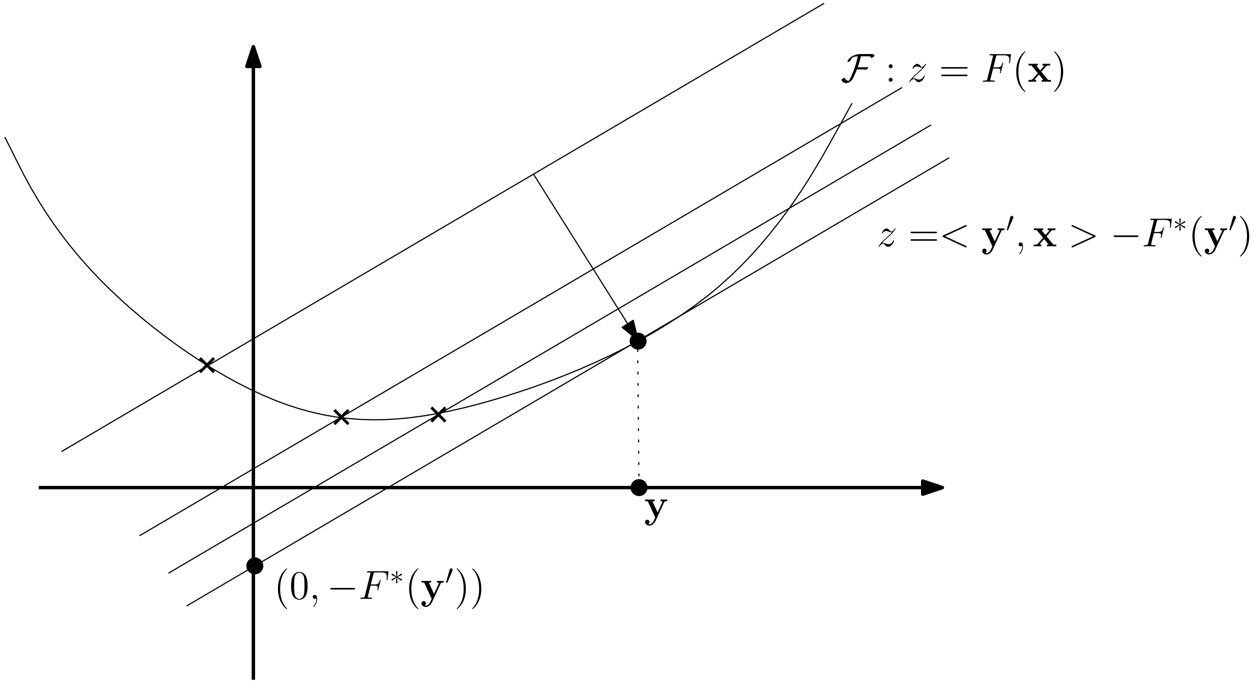
Legendre transformation: Convex conjugates

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: $(F, \mathcal{X}) \Leftrightarrow (F^*, \mathcal{X}^*)$


$\mathcal{F} : z = F(\mathbf{x})$
$z =< \mathbf{y}', \mathbf{x} > -F^*(\mathbf{y}')$
$\mathbf{y}$
$(0, -F^*(\mathbf{y}'))$

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

Approximating the functional inverse f^{-1}

For non-closed inverse formula, use Householder's iterative scheme:
For a given f, we seek f^{-1} such that

$$\forall x, (f \circ f^{-1})(x) = f(f^{-1}(x)) = x.$$

$\boxed{\text{Problem: } f^{-1}(x)?.}$

For a given (fixed) x, consider the function $g(y) = f(y) - x$. The root y^* of g yields $f^{-1}(x)$.

→ Automatic derivatives $f^{(d)}$ are always easy to get.

Iterative update :

$$y_{k+1} = y_k + (d+1)\frac{(1/f)^{(d)}(y_k)}{(1/f)^{(d+1)}(y_k)}$$

(Extend well-known Newton $d = 0$ and Halley's method $d = 1$)

Convergence of **order** $d + 1$.

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)

Convex conjugate generators

Dual divergences: Logistic loss/binary relative entropy		
$F(\theta) = \log(1 + \exp\theta)$	$B_F(\theta\vert\vert\theta') = \log\frac{1+\exp\theta}{1+\exp\theta'} - (\theta - \theta')\frac{\exp\theta'}{1+\exp\theta'}$	$f(\theta) = \frac{\exp\theta}{1+\exp\theta} = \mu$
$F^*(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$	$D_{F^*}(\mu'\vert\vert\mu) = \mu'\log\frac{\mu'}{\mu} + (1-\mu)\log\frac{1-\mu'}{1-\mu}$	$f^*(\mu) = \log\frac{\mu}{1-\mu} = \theta$

Dual divergences: Exponential loss/Unnormalized Shannon entropy		
$F(\theta) = \exp\theta$	$B_F(\theta\vert\vert\theta') = \exp\theta - \exp\theta' - (\theta - \theta')\exp\theta'$	$f(\theta) = \exp\theta = \mu$
$F^*(\mu) = \mu\log\mu - \mu$	$D_{F^*}(\mu'\vert\vert\mu) = \mu'\log\frac{\mu'}{\mu} + \mu - \mu'$	$f^*(\mu) = \log\mu = \theta$

Symmetrized Bregman divergence S_F...

Many ways to symmetrize [Chen'08]. Consider the following one:

$$\begin{aligned} S_F(p;q) &= \frac{1}{2}(B_F(p||q)+B_F(q||p)) \\ &= \frac{1}{2}(B_F(p||q)+B_{F^*}(\nabla F(p)||\nabla F(q))) \end{aligned}$$

...is not a Bregman divergence except for squared Mahalanobis/quadratic distances.

Lifting:
$p \mapsto \tilde{p} = (p, \nabla F(p))$, and $\tilde{F} = (\frac{1}{2}F, \frac{1}{2}F^*)$ then
$B_{\tilde{F}}(p||q) = B_{\tilde{F}}(q||p) = S_F(p;q)$.
But for $\mathcal{X}, \mathcal{X}^* \subset \mathbb{R}^d$ convex domains, $\tilde{\mathcal{X}} = \{(x, \nabla F(x) \mid x \in \mathcal{X}\}$ is usually **not convex** in $\mathbb{R}^{2d}$.
($\tilde{\mathcal{X}}$ is a d-dimensional hypersurface in $2d$-dimensional space)
(Eg., take $f(x) = -\log x$, $(x, -1/x)$ is not convex in $\mathbb{R}_*^+ \times \mathbb{R}_*^-$)

Weighted coupled Bregman k-means

Consider the following **potential function** handling *asymmetric* Bregman divergences:

$$\Delta_{F,\alpha}(c^*||x||c) = (1-\alpha)B_F(c^*||x) + \alpha B_F(x||c)$$

- $\alpha = 0$: Left-sided Bregman divergence
- $\alpha = 1$: Right-sided Bregman divergence
- $\alpha = \frac{1}{2}$: Symmetrized Bregman divergence for $c^* = c$.

Mixed Bregman Clustering with Approximation Guarantees [ECML'08]

Weighted coupled Bregman k-means

For a cluster $\mathcal{C}$, the **optimal pair** (a^*, b^*) of centers minimizing the potential:

$$L_{F,\alpha}(\mathcal{C}) = \sum_{x\in\mathcal{C}} \Delta_{F,\alpha}(a||x||b) = (1-\alpha)\sum_{x\in\mathcal{C}} B_F(a||x) + \alpha \sum_{x\in\mathcal{C}} B_F(x||b)$$

is

- Right-centered centroid:

$$b^* = \frac{1}{|\mathcal{C}|}\sum_{p\in\mathcal{C}} p$$

- Left-centered centroid:

$$a^* = \nabla F^{-1}\left(\frac{1}{|\mathcal{C}|}\sum_{p\in\mathcal{C}} \nabla F(p)\right)$$

Weighted coupled Bregman k-means clustering

Loss/potential function optimization for k-clustering:

- Primal potential function:

$$L_{F,\alpha}(\mathcal{P},\mathcal{C}) = \sum_{p\in\mathcal{P}} \min_{(a,b)\in\mathcal{C}} \Delta_{F,\alpha}(a||x||b)$$

  Coupled k-means clustering:

$$L^*_{F,\alpha} = \min_{\mathcal{C}\in\mathcal{X}^2,|\mathcal{C}|=k} L_{F,\alpha}(\mathcal{P},\mathcal{C})$$

  ...is dually equivalent too...

- Dual potential function:

$$L_{F,\alpha}(\mathcal{P},\mathcal{C}) = \sum_{p\in\mathcal{P}} \min_{(a,b)\in\mathcal{C}} \Delta_{F,1-\alpha}(b||x||a)$$

  (notice that pairs of centroids are swapped and $\alpha \to 1-\alpha$)

D^2-initialization for coupled Bregman clustering

Compute k (pairs of) center clusters:

- Let $\mathcal{C} \leftarrow \{x, x\}$ for $x \in \mathcal{S}$
- Repeat $k-1$ times:
  Pick point $x \in \mathcal{S}$ with probability:

$$\frac{\Delta_{F,\alpha}(c_x||x||c_x)}{\sum_{y\in\mathcal{P}} \Delta_{F,\alpha}(c_y||y||c_y)},$$

where $(c_x, c_x) = \arg\min \Delta_{F,\alpha}(z||x||z)$

$\longrightarrow$ **α-seeding** with guaranteed approximation factor depending on the Hessian *"spread"* ρ_F:

$$L^*_{F,\alpha} \leq L_{F,\alpha} \leq 8{\rho_F}^2(2+\log k)L^*_{F,\alpha}$$

$$\rho_F = \sup_{u,v,s,t\in\mathcal{X}} \frac{(u-v)^T \nabla F^2(s)(u-v)}{(u-v)^T \nabla F^2(t)(u-v)} < \infty$$

...on compact domains $\mathcal{X}$ (finite point sets). $\rho_{F(x)=x^Tx} = 1$.

($\longrightarrow$ Avoid to compute explicitly symmetrized Bregman centroids)

[Nielsen'07] On the Centroids of Symmetrized Bregman Divergences, arXiv:0711.3242.

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 $p \rightarrow \pm\infty$

Bijection: Bregman divergences B_F and ∇F-means

$$\boxed{\text{Bregman divergence } B_F \leftrightarrow \nabla F\text{-means}}$$

Equivalence classes of generalized means :
$M(\mathcal{S}; f) = M(\mathcal{S}; af + b)\ \forall a \in \mathbb{R}^+_*$ and $\forall b \in \mathbb{R}$
Equivalence classes of Bregman divergences :
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 $B_F(p||q) = B_G(p||q)$.

(Bregman divergences defined for generators modulo affine terms $ax + b$)

Furthermore, Bregman divergences are a **linear operator** :

$$B_{F_1+\lambda F_2}(p||q) = B_{F_1}(p||q) + \lambda B_{F_2}(p||q) \forall \lambda \in \mathbb{R}^+$$

($F_1 + \lambda F_2$ is a strictly convex and differentiable function)

Bijection: Bregman divergences B_F and ∇F-means

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 $\frac{n}{\sum_{j=1}^{n} \frac{1}{x_j}}$

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)

Bregman soft clustering
— Expectation-maximization made easy —

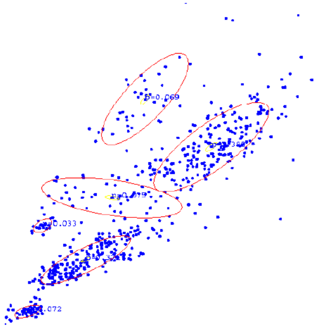
Soft clustering & EM algorithm

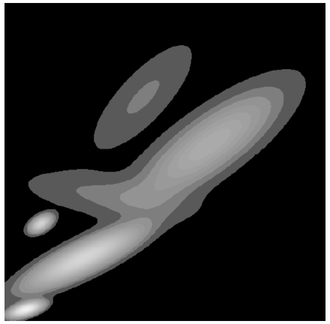
Soft clustering: each point belongs to all clusters according to a weight distribution (=density). → statistical modeling

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

$\rightarrow$ 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)

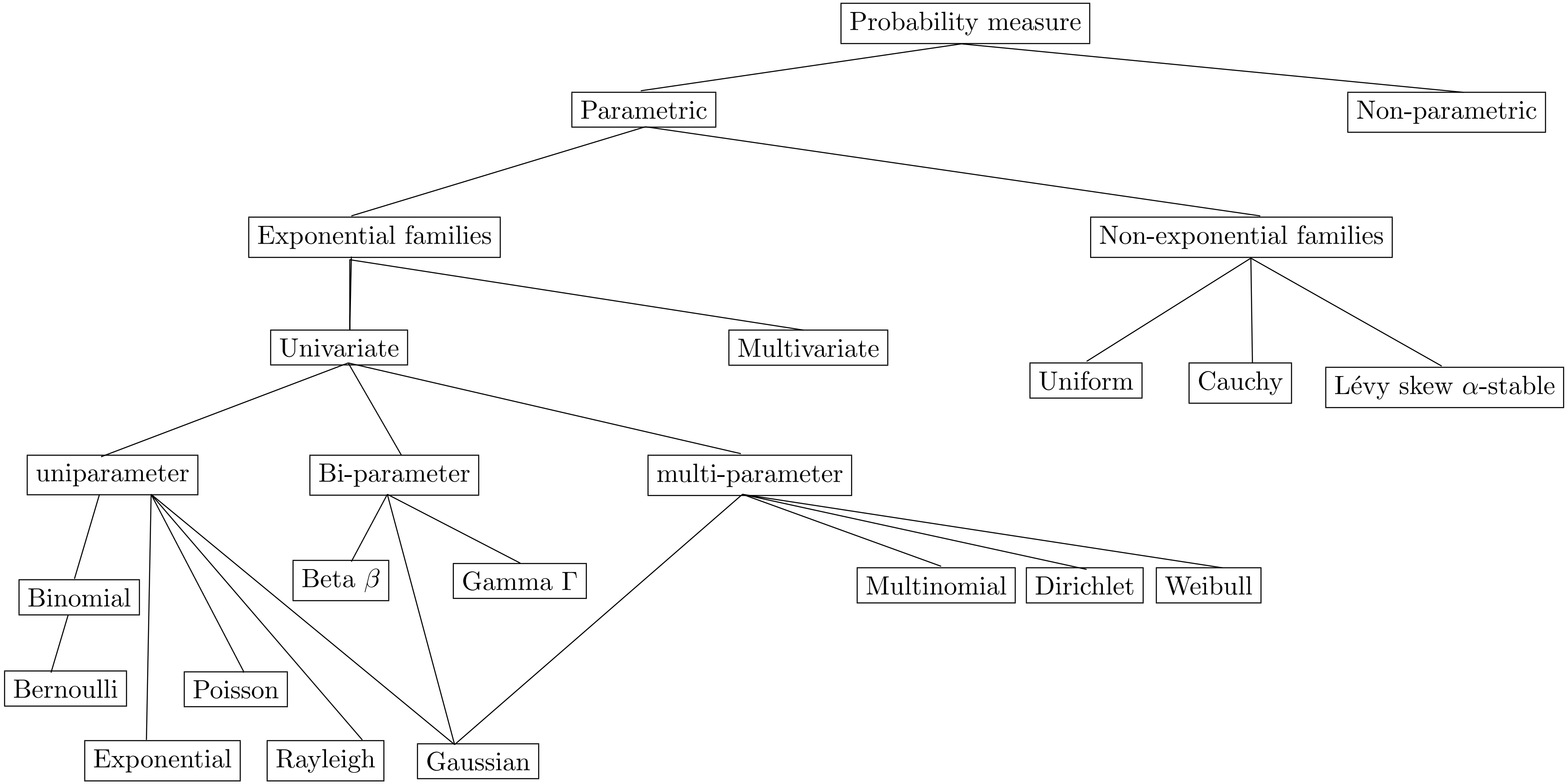
Log normalizer : From $\int_{x\in\mathcal{X}} p_F(x|\theta)\mathrm{d}x = 1$

