

Article

On Clustering Histograms with k -Means by Using Mixed α -Divergences

Frank Nielsen ^{1,2,*}, Richard Nock ³ and Shun-ichi Amari ⁴

¹ Sony Computer Science Laboratories, Inc, Tokyo 141-0022, Japan

² École Polytechnique, 91128 Palaiseau Cedex, France

³ NICTA and The Australian National University, Locked Bag 9013, Alexandria NSW 1435, Australia

⁴ RIKEN Brain Science Institute, 2-1 Hirosawa Wako City, Saitama 351-0198, Japan;

E-Mail: amari@brain.riken.jp

* Author to whom correspondence should be addressed; E-Mail: Frank.Nielsen@acm.org;
Tel.:+81-3-5448-4380.

Received: 15 May 2014; in revised form: 10 June 2014 / Accepted: 13 June 2014 /

Published: 17 June 2014

Abstract: Clustering sets of histograms has become popular thanks to the success of the generic method of bag-of-X used in text categorization and in visual categorization applications. In this paper, we investigate the use of a parametric family of distortion measures, called the α -divergences, for clustering histograms. Since it usually makes sense to deal with symmetric divergences in information retrieval systems, we symmetrize the α -divergences using the concept of mixed divergences. First, we present a novel extension of k -means clustering to mixed divergences. Second, we extend the k -means++ seeding to mixed α -divergences and report a guaranteed probabilistic bound. Finally, we describe a soft clustering technique for mixed α -divergences.

Keywords: bag-of-X; α -divergence; Jeffreys divergence; centroid; k -means clustering; k -means seeding

1. Introduction: Motivation and Background

1.1. Clustering Histograms in the Bag-of-Word Modeling Paradigm

A common task of information retrieval (IR) systems is to classify documents into categories. Given a training set of documents labeled with categories, one asks to classify new incoming documents. Text

categorisation [1,2] proceeds by first defining a dictionary of words from a corpus. It then models each document by a word count yielding a word distribution histogram per document (see the University of California, Irvine, UCI, machine learning repository for such data-sets [3]). The importance of the words in the dictionary can be weighted by the term frequency-inverse document frequency [2] (tf-idf) that takes into account both the frequency of the words in a given document, but also of the frequency of the words in all documents: Namely, the tf-idf weight for a given word in a given document is the product of the frequency of that word in the document times the logarithm of the ratio of the number of documents divided by the document frequency of the word [2]. Defining a proper distance between histograms allows one to:

- Classify a new on-line document: We first calculate its word distribution histogram signature and seek for the labeled document, which has the most similar histogram to deduce its category tag.
- Find the initial set of categories: we cluster all document histograms and assign a category per cluster.

This text classification method based on the representation of the bag-of -words (BoWs) has also been instrumental in computer vision for efficient object categorization [4] and recognition in natural images [5]. This paradigm is called bag-of-features [6] (BoFs) in the general case. It first requires one to create a dictionary of “visual words” by quantizing keypoints (e.g., affine invariant descriptors of image patches) of the training database. Quantization is performed using the k -means [7–9] algorithm that partitions n data $\mathcal{X} = \{x_1, \dots, x_n\}$ into k pairwise disjoint clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$, where each data element belongs to the closest cluster center (*i.e.*, the cluster prototype). From a given initialization, batched k -means first assigns data points to their closest centers and then updates the cluster centers and reiterates this process until convergence is met to a local minimum (not necessarily the global minimum) after a provably finite number of steps. Csurka *et al.* [4] used the squared Euclidean distance for building the visual vocabulary. Depending on the chosen features, other distances have proven useful. For example, the symmetrized Kullback–Leibler (KL) divergence was shown to perform experimentally better than the Euclidean or squared Euclidean distances for a compressed histogram of gradient descriptors [10] (CHoGs), even if it is not a metric distance, since it fails to satisfy the triangular inequality. To summarize, k -means histogram clustering with respect to the symmetrized KL (called Jeffreys divergence J) can be used to quantize both visual words and document categories. Nowadays, the seminal bag-of-word method has been generalized fruitfully to various settings using the generic bag-of-X paradigm, like the bag-of-textons [6], the bag-of-readers [11], *etc.* Bag-of-X represents each data (e.g., document, image, *etc.*) as an histogram of codeword count indices. Furthermore, the semantic space [12] paradigm has been recently explored to overcome two drawbacks of the bag-of-X paradigms: the high-dimensionality of the histograms (number of bins) and difficult human interpretation of the codewords due to the lack of semantic information. In semantic space, modeling relies on semantic multinomials that are discrete frequency histograms; see [12].

In summary, clustering histograms with respect to symmetric distances (like the symmetrized KL divergence) is playing an increasing role. It turns out that the symmetrized KL divergence belongs to a 1-parameter family of divergences, called symmetrized α -divergences, or Jeffreys α -divergence [13].

1.2. Contributions

Since divergences $D(p : q)$ are usually asymmetric distortion measures between two objects p and q , one has to often consider two kinds of centroids obtained by carrying the minimization process either on the left argument or on the right argument of the divergences; see [14]. In theory, it is enough to consider only one type of centroid, say the right centroid, since the left centroid with respect to a divergence $D(p : q)$ is equivalent to the right centroid with respect to the mirror divergence $D'(p : q) = D(q : p)$. In this paper, we consider mixed divergences [15] that allow one to handle in a unified way the arithmetic symmetrization $S(p, q) = \frac{1}{2}(D(p : q) + D(q : p))$ of a given divergence $D(p : q)$ with both the sided divergences: $D(p : q)$ and its mirror divergence $D'(p : q)$. The mixed α -divergence is the mixed divergence obtained for the α -divergence. We term α -clustering the clustering with respect to α -divergences and mixed α -clustering the clustering w.r.t. mixed α -divergences [16]. Our main contributions are to extend the celebrated batched k -means [7–9] algorithm to mixed divergences by associating two dual centroids per cluster and to generalize the probabilistically guaranteed good seeding of k -means++ [17] to mixed α -divergences. The mixed α -seedings provide guaranteed probabilistic clustering bounds by picking up seeds from the data and do not require explicitly computing of centroids. Therefore, it follows a fast clustering technique in practice, even when cluster centers are not available in closed form. We also consider clustering histograms by explicitly building the symmetrized α -centroids and end up with a variational k -means when the centroids are not available in closed-form. Finally, we investigate soft mixed α -clustering and discuss topics related to α -clustering. Note that clustering with respect to non-symmetrized α -divergences has been recently investigated independently in [18] and proven useful in several applications.

1.3. Outline of the Paper

The paper is organized as follows: Section 2 introduces the notion of mixed divergences, presents an extension of k -means to mixed divergences and recalls some properties of α -divergences. Section 3 describes the α -seeding techniques and reports a probabilistically-guaranteed bound on the clustering quality. Section 4 investigates the various sided/symmetrized/mixed calculations of the α -centroids. Section 5 presents the soft α -clustering with respect to α -mixed divergences. Finally, Section 6 summarises the contributions, discusses related topics and hints at further perspectives. The paper is followed by two appendices. Appendix B studies several properties of α -divergences that are used to derive the guaranteed probabilistic performance of the α -seeding. Appendix C proves that α -sided centroids are quasi-arithmetic means for the power generator functions.

2. Mixed Centroid-Based k -Means Clustering

2.1. Divergences, Centroids and k -Means

Consider a set \mathcal{H} of n histograms h_1, \dots, h_n , each with d bins, with all positive real-valued bins: $h_j^i > 0, \forall 1 \leq i \leq d, 1 \leq j \leq n$. A histogram h is called a frequency histogram when its bins sums

up to one: $w(h) = w_h = \sum_i h^i = 1$. Otherwise, it is called a positive histogram that can eventually be normalized to a frequency histogram:

$$\tilde{h} \doteq \frac{h}{w(h)}. \tag{1}$$

The frequency histograms belong to the $(d-1)$ -dimensional open probability simplex Δ_d :

$$\Delta_d \doteq \left\{ (x^1, \dots, x^d) \in \mathbb{R}^d \mid \forall i, x^i > 0, \text{ and } \sum_{i=1}^d x^i = 1 \right\}. \tag{2}$$

That is, although frequency histograms have d bins, the constraint that those bin values should sum up to one yields $d-1$ degrees of freedom. In probability theory, the frequency or counting of histograms either model discrete multinomial probabilities or discrete positive measures (also called positive arrays [19]).

The celebrated k -means clustering [8,9] is one of the most famous clustering techniques that has been generalized in many ways [20,21]. In information geometry [22], a divergence $D(p : q)$ is a smooth C^3 differentiable dissimilarity measure that is not necessarily symmetric ($D(p : q) \neq D(q : p)$, hence the notation “:” instead of the classical “,” reserved for metric distances), but is non-negative and satisfies the separability property: $D(p : q) = 0$ iff $p = q$. More precisely, let $\partial_i D(x : y) = \frac{\partial}{\partial x^i} D(x : y)$, $\partial_{,i} D(x : y) = \frac{\partial}{\partial y^i} D(x : y)$. Then, we require $\partial_i D(x : x) = \partial_{,i} D(x : x) = 0$ and $-\partial_i \partial_{,j} D(x : y)$ positive definite for defining a divergence. For a distance function $D(\cdot : \cdot)$, we denote by $D(x : \mathcal{H})$ the weighted average distance of x to a set a weighted histograms:

$$D(x : \mathcal{H}) \doteq \sum_{j=1}^n w_j D(x : h_j). \tag{3}$$

An important class of divergences on frequency histograms is the f -divergences [23–25] defined for a convex generator f (with $f(1) = f'(1) = 0$ and $f''(1) = 1$):

$$I_f(p : q) \doteq \sum_{i=1}^d q^i f\left(\frac{p^i}{q^i}\right).$$

Those divergences preserve information monotonicity [19] under any arbitrary transition probability (Markov morphisms). f -divergences can be extended to positive arrays [19].

The k -means algorithm on a set of weighted histograms can be tailored to any divergence as follows: First, we initialize the k cluster centers $\mathcal{C} = \{c_1, \dots, c_k\}$ (say, by picking up randomly arbitrary distinct seeds). Then, we iteratively repeat until convergence the following two steps:

- Assignment: Assign each histogram h_j to its closest cluster center:

$$l(h_j) \doteq \arg \min_{l=1}^k D(h_j : c_l).$$

This yields a partition of the histogram set $\mathcal{H} = \cup_{l=1}^k \mathcal{A}_l$, where \mathcal{A}_l denotes the set of histograms of the l -th cluster: $\mathcal{A}_l = \{h_j \mid l(h_j) = l\}$.

- Center relocation: Update the cluster centers by taking their centroids:

$$c_l \doteq \arg \min_x \sum_{h_j \in \mathcal{A}_l} w_j D(h_j : x).$$

Throughout this paper, centroid shall be understood in the broader sense of a barycenter when weights are non-uniform.

