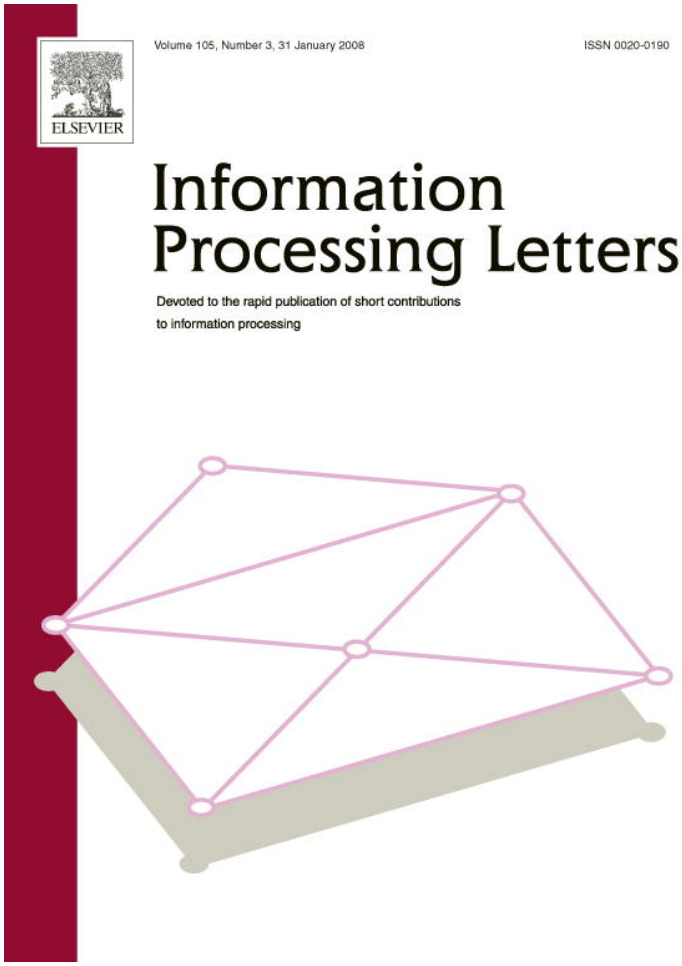

Available online at www.sciencedirect.com
ScienceDirect
Information Processing Letters 105 (2008) 93–97
Information Processing Letters
www.elsevier.com/locate/ipl

On the smallest enclosing information disk

Frank Nielsen [a,*], Richard Nock [b]

[a] *Sony Computer Science Laboratories, Inc., 3-14-13 Higashi Gotanda, Shinagawa-Ku, Tokyo 141-0022, Japan*
[b] *Université des Antilles-Guyane, Campus de Schoelcher, BP 7209, 97275 Schoelcher, Martinique, France*

Received 11 April 2006
Available online 19 August 2007
Communicated by F. Dehne

Abstract

We present a generalization of Welzl's smallest enclosing disk algorithm [E. Welzl, Smallest enclosing disks (balls and ellipsoids), in: New Results and New Trends in Computer Science, in: Lecture Notes in Computer Science, vol. 555, Springer, 1991, pp. 359–370] for point sets lying in information-geometric spaces. Given a set of points equipped with a Bregman divergence as a (dis)similarity measure, we investigate the problem of finding its unique (circum)center defined as the point minimizing the maximum divergence to the point set. As an application, we show how to solve a statistical model estimation problem by computing the center of a finite set of univariate normal distributions.