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

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:

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

- Variance:

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

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

Finite expectations & variances (proof)

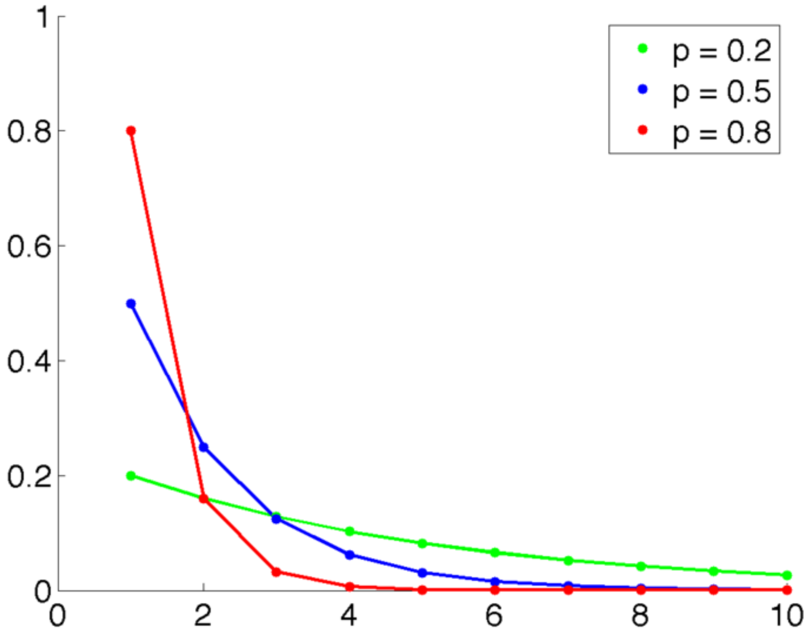
$$\begin{aligned}
\nabla_\theta \int \exp(\langle \theta, t(x)\rangle - F(\theta) + C(x))\mathrm{d}x &= \nabla_\theta 1\\
\int \exp(\langle \theta, t(x)\rangle - F(\theta) + C(x))(t(x) - \nabla F(\theta))\mathrm{d}x &= 0\\
\Rightarrow E[t(X)] - \nabla_\theta F(\theta) &= 0
\end{aligned}$$

$$\begin{aligned}
\nabla^2_\theta \int \exp(\langle \theta, t(x)\rangle - F(\theta) + C(x))\mathrm{d}x &= \nabla^2_\theta 1\\
\int p(x;\theta)\left((t(x) - \nabla_\theta F(\theta))(t(x) - \nabla_\theta F(\theta))^T - \nabla^2_\theta F(\theta)\right)\mathrm{d}x &= 0\\
\mathcal{E}[(t(X) - \nabla_\theta F(\theta))(t(X) - \nabla_F \theta(\theta))^T] - \nabla^2_\theta F(\theta) &= 0\\
\Rightarrow \mathrm{Cov}[t(X)t(X)^T] - \nabla^2_\theta F(\theta) &= 0
\end{aligned}$$

Example of exponential families: Geometric distribution

X: # of Bernoulli trials to get **one success** (p: success probability)

$$\Pr(X = k) = (1-p)^{k-1}p$$



with

$$E[X] = \frac{1}{p}$$

$$\text{var}[X] = \frac{1-p}{p^2}$$

Geometric distribution: An exponential family

$$\begin{aligned} \Pr(X=k) &= (1-p)^{k-1}p \\ &= \exp^{\langle t(x),\theta\rangle - F(\theta) + k(x)} \end{aligned}$$

Natural parameter	$\theta = \log(1-p)$
Log normalizer	$F(\theta) = \log \frac{\exp^\theta}{1-\exp^\theta}$
Sufficient statistic	$t(x) = x$
Carrier measure	$k(x) = 0$ (counting)
Expectation	$\mu = \nabla F(\theta) = 1 + \frac{e^\theta}{1-e^\theta} \equiv \frac{1}{p}$
Variance	$\sigma^2 = \nabla^2 F(\theta) = \frac{e^\theta}{(1-e^\theta)^2} \equiv \frac{1-p}{p^2}$

Example of exponential families: Gaussian distributions

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

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

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: $\tilde{x} = (x, -\frac{1}{2}xx^T)$ (→ 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}.)

MLE of exp. fam.: Centroid of sufficient statistics

$$\Pr(X = x_i) = \exp(\langle \theta, t(x_i) \rangle - F(\theta) + C(x_i))$$

Log-likelihood:

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

Gradient wrt. to $\theta \Longrightarrow \sum_{i=1}^{n} t(x_i) - n\nabla F(\theta) = 0$.

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

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

Global minimum: Derivating twice the log-likelihood function: $-n\nabla^2 F(\theta) \leq 0$.

Since F is convex, log-likelihood is concave ⇒ **unique global minimizer**

Examples of MLEs of exponential families

Probability distribution	Maximum likelihood estimator $\hat{\theta}$
Exponential family $E_F(\theta)$	$\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$ $\hat{\theta} = \nabla^{-1}F(\frac{1}{n}\sum_{i=1}^{n} t(x_i))$
Closed form solutions (easy cases)	
Bernoulli q	$\hat{q} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Exponential λ	$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$
Poisson λ	$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Rayleigh σ	$\hat{\sigma} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}$
Normal $\mu, \sigma)$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$
Lognormal (μ, σ)	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$ $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\log x_i - \hat{\mu})^2}$
Non-analytic cases (hard cases)	
Gamma (α, β)	$\alpha\beta = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\log x_i$
Beta (α, β)	$\frac{\partial B(\alpha,\beta)}{\partial\alpha} = B(\alpha,\beta)\sum_{i=1}^{n}\log x_i$ $\frac{\partial B(\alpha,\beta)}{\partial\beta} = B(\alpha,\beta)\sum_{i=1}^{n}\log 1 - x_i$

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):

$$\boxed{\mathrm{KL}(p_F(x;\theta_1)||p_F(x;\theta_2)) = B_F(\theta_2||\theta_1)}$$

$$\begin{aligned}
\mathrm{KL}(\theta_p||\theta_q) &= \int_x p(x|\theta_p)\log\frac{p(x|\theta_p)}{p(x|\theta_q)}\mathrm{d}x \\
&= \int_x p(x|\theta_p)(F(\theta_q)-F(\theta_p)+\langle\theta_p-\theta_q, t(x)\rangle)\mathrm{d}x \\
&= \int_x p(x|\theta_p)\left(B_F(\theta_q||\theta_p)+\langle\theta_q-\theta_p,\nabla F(\theta_p)\rangle+\langle\theta_p-\theta_q,t(x)\rangle\right)\mathrm{d}x \\
&= B_F(\theta_q||\theta_p)+\int_x p(x|\theta_p)\langle\theta_q-\theta_p,\nabla F(\theta_p)-t(x)\rangle)\mathrm{d}x \\
&= B_F(\theta_q||\theta_p)-\int_x p(x|\theta_p)\langle\theta_q-\theta_p,t(x)\rangle\mathrm{d}x+\langle\theta_q-\theta_p,\nabla F(\theta_p)\rangle \\
&\overset{(\mathrm{Eq.}\ *)}{=} B_F(\theta_q||\theta_p)
\end{aligned}$$

since $\nabla F(\theta) = \left[\int_x t(x)\exp\{\langle\theta,t(x)\rangle - F(\theta)+C(x)\}\mathrm{d}x\right]$ (*)

KL of multivariate normals

Easy to recover KL divergence of exponential families (only need ∇F):

$$\boxed{\begin{array}{l}\mathrm{KL}(p(x;\mu_i,\Sigma_i)||p(x;\mu_j,\Sigma_j)) = \frac{1}{2}\log|\Sigma_i^{-1}\Sigma_j| + \\ \frac{1}{2}\mathrm{Tr}\left((\Sigma_i^{-1}\Sigma_j)^{-1}\right) - \frac{d}{2} + \frac{1}{2}(\mu_i-\mu_j)^T\Sigma_j^{-1}(\mu_i-\mu_j)\end{array}}$$

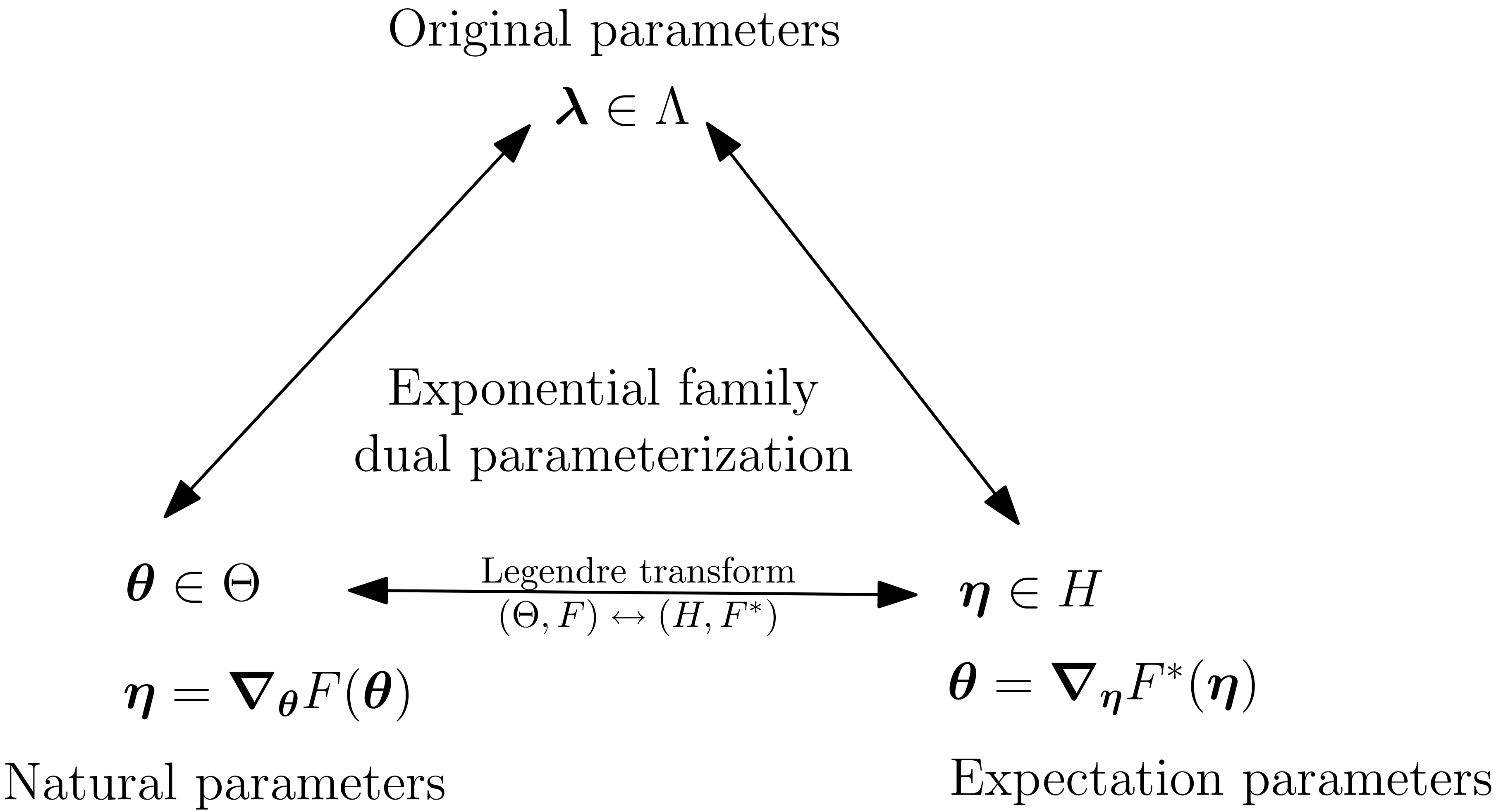
$$\mathrm{KL}(p(x;\mu_i,\Sigma_i)||p(x;\mu_j,\Sigma_j)) = B_F(\tilde{\Theta}_j||\tilde{\Theta}_i)$$

$$B_F(p||q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle$$

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

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

Parameterization of exponential families



Original parameters

$\boldsymbol{\lambda} \in \Lambda$

Exponential family dual parameterization

$\boldsymbol{\theta} \in \Theta$

Legendre transform

$(\Theta, F) \leftrightarrow (H, F^*)$

$\boldsymbol{\eta} \in H$

$\boldsymbol{\eta} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$

$\boldsymbol{\theta} = \boldsymbol{\nabla}_{\boldsymbol{\eta}} F^*(\boldsymbol{\eta})$

Natural parameters

Expectation parameters

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

Bijection: Exponential families $\Leftrightarrow$ Bregman divergences

Regular exponential family $\Longleftrightarrow$ regular Bregman divergence:

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

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

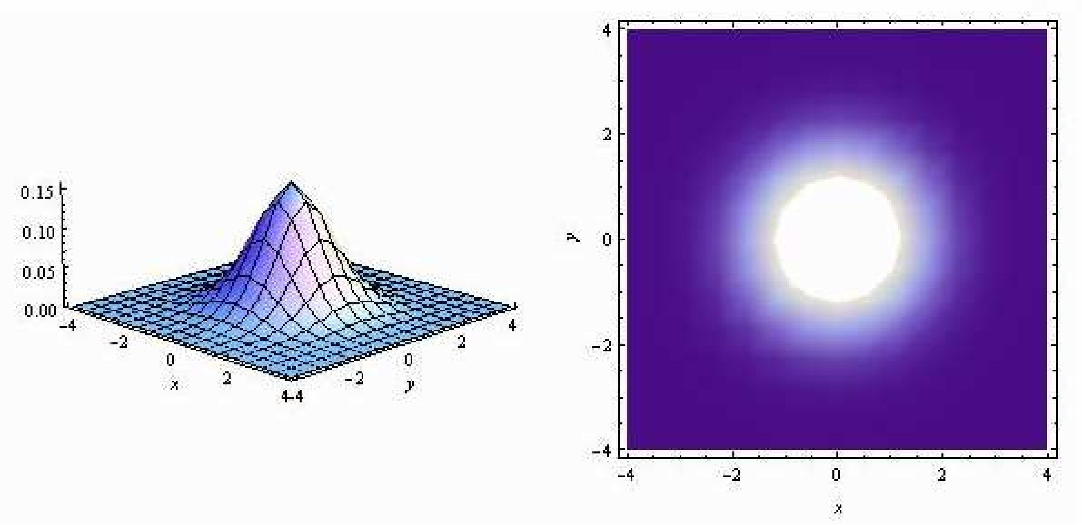
Note: For some generators F, the Legendre dual F^* may not have closed-form solutions. Use Householder formula to approximate the $\nabla F(\theta)$ for a given θ (root finding).