2.2. Mixed Divergences and Mixed k -Means Clustering

Since divergences are potentially asymmetric, we can define two-sided k -means or always consider a right-sided k -means, but then define another sided divergence $D'(p : q) = D(q : p)$. We can also consider the symmetrized k -means with respect to the symmetrized divergence: $S(p, q) = D(p : q) + D(q : p)$. Eventually, we may skew the symmetrization with a parameter $\lambda \in [0, 1]$: $S_\lambda(p, q) = \lambda D(p : q) + (1 - \lambda) D(q : p)$ (and consider other averaging schemes instead of the arithmetic mean).

In order to handle those sided and symmetrized k -means under the same framework, let us introduce the notion of mixed divergences [15] as follows:

Definition 1 (Mixed divergence).

$$M_\lambda(p : q : r) \doteq \lambda D(p : q) + (1 - \lambda) D(q : r), \tag{4}$$

for $\lambda \in [0, 1]$.

A mixed divergence includes the sided divergences for $\lambda \in \{0, 1\}$ and the symmetrized (arithmetic mean) divergence for $\lambda = \frac{1}{2}$.

We generalize k -means clustering to mixed k -means clustering [15] by considering two centers per cluster (for the special cases of $\lambda = 0, 1$, it is enough to consider only one). Algorithm 1 sketches the generic mixed k -means algorithm. Note that a simple initialization consists of choosing randomly the k distinct seeds from the dataset with $l_i = r_i$.

Algorithm 1: Mixed divergence-based k -means clustering.

Input: Weighted histogram set \mathcal{H} , divergence $D(\cdot, \cdot)$, integer $k > 0$, real $\lambda \in [0, 1]$;

Initialize left-sided/right-sided seeds $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k$;

repeat

 //Assignment

for $i = 1, 2, \dots, k$ **do**

$\mathcal{C}_i \leftarrow \{h \in \mathcal{H} : i = \arg \min_j M_\lambda(l_j : h : r_j)\}$;

 // Dual-sided centroid relocation

for $i = 1, 2, \dots, k$ **do**

$r_i \leftarrow \arg \min_x D(\mathcal{C}_i : x) = \sum_{h \in \mathcal{C}_i} w_j D(h : x)$;

$l_i \leftarrow \arg \min_x D(x : \mathcal{C}_i) = \sum_{h \in \mathcal{C}_i} w_j D(x : h)$;

until convergence;

Output: Partition of \mathcal{H} into k clusters following \mathcal{C} ;

Notice that the mixed k -means clustering is different from the k -means clustering with respect to the symmetrized divergences S_λ that considers only one centroid per cluster.

2.3. Sided, Symmetrized and Mixed α -Divergences

For $\alpha \neq \pm 1$, we define the family of α -divergences [26] on positive arrays [27] as:

$$\begin{aligned}
 D_\alpha(p : q) &\doteq \sum_{i=1}^d \frac{4}{1-\alpha^2} \left(\frac{1-\alpha}{2} p^i + \frac{1+\alpha}{2} q^i - (p^i)^{\frac{1-\alpha}{2}} (q^i)^{\frac{1+\alpha}{2}} \right), \\
 &= D_{-\alpha}(q : p), \alpha \in \mathbb{R} \setminus \{0, 1\},
 \end{aligned}
 \tag{5}$$

with the limit cases $D_{-1}(p : q) = \text{KL}(p : q)$ and $D_1(p : q) = \text{KL}(q : p)$, where KL is the extended Kullback–Leibler divergence:

$$\text{KL}(p : q) \doteq \sum_{i=1}^d p^i \log \frac{p^i}{q^i} + q^i - p^i.
 \tag{6}$$

Divergence D_0 is the squared Hellinger symmetric distance (scaled by a multiplicative factor of four) extended to positive arrays:

$$D_0(p : q) = 2 \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx = 4H^2(p, q),
 \tag{7}$$

with the Hellinger distance:

$$H(p, q) = \sqrt{\frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx}.
 \tag{8}$$

Note that α -divergences are defined for the full range of α values: $\alpha \in \mathbb{R}$. Observe that α -divergences of Equation (5) are homogeneous of degree one: $D_\alpha(\lambda p : \lambda q) = \lambda D_\alpha(p : q)$ for $\lambda > 0$.

When histograms p and q are both frequency histograms, we have:

$$\begin{aligned}
 D_\alpha(\tilde{p} : \tilde{q}) &= \frac{4}{1-\alpha^2} \left(1 - \sum_{i=1}^d (\tilde{p}^i)^{\frac{1-\alpha}{2}} (\tilde{q}^i)^{\frac{1+\alpha}{2}} \right), \\
 &= D_{-\alpha}(\tilde{q} : \tilde{p}), \alpha \in \mathbb{R} \setminus \{0, 1\},
 \end{aligned}
 \tag{9}$$

and the extended Kullback–Leibler divergence reduces to the traditional Kullback–Leibler divergence: $\text{KL}(\tilde{p} : \tilde{q}) = \sum_{i=1}^d \tilde{p}^i \log \frac{\tilde{p}^i}{\tilde{q}^i}$.

The Kullback–Leibler divergence between frequency histograms \tilde{p} and \tilde{q} ($\alpha = \pm 1$) is interpreted as the cross-entropy minus the Shannon entropy:

$$\text{KL}(\tilde{p} : \tilde{q}) \doteq H^\times(\tilde{p} : \tilde{q}) - H(\tilde{p}).$$

Often, \tilde{p} denotes the true model (hidden by nature), and \tilde{q} is the estimated model from observations. However, in information retrieval, both \tilde{p} and \tilde{q} play the same symmetrical role, and we prefer to deal with a symmetric divergence.

The Pearson and Neyman χ^2 distances are obtained for $\alpha = -3$ and $\alpha = 3$, respectively:

$$D_3(\tilde{p} : \tilde{q}) = \frac{1}{2} \sum_i \frac{(\tilde{q}^i - \tilde{p}^i)^2}{\tilde{p}^i},
 \tag{10}$$

$$D_{-3}(\tilde{p} : \tilde{q}) = \frac{1}{2} \sum_i \frac{(\tilde{q}^i - \tilde{p}^i)^2}{\tilde{q}^i}.
 \tag{11}$$

The α -divergences belong to the class of Csiszár f -divergences with the following generator:

$$f(t) = \begin{cases} \frac{4}{1-\alpha^2}(1 - t^{(1+\alpha)/2}), & \text{if } \alpha \neq \pm 1, \\ t \ln t, & \text{if } \alpha = 1, \\ -\ln t, & \text{if } \alpha = -1 \end{cases} \tag{12}$$

Remark 1. Historically, the α -divergences have been introduced by Chernoff [28,29] in the context of hypothesis testing. In Bayesian binary hypothesis testing, we are asked to decide whether an observation belongs to one class or the other class, based on prior w_1 and w_2 and class-conditional probabilities p_1 and p_2 . The average expected error of the best decision maximum a posteriori (MAP) rule is called the probability of error, denoted by P_e . When prior probabilities are identical ($w_1 = w_2 = \frac{1}{2}$), we have $P_e(p_1, p_2) = \frac{1}{2} \int \min(p_1(x), p_2(x))dx$. Let $S(p, q) = \int \min(p(x), q(x))dx$ denote the intersection similarity measure, with $0 < S \leq 1$ (generalizing the histogram intersection distance often used in computer vision [30]). S is bounded by the α -Chernoff affinity coefficient:

$$S(p, q) \leq C_\beta(p, q) = \int p^\beta(x)q^{1-\beta}(x)dx,$$

for all $\beta \in [0, 1]$. We can convert the affinity coefficient $0 < C_\beta \leq 1$ into a divergence D_β by simply taking $D_\beta = 1 - C_\beta$. Since the absolute value of divergences does not matter, we can rescale appropriately the divergence. One nice rescaling is by multiplying by $\frac{1}{\beta(1-\beta)}$: $D_\beta = \frac{1}{\beta(1-\beta)}(1 - C_\beta)$. This lets coincide the parameterized divergence with the fundamental Kullback–Leibler divergence for the limit values $\beta \in \{0, 1\}$. Last, by choosing $\beta = \frac{1-\alpha}{2}$, it yields the well-known expression of the α -divergences.

Interestingly, the α -divergences can be interpreted as a generalized α -Kullback–Leibler divergence [26] with deformed logarithms.

Next, we introduce the mixed α -divergence of a histogram x to two histograms p and q as follows:

Definition 2 (Mixed α -divergence). The mixed α -divergence of a histogram x to two histograms p and q is defined by:

$$\begin{aligned} M_{\lambda,\alpha}(p : x : q) &= \lambda D_\alpha(p : x) + (1 - \lambda)D_\alpha(x : q), \\ &= \lambda D_{-\alpha}(x : p) + (1 - \lambda)D_{-\alpha}(q : x), \\ &= M_{1-\lambda,-\alpha}(q : x : p), \end{aligned} \tag{13}$$

The α -Jeffreys symmetrized divergence is obtained for $\lambda = \frac{1}{2}$:

$$S_\alpha(p, q) = M_{\frac{1}{2},\alpha}(q : p : q) = M_{\frac{1}{2},\alpha}(p : q : p).$$

The skew symmetrized α -divergence is defined by:

$$S_{\lambda,\alpha}(p : q) = \lambda D_\alpha(p : q) + (1 - \lambda)D_\alpha(q : p).$$

2.4. Notations and Hard/Soft Clusterings

Throughout the paper, superscript index i denotes the histogram bin numbers and subscript index j the histogram numbers. Index l is used to iterate on the clusters. The left-sided, right-sided and symmetrized histogram positive and frequency α -centroids are denoted by $l_\alpha, r_\alpha, s_\alpha$ and $\tilde{l}_\alpha, \tilde{r}_\alpha, \tilde{s}_\alpha$, respectively.

In this paper, we investigate the following kinds of clusterings for sets of histograms:

Hard clustering. Each histogram belongs to exactly one cluster:

- k -means with respect to mixed divergences $M_{\lambda,\alpha}$.
- k -means with respect to symmetrized divergences $S_{\lambda,\alpha}$.
- Randomized seeding for mixed/symmetrized k -means by extending k -means++ with guaranteed probabilistic bounds for α -divergences.

Soft clustering. Each histogram belongs to all clusters according to some weight distribution: the soft mixed α -clustering.

3. Coupled k -Means++ α -Seeding

It is well-known that the Lloyd k -means clustering algorithm monotonically decreases the loss function and stops after a finite number of iterations into a local optimal. Optimizing globally the k -means loss is NP-hard [17] when $d > 1$ and $k > 1$. In practice, the performance of the k -means algorithm heavily relies on the initialization. A breakthrough was obtained by the k -means++ seeding [17], which guarantees in expectation a good starting partition. We extend this scheme to the coupled α -clustering. However, we point out that although k -means++ prove popular and are often used in practice with very good results; it has been recently pointed out that “worst case” configurations exist and even in small dimensions, on which the algorithm cannot beat significantly its expected approximability with a high probability [31]. Still, the expected approximability ratio, roughly in $O(\log(k))$, is very good, as long as the number of clusters is not too large.

