On the symmetrical Kullback-Leibler Jeffreys centroids

Frank Nielsen
Sony Computer Science Laboratories, Inc.
3-14-13 Higashi Gotanda, 141-0022 Shinagawa-ku, Tokyo, Japan

Abstract

Due to the success of the bag-of-word modeling paradigm, clustering histograms has become an important ingredient of modern information processing. Clustering histograms can be performed using the celebrated k-means centroid-based algorithm. From the viewpoint of applications, it is usually required to deal with symmetric distances. In this letter, we consider the Jeffreys divergence that symmetrizes the Kullback-Leibler divergence, and investigate the computation of Jeffreys centroids. We first prove that the Jeffreys centroid can be expressed *analytically* using the Lambert W function for *positive* histograms. We then show how to obtain a fast guaranteed approximation when dealing with *frequency* histograms. Finally, we conclude with some remarks on the k-means histogram clustering.

1 Introduction: Motivation and prior work

1.1 Motivation: The Bag-of-Word modeling paradigm

Classifying documents into categories is a common task of information retrieval systems: Given a training set of documents labeled with categories, one asks to classify incoming new documents. Text categorization [1] proceeds by first defining a dictionary of words (the corpus). It then models each document by a *word count* yielding a word histogram per document. Defining a proper distance $d(\cdot,\cdot)$ between histograms allows one to:

- Classify a new on-line document: we first calculate its histogram signature and then seek for the labeled document which has the *most similar* histogram to deduce its tag (*e.g.*, using a nearest neighbor classifier).
- Find the initial set of categories: we cluster all document histograms and assign a category per cluster.

It has been shown experimentally that the Jeffreys divergence (symmetrizing the Kullback-Leibler divergence) achieves better performance than the traditional *tf-idf*

method [1]. This text classification method based on the Bag of Words (BoWs) representation has also been instrumental in computer vision for efficient object categorization [2] and recognition in natural images. It first requires to create a dictionary of "visual words" by quantizing keypoints of the training database. Quantization is then performed using the k-means algorithm [8] that partitions n data points $\mathcal{X} = \{x_1, ..., x_n\}$ into k clusters $\mathcal{C}_1, ..., \mathcal{C}_k$ where each data element belongs to the closest cluster center. From a given initialization, Lloyd's batched k-means first assigns points to their closest cluster centers, and then update the cluster centers, and reiterate this process until convergence is met after a finite number of steps. When the distance function $d(x, y)$ is chosen as the squared Euclidean distance $d(x, y) = \|x - y\|^2$, the cluster centroids updates to their centers of mass. Csurka et al. [2] used the squared Euclidean distance for building the visual vocabulary. Depending on the considered features, other distances have proven useful: For example, the Jeffreys divergence was shown to perform experimentally better than the Euclidean or squared Euclidean distances for Compressed Histogram of Gradient descriptors [3]. To summarize, k-means histogram clustering with respect to the Jeffreys divergence can be used to both quantize visual words to create a dictionary and to cluster document words for assigning initial categories.