[JMLR'05] Clustering with Bregman divergences.

Visualizing the density/distance bijection

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

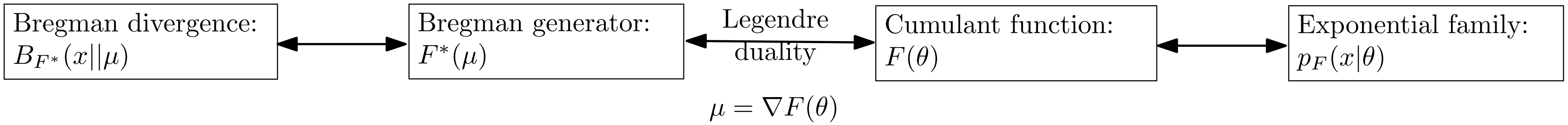
The **expectation** is $\int p_F(x;\theta)\mathrm{d}x = \nabla F(\theta) \overset{\text{equal}}{=} \mu$.
The **iso-density** is $p_F(x;\theta) = \lambda \Longleftrightarrow c_{F^*}(x) \exp -B_{F^*}(x||\nabla F(\theta)) = \lambda$.
Bregman divergences are **always convex** in first argument.
$\mu = \nabla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

Bregman generator: $F^*(\mu)$

Legendre duality

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

Cumulant function: $F(\theta)$

Exponential family: $p_F(x|\theta)$

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

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

Examples: Exponential families $\Leftrightarrow$ Bregman divergences

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

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

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

Experimental results of Bregman clustering

- **Text documents** modeled using **multinomial distributions** (∈ exponential family):

corresponding Bregman divergence = KL-divergence between word distributions

- Speech power spectra follow exponential family densities: $\Pr(x) = \lambda \exp -\lambda x$:

  corresponding Bregman divergence = Itakura-Saito divergence

- Experiments: Datasets drawn from know exponential family distributions using the corresponding Bregman divergence as well as non-matching divergences.

Evaluation using normalized mutual information (NMI)

Mutual information :

$$\mathrm{MI}(P,Q) = H(P) + H(Q) - H(P,Q),$$

with **joint entropy** :

$$\max\{H(P),H(Q)\} \leq H(P,Q) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\log_2 p(x,y)\mathrm{d}x\mathrm{d}y \leq H(P){+}H(Q)$$

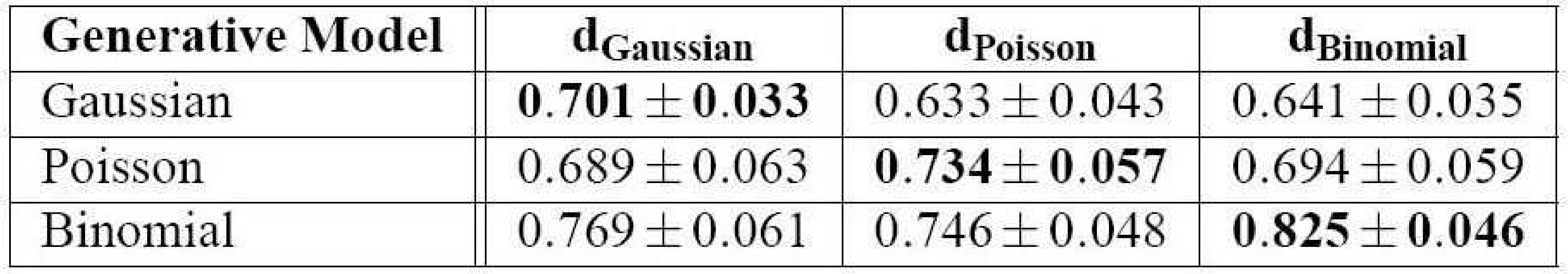
Normalized mutual information :

$$\mathrm{NMI}(P,Q) = \frac{H(P)+H(Q)}{H(P,Q)}$$

NMI between **predicted** and **original** clusters (based on the ground-truth mixture components).

Appropriate Bregman divergence?

- Characterize data generative process using a mixture of exponential family distributions.
  → use corresponding Bregman divergence...
- ...but this is just a rule of thumb... The divergence should capture the similarity properties desirable in **applications** , and not necessarily depend on how the data was actually generated.



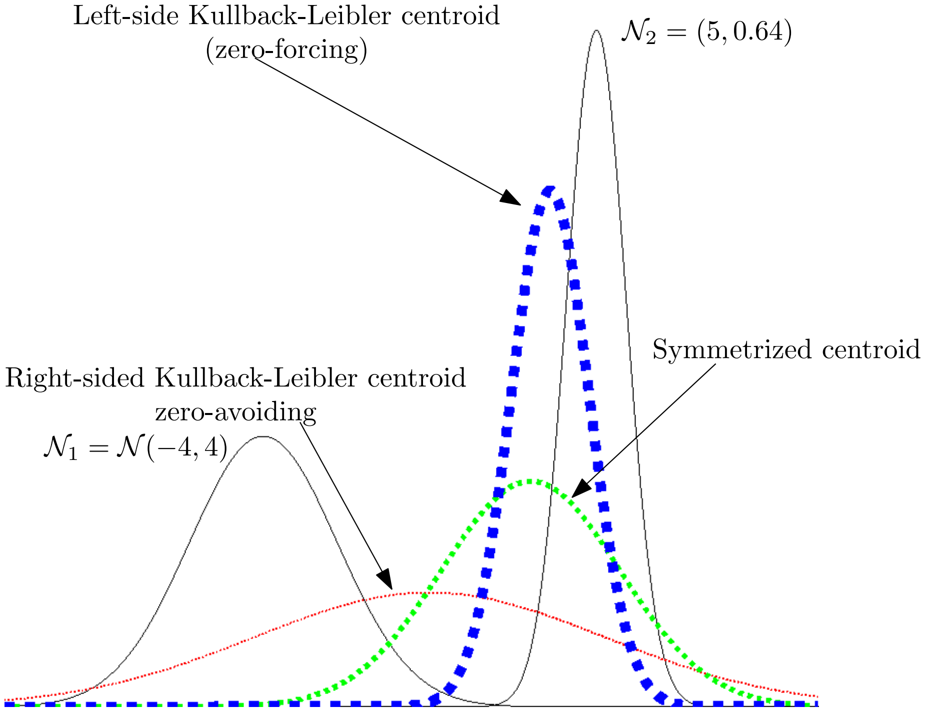
Generative Model	$\mathbf{d}_{\textbf{Gaussian}}$	$\mathbf{d}_{\textbf{Poisson}}$	$\mathbf{d}_{\textbf{Binomial}}$
Gaussian	**0.701 ± 0.033**	0.633 ± 0.043	0.641 ± 0.035
Poisson	0.689 ± 0.063	**0.734 ± 0.057**	0.694 ± 0.059
Binomial	0.769 ± 0.061	0.746 ± 0.048	**0.825 ± 0.046**

[JMLR'05] Clustering with Bregman divergences.

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).

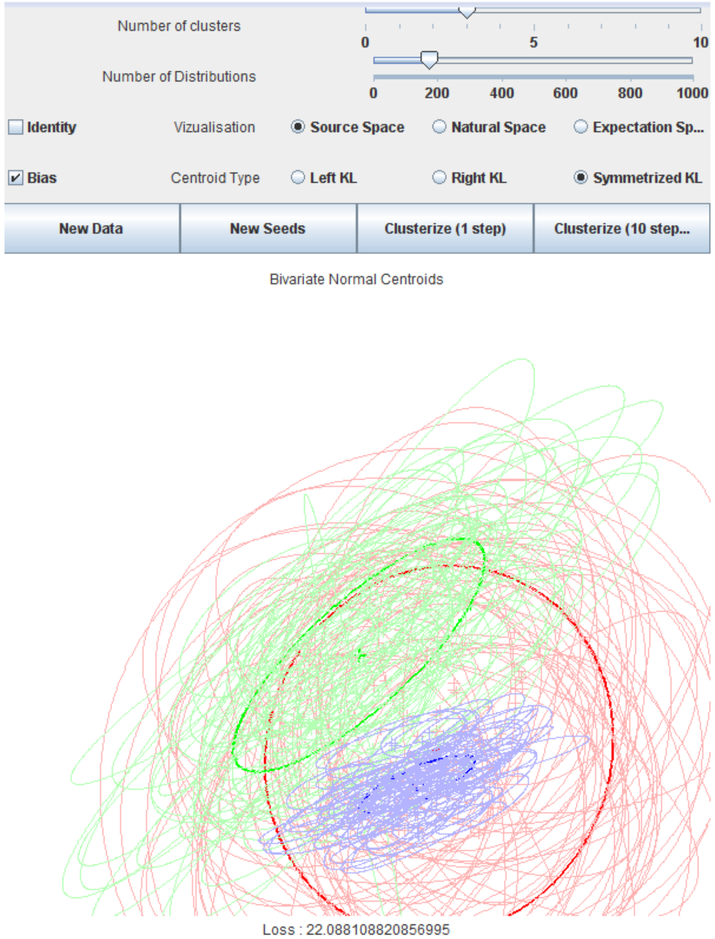


Summary of keypoints

- **Bregman divergences** : $B_F(p||q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle$
- unifies quadratic distances ($F = x^T x$) with relative entropy ($F(x) = \sum_i x_i \log x_i$)
- **Bregman k-means++** is most general Lloyd's k-means++ generalization (with centroid relocation)
- Dual **convex conjugates** F, F^* (from Legendre transformation: $\nabla F^* = (\nabla F)^{-1}$
- Dual Bregman divergences $B_F(p||q) = B_{F^*}(\nabla F(q)||\nabla F(p))\ \forall (p,q) \in \mathcal{X} \times \mathcal{X}$
- **Exponential families** have probability measures (Gaussian, multinomial) $p_F(x|\theta) = \exp\left(\langle t(x), \theta \rangle - F(\theta) + k(x)\right)$
- **Kullback-Leibler divergence** $\int p(x) \log \frac{p(x)}{q(x)}$ for exponential families: $\mathrm{KL}(p_F(x;\theta_1)||p_F(x;\theta_2)) = B_F(\theta_2||\theta_1)$
- **Bijection exponential families** $\leftrightarrow$ Bregman divergences: $\log p_F(x;\theta) = -B_{F^*}(x||\nabla F(\theta)) + \log c_{F^*}(x)$ (Expectation maximization made easy: Bregman soft clustering)

Bregman k-means: Simplifying GMMs (Quantization)

Applications: Simply bivariate Gaussian mixture models ($n \rightarrow k$ components, $d = 2 + 3 = 5$)
$\rightarrow$ k-means wrt. Kullback-Leibler or SKL in natural parameter space
Loss function=Minimize the Bregman information - Bregman information of the centers
(for KL, Bregman information related to mutual information)



Number of clusters 0 5 10

Number of Distributions 0 200 400 600 800 1000

□ Identity Vizualisation ● Source Space ○ Natural Space ○ Expectation Sp...

☑ Bias Centroid Type ○ Left KL ○ Right KL ● Symmetrized KL

Bivariate Normal Centroids

Loss : 22.088108820856995

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

Properties of Bregman divergences

Bregman divergences: A reminder

Bregman divergences B_F defined on **types** : vector spaces, matrix spaces, functional spaces, etc.

- Column **vectors**

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

($F(x) = x^T x \rightarrow L_2^2$, $F(x) = x \log x \rightarrow$ Kullback-Leibler (KL), $F(x) = -\log x \rightarrow$ Itakura-Saito divergence, etc.)

- Psd. **matrices** (eg., density Hermitian matrices)

$$\begin{aligned} B_F(P||Q) &= F(P) - F(Q) - \langle P - Q, \nabla F(Q) \rangle \\ &= F(P) - F(Q) - \mathrm{Tr}((P-Q) \nabla F(Q)^T) \end{aligned}$$

($F(X) = X \log X \rightarrow$ quantum relative entropy, $F(X) = -\log \det X \rightarrow$ LogDet/Stein distances, etc.)

Bregman generators F defined up to an *affine* factor $ax + b$.

Homogeneous Bregman divergences

Homogeneity condition of degree α.
For $\lambda > 0$,

$$B_F(\lambda p || \lambda q) = \lambda^\alpha B_F(p||q)$$

Three remarkable Bregman divergences:

α	$F(x)$	Bregman divergence
$\alpha = 2$	$F(x) = x^2$	square Euclidean distance
$\alpha = 1$	$F(x) = x \log x$	Kullback-Leibler, I-divergence
$\alpha = 0$	$F(x) = -\log x$	Itakura-Saito divergence

Modulo **affine terms** ($B_{F+ax+b} = B_F$).

Smooth family B_{F_α} of **power Bregman divergences**
(from Burg to negative Shannon entropy).