Algorithm 2: Mixed α -seeding; MAS($\mathcal{H}, k, \lambda, \alpha$)

Input: Weighted histogram set \mathcal{H} , integer $k \geq 1$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} \leftarrow h_j$ with uniform probability ;

for $i = 2, 3, \dots, k$ **do**

Pick at random histogram $h \in \mathcal{H}$ with probability:

$$\pi_{\mathcal{H}}(h) \doteq \frac{w_h M_{\lambda,\alpha}(c_h : h : c_h)}{\sum_{y \in \mathcal{H}} w_y M_{\lambda,\alpha}(c_y : y : c_y)} , \tag{14}$$

//where $(c_h, c_h) \doteq \arg \min_{(z,z) \in \mathcal{C}} M_{\lambda,\alpha}(z : h : z)$;

$\mathcal{C} \leftarrow \mathcal{C} \cup \{(h, h)\}$;

Output: Set of initial cluster centers \mathcal{C} ;

Algorithm 2 provides our adaptation of k -means++ seeding [15,17]. It works for all three of our sided/symmetrized and mixed clustering settings:

- Pick $\lambda = 1$ for the left-sided centroid initialization,
- Pick $\lambda = 0$ for the right-sided centroid initialization (a left-sided initialization for $-\alpha$),
- with arbitrary λ , for the λ - J_α (skew Jeffreys) centroids or mixed λ centroids. Indeed, the initialization is the same (see the MAS procedure in Algorithm 2).

Our proof follows and generalizes the proof described for the case of mixed Bregman seeding [15] (Lemma 2). In fact, our proof is more precise, as it quantifies the expected potential with respect to the optimum only, whereas in [15], the optimal potential is averaged with a dual optimal potential, which depends on the optimal centers, but may be larger than the optimum sought.

Theorem 1. *Let $C_{\lambda,\alpha}$ denote for short the cost function related to the clustering type chosen (left-, right-, skew Jeffreys or mixed) in MAS and $C_{\lambda,\alpha}^{opt}$ denote the optimal related clustering in k clusters, for $\lambda \in [0, 1]$ and $\alpha \in (-1, 1)$. Then, on average, with respect to distribution (14), the initial clustering of MAS satisfies:*

$$E_\pi[C_{\lambda,\alpha}] \leq 4 \begin{cases} f(\lambda)g(k)h^2(\alpha)C_{\lambda,\alpha}^{opt} & \text{if } \lambda \in (0, 1) \\ g(k)z(\alpha)h^4(\alpha)C_{\lambda,\alpha}^{opt} & \text{otherwise} \end{cases} \quad (15)$$

Here, $f(\lambda) = \max\{\frac{1-\lambda}{\lambda}, \frac{\lambda}{1-\lambda}\}$, $g(k) = 2(2 + \log k)$, $z(\alpha) = \left(\frac{1+|\alpha|}{1-|\alpha|}\right)^{\frac{8|\alpha|^2}{(1-|\alpha|)^2}}$, $h(\alpha) = \max_i p_i^{|\alpha|} / \min_i p_i^{|\alpha|}$; the min is defined on strictly positive coordinates, and π denotes the picking distribution of Algorithm 2.

Remark 2. *The bound is particularly good when λ is close to 1/2, and in particular for the α -Jeffreys clustering, as in these cases, the only additional penalty compared to the Euclidean case [17] is $h^2(\alpha)$, a penalty that relies on an optimal triangle inequality for α -divergences that we provide in Lemma 8 below.*

Remark 3. *This guaranteed initialization is particularly useful for α -Jeffreys clustering, as there is no closed form solution for the centroids (except when $\alpha = \pm 1$, see [32]).*

Algorithm 3: Mixed α -hard clustering: MAhC(\mathcal{H} , k , λ , α)

Input: Weighted histogram set \mathcal{H} , integer $k > 0$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \text{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

//Assignment

for $i = 1, 2, \dots, k$ **do**

$\mathcal{A}_i \leftarrow \{h \in \mathcal{H} : i = \arg \min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

// Centroid relocation

for $i = 1, 2, \dots, k$ **do**

$r_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

$l_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

until convergence;

Output: Partition of \mathcal{H} in k clusters following \mathcal{C} ;

Algorithm 3 presents the general hard mixed k -means clustering, which can be adapted also to left- ($\lambda = 1$) and right- ($\lambda = 0$) α -clustering.

For skew Jeffreys centers, since the centroids are not available in closed form [32], we adopt a variational approach of k -means by updating iteratively the centroid in each cluster (thus improving the overall loss function without computing the optimal centroids that would eventually require infinitely many iterations).

4. Sided, Symmetrized and Mixed α -Centroids

The k -means clustering requires assigning data elements to their closest cluster center and then updating those cluster centers by taking their centroids. This section investigates the centroid computations for the sided, symmetrized and mixed α -divergences.

Note that the mixed α -seeding presented in Section 3 does not require computing centroids and, yet, guarantees probabilistically a good clustering partition.

Since mixed α -divergences are f -divergences, we start with the generic f -centroids.

4.1. Csiszár f -Centroids

The centroids induced by f -divergences of a set of positive measures (that relaxes the normalisation constraint) have been studied by Ben-Tal *et al.* [33]. Those entropic centroids are shown to be unique, since f -divergences are convex statistical distances in both arguments. Let E_f denote the energy to minimize when considering f -divergences:

$$E_f \doteq \min_{x \in \mathcal{X}} I_f(\mathcal{H} : x) = \sum_{j=1}^n w_j I_f(h_j : x), \tag{16}$$

$$= \min_{x \in \mathcal{X}} \sum_{j=1}^n w_j \sum_{i=1}^d p_j^i f\left(\frac{c^i}{h_j^i}\right). \tag{17}$$

When the domain is the open probability simplex $\mathcal{X} = \Delta_d$, we get a constrained optimisation problem to solve. We transform this constrained minimisation problem (*i.e.*, $x \in \Delta_d$) into an equivalent unconstrained minimisation problem by using the Lagrange multiplier, γ :

$$\min_{x \in \mathbb{R}^d} \sum_{j=1}^n w_j I_f(h_j : c) + \gamma \left(\sum_{i=1}^d x^i - 1 \right). \tag{18}$$

Taking the derivatives according to x^i , we get:

$$\forall i \in \{1, \dots, d\}, \sum_{j=1}^n w_j f'\left(\frac{x^i}{h_j^i}\right) - \gamma = 0. \tag{19}$$

We now consider this equation for α -divergences and symmetrized α -divergences, both f -divergences.

4.2. Sided Positive and Frequency α -Centroids

The positive sided α -centroids for a set of weighted histograms were reported in [34] using the representation Bregman divergence. We summarise the results in the following theorem:

Theorem 2 (Sided positive α -centroids [34]). *The left-sided l_α and right-sided r_α positive weighted α -centroid coordinates of a set of n positive histograms h_1, \dots, h_n are weighted α -means:*

$$r_\alpha^i = f_\alpha^{-1} \left(\sum_{j=1}^n w_j f_\alpha(h_j^i) \right), l_\alpha^i = r_{-\alpha}^i$$

with $f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & \alpha \neq \pm 1, \\ \log x & \alpha = 1. \end{cases}$

Furthermore, the frequency-sided α -centroids are simply the normalized-sided α -centroids.

Theorem 3 (Sided frequency α -centroids [16]). *The coordinates of the sided frequency α -centroids of a set of n weighted frequency histograms are the normalised weighted α -means.*

Table 1 summarizes the results concerning the sided positive and frequency α -centroids.

Table 1. Positive and frequency α -centroids: the frequency α -centroids are normalized positive α -centroids, where $w(h)$ denotes the cumulative sum of the histogram bins. The arithmetic mean is obtained for $r_{-1} = l_1$ and the geometric mean for $r_1 = l_{-1}$.

	Positive centroid	Frequency centroid
Right-sided centroid	$r_\alpha^i = \begin{cases} (\sum_{j=1}^n w_j (h_j^i)^{\frac{1-\alpha}{2}})^{\frac{2}{1-\alpha}} & \alpha \neq 1 \\ r_1^i = \prod_{j=1}^n (h_j^i)^{w_j} & \alpha = 1 \end{cases}$	$\tilde{r}_\alpha^i = \frac{r_\alpha^i}{w(\tilde{r}_\alpha)}$
Left-sided centroid	$l_\alpha^i = r_{-\alpha}^i = \begin{cases} (\sum_{j=1}^n w_j (h_j^i)^{\frac{1+\alpha}{2}})^{\frac{2}{1+\alpha}} & \alpha \neq -1 \\ l_{-1}^i = \prod_{j=1}^n (h_j^i)^{w_j} & \alpha = -1 \end{cases}$	$\tilde{l}_\alpha^i = \tilde{r}_{-\alpha}^i = \frac{r_{-\alpha}^i}{w(\tilde{r}_{-\alpha})}$

4.3. Mixed α -Centroids

The mixed α -centroids for a set of n weighted histograms is defined as the minimizer of:

$$\sum_j w_j M_{\lambda,\alpha}(l : h_j : r). \tag{20}$$

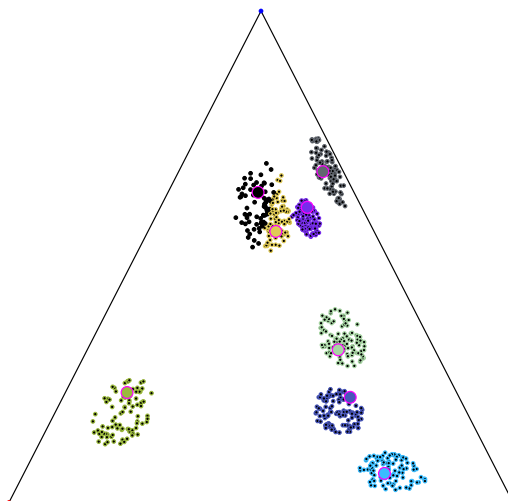
We state the theorem generalizing [15]:

Theorem 4. *The two mixed α -centroids are the left-sided and right-sided α -centroids.*

Figure 1 depicts some clustering result with our α -clustering software. We remark that the clusters found are all approximately subclusters of the “distinct” clusters that appear on the figure. When those distinct clusters are actually the optimal clusters—which is likely to be the case when they are separated by large minimal distance to other clusters—this is clearly a desirable qualitative property as long as the

number of experimental clusters is not too large compared to the number of optimal clusters. We remark also that in the experiment displayed, there is no closed form solution for the cluster centers.

Figure 1. Snapshot of the α -clustering software. Here, $n = 800$ frequency histograms of three bins with $k = 8$, and $\alpha = 0.7$ and $\lambda = \frac{1}{2}$.



