entropy

Article

Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities

Frank Nielsen [1,2,*] and Ke Sun [3]

[1] Computer Science Department LIX, École Polytechnique, 91128 Palaiseau Cedex, France
[2] Sony Computer Science Laboratories Inc, Tokyo 141-0022, Japan
[3] King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia; sunk.edu@gmail.com
* Correspondence: Frank.Nielsen@acm.org; Tel.: +33-1-7757-8070

Academic Editor: Antonio M. Scarfone
Received: 20 October 2016; Accepted: 5 December 2016; Published: 9 December 2016

Abstract: Information-theoretic measures, such as the entropy, the cross-entropy and the Kullback–Leibler divergence between two mixture models, are core primitives in many signal processing tasks. Since the Kullback–Leibler divergence of mixtures provably does not admit a closed-form formula, it is in practice either estimated using costly Monte Carlo stochastic integration, approximated or bounded using various techniques. We present a fast and generic method that builds algorithmically closed-form lower and upper bounds on the entropy, the cross-entropy, the Kullback–Leibler and the α-divergences of mixtures. We illustrate the versatile method by reporting our experiments for approximating the Kullback–Leibler and the α-divergences between univariate exponential mixtures, Gaussian mixtures, Rayleigh mixtures and Gamma mixtures.

Keywords: information geometry; mixture models; α-divergences; log-sum-exp bounds

1. Introduction

Mixture models are commonly used in signal processing. A typical scenario is to use mixture models [1–3] to smoothly model histograms. For example, Gaussian Mixture Models (GMMs) can be used to convert grey-valued images into binary images by building a GMM fitting the image intensity histogram and then choosing the binarization threshold as the average of the Gaussian means [1]. Similarly, Rayleigh Mixture Models (RMMs) are often used in ultrasound imagery [2] to model histograms, and perform segmentation by classification. When using mixtures, a fundamental primitive is to define a proper statistical distance between them. The Kullback–Leibler (KL) divergence [4], also called relative entropy or information discrimination, is the most commonly used distance. Hence the main target of this paper is to faithfully measure the KL divergence. Let $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{i=1}^{k'} w'_i p'_i(x)$ be two finite statistical density mixtures of k and k' components, respectively. Notice that the Cumulative Density Function (CDF) of a mixture is like its density also a convex combinations of the component CDFs. However, beware that a mixture is not a sum of random variables (RVs). Indeed, sums of RVs have convolutional densities. In statistics, the mixture components $p_i(x)$ are often parametric: $p_i(x) = p(x;\theta_i)$, where θ_i is a vector of parameters. For example, a mixture of Gaussians (MoG also used as a shortcut instead of GMM) has each component distribution parameterized by its mean μ_i and its covariance matrix Σ_i (so that the parameter vector is $\theta_i = (\mu_i, \Sigma_i)$). Let $\mathcal{X} = \{x \in \mathbb{R} \,:\, p(x;\theta) > 0\}$ be the support of the component distributions. Denote by $H_\times(m, m') = -\int_{\mathcal{X}} m(x) \log m'(x) \mathrm{d}x$ the cross-entropy [4] between two continuous mixtures of densities m and m', and denote by $H(m) = H_\times(m, m) = \int_\mathcal{X} m(x) \log \frac{1}{m(x)} \mathrm{d}x = -\int_\mathcal{X} m(x) \log m(x) \mathrm{d}x$ the Shannon entropy [4]. Then the Kullback–Leibler divergence between m and m' is given by:

$$\mathrm{KL}(m : m') = H_\times(m, m') - H(m) = \int_\mathcal{X} m(x) \log \frac{m(x)}{m'(x)} \mathrm{d}x \geq 0. \tag{1}$$

The notation ":" is used instead of the usual comma "," notation to emphasize that the distance is not a metric distance since neither is it symmetric ($\mathrm{KL}(m : m') \neq \mathrm{KL}(m' : m)$), nor does it satisfy the triangular inequality [4] of metric distances ($\mathrm{KL}(m : m') + \mathrm{KL}(m' : m'') \not\geq \mathrm{KL}(m : m'')$). When the natural base of the logarithm is chosen, we get a differential entropy measure expressed in nat units. Alternatively, we can also use the base-2 logarithm ($\log_2 x = \frac{\log x}{\log 2}$) and get the entropy expressed in bit units. Although the KL divergence is available in closed-form for many distributions (in particular as equivalent Bregman divergences for exponential families [5], see Appendix C), it was proven that the Kullback–Leibler divergence between two (univariate) GMMs is not analytic [6] (see also the particular case of a GMM of two components with the same variance that was analyzed in [7]). See Appendix A for an analysis. Note that the differential entropy may be negative. For example, the differential entropy of a univariate Gaussian distribution is $\log(\sigma\sqrt{2\pi e})$, and is therefore negative when the standard variance $\sigma < \frac{1}{\sqrt{2\pi e}} \approx 0.242$. We consider continuous distributions with entropies well-defined (entropy may be undefined for singular distributions like Cantor's distribution [8]).

1.1. Prior Work

Many approximation techniques have been designed to beat the computationally intensive Monte Carlo (MC) stochastic estimation: $\widehat{\mathrm{KL}}_s(m : m') = \frac{1}{s} \sum_{i=1}^{s} \log \frac{m(x_i)}{m'(x_i)}$ with $x_1, \ldots, x_s \sim m(x)$ (s independently and identically distributed (i.i.d.) samples $x_1, \ldots, x_s$). The MC estimator is asymptotically consistent, $\lim_{s \to \infty} \widehat{\mathrm{KL}}_s(m : m') = \mathrm{KL}(m : m')$, so that the "true value" of the KL of mixtures is estimated in practice by taking a very large sample (say, $s = 10^9$). However, we point out that the MC estimator gives as output a stochastic *approximation*, and therefore does not guarantee deterministic bounds (confidence intervals may be used). Deterministic lower and upper bounds of the integral can be obtained by various numerical integration techniques using quadrature rules. We refer to [9–12] for the current state-of-the-art approximation techniques and bounds on the KL of GMMs. The latest work for computing the entropy of GMMs is [13]. It considers arbitrary finely tuned bounds of the entropy of isotropic Gaussian mixtures (a case encountered when dealing with KDEs, kernel density estimators). However, there is a catch in the technique of [13]: It relies on solving the unique roots of some log-sum-exp equations (See Theorem 1 of [13], p. 3342) that do not admit a closed-form solution. Thus it is a hybrid method that contrasts with our combinatorial approach. Bounds of the KL divergence between mixture models can be generalized to bounds of the likelihood function of mixture models [14], because log-likelihood is just the KL between the empirical distribution and the mixture model up to a constant shift.

In information geometry [15], a mixture family of linearly independent probability distributions $p_1(x), ..., p_k(x)$ is defined by the convex combination of those non-parametric component distributions: $m(x; \eta) = \sum_{i=1}^{k} \eta_i p_i(x)$ with $\eta_i > 0$ and $\sum_{i=1}^{k} \eta_i = 1$. A mixture family induces a dually flat space where the Kullback–Leibler divergence is equivalent to a Bregman divergence [5,15] defined on the η-parameters. However, in that case, the Bregman convex generator $F(\eta) = \int m(x; \eta) \log m(x; \eta) \mathrm{d}x$ (the Shannon information) is not available in closed-form. Except for the family of multinomial distributions that is both a mixture family (with closed-form $\mathrm{KL}(m : m') = \sum_{i=1}^{k} m_i \log \frac{m_i}{m'_i}$, the discrete KL [4]) and an exponential family [15].

1.2. Contributions

In this work, we present a simple and efficient method that builds algorithmically a closed-form formula that guarantees both deterministic lower and upper bounds on the KL divergence within an additive factor of $\log k + \log k'$. We then further refine our technique to get improved adaptive bounds. For univariate GMMs, we get the non-adaptive bounds in $O(k \log k + k' \log k')$ time, and the adaptive bounds in $O(k^2 + k'^2)$ time. To illustrate our generic technique, we demonstrate it based on Exponential Mixture Models (EMMs), Gamma mixtures, RMMs and GMMs. We extend our preliminary results on KL divergence [16] to other information theoretical measures such as the differential entropy and α-divergences.

1.3. Paper Outline

The paper is organized as follows. Section 2 describes the algorithmic construction of the formula using piecewise log-sum-exp inequalities for the cross-entropy and the Kullback–Leibler divergence. Section 3 instantiates this algorithmic principle to the entropy and discusses related works. Section 4 extends the proposed bounds to the family of alpha divergences. Section 5 discusses an extension of the lower bound to f-divergences. Section 6 reports our experimental results on several mixture families. Finally, Section 7 concludes this work by discussing extensions to other statistical distances. Appendix A proves that the Kullback–Leibler divergence of mixture models is not analytic [6]. Appendix B reports the closed-form formula for the KL divergence between scaled and truncated distributions of the same exponential family [17] (that include Rayleigh, Gaussian and Gamma distributions among others). Appendix C shows that the KL divergence between two mixtures can be approximated by a Bregman divergence.

2. A Generic Combinatorial Bounding Algorithm Based on Density Envelopes

Let us bound the cross-entropy $H_\times(m : m')$ by deterministic lower and upper bounds, $L_\times(m : m') \leq H_\times(m : m') \leq U_\times(m : m')$, so that the bounds on the Kullback–Leibler divergence $\mathrm{KL}(m : m') = H_\times(m : m') - H_\times(m : m)$ follows as:

$$L_\times(m : m') - U_\times(m : m) \leq \mathrm{KL}(m : m') \leq U_\times(m : m') - L_\times(m : m). \tag{2}$$

Since the cross-entropy of two mixtures $\sum_{i=1}^{k} w_i p_i(x)$ and $\sum_{j=1}^{k'} w'_j p'_j(x)$:

$$H_\times(m : m') = -\int_{\mathcal{X}} \left(\sum_{i=1}^{k} w_i p_i(x) \right) \log \left(\sum_{j=1}^{k'} w'_j p'_j(x) \right) \mathrm{d}x \tag{3}$$

has a log-sum term of positive arguments, we shall use bounds on the log-sum-exp (lse) function [18,19]:

$$\mathrm{lse}\left(\{x_i\}_{i=1}^{l} \right) = \log \left(\sum_{i=1}^{l} e^{x_i} \right).$$

We have the following basic inequalities:

$$\max\{x_i\}_{i=1}^{l} < \mathrm{lse}\left(\{x_i\}_{i=1}^{l} \right) \leq \log l + \max\{x_i\}_{i=1}^{l}. \tag{4}$$

The left-hand-side (LHS) strict inequality holds because $\sum_{i=1}^{l} e^{x_i} > \max\{e^{x_i}\}_{i=1}^{l} = \exp\left(\max\{x_i\}_{i=1}^{l} \right)$ since $e^x > 0, \forall x \in \mathbb{R}$. The right-hand-side (RHS) inequality follows from the fact that $\sum_{i=1}^{l} e^{x_i} \leq l \max\{e^{x_i}\}_{i=1}^{l} = l \exp(\max\{x_i\}_{i=1}^{l})$, and equality holds if and only if $x_1 = \cdots = x_l$. The lse function is convex but not strictly convex, see exercise 7.9 [20]. It is known [21] that the conjugate of the lse function is the negative entropy restricted to the probability simplex. The lse function enjoys the following translation identity property: $\mathrm{lse}\left(\{x_i\}_{i=1}^{l} \right) = c + \mathrm{lse}\left(\{x_i - c\}_{i=1}^{l} \right), \forall c \in \mathbb{R}$. Similarly, we can also lower bound the lse function by $\log l + \min\{x_i\}_{i=1}^l$. We write equivalently that for l positive numbers $x_1, \ldots, x_l$,

$$\max\left\{\log\max\{x_i\}_{i=1}^l, \log l + \log\min\{x_i\}_{i=1}^l\right\} \leq \log\sum_{i=1}^l x_i \leq \log l + \log\max\{x_i\}_{i=1}^l. \tag{5}$$

In practice, we seek matching lower and upper bounds that minimize the bound gap. The gap of that ham-sandwich inequality in Equation (5) is $\min\{\log\frac{\max_i x_i}{\min_i x_i}, \log l\}$, which is upper bounded by $\log l$.