Bregman divergence as exact Taylor remainder

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=0.02720232223423058,
vareps=0.5301484973611689, vareps belongs to [p,q]
```

Since ∇F is monotonously increasing, easy to get upper/lower bounds on $B_F(p||q)$.

Bregman dual bisectors: Hyperplane/hypersurface

Right-sided bisector: → 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: → Hypersurface

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

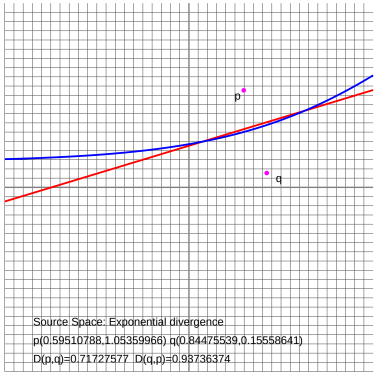
$$H'_F : \langle \nabla F(x), q - p\rangle + F(p) - F(q) = 0$$

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

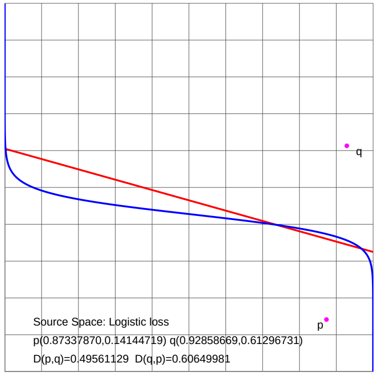
Visualizing Bregman bisectors

Primal coordinates θ

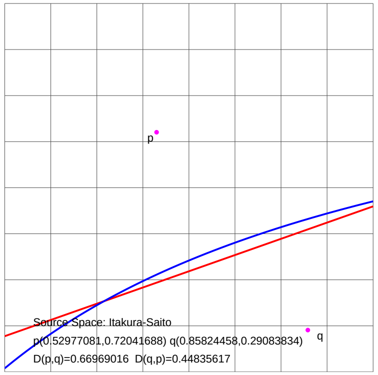
natural parameters



Source Space: Exponential divergence
p(0.59510788,1.05359966) q(0.84475539,0.15558641)
D(p,q)=0.71727577 D(q,p)=0.93736374



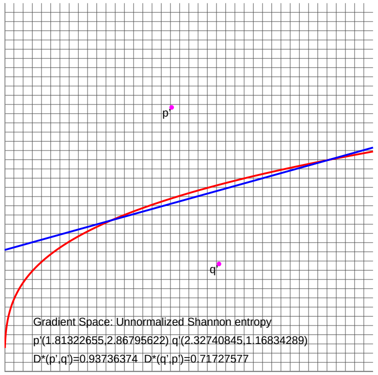
Source Space: Logistic loss
p(0.87337870,0.14144719) q(0.92858669,0.61296731)
D(p,q)=0.49561129 D(q,p)=0.60649981



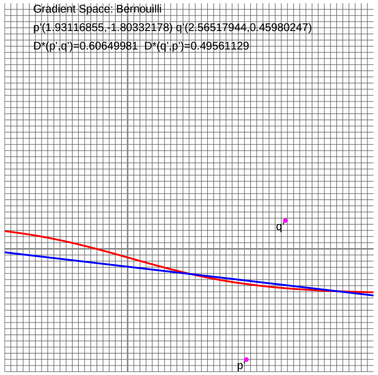
Source Space: Itakura-Saito
p(0.52977081,0.72041688) q(0.85824458,0.29083834)
D(p,q)=0.66969016 D(q,p)=0.44835617

Dual coordinates η

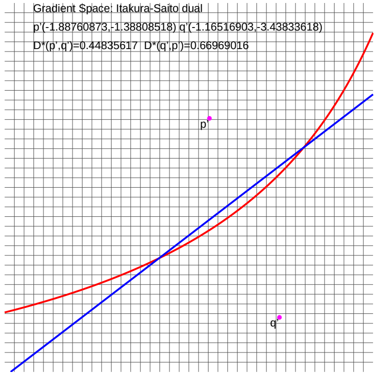
expectation parameters



Gradient Space: Unnormalized Shannon entropy
p'(1.81322655,2.86795622) q'(2.32740845,1.16834289)
D*(p',q')=0.93736374 D*(q',p')=0.71727577



Gradient Space: Bernouilli
p'(1.93116855,-1.80332178) q'(2.56517944,0.45980247)
D*(p',q')=0.60649981 D*(q',p')=0.49561129



Gradient Space: Itakura-Saito dual
p'(-1.88760873,-1.38808518) q'(-1.16516903,-3.43833618)
D*(p',q')=0.44835617 D*(q',p')=0.66969016

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:

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

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

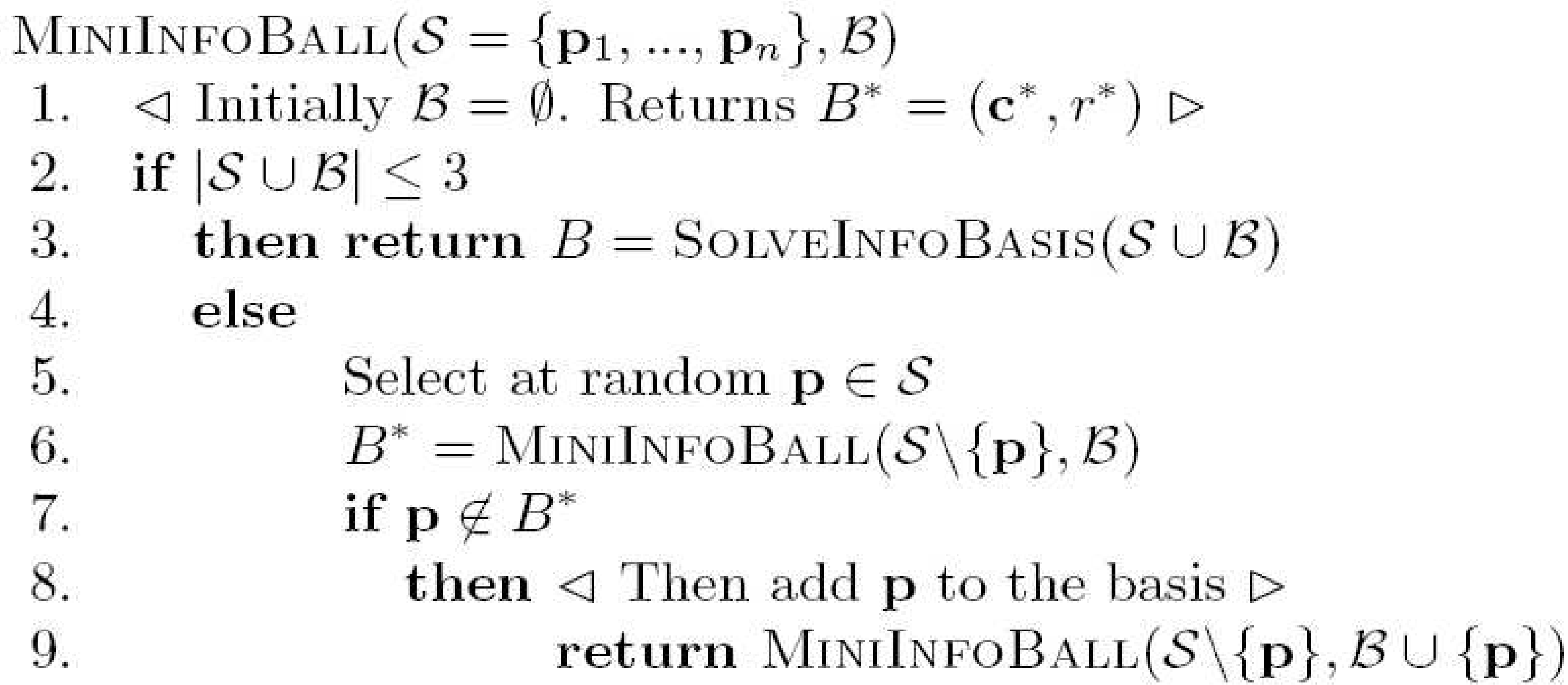
Fit the **LP-type** framework:

- **Monotonicity** : For any $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} \subseteq \mathcal{G} \subset \mathcal{X}$, $r^*(\mathcal{F}) \leq r^*(\mathcal{G})$
- **Locality** : For any $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} \subseteq \mathcal{G} \subset \mathcal{X}$ with $r^*(\mathcal{F}) = r^*(\mathcal{G})$ and any point $p \in \mathcal{X}$: $r^*(\mathcal{G}) < r^*(\mathcal{G} \cup \{p\}) \rightarrow r^*(\mathcal{F}) < r^*(\mathcal{F} \cup \{p\})$

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

Bregman MiniBall (extending Welzl's miniball)

Technicalities: $\mathcal{X}$ is an **open convex set** (→ replace $\min\max$ by $\inf\sup$).
Combinatorial basis ranging from 2 to $d+1$.
(need efficient procedures for solving recursion's terminal cases)

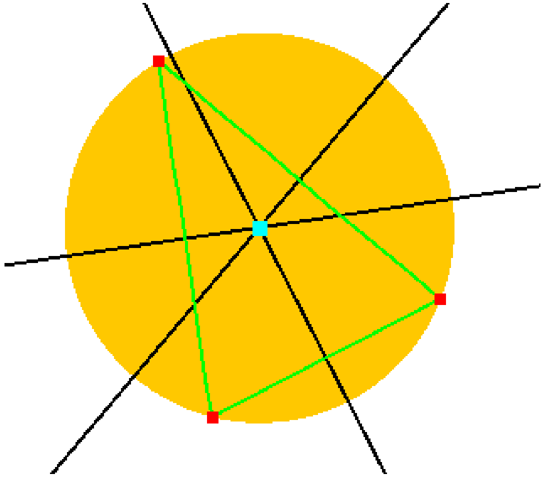