4.4. Symmetrized Jeffreys-Type α -Centroids

The Kullback–Leibler divergence can be symmetrized in various ways: Jeffreys divergence, Jensen–Shannon divergence and Chernoff information, just to mention a few. Here, we consider the following symmetrization of α -divergences extending Jeffreys J -divergence:

$$\begin{aligned}
 S_\alpha(p, q) &= \frac{1}{2} (D_\alpha(p : q) + D_\alpha(q : p)) = S_{-\alpha}(p, q), \\
 &= M_{\frac{1}{2}}(p : q : p),
 \end{aligned}
 \tag{21}$$

For $\alpha = \pm 1$, we get half of Jeffreys divergence:

$$S_{\pm 1}(p, q) = \frac{1}{2} \sum_{i=1}^d (p^i - q^i) \log \frac{p^i}{q^i}$$

In particular, when p and q are frequency histograms, we have for $\alpha \neq \pm 1$:

$$J_\alpha(\tilde{p} : \tilde{q}) = \frac{8}{1 - \alpha^2} \left(1 + \sum_{i=1}^d H_{\frac{1-\alpha}{2}}(\tilde{p}^i, \tilde{q}^i) \right),
 \tag{22}$$

where $H_{\frac{1-\alpha}{2}}(a, b)$ a symmetric Heinz mean [35,36]:

$$H_\beta(a, b) = \frac{a^\beta b^{1-\beta} + a^{1-\beta} b^\beta}{2}.$$

Heinz means interpolate the arithmetic and geometric means and satisfies the inequality:

$$\sqrt{ab} = H_{\frac{1}{2}}(a, b) \leq H_\alpha(a, b) \leq H_0(a, b) = \frac{a + b}{2}.$$

(Another interesting property of Heinz means is the integral representation of the logarithmic mean: $L(x, y) = \frac{x-y}{\log x - \log y} = \int_0^1 H_\beta(x, y) d\beta$. This allows one to prove easily that $\sqrt{xy} \leq L(x, y) \leq \frac{x+y}{2}$.)

The J_α -divergence is a Csiszár f -divergence [24,25].

Observe that it is enough to consider $\alpha \in [0, \infty)$ and that the symmetrized α -divergence for positive and frequency histograms coincide only for $\alpha = \pm 1$.

For $\alpha = \pm 1$, $S_\alpha(p, q)$ tends to the Jeffreys divergence:

$$J(p, q) = \text{KL}(p, q) + \text{KL}(q, p) = \sum_{i=1}^d (p^i - q^i)(\log p^i - \log q^i). \tag{23}$$

The Jeffreys divergence writes mathematically the same for frequency histograms:

$$J(\tilde{p}, \tilde{q}) = \text{KL}(\tilde{p}, \tilde{q}) + \text{KL}(\tilde{q}, \tilde{p}) = \sum_{i=1}^d (\tilde{p}^i - \tilde{q}^i)(\log \tilde{p}^i - \log \tilde{q}^i). \tag{24}$$

We state the results reported in [32]:

Theorem 5 (Jeffreys positive centroid [32]). *The Jeffreys positive centroid $c = (c^1, \dots, c^d)$ of a set $\{h_1, \dots, h_n\}$ of n weighted positive histograms with d bins can be calculated component-wise exactly using the Lambert W analytic function:*

$$c^i = \frac{a^i}{W(\frac{a^i}{g^i}e)},$$

where $a^i = \sum_{j=1}^n \pi_j h_j^i$ denotes the coordinate-wise arithmetic weighted means and $g^i = \prod_{j=1}^n (h_j^i)^{\pi_j}$ the coordinate-wise geometric weighted means.

The Lambert analytic function W [37] (positive branch) is defined by $W(x)e^{W(x)} = x$ for $x \geq 0$.

Theorem 6 (Jeffreys frequency centroid [32]). *Let \tilde{c} denote the Jeffreys frequency centroid and $\tilde{c}' = \frac{c}{w_c}$ the normalised Jeffreys positive centroid. Then, the approximation factor $\alpha_{\tilde{c}'} = \frac{S_1(\tilde{c}', \tilde{\mathcal{H}})}{S_1(\tilde{c}, \tilde{\mathcal{H}})}$ is such that $1 \leq \alpha_{\tilde{c}'} \leq \frac{1}{w_c}$ (with $w_c \leq 1$).*

Therefore, we shall consider $\alpha \neq \pm 1$ in the remainder.

We state the following lemma generalizing the former results in [38] that were tailored to the symmetrized Kullback–Leibler divergence or the symmetrized Bregman divergence [14]:

Lemma 1 (Reduction property). *The symmetrized J_α -centroid of a set of n weighted histograms amount to computing the symmetrized α -centroid for the weighted α -mean and $-\alpha$ -mean:*

$$\min_x J_\alpha(x, \mathcal{H}) = \min_x (D_\alpha(x : r_\alpha) + D_\alpha(l_\alpha : x)).$$

Proof. It follows that the minimization problem $\min_x S_\alpha(x, \mathcal{H}) = \sum_{j=1}^n w_j S_\alpha(x, h_j)$ reduces to the following minimization:

$$\min \sum_{i=1}^d x^i - (x^i)^{\frac{1+\alpha}{2}} \bar{h}_\alpha^i - (x^i)^{\frac{1-\alpha}{2}} \bar{h}_{-\alpha}^i. \tag{25}$$

This is equivalent to minimizing:

$$\begin{aligned} &\equiv \sum_{i=1}^d x^i - (x^i)^{\frac{1+\alpha}{2}} ((\bar{h}_\alpha^i)^{\frac{2}{1-\alpha}})^{\frac{1-\alpha}{2}} - \\ &\quad (x^i)^{\frac{1-\alpha}{2}} ((\bar{h}_{-\alpha}^i)^{\frac{2}{1+\alpha}})^{\frac{1+\alpha}{2}}, \\ &\equiv \sum_{i=1}^d x^i - (x^i)^{\frac{1+\alpha}{2}} (r_\alpha^i)^{\frac{1-\alpha}{2}} - (x^i)^{\frac{1-\alpha}{2}} (l_\alpha^i)^{\frac{1+\alpha}{2}} \\ &\equiv D_\alpha(x : r_\alpha) + D_\alpha(l_\alpha : x). \end{aligned}$$

Note that $\alpha = \pm 1$, the lemma states that the minimization problem is equivalent to minimizing $KL(a : x) + KL(x : g)$ with respect to x , where $a = l_1$ and $g = r_1$ denote the arithmetic and geometric means, respectively. \square

The lemma states that the optimization problem with n weighted histograms is equivalent to the optimization with only two equally weighted histograms.

The positive symmetrized α -centroid is equivalent to computing a representation symmetrized Bregman centroid [14,34].

The frequency symmetrized α -centroid asks to minimize the following problem:

$$\min_{\tilde{x} \in \Delta_d} \sum_j w_j S_\alpha(\tilde{x}, \tilde{h}_j).$$

Instead of seeking for \tilde{x} in the probability simplex, we can optimize on the unconstrained domain \mathbb{R}^{d-1} by using a reparameterization. Indeed, frequency histograms belong to the exponential families [39] (multinomials).

Exponential families also include many other continuous distributions, like the Gaussian, Beta or Dirichlet distributions. It turns out the α -divergences can be computed in closed-form for members of the same exponential family:

Lemma 2. *The α -divergence for distributions belonging to the same exponential families amounts to computing a divergence on the corresponding natural parameters:*

$$A_\alpha(p : q) = \frac{4}{1 - \alpha^2} \left(1 - e^{-J_F^{(\frac{1-\alpha}{2})}(\theta_p:\theta_q)} \right),$$

where $J_F^\beta(\theta_1 : \theta_2) = \beta F(\theta_1) + (1 - \beta)F(\theta_2) - F(\beta\theta_1 + (1 - \beta)\theta_2)$ is a skewed Jensen divergence defined for the log-normaliser F of the family.

The proof follows from the fact that $\int p^\alpha(x)q^{1-\alpha}(x)dx = e^{-J_F^{(\alpha)}(\theta_p:\theta_q)}$; see [40].

First, we convert a frequency histogram \tilde{h} to its natural parameter θ with $\theta^i = \log \frac{\tilde{h}^i}{\tilde{h}^d}$; see [39]. The log-normaliser is a non-separable convex function $F(\theta) = \log(1 + \sum_i e^{\theta_i})$. To convert back a multinomial to a frequency histogram with d bins, we first set $\tilde{h}^d = \frac{1}{1 + \sum_{i=1}^{d-1} e^{\theta_i}}$ and then retrieve the other bin values as $\tilde{h}^i = \tilde{h}^d e^{\theta_i}$.

The centroids with respect to skewed Jensen divergences has been investigated in [13,40].

Remark 4. Note that for the special case of $\alpha = 0$ (squared Hellinger centroid), the sided and symmetrized centroids coincide. In that case, the coordinates s_0^i of the squared Hellinger centroid are:

$$s_0^i = \left(\sum_{j=1}^n w_j \sqrt{h_j^i} \right)^2, 1 \leq i \leq d.$$

Remark 5. The symmetrized positive α -centroids can be solved in special cases ($\alpha = \pm 3, \alpha = \pm 1$ corresponding to the symmetrized χ^2 and Jeffreys positive centroids). For frequency centroids, when dealing with binary histograms ($d = 2$), we have only one degree of freedom and can solve the binary frequency centroids. Binary histograms (and mixtures thereof) are used in computer vision and pattern recognition [41].

Remark 6. Since α -divergences are Csiszár f -divergences and f -divergences can always be symmetrized by taking generator $s(t) = f(t) + tf(\frac{1}{t})$, we deduce that symmetrized α -divergences S_α are f -divergences for the generator:

$$f(t) = -\log((1 - \alpha) + \alpha t) - t \log \left((1 - \alpha) + \frac{\alpha}{t} \right). \tag{26}$$

Hence, S_α divergences are convex in both arguments, and the s_α centroids are unique.

5. Soft Mixed α -Clustering

Algorithm 4 reports the general clustering with soft membership, which can be adapted to left ($\lambda_{\text{init}} = 1$), right ($\lambda_{\text{init}} = 0$) or mixed clustering. We have not considered a weighted histogram set in order not load the notations and because the extension is straightforward.

Again, for skew Jeffreys centers, we shall adopt a variational approach. Notice that the soft clustering approach learns all parameters, including λ (if not constrained to zero or one) and $\alpha \in \mathbb{R}$. This is not the case for Matsuyama’s α -expectation maximization (EM) algorithm [42] in which α is fixed beforehand (and, thus, not learned).