Keywords: Computational information geometry; Smallest enclosing ball; Divergence; Algorithms

1. Introduction

Given a set $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ of n vector points, we are interested in computing a simplified description that fits well $\mathcal{S}$ called its *center* $\mathbf{c}^*$. Two optimization criteria are usually considered for finding $\mathbf{c}^*$: (MINAVG) minimizes the *average distortion* $\mathbf{c}^* = \operatorname{argmin}_{\mathbf{c}} \sum_i d(\mathbf{c}, \mathbf{s}_i)$, or (MINMAX) minimizes the *maximal distortion* $\mathbf{c}^* = \operatorname{argmin}_{\mathbf{c}} \max_i d(\mathbf{c}, \mathbf{s}_i)$. These problems have been widely studied in computational geometry (1-center problem), computational statistics (1-point estimator), and machine learning (1-class classification). It is known that for *squared* Euclidean distance (L_2^2) the *centroid* is the center of MINAVG(L_2^2) [2]. For the Euclidean distance L_2, the circumcenter of $\mathcal{S}$ is the center of MINMAX(L_2), and the Fermat–Weber point is the center of MINAVG(L_2). Welzl [5] developed a simple and elegant recursive $\tilde{O}(n)$ randomized algorithm that is often used in practice. On the Euclidean plane $\mathbb{E}^2$, the distance measure $d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|$ defines a metric space (the L_2 norm). In a metric space, the distance function has important properties: (1) $d(\mathbf{p}, \mathbf{q}) \geqslant 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$, (2) symmetry $d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p})$, and (3) triangle inequality: $d(\mathbf{p}, \mathbf{r}) \leqslant d(\mathbf{p}, \mathbf{q}) + d(\mathbf{q}, \mathbf{r})$. A disk $\mathcal{B} = \text{Disk}(\mathbf{c}, r)$ of center $\mathbf{c}$ and radius r is defined as the set of points that are within distance r from the center: $\mathcal{B} = \{\mathbf{x} \in \mathbb{E}^2 \mid d(\mathbf{c}, \mathbf{x}) \leqslant r\}$. In computational machine learning, it is seldomly the case that the L_2 geometric distance reflects the distortion between two d-dimensional data elements. A general distortion framework, known as *Bregman divergences*, is

* Corresponding author. Tel.: +81 3 5448 4380; fax: +81 3 5448 4273.
E-mail addresses: Frank.Nielsen@acm.org (F. Nielsen), Richard.Nock@martinique.univ-ag.fr (R. Nock).

0020-0190/$ – see front matter
doi:10.1016/j.ipl.2007.08.007

Table 1
Common Bregman divergences D_F

Name	Domain	$F(\mathbf{s})$	$D_F(\mathbf{c},\mathbf{s})$
Squared Euclidean	$\mathbb{R}^d$	$\sum_{j=1}^d s_j^2$	$\sum_{j=1}^d (c_j - s_j)^2$
I-divergence extended KL	$(\mathbb{R}^{+,*})^d$	$\sum_{j=1}^d s_j \log s_j - s_j$	$\sum_{j=1}^d c_j \log(c_j/s_j) - c_j + s_j$
Itakura–Saito	$(\mathbb{R}^{+,*})^d$	$-\sum_{j=1}^d \log s_j$	$\sum_{j=1}^d (c_j/s_j) - \log(c_j/s_j) - 1$

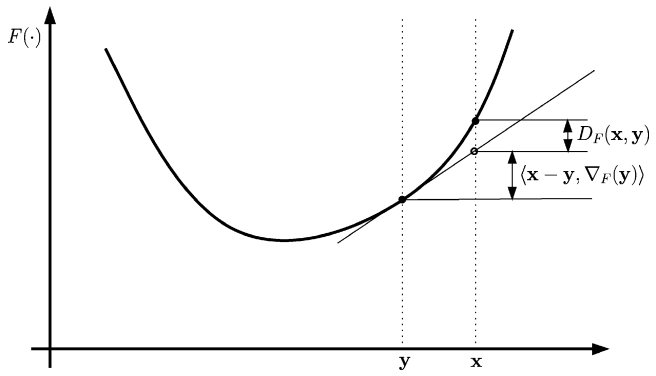

Fig. 1. Visualizing convex and differentiable function F and its corresponding Bregman divergence $D_F(\cdot,\cdot)$.

rather used. Informally speaking, a Bregman divergence D_F is the tail of a Taylor expansion of a strictly convex and differentiable function F: $D_F(\mathbf{x},\mathbf{y}) = F(\mathbf{x}) - F(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \nabla_F(\mathbf{y})\rangle$, where ∇_F denotes the gradient operator, and $\langle\cdot,\cdot\rangle$ the inner product (dot product). Bregman divergences are parameterized families of distortions defined on a convex domain $\mathcal{X} \subseteq \mathbb{R}^d$ for strictly convex and differentiable functions F on $\mathrm{int}(\mathcal{X})$ (see Fig. 1 and Table 1).