Let $w_h = \sum_{i=1}^{d} h^i$ denote the cumulative sum of the bin values of histogram h. We distinguish between *positive histograms* and *frequency histograms*. A frequency histogram $\tilde{h}$ is a unit histogram (*i.e.*, the cumulative sum $w_{\tilde{h}}$ of its bins adds up to one). In statistics, those positive and frequency histograms correspond respectively to *positive discrete* and *multinomial* distributions when all bins are non-empty. Let $\mathcal{H} = \{h_1, ..., h_n\}$ be a collection of n histograms with d positive-valued bins. By notational convention, we use the superscript and the subscript to indicate the bin number and the histogram number, respectively. Without loss of generality, we assume that all bins are non-empty[1]: $h_j^i \geq 0, 1 \leq j \leq n, 1 \leq i \leq d$. To measure the distance between two such histograms p and q, we rely on the *relative entropy*. The *extended KL divergence* [8] between two positive (but not necessarily normalized) histograms p and q is defined by $\mathrm{KL}(p : q) = \sum_{i=1}^{d} p^i \log \frac{p^i}{q^i} + q^i - p^i$. Observe that this information-theoretic dissimilarity measure is not symmetric nor does it satisfy the triangular inequality property of metrics. Let $\tilde{p} = \frac{p}{\sum_{i=1}^{d} p^i}$ and $\tilde{q} = \frac{q}{\sum_{i=1}^{d} q^i}$ denote the corresponding normalized frequency histograms. In the remainder, the $\tilde{}$ denotes this normalization operator. The extended KL divergence formula applied to normalized histograms yields the traditional KL divergence [8]: $\mathrm{KL}(\tilde{p} : \tilde{q}) = \sum_{i=1}^{d} \tilde{p}^i \log \frac{\tilde{p}^i}{\tilde{q}^i}$ since $\sum_{i=1}^{d} \tilde{q}^i - \tilde{p}^i = \sum_{i=1}^{d} \tilde{q}^i - \sum_{i=1}^{d} \tilde{p}^i = 1 - 1 = 0$. The KL divergence is interpreted as the *relative entropy* between $\tilde{p}$ and $\tilde{q}$: $\mathrm{KL}(\tilde{p} : \tilde{q}) = H^{\times}(\tilde{p} : \tilde{q}) - H(\tilde{p})$, where $H^{\times}(\tilde{p} : \tilde{q}) = \sum_{i=1}^{d} \tilde{p}^i \log \frac{1}{\tilde{q}^i}$ denotes the *cross-entropy* and $H(\tilde{p}) = H^{\times}(\tilde{p} : \tilde{p}) = \sum_{i=1}^{d} \tilde{p}^i \log \frac{1}{\tilde{p}^i}$ is the *Shannon entropy*. This distance is explained as the expected extra number of bits per datum that must be transmitted when using the "wrong" distribution $\tilde{q}$ instead of the true distribution $\tilde{p}$. Often $\tilde{p}$ is hidden by nature and need to be hypothesized while $\tilde{q}$ is estimated. When clustering histograms, all histograms play the *same role*, and it is therefore better to consider the

[1]Otherwise, we may add an arbitrary small quantity $\epsilon > 0$ to all bins. When frequency histograms are required, we then re-normalize.

Jeffreys [4] divergence $J(p,q) = \mathrm{KL}(p:q) + \mathrm{KL}(q:p)$ that symmetrizes the KL divergence:

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

Observe that the formula for Jeffreys divergence holds for arbitrary positive histograms (including frequency histograms).

This letter is devoted to compute efficiently the *Jeffreys centroid* c of a set $\mathcal{H} = \{h_1, ..., h_n\}$ of weighted histograms defined as:

$$c = \arg\min_x \sum_{j=1}^{n} \pi_j J(h_j, x), \tag{2}$$

where the π_j's denote the histogram positive weights (with $\sum_{j=1}^{n} \pi_j = 1$). When all histograms $h_j \in \mathcal{H}$ are normalized, we require the minimization of x to be carried out over Δ_d, the $(d-1)$-dimensional probability simplex. This yields the *Jeffreys frequency centroid* $\tilde{c}$. Otherwise, for positive histograms $h_j \in \mathcal{H}$, the minimization of x is done over the positive orthant $\mathbb{R}^d_+$, to get the *Jeffreys positive centroid* c. Since the J-divergence is convex in both arguments, both the Jeffreys positive and frequency centroids are unique.