```
MiniInfoBall($\mathcal{S} = \{\mathbf{p}_1, ..., \mathbf{p}_n\}, \mathcal{B}$)
    ◁ Initially $\mathcal{B} = \emptyset$. Returns $B^* = (\mathbf{c}^*, r^*)$ ▷
    if $|\mathcal{S} \cup \mathcal{B}| \leq 3$
        then return $B = $ SolveInfoBasis($\mathcal{S} \cup \mathcal{B}$)
        else
            Select at random $\mathbf{p} \in \mathcal{S}$
            $B^* = $ MiniInfoBall($\mathcal{S} \backslash \{\mathbf{p}\}, \mathcal{B}$)
            if $\mathbf{p} \notin B^*$
                then ◁ Then add $\mathbf{p}$ to the basis ▷
                    return MiniInfoBall($\mathcal{S} \backslash \{\mathbf{p}\}, \mathcal{B} \cup \{\mathbf{p}\}$)
```

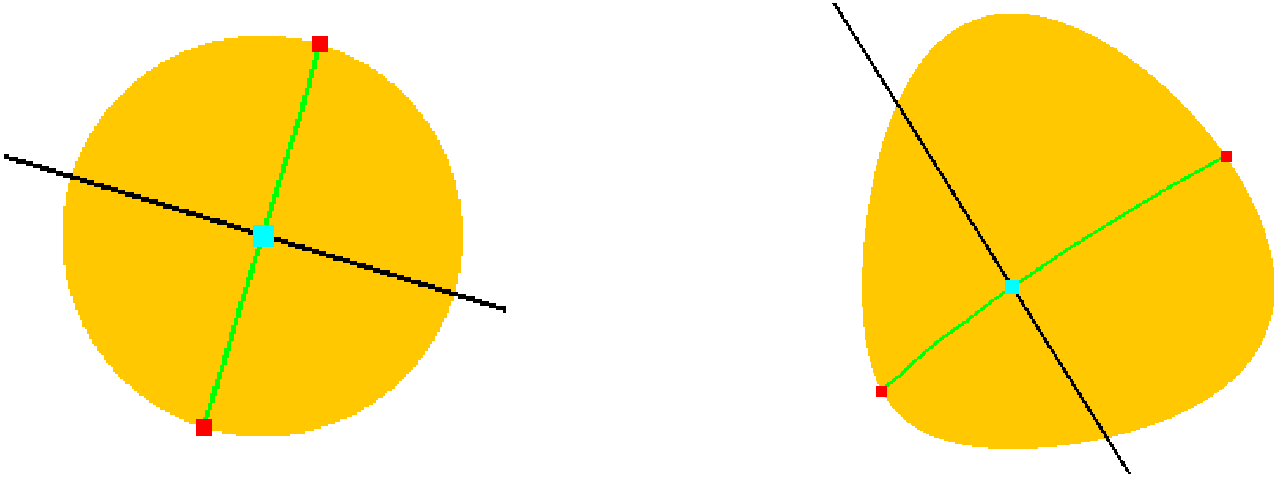
[Welzl'91] Smallest enclosing disks (balls and ellipsoids)

MINIBALL: **Solving terminal cases (in 2D)**

- $b = 3$ $(p, q$ and $r)$: **Exact** computation as the intersection of the three bisectors (2 of them are enough):



- $b = 2$ $(p$ and $q)$: Intersection of the geodesic Γ_{pq} with the bisector $H_F(p,q)$ (**approximate**, geodesic bisection walk).



Geodesics in Euclidean/Bregman spaces

In Euclidean geometry, the geodesic linking P and Q is the **line segment** joining these endpoints. Using any Cartesian coordinate system, $\Gamma_{pq} = \{x \mid x = (1-\lambda)p + \lambda q, \lambda \in [0,1]\}$. In Euclidean spaces, geodesics depicts shortest paths. ($\Gamma_{pq} = \{x \mid \mathrm{LERP}(\lambda, p, q)\}$)

In Bregman geometry (=dually flat spaces), we have two **dual (biorthogonal) coordinate systems** : θ and η. Geodesics are defined with respect to underlying **dual affine connections** $\prod$ and $\prod^*$: They are not classic Riemannian geodesics.

$$\Gamma(P,Q) = \{X \mid \theta(X) = (1-\lambda)\theta(P) + \lambda\theta(Q), \lambda \in [0,1]\}$$

$$\Gamma^*(P,Q) = \{X \mid \eta(X) = (1-\lambda)\eta(P) + \lambda\eta(Q), \lambda \in [0,1]\}$$

$\Gamma(P,Q)$ and $\Gamma^*(P,Q)$ do not usually coincide (they do for L_2^2)

Visualizing geodesics in coordinate systems...

Consider the $\theta = \nabla F^*(\eta)$ and $\eta = \nabla F(\theta)$ coordinate systems with $\nabla F^* = \nabla F^{-1}$ (from Legendre transformation).

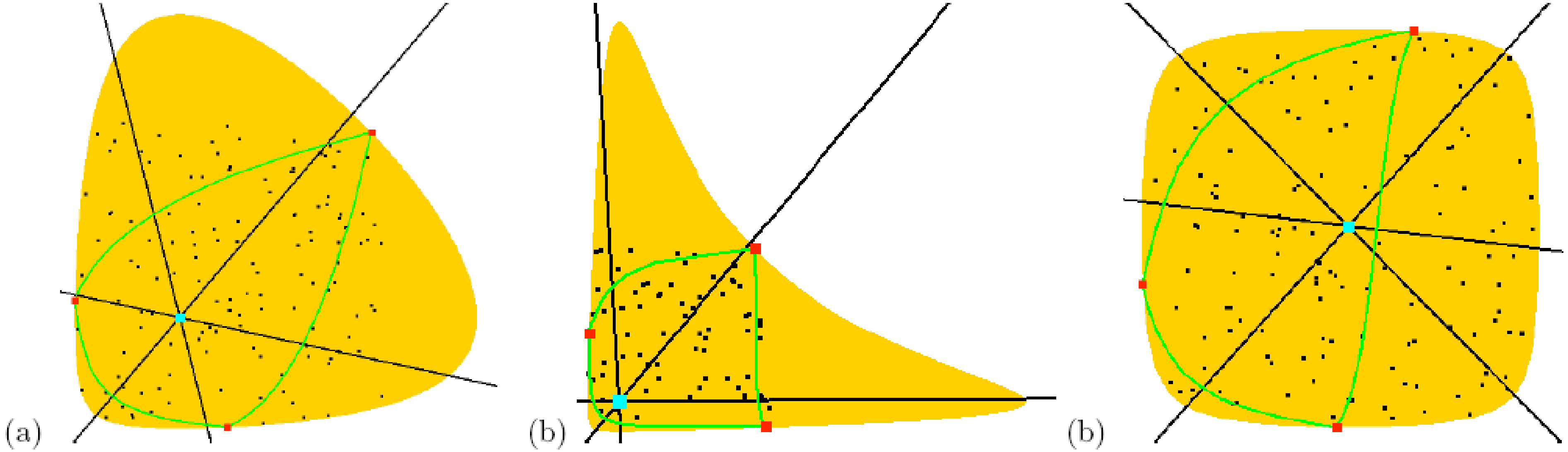
$$\Gamma(P,Q) = \{X \mid \theta(X) = (1-\lambda)\theta(P) + \lambda\theta(Q), \lambda \in [0,1]\}$$

$\Gamma(P,Q) = \{x \mid x = (1-\lambda)p + \lambda q, \lambda \in [0,1]\}$ in primal coordinate system $\theta \in \Theta \subset \mathbb{R}^d$.

- In primal θ coordinate system, the two geodesics are written as:
$$\Gamma = \{(1-\lambda)p + \lambda q | \lambda \in [0,1]\}$$
$$\Gamma^* = \left\{\nabla F^{-1}\left((1-\lambda)\nabla F(p) + \lambda \nabla F(q)\right) | \lambda \in [0,1]\right\}$$
- In dual η coordinate system, the two geodesics are written as:
$$\Gamma^* = \{(1-\lambda)p' + \lambda q' | \lambda \in [0,1]\}$$
$$\Gamma = \left\{\nabla F^{*-1}\left((1-\lambda)\nabla F^*(p) + \lambda \nabla F^*(q)\right) | \lambda \in [0,1]\right\}$$

Bregman MINIBALL**: Demo**

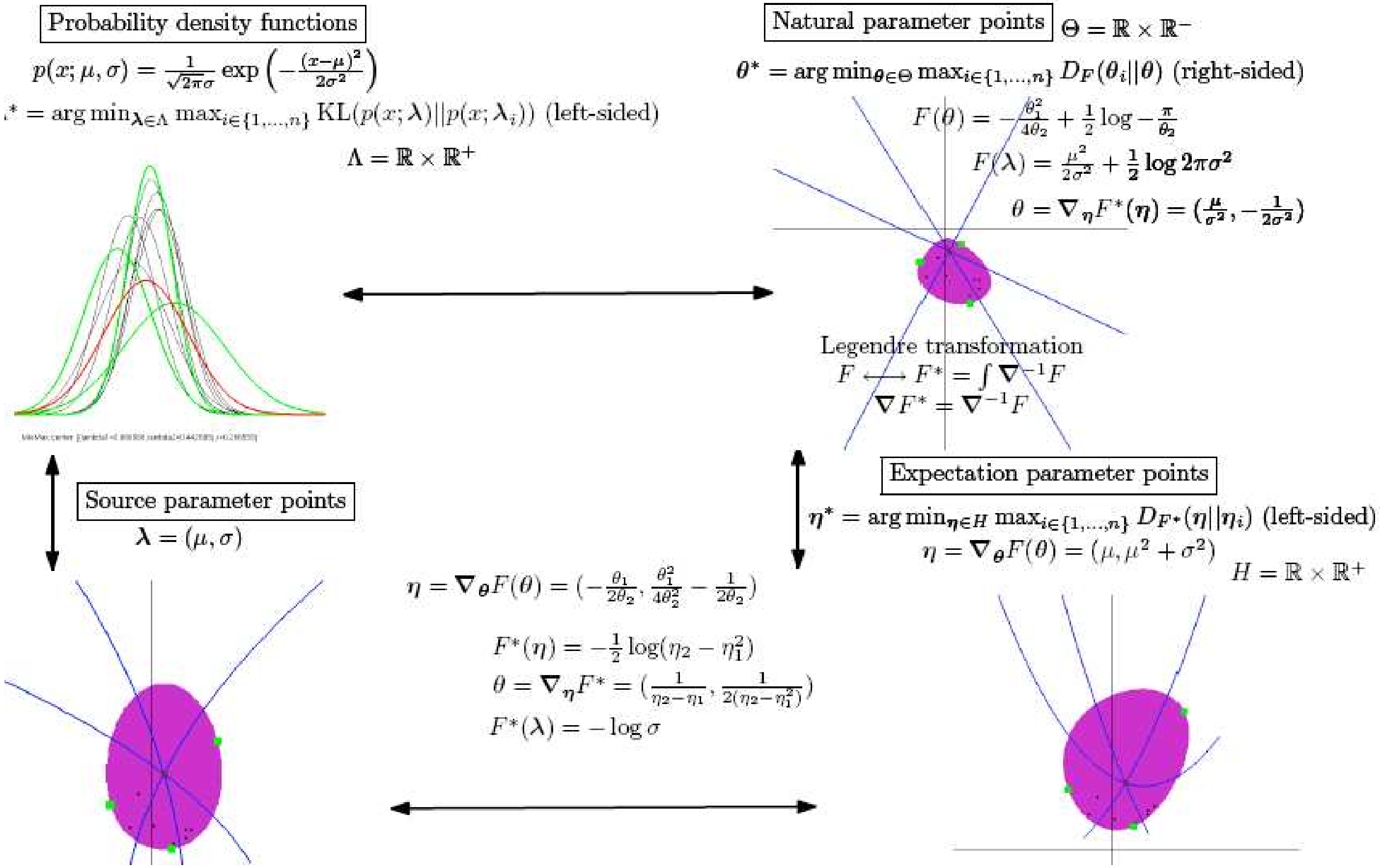
[DEMO]
http://www.sonycsl.co.jp/person/nielsen/BregmanBall/MINIBALL/



[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)



Probability density functions

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$.^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda}\max_{i\in\{1,\dots,n\}}\mathrm{KL}(p(x;\boldsymbol{\lambda})||p(x;\boldsymbol{\lambda}_i))\ \text{(left-sided)}$$

$$\Lambda = \mathbb{R}\times\mathbb{R}^+$$

Natural parameter points $\Theta = \mathbb{R}\times\mathbb{R}^-$

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta}\max_{i\in\{1,\dots,n\}} D_F(\boldsymbol{\theta}_i||\boldsymbol{\theta})\ \text{(right-sided)}$$

$$F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log -\frac{\pi}{\theta_2}$$

$$F(\lambda) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2$$

$$\theta = \nabla_{\eta}F^*(\boldsymbol{\eta}) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$$

Legendre transformation

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

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

Source parameter points

$$\boldsymbol{\lambda} = (\mu,\sigma)$$

Expectation parameter points

$$\boldsymbol{\eta}^* = \arg\min_{\boldsymbol{\eta}\in H}\max_{i\in\{1,\dots,n\}} D_{F^*}(\boldsymbol{\eta}||\boldsymbol{\eta}_i)\ \text{(left-sided)}$$

$$\eta = \nabla_{\boldsymbol{\theta}}F(\theta) = (\mu, \mu^2+\sigma^2)$$

$$H = \mathbb{R}\times\mathbb{R}^+$$

$$\boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}}F(\theta) = (-\frac{\theta_1}{2\theta_2}, \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2})$$

$$F^*(\eta) = -\frac{1}{2}\log(\eta_2 - \eta_1^2)$$

$$\theta = \nabla_{\eta}F^* = (\frac{1}{\eta_2-\eta_1}, \frac{1}{2(\eta_2-\eta_1^2)})$$

$$F^*(\lambda) = -\log\sigma$$

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

Bregman MINIBALL: Quantum information geometry

Some open questions of *quantum information theory* (QIT) turns out to be **geometric** problems:
For example: ... Take a unit sphere $\mathbb{S}$, deforms it by an **affine transform** A, and compute the smallest enclosing ball of $A\mathbb{S}$ with respect to quantum relative entropy (a Bregman divergence).
The **Holevo capacity** is the radius of the smallest ball enclosing the ellipsoid.

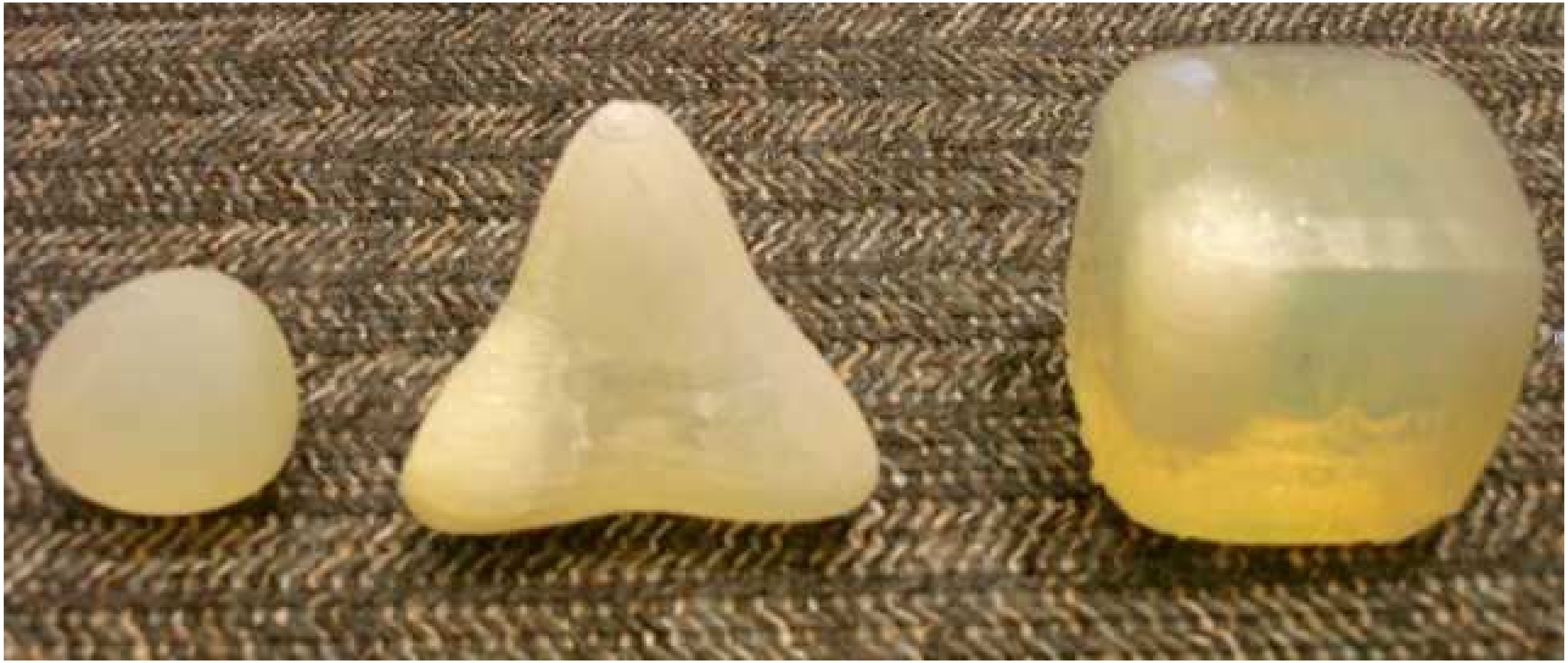
For 1-qubit systems, → at most $d + 1 = 4$ critical points of the ellipsoid $A\mathbb{S}$ (=maximum size of combinatorial basis)

← Geometry is critical for solving the additivity *conjecture* in QIT.

[ISIT'08] Quantum Voronoi diagrams and Holevo channel capacity for 1-qubit quantum states.

[Shor'00] Quantum Information Theory: Results and Open Problems.

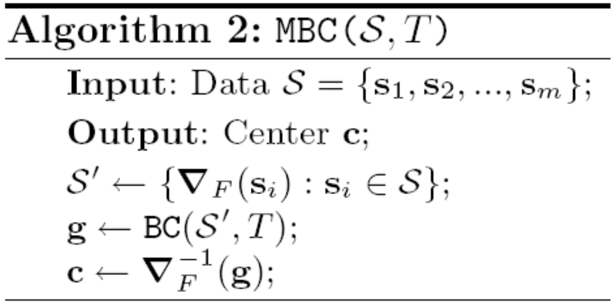
Printing 3D Bregman balls using lithography



(primal coordinate system)

Fitting the smallest Bregman enclosing ball

Randomized linear-time MiniBall fails in practice as soon as $d \simeq 30$.
Extend Badoiu & Clarkson's simple **core-set** algorithm ($O(\frac{dn}{\epsilon^2})$).



