COMBINATORIAL BOUNDS ON THE α-DIVERGENCE OF UNIVARIATE MIXTURE MODELS

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

[1] École Polytechnique [2] Sony Computer Science Laboratories Inc.
[3] King Abdullah University of Science and Technology

ABSTRACT

We derive lower- and upper-bounds of α-divergence between univariate mixture models with components in the exponential family. Three pairs of bounds are presented in order with increasing quality and increasing computational cost. They are verified empirically through simulated Gaussian mixture models. The presented methodology generalizes to other divergence families relying on Hellinger-type integrals.

Index Terms— α-divergence, information geometry

1. INTRODUCTION

Information divergence [1] is a fundamental concept in signal processing and machine learning. For example, any statistical parametric learning can be regarded as a projection from the empirical distribution to a parameter manifold according to some divergence. It is therefore important to compute accurately these divergence measures.

This paper studies divergences between a pair of given univariate mixture models $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{j=1}^{k'} w'_j p'_j(x)$ defined on the support $\mathcal{X}$. Their α-divergence [2, 3, 4, 1] is *a family* of information divergence parametrized by $\alpha \in \Re \setminus \{0, 1\}$, defined as

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

which clearly satisfies $D_\alpha\left(m : m'\right) = D_{1-\alpha}\left(m' : m\right)$. Let $\alpha \to 1$, we get the Kullback-Leibler (KL) divergence (see [5] for a proof):

$$\lim_{\alpha\to 1} D_\alpha(m : m') = \mathrm{KL}(m : m') = \int_{\mathcal{X}} m(x)\log\frac{m(x)}{m'(x)}\mathrm{d}x, \tag{2}$$

and $\alpha \to 0$ gives the reverse KL divergence: $\lim_{\alpha\to 0} D_\alpha(m : m') = \mathrm{KL}(m' : m)$. Other interesting values [3] includes $\alpha = 1/2$ (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, however, $D_\alpha(m : m')$ only satisfies for any $m(x)$ and $m'(x)$ $D_\alpha\left(m : m'\right) \geq 0$ with $D_\alpha\left(m : m'\right) = 0$ if and only if $m = m'$. It is neither symmetric nor admitting the triangle inequality. Minimization of α-divergence allows one to trade-off mode fitting versus support fitting [6]. The minimizer of α-divergence including the maximum likelihood estimator (MLE) as a special case has interesting connections with transcendental number theory [7].

To compute $D_\alpha\left(m : m'\right)$ for given $m(x)$ and $m'(x)$ reduces to evaluate the Hellinger integral [8, 9]

$$I_\alpha(m : m') = \int_{\mathcal{X}} m(x)^\alpha m'(x)^{1-\alpha}\mathrm{d}x, \tag{3}$$

which in general does not have a closed form, as it was known that the α-divergence of mixture models is not analytic [10]. Moreover, $I_\alpha(m : m')$ may diverge making the α-divergence unbounded. Once $I_\alpha(m : m')$ can be solved, the Rényi and Tsallis divergences [5] and in general Sharma-Mittal divergence [11] can be easily computed. Therefore the results presented here *directly extend* to those divergences.

The Monto-Carlo (MC) stochastic estimation of $I_\alpha(m : m')$ can be computed as $\hat{I}_\alpha^n\left(m : m'\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{m'(x_i)}{m(x_i)}\right)^{1-\alpha}$, where $x_1, \ldots, x_n \sim m(x)$ are independent and identically distributed (i.i.d.). It is consistent so that $\lim_{n\to\infty}\hat{I}_\alpha^n\left(m : m'\right) = I_\alpha\left(m : m'\right)$. However, it requires a large sample and does not guarantee deterministic bounds. The techniques described in [12] work in practice for very close distributions, and do not apply between mixture models. Recently, a series of maximum entropy upper bounds for the differential entropy was reported in [13]. Our recent work [14] provides tight combinatorial bounds of $\mathrm{KL}(m : m')$ based on the log-sum-exp inequalities. Based on similar methodologies, this paper extends these bounds to the family of α-divergence with numerical simulations.

The rest of this paper is organized as follows. Section 2 derives the basic bounds of $D_\alpha(m : m')$. Section 3 improves these bounds at the cost of higher computational complexity. Section 4 proposes another technique to further improve the bounds. Section 5 performs an empirical study on Gaussian mixture models. Section 6 concludes.

2. BASIC BOUNDS

For a pair of given $m(x)$ and $m'(x)$, we only need to derive bounds of $I_\alpha(m : m')$ in eq. (3) so that $L_\alpha(m : m') \leq I_\alpha(m : m') \leq U_\alpha(m : m')$. Then the bounds of $D_\alpha(m : m')$ are obtained by a linear transformation of the range $[L_\alpha(m : m'),\ U_\alpha(m : m')]$. We will assume $\alpha \geq 1/2$. Otherwise we can bound $D_\alpha(m : m')$ by considering the equivalent $D_{1-\alpha}(m' : m)$.