A mixture model $\sum_{j=1}^{k'} w'_j p'_j(x)$ must satisfy

$$\begin{aligned}\max\left\{\max\{\log w'_j p'_j(x)\}_{j=1}^{k'},\ \log k' + \min\{\log w'_j p'_j(x)\}_{j=1}^{k'}\right\}\\ \leq \log\left(\sum_{j=1}^{k'} w'_j p'_j(x)\right) \leq \log k' + \max\{\log w'_j p'_j(x)\}_{j=1}^{k'}\end{aligned} \tag{6}$$

point-wisely for any $x \in \mathcal{X}$. Therefore we shall bound the integral term $\int_{\mathcal{X}} m(x)\log\left(\sum_{j=1}^{k'} w'_j p'_j(x)\right)\mathrm{d}x$ in Equation (3) using piecewise lse inequalities where the min and max are kept unchanged. We get

$$L_\times(m : m') = -\int_{\mathcal{X}} m(x)\max\{\log w'_j p'_j(x)\}_{j=1}^{k'}\mathrm{d}x - \log k', \tag{7}$$

$$U_\times(m : m') = -\int_{\mathcal{X}} m(x)\max\left\{\min\{\log w'_j p'_j(x)\}_{j=1}^{k'} + \log k',\ \max\{\log w'_j p'_j(x)\}_{j=1}^{k'}\right\}\mathrm{d}x. \tag{8}$$

In order to calculate $L_\times(m : m')$ and $U_\times(m : m')$ efficiently using closed-form formula, let us compute the upper and lower envelopes of the k' real-valued functions $\{w'_j p'_j(x)\}_{j=1}^{k'}$ defined on the support $\mathcal{X}$, that is, $\mathcal{E}_{\mathrm{U}}(x) = \max\{w'_j p'_j(x)\}_{j=1}^{k'}$ and $\mathcal{E}_{\mathrm{L}}(x) = \min\{w'_j p'_j(x)\}_{j=1}^{k'}$. These envelopes can be computed exactly using techniques of computational geometry [22,23] provided that we can calculate the roots of the equation $w'_r p'_r(x) = w'_s p'_s(x)$, where $w'_r p'_r(x)$ and $w'_s p'_s(x)$ are a pair of weighted components. (Although this amounts to solve quadratic equations for Gaussian or Rayleigh distributions, the roots may not always be available in closed form, e.g. in the case of Weibull distributions.)

Let the envelopes be combinatorially described by ℓ elementary interval pieces in the form $I_r = (a_r, a_{r+1})$ partitioning the support $\mathcal{X} = \uplus_{r=1}^{\ell} I_r$ (with $a_1 = \min\mathcal{X}$ and $a_{\ell+1} = \max\mathcal{X}$). Observe that on each interval I_r, the maximum of the functions $\{w'_j p'_j(x)\}_{j=1}^{k'}$ is given by $w'_{\delta(r)} p'_{\delta(r)}(x)$, where $\delta(r)$ indicates the weighted component dominating all the others, i.e., the arg max of $\{w'_j p'_j(x)\}_{j=1}^{k'}$ for any $x \in I_r$, and the minimum of $\{w'_j p'_j(x)\}_{j=1}^{k'}$ is given by $w'_{\epsilon(r)} p'_{\epsilon(r)}(x)$.

To fix ideas, when mixture components are univariate Gaussians, the upper envelope $\mathcal{E}_{\mathrm{U}}(x)$ amounts to find equivalently the lower envelope of k' parabolas (see Figure 1) which has linear complexity, and can be computed in $O(k'\log k')$-time [24], or in output-sensitive time $O(k'\log\ell)$ [25], where ℓ denotes the number of parabola segments in the envelope. When the Gaussian mixture components have all the same weight and variance (e.g., kernel density estimators), the upper envelope amounts to find a lower envelope of cones: $\min_j |x - \mu'_j|$ (a Voronoi diagram in arbitrary dimension).

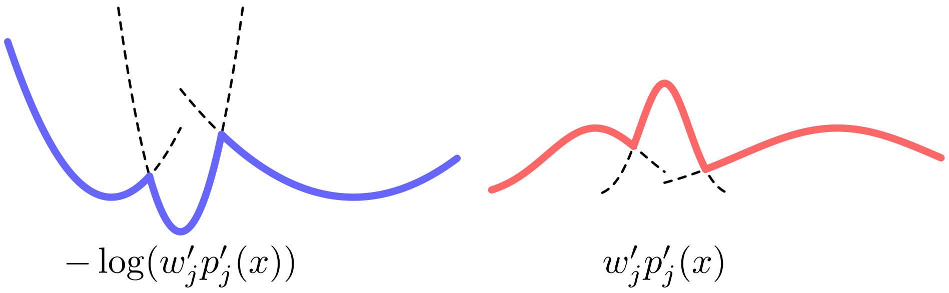


Figure 1. Lower envelope of parabolas corresponding to the upper envelope of weighted components of a Gaussian mixture with $k' = 3$ components.

To proceed once the envelopes have been built, we need to calculate two types of definite integrals on those elementary intervals: (i) the probability mass in an interval $\int_a^b p(x)\mathrm{d}x = \Phi(b) - \Phi(a)$ where Φ denotes the Cumulative Distribution Function (CDF); and (ii) the partial cross-entropy $-\int_a^b p(x) \log p'(x)\mathrm{d}x$ [26]. Thus let us define these two quantities:

$$C_{i,j}(a,b) = -\int_a^b w_i p_i(x) \log(w'_j p'_j(x))\mathrm{d}x, \tag{9}$$

$$M_i(a,b) = -\int_a^b w_i p_i(x)\mathrm{d}x. \tag{10}$$

By Equations (7) and (8), we get the bounds of $H_\times(m : m')$ as

$$L_\times(m : m') = \sum_{r=1}^{\ell} \sum_{s=1}^{k} C_{s,\delta(r)}(a_r, a_{r+1}) - \log k',$$
$$U_\times(m : m') = \sum_{r=1}^{\ell} \sum_{s=1}^{k} \min\left\{ C_{s,\delta(r)}(a_r, a_{r+1}),\ C_{s,\epsilon(r)}(a_r, a_{r+1}) - M_s(a_r, a_{r+1}) \log k' \right\}. \tag{11}$$

The size of the lower/upper bound formula depends on the envelope complexity ℓ, the number k of mixture components, and the closed-form expressions of the integral terms $C_{i,j}(a,b)$ and $M_i(a,b)$. In general, when a pair of weighted component densities intersect in at most p points, the envelope complexity is related to the Davenport–Schinzel sequences [27]. It is quasi-linear for bounded $p = O(1)$, see [27].

Note that in symbolic computing, the Risch semi-algorithm [28] solves the problem of computing indefinite integration in terms of elementary functions provided that there exists an oracle (hence the term "semi-algorithm") for checking whether an expression is equivalent to zero or not (however it is unknown whether there exists an algorithm implementing the oracle or not).

We presented the technique by bounding the cross-entropy (and entropy) to deliver lower/upper bounds on the KL divergence. When only the KL divergence needs to be bounded, we rather consider the ratio term $\frac{m(x)}{m'(x)}$. This requires to partition the support $\mathcal{X}$ into elementary intervals by overlaying the critical points of both the lower and upper envelopes of $m(x)$ and $m'(x)$, which can be done in linear time. In a given elementary interval, since $\max\{k \min_i\{w_i p_i(x)\}, \max_i\{w_i p_i(x)\}\} \leq m(x) \leq k \max_i\{w_i p_i(x)\}$, we then consider the inequalities:

$$\frac{\max\{k \min_i\{w_i p_i(x)\}, \max_i\{w_i p_i(x)\}\}}{k' \max_j\{w'_j p'_j(x)\}} \leq \frac{m(x)}{m'(x)} \leq \frac{k \max_i\{w_i p_i(x)\}}{\max\{k' \min_j\{w'_j p'_j(x)\}, \max_j\{w'_j p'_j(x)\}\}}. \tag{12}$$

We now need to compute definite integrals of the form $\int_a^b w_1 p(x;\theta_1) \log \frac{w_2 p(x;\theta_2)}{w_3 p(x;\theta_3)} \mathrm{d}x$ (see Appendix B for explicit formulas when considering scaled and truncated exponential families [17]). (Thus for exponential families, the ratio of densities removes the auxiliary carrier measure term.)

We call these bounds CELB and CEUB for Combinatorial Envelope Lower and Upper Bounds, respectively.

2.1. Tighter Adaptive Bounds

We shall now consider shape-dependent bounds improving over the additive $\log k + \log k'$ non-adaptive bounds. This is made possible by a decomposition of the lse function explained as follows. Let $t_i(x_1, \ldots, x_k) = \log\left(\sum_{j=1}^{k} e^{x_j - x_i}\right)$. By translation identity of the lse function,

$$\mathrm{lse}(x_1, \ldots, x_k) = x_i + t_i(x_1, \ldots, x_k) \tag{13}$$

for all $i \in [k]$. Since $e^{x_j - x_i} = 1$ if $j = i$, and $e^{x_j - x_i} > 0$, we have necessarily $t_i(x_1, \ldots, x_k) > 0$ for any $i \in [k]$. Since Equation (13) is an identity for all $i \in [k]$, we minimize the residual $t_i(x_1, \ldots, x_k)$ by maximizing x_i. Denoting by $x_{(1)}, \ldots, x_{(k)}$ the sequence of numbers sorted in non-decreasing order, the decomposition

$$\mathrm{lse}(x_1, \ldots, x_k) = x_{(k)} + t_{(k)}(x_1, \ldots, x_k) \tag{14}$$

yields the smallest residual. Since $x_{(j)} - x_{(k)} \leq 0$ for all $j \in [k]$, we have

$$t_{(k)}\left(x_1, \ldots, x_k\right) = \log\left(1 + \sum_{j=1}^{k-1} e^{x_{(j)} - x_{(k)}}\right) \leq \log k.$$

This shows the bounds introduced earlier can indeed be improved by a more accurate computation of the residual term $t_{(k)}\left(x_1, \ldots, x_k\right)$.

When considering 1D GMMs, let us now bound $t_{(k)}(x_1, \ldots, x_k)$ in a combinatorial range $I_r = (a_r, a_{r+1})$. Let $\delta = \delta(r)$ denote the index of the dominating weighted component in this range. Then,

$$\forall x \in I_r, \forall i, \quad \exp\left(-\log\sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log w_i\right) \leq \exp\left(-\log\sigma_\delta - \frac{(x-\mu_\delta)^2}{2\sigma_\delta^2} + \log w_\delta\right).$$

Thus we have:

$$\log m(x) = \log\frac{w_\delta}{\sigma_\delta\sqrt{2\pi}} - \frac{(x-\mu_\delta)^2}{2\sigma_\delta^2} + \log\left(1 + \sum_{i\neq\delta}\exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} + \log\frac{w_i}{\sigma_i} + \frac{(x-\mu_\delta)^2}{2\sigma_\delta^2} - \log\frac{w_\delta}{\sigma_\delta}\right)\right).$$

Now consider the ratio term:

$$\rho_{i,\delta}(x) = \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} + \log\frac{w_i\sigma_\delta}{w_\delta\sigma_i} + \frac{(x-\mu_\delta)^2}{2\sigma_\delta^2}\right).$$

It is maximized in $I_r = (a_r, a_{r+1})$ by maximizing equivalently the following quadratic equation:

$$l_{i,\delta}(x) = -\frac{(x-\mu_i)^2}{2\sigma_i^2} + \log\frac{w_i\sigma_\delta}{w_\delta\sigma_i} + \frac{(x-\mu_\delta)^2}{2\sigma_\delta^2}.$$

Setting the derivative to zero ($l'_{i,\delta}(x) = 0$), we get the root (when $\sigma_i \neq \sigma_\delta$)

$$x_{i,\delta} = \left(\frac{\mu_\delta}{\sigma_\delta^2} - \frac{\mu_i}{\sigma_i^2}\right) \Big/ \left(\frac{1}{\sigma_\delta^2} - \frac{1}{\sigma_i^2}\right).$$

If $x_{i,\delta} \in I_r$, the ratio $\rho_{i,\delta}(x)$ can be bounded in the slab I_r by considering the extreme values of the three element set $\{\rho_{i,\delta}(a_r), \rho_{i,\delta}(x_{i,\delta}), \rho_{i,\delta}(a_{r+1})\}$. Otherwise $\rho_{i,\delta}(x)$ is monotonic in I_r, its bounds in I_r

are given by $\{\rho_{i,\delta}(a_r), \rho_{i,\delta}(a_{r+1})\}$. In any case, let $\rho^{\min}_{i,\delta}(r)$ and $\rho^{\max}_{i,\delta}(r)$ represent the resulting lower and upper bounds of $\rho_{i,\delta}(x)$ in I_r. Then t_δ is bounded in the range I_r by:

$$0 < \log\left(1 + \sum_{i \neq \delta} \rho^{\min}_{i,\delta}(r)\right) \leq t_\delta \leq \log\left(1 + \sum_{i \neq \delta} \rho^{\max}_{i,\delta}(r)\right) \leq \log k.$$

In practice, we always get better bounds using the shape-dependent technique at the expense of computing overall $O(k^2)$ intersection points of the pairwise densities. We call those bounds CEALB and CEAUB for Combinatorial Envelope Adaptive Lower Bound and Combinatorial Envelope Adaptive Upper Bound.

Let us illustrate one scenario where this adaptive technique yields very good approximations. Consider a GMM with all variance σ^2 tending to zero (a mixture of k Diracs). Then in a combinatorial slab I_r, we have $\rho^{\max}_{i,\delta}(r) \to 0$ for all $i \neq \delta$, and therefore we get tight bounds.

As a related technique, we could also upper bound $\int_{a_r}^{a_{r+1}} \log m(x) \mathrm{d}x$ by $(a_{r+1} - a_r) \log m(a_r, a_{r+1})$ where $m(x, x')$ denotes the maximal value of the mixture density in the range (x, x'). This maximal value is either found at the slab extremities, or is a mode of the GMM. It then requires to find the modes of a GMM [29,30], for which no analytical solution is known in general.