1.2 Prior work and contributions

On one hand, clustering histograms has been studied using various distances and clustering algorithms. Besides the classical Minkowski ℓ_p-norm distances, hierarchical clustering with respect to the χ^2 distance has been investigated in [7]. Banerjee et al. [8] generalized k-means to Bregman k-means thus allowing to cluster distributions of the same exponential families with respect to the KL divergence. Mignotte [9] used k-means with respect to the Bhattacharyya distance [10] on histograms of various color spaces to perform image segmentation. On the other hand, Jeffreys k-means has not been yet extensively studied as the centroid computations are non-trivial: In 2002, Veldhuis [11] reported an iterative Newton-like algorithm to approximate arbitrarily finely the Jeffreys frequency centroid $\tilde{c}$ of a set of frequency histograms that requires two nested loops. Nielsen and Nock [12] considered the information-geometric structure of the manifold of multinomials (frequency histograms) to report a simple geodesic bisection search algorithm (*i.e.*, replacing the two nested loops of [11] by one single loop). Indeed, the family of frequency histograms belongs to the exponential families [8], and the Jeffreys frequency centroid amount to compute equivalently a symmetrized Bregman centroid [12].

To overcome the explicit computation of the Jeffreys centroid, Nock et al. [13] generalized the Bregman k-means [8] and k-means++ seeding using *mixed Bregman divergences*: They consider two dual centroids $\tilde{c}_m$ and $\tilde{c}^*_m$ attached per cluster, and use the following divergence depending on these two centers: $\Delta\mathrm{KL}(\tilde{c}_m : x : \tilde{c}^*_m) = \mathrm{KL}(\tilde{c}_m : x) + \mathrm{KL}(x : \tilde{c}^*_m)$. However, note that this mixed Bregman 2-centroid-per-cluster clustering is *different* from the Jeffreys k-means clustering that relies on one centroid per cluster.

This letter is organized as follows: Section 2 reports a closed-form expression of the positive Jeffreys centroid for a set of positive histograms. Section 3 studies the guaranteed tight approximation factor obtained when normalizing the positive Jeffreys centroid, and further describes a simple bisection algorithm to arbitrarily finely approximate the optimal Jeffreys frequency centroid. Section 4 reports on our experimental results that show that our normalized approximation is in practice tight enough to avoid doing the bisection process. Finally, Section 5 concludes this work.

2 Jeffreys positive centroid

We consider a set $\mathcal{H} = \{h_1, ..., h_n\}$ of n positive weighted histograms with d non-empty bins ($h_j \in \mathbb{R}^d_+$, $\pi_j > 0$ and $\sum_j \pi_j = 1$). The *Jeffreys positive centroid* c is defined by:

$$c = \arg\min_{x \in \mathbb{R}^d_+} J(\mathcal{H}, x) = \arg\min_{x \in \mathbb{R}^d_+} \sum_{j=1}^{n} \pi_j J(h_j, x). \tag{3}$$

We state the first result:

Theorem 1 *The Jeffreys positive centroid $c = (c^1, ..., c^d)$ of a set $\{h_1, ..., h_n\}$ 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.*

Proof We seek for $x \in \mathbb{R}^d_+$ that minimizes Eq. 3. After expanding Jeffreys divergence formula of Eq. 1 in Eq. 3 and removing all additive terms independent of x, we find the following equivalent minimization problem:

$$\min_{x \in \mathbb{R}^d_+} \sum_{i=1}^{d} x^i \log \frac{x^i}{g^i} - a^i \log x^i.$$

This optimization can be performed coordinate-wise, independently. For each coordinate, dropping the superscript notation and setting the derivative to zero, we have to solve $\log \frac{x}{g} + 1 - \frac{a}{x} = 0$, which yields $x = \frac{a}{W(\frac{a}{g}e)}$, where $W(\cdot)$ denotes the Lambert W function [14].

Lambert function[2] W is defined by $W(x)e^{W(x)} = x$ for $x \geq 0$. That is, the Lambert function is the functional inverse of $f(x) = xe^x = y$: $x = W(y)$. Although function W may seem non-trivial at first sight, it is a popular elementary analytic function similar to the logarithm or exponential functions. In practice, we get a fourth-order convergence algorithm to estimate it by implementing Halley's numerical root-finding method. It requires fewer than 5 iterations to reach machine accuracy using the IEEE 754 floating point standard [14]. Notice that the Lambert W function plays a particular role in information theory [15].