```
Algorithm 2: MBC($\mathcal{S}, T$)
    Input: Data $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$;
    Output: Center $\mathbf{c}$;
    $\mathcal{S}' \leftarrow \{\nabla_F(\mathbf{s}_i) : \mathbf{s}_i \in \mathcal{S}\}$;
    $\mathbf{g} \leftarrow$ BC($\mathcal{S}', T$);
    $\mathbf{c} \leftarrow \nabla_F^{-1}(\mathbf{g})$;
```

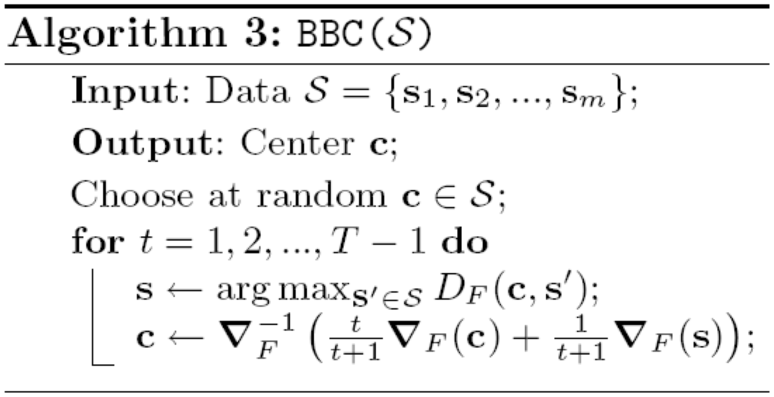
→ = BC algorithm in η-coordinate system.

→ Applications in machine learning: Core Ball Machines, etc.

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

[IJCGA'09] Approximating smallest enclosing balls with applications to machine learning

Fitting the smallest Bregman enclosing ball: Bregman BC



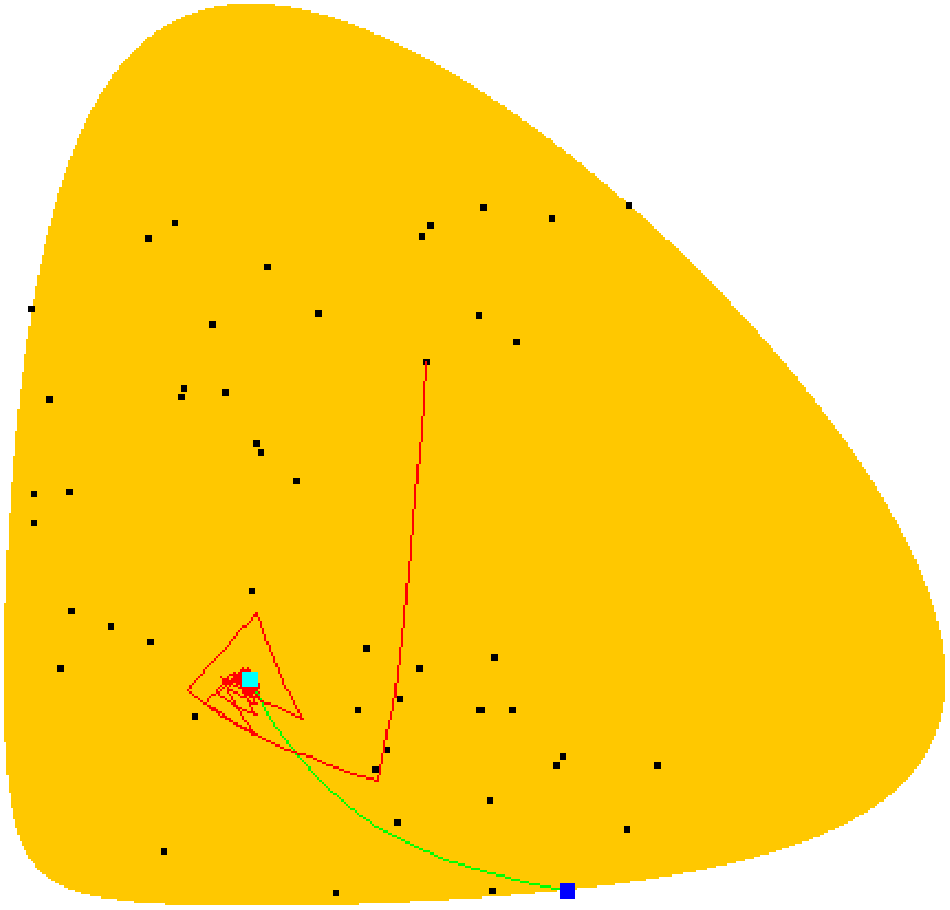
```
Algorithm 3: BBC($\mathcal{S}$)
    Input: Data $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$;
    Output: Center $\mathbf{c}$;
    Choose at random $\mathbf{c} \in \mathcal{S}$;
    for $t = 1, 2, ..., T-1$ do
        $\mathbf{s} \leftarrow \arg\max_{\mathbf{s}' \in \mathcal{S}} D_F(\mathbf{c}, \mathbf{s}')$;
        $\mathbf{c} \leftarrow \boldsymbol{\nabla}_F^{-1}\left(\frac{t}{t+1}\boldsymbol{\nabla}_F(\mathbf{c}) + \frac{1}{t+1}\boldsymbol{\nabla}_F(\mathbf{s})\right)$;
```

Panigrahy's method in $O(\frac{dn}{\epsilon})$ can be extended to Bregman divergences as well.

[CoRR'04] Minimum Enclosing Polytope in High Dimensions CoRR cs.CG/0407020: (2004)

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

Bregman core-sets: Demo



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

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

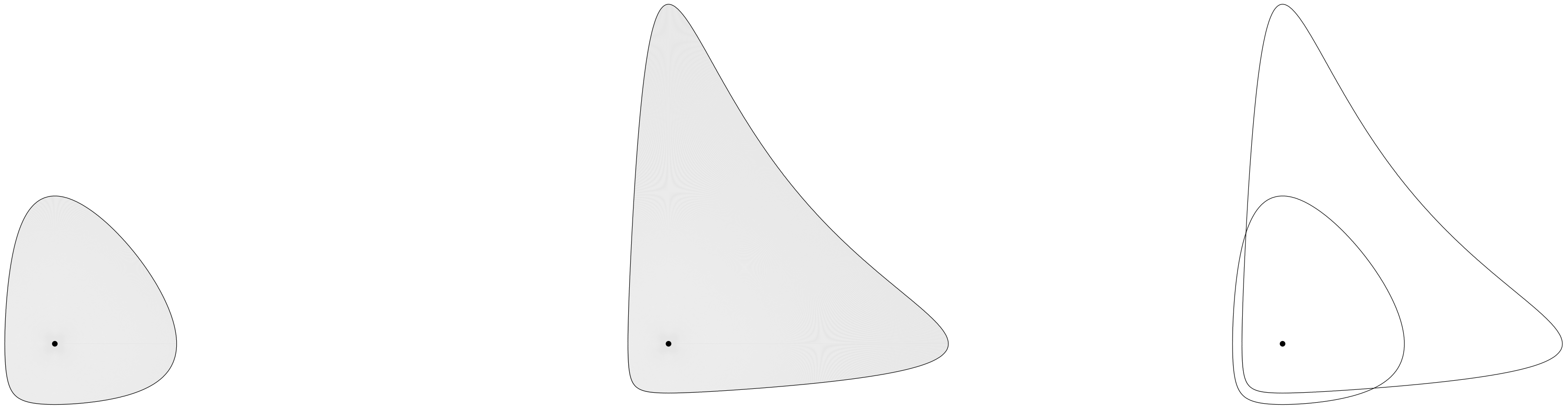
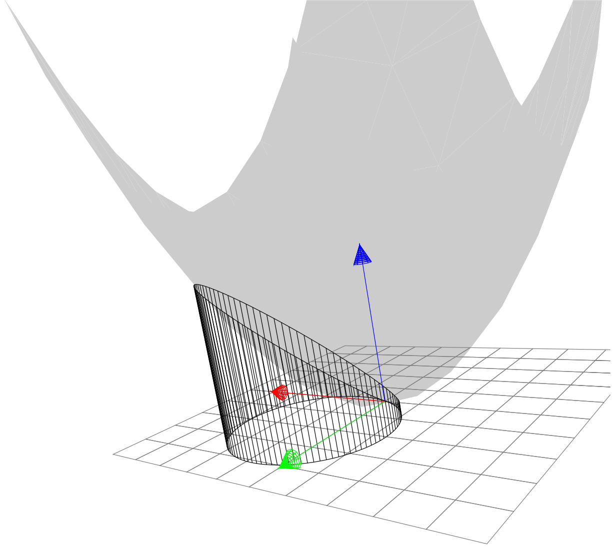
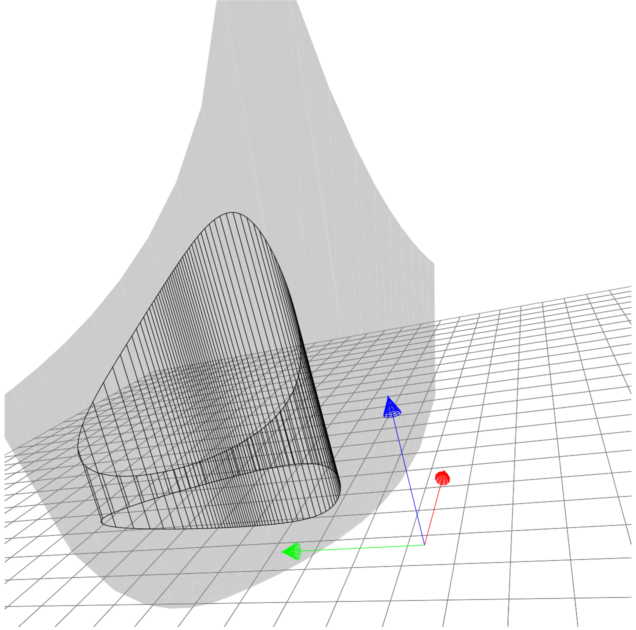


Illustration for Itakura-Saito divergence, $F(x) = -\log x$

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

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

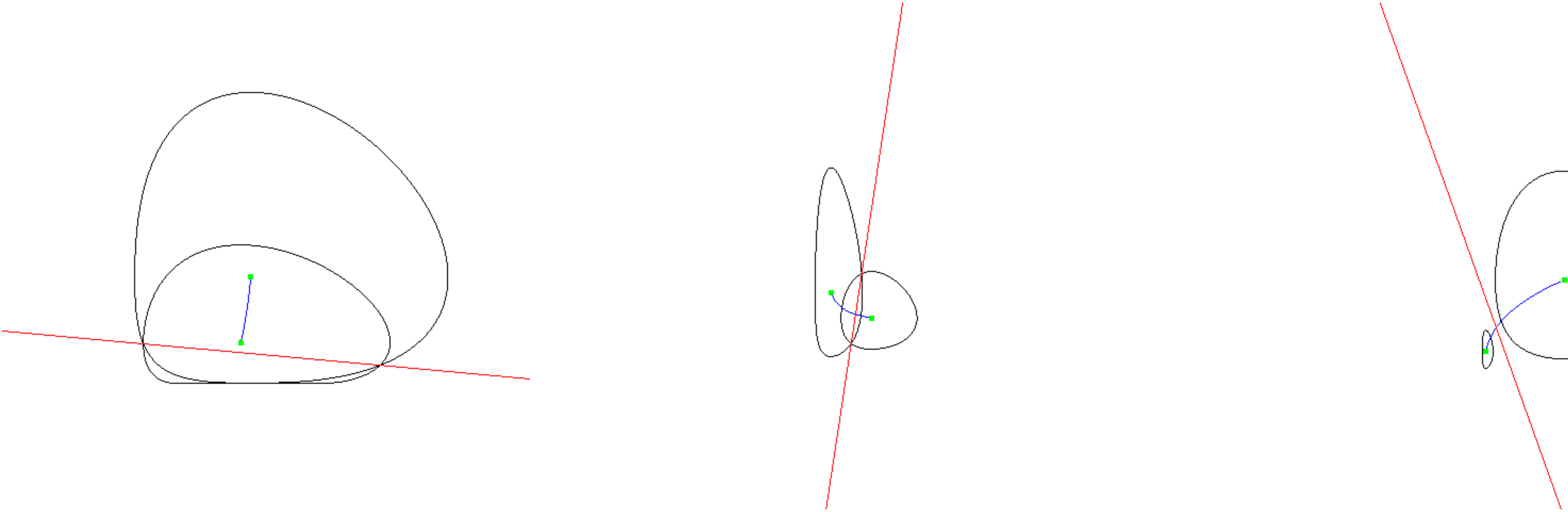
$\texttt{InSphere}(\mathbf{x}; \mathbf{p}_0, ..., \mathbf{p}_d)$ is negative, null or positive depending on whether x lies inside, on, or outside σ.

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.

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

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

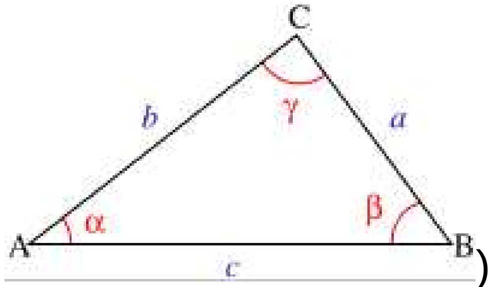
Bregman: Three-point property and Bregman projection

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

4-point property

3-point property:

$$B_F(p||q) + B_F(q||r) = B_F(p||r) + \langle p - q, \nabla F(r) - \nabla F(q) \rangle$$

...is a special case of the **4-point property** :

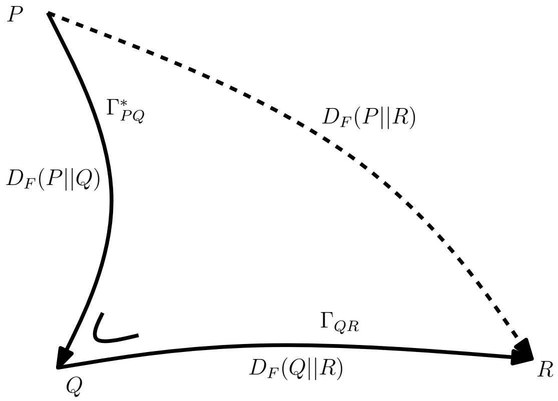
$$B_F(p||q) + B_F(s||r) - B_F(p||r) - B_F(s||q) = \langle r - q, \nabla F(p) - \nabla F(s) \rangle$$

S. Della Pietra, V. J. Della Pietra, J. D. Lafferty: Inducing Features of Random Fields.
IEEE Trans. Pattern Anal. Mach. Intell. 19(4): 380-393 (1997)

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** $\mathcal{W}$.

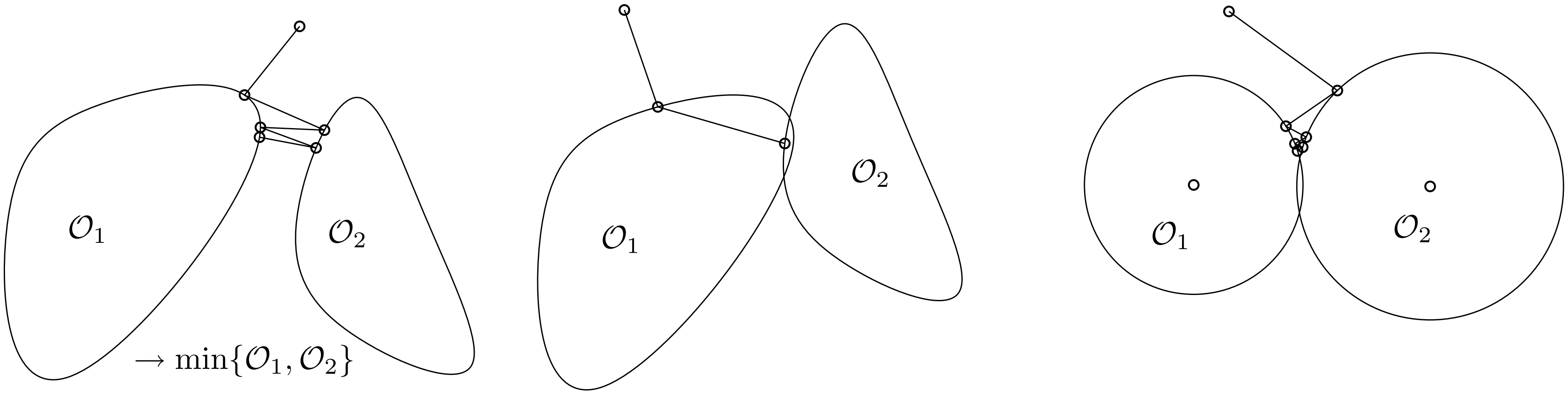


[Bregman'66] The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming.

Bregman projection: Relaxation methods

Problem: Given a set of convex objects $\mathcal{O}_1, ..., \mathcal{O}_n$, report a point in the common intersection $\cap_{i=1}^{n} \mathcal{O}_i$, if it exists.

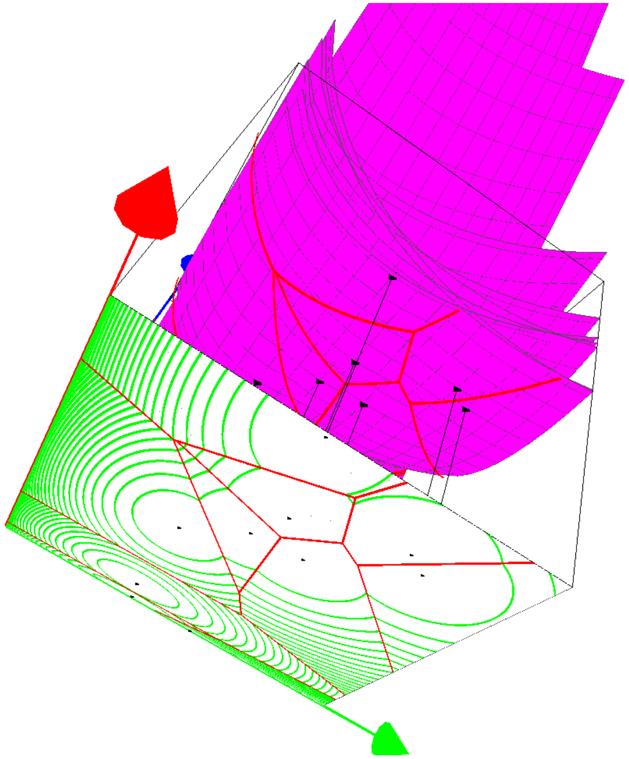
Relaxation method:
Start with *any* seed point $p_0 \in \mathcal{X}$, an consider the cyclic sequence of "constraints" $\mathcal{O}_1, ..., \mathcal{O}_n$. At iteration i, project p_{i-1} onto $\mathcal{O}_{1+(i-1 \bmod n)}$.
→ converge towards a common point **in the limit** .



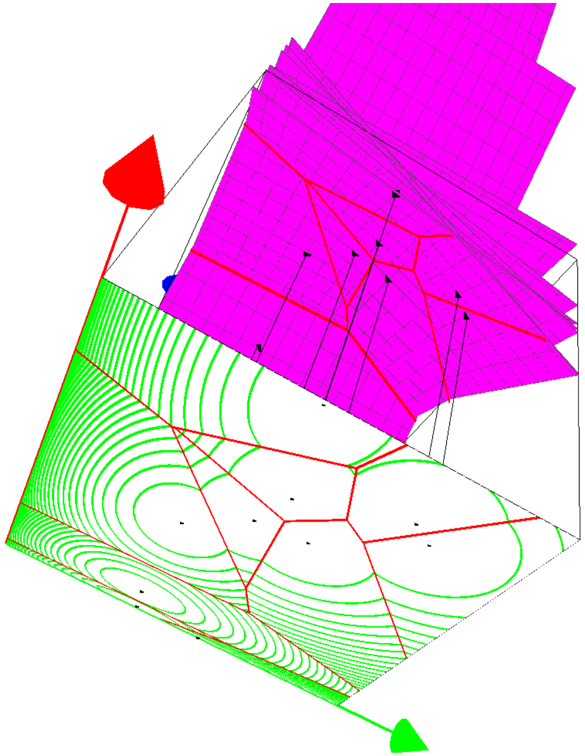
Alternating projections:
`http://www.lix.polytechnique.fr/~nielsen/BregmanProjection/`

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 $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.

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 $\mathcal{X}$)

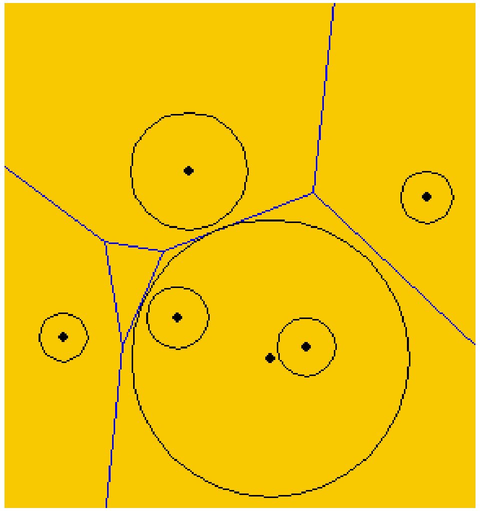
Power distance of x to $\mathrm{Ball}(p, r)$: $||p - x||^2 - r^2$.

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** :

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

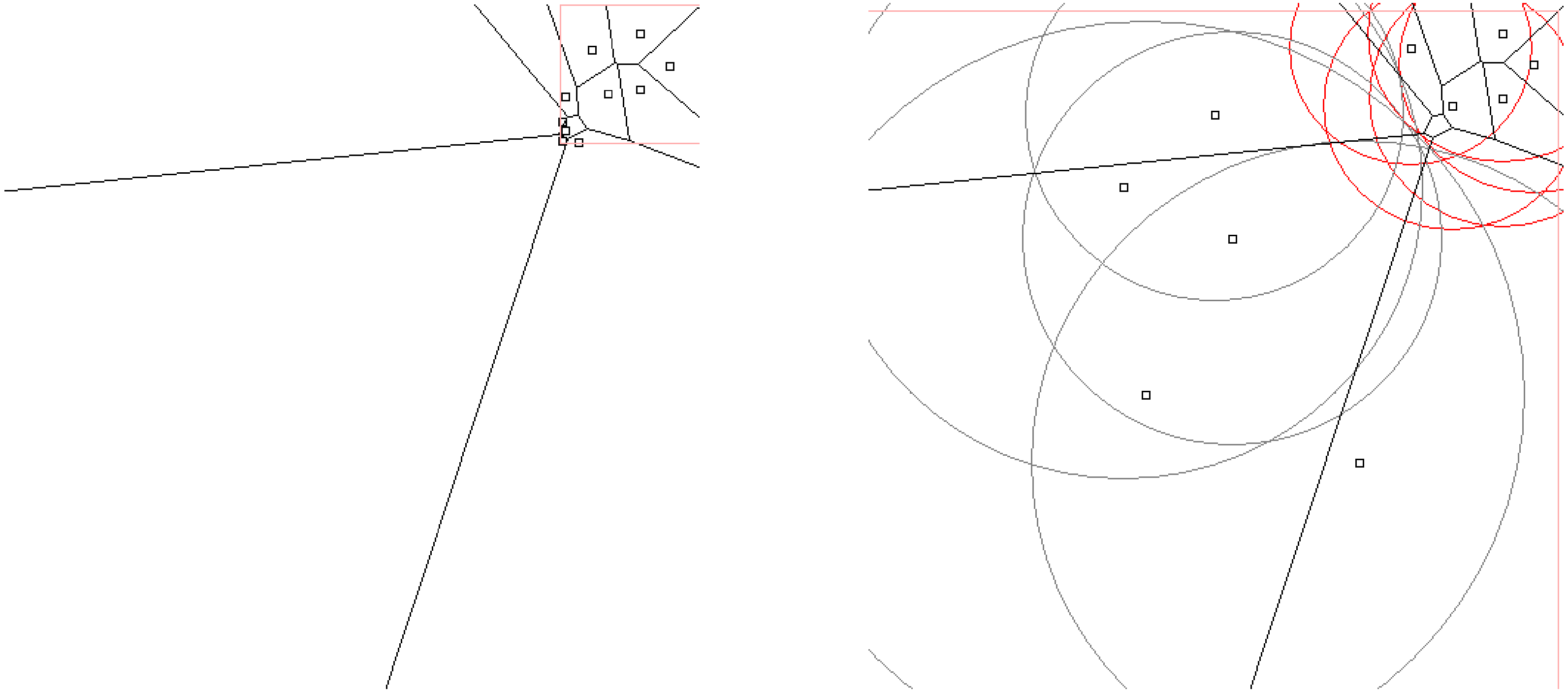
$$\boxed{\text{Affine Bregman Voronoi diagram} \Leftarrow \text{Power diagram}}$$



[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.)

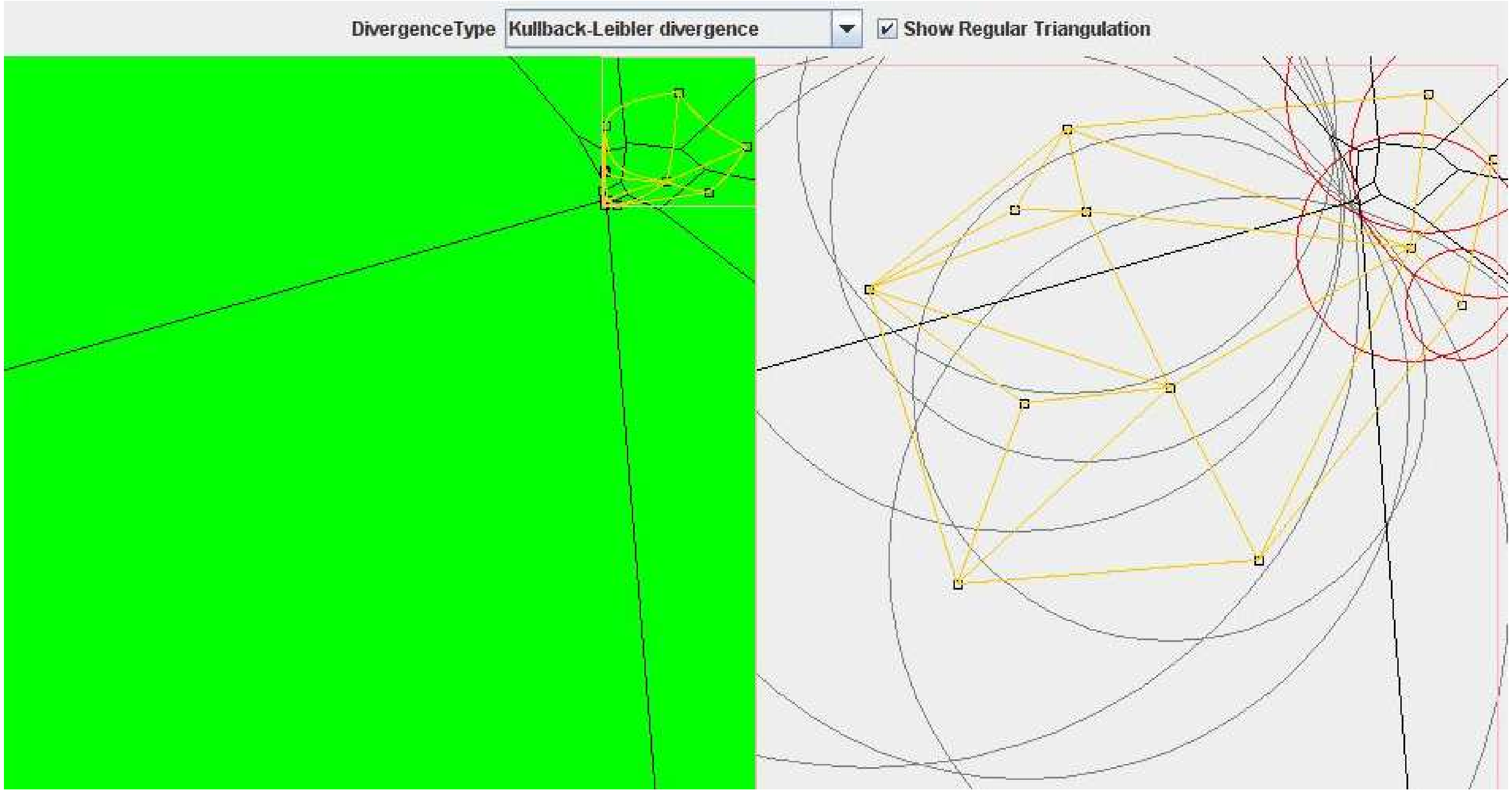
`http://www.csl.sony.co.jp/person/nielsen/BVDapplet/`

Bregman Delaunay/geodesic triangulations

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