2.2. Another Derivation Using the Arithmetic-Geometric Mean Inequality

Let us start by considering the inequality of arithmetic and geometric weighted means (AGI, Arithmetic-Geometric Inequality) applied to the mixture component distributions:

$$m(x) = \sum_{i=1}^{k} w_i p(x;\theta_i) \geq \prod_{i=1}^{k} p(x;\theta_i)^{w_i}$$

with equality holds iff. $\theta_1 = \ldots = \theta_k$.

To get a tractable formula with a positive remainder of the log-sum term $\log m(x)$, we need to have the log argument greater or equal to 1, and thus we shall write the positive remainder:

$$R(x) = \log\left(\frac{m(x)}{\prod_{i=1}^{k} p(x;\theta_i)^{w_i}}\right) \geq 0.$$

Therefore, we can decompose the log-sum into a tractable part and a remainder as:

$$\log m(x) = \sum_{i=1}^{k} w_i \log p(x;\theta_i) + \log\left(\frac{m(x)}{\prod_{i=1}^{k} p(x;\theta_i)^{w_i}}\right). \tag{15}$$

For exponential families, the first term can be integrated accurately. For the second term, we notice that $\prod_{i=1}^{k} p(x;\theta_i)^{w_i}$ is a distribution in the same exponential family. We denote $p(x;\theta_0) = \prod_{i=1}^{k} p(x;\theta_i)^{w_i}$. Then

$$R(x) = \log\left(\sum_{i=1}^{k} w_i \frac{p(x;\theta_i)}{p(x;\theta_0)}\right)$$

As the ratio $p(x;\theta_i)/p(x;\theta_0)$ can be bounded above and below using techniques in Section 2.1, $R(x)$ can be correspondingly bounded. Notice the similarity between Equations (14) and (15). The key difference with the adaptive bounds is that, here we choose $p(x;\theta_0)$ instead of the dominating component in $m(x)$ as the "reference distribution" in the decomposition. This subtle difference is not presented in detail in our experimental studies but discussed here for completeness. Essentially, the gap of the bounds is up to the difference between the geometric average and the arithmetic average. In the extreme case that all mixture components are identical, this gap will reach zero. Therefore we expect good quality bounds with a small gap when the mixture components are similar as measured by KL divergence.

2.3. Case Studies

In the following, we instantiate the proposed method for several prominent cases on the mixture of exponential family distributions.

2.3.1. The Case of Exponential Mixture Models

An exponential distribution has density $p(x;\lambda) = \lambda \exp(-\lambda x)$ defined on $\mathcal{X} = [0, \infty)$ for $\lambda > 0$. Its CDF is $\Phi(x;\lambda) = 1 - \exp(-\lambda x)$. Any two components $w_1 p(x;\lambda_1)$ and $w_2 p(x;\lambda_2)$ (with $\lambda_1 \neq \lambda_2$) have a unique intersection point

$$x^\star = \frac{\log(w_1\lambda_1) - \log(w_2\lambda_2)}{\lambda_1 - \lambda_2} \tag{16}$$

if $x^\star \geq 0$; otherwise they do not intersect. The basic formulas to evaluate the bounds are

$$\begin{aligned} C_{i,j}(a,b) &= \log\left(\lambda'_j w'_j\right) M_i(a,b) + w_i \lambda'_j \left[\left(a + \frac{1}{\lambda_i}\right) e^{-\lambda_i a} - \left(b + \frac{1}{\lambda_i}\right) e^{-\lambda_i b}\right], && (17) \\ M_i(a,b) &= -w_i \left(e^{-\lambda_i a} - e^{-\lambda_i b}\right). && (18) \end{aligned}$$

2.3.2. The Case of Rayleigh Mixture Models

A Rayleigh distribution has density $p(x;\sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$, defined on $\mathcal{X} = [0, \infty)$ for $\sigma > 0$. Its CDF is $\Phi(x;\sigma) = 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right)$. Any two components $w_1 p(x;\sigma_1)$ and $w_2 p(x;\sigma_2)$ (with $\sigma_1 \neq \sigma_2$) must intersect at $x_0 = 0$ and can have at most one other intersection point

$$x^\star = \sqrt{\log \frac{w_1 \sigma_2^2}{w_2 \sigma_1^2} \Big/ \left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right)} \tag{19}$$

if the square root is well defined and $x^\star > 0$. We have

$$\begin{aligned} C_{i,j}(a,b) &= \log \frac{w'_j}{(\sigma'_j)^2} M_i(a,b) + \frac{w_i}{2(\sigma'_j)^2} \left[(a^2 + 2\sigma_i^2) e^{-\frac{a^2}{2\sigma_i^2}} - (b^2 + 2\sigma_i^2) e^{-\frac{b^2}{2\sigma_i^2}}\right] \\ &\quad - w_i \int_a^b \frac{x}{\sigma_i^2} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) \log x \mathrm{d}x, && (20) \\ M_i(a,b) &= -w_i \left(e^{-\frac{a^2}{2\sigma_i^2}} - e^{-\frac{b^2}{2\sigma_i^2}}\right). && (21) \end{aligned}$$

The last term in Equation (20) does not have a simple closed form (it requires the exponential integral, Ei). One need a numerical integrator to compute it.

2.3.3. The Case of Gaussian Mixture Models

The Gaussian density $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$ has support $\mathcal{X} = \mathbb{R}$ and parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Its CDF is $\Phi(x;\mu,\sigma) = \frac{1}{2}\left[1 + \operatorname{erf}(\frac{x-\mu}{\sqrt{2}\sigma})\right]$, where erf is the Gauss error function. The intersection point $x^\star$ of two components $w_1 p(x;\mu_1,\sigma_1)$ and $w_2 p(x;\mu_2,\sigma_2)$ can be obtained by solving the quadratic equation $\log(w_1 p(x;\mu_1,\sigma_1)) = \log(w_2 p(x;\mu_2,\sigma_2))$, which gives at most two solutions. As shown in Figure 1, the upper envelope of Gaussian densities corresponds to the lower envelope of parabolas. We have

$$C_{i,j}(a,b) = M_i(a,b)\left(\log w'_j - \log \sigma'_j - \frac{1}{2}\log(2\pi) - \frac{1}{2(\sigma'_j)^2}\left((\mu'_j-\mu_i)^2+\sigma_i^2\right)\right)$$
$$+\frac{w_i\sigma_i}{2\sqrt{2\pi}(\sigma'_j)^2}\left[(a+\mu_i-2\mu'_j)e^{-\frac{(a-\mu_i)^2}{2\sigma_i^2}} - (b+\mu_i-2\mu'_j)e^{-\frac{(b-\mu_i)^2}{2\sigma_i^2}}\right], \tag{22}$$

$$M_i(a,b) = -\frac{w_i}{2}\left(\operatorname{erf}\left(\frac{b-\mu_i}{\sqrt{2}\sigma_i}\right) - \operatorname{erf}\left(\frac{a-\mu_i}{\sqrt{2}\sigma_i}\right)\right). \tag{23}$$

2.3.4. The Case of Gamma Distributions

For simplicity, we only consider gamma distributions with the shape parameter $k > 0$ fixed and the scale $\lambda > 0$ varying. The density is defined on $(0, \infty)$ as $p(x; k, \lambda) = \frac{x^{k-1}e^{-\frac{x}{\lambda}}}{\lambda^k\Gamma(k)}$, where $\Gamma(\cdot)$ is the gamma function. Its CDF is $\Phi(x; k, \lambda) = \gamma(k, x/\lambda)/\Gamma(k)$, where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function. Two weighted gamma densities $w_1 p(x; k, \lambda_1)$ and $w_2 p(x; k, \lambda_2)$ (with $\lambda_1 \neq \lambda_2$) intersect at a unique point

$$x^\star = \left(\log\frac{w_1}{\lambda_1^k} - \log\frac{w_2}{\lambda_2^k}\right) / \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \tag{24}$$

if $x^\star > 0$; otherwise they do not intersect. From straightforward derivations,

$$C_{i,j}(a,b) = \log\frac{w'_j}{(\lambda'_j)^k\Gamma(k)} M_i(a,b) + w_i\int_a^b \frac{x^{k-1}e^{-\frac{x}{\lambda_i}}}{\lambda_i^k\Gamma(k)}\left(\frac{x}{\lambda'_j} - (k-1)\log x\right)dx, \tag{25}$$

$$M_i(a,b) = -\frac{w_i}{\Gamma(k)}\left(\gamma\left(k, \frac{b}{\lambda_i}\right) - \gamma\left(k, \frac{a}{\lambda_i}\right)\right). \tag{26}$$

Similar to the case of Rayleigh mixtures, the last term in Equation (25) relies on numerical integration.

3. Upper-Bounding the Differential Entropy of a Mixture

First, consider a finite parametric mixture model $m(x) = \sum_{i=1}^k w_i p(x; \theta_i)$. Using the chain rule of the entropy, we end up with the well-known lemma:

Lemma 1. *The entropy of a d-variate mixture is upper bounded by the sum of the entropy of its marginal mixtures: $H(m) \leq \sum_{i=1}^d H(m_i)$, where m_i is the 1D marginal mixture with respect to variable x_i.*

Since the 1D marginals of a multivariate GMM are univariate GMMs, we thus get a loose upper bound. A generic sample-based probabilistic bound is reported for the entropies of distributions with given support [31]: The method builds probabilistic upper and lower piecewisely linear CDFs based on an i.i.d. finite sample set of size n and a given deviation probability threshold. It then builds algorithmically between those two bounds the maximum entropy distribution [31] with a so-called string-tightening algorithm.

Instead, we proceed as follows: Consider finite mixtures of component distributions defined on the full support $\mathbb{R}^d$ that have finite component means and variances (like exponential families). Then we shall use the fact that the maximum entropy distribution with prescribed mean and variance is a Gaussian distribution, and conclude the upper bound by plugging the mixture mean and variance in the differential entropy formula of the Gaussian distribution. In general, the maximum entropy with moment constraints yields as a solution an exponential family.

Without loss of generality, consider GMMs in the form $m(x) = \sum_{i=1}^k w_i p(x;\mu_i,\Sigma_i)$ ($\Sigma_i = \sigma_i^2$ for univariate Gaussians). The mean $\bar{\mu}$ of the mixture is $\bar{\mu} = \sum_{i=1}^k w_i\mu_i$ and the variance is $\bar{\sigma}^2 = E[m^2] - E[m]^2$. Since $E[m^2] = \sum_{i=1}^k w_i \int x^2 p(x;\mu_i,\Sigma_i)\mathrm{d}x = \sum_{i=1}^k w_i \left(\mu_i^2 + \sigma_i^2\right)$, we deduce that

$$\bar{\sigma}^2 = \sum_{i=1}^k w_i(\mu_i^2 + \sigma_i^2) - \left(\sum_{i=1}^k w_i\mu_i\right)^2 = \sum_{i=1}^k w_i\left[(\mu_i - \bar{\mu})^2 + \sigma_i^2\right].$$

The entropy of a random variable with a prescribed variance $\bar{\sigma}^2$ is maximal for the Gaussian distribution with the same variance $\bar{\sigma}^2$, see [4]. Since the differential entropy of a Gaussian is $\log(\bar{\sigma}\sqrt{2\pi e})$, we deduce that the entropy of the GMM is upper bounded by

$$H(m) \leq \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log\sum_{i=1}^k w_i\left[(\mu_i - \bar{\mu})^2 + \sigma_i^2\right].$$

This upper bound can be easily generalized to arbitrary dimensionality. We get the following lemma:

Lemma 2. *The entropy of a d-variate GMM* $m(x) = \sum_{i=1}^k w_i p(x;\mu_i,\Sigma_i)$ *is upper bounded by* $\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\Sigma$, *where* $\Sigma = \sum_{i=1}^k w_i(\mu_i\mu_i^\top + \Sigma_i) - \left(\sum_{i=1}^k w_i\mu_i\right)\left(\sum_{i=1}^k w_i\mu_i^\top\right)$.

In general, exponential families have finite moments of any order [17]: In particular, we have $E[t(X)] = \nabla F(\theta)$ and $V[t(X)] = \nabla^2 F(\theta)$. For Gaussian distribution, we have the sufficient statistics $t(x) = (x, x^2)$ so that $E[t(X)] = \nabla F(\theta)$ yields the mean and variance from the log-normalizer. It is easy to generalize Lemma 2 to mixtures of exponential family distributions.

Note that this bound (called the Maximum Entropy Upper Bound in [13], MEUB) is tight when the GMM approximates a single Gaussian. It is fast to compute compared to the bound reported in [9] that uses Taylor' s expansion of the log-sum of the mixture density.

A similar argument cannot be applied for a lower bound since a GMM with a given variance may have entropy tending to $-\infty$. For example, assume the 2-component mixture's mean is zero, and that the variance approximates 1 by taking $m(x) = \frac{1}{2}G(x;-1,\epsilon) + \frac{1}{2}G(x;1,\epsilon)$ where G denotes the Gaussian density. Letting $\epsilon \to 0$, we get the entropy tending to $-\infty$.

We remark that our log-sum-exp inequality technique yields a $\log 2$ additive approximation range in the case of a Gaussian mixture with two components. It thus generalizes the bounds reported in [7] to GMMs with arbitrary variances that are not necessarily equal.