Assuming we model the prior for histograms by:

$$p_{\lambda,\alpha,j}(h_i) \propto \lambda \exp -D_\alpha(l_j : h_i) + (1 - \lambda) \exp -D_\alpha(h_i : r_j), \tag{27}$$

the negative log-likelihood involves the α -depending quantity:

$$\begin{aligned} & \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) \log p_{\lambda,\alpha,j}(h_i) \geq \\ & \sum_{j=1}^k \sum_{i=1}^m M_{\lambda,\alpha}(l_j : h_i : r_j) p(j|h_i), \end{aligned} \tag{28}$$

because of the concavity of the logarithm function. Therefore, the maximization step for α involves finding:

$$\arg \max_{\alpha} \sum_{j=1}^k \sum_{i=1}^m M_{\lambda,\alpha}(l_j : h_i : r_j) p(j|h_i). \tag{29}$$

Algorithm 4: Mixed α -soft clustering; MAS $C(\mathcal{H}, k, \lambda, \alpha)$

Input: Histogram set \mathcal{H} with $|\mathcal{H}| = m$, integer $k > 0$, real $\lambda \leftarrow \lambda_{\text{init}} \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k \leftarrow \text{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

 //Expectation

for $i = 1, 2, \dots, m$ **do**

for $j = 1, 2, \dots, k$ **do**

$$p(j|h_i) = \frac{\pi_j \exp(-M_{\lambda,\alpha}(l_j:h_i:r_j))}{\sum_{j'} \pi_{j'} \exp(-M_{\lambda,\alpha}(l_{j'}:h_i:r_{j'}))};$$

 //Maximization

for $j = 1, 2, \dots, k$ **do**

$$\pi_j \leftarrow \frac{1}{m} \sum_i p(j|h_i);$$

$$l_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1+\alpha}{2}} \right)^{\frac{2}{1+\alpha}};$$

$$r_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1-\alpha}{2}} \right)^{\frac{2}{1-\alpha}};$$

 //Alpha - Lambda

$$\alpha \leftarrow \alpha - \eta_1 \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) \frac{\partial}{\partial \alpha} M_{\lambda,\alpha}(l_j : h_i : r_j);$$

if $\lambda_{\text{init}} \neq 0, 1$ **then**

$$\lambda \leftarrow \lambda - \eta_2 \left(\sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_{\alpha}(l_j : h_i) - \right.$$

$$\left. \sum_{j=1}^k \sum_{i=1}^m p(j|h_i) D_{\alpha}(h_i : r_j) \right);$$

 //for some small η_1, η_2 ; ensure that $\lambda \in [0, 1]$.

until convergence;

Output: Soft clustering of \mathcal{H} according to k densities $p(j|\cdot)$ following \mathcal{C} ;

No closed-form solution are known, so we compute the gradient update in Algorithm 4 with:

$$\frac{\partial M_{\lambda,\alpha}(l_j : h_i : r_j)}{\partial \alpha} = \lambda \frac{\partial D_{\alpha}(l_j : h_i)}{\partial \alpha} + (1 - \lambda) \frac{\partial D_{\alpha}(h_i : r_j)}{\partial \alpha}, \tag{30}$$

$$\frac{\partial D_{\alpha}(p : q)}{\partial \alpha} = \frac{2}{(1 - \alpha)^2} \times \left(q - \left(\frac{1 - \alpha}{1 + \alpha} \right)^2 p + p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}} \left(\frac{4\alpha}{1 - \alpha^2} - \ln \left(\frac{q}{p} \right) \right) \right). \tag{31}$$

The update in λ is easier as:

$$\frac{\partial M_{\lambda,\alpha}(l_j : h_i : r_j)}{\partial \lambda} = D_{\alpha}(l_j : h_i) - D_{\alpha}(h_i : r_j). \tag{32}$$

Maximizing the likelihood in λ would imply choosing $\lambda \in \{0, 1\}$ (a hard choice for left/right centers), yet we prefer the soft update for the parameter, like for α .

6. Conclusions

The family of α -divergences plays a fundamental role in information geometry: These statistical distortion measures are the canonical divergences of dual spaces on probability distribution manifolds with constant curvature $\kappa = \frac{1-\alpha^2}{4}$ and the canonical divergences of dually flat manifolds for positive distribution manifolds [19].

In this work, we have presented three techniques for clustering (positive or frequency) histograms using k -means:

- (1) Sided left or right α -centroid k -means,
- (2) Symmetrized Jeffreys-type α -centroid (variational) k -means, and
- (3) Coupled k -means with respect to mixed α -divergences relying on dual α -centroids.

Sided and mixed dual centroids are always available in closed-forms and are therefore highly attractive from the standpoint of implementation. Symmetrized Jeffreys centroids are in general not available in closed-form and require one to implement a variational k -means by updating incrementally the cluster centroids in order to monotonically decrease the k -means loss function. From the clustering standpoint, this appears not to be a problem when guaranteed expected approximations to the optimal clustering are enough.

Indeed, we also presented and analyzed an extension of k -means++ [17] for seeding those k -means algorithms. The mixed α -seeding initializations do not require one to calculate centroids and behaves like a discrete k -means by picking up the seeds among the data. We reported guaranteed probabilistic clustering bounds. Thus, it yields a fast hard/soft data partitioning technique with respect to mixed or symmetrized α -divergences. Recently, the advantage of clustering using α -divergences by tuning α in applications has been demonstrated in [18]. We thus expect the computationally fast mixed α -seeding with guaranteed performance to be useful in a growing number of applications.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Author Contributions

All authors contributed equally to the design of the research. The research was carried out by all authors. Frank Nielsen and Richard Nock wrote the paper. Frank Nielsen implemented the algorithms and performed experiments. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interests.

Appendix

A. Proof Sketch of Theorem 1

We give here the key results allowing one to obtain the proof of the Theorem, following the proof scheme of [15]. In order not to load the notations, weights are considered uniform. The extension to non-uniform weights is immediate as it boils down to duplicate histograms in the histogram set and does not change the approximation result.

Let $\mathcal{A} \subseteq \mathcal{H}$ be an arbitrary cluster of \mathcal{C}_{opt} . Let us define $U_{\mathcal{A}}$ and $\pi_{\mathcal{A}}$ as the uniform and biased distributions conditioned to \mathcal{A} . The key to the proof is to relate the expected potential of \mathcal{A} under $U_{\mathcal{A}}$ and $\pi_{\mathcal{A}}$ to its contribution to the optimal potential.

Lemma 3. *Let $\mathcal{A} \subseteq \mathcal{H}$ be an arbitrary cluster of \mathcal{C}_{opt} . Then:*

$$\begin{aligned} \mathbf{E}_{c \sim U_{\mathcal{A}}}[M_{\lambda, \alpha}(\mathcal{A}, c)] &= M_{\text{opt}, \lambda, \alpha}(\mathcal{A}) + M_{\text{opt}, \lambda, -\alpha}(\mathcal{A}) \\ &= \mathbf{E}_{c \sim U_{\mathcal{A}}}[M_{\lambda, -\alpha}(\mathcal{A}, c)] \end{aligned}$$

where $U_{\mathcal{A}}$ is the uniform distribution over \mathcal{A} .

Proof. α -coordinates have the property that for any subset $\mathcal{A} \subseteq \mathcal{H}$, $(1/|\mathcal{A}|) \sum_{p \in \mathcal{A}} u_{\alpha}(p) = u_{\alpha}(r_{\alpha, \mathcal{A}})$. Hence, we have:

$$\begin{aligned} \forall c \in \mathcal{A} \ , \sum_{p \in \mathcal{A}} D_{\alpha}(p : c) &= \sum_{p \in \mathcal{A}} D_{\varphi_{\alpha}}(u_{\alpha}(p) : u_{\alpha}(c)) \\ &= \sum_{p \in \mathcal{A}} D_{\varphi_{\alpha}}(u_{\alpha}(p) : u_{\alpha}(r_{\alpha, \mathcal{A}})) + |\mathcal{A}| D_{\varphi_{\alpha}}(u_{\alpha}(r_{\alpha, \mathcal{A}}) : u_{\alpha}(c)) \\ &= \sum_{p \in \mathcal{A}} D_{\alpha}(p : r_{\alpha, \mathcal{A}}) + |\mathcal{A}| D_{\alpha}(r_{\alpha, \mathcal{A}} : c) \ . \end{aligned} \tag{33}$$

Because $D_{\alpha}(p : q) = D_{-\alpha}(q : p)$ and $l_{\alpha} = r_{-\alpha}$, we obtain:

$$\begin{aligned} \forall c \in \mathcal{A} \ , \sum_{p \in \mathcal{A}} D_{\alpha}(c : p) &= \sum_{p \in \mathcal{A}} D_{-\alpha}(p : c) \\ &= \sum_{p \in \mathcal{A}} D_{-\alpha}(p : r_{-\alpha, \mathcal{A}}) + |\mathcal{A}| D_{-\alpha}(r_{-\alpha, \mathcal{A}} : c) \\ &= \sum_{p \in \mathcal{A}} D_{\alpha}(l_{\alpha, \mathcal{A}} : p) + |\mathcal{A}| D_{\alpha}(c : l_{\alpha, \mathcal{A}}) \ . \end{aligned} \tag{34}$$

It comes now from (33) and (34) that:

$$\begin{aligned}
 & \mathbf{E}_{c \sim U_{\mathcal{A}}}[M_{\lambda, \alpha}(\mathcal{A}, c)] \\
 &= \frac{1}{|\mathcal{A}|} \sum_{c \in \mathcal{A}} \sum_{p \in \mathcal{A}} \{ \lambda D_{\alpha}(c : p) + (1 - \lambda) D_{\alpha}(p : c) \} \\
 &= (1 - \lambda) \sum_{p \in \mathcal{A}} D_{\alpha}(p : r_{\alpha, \mathcal{A}}) + (1 - \lambda) \sum_{p \in \mathcal{A}} D_{\alpha}(r_{\alpha, \mathcal{A}} : p) \\
 &\quad + \lambda \sum_{p \in \mathcal{A}} D_{\alpha}(l_{\alpha, \mathcal{A}} : p) + \lambda \sum_{p \in \mathcal{A}} D_{\alpha}(p : l_{\alpha, \mathcal{A}}) \\
 &= (1 - \lambda) M_{\text{opt}, 0, \alpha}(\mathcal{A}) + \lambda M_{\text{opt}, 1, \alpha}(\mathcal{A}) \\
 &\quad + (1 - \lambda) M_{\text{opt}, 0, -\alpha}(\mathcal{A}) + \lambda M_{\text{opt}, 1, -\alpha}(\mathcal{A}) \\
 &= M_{\text{opt}, \lambda, \alpha}(\mathcal{A}) + M_{\text{opt}, \lambda, -\alpha}(\mathcal{A}) .
 \end{aligned} \tag{35}$$

This gives the left-hand side equality of the Lemma. The right-hand side follows from the fact that $\mathbf{E}_{c \sim U_{\mathcal{A}}}[M_{\lambda, -\alpha}(\mathcal{A}, c)] = M_{\text{opt}, 1-\lambda, \alpha}(\mathcal{A}) + M_{\text{opt}, 1-\lambda, -\alpha}(\mathcal{A})$. \square

Instead of $M_{\text{opt}, \lambda, \alpha}(\mathcal{A}) + M_{\text{opt}, \lambda, -\alpha}(\mathcal{A})$, we want a term depending solely on $M_{\text{opt}, \lambda, \alpha}(\mathcal{A})$ as it is the “true” optimum. We now give two lemmata that shall be useful in obtaining this upper bound. The first is of independent interest, as it shows that any α -divergence is a scaled, squared Hellinger distance between geometric means of points.