[2]We consider only the branch W_0 [14] since arguments of the function are always positive.

3 Jeffreys frequency centroid

We consider a set $\tilde{H}$ of n frequency histograms: $\tilde{H} = \{\tilde{h}_1, ..., \tilde{h}_n\}$.

3.1 A guaranteed approximation

If we relax x to the positive orthant $\mathbb{R}^d_+$ instead of the probability simplex, we get the optimal positive Jeffreys centroid c, with $c^i = \frac{a^i}{W(\frac{a^i}{g^i}e)}$ (Theorem 1). Normalizing this positive Jeffreys centroid to get $\tilde{c}' = \frac{c}{w_c}$ *does not* yield the Jeffreys frequency centroid $\tilde{c}$ that requires dedicated iterative optimization algorithms [11, 12]. In this section, we consider approximations of the Jeffreys frequency histograms. Veldhuis [11] approximated the Jeffreys frequency centroid $\tilde{c}$ by $\tilde{c}'' = \frac{\tilde{a}+\tilde{g}}{2}$, where $\tilde{a}$ and $\tilde{g}$ denotes the normalized weighted arithmetic and geometric means, respectively. The normalized geometric weighted mean $\tilde{g} = (\tilde{g}^1, ..., \tilde{g}^d)$ is defined by $\tilde{g}^i = \frac{\prod_{j=1}^n (\tilde{h}^i_j)^{\pi_j}}{\sum_{i=1}^d \prod_{j=1}^n (\tilde{h}^i_j)^{\pi_j}}, i \in \{1, ..., d\}$. Since $\sum_{i=1}^d \sum_{j=1}^n \pi_j \tilde{h}^i_j = \sum_{j=1}^n \pi_j \sum_{i=1}^d \tilde{h}^i_j = \sum_{j=1}^n \pi_j = 1$, the normalized arithmetic weighted mean has coordinates: $\tilde{a}^i = \sum_{j=1}^n \pi_j \tilde{h}^i_j$.

We consider approximating the Jeffreys frequency centroid by normalizing the Jeffreys positive centroid c: $\tilde{c}' = \frac{c}{w_c}$.

We start with a simple lemma:

Lemma 1 *The cumulative sum w_c of the bin values of the Jeffreys positive centroid c of a set of frequency histograms is less or equal to one: $0 < w_c \leq 1$.*

Proof Consider the frequency histograms $\tilde{H}$ as positive histograms. It follows from Theorem 1 that the Jeffreys positive centroid c is such that $w_c = \sum_{i=1}^d c^i = \sum_{i=1}^d \frac{a^i}{W(\frac{a^i}{g^i}e)}$. Now, the arithmetic-geometric mean inequality states that $a^i \geq g^i$ where a^i and g^i denotes the coordinates of the arithmetic and geometric positive means. Therefore $W(\frac{a^i}{g^i}e) \geq 1$ and $c^i \leq a^i$. Thus $w_c = \sum_{i=1}^d c^i \leq \sum_{i=1}^d a^i = 1$.

We consider approximating Jeffreys frequency centroid on the probability simplex Δ_d by using the normalization of the Jeffreys positive centroid: $\tilde{c}' = \frac{a^i}{wW(\frac{a^i}{g^i}e)}$, with $w = \sum_{i=1}^d \frac{a^i}{W(\frac{a^i}{g^i}e)}$. To study the quality of this approximation, we use the following lemma:

Lemma 2 *For any histogram x and frequency histogram $\tilde{h}$, we have $J(x, \tilde{h}) = J(\tilde{x}, \tilde{h}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{h}) + \log w_x)$, where w_x denotes the normalization factor ($w_x = \sum_{i=1}^d x^i$).*