For Bregman divergences, there exist two types of Bregman balls depending on the argument position of the center [4]:

$$\mathcal{B}_{\mathbf{c},r} = \left\{\mathbf{x}\in\mathcal{X}\colon\ D_F(\mathbf{c},\mathbf{x}) \leqslant r\right\}$$

and

$$\mathcal{B}'_{\mathbf{c},r} = \left\{\mathbf{x}\in\mathcal{X}\colon\ D_F(\mathbf{x},\mathbf{c}) \leqslant r\right\},$$

that are not necessarily convex nor identical. We can further define a third-type disk by taking the symmetric divergence

$$D'_F(\mathbf{x},\mathbf{c}) = D'_F(\mathbf{c},\mathbf{x}) = \frac{D_F(\mathbf{x},\mathbf{c}) + D_F(\mathbf{c},\mathbf{x})}{2}.$$

In the reminder, we consider only the first-type disks $\mathcal{B}_{\mathbf{c},r}$. (Computing second-type disks can be transformed into first-type disks using the Legendre transformation.) We have shown in [4] that Bregman smallest enclosing balls are unique, thus generalizing the former results of Welzl for balls/ellipsoids [5]. We denote by $\mathcal{B}^*\mathcal{S}$ the smallest enclosing ball of set $\mathcal{S}$. Moreover, let $c^*(\mathcal{S})$ and $r^*(\mathcal{S})$ denote the center and radius of the smallest enclosing ball $\mathcal{B}^*(\mathcal{S})$ of $\mathcal{S}$.

2. LP-type and basis procedures

The randomized linear-time algorithm of Welzl [5] for finding the smallest enclosing ellipsoid is a particular case of a broader class of algorithms that solve linear programming-type (LP-type) problems [3]. Finding the smallest Bregman ball is LP-type because it satisfies the two sufficient and necessary LP-type axioms [3]:

Monotonicity. For any $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F}\subseteq\mathcal{G}\subseteq\mathcal{X}$, $r^*(\mathcal{F}) \leqslant r^*(\mathcal{G})$.

Locality. For any $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F}\subseteq\mathcal{G}\subseteq\mathcal{X}$ with $r^*(\mathcal{F}) = r^*(\mathcal{G})$, and any point $\mathbf{p}\in\mathcal{X}$, $r^*(\mathcal{G}) < r^*(G\cup\{\mathbf{p}\}) \to r^*(\mathcal{F}) < r^*(F\cup\{\mathbf{p}\})$.

The latter locality property holds because of the uniqueness of Bregman balls. Thus, we are able to use Welzl's abstract randomized recursive algorithm [5]:

```
MINIINFOBALL S = {p_1, ..., p_n}, B
1. ◁ Initially B = ∅. Returns B* = (c*, r*) ▷
2. if |S ∪ B| ⩽ 3
3.    then return B = SOLVEINFOBASIS(S ∪ B)
4.    else
5.       Select at random p ∈ S
6.       B* = MINIINFOBALL(S\{p}, B)
7.       if p ∉ B*
8.          then ◁ Then add p to the basis ▷
9.             return MINIINFOBALL (S\{p}, B ∪ {p})
```

We still need to solve the basis problem: solving the smallest enclosing disk of (at most) three points $\mathcal{B}$. We do this by computing all enclosing disks of $\mathcal{B}$ generated by either two or three points of $\mathcal{B}$ on their boundaries, and choose the disk that has minimum radius (i.e., minimum divergence). For computing *exactly* the center of a Bregman disk passing through three points, we first define the Bregman bisectors. Let $\mathrm{Bisector}(\mathbf{p},\mathbf{q}) = \{\mathbf{c}\in\mathcal{X} \mid D_F(\mathbf{c},\mathbf{p}) = D_F(\mathbf{c},\mathbf{q})\}$ be the Bregman bisector of locii $\mathbf{p}$ and $\mathbf{q}$. That is, $\mathrm{Bisector}(\mathbf{p},\mathbf{q})$ represents the set of