To see the bound gap, we have

$$\begin{aligned} &-\sum_r \int_{I_r} m(x)\left(\log k + \log\max_i w_i p_i(x)\right)\mathrm{d}x \leq H(m) \\ &\quad\leq -\sum_r \int_{I_r} m(x)\max\left\{\log\max_i w_i p_i(x),\, \log k + \log\min_i w_i p_i(x)\right\}\mathrm{d}x. \end{aligned} \tag{27}$$

Therefore the gap is at most

$$\begin{aligned} \Delta &= \min\left\{\sum_r \int_{I_r} m(x)\log\frac{\max_i w_i p_i(x)}{\min_i w_i p_i(x)}\mathrm{d}x,\ \log k\right\} \\ &= \min\left\{\sum_s\sum_r \int_{I_r} w_s p_s(x)\log\frac{\max_i w_i p_i(x)}{\min_i w_i p_i(x)}\mathrm{d}x,\ \log k\right\}. \end{aligned} \tag{28}$$

Thus to compute the gap error bound of the differential entropy, we need to integrate terms in the form

$$\int w_a p_a(x)\log\frac{w_b p_b(x)}{w_c p_c(x)}\mathrm{d}x.$$

See Appendix B for a closed-form formula when dealing with exponential family components.

4. Bounding the α-Divergence

The α-divergence [15,32–34] between $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{i=1}^{k'} w'_i p'_i(x)$ is defined as

$$D_\alpha\left(m : m'\right) = \frac{1}{\alpha(1-\alpha)}\left(1 - \int_{\mathcal{X}} m(x)^{\alpha} m'(x)^{1-\alpha} \mathrm{d}x\right), \tag{29}$$

which clearly satisfies $D_\alpha\left(m : m'\right) = D_{1-\alpha}\left(m' : m\right)$. The α-divergence is *a family* of information divergences parametrized by $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Let $\alpha \to 1$, we get the KL divergence (see [35] for a proof):

$$\lim_{\alpha \to 1} D_\alpha(m : m') = \mathrm{KL}(m : m') = \int_{\mathcal{X}} m(x) \log \frac{m(x)}{m'(x)} \mathrm{d}x, \tag{30}$$

and $\alpha \to 0$ gives the reverse KL divergence:

$$\lim_{\alpha \to 0} D_\alpha(m : m') = \mathrm{KL}(m' : m).$$

Other interesting values [33] include $\alpha = 1/2$ (squared Hellinger distance), $\alpha = 2$ (Pearson Chi-square distance), $\alpha = -1$ (Neyman Chi-square distance), etc. Notably, the Hellinger distance is a valid distance metric which satisfies non-negativity, symmetry, and the triangle inequality. In general, $D_\alpha(m : m')$ only satisfies non-negativity so that $D_\alpha\left(m : m'\right) \geq 0$ for any $m(x)$ and $m'(x)$. It is neither symmetric nor admitting the triangle inequality. Minimization of α-divergences allows one to choose a trade-off between mode fitting and support fitting of the minimizer [36]. The minimizer of α-divergences including MLE as a special case has interesting connections with transcendental number theory [37].

To compute $D_\alpha\left(m : m'\right)$ for given $m(x)$ and $m'(x)$ reduces to evaluate the Hellinger integral [38,39]:

$$H_\alpha(m : m') = \int_{\mathcal{X}} m(x)^{\alpha} m'(x)^{1-\alpha} \mathrm{d}x, \tag{31}$$

which in general does not have a closed form, as it was known that the α-divergence of mixture models is not analytic [6]. Moreover, $H_\alpha(m : m')$ may diverge making the α-divergence unbounded. Once $H_\alpha(m : m')$ can be solved, the Rényi and Tsallis divergences [35] and in general Sharma–Mittal divergences [40] can be easily computed. Therefore the results presented here directly extend to those divergence families.

Similar to the case of KL divergence, the Monte Carlo stochastic estimation of $H_\alpha(m : m')$ can be computed either as

$$\hat{H}_\alpha^n\left(m : m'\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{m'(x_i)}{m(x_i)}\right)^{1-\alpha},$$

where $x_1, \ldots, x_n \sim m(x)$ are i.i.d. samples, or as

$$\hat{H}_\alpha^n\left(m : m'\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{m(x_i)}{m'(x_i)}\right)^{\alpha},$$

where $x_1, \ldots, x_n \sim m'(x)$ are i.i.d. In either case, it is consistent so that $\lim_{n \to \infty} \hat{H}_\alpha^n\left(m : m'\right) = H_\alpha\left(m : m'\right)$. However, MC estimation requires a large sample and does not guarantee deterministic bounds. The techniques described in [41] work in practice for very close distributions, and do not apply between mixture models. We will therefore derive combinatorial bounds for $H_\alpha(m : m')$. The structure of this Section is parallel with Section 2 with necessary reformulations for a clear presentation.

4.1. Basic Bounds

For a pair of given $m(x)$ and $m'(x)$, we only need to derive bounds of $H_\alpha(m : m')$ in Equation (31) so that $L_\alpha(m : m') \leq H_\alpha(m : m') \leq U_\alpha(m : m')$. Then the α-divergence $D_\alpha(m : m')$ can be bounded by a linear transformation of the range $[L_\alpha(m : m'),\ U_\alpha(m : m')]$. In the following we always assume without loss of generality $\alpha \geq 1/2$. Otherwise we can bound $D_\alpha(m : m')$ by considering equivalently the bounds of $D_{1-\alpha}(m' : m)$.

Recall that in each elementary slab I_r, we have

$$\max\left\{kw_{\epsilon(r)}p_{\epsilon(r)}(x), w_{\delta(r)}p_{\delta(r)}(x)\right\} \leq m(x) \leq kw_{\delta(r)}p_{\delta(r)}(x). \tag{32}$$

Notice that $kw_{\epsilon(r)}p_{\epsilon(r)}(x)$, $w_{\delta(r)}p_{\delta(r)}(x)$, and $kw_{\delta(r)}p_{\delta(r)}(x)$ are all single component distributions up to a scaling coefficient. The general thinking is to bound the multi-component mixture $m(x)$ by single component distributions in each elementary interval, so that the integral in Equation (31) can be computed in a piecewise manner.

For the convenience of notation, we rewrite Equation (32) as

$$c_{\nu(r)}p_{\nu(r)}(x) \leq m(x) \leq c_{\delta(r)}p_{\delta(r)}(x), \tag{33}$$

where

$$c_{\nu(r)}p_{\nu(r)}(x) := kw_{\epsilon(r)}p_{\epsilon(r)}(x) \quad \text{or} \quad w_{\delta(r)}p_{\delta(r)}(x), \tag{34}$$

$$c_{\delta(r)}p_{\delta(r)}(x) := kw_{\delta(r)}p_{\delta(r)}(x). \tag{35}$$

If $1/2 \leq \alpha < 1$, then both x^α and $x^{1-\alpha}$ are monotonically increasing on $\mathbb{R}^+$. Therefore we have

$$A^\alpha_{\nu(r),\nu'(r)}(I_r) \leq \int_{I_r} m(x)^\alpha m'(x)^{1-\alpha}\mathrm{d}x \leq A^\alpha_{\delta(r),\delta'(r)}(I_r), \tag{36}$$

where

$$A^\alpha_{i,j}(I) = \int_I (c_i p_i(x))^\alpha \left(c'_j p'_j(x)\right)^{1-\alpha} \mathrm{d}x, \tag{37}$$

and I denotes an interval $I = (a, b) \subset \mathbb{R}$. The other case $\alpha > 1$ is similar by noting that x^α and $x^{1-\alpha}$ are monotonically increasing and decreasing on $\mathbb{R}^+$, respectively. In conclusion, we obtain the following bounds of $H_\alpha(m : m')$:

$$\text{If } 1/2 \leq \alpha < 1,\ L_\alpha(m : m') = \sum_{r=1}^{\ell} A^\alpha_{\nu(r),\nu'(r)}(I_r), \quad U_\alpha(m : m') = \sum_{r=1}^{\ell} A^\alpha_{\delta(r),\delta'(r)}(I_r); \tag{38}$$

$$\text{if } \alpha > 1,\ L_\alpha(m : m') = \sum_{r=1}^{\ell} A^\alpha_{\nu(r),\delta'(r)}(I_r), \quad U_\alpha(m : m') = \sum_{r=1}^{\ell} A^\alpha_{\delta(r),\nu'(r)}(I_r). \tag{39}$$

The remaining problem is to compute the definite integral $A^\alpha_{i,j}(I)$ in the above equations. Here we assume all mixture components are in the same exponential family so that $p_i(x) = p(x; \theta_i) = h(x)\exp\left(\theta_i^\top t(x) - F(\theta_i)\right)$, where $h(x)$ is a base measure, $t(x)$ is a vector of sufficient statistics, and the function F is known as the cumulant generating function. Then it is straightforward from Equation (37) that

$$A^\alpha_{i,j}(I) = c_i^\alpha (c'_j)^{1-\alpha} \int_I h(x) \exp\left(\left(\alpha\theta_i + (1-\alpha)\theta'_j\right)^\top t(x) - \alpha F(\theta_i) - (1-\alpha)F(\theta'_j)\right)\mathrm{d}x. \tag{40}$$

If $1/2 \leq \alpha < 1$, then $\bar{\theta} = \alpha\theta_i + (1-\alpha)\theta'_j$ belongs to the natural parameter space $\mathcal{M}_\theta$. Therefore $A^\alpha_{i,j}(I)$ is bounded and can be computed from the CDF of $p(x;\bar{\theta})$ as

$$A^\alpha_{i,j}(I) = c_i^\alpha (c'_j)^{1-\alpha} \exp(F(\bar{\theta}) - \alpha F(\theta_i) - (1-\alpha)F(\theta'_j)) \int_I p(x;\bar{\theta})\,\mathrm{d}x. \tag{41}$$

The other case $\alpha > 1$ is more difficult: if $\bar{\theta} = \alpha\theta_i + (1-\alpha)\theta'_j$ still lies in $\mathcal{M}_\theta$, then $A^\alpha_{i,j}(I)$ can be computed by Equation (41). Otherwise we try to solve it by a numerical integrator. This is not ideal as the integral may diverge, or our approximation may be too loose to conclude. We point the reader to [42] and Equations (61)–(69) in [35] for related analysis with more details. As computing $A^\alpha_{i,j}(I)$ only requires $O(1)$ time, the overall computational complexity (without considering the envelope computation) is $O(\ell)$.

4.2. Adaptive Bounds

This section derives the shape-dependent bounds which improve the basic bounds in Section 4.1. We can rewrite a mixture model $m(x)$ in a slab I_r as

$$m(x) = w_{\zeta(r)} p_{\zeta(r)}(x) \left(1 + \sum_{i \neq \zeta(r)} \frac{w_i p_i(x)}{w_{\zeta(r)} p_{\zeta(r)}(x)}\right), \tag{42}$$

where $w_{\zeta(r)} p_{\zeta(r)}(x)$ is a weighted component in $m(x)$ serving as a *reference*. We only discuss the case that the reference is chosen as the dominating component, i.e., $\zeta(r) = \delta(r)$. However it is worth to note that the proposed bounds do not depend on this particular choice. Therefore the ratio

$$\frac{w_i p_i(x)}{w_{\zeta(r)} p_{\zeta(r)}(x)} = \frac{w_i}{w_{\zeta(r)}} \exp\left(\left(\theta_i - \theta_{\zeta(r)}\right)^\top t(x) - F(\theta_i) + F(\theta_{\zeta(r)})\right) \tag{43}$$

can be bounded in a sub-range of $[0,1]$ by analyzing the extreme values of $t(x)$ in the slab I_r. This can be done because $t(x)$ usually consists of polynomial functions with finite critical points which can be solved easily. Correspondingly the function $\left(1 + \sum_{i \neq \zeta(r)} \frac{w_i p_i(x)}{w_{\zeta(r)} p_{\zeta(r)}(x)}\right)$ in I_r can be bounded in a subrange of $[1,k]$, denoted as $[\omega_{\zeta(r)}(I_r), \Omega_{\zeta(r)}(I_r)]$. Hence

$$\omega_{\zeta(r)}(I_r) w_{\zeta(r)} p_{\zeta(r)}(x) \leq m(x) \leq \Omega_{\zeta(r)}(I_r) w_{\zeta(r)} p_{\zeta(r)}(x). \tag{44}$$

This forms better bounds of $m(x)$ than Equation (32) because each component in the slab I_r is analyzed more accurately. Therefore, we refine the fundamental bounds of $m(x)$ by replacing the Equations (34) and (35) with

$$c_{\nu(r)} p_{\nu(r)}(x) := \omega_{\zeta(r)}(I_r) w_{\zeta(r)} p_{\zeta(r)}(x), \tag{45}$$

$$c_{\delta(r)} p_{\delta(r)}(x) := \Omega_{\zeta(r)}(I_r) w_{\zeta(r)} p_{\zeta(r)}(x). \tag{46}$$

Then, the improved bounds of H_α are given by Equations (38) and (39) according to the above replaced definition of $c_{\nu(r)} p_{\nu(r)}(x)$ and $c_{\delta(r)} p_{\delta(r)}(x)$.

To evaluate $\omega_{\zeta(r)}(I_r)$ and $\Omega_{\zeta(r)}(I_r)$ requires iterating through all components in each slab. Therefore the computational complexity is increased to $O\left(\ell(k+k')\right)$.