[Rajan'94] Optimality of the Delaunay triangulation in $\mathbb{R}^d$, Disc. & Comp. Geom.

Dual regular triangulation from power diagram

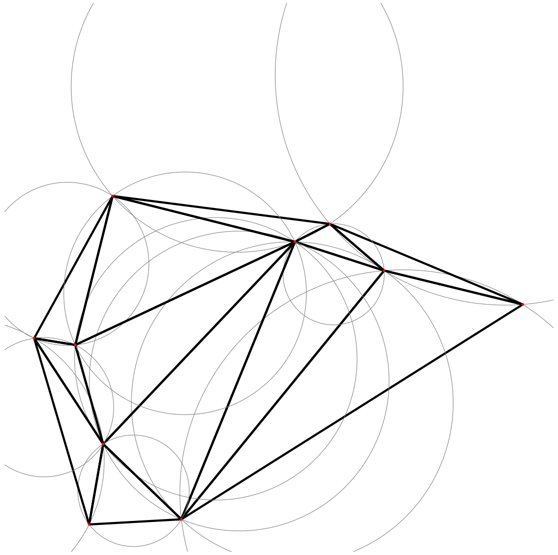


DivergenceType: Kullback-Leibler divergence

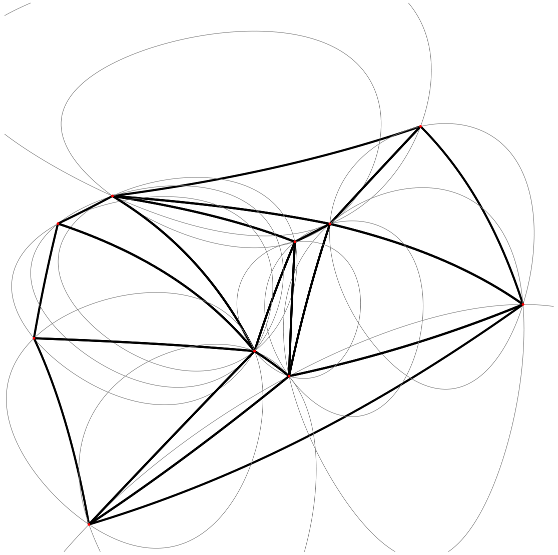
☑ Show Regular Triangulation

Kullback-Leibler divergence extended on **positive measures** (=I-divergence on $\mathbb{R}^2_{+*}$)

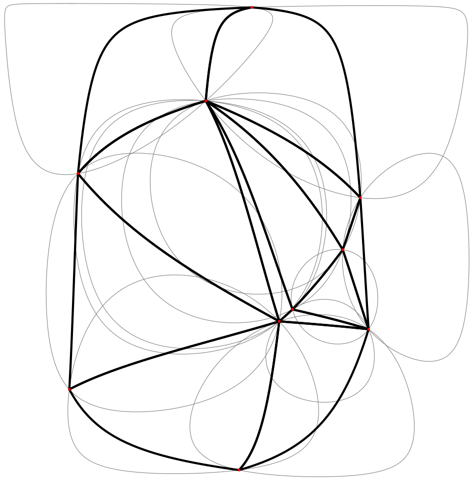
Bregman Delaunay triangulations



Ordinary Delaunay



Exponential loss



Hellinger-like divergence

- empty Bregman sphere property,
- geodesic triangles.

Bregman Voronoi/regular triangulations

$$\begin{tikzcd}
\boxed{\text{Primal space}} & & \boxed{\text{Dual gradient space}} \\
\theta & & \eta \\
\text{Affine BVD}(\mathcal{P}) \arrow[d, leftrightarrow] & = & \text{Laguerre/power diagram on } \nabla F(\mathcal{P}) \arrow[d, leftrightarrow] \\
\text{Geodesic } \mathrm{BT}(\mathcal{P}) \arrow[rr, leftrightarrow] & & \text{Regular triangulation on } \nabla F(\mathcal{P})
\end{tikzcd}$$

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 $\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$:

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

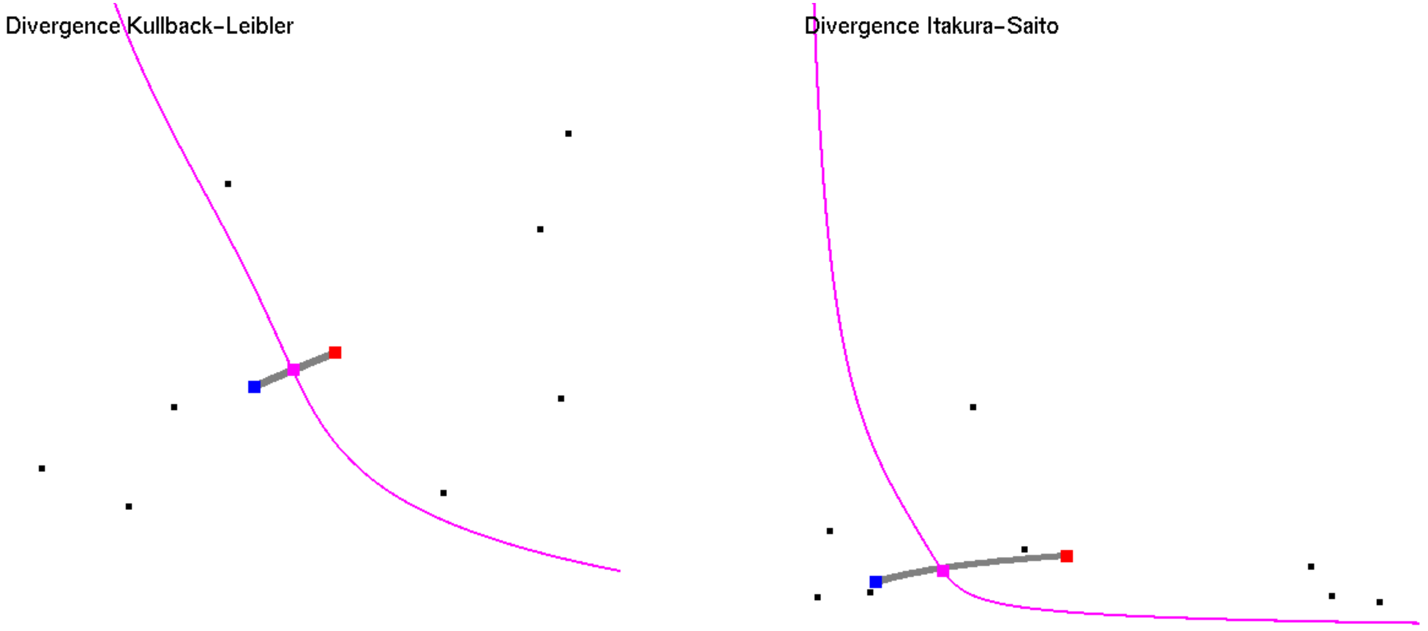
$$\begin{aligned}
\mathrm{AVG}_F(\mathcal{P}||q) &= \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + B_F(\bar{p}||q) \\
\mathrm{AVG}_F(q||\mathcal{P}) &= \mathrm{AVG}_{F^*}({\mathcal{P}_F}'||q') \\
&= (\sum_{i=1}^n \frac{1}{n} F^*(p'_i) - F^*(\bar{p'})) + B_{F^*}(\bar{p'_F}||q'_F)
\end{aligned}$$

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 $\nabla F^* \circ \nabla F(q) = q$.

$\arg\min_{c\in\mathcal{X}} \frac{1}{2}(\mathrm{AVG}_F(\mathcal{P}||q) + \mathrm{AVG}_F(q||\mathcal{P})) \iff$
$\arg\min_{q\in\mathcal{X}} B_F(c_R^F||q) + B_F(q||c_L^F)$ (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 :

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

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

Dually flat spaces (differential geometry)

$\{\mathcal{M}, g, \nabla, \nabla^*\}$: Dually flatspace

$$\langle p, q \rangle = \langle \prod p, \prod^* q \rangle = \sum_{ij} g_{ij} p_i q_i.$$

$$\mathrm{d}s^2 = \sum_{ij} g_{ij}(\theta)\mathrm{d}\theta_i \mathrm{d}\theta_j = \mathrm{d}\theta^T \nabla^2 F(\theta) \mathrm{d}\theta$$

Euclidean geometry : $G = I$, identity matrix. ($\rightarrow$ self-dual)

Riemannian geometry : $\nabla = \nabla^*$: Levi-Civita connection.

Computational geometry in dually flat spaces

Geometry & statistics handled in a unified framework:

ECML'05 Richard Nock, Frank Nielsen: *Fitting the Smallest Enclosing Bregman Ball*. ECML 2005: 649-656

IPL'08 Frank Nielsen, Richard Nock: *On the smallest enclosing information disk*. Inf. Process. Lett. 105(3): 93-97 (2008)

SHA'08 Frank Nielsen: *An Interactive Tour of Voronoi Diagrams on the GPU*, ShaderX^6, 2008.

SODA'07 Frank Nielsen, Jean-Daniel Boissonnat, Richard Nock: *On Bregman Voronoi diagrams*, 2007: 746 - 755. (arXiv 0709.2196)

ISIT'08 Frank Nielsen, Richard Nock: *Quantum Voronoi Diagrams and Holevo Channel Capacity for 1-Qubit Quantum States*. ISIT 2008.

ICPR'08 Frank Nielsen, Richard Nock: *Bregman Sided and Symmetrized Centroids*. ICPR 2008. (arXiv:0711.3242)

IJCGA'09 Frank Nielsen, Richard Nock: *Approximating Smallest Enclosing Balls with Application to Machine Learning*, International Journal on Computational Geometry and Applications , 2009.

Computational geometry in dually flat spaces

Geometry & machine learning:

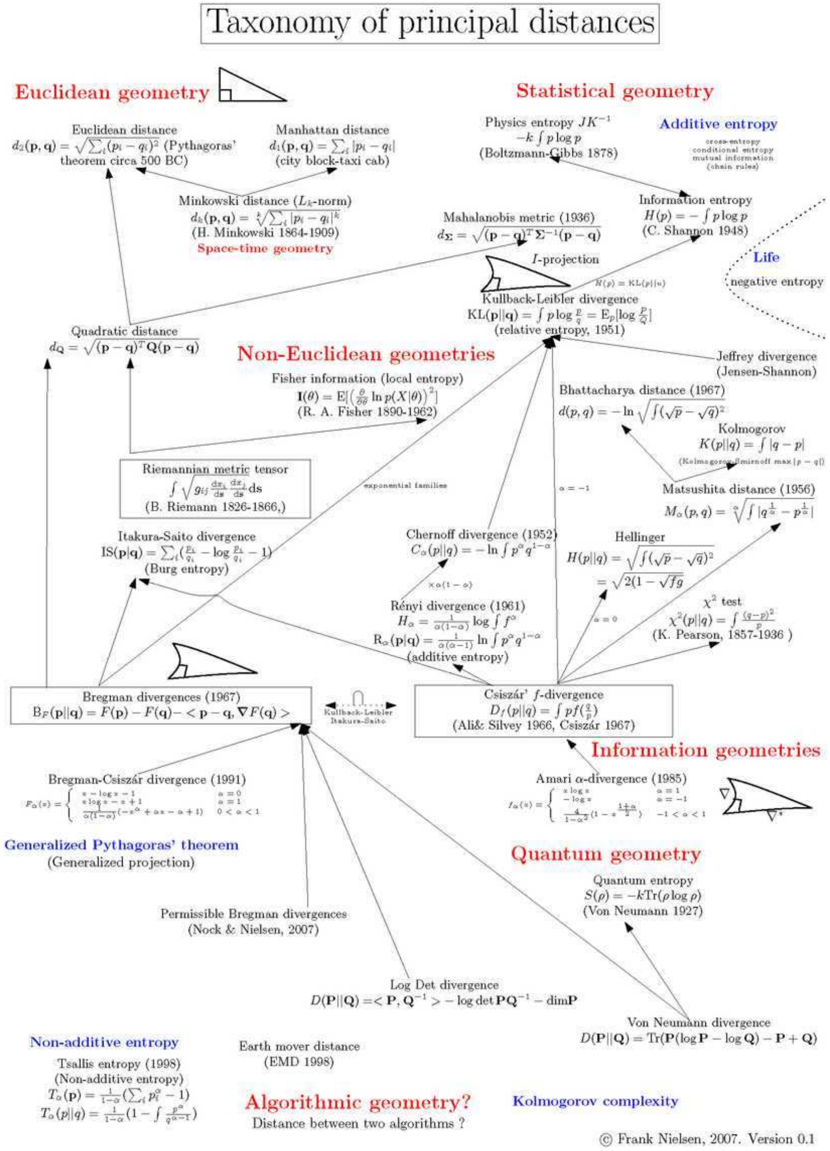
ETVC'09 Richard Nock, Frank Nielsen: *Intrinsic Geometries in Learning*, Emerging Trends in Visual Computing: 175-215, 2009. Springer Verlag LNCS 5416.

PAMI'09 Richard Nock, Frank Nielsen: *Bregman Divergences and Surrogates for Learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

PAMI'06 Richard Nock and Frank Nielsen: *On Weighting Clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence (28)-8, pp 1223-1235, 2006.

NIPS'08 Richard Nock and Frank Nielsen: *On the Efficient Minimization of Classification-Calibrated Surrogates*, Advances in Neural Information Processing Systems (NIPS*21)

A historical retrospective of distances



Taxonomy of principal distances

Euclidean geometry

Euclidean distance
$d_2(\mathbf{p},\mathbf{q}) = \sqrt{\sum_i (p_i - q_i)^2}$ (Pythagoras' theorem circa 500 BC)

Manhattan distance
$d_1(\mathbf{p},\mathbf{q}) = \sum_i |p_i - q_i|$
(city block-taxi cab)

Minkowski distance (L_k-norm)