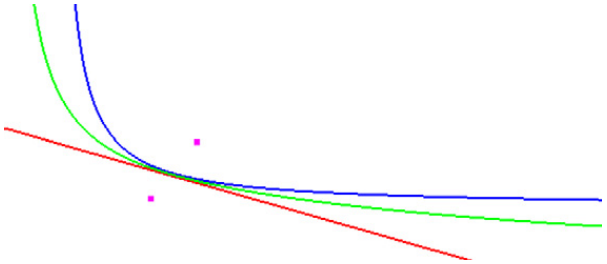

Fig. 2. Three Itakura-Saito bisectors: first-type (red), second-type (blue) and third-type (green). The first-type Bregman bisector is *always* a linear separator. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

points that have the same divergence to $\mathbf{p}$ and $\mathbf{q}$. We observe that the first-type Bregman bisector is linear. (But not necessarily the second- nor the third-type, as depicted in Fig. 2.) Proof: We write $D_F(\mathbf{c},\mathbf{p}) = D_F(\mathbf{c},\mathbf{q})$. That is,

$$F(\mathbf{c}) - F(\mathbf{p}) - \big\langle(\mathbf{c}-\mathbf{p}), \nabla_F(\mathbf{p})\big\rangle = F(\mathbf{c}) - F(\mathbf{q}) - \big\langle(\mathbf{c}-\mathbf{q}), \nabla_F(\mathbf{q})\big\rangle,$$

$$\big\langle\mathbf{c}, \big(\nabla_F(\mathbf{p}) - \nabla_F(\mathbf{q})\big)\big\rangle + F(\mathbf{p}) - F(\mathbf{q}) + \big\langle\mathbf{q}, \nabla_F(\mathbf{q})\big\rangle - \big\langle\mathbf{p}, \nabla_F(\mathbf{p})\big\rangle = 0.$$

This is a linear equation in $\mathbf{c}$. Thus, the bisector $\mathrm{Bisector}(\mathbf{p},\mathbf{q}) = \{\mathbf{x} \mid \langle\mathbf{x}, \mathbf{d_{pq}}\rangle + k_{\mathbf{pq}} = 0\}$ (with $\mathbf{d_{pq}} = \nabla_F(\mathbf{p}) - \nabla_F(\mathbf{q})$ a vector and $k_{\mathbf{pq}} = F(\mathbf{p}) - F(\mathbf{q}) + \langle\mathbf{q}, \nabla_F(\mathbf{q})\rangle - \langle\mathbf{p}, \nabla_F(\mathbf{p})\rangle$ a constant) is geometrically an hyperplane (e.g., a line for 2D vectors).