4.3. Variance-Reduced Bounds

This section further improves the proposed bounds based on variance reduction [43]. By assumption, $\alpha \geq 1/2$, then $m(x)^\alpha m'(x)^{1-\alpha}$ is more similar to $m(x)$ rather than $m'(x)$. The ratio

$m(x)^{\alpha}m'(x)^{1-\alpha}/m(x)$ is likely to have a small variance when x varies inside a slab I_r, especially when α is close to 1. We will therefore bound this ratio term in

$$\int_{I_r} m(x)^{\alpha}m'(x)^{1-\alpha}\mathrm{d}x = \int_{I_r} m(x)\left(\frac{m(x)^{\alpha}m'(x)^{1-\alpha}}{m(x)}\right)\mathrm{d}x = \sum_{i=1}^{k}\int_{I_r} w_i p_i(x)\left(\frac{m'(x)}{m(x)}\right)^{1-\alpha}\mathrm{d}x. \tag{47}$$

No matter $\alpha < 1$ or $\alpha > 1$, the function $x^{1-\alpha}$ must be monotonic on $\mathbb{R}^{+}$. In each slab I_r, $(m'(x)/m(x))^{1-\alpha}$ ranges between these two functions:

$$\left(\frac{c'_{\nu'(r)}p'_{\nu'(r)}(x)}{c_{\delta(r)}p_{\delta(r)}(x)}\right)^{1-\alpha} \quad \text{and} \quad \left(\frac{c'_{\delta'(r)}p'_{\delta'(r)}(x)}{c_{\nu(r)}p_{\nu(r)}(x)}\right)^{1-\alpha}, \tag{48}$$

where $c_{\nu(r)}p_{\nu(r)}(x)$, $c_{\delta(r)}p_{\delta(r)}(x)$, $c'_{\nu'(r)}p'_{\nu'(r)}(x)$ and $c'_{\delta'(r)}p'_{\delta'(r)}(x)$ are defined in Equations (45) and (46). Similar to the definition of $A^{\alpha}_{i,j}(I)$ in Equation (37), we define

$$B^{\alpha}_{i,j,l}(I) = \int_{I} w_i p_i(x)\left(\frac{c'_l p'_l(x)}{c_j p_j(x)}\right)^{1-\alpha}\mathrm{d}x. \tag{49}$$

Therefore we have,

$$\begin{aligned} &L_{\alpha}(m:m') = \min\mathcal{S}, \quad U_{\alpha}(m:m') = \max\mathcal{S}, \\ &\mathcal{S} = \left\{\sum_{r=1}^{\ell}\sum_{i=1}^{k} B^{\alpha}_{i,\delta(r),\nu'(r)}(I_r),\ \sum_{r=1}^{\ell}\sum_{i=1}^{k} B^{\alpha}_{i,\nu(r),\delta'(r)}(I_r)\right\}. \end{aligned} \tag{50}$$

The remaining problem is to evaluate $B^{\alpha}_{i,j,l}(I)$ in Equation (49). Similar to Section 4.1, assuming the components are in the same exponential family with respect to the natural parameters θ, we get

$$B^{\alpha}_{i,j,l}(I) = w_i\frac{c_l'^{1-\alpha}}{c_j^{1-\alpha}}\exp\left(F(\bar{\theta}) - F(\theta_i) - (1-\alpha)F(\theta'_l) + (1-\alpha)F(\theta_j)\right)\int_I p(x;\bar{\theta})\mathrm{d}x. \tag{51}$$

If $\bar{\theta} = \theta_i + (1-\alpha)\theta'_l - (1-\alpha)\theta_j$ is in the natural parameter space, $B^{\alpha}_{i,j,l}(I)$ can be computed from the CDF of $p(x;\bar{\theta})$; otherwise $B^{\alpha}_{i,j,l}(I)$ can be numerically integrated by its definition in Equation (49). The computational complexity is the same as the bounds in Section 4.2, i.e., $O(\ell\,(k+k'))$.

We have introduced three pairs of deterministic lower and upper bounds that enclose the true value of α-divergence between univariate mixture models. Thus the gap between the upper and lower bounds provides the additive approximation factor of the bounds. We conclude by emphasizing that the presented methodology can be easily generalized to other divergences [35,40] relying on Hellinger-type integrals $H_{\alpha,\beta}(p:q) = \int p(x)^{\alpha}q(x)^{\beta}\mathrm{d}x$ like the γ-divergence [44] as well as entropy measures [45].

5. Lower Bounds of the f-Divergence

The f-divergence between two distributions $m(x)$ and $m'(x)$ (not necessarily mixtures) is defined for a *convex generator* f by:

$$D_f(m:m') = \int m(x)f\left(\frac{m'(x)}{m(x)}\right)\mathrm{d}x.$$

If $f(x) = -\log x$, then $D_f(m:m') = \mathrm{KL}(m:m')$.

Let us partition the support $\mathcal{X} = \uplus_{r=1}^{\ell} I_r$ arbitrarily into elementary ranges, which *do not necessarily correspond to the envelopes*. Denote by M_I the probability mass of a mixture $m(x)$ in the range I: $M_I = \int_I m(x)dx$. Then

$$D_f(m : m') = \sum_{r=1}^{\ell} M_{I_r} \int_{I_r} \frac{m(x)}{M_{I_r}} f\left(\frac{m'(x)}{m(x)}\right) \mathrm{d}x.$$

Note that in range I_r, $\frac{m(x)}{M_{I_r}}$ is a unit weight distribution. Thus by Jensen's inequality $f(E[X]) \leq E[f(X)]$, we get

$$D_f(m : m') \geq \sum_{r=1}^{\ell} M_{I_r} f\left(\int_{I_r} \frac{m(x)}{M_{I_r}} \frac{m'(x)}{m(x)} \mathrm{d}x\right) = \sum_{r=1}^{\ell} M_{I_r} f\left(\frac{M'_{I_r}}{M_{I_r}}\right). \tag{52}$$

Notice that the RHS of Equation (52) is the f-divergence between $(M_{I_1}, \cdots, M_{I_\ell})$ and $(M'_{I_1}, \cdots, M'_{I_\ell})$, denoted by $D_f^{\mathcal{I}}(m : m')$. In the special case that $\ell = 1$ and $I_1 = \mathcal{X}$, the above Equation (52) turns out to be the usual Gibbs' inequality: $D_f(m : m') \geq f(1)$, and Csiszár generator is chosen so that $f(1) = 0$. In conclusion, for a fixed (coarse-grained) countable partition of $\mathcal{X}$, we recover the well-know information monotonicity [46] of the f-divergences:

$$D_f(m : m') \geq D_f^{\mathcal{I}}(m : m') \geq 0.$$

In practice, we get closed-form lower bounds when $M_I = \int_a^b m(x)\mathrm{d}x = \Phi(b) - \Phi(a)$ is available in closed-form, where $\Phi(\cdot)$ denote the CDF. In particular, if $m(x)$ is a mixture model, then its CDF can be computed by linearly combining the CDFs of its components.

To wrap up, we have proved that coarse-graining by making a finite partition of the support $\mathcal{X}$ yields a lower bound on the f-divergence by virtue of the information monotonicity. Therefore, instead of doing Monte Carlo stochastic integration:

$$\hat{D}_f^n(m : m') = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{m'(x_i)}{m(x_i)}\right),$$

with $x_1, \ldots, x_n \sim_{\text{i.i.d.}} m(x)$, it could be better to sort those n samples and consider the coarse-grained partition:

$$\mathcal{I} = (-\infty, x_{(1)}] \cup \left(\uplus_{i=1}^{n-1} (x_{(i)}, x_{(i+1)}]\right) \cup (x_{(n)}, +\infty)$$

to get a *guaranteed lower bound* on the f-divergence. We will call this bound CGQLB for Coarse Graining Quantization Lower Bound.

Given a budget of n splitting points on the range $\mathcal{X}$, it would be interesting to find the best n points that maximize the lower bound $D_f^{\mathcal{I}}(m : m')$. This is ongoing research.

6. Experiments

We perform an empirical study to verify our theoretical bounds. We simulate four pairs of mixture models $\{(\mathtt{EMM}_1, \mathtt{EMM}_2), (\mathtt{RMM}_1, \mathtt{RMM}_2), (\mathtt{GMM}_1, \mathtt{GMM}_2), (\mathtt{GaMM}_1, \mathtt{GaMM}_2)\}$ as the test subjects. The component type is implied by the model name, where GaMM stands for Gamma mixtures. The components of each mixture model are given as follows.

1. $\mathtt{EMM}_1$'s components, in the form (λ_i, w_i), are given by $(0.1, 1/3)$, $(0.5, 1/3)$, $(1, 1/3)$; $\mathtt{EMM}_2$'s components are $(2, 0.2)$, $(10, 0.4)$, $(20, 0.4)$.
2. $\mathtt{RMM}_1$'s components, in the form (σ_i, w_i), are given by $(0.5, 1/3)$, $(2, 1/3)$, $(10, 1/3)$; $\mathtt{RMM}_2$ consists of $(5, 0.25)$, $(60, 0.25)$, $(100, 0.5)$.

3. $\mathtt{GMM}_1$'s components, in the form (μ_i, σ_i, w_i), are $(-5, 1, 0.05)$, $(-2, 0.5, 0.1)$, $(5, 0.3, 0.2)$, $(10, 0.5, 0.2)$, $(15, 0.4, 0.05)$, $(25, 0.5, 0.3)$, $(30, 2, 0.1)$; $\mathtt{GMM}_2$ consists of $(-16, 0.5, 0.1)$, $(-12, 0.2, 0.1)$, $(-8, 0.5, 0.1)$, $(-4, 0.2, 0.1)$, $(0, 0.5, 0.2)$, $(4, 0.2, 0.1)$, $(8, 0.5, 0.1)$, $(12, 0.2, 0.1)$, $(16, 0.5, 0.1)$.
4. $\mathtt{GaMM}_1$'s components, in the form (k_i, λ_i, w_i), are $(2, 0.5, 1/3)$, $(2, 2, 1/3)$, $(2, 4, 1/3)$; $\mathtt{GaMM}_2$ consists of $(2, 5, 1/3)$, $(2, 8, 1/3)$, $(2, 10, 1/3)$.

We compare the proposed bounds with Monte Carlo estimation with different sample sizes in the range $\{10^2, 10^3, 10^4, 10^5\}$. For each sample size configuration, we report the 0.95 confidence interval by Monte Carlo estimation using the corresponding number of samples. Figure 2a–d shows the input signals as well as the estimation results, where the proposed bounds CELB, CEUB, CEALB, CEAUB, CGQLB are presented as horizontal lines, and the Monte Carlo estimations over different sample sizes are presented as error bars. We can loosely consider the average Monte Carlo output with the largest sample size (10^5) as the underlying truth, which is clearly inside our bounds. This serves as an empirical justification on the correctness of the bounds.

A key observation is that the bounds can be *very tight*, especially when the underlying KL divergence has a large magnitude, e.g., $\mathrm{KL}(\mathtt{RMM}_2 : \mathtt{RMM}_1)$. This is because the gap between the lower and upper bounds is always guaranteed to be within $\log k + \log k'$. Because KL is unbounded [4], in the general case two mixture models may have a large KL. Then our approximation gap is relatively very small. On the other hand, we also observed that the bounds in certain cases, e.g., $\mathrm{KL}(\mathtt{EMM}_2 : \mathtt{EMM}_1)$, are not as tight as the other cases. When the underlying KL is small, the bounds are not as informative as the general case.

Comparatively, there is a significant improvement of the shape-dependent bounds (CEALB and CEAUB) over the combinatorial bounds (CELB and CEUB). In all investigated cases, the adaptive bounds can roughly shrink the gap by half of its original size at the cost of additional computation.

Note that, the bounds are accurate and must contain the true value. Monte Carlo estimation gives no guarantee on where the true value is. For example, in estimating $\mathrm{KL}(\mathtt{GMM}_1 : \mathtt{GMM}_2)$, Monte Carlo estimation based on 10^4 samples can go beyond our bounds! It therefore suffers from a larger estimation error.