Proof It follows from the definition of Jeffreys divergence and the fact that $x^i = w_x \tilde{x}^i$ that $J(x, \tilde{h}) = \sum_{i=1}^d (w_x \tilde{x}^i - \tilde{h}^i) \log \frac{w_x \tilde{x}^i}{\tilde{h}^i}$. Expanding and mathematically rewriting the rhs. yields $J(x, \tilde{h}) = \sum_{i=1}^d (w_x \tilde{x}^i \log \frac{\tilde{x}^i}{\tilde{h}^i} + w_x \tilde{x}^i \log w_x + \tilde{h}^i \log \frac{\tilde{h}^i}{\tilde{x}^i} - \tilde{h}^i \log w_x) = (w_x - 1) \log w_x + J(\tilde{x}, \tilde{h}) + (w_x - 1) \sum_{i=1}^d \tilde{x}^i \log \frac{\tilde{x}^i}{\tilde{h}^i} = J(\tilde{x}, \tilde{h}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{h}) + \log w_x)$, since $\sum_{i=1}^d \tilde{h}^i = \sum_{i=1}^d \tilde{x}^i = 1$.

The lemma can be extended to a set of weighted frequency histograms $\tilde{\mathcal{H}}$:

$$J(x, \tilde{H}) = J(\tilde{x}, \tilde{H}) + (w_x - 1)(\mathrm{KL}(\tilde{x} : \tilde{H}) + \log w_x),$$

where $J(x, \tilde{H}) = \sum_{j=1}^{n} \pi_j J(x, \tilde{h}_j)$ and $\mathrm{KL}(\tilde{x} : \tilde{H}) = \sum_{j=1}^{n} \pi_j \mathrm{KL}(\tilde{x}, \tilde{h}_j)$ (with $\sum_{j=1}^{n} \pi_j = 1$).

We state the second theorem concerning our guaranteed approximation:

Theorem 2 *Let $\tilde{c}$ denote the Jeffreys frequency centroid and $\tilde{c}' = \frac{c}{w_c}$ the normalized Jeffreys positive centroid. Then the approximation factor $\alpha_{\tilde{c}'} = \frac{J(\tilde{c}', \tilde{H})}{J(\tilde{c}, \tilde{H})}$ is such that $1 \leq \alpha_{\tilde{c}'} \leq 1 + (\frac{1}{w_c} - 1)\frac{\mathrm{KL}(c, \tilde{H})}{J(c, \tilde{H})} \leq \frac{1}{w_c}$ (with $w_c \leq 1$).*

Proof We have $J(c, \tilde{H}) \leq J(\tilde{c}, \tilde{H}) \leq J(\tilde{c}', \tilde{H})$. Using Lemma 2, since $J(\tilde{c}', \tilde{H}) = J(c, \tilde{H}) + (1 - w_c)(\mathrm{KL}(\tilde{c}', \tilde{H}) + \log w_c))$ and $J(c, \tilde{H}) \leq J(\tilde{c}, \tilde{H})$, it follows that $1 \leq \alpha_{\tilde{c}'} \leq 1 + \frac{(1-w_c)(\mathrm{KL}(\tilde{c}', \tilde{H}) + \log w_c)}{J(\tilde{c}, \tilde{H})}$. We also have $\mathrm{KL}(\tilde{c}' : \tilde{H}) = \frac{1}{w_c}\mathrm{KL}(c, \tilde{H}) - \log w_c$ (by expanding the KL expression and using the fact that $w_c = \sum_i c^i$). Therefore $\alpha_{\tilde{c}'} \leq 1 + \frac{(1-w_c)\mathrm{KL}(c, \tilde{H})}{w_c J(\tilde{c}, \tilde{H})}$. Since $J(\tilde{c}, \tilde{H}) \geq J(c, \tilde{H})$ and $\mathrm{KL}(c, \tilde{H}) \leq J(c, \tilde{H})$, we finally obtain $\alpha_{\tilde{c}'} \leq \frac{1}{w_c}$.

When $w_c = 1$ the bound is tight. Experimental results described in the next section shows that this normalized Jeffreys positive centroid $\tilde{c}'$ almost coincide with the Jeffreys frequency centroid.