It follows that for *any* Bregman divergence, the exact circumcenter of the Bregman disk passing through three points $\mathbf{p}_1, \mathbf{p}_2$ and $\mathbf{p}_3$ can be computed *exactly* as the intersection point of any two bisectors. We get: $\mathbf{c}^* = \mathsf{l}_{12} \times \mathsf{l}_{13} = \mathsf{l}_{12} \times \mathsf{l}_{23} = \mathsf{l}_{13} \times \mathsf{l}_{23}$, where l_{ij} is the projective point associated to the linear bisector $\mathrm{Bisector}(\mathbf{p}_i, \mathbf{p_j})$ and $\times$ denote the vector cross-product operation. That is, the "circumcenter" of three points is the Bregman *trisector*, as shown in Fig. 3. Observe that although we compute exactly the circumcenter, the border of the Bregman ball is rasterized approximately (require to solve a convex optimization). To solve for the exact circumcenter $\mathbf{c}^*$ of the smallest Bregman disk passing through two points $\mathbf{p}$ and $\mathbf{q}$, we consider $\mathbf{c}^* \in \mathrm{Bisector}(\mathbf{p},\mathbf{q})$ and minimize $D_F(\mathbf{c},\mathbf{p})$. For Mahalanobis distance, $\mathbf{c}^*$ is simply the intersection of the bisector with the line passing through $\mathbf{p}$ and $\mathbf{q}$ (for L_2, it is simply $(\mathbf{p}+\mathbf{q})/2$). This is not always the case for Bregman divergences (e.g., Kullback–Leibler or Itakura–Saito divergences). However, for *any* Bregman divergence, we get a convex optimization problem that can be solved approximately within b bits (machine precision) in $\mathrm{O}(b)$ time. In fact, we do not need to compute the circumcenter of the disk, but rather implicitly represent and manipulate the disk using its combinatorial basis (two or three points). To decide whether a point $\mathbf{p}$ falls inside, on, or outside the Bregman disk defined by two points $\mathbf{p}_1$ and $\mathbf{p}_2$, We compute the exact radius of the disk $\mathrm{Disk}(\mathbf{p}_1,\mathbf{p}_2,\mathbf{p})$ and check whether that radius is strictly larger than the radius of the smallest disk $\mathrm{Disk}(\mathbf{p}_1,\mathbf{p}_2)$ or not. This becomes a *decision problem* (better known as the *In-Sphere* predicate) that can be solved only within some prescribed precision (bit complexity model). (Note that Bregman co-circularities can be detected exactly by checking the centers of any 3-point subsets.)

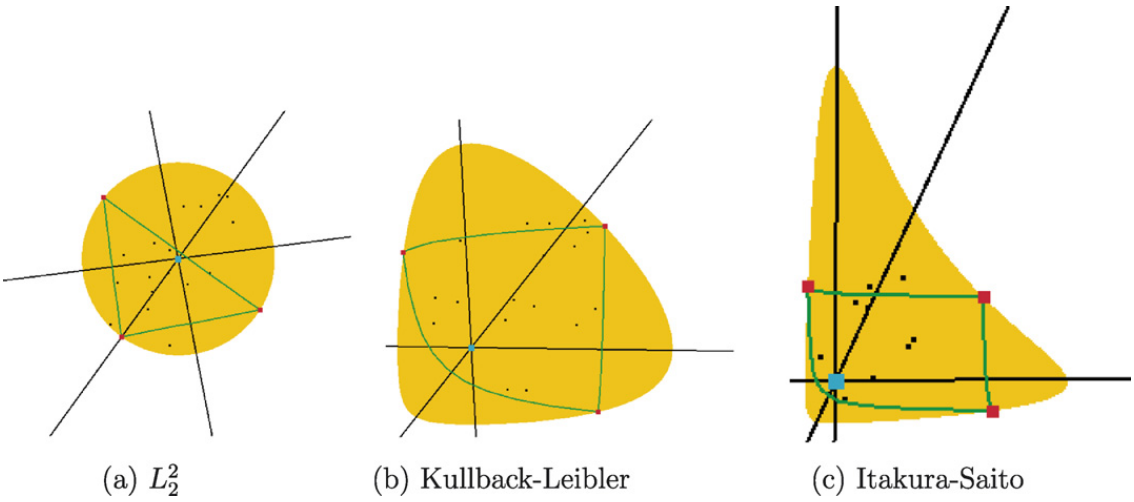

Fig. 3. Examples of smallest enclosing information disks, all chosen with basis size 3. Observe that the center may fall outside the convex hull of the support points. Snapshots from http://www.sonycsl.co.jp/person/nielsen/BregmanBall/MINIBALL/.