Lemma 4. For any p, q and $\alpha \neq 1$, there exists $r \in [p, q]$, such that $(1 - \alpha)^2 D_{\alpha}(p : q) = D_0(p^{1-\alpha} r^{\alpha} : q^{1-\alpha} r^{\alpha})$.

Proof. By the definition of Bregman divergences, for any x, y , there exists some $z \in [x, y]$, such that:

$$\begin{aligned}
 D_{\varphi_{\alpha}}(x : y) &= \frac{1}{2} (x - y)^2 \varphi''_{\alpha}(z) \\
 &= \frac{1}{2} (x - y)^2 \left(1 + \frac{1 - \alpha}{2} z \right)^{\frac{2\alpha}{1-\alpha}} ,
 \end{aligned}$$

and since u_{α} is continuous and strictly increasing, for any p, q , there exists some $r \in [p, q]$, such that:

$$\begin{aligned}
 D_{\alpha}(p : q) &= D_{\varphi_{\alpha}}(u_{\alpha}(p) : u_{\alpha}(q)) \\
 &= \frac{1}{2} (u_{\alpha}(p) - u_{\alpha}(q))^2 \left(1 + \frac{1 - \alpha}{2} u_{\alpha}(r) \right)^{\frac{2\alpha}{1-\alpha}} \\
 &= \frac{2}{(1 - \alpha)^2} \left(p^{\frac{1-\alpha}{2}} - q^{\frac{1-\alpha}{2}} \right)^2 r^{\alpha} \\
 &= \frac{2}{(1 - \alpha)^2} \left(p^{1-\alpha} + q^{1-\alpha} - 2(pq)^{\frac{1-\alpha}{2}} \right) r^{\alpha} \\
 &= \frac{1}{(1 - \alpha)^2} D_0(p^{1-\alpha} r^{\alpha} : q^{1-\alpha} r^{\alpha}) .
 \end{aligned}$$

\square

Lemma 5. Let discrete random variable x take non-negative values x_1, x_2, \dots, x_m with uniform probabilities. Then, for any $\beta > -1$, we have $\text{var}(x^{1+\beta}/u^\beta) \leq \text{var}(x)$, with $u \doteq (1 + \beta)^\beta \max_i x_i$.

Proof. First, $\forall \beta > -1$, remark that for any x , function $f(x) = x(u^\beta - x^\beta)$ is increasing for $x \leq u/(1 + \beta)^\beta$. Hence, assuming that the x_i s are put in non-increasing order without loss of generality, we have $f(x_i) \geq f(x_j)$, and so, $x_i(u^\beta - x_i^\beta) \geq x_j(u^\beta - x_j^\beta), \forall i \geq j$, as long as $x_i \leq u/(1 + \beta)^\beta$. Choosing $u = x_1(1 + \beta)^\beta$ yields, after reordering and putting the exponent, $(x_i^{1+\beta} - x_j^{1+\beta})^2 \leq (x_i u^\beta - x_j u^\beta)^2$. Hence:

$$\begin{aligned} & \frac{1}{m} \sum_i x_i^{2(1+\beta)} - \left(\frac{1}{m} \sum_i x_i^{(1+\beta)} \right)^2 \\ &= \frac{1}{2m^2} \sum_{i,j} (x_i^{1+\beta} - x_j^{1+\beta})^2 \\ &\leq \frac{1}{2m^2} \sum_{i,j} (x_i u^\beta - x_j u^\beta)^2 \\ &= \frac{u^{2\beta}}{2m^2} \sum_{i,j} (x_i - x_j)^2 \\ &= u^{2\beta} \left(\frac{1}{m} \sum_i x_i^2 - \left(\frac{1}{m} \sum_i x_i \right)^2 \right) . \end{aligned}$$

Dividing by $u^{2\beta}$ the leftmost and rightmost terms and using the fact that $\text{var}(\lambda x) = \lambda^2 \text{var}(x)$ yields the statement of the Lemma. \square

We are now ready to upper bound $M_{\text{opt},\lambda,-\alpha}(\mathcal{A})$ as a function of $M_{\text{opt},\lambda,\alpha}(\mathcal{A})$.

Lemma 6. For any cluster \mathcal{A} of \mathcal{C}_{opt} ,

$$M_{\text{opt},\lambda,-\alpha}(\mathcal{A}) \leq M_{\text{opt},\lambda,\alpha}(\mathcal{A}) \times \begin{cases} f(\lambda) & \text{if } \lambda \in (0, 1) \\ z(\alpha)h^2(\alpha) & \text{otherwise} \end{cases} ,$$

where $z(\alpha)$, $f(\lambda)$ and $h(\alpha)$ are defined in Theorem 1.

Proof. The case $\lambda \neq 0, 1$ is fast, as we have by definition:

$$\begin{aligned} M_{\text{opt},\lambda,-\alpha}(\mathcal{A}) &= \sum_{p \in \mathcal{A}} \lambda D_{-\alpha}(l_{-\alpha,\mathcal{A}} : p) + (1 - \lambda) D_{-\alpha}(p : r_{-\alpha,\mathcal{A}}) \\ &= \sum_{p \in \mathcal{A}} \lambda D_{\alpha}(p : l_{-\alpha,\mathcal{A}}) + (1 - \lambda) D_{\alpha}(r_{-\alpha,\mathcal{A}} : p) \\ &= \sum_{p \in \mathcal{A}} \lambda D_{\alpha}(p : r_{\alpha,\mathcal{A}}) + (1 - \lambda) D_{\alpha}(l_{\alpha,\mathcal{A}} : p) \\ &\leq \max \left\{ \frac{1 - \lambda}{\lambda}, \frac{\lambda}{1 - \lambda} \right\} M_{\text{opt},\lambda,\alpha}(\mathcal{A}) \\ &= f(\lambda) M_{\text{opt},\lambda,\alpha}(\mathcal{A}) . \end{aligned}$$

Suppose now that $\lambda = 0$ and $\alpha \geq 0$. Because $M_{\text{opt},0,-\alpha}(\mathcal{A}) = \sum_{p \in \mathcal{A}} D_{-\alpha}(p : r_{-\alpha,\mathcal{A}}) = \sum_{p \in \mathcal{A}} D_{\alpha}(l_{\alpha,\mathcal{A}} : p) = M_{\text{opt},1,\alpha}(\mathcal{A})$, what we wish to do is upper bound $\sum_{p \in \mathcal{A}} D_{\alpha}(l_{\alpha,\mathcal{A}} : p) =$

$M_{\text{opt},1,\alpha}(\mathcal{A})$ as a function of $\sum_{p \in \mathcal{A}} D_\alpha(p : r_{\alpha,\mathcal{A}}) = M_{\text{opt},0,\alpha}(\mathcal{A})$. We use Lemmas 4 and 5 in the following derivations, using $r(p)$ to refer to the r in Lemma 4, assuming $\alpha \geq 0$. We also note $\text{var}_{\mathcal{A}}(f(p))$ as the variance, under the uniform distribution over \mathcal{A} , of discrete random variable $f(p)$, for $p \in \mathcal{A}$. We have:

$$\begin{aligned}
 & \sum_{p \in \mathcal{A}} D_\alpha(l_{\alpha,\mathcal{A}} : p) \\
 &= \sum_{p \in \mathcal{A}} D_{-\alpha}(p : l_{\alpha,\mathcal{A}}) \\
 &= \frac{1}{(1 + \alpha)^2} \sum_{p \in \mathcal{A}} r(p)^{-\alpha} D_0(p^{1+\alpha} : l_{\alpha,\mathcal{A}}^{1+\alpha}) \\
 &\leq \frac{1}{(1 + \alpha)^2 \min_{\mathcal{A}} p^\alpha} \sum_{p \in \mathcal{A}} D_0(p^{1+\alpha} : l_{\alpha,\mathcal{A}}^{1+\alpha}) \\
 &= \frac{1}{(1 + \alpha)^2 \min_{\mathcal{A}} p^\alpha} \sum_{p \in \mathcal{A}} \left(p^{1+\alpha} + l_{\alpha,\mathcal{A}}^{1+\alpha} - 2p^{\frac{1+\alpha}{2}} l_{\alpha,\mathcal{A}}^{\frac{1+\alpha}{2}} \right) \\
 &= \frac{|\mathcal{A}|}{(1 + \alpha)^2 \min_{\mathcal{A}} p^\alpha} \left(\frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} p^{1+\alpha} - \left(\frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} p^{\frac{1+\alpha}{2}} \right)^2 \right) \\
 &= \frac{|\mathcal{A}| \text{var}_{\mathcal{A}}(p^{\frac{1+\alpha}{2}})}{(1 + \alpha)^2 \min_{\mathcal{A}} p^\alpha} . \tag{36}
 \end{aligned}$$

We have used the expression of left centroid $l_{\alpha,\mathcal{A}}^{1+\alpha}$ to simplify the expressions. Now, picking $x_i = p_i^{\frac{1-\alpha}{2}}$, $\beta = 2\alpha/(1 - \alpha)$ and $u = \left(\frac{1+\alpha}{1-\alpha}\right)^{\frac{2\alpha}{1-\alpha}} \max_{\mathcal{A}} p^{\frac{1-\alpha}{2}}$ in Lemma 5 yields:

$$\begin{aligned}
 & \text{var}_{\mathcal{A}}(p^{\frac{1+\alpha}{2}}) \\
 &= u^{2\beta} \text{var}_{\mathcal{A}}(p^{\frac{1+\alpha}{2}} / u^\beta) \\
 &= u^{2\beta} \text{var}_{\mathcal{A}}\left(p^{\frac{1-\alpha}{2}} p^\alpha / u^\beta\right) \\
 &= u^{2\beta} \text{var}(x^{1+\beta} / u^\beta) \\
 &\leq u^{2\beta} \text{var}(x) \\
 &= u^{2\beta} \text{var}_{\mathcal{A}}\left(p^{\frac{1-\alpha}{2}}\right) \\
 &= \left(\frac{1 + \alpha}{1 - \alpha}\right)^{\frac{8\alpha^2}{(1-\alpha)^2}} \max_{\mathcal{A}} p^{2\alpha} \text{var}_{\mathcal{A}}\left(p^{\frac{1-\alpha}{2}}\right) . \tag{37}
 \end{aligned}$$