3.2 Arbitrary fine approximation by bisection search

It has been shown in [11, 12] that minimizing Eq. 2 over the probability simplex Δ_d amounts to minimize the following equivalent problem:

$$\tilde{c} = \arg\min_{\tilde{x} \in \Delta_d} \mathrm{KL}(\tilde{a} : \tilde{x}) + \mathrm{KL}(\tilde{x} : \tilde{g}), \tag{4}$$

Nevertheless, instead of using the two-nested loops of Veldhuis' Newton scheme [11], we can design a single loop optimization algorithm. We consider the Lagrangian function obtained by enforcing the normalization constraint $\sum_i c^i = 1$ similar to [11]. For each coordinate, setting the derivative with respect to $\tilde{c}^i$, we get $\log \frac{\tilde{c}^i}{\tilde{g}^i} + 1 - \frac{\tilde{a}^i}{\tilde{c}^i} + \lambda = 0$, which solves as $\tilde{c}^i = \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)}$. By multiplying these d constraints with $\tilde{c}^i$ respectively and summing up, we deduce that $\lambda = -\mathrm{KL}(\tilde{c} : \tilde{g}) \leq 0$ (also noticed by [11]). From the constraints that all c_i's should be less than one, we bound λ as follows: $c^i = \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)} \leq 1$, which solves for equality when $\lambda = \log(e^{\tilde{a}^i}\tilde{g}^i) - 1$. Thus we seek for $\lambda \in [\max_i \log(e^{\tilde{a}^i}\tilde{g}^i) - 1, 0]$. Since $s = \sum_i c^i = 1$, we have the following cumulative sum equation depending on the unknown parameter λ: $s(\lambda) = \sum_i c^i(\lambda) = \sum_{i=1}^{d} \frac{\tilde{a}^i}{W\left(\frac{\tilde{a}^i e^{\lambda+1}}{\tilde{g}^i}\right)}$. This is a monotonously decreasing function with $s(0) \leq 1$. We can thus perform a simple bisection search to approximate the optimal value of λ, and therefore deduce an arbitrary fine approximation of the Jeffreys frequency centroid.

	α_c(opt. positive)	$\alpha_{\tilde{c}'}$(n′lized approx.)	$w_c \leq 1$(n′lizing coeff.)	$\alpha_{\tilde{c}''}$(Veldhuis)
avg	0.9648680345638155	**1.0002205080964255**	0.9338228644308926	1.065590178484613
min	0.906414219584823	**1.0000005079528809**	0.8342819488534723	1.0027707382095195
max	0.9956399220678585	**1.0000031489541772**	0.9931975105809021	1.3582296675397754

Table 1: Experimental performance ratio and statistics for the 30000+ images of the Caltech-256 database. Observe that $\alpha_c = \frac{J(\mathcal{H},c)}{J(\mathcal{H},\tilde{c})} \leq 1$ since the positive Jeffreys centroid (available in closed-form) minimizes the average Jeffreys divergence criterion. Our guaranteed normalized approximation $\tilde{c}'$ is almost optimal. Veldhuis' simple half normalized arithmetic-geometric approximation performs on average with a 6.56% error but can be far from the optimal in the worst-case (35.8%).