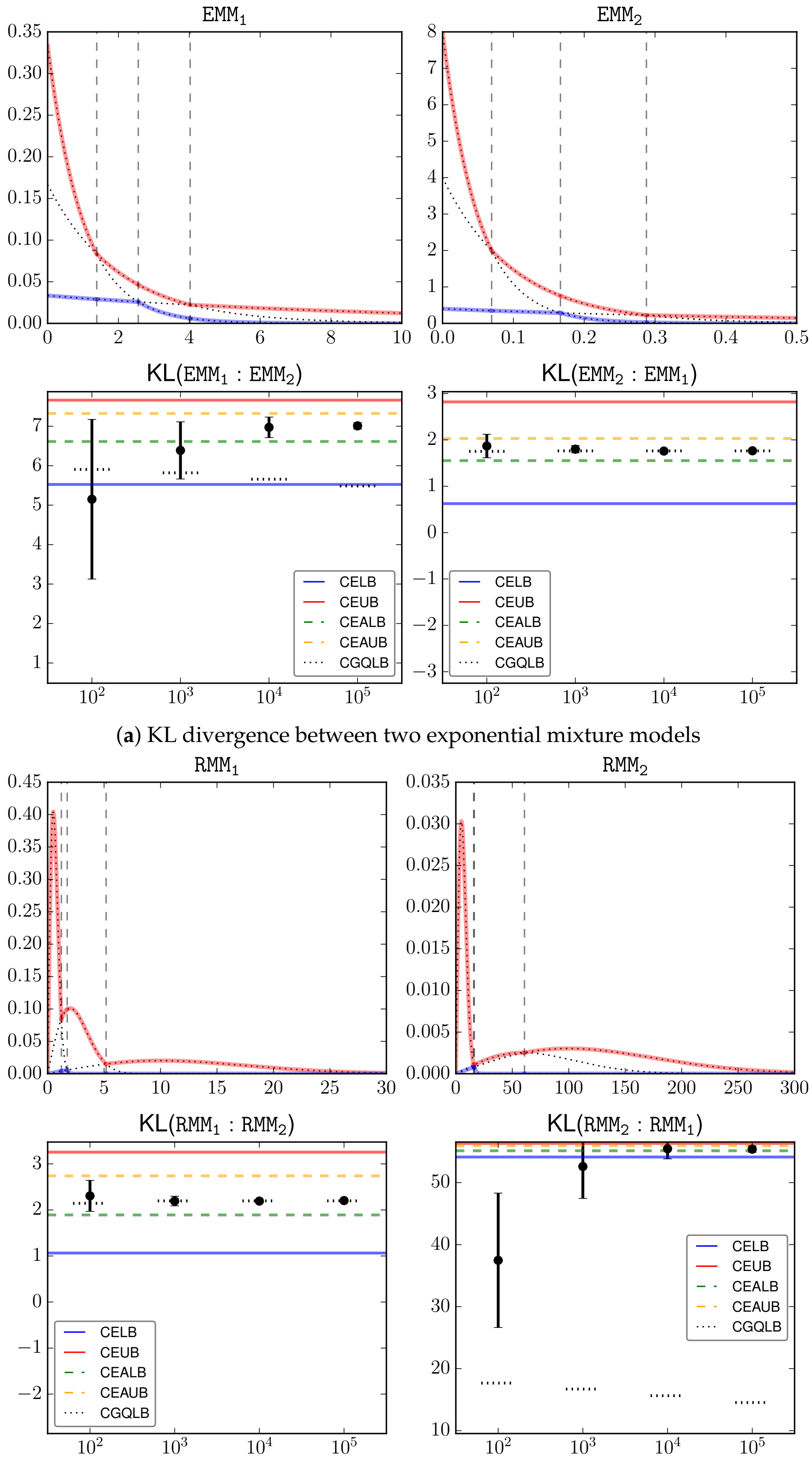
CGQLB as a simple-to-implement technique shows surprising good performance in several cases, e.g., $\mathrm{KL}(\mathtt{RMM}_1, \mathtt{RMM}_2)$. Although it requires a large number of samples, we can observe that increasing sample size has limited effect on improving this bound. Therefore, in practice, one may intersect the range defined by CEALB and CEAUB with the range defined by CGQLB with a small sample size (e.g., 100) to get better bounds.

We simulates a set of Gaussian mixture models besides the above $\mathtt{GMM}_1$ and $\mathtt{GMM}_2$. Figure 3 shows the GMM densities as well as their differential entropy. A detailed explanation of the components of each GMM model is omitted for brevity.

The key observation is that CEUB (CEAUB) is *very tight* in most of the investigated cases. This is because that the upper envelope that is used to compute CEUB (CEAUB) gives a very good estimation of the input signal.

Notice that MEUB only gives an upper bound of the differential entropy as discussed in Section 3. In general the proposed bounds are tighter than MEUB. However, this is not the case when the mixture components are merged together and approximate one single Gaussian (and therefore its entropy can be well approximated by the Gaussian entropy), as shown in the last line of Figure 3.

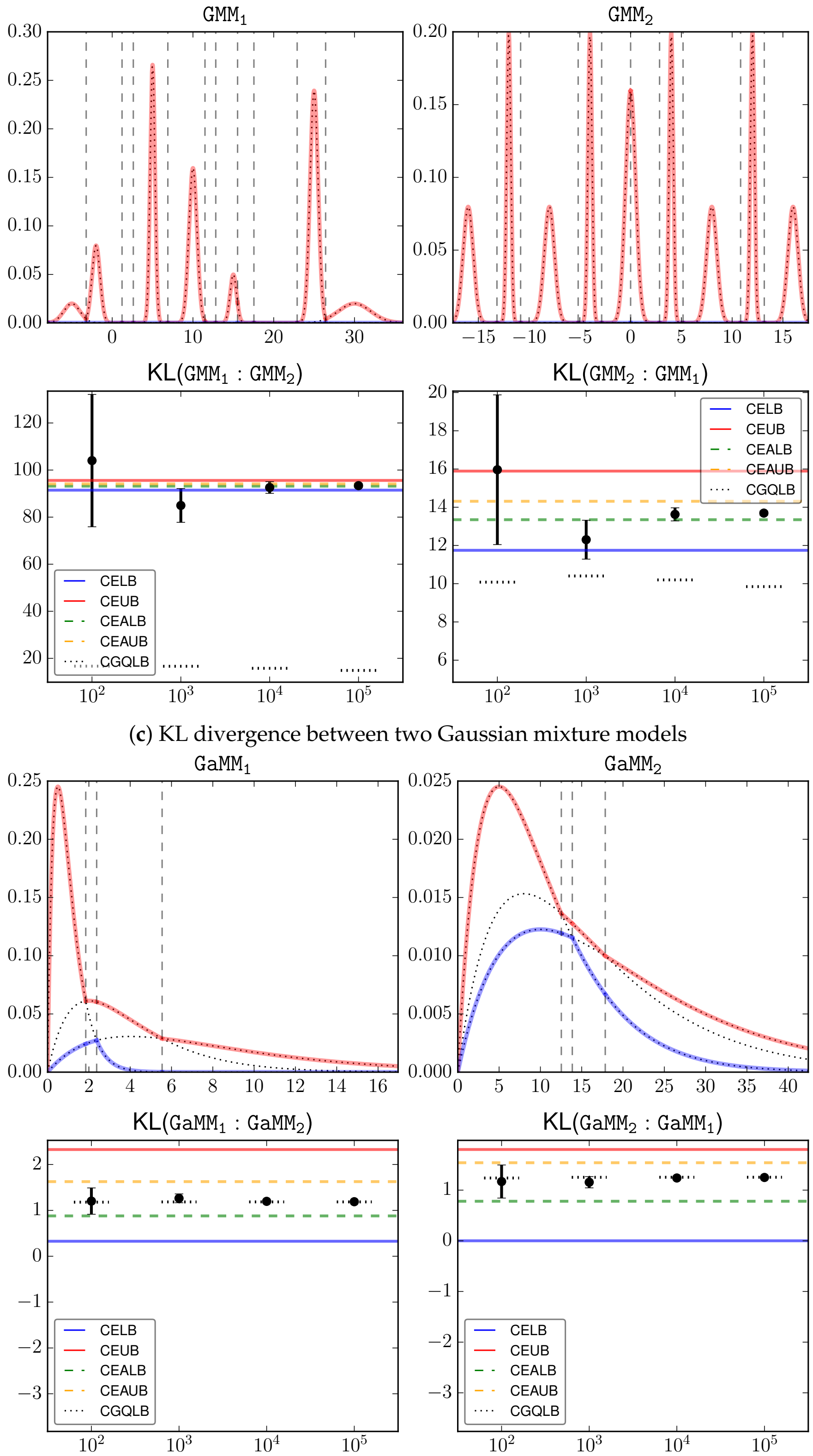
For α-divergence, the bounds introduced in Sections 4.1–4.3 are denoted as “Basic”, “Adaptive” and “VR”, respectively. Figure 4 visualizes these GMMs and plots the estimations of their α-divergences against α. The red lines mean the upper envelope. The dashed vertical lines mean the elementary intervals. The components of $\mathtt{GMM}_1$ and $\mathtt{GMM}_2$ are more separated than $\mathtt{GMM}_3$ and $\mathtt{GMM}_4$. Therefore these two pairs present different cases. For a clear presentation, only VR (which is expected to be better than Basic and Adaptive) is shown. We can see that, visually in the big scale, VR tightly surrounds the true value.



(**a**) KL divergence between two exponential mixture models

(**b**) KL divergence between two Rayleigh mixture models

Figure 2. *Cont.*



(**c**) KL divergence between two Gaussian mixture models

(**d**) KL divergence between two Gamma mixture models

Figure 2. Lower and upper bounds on the KL divergence between mixture models. The y-axis means KL divergence. Solid/dashed lines represent the combinatorial/adaptive bounds, respectively. The error-bars show the 0.95 confidence interval by Monte Carlo estimation using the corresponding sample size (x-axis). The narrow dotted bars show the CGQLB estimation w.r.t. the sample size.

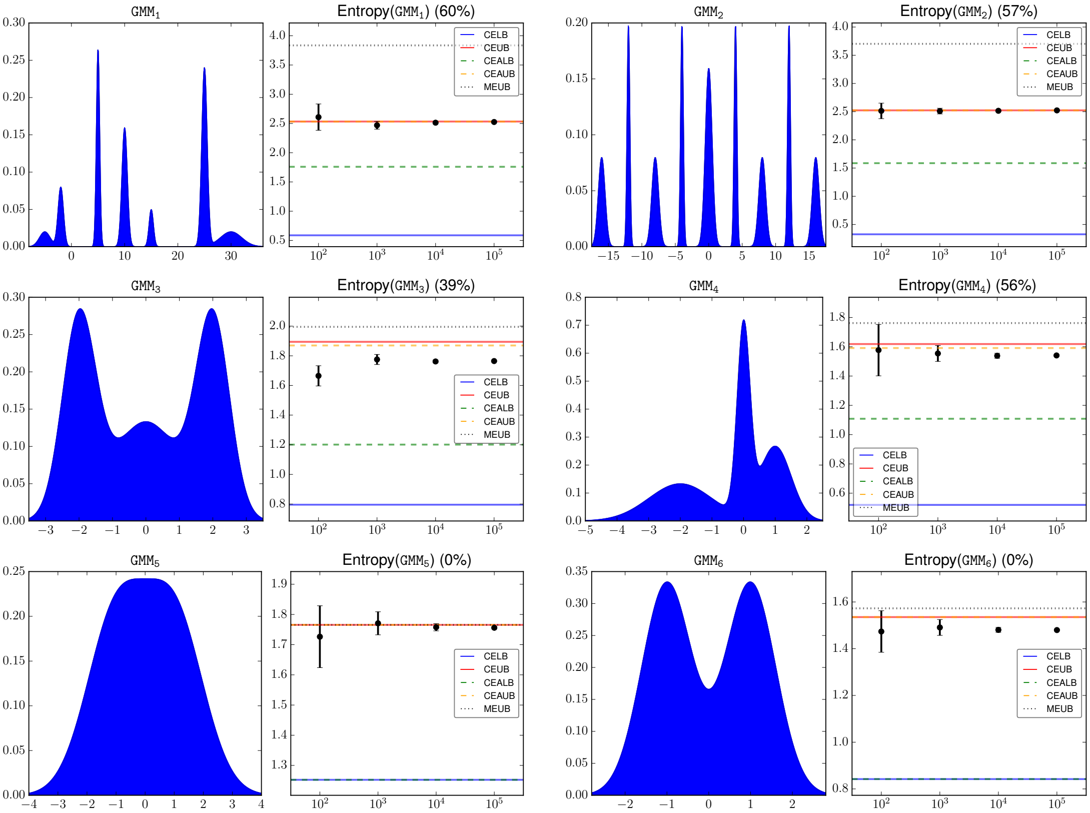


Figure 3. Lower and upper bounds on the differential entropy of Gaussian mixture models. On the left of each subfigure is the simulated GMM signal. On the right of each subfigure is the estimation of its differential entropy. Note that a subset of the bounds coincide with each other in several cases.

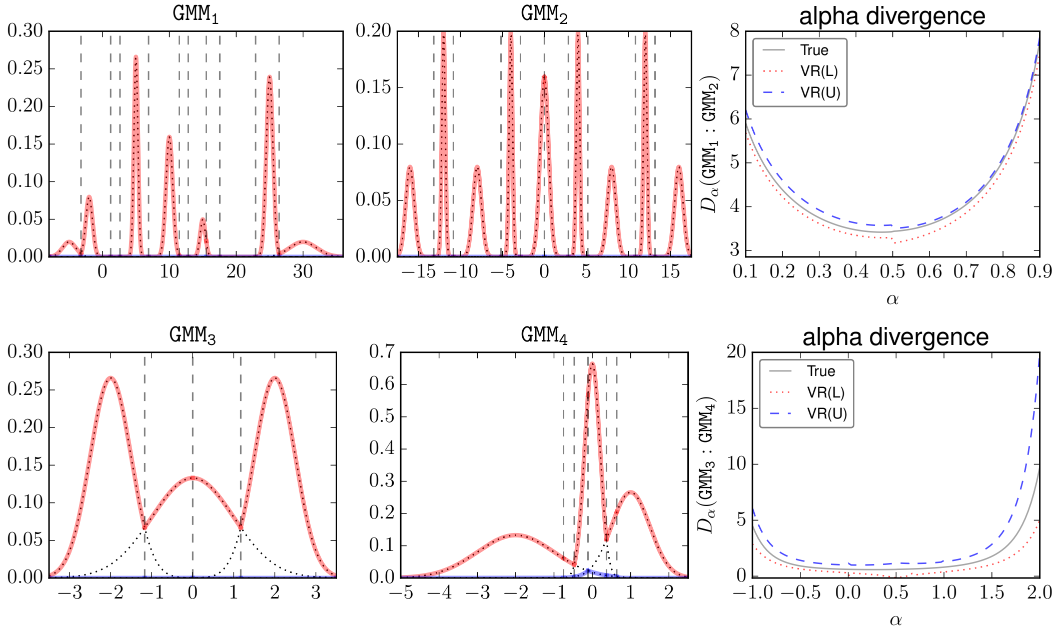


Figure 4. Two pairs of Gaussian Mixture Models and their α-divergences against different values of α. The "true" value of D_α is estimated by MC using 10^4 random samples. VR(L) and VR(U) denote the variation reduced lower and upper bounds, respectively. The range of α is selected for each pair for a clear visualization.