$d_k(\mathbf{p},\mathbf{q}) = \sqrt[k]{\sum_i |p_i - q_i|^k}$
(H. Minkowski 1864-1909)
Space-time geometry

Statistical geometry

Physics entropy JK^{-1}
$-k \int p \log p$
(Boltzmann-Gibbs 1878)

Additive entropy

cross-entropy
conditional entropy
mutual information
(chain rules)

Information entropy
$H(p) = -\int p \log p$
(C. Shannon 1948)

Mahalanobis metric (1936)
$d_{\mathbf{\Sigma}} = \sqrt{(\mathbf{p}-\mathbf{q})^T \mathbf{\Sigma}^{-1} (\mathbf{p}-\mathbf{q})}$

I-projection

$H(p) = \mathrm{KL}(p||u)$

Life

negative entropy

Kullback-Leibler divergence
$\mathrm{KL}(\mathbf{p}||\mathbf{q}) = \int p \log \frac{p}{q} = \mathrm{E}_p[\log \frac{P}{Q}]$
(relative entropy, 1951)

Quadratic distance
$d_{\mathbf{Q}} = \sqrt{(\mathbf{p}-\mathbf{q})^T \mathbf{Q} (\mathbf{p}-\mathbf{q})}$

Non-Euclidean geometries

Fisher information (local entropy)
$\mathbf{I}(\theta) = \mathrm{E}[\left(\frac{\partial}{\partial\theta} \ln p(X|\theta)\right)^2]$
(R. A. Fisher 1890-1962)

Jeffrey divergence
(Jensen-Shannon)

Bhattacharya distance (1967)
$d(p,q) = -\ln \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$

Kolmogorov
$K(p||q) = \int |q - p|$
(Kolmogorov-Smirnoff $\max |p - q|$)

Riemannian metric tensor
$\int \sqrt{g_{ij} \frac{\mathrm{d}x_i}{\mathrm{d}s} \frac{\mathrm{d}x_j}{\mathrm{d}s}} \mathrm{d}s$
(B. Riemann 1826-1866,)

exponential families

$\alpha = -1$

Matsushita distance (1956)
$M_\alpha(p,q) = \sqrt[\alpha]{\int |q^{\frac{1}{\alpha}} - p^{\frac{1}{\alpha}}|}$

Itakura-Saito divergence
$\mathrm{IS}(\mathbf{p}|\mathbf{q}) = \sum_i (\frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1)$
(Burg entropy)

Chernoff divergence (1952)
$C_\alpha(p||q) = -\ln \int p^\alpha q^{1-\alpha}$

$\times \alpha(1-\alpha)$

Hellinger
$H(p||q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
$= \sqrt{2(1 - \sqrt{fg}}$

Rényi divergence (1961)
$H_\alpha = \frac{1}{\alpha(1-\alpha)} \log \int f^\alpha$
$\mathrm{R}_\alpha(\mathbf{p}|\mathbf{q}) = \frac{1}{\alpha(\alpha-1)} \ln \int p^\alpha q^{1-\alpha}$
(additive entropy)

$\alpha = 0$

χ^2 test
$\chi^2(p||q) = \int \frac{(q-p)^2}{p}$
(K. Pearson, 1857-1936)

Bregman divergences (1967)
$\mathrm{B}_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - <\mathbf{p}-\mathbf{q}, \nabla F(\mathbf{q})>$

$\cap$
Kullback-Leibler
Itakura-Saito

Csiszár' f-divergence
$D_f(p||q) = \int p f(\frac{q}{p})$
(Ali& Silvey 1966, Csiszár 1967)

Information geometries

Bregman-Csiszár divergence (1991)
$F_\alpha(x) = \begin{cases} x - \log x - 1 & \alpha = 0 \\ x \log x - x + 1 & \alpha = 1 \\ \frac{1}{\alpha(1-\alpha)}(-x^\alpha + \alpha x - \alpha + 1) & 0 < \alpha < 1 \end{cases}$

Amari α-divergence (1985)
$f_\alpha(x) = \begin{cases} x \log x & \alpha = 1 \\ -\log x & \alpha = -1 \\ \frac{4}{1-\alpha^2}(1 - x^{\frac{1+\alpha}{2}}) & -1 < \alpha < 1 \end{cases}$

∇

∇^*

Generalized Pythagoras' theorem
(Generalized projection)

Quantum geometry

Quantum entropy
$S(\rho) = -k\mathrm{Tr}(\rho \log \rho)$
(Von Neumann 1927)

Permissible Bregman divergences
(Nock & Nielsen, 2007)

Log Det divergence
$D(\mathbf{P}||\mathbf{Q}) = <\mathbf{P}, \mathbf{Q}^{-1}> - \log\det \mathbf{P}\mathbf{Q}^{-1} - \dim\mathbf{P}$

Von Neumann divergence
$D(\mathbf{P}||\mathbf{Q}) = \mathrm{Tr}(\mathbf{P}(\log \mathbf{P} - \log \mathbf{Q}) - \mathbf{P} + \mathbf{Q})$

Non-additive entropy

Tsallis entropy (1998)
(Non-additive entropy)
$T_\alpha(\mathbf{p}) = \frac{1}{1-\alpha}(\sum_i p_i^\alpha - 1)$
$T_\alpha(p||q) = \frac{1}{1-\alpha}(1 - \int \frac{p^\alpha}{q^{\alpha-1}})$

Earth mover distance
(EMD 1998)

Algorithmic geometry?
Distance between two algorithms ?

Kolmogorov complexity

© Frank Nielsen, 2007. Version 0.1

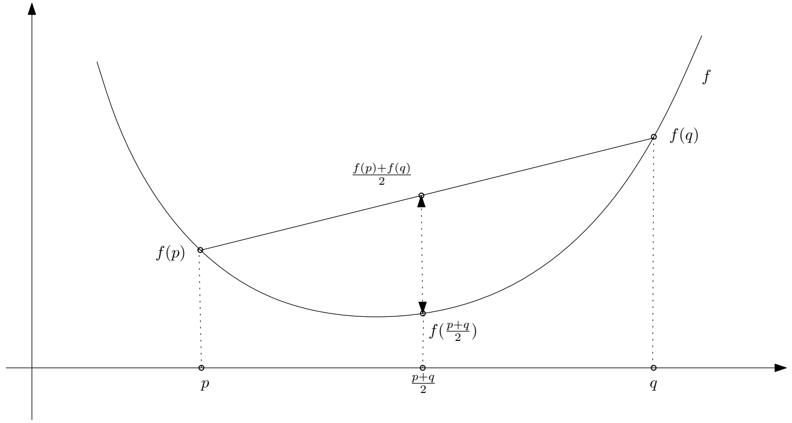
→ **Sea of distances** covered by classes of **parameterized divergences**.

Sea of distances covered by parameterized distortions

Three prominent classes of divergences:

- **Bregman divergences** $B_F(p||q) = F(p) - F(q) - \langle p - q, \nabla_F(q)\rangle = B_{F^*}(\nabla F(q)||\nabla F(p))$ for $\nabla F^* = (\nabla F)^{-1}$ (equiv. to a subclass of KL for exponential families)
- **Csiszár f-divergences** (variational, chi-squared, Hellinger-Matsusita, Chernoff, etc.) $C_f(p||q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x = C_{f^*}(q||p)$, for $f^*(y) = yf(\frac{1}{y})$. → easy to symmetrize.
- **Burbea-Rao divergences** (Jensen remainder):

$$J_f(p;q) = \frac{f(p)+f(q)}{2} - f\left(\frac{p+q}{2}\right) \geq 0$$



Thank you

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