4 Experimental results and discussion

We used a multi-precision floating point (`http://www.apfloat.org/`) package to handle calculations and control arbitrary precisions. We chose the Caltech-256 database [16] consisting of 30607 images labeled into 256 categories to perform experiments: We consider the set of intensity[3] histograms $\mathcal{H}$. For each of the 256 category, we consider the set of histograms falling inside this category and compute the exact Jeffreys positive centroids c, its normalization Jeffreys approximation $\tilde{c}'$ and optimal frequency centroids $\tilde{c}$. We also consider the average of the arithmetic and geometric normalized means $\tilde{c}'' = \frac{\tilde{a}+\tilde{g}}{2}$. We evaluate the average, minimum and maximum ratio $\alpha_x = \frac{J(\mathcal{H},x)}{J(\mathcal{H},\tilde{c})}$ for $x \in \{c, \tilde{c}', \tilde{c}''\}$. The results are reported in Table 1. Furthermore, to study the best/worst/average performance of the the normalized Jeffreys positive centroid $\tilde{c}'$, we ran 10^6 trials as follows: We draw two random binary histograms ($d = 2$), calculate a fine precision approximation of $\tilde{c}$ using numerical optimization, and calculate the approximation obtained by using the normalized closed-form centroid $\tilde{c}'$. We gather statistics on the ratio $\alpha = \frac{J(\tilde{c}')}{J(\tilde{c})} \geq 1$. We find experimentally the following performance: $\bar{\alpha} \sim 1.0000009, \alpha_{\max} \sim 1.00181506, \alpha_{\min} = 1.000000$. Although $\tilde{c}'$ is almost matching $\tilde{c}$ in those two real-world and synthetic experiments, it remains open to express analytically and exactly its worst-case performance.

5 Conclusion

We summarize the two main contributions of this paper: (1) we proved that the Jeffreys positive centroid admits a closed-form formula expressed using the Lambert W function, and (2) we proved that normalizing this Jeffreys positive centroid yields a tight guaranteed approximation to the Jeffreys frequency centroid. We noticed experimentally that the closed-form normalized Jeffreys positive centroid almost coincide with the Jeffreys frequency

[3]Converting RGB color pixels to $0.3R + 0.596G + 0.11B$ I grey pixels.

centroid, and can therefore be used in Jeffreys k-means clustering. Notice that since the k-means assignment/relocate algorithm monotonically converges even if instead of computing the exact cluster centroids we update it with provably better centroids (i.e., by applying one bisection iteration of Jeffreys frequency centroid computation), we end up with a converging *variational* Jeffreys frequency k-means that requires to implement a stopping criterion. Jeffreys divergence is not the only way to symmetrize the Kullback-Leibler divergence. Other KL symmetrizations include the Jensen-Shannon divergence [5], the Chernoff divergence [6], and a smooth family of symmetric divergences including the Jensen-Shannon and Jeffreys divergences [17].

References

[1] B. Bigi, "Using Kullback-Leibler distance for text categorization," in *Proceedings of the 25th European conference on IR research (ECIR)*, Springer-Verlag, 2003, pp. 305–319.

[2] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," *Workshop on Statistical Learning in Computer Vision (ECCV)*, pp. 1–22, 2004.

[3] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. A. Reznik, R. Grzeszczuk, and B. Girod, "Compressed histogram of gradients: A low-bitrate descriptor," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 384–399, 2012.

[4] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London*, vol. 186, no. 1007, pp. 453–461, 1946.

[5] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, 1991.

[6] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[7] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (TAI)*. 1995, pp. 388–391.

[8] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[9] M. Mignotte, "Segmentation by fusion of histogram-based k-means clusters in different color spaces," *IEEE Transactions on Image Processing (TIP)*, vol. 17, no. 5, pp. 780–787, 2008.

[10] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.

[11] R. N. J. Veldhuis, "The centroid of the symmetrical Kullback-Leibler distance," *IEEE signal processing letters*, vol. 9, no. 3, pp. 96–99, 2002.

[12] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2048–2059, June 2009.

[13] R. Nock, P. Luosto, and J. Kivinen, "Mixed Bregman clustering with approximation guarantees," in *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*. Springer-Verlag, 2008, pp. 154–169.

[14] D. A. Barry, P. J. Culligan-Hensley, and S. J. Barry, "Real values of the W-function," *ACM Trans. Math. Softw.*, vol. 21, no. 2, pp. 161–171, 1995.

[15] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Problems of Information Transmission*, vol 34, no 2, pp. 142–146, 1998.

[16] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007.

[17] F. Nielsen, "A family of statistical symmetric divergences based on Jensen's inequality," CoRR 1009.4004, 2010.