Plugging this in (36) yields:

$$\begin{aligned}
 & \sum_{p \in \mathcal{A}} D_\alpha(l_{\alpha, \mathcal{A}} : p) \\
 & \leq \left(\frac{1 + \alpha}{1 - \alpha} \right)^{\frac{8\alpha^2}{(1-\alpha)^2}} \frac{|\mathcal{A}| \max_{\mathcal{A}} p^{2\alpha} \text{var}_{\mathcal{A}} \left(p^{\frac{1-\alpha}{2}} \right)}{(1 + \alpha)^2 \min_{\mathcal{A}} p^\alpha} \\
 & = \left(\frac{1 + \alpha}{1 - \alpha} \right)^{\frac{8\alpha^2}{(1-\alpha)^2} - 2} \left(\frac{\max_{\mathcal{A}} p}{\min_{\mathcal{A}} p} \right)^{2\alpha} \times \frac{|\mathcal{A}| \min_{\mathcal{A}} p^\alpha \text{var}_{\mathcal{A}} \left(p^{\frac{1-\alpha}{2}} \right)}{(1 - \alpha)^2} \\
 & = \left(\frac{1 + \alpha}{1 - \alpha} \right)^{\frac{8\alpha^2}{(1-\alpha)^2} - 2} \left(\frac{\max_{\mathcal{A}} p}{\min_{\mathcal{A}} p} \right)^{2\alpha} \times \frac{\min_{\mathcal{A}} p^\alpha}{(1 - \alpha)^2} \sum_{p \in \mathcal{A}} D_0(p^{1-\alpha} : r_{\alpha, \mathcal{A}}^{1-\alpha}) \tag{38} \\
 & \leq \left(\frac{1 + \alpha}{1 - \alpha} \right)^{\frac{8\alpha^2}{(1-\alpha)^2} - 2} \left(\frac{\max_{\mathcal{A}} p}{\min_{\mathcal{A}} p} \right)^{2\alpha} \times \frac{1}{(1 - \alpha)^2} \sum_{p \in \mathcal{A}} r(p)^\alpha D_0(p^{1-\alpha} : r_{\alpha, \mathcal{A}}^{1-\alpha}) \\
 & = \left(\frac{1 + \alpha}{1 - \alpha} \right)^{\frac{8\alpha^2}{(1-\alpha)^2} - 2} \left(\frac{\max_{\mathcal{A}} p}{\min_{\mathcal{A}} p} \right)^{2\alpha} \times \sum_{p \in \mathcal{A}} D_\alpha(p : r_{\alpha, \mathcal{A}}) \\
 & \leq z(\alpha) \left(\frac{\max_{\mathcal{A}} p}{\min_{\mathcal{A}} p} \right)^{2\alpha} \times \sum_{p \in \mathcal{A}} D_\alpha(p : r_{\alpha, \mathcal{A}}) . \tag{39}
 \end{aligned}$$

Here, (38) follows the path backwards of derivations that lead to (36). The cases $\lambda = 1$ or $\alpha < 0$ are obtained using the same chains of derivations and achieve the proof of Lemma 6. \square

Lemma 6 can be directly used to refine the bound of Lemma 3 in the uniform distribution. We give the Lemma for the biased distribution, directly integrating the refinement of the bound.

Lemma 7. *Let \mathcal{A} be an arbitrary cluster of \mathcal{C}_{opt} and \mathcal{C} an arbitrary clustering. If we add a random couple (c, c) to \mathcal{C} , chosen from \mathcal{A} with π as in Algorithm 2, then:*

$$\begin{aligned}
 & \mathbf{E}_{c \sim \pi_{\mathcal{A}}} [M_{\lambda, \alpha}(\mathcal{A}, c)] \\
 & \leq 4 \begin{cases} f(\lambda)h^2(\alpha)M_{\text{opt}, \lambda, \alpha}(\mathcal{A}) & \text{if } \lambda \in (0, 1) \\ z(\alpha)h^4(\alpha)M_{\text{opt}, \lambda, \alpha}(\mathcal{A}) & \text{otherwise} \end{cases} , \tag{40}
 \end{aligned}$$

where $f(\lambda)$ and $h(\alpha)$ are defined in Theorem 1.

Proof. The proof essentially follows the proof of Lemma 3 in [15]. To complete it, we need a triangle inequality involving α -divergences. We give it here.

Lemma 8. *For any p, q, r and α , we have:*

$$\sqrt{D_\alpha(p : q)} \leq \left(\frac{\max_i \{p_i, q_i, r_i\}}{\min_i \{p_i, q_i, r_i\}} \right)^{|\alpha|} \left(\sqrt{D_\alpha(p : r)} + \sqrt{D_\alpha(r : q)} \right) \tag{41}$$

(where the min is over strictly positive values)

Remark: take $\alpha = 0$; we find the triangle inequality for the squared Hellinger distance.

Proof. Using the proof of Lemma 2 in [15] for Bregman divergence D_{φ_α} , we get:

$$\begin{aligned} & \sqrt{D_{\varphi_\alpha}(x : z)} \\ & \leq \rho(\alpha) \left(\sqrt{D_{\varphi_\alpha}(x : y)} + \sqrt{D_{\varphi_\alpha}(y : z)} \right), \end{aligned} \tag{42}$$

where:

$$\rho(\alpha) = \max_{u,v} \frac{\left(1 + \frac{1-\alpha}{2}u\right)^{\frac{2\alpha}{1-\alpha}}}{\left(1 + \frac{1-\alpha}{2}v\right)^{\frac{2\alpha}{1-\alpha}}}. \tag{43}$$

Taking $x = u_\alpha(p)$, $y = u_\alpha(q)$, $z = u_\alpha(r)$ yields $\rho(\alpha) = \max_{s,t \in \{p_i, q_i, r_i\}} (s/t)^{|\alpha|}$ and the statement of Lemma 8. \square

The rest of the proof of Lemma 7 follows the proof of Lemma 3 in [15]. \square

We get all of the ingredients to our proof, and there remains to use Lemma 4 in [15] to achieve the proof of Theorem 1.

B. Properties of α -Divergences

For positive arrays p and q , the α -divergence $D_\alpha(p : q)$ can be defined as an equivalent representational Bregman divergence [19,34] $B_{\varphi_\alpha}(u_\alpha(p) : u_\alpha(q))$ over the (u_α, v_α) -structure [43] with:

$$\varphi_\alpha(x) \doteq \frac{2}{1 + \alpha} \left(1 + \frac{1 - \alpha}{2}x\right)^{\frac{2}{1-\alpha}}, \tag{44}$$

$$u_\alpha(p) \doteq \frac{2}{1 - \alpha} \left(p^{\frac{1-\alpha}{2}} - 1\right), \tag{45}$$

$$v_\alpha(p) \doteq \frac{2}{1 + \alpha} p^{\frac{1+\alpha}{2}}, \tag{46}$$

where we assume that $\alpha \neq \pm 1$. Otherwise, for $\alpha = \pm 1$, we compute $D_\alpha(p : q)$ by taking the sided Kullback–Leibler divergence extended to positive arrays.

In the proof of Theorem 1, we have used two properties of α -divergences of independent interest:

- any α -divergence can be explained as a scaled squared Hellinger distance between geometric means of its arguments and a point that belong to their segment (Lemma 4);
- any α -divergence satisfies a generalized triangle inequality (Lemma 8). Notice that this Lemma is optimal in the sense that for $\alpha = 0$, it is possible to recover the triangle inequality of the Hellinger distance.

The following lemma shows how to bound the mixed divergence as a function of an α -divergence.

Lemma 9. For any positive arrays l, h, r and $\alpha \neq \pm 1$, define $\eta \doteq \lambda(1 - \alpha)/(1 - \alpha(2\lambda - 1)) \in [0, 1]$, g_η with $g_\eta^i \doteq (l^i)^\eta (r^i)^{1-\eta}$ and a_η with $a_\eta^i \doteq \eta l^i + (1 - \eta)r^i$. Then, we have:

$$M_{\lambda,\alpha}(l : h : r) \leq \frac{1 - \alpha^2(2\lambda - 1)^2}{1 - \alpha^2} D_{\alpha(2\lambda-1)}(g_\eta : h) + \frac{2(1 - \alpha(2\lambda - 1))}{1 - \alpha^2} \sum_i (a_\eta^i - g_\eta^i) .$$

Proof. For all index i , we have:

$$M_{\lambda,\alpha}(l^i : h^i : r^i) = \lambda D_\alpha(l^i : h^i) + (1 - \lambda) D_\alpha(h^i : r^i) = \frac{4}{1 - \alpha^2} \left(\frac{\lambda(1 - \alpha)}{2} l^i + \frac{(1 - \lambda)(1 + \alpha)}{2} r^i + \frac{1 + \alpha(2\lambda - 1)}{2} h^i \right) \tag{47}$$

$$- \lambda(l^i)^{\frac{1-\alpha}{2}} (h^i)^{\frac{1+\alpha}{2}} - (1 - \lambda)(r^i)^{\frac{1+\alpha}{2}} (h^i)^{\frac{1-\alpha}{2}} . \tag{48}$$

The arithmetic-geometric-harmonic (AGH) inequality implies:

$$\begin{aligned} \lambda(l^i)^{\frac{1-\alpha}{2}} (h^i)^{\frac{1+\alpha}{2}} + (1 - \lambda)(r^i)^{\frac{1+\alpha}{2}} (h^i)^{\frac{1-\alpha}{2}} &\geq (l^i)^{\frac{\lambda(1-\alpha)}{2}} (r^i)^{\frac{(1-\lambda)(1+\alpha)}{2}} (h^i)^{\frac{1+\alpha(2\lambda-1)}{2}} \\ &= \left((l^i)^{\frac{\lambda(1-\alpha)}{1-\alpha(2\lambda-1)}} (r^i)^{\frac{(1-\lambda)(1+\alpha)}{1-\alpha(2\lambda-1)}} \right)^{\frac{1-\alpha(2\lambda-1)}{2}} (h^i)^{\frac{1+\alpha(2\lambda-1)}{2}} \\ &= ((l^i)^\eta (r^i)^{1-\eta})^{\frac{1-\alpha(2\lambda-1)}{2}} (h^i)^{\frac{1+\alpha(2\lambda-1)}{2}} \\ &= (g_\eta^i)^{\frac{1-\alpha(2\lambda-1)}{2}} (h^i)^{\frac{1+\alpha(2\lambda-1)}{2}} . \end{aligned}$$

It follows that (48) yields:

$$\begin{aligned} M_{\lambda,\alpha}(l^i : h^i : r^i) &\leq \frac{4}{1 - \alpha^2} \left(\frac{1 - \alpha(2\lambda - 1)}{2} (\eta l^i + (1 - \eta) r^i) + \frac{1 + \alpha(2\lambda - 1)}{2} h^i - (g_\eta^i)^{\frac{1-\alpha(2\lambda-1)}{2}} (h^i)^{\frac{1+\alpha(2\lambda-1)}{2}} \right) \tag{49} \\ &= \frac{1 - \alpha^2(2\lambda - 1)^2}{1 - \alpha^2} D_{\alpha(2\lambda-1)}(g_\eta^i : h^i) + \frac{2(1 - \alpha(2\lambda - 1))}{1 - \alpha^2} (a_\eta^i - g_\eta^i) , \tag{50} \end{aligned}$$

out of which we get the statement of the Lemma. \square

C. Sided α -Centroids

For the sake of completeness, we prove the following theorem:

Theorem 7 (Sided positive α -centroids [34]). *The left-sided l_α and right-sided r_α positive weighted α -centroid coordinates of a set of n positive histograms h_1, \dots, h_n are weighted α -means:*

$$r_\alpha^i = f_\alpha^{-1} \left(\sum_{j=1}^n w_j f_\alpha(h_j^i) \right), l_\alpha^i = r_{-\alpha}^i$$

with:

$$f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & \alpha \neq \pm 1, \\ \log x & \alpha = 1. \end{cases}$$

Proof. We distinguish three cases: $\alpha \neq \pm 1$, $\alpha = -1$ and $\alpha = 1$.