For a more quantitative comparison, Table 1 shows the estimated α-divergence by MC, Basic, Adaptive, and VR. As D_α is defined on $\mathbb{R} \setminus \{0, 1\}$, the KL bounds CE(A)LB and CE(A)UB are presented for $\alpha = 0$ or 1. Overall, we have the following order of gap size: Basic > Adaptive > VR, and VR is recommended in general for bounding α-divergences. There are certain cases that the upper VR bound is looser than Adaptive. In practice one can compute the intersection of these bounds as well as the trivial bound $D_\alpha(m : m') \geq 0$ to get the best estimation.

Table 1. The estimated D_α and its bounds. The 95% confidence interval is shown for MC.

	α	MC(10^2)	MC(10^3)	MC(10^4)	Basic		Adaptive		VR	
					L	U	L	U	L	U
GMM$_1$ & GMM$_2$	0	15.96 ± 3.9	12.30 ± 1.0	13.63 ± 0.3	11.75	15.89	12.96	14.63		
	0.01	13.36 ± 2.9	10.63 ± 0.8	11.66 ± 0.3	-700.50	11.73	-77.33	11.73	11.40	12.27
	0.5	3.57 ± 0.3	3.47 ± 0.1	3.47 ± 0.07	-0.60	3.42	3.01	3.42	3.17	3.51
	0.99	40.04 ± 7.7	37.22 ± 2.3	38.58 ± 0.8	-333.90	39.04	5.36	38.98	38.28	38.96
	1	104.01 ± 28	84.96 ± 7.2	92.57 ± 2.5	91.44	95.59	92.76	94.41		
GMM$_3$ & GMM$_4$	0	0.71 ± 0.2	0.63 ± 0.07	0.62 ± 0.02	0.00	1.76	0.00	1.16		
	0.01	0.71 ± 0.2	0.63 ± 0.07	0.62 ± 0.02	-179.13	7.63	-38.74	4.96	0.29	1.00
	0.5	0.82 ± 0.3	0.57 ± 0.1	0.62 ± 0.04	-5.23	0.93	-0.71	0.85	-0.18	1.19
	0.99	0.79 ± 0.3	0.76 ± 0.1	0.80 ± 0.03	-165.72	12.10	-59.76	9.11	0.37	1.28
	1	0.80 ± 0.3	0.77 ± 0.1	0.81 ± 0.03	0.00	1.82	0.31	1.40		

Note the similarity between KL in Equation (30) and the expression in Equation (47). We give without a formal analysis that: CEAL(U)B is equivalent to VR at the limit $\alpha \to 0$ or $\alpha \to 1$. Experimentally as we slowly set $\alpha \to 1$, we can see that VR is consistent with CEAL(U)B.

7. Concluding Remarks and Perspectives

We have presented a fast versatile method to compute bounds on the Kullback–Leibler divergence between mixtures by building algorithmic formulae. We reported on our experiments for various mixture models in the exponential family. For univariate GMMs, we get a guaranteed bound of the KL divergence of two mixtures m and m' with k and k' components within an additive approximation factor of $\log k + \log k'$ in $O\left((k + k')\log(k + k')\right)$-time. Therefore, the larger the KL divergence, the better the bound when considering a multiplicative $(1 + \alpha)$-approximation factor, since $\alpha = \frac{\log k + \log k'}{\mathrm{KL}(m:m')}$. The adaptive bounds are guaranteed to yield better bounds at the expense of computing potentially $O\left(k^2 + (k')^2\right)$ intersection points of pairwise weighted components.

Our technique also yields the bound for the Jeffreys divergence (the symmetrized KL divergence: $J(m, m') = \mathrm{KL}(m : m') + \mathrm{KL}(m' : m)$) and the Jensen–Shannon divergence [47] (JS):

$$\mathrm{JS}(m, m') = \frac{1}{2}\left(\mathrm{KL}\left(m : \frac{m + m'}{2}\right) + \mathrm{KL}\left(m' : \frac{m + m'}{2}\right)\right),$$

since $\frac{m+m'}{2}$ is a mixture model with $k + k'$ components. One advantage of this statistical distance is that it is symmetric, always bounded by $\log 2$, and its square root yields a metric distance [48]. The log-sum-exp inequalities may also be used to compute some Rényi divergences [35]:

$$R_\alpha(m, p) = \frac{1}{\alpha - 1}\log\left(\int m(x)^\alpha p(x)^{1-\alpha}\right)\mathrm{d}x,$$

when α is an integer, $m(x)$ a mixture and $p(x)$ a single (component) distribution. Getting fast guaranteed tight bounds on statistical distances between mixtures opens many avenues. For example, we may consider building hierarchical mixture models by merging iteratively two mixture components so that those pairs of components are chosen so that the KL distance between the full mixture and the simplified mixture is minimized.

In order to be useful, our technique is unfortunately limited to univariate mixtures: indeed, in higher dimensions, we can still compute the maximization diagram of weighted components (an additively weighted Bregman–Voronoi diagram [49,50] for components belonging to the same exponential family). However, it becomes more complex to compute in the elementary Voronoi cells V, the functions $C_{i,j}(V)$ and $M_i(V)$ (in 1D, the Voronoi cells are segments). We may obtain hybrid algorithms by approximating or estimating these functions. In 2D, it is thus possible to obtain lower and upper bounds on the Mutual Information [51] (MI) when the joint distribution $m(x,y)$ is a 2D mixture of Gaussians:

$$I(M;M') = \int m(x,y) \log \frac{m(x,y)}{m(x)m'(y)} \mathrm{d}x\mathrm{d}y.$$

Indeed, the marginal distributions $m(x)$ and $m'(y)$ are univariate Gaussian mixtures.

A Python code implementing those computational-geometric methods for reproducible research is available online [52].

Acknowledgments: The authors gratefully thank the referees for their comments. This work was carried out while Ke Sun was visiting Frank Nielsen at Ecole Polytechnique, Palaiseau, France.

Author Contributions: Frank Nielsen and Ke Sun contributed to the theoretical results as well as to the writing of the article. Ke Sun implemented the methods and performed the numerical experiments. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Kullback–Leibler Divergence of Mixture Models Is Not Analytic [6]

Ideally, we aim at getting a finite length closed-form formula to compute the KL divergence of finite mixture models. However, this is provably mathematically intractable [6] because of the log-sum term in the integral, as we shall prove below.

Analytic expressions encompass closed-form formula and may include special functions (e.g., Gamma function) but do not allow to use limits or integrals. An analytic function $f(x)$ is a C^∞ function (infinitely differentiable) such that around any point x_0 the k-order Taylor series $T_k(x) = \sum_{i=0}^{k} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i$ converges to $f(x)$: $\lim_{k\to\infty} T_k(x) = f(x)$ for x belonging to a neighborhood $N_r(x_0) = \{x : |x-x_0| \leq r\}$ of x_0, where r is called the radius of convergence. The analytic property of a function is equivalent to the condition that for each $k \in \mathbb{N}$, there exists a constant c such that $\left|\frac{\mathrm{d}^k f}{\mathrm{d}x^k}(x)\right| \leq c^{k+1}k!$.

To prove that the KL of mixtures is not analytic (hence does not admit a closed-form formula), we shall adapt the proof reported in [6] (in Japanese, we thank Professor Aoyagi for sending us his paper [6]). We shall prove that $\mathrm{KL}(p:q)$ is not analytic for $p(x) = G(x;0,1)$ and $q(x;w) = (1-w)G(x;0,1) + wG(x;1,1)$, where $w \in (0,1)$, and $G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is the density of a univariate Gaussian of mean μ and standard deviation σ. Let $D(w) = \mathrm{KL}(p(x):q(x;w))$ denote the KL divergence between these two mixtures (p has a single component and q has two components).

We have

$$\log\frac{p(x)}{q(x;w)} = \log\frac{\exp\left(-\frac{x^2}{2}\right)}{(1-w)\exp\left(-\frac{x^2}{2}\right) + w\exp\left(-\frac{(x-1)^2}{2}\right)} = -\log(1 + w(e^{x-\frac{1}{2}} - 1)). \tag{A1}$$

Therefore

$$\frac{\mathrm{d}^k D}{\mathrm{d}w^k} = \frac{(-1)^k}{k}\int p(x)(e^{x-\frac{1}{2}} - 1)\mathrm{d}x.$$

Let x_0 be the root of the equation $e^{x-\frac{1}{2}} - 1 = e^{\frac{x}{2}}$ so that for $x \geq x_0$, we have $e^{x-\frac{1}{2}} - 1 \geq e^{\frac{x}{2}}$. It follows that:

$$\left|\frac{\mathrm{d}^k D}{\mathrm{d}w^k}\right| \geq \frac{1}{k}\int_{x_0}^{\infty} p(x)e^{\frac{kx}{2}}\mathrm{d}x = \frac{1}{k}e^{\frac{k^2}{8}}A_k$$

with $A_k = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x-\frac{k}{2}}{2})\mathrm{d}x$. When $k \to \infty$, we have $A_k \to 1$. Consider $k_0 \in \mathbb{N}$ such that $A_{k_0} > 0.9$. Then the radius of convergence r is such that:

$$\frac{1}{r} \geq \lim_{k\to\infty} \left(\frac{1}{kk!} 0.9 \exp\left(\frac{k^2}{8}\right)\right)^{\frac{1}{k}} = \infty.$$

Thus the convergence radius is $r = 0$, and therefore the KL divergence is not an analytic function of the parameter w. The KL of mixtures is an example of a non-analytic smooth function. (Notice that the absolute value is not analytic at 0.)

Appendix B. Closed-Form Formula for the Kullback–Leibler Divergence between Scaled and Truncated Exponential Families

When computing approximation bounds for the KL divergence between two mixtures $m(x)$ and $m'(x)$, we end up with the task of computing $\int_{\mathcal{D}} w_a p_a(x) \log \frac{w'_b p'_b(x)}{w'_c p'_c(x)} \mathrm{d}x$ where $\mathcal{D} \subseteq \mathcal{X}$ is a subset of the full support $\mathcal{X}$. We report a generic formula for computing these formula when the mixture (scaled and truncated) components belong to the same exponential family [17]. An exponential family has canonical log-density written as $l(x;\theta) = \log p(x;\theta) = \theta^\top t(x) - F(\theta) + k(x)$, where $t(x)$ denotes the sufficient statistics, $F(\theta)$ the log-normalizer (also called cumulant function or partition function), and $k(x)$ an auxiliary carrier term.

Let $\mathrm{KL}(w_1 p_1 : w_2 p_2 : w_3 p_3) = \int_{\mathcal{X}} w_1 p_1(x) \log \frac{w_2 p_2(x)}{w_3 p_3(x)} \mathrm{d}x = H_\times(w_1 p_1 : w_3 p_3) - H_\times(w_1 p_1 : w_2 p_2)$. Since it is a difference of two cross-entropies, we get for three distributions belonging to the same exponential family [26] the following formula:

$$\mathrm{KL}(w_1 p_1 : w_2 p_2 : w_3 p_3) = w_1 \log \frac{w_2}{w_3} + w_1 (F(\theta_3) - F(\theta_2) - (\theta_3 - \theta_2)^\top \nabla F(\theta_1)).$$

Furthermore, when the support is restricted, say to support range $\mathcal{D} \subseteq \mathcal{X}$, let $m_{\mathcal{D}}(\theta) = \int_{\mathcal{D}} p(x;\theta)\mathrm{d}x$ denote the mass and $p(\tilde{x;\theta}) = \frac{p(x;\theta)}{m_{\mathcal{D}}(\theta)}$ the normalized distribution. Then we have:

$$\int_{\mathcal{D}} w_1 p_1(x) \log \frac{w_2 p_2(x)}{w_3 p_3(x)} \mathrm{d}x = m_{\mathcal{D}}(\theta_1)(\mathrm{KL}(w_1 \tilde{p}_1 : w_2 \tilde{p}_2 : w_3 \tilde{p}_3)) - \log \frac{w_2 m_{\mathcal{D}}(\theta_3)}{w_3 m_{\mathcal{D}}(\theta_2)}.$$

When $F_{\mathcal{D}}(\theta) = F(\theta) - \log m_{\mathcal{D}}(\theta)$ is strictly convex and differentiable then $p(\tilde{x;\theta})$ is an exponential family and the closed-form formula follows straightforwardly. Otherwise, we still get a closed-form but need more derivations. For univariate distributions, we write $\mathcal{D} = (a,b)$ and $m_{\mathcal{D}}(\theta) = \int_a^b p(x;\theta)\mathrm{d}x = P_\theta(b) - P_\theta(a)$ where $P_\theta(a) = \int^a p(x;\theta)\mathrm{d}x$ denotes the cumulative distribution function.