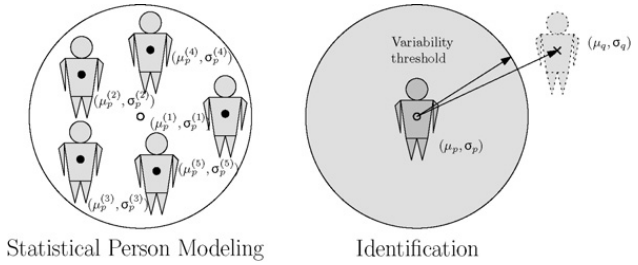

Fig. 4. Identification/change detection by model selection: each observation of a person P yields a statistical Normal distribution $(\mu_p^{(i)}, \sigma_p^{(i)})$. A person is modeled using a Normal distribution (μ_p, σ_p) by minimizing the MinMax KL divergence and a variability parameter r: the radius of the smallest enclosing Bregman disk. A person Q is different from P iff $\mathrm{KL}(P\|Q) > r$.

3. An application example

Let $\mathcal{N}$ denote the statistical exponential family of univariate Normal distributions. A Normal probability distribution $N(\mu,\sigma) \in \mathcal{N}$ with mean μ and variance σ^2 (σ is called the standard deviation) has a corresponding probability density function (pdf)

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

(with $\forall x\ N(x|\mu,\sigma) \geqslant 0$ and $\int_{-\infty}^{\infty} N(x|\mu,\sigma)\,\mathrm{d}x = 1$). Let $\mathcal{S} = \{N_1,\ldots,N_n\}$ be a set of n univariate Normal distributions $N_i = N(\mu_i,\sigma_i)$. Estimating the population center of $\mathcal{S}$ amounts to find the Normal distribution $N^*(\mu^*,\sigma^*) \in \mathcal{N}$ such that the maximum divergence of N^* to any $N_i \in \mathcal{S}$ is minimized. That is, the population mean μ^* and population variance σ^{*2} defines the center of the smallest enclosing disk of the 2D set $\mathcal{P} = \{\mathbf{p}_1,\ldots,\mathbf{p}_n\}$ with $\mathbf{p}_i = (\mu_i,\sigma_i)$ for all $i \in \{1,\ldots,n\}$. This statistical model selection is useful for person/machine identification or change detection algorithms that require to take into account variability of distributions, as depicted in Fig. 4. We need to choose an appropriate divergence D_F for Normal distributions. The entropy $H(f)$ of a pdf f is defined as $H(f) = \int_x f(x)\log_2\frac{1}{f(x)}$, and mathematically represents the amount of information in bits. The relative entropy KL, also known as the Kullback–Leibler divergence [1], measures the dissimilarity of two probability distributions f and g: $\mathrm{KL}(f\|g) = \int_x f(x)\log_2\frac{f(x)}{g(x)}$. The relative entropy corresponds to the average number of additional bits required for coding f when using an optimal code for g. The measure is therefore not symmetric nor does the triangle inequality hold. It turns out that the relative entropy KL is a good distortion measure of distributions, and belongs to the family of Bregman divergences. In fact, we can rewrite the pdf of the Normal distribution as

$$N(x|\mu,\sigma) = \frac{1}{Z(\boldsymbol{\theta})}\exp\{\langle\boldsymbol{\theta},\mathbf{f}(x)\rangle\}, \quad \text{with}$$

$$Z(\boldsymbol{\theta}) = \sqrt{2\pi}\sigma\exp\left\{\frac{\mu^2}{2\sigma^2}\right\} = \sqrt{-\frac{\theta_1}{2}}\exp\left\{-\frac{\theta_2^2}{4\theta_1}\right\},$$

$\mathbf{f}(x) = [x^2\ x]^{\mathrm{T}}$ is called the *sufficient statistics* and $\boldsymbol{\theta} = [-\frac{1}{2\sigma^2}\ \frac{\mu}{\sigma^2}]^{\mathrm{T}}$ is called the *natural parameters* of the *statistical exponential family* of Normal distributions. Exponential families contain many famous distributions such as Poisson, Gaussian and multinomial distributions, and have been thoroughly studied in Information Geometry [1]. The Kullback–Leibler divergence between any two models $\mathbf{p}$ and $\mathbf{q}$ of an exponential family is obtained from the Bregman divergence by choosing $F(\boldsymbol{\theta}) = -\log Z(\boldsymbol{\theta})$, and swapping arguments. This yields