First, consider the general case $\alpha \neq \pm 1$. We have to minimize:

$$R_\alpha(x, \mathcal{H}) = \frac{4}{1 - \alpha^2} \sum_{j=1}^n w_j \times \sum_{i=1}^d \left(\frac{1 - \alpha}{2} h_j^i + \frac{1 + \alpha}{2} x^i - (h_j^i)^{\frac{1-\alpha}{2}} (x^i)^{\frac{1+\alpha}{2}} \right).$$

Removing all additive terms independent of x^i and the overall constant multiplicative factor $\frac{4}{1-\alpha^2} \neq 0$, we get the following equivalent minimisation problem:

$$R'_\alpha(x, \mathcal{H}) = \sum_{i=1}^d \frac{1 + \alpha}{2} x^i - (x^i)^{\frac{1+\alpha}{2}} \underbrace{\left(\sum_{j=1}^n w_j (h_j^i)^{\frac{1-\alpha}{2}} \right)}_{\bar{h}_\alpha^i}, \tag{51}$$

where \bar{h}_α^i denote the following aggregation term:

$$\bar{h}_\alpha^i = \sum_{j=1}^n w_j (h_j^i)^{\frac{1-\alpha}{2}}.$$

Setting coordinate-wise the derivative to zero of Equation (51) (i.e., $\nabla_x R'(x, \mathcal{H}) = 0$), we get:

$$\frac{1 + \alpha}{2} - \frac{1 + \alpha}{2} (x^i)^{\frac{\alpha-1}{2}} \bar{h}_\alpha^i = 0$$

Thus, we find that the coordinates of the right-sided α -centroids are:

$$c_\alpha^i = (\bar{h}_\alpha^i)^{\frac{2}{1-\alpha}} = \left(\sum_{j=1}^n w_j (h_j^i)^{\frac{1-\alpha}{2}} \right)^{\frac{2}{1-\alpha}} = \hat{h}_\alpha^i.$$

We recognise the expression of a quasi-arithmetic mean for the strictly monotonous generator $f_\alpha(x)$:

$$r_\alpha^i = f_\alpha^{-1} \left(\sum_{j=1}^n w_j f_\alpha(h_j^i) \right), \tag{52}$$

with:

$$f_\alpha(x) = x^{\frac{1-\alpha}{2}}, \quad f_\alpha^{-1}(x) = x^{\frac{2}{1-\alpha}}, \alpha \neq \pm 1.$$

Therefore, we conclude that the coordinates of the positive α -centroid are the weighted α -means of the histogram coordinates (for $\alpha \neq \pm 1$). Quasi-arithmetic means are also called in the literature quasi-linear means or f -means.

When $\alpha = -1$, we search for the right-sided extended Kullback–Leibler divergence centroid by minimising:

$$R_{-1}(x; \tilde{\mathcal{H}}) = \sum_{j=1}^n w_j \sum_{i=1}^d h_j^i \log \frac{h_j^i}{x^i} + x^i - h_j^i.$$

It is equivalent to minimizing:

$$R'_{-1}(x; \tilde{\mathcal{H}}) = \sum_{i=1}^d x^i - \underbrace{\left(\sum_{j=1}^n w_j h_j^i \right)}_a \log x^i,$$

where a denotes the arithmetic mean. Solving coordinate-wise, we get $c^i = a^i = \sum_{j=1}^n w_j h_j^i$.

When $\alpha = 1$, the right-sided reverse extended KL centroid is a left-sided extended KL centroid. The minimisation problem is:

$$R_1(x; \tilde{\mathcal{H}}) = \sum_{j=1}^n w_j \sum_{i=1}^d x^i \log \frac{x^i}{h_j^i} + h_j^i - x^i.$$

Since $\sum_j w_j = 1$, we solve coordinate-wise and find $\log x = \sum_j w_j \log h_j$. That is, r_1^i is the geometric mean:

$$r_1^i = \prod_{j=1}^n (h_j^i)^{w_j}.$$

Both the arithmetic mean and the geometric mean are power means in the limit case (and hence quasi-arithmetic means). Thus,

$$r_\alpha^i = f_\alpha^{-1} \left(\sum_{j=1}^n w_j f_\alpha(h_j^i) \right), \quad (53)$$

with:

$$f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & \alpha \neq \pm 1, \\ \log x & \alpha = 1. \end{cases}$$

□

References

1. Baker, L.D.; McCallum, A.K. Distributional clustering of words for text classification. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; ACM: New York, NY, USA, 1998; pp. 96–103.
2. Bigi, B. Using Kullback–Leibler distance for text categorization. In Proceedings of the 25th European conference on IR research (ECIR), Pisa, Italy, 14–16 April 2003; Springer-Verlag: Berlin/Heidelberg, Germany, 2003; ECIR'03, pp. 305–319.
3. Bag of Words Data Set. Available online: <http://archive.ics.uci.edu/ml/datasets/Bag+of+Words> (accessed on 17 June 2014).
4. Csurka, G.; Bray, C.; Dance, C.; Fan, L. *Visual Categorization with Bags of Keypoints*; Workshop on Statistical Learning in Computer Vision (ECCV); Xerox Research Centre Europe: Meylan, France, 2004, pp. 1–22.
5. Jégou, H.; Douze, M.; Schmid, C. Improving Bag-of-Features for Large Scale Image Search. *Int. J. Comput. Vis.* **2010**, *87*, 316–336.
6. Yu, Z.; Li, A.; Au, O.; Xu, C. Bag of textons for image segmentation via soft clustering and convex shift. In Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 781–788.
7. Steinhaus, H. Sur la division des corp matériels en parties. *Bull. Acad. Polon. Sci.* **1956**, *1*, 801–804. (in French)

8. Lloyd, S.P. *Least Squares Quantization in PCM*; Technical Report RR-5497; Bell Laboratories: Murray Hill, NJ, USA, 1957.
9. Lloyd, S.P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.
10. Chandrasekhar, V.; Takacs, G.; Chen, D.M.; Tsai, S.S.; Reznik, Y.A.; Grzeszczuk, R.; Girod, B. Compressed histogram of gradients: A low-bitrate descriptor. *Int. J. Comput. Vis.* **2012**, *96*, 384–399.
11. Nock, R.; Nielsen, F.; Briys, E. Non-linear book manifolds: Learning from associations the dynamic geometry of digital libraries. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, 2013; pp. 313–322.
12. Kwitt, R.; Vasconcelos, N.; Rasiwasia, N.; Uhl, A.; Davis, B.C.; Häfner, M.; Wrba, F. Endoscopic image analysis in semantic space. *Med. Image Anal.* **2012**, *16*, 1415–1422.
13. Nielsen, F. A family of statistical symmetric divergences based on Jensen’s inequality. **2010**, arXiv:1009.4004.
14. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904.
15. Nock, R.; Luosto, P.; Kivinen, J. Mixed Bregman clustering with approximation guarantees. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 15–19 September 2008; Springer-Verlag: Berlin/Heidelberg, Germany, 2008; pp. 154–169.
16. Amari, S. Integration of Stochastic Models by Minimizing α -Divergence. *Neural Comput.* **2007**, *19*, 2780–2796.
17. Arthur, D.; Vassilvitskii, S. k -means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, LA, USA, 7–9 January 2007; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007; pp. 1027–1035.
18. Olszewski, D.; Ster, B. Asymmetric clustering using the alpha-beta divergence. *Pattern Recognit.* **2014**, *47*, 2031–2041.
19. Amari, S. Alpha-divergence is unique, belonging to both f -divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
20. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
21. Teboulle, M. A unified continuous optimization framework for center-based clustering methods. *J. Mach. Learn. Res.* **2007**, *8*, 65–102.
22. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
23. Morimoto, T. Markov Processes and the H -theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331.
24. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 131–142.
25. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studi. Sci. Math. Hung.* **1967**, *2*, 229–318.

26. Cichocki, A.; Cruces, S.; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
27. Zhu, H.; Rohwer, R. Measurements of generalisation based on information geometry. In *Mathematics of Neural Networks*; Operations Research/Computer Science Interfaces Series; Ellacott, S., Mason, J., Anderson, I., Eds.; Springer: New York, NY, USA, 1997; Volume 8, pp. 394–398.
28. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507.
29. Nielsen, F. An information-geometric characterization of Chernoff information. *IEEE Signal Process. Lett.* **2013**, *20*, 269–272.
30. Wu, J.; Rehg, J. Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In Proceedings of 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 630–637.
31. Bhattacharya, A.; Jaiswal, R.; Ailon, N. A tight lower bound instance for k -means++ in constant dimension. In *Theory and Applications of Models of Computation*; Lecture Notes in Computer Science; Gopal, T., Agrawal, M., Li, A., Cooper, S., Eds.; Springer International Publishing: New York, NY, USA, 2014; Volume 8402, pp. 7–22.
32. Nielsen, F. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Process. Lett.* **2013**, *20*, 657–660.
33. Ben-Tal, A.; Charnes, A.; Teboulle, M. Entropic means. *J. Math. Anal. Appl.* **1989**, *139*, 537–551.
34. Nielsen, F.; Nock, R. The dual Voronoi diagrams with respect to representational Bregman divergences. In Proceedings of International Symposium on Voronoi Diagrams (ISVD), Copenhagen, Denmark, 23–26 June 2009; pp. 71–78.
35. Heinz, E. Beiträge zur Störungstheorie der Spektralzerlegung. *Math. Anna.* **1951**, *123*, 415–438. (in German)
36. Besenyei, A. On the invariance equation for Heinz means. *Math. Inequal. Appl.* **2012**, *15*, 973–979.
37. Barry, D.A.; Culligan-Hensley, P.J.; Barry, S.J. Real values of the W -function. *ACM Trans. Math. Softw.* **1995**, *21*, 161–171.
38. Veldhuis, R.N.J. The centroid of the symmetrical Kullback–Leibler distance. *IEEE Signal Process. Lett.* **2002**, *9*, 96–99.
39. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. **2009**, arXiv.org: 0911.4863.
40. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
41. Romberg, S.; Lienhart, R. Bundle min-hashing for logo recognition. In Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, Dallas, TX, USA, 16–19 April 2013; ACM: New York, NY, USA, 2013; pp. 113–120.
42. Matsuyama, Y. The alpha-EM algorithm: Surrogate likelihood maximization using alpha-logarithmic information measures. *IEEE Trans. Inf. Theory* **2003**, *49*, 692–706.

43. Amari, S.I. New developments of information geometry (26): Information geometry of convex programming and game theory. In *Mathematical Sciences (suurikagaku)*; Number 605; The Science Company: Denver, CO, USA, 2013; pp. 65–74. (In Japanese)

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).