The usual formula for truncated and scaled Kullback–Leibler divergence is:

$$\mathrm{KL}_{\mathcal{D}}(wp(x;\theta) : w' p(x;\theta')) = w m_{\mathcal{D}}(\theta) \left(\log \frac{w}{w'} + B_F(\theta' : \theta) \right) + w(\theta' - \theta)^\top \nabla m_{\mathcal{D}}(\theta), \qquad \text{(B1)}$$

where $B_F(\theta' : \theta)$ is a Bregman divergence [5]:

$$B_F(\theta' : \theta) = F(\theta') - F(\theta) - (\theta' - \theta)^\top \nabla F(\theta).$$

This formula extends the classic formula [5] for full regular exponential families (by setting $w = w' = 1$ and $m_{\mathcal{D}}(\theta) = 1$ with $\nabla m_{\mathcal{D}}(\theta) = 0$).

Similar formulæ are available for the cross-entropy and entropy of exponential families [26].

Appendix C. On the Approximation of KL between Smooth Mixtures by a Bregman Divergence [5]

Clearly, since Bregman divergences are always finite while KL divergences may diverge, we need extra conditions to assert that the KL between mixtures can be approximated by Bregman divergences.

We require that the Jeffreys divergence between mixtures be finite in order to approximate the KL between mixtures by a Bregman divergence. We loosely derive this observation (Careful derivations will be reported elsewhere) using two different approaches:

- First, continuous mixture distributions have smooth densities that can be arbitrarily closely approximated using a single distribution (potentially multi-modal) belonging to the Polynomial Exponential Families [53,54] (PEFs). A polynomial exponential family of order D has log-likelihood $l(x;\theta) \propto \sum_{i=1}^{D} \theta_i x^i$: Therefore, a PEF is an exponential family with polynomial sufficient statistics $t(x) = (x, x^2, \ldots, x^D)$. However, the log-normalizer $F_D(\theta) = \log \int \exp(\theta^\top t(x))\mathrm{d}x$ of a D-order PEF is not available in closed-form: It is computationally intractable. Nevertheless, the KL between two mixtures $m(x)$ and $m'(x)$ can be theoretically approximated closely by a Bregman divergence between the two corresponding PEFs: $\mathrm{KL}(m(x) : m'(x)) \simeq \mathrm{KL}(p(x;\theta) : p(x;\theta')) = B_{F_D}(\theta' : \theta)$, where θ and θ' are the natural parameters of the PEF family $\{p(x;\theta)\}$ approximating $m(x)$ and $m'(x)$, respectively (i.e., $m(x) \simeq p(x;\theta)$ and $m'(x) \simeq p(x;\theta')$). Notice that the Bregman divergence of PEFs has necessarily finite value but the KL of two smooth mixtures can potentially diverge (infinite value), hence the conditions on Jeffreys divergence to be finite.

- Second, consider two finite mixtures $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{j=1}^{k'} w'_j p'_j(x)$ of k and k' components (possibly with heterogeneous components $p_i(x)$'s and $p'_j(x)$'s), respectively. In information geometry, a mixture family is the set of convex combination of fixed component densities. Thus in statistics, a mixture is understood as a convex combination of parametric components while in information geometry a mixture family is the set of convex combination of fixed components. Let us consider the mixture families $\{g(x;(w,w'))\}$ generated by the $D = k + k'$ fixed components $p_1(x), \ldots, p_k(x), p'_1(x), \ldots, p'_{k'}(x)$:

$$\left\{ g(x;(w,w')) = \sum_{i=1}^{k} w_i p_i(x) + \sum_{j=1}^{k'} w'_j p'_j(x) \ : \ \sum_{i=1}^{k} w_i + \sum_{j=1}^{k'} w'_j = 1 \right\}$$

  We can approximate arbitrarily finely (with respect to total variation) mixture $m(x)$ for any $\epsilon > 0$ by $g(x;\alpha) \simeq (1-\epsilon)m(x) + \epsilon m'(x)$ with $\alpha = ((1-\epsilon)w, \epsilon w')$ (so that $\sum_{i=1}^{k+k'} \alpha_i = 1$) and $m'(x) \simeq g(x;\alpha') = \epsilon m(x) + (1-\epsilon)m'(x)$ with $\alpha' = (\epsilon w, (1-\epsilon)w')$ (and $\sum_{i=1}^{k+k'} \alpha'_i = 1$). Therefore $\mathrm{KL}(m(x) : m'(x)) \simeq \mathrm{KL}(g(x;\alpha) : g(x;\alpha')) = B_{F^*}(\alpha : \alpha')$, where $F^*(\alpha) = \int g(x;\alpha)\log g(x;\alpha)\mathrm{d}x$ is the Shannon information (negative Shannon entropy) for the composite mixture family. Again, the Bregman divergence $B_{F^*}(\alpha : \alpha')$ is necessarily finite but $\mathrm{KL}(m(x) : m'(x))$ between mixtures may be potentially infinite when the KL integral diverges (hence, the condition on Jeffreys divergence finiteness). Interestingly, this Shannon information can be arbitrarily closely approximated when considering isotropic Gaussians [13]. Notice that the convex conjugate $F(\theta)$ of the continuous Shannon neg-entropy $F^*(\eta)$ is the log-sum-exp function on the inverse soft map.

References

1. Huang, Z.K.; Chau, K.W. A new image thresholding method based on Gaussian mixture model. *Appl. Math. Comput.* **2008**, *205*, 899–907.
2. Seabra, J.; Ciompi, F.; Pujol, O.; Mauri, J.; Radeva, P.; Sanches, J. Rayleigh mixture model for plaque characterization in intravascular ultrasound. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 1314–1324.
3. Julier, S.J.; Bailey, T.; Uhlmann, J.K. Using Exponential Mixture Models for Suboptimal Distributed Data Fusion. In Proceedings of the 2006 IEEE Nonlinear Statistical Signal Processing Workshop, Cambridge, UK, 13–15 September 2006; IEEE: New York, NY, USA, 2006; pp. 160–163.
4. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
5. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.

6. Watanabe, S.; Yamazaki, K.; Aoyagi, M. *Kullback Information of Normal Mixture is Not an Analytic Function*; Technical Report of IEICE Neurocomputing; The Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, 2004; pp. 41–46. (In Japanese)
7. Michalowicz, J.V.; Nichols, J.M.; Bucholtz, F. Calculation of differential entropy for a mixed Gaussian distribution. *Entropy* **2008**, *10*, 200–206.
8. Pichler, G.; Koliander, G.; Riegler, E.; Hlawatsch, F. Entropy for singular distributions. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 2484–2488.
9. Huber, M.F.; Bailey, T.; Durrant-Whyte, H.; Hanebeck, U.D. On entropy approximation for Gaussian mixture random vectors. In Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, Seoul, Korea, 20–22 August 2008; IEEE: New York, NY, USA, 2008; pp. 181–188.
10. Yamada, M.; Sugiyama, M. Direct importance estimation with Gaussian mixture models. *IEICE Trans. Inf. Syst.* **2009**, *92*, 2159–2162.
11. Durrieu, J.L.; Thiran, J.P.; Kelly, F. Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian Mixture Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; IEEE: New York, NY, USA, 2012; pp. 4833–4836.
12. Schwander, O.; Marchand-Maillet, S.; Nielsen, F. Comix: Joint estimation and lightspeed comparison of mixture models. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, 20–25 March 2016; pp. 2449–2453.
13. Moshksar, K.; Khandani, A.K. Arbitrarily Tight Bounds on Differential Entropy of Gaussian Mixtures. *IEEE Trans. Inf. Theory* **2016**, *62*, 3340–3354.
14. Mezuman, E.; Weiss, Y. A Tight Convex Upper Bound on the Likelihood of a Finite Mixture. *arXiv* **2016**, arXiv:1608.05275.
15. Amari, S.-I. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016; Volume 194.
16. Nielsen, F.; Sun, K. Guaranteed Bounds on the Kullback–Leibler Divergence of Univariate Mixtures. *IEEE Signal Process. Lett.* **2016**, *23*, 1543–1546.
17. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
18. Calafiore, G.C.; El Ghaoui, L. *Optimization Models*; Cambridge University Press: Cambridge, UK, 2014.
19. Shen, C.; Li, H. On the dual formulation of boosting algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 2216–2231.
20. Beck, A. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2014.
21. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
22. De Berg, M.; van Kreveld, M.; Overmars, M.; Schwarzkopf, O.C. *Computational Geometry*; Springer: Heidelberg, Germany, 2000.
23. Setter, O.; Sharir, M.; Halperin, D. *Constructing Two-Dimensional Voronoi Diagrams via Divide-and-Conquer of Envelopes in Space*; Springer: Heidelberg, Germany, 2010.
24. Devillers, O.; Golin, M.J. Incremental algorithms for finding the convex hulls of circles and the lower envelopes of parabolas. *Inf. Process. Lett.* **1995**, *56*, 157–164.
25. Nielsen, F.; Yvinec, M. An output-sensitive convex hull algorithm for planar objects. *Int. J. Comput. Geom. Appl.* **1998**, *8*, 39–65.
26. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, China, 26–29 September 2010; IEEE: New York, NY, USA, 2010; pp. 3621–3624.
27. Sharir, M.; Agarwal, P.K. *Davenport-Schinzel Sequences and Their Geometric Applications*; Cambridge University Press: Cambridge, UK, 1995.
28. Bronstein, M. Algorithms and computation in mathematics. In *Symbolic Integration. I. Transcendental Functions*; Springer: Berlin, Germany, 2005.
29. Carreira-Perpinan, M.A. Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1318–1323.

30. Aprausheva, N.N.; Sorokin, S.V. Exact equation of the boundary of unimodal and bimodal domains of a two-component Gaussian mixture. *Pattern Recognit. Image Anal.* **2013**, *23*, 341–347.
31. Learned-Miller, E.; DeStefano, J. A probabilistic upper bound on differential entropy. *IEEE Trans. Inf. Theory* **2008**, *54*, 5223–5230.
32. Amari, S.-I. α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
33. Cichocki, A.; Amari, S.I. Families of Alpha- Beta- and Gamma-Divergences: Flexible and Robust Measures of Similarities. *Entropy* **2010**, *12*, 1532–1568.
34. Póczos, B.; Schneider, J. On the Estimation of α-Divergences. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 609–617.
35. Nielsen, F.; Nock, R. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv* **2011**, arXiv:1105.3259.
36. Minka, T. *Divergence Measures and Message Passing*; Technical Report MSR-TR-2005-173; Microsoft Research: Cambridge, UK, 2005.
37. Améndola, C.; Drton, M.; Sturmfels, B. Maximum Likelihood Estimates for Gaussian Mixtures Are Transcendental. *arXiv* **2015**, arXiv:1508.06958.
38. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.* **1909**, *136*, 210–271. (In German)
39. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820.
40. Nielsen, F.; Nock, R. A closed-form expression for the Sharma-Mittal entropy of exponential families. *J. Phys. A Math. Theor.* **2012**, *45*, 032003.
41. Nielsen, F.; Nock, R. On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences. *IEEE Signal Process. Lett.* **2014**, *21*, 10–13.
42. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
43. Jarosz, W. Efficient Monte Carlo Methods for Light Transport in Scattering Media. Ph.D. Thesis, University of California, San Diego, CA, USA, 2008.
44. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
45. Havrda, J.; Charvát, F. Quantification method of classification processes. Concept of structural α-entropy. *Kybernetika* **1967**, *3*, 30–35.
46. Liang, X. A Note on Divergences. *Neural Comput.* **2016**, *28*, 2045–2062.
47. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
48. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860.
49. Nielsen, F.; Boissonnat, J.D.; Nock, R. On Bregman Voronoi diagrams. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007; pp. 746–755.
50. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi diagrams. *Discret. Comput. Geom.* **2010**, *44*, 281–307.
51. Foster, D.V.; Grassberger, P. Lower bounds on mutual information. *Phys. Rev. E* **2011**, *83*, 010101.
52. Nielsen, F.; Sun, K. PyKLGMM: Python Software for Computing Bounds on the Kullback-Leibler Divergence between Mixture Models. 2016. Available online: https://www.lix.polytechnique.fr/~nielsen/KLGMM/ (accessed on 6 December 2016).
53. Cobb, L.; Koppstein, P.; Chen, N.H. Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. Am. Stat. Assoc.* **1983**, *78*, 124–130.
54. Nielsen, F.; Nock, R. Patch matching with polynomial exponential families and projective divergences. In Proceedings of the 9th International Conference Similarity Search and Applications (SISAP), Tokyo, Japan, 24–26 October 2016.