$$\begin{aligned}\mathrm{KL}(\boldsymbol{\theta}_q\|\boldsymbol{\theta}_p) &= D_F(\boldsymbol{\theta}_p,\boldsymbol{\theta}_q)\\ &= \langle(\boldsymbol{\theta}_p-\boldsymbol{\theta}_q),\boldsymbol{\theta}_p[\mathbf{f}]\rangle + \log\frac{Z(\boldsymbol{\theta}_q)}{Z(\boldsymbol{\theta}_p)},\end{aligned}$$

with:

$$\boldsymbol{\theta}_p[\mathbf{f}] = \begin{bmatrix}\int_x \frac{x^2}{Z(\boldsymbol{\theta}_p)}\exp\{\langle\boldsymbol{\theta}_p,\mathbf{f}(x)\rangle\}\\ \int_x \frac{x}{Z(\boldsymbol{\theta}_p)}\exp\{\langle\boldsymbol{\theta}_p,\mathbf{f}(x)\rangle\}\end{bmatrix} = \begin{bmatrix}\mu_p^2+\sigma_p^2\\ \mu_p\end{bmatrix}.$$

The equation of the linear bisector is:

$$\langle(\boldsymbol{\theta}_p-\boldsymbol{\theta}_q),\boldsymbol{\theta}_c[\mathbf{f}]\rangle + \log\frac{Z(\boldsymbol{\theta}_p)}{Z(\boldsymbol{\theta}_q)} = 0.$$

Thus, we choose to change variables as $(\mu,\sigma) \to (\mu^2+\sigma^2,\mu) = (x,y)$ to get the proper coordinate system on which to apply Welzl's algorithm [5]. It comes that

$$Z(x,y) = \sqrt{x-y^2}\exp\left\{\frac{y^2}{2(x-y^2)}\right\},$$

$$\log Z(x,y) = \log\sqrt{x-y^2} + \frac{y^2}{2(x-y^2)} \quad \text{and}$$

$$\nabla_F(x,y) = \left(\frac{1}{2(x-y^2)} - \frac{y^2}{2(x-y^2)}, \frac{y^3}{(x-y^2)^2}\right).$$

Once the center (x^*,y^*) of the smallest enclosing ball is computed, we retrieve the corresponding (μ^*,σ^*) parameters as $(y^*, x^*-(y^*)^2)$. For example, the ball passing through exactly three Normal "statistical" points $(\mu_1 = 2,\sigma_1 = 1)$, $(\mu_2 = 2,\sigma_2 = \frac{3}{2})$ and $(\mu_3 = 3,\sigma_3 = 1)$ has center $(\mu^*,\sigma^*) \simeq (2.67446, 1.08313)$ and radius $r^* \simeq 0.801357$. Note that for two normal distributions $N_i(\mu_i,\sigma_i)$ and $N_j(\mu_j,\sigma_j)$, the relative Kullback–Leibler entropy $\mathrm{KL}(N_i\|N_j)$ admits the following closed-form solution

$$\mathrm{KL}(N_i\|N_j) = \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_i^2} + 2\log_2\frac{\sigma_i}{\sigma_j} - 1 + \frac{(\mu_j-\mu_i)^2}{\sigma_i^2}\right).$$

Thus, if we assume the standard deviations identical, this KL-divergence becomes simply a weighted squared Euclidean distance.

References

[1] S.-I. Amari, H. Nagaoka, Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191, AMS, ISBN 0821805312, 2000.

[2] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, Journal of Machine Learning Research 6 (2005) 1705–1749.

[3] J. Matoušek, M. Sharir, E. Welzl, A subexponential bound for linear programming, Algorithmica 16 (4–5) (1996) 498–516.

[4] R. Nock, F. Nielsen, Fitting the smallest enclosing Bregman ball, in: European Conference on Machine Learning, in: Lecture Notes in Computer Science, vol. 3720, Springer, 2005, pp. 649–656.

[5] E. Welzl, Smallest enclosing disks (balls and ellipsoids), in: New Results and New Trends in Computer Science, in: Lecture Notes in Computer Science, vol. 555, Springer, 1991, pp. 359–370.