Our general thinking [14] to solve the integral in eq. (3) is to bound the multi-component mixtures by single component distributions in each elementary interval. The mixture model $m(x)$ (same for $m'(x)$) must satisfy $w_{\epsilon(x)}p_{\epsilon(x)}(x) \leq w_i p_i(x) \leq w_{\delta(x)}p_{\delta(x)}(x)$ for all $i = 1, \ldots, k$, where $\epsilon(x) = \operatorname{argmin}_i w_i p_i(x)$ ($\delta(x) = \operatorname{argmax}_i w_i p_i(x)$) denotes the index of the weighted component that is under (above) all the other components at x. Using tools of computational geometry [15, 16], we can compute the *upper and lower envelopes* of the weighted components $\{w_i p_i(x)\}_{i=1}^k$. We regard these computation as a black box (see [16] and our implementation [17] for details) which gives the following result: the support $\mathcal{X}$ is split into ℓ pieces of *elementary intervals* $I_1, \ldots, I_\ell$, where $I_s = (x_s, x_{s+1}), \forall s = 1, \ldots, \ell$, so that $\min \mathcal{X} = x_1 < x_2 < \ldots < x_\ell < x_{\ell+1} = \max \mathcal{X}$, and $\mathcal{X} = \uplus_{s=1}^{\ell} I_s$. Moreover, in each interval I_s, $\delta(x) = \delta_s$ and $\epsilon(x) = \epsilon_s$ are both constants. The size ℓ is related to the Davenport-Schinzel sequence [15]. Therefore we call the bounds "combinatorial" as they are obtained by summing up atomic formulae corresponding to elementary integrals induced by the combinatorial complexity of the weighted density envelopes.

In each I_s, we have

$$c_{\nu_s}p_{\nu_s}(x) \leq m(x) \leq c_{\delta_s}p_{\delta_s}(x), \tag{4}$$

where

$$\boxed{\begin{array}{ll} c_{\nu_s}p_{\nu_s}(x) := k w_{\epsilon_s}p_{\epsilon_s}(x) \quad \text{or} \quad w_{\delta_s}p_{\delta_s}(x), & (5) \\ c_{\delta_s}p_{\delta_s}(x) := k w_{\delta_s}p_{\delta_s}(x). & (6) \end{array}}$$

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

$$A^\alpha_{\nu_s,\nu'_s}(I_s) \leq \int_{I_s} m(x)^\alpha m'(x)^{1-\alpha}\mathrm{d}x \leq A^\alpha_{\delta_s,\delta'_s}(I_s), \tag{7}$$

where

$$A^\alpha_{i,j}(I) = \int_I (c_i p_i(x))^\alpha \left(c'_j p'_j(x)\right)^{1-\alpha} \mathrm{d}x, \tag{8}$$

and I denotes an interval $I = (a, b) \subset \Re$. The other case $\alpha > 1$ is similar by noting that x^α and $x^{1-\alpha}$ are monotonically increasing and decreasing on $\Re^+$, respectively. In conclusion, we have the following bounds of $I_\alpha(m : m')$, where L_α and U_α are shorthands for $L_\alpha(m : m')$ and $U_\alpha(m : m')$:

$$\text{If } 1/2 \leq \alpha < 1, L_\alpha = \sum_{s=1}^{\ell} A^\alpha_{\nu_s,\nu'_s}(I_s),\ U_\alpha = \sum_{s=1}^{\ell} A^\alpha_{\delta_s,\delta'_s}(I_s); \tag{9}$$

$$\text{if } \alpha > 1, \qquad L_\alpha = \sum_{s=1}^{\ell} A^\alpha_{\nu_s,\delta'_s}(I_s),\ U_\alpha = \sum_{s=1}^{\ell} A^\alpha_{\delta_s,\nu'_s}(I_s). \tag{10}$$

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 convex function F is known as the cumulant generating function. Then it is straightforward from eq. (8) that

$$A^\alpha_{i,j}(I) = c_i^\alpha (c'_j)^{1-\alpha} \int_I h(x) \exp\Bigg(\left(\alpha\theta_i + (1-\alpha)\theta'_j\right)^\top t(x) - \alpha F(\theta_i) - (1-\alpha)F(\theta'_j)\Bigg)\mathrm{d}x. \tag{11}$$

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 cumulative distribution function (CDF) of $p(x; \bar\theta)$ as $A^\alpha_{i,j}(I) = c_i^\alpha (c'_j)^{1-\alpha} \exp(F\left(\bar\theta\right) - \alpha F(\theta_i) - (1-\alpha)F(\theta'_j)) \int_I p\left(x; \bar\theta\right) \mathrm{d}x$. The other case $\alpha > 1$ is more difficult: if $\bar\theta$ still lies in $\mathcal{M}_\theta$, $A^\alpha_{i,j}(I)$ can be computed in the same way. 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 [18] and eqs.(61-69) in [5] for related analysis with more details.

For simplicity, we only measure the computational complexity after envelope computation. As computing $A^\alpha_{i,j}(I)$ requires $O(1)$ time, the overall complexity is $O(\ell)$.

3. ADAPTIVE BOUNDS

This section derives the shape-dependent bounds which improve the basic bounds in section 2. We can rewrite a mixture model $m(x)$ in a slab I_s as

$$m(x) = w_{\zeta_s}p_{\zeta_s}(x)\left(1 + \sum_{i \neq \zeta_s} \frac{w_i p_i(x)}{w_{\zeta_s}p_{\zeta_s}(x)}\right), \tag{12}$$

where $w_{\zeta_s}p_{\zeta_s}(x)$ is a weighted component serving as a *reference*. In this paper we only discuss the case that the reference is chosen as the dominating component, i.e., $\zeta_s = \delta_s$. Therefore the ratio

$$\frac{w_i p_i(x)}{w_{\zeta_s}p_{\zeta_s}(x)} = \frac{w_i}{w_{\zeta_s}} \exp\left(\left(\theta_i - \theta_{\zeta_s}\right)^\top t(x) - F(\theta_i) + F(\theta_{\zeta_s})\right) \tag{13}$$

can be bounded in a subrange of $[0,1]$ by analysing the extreme values of $t(x)$ in the slab I_s. This can be done because $t(x)$ usually consists of polynomial functions with finite critical points which can be solved easily. Correspondingly the function $\left(1+\sum_{i\neq\zeta_s}\frac{w_i p_i(x)}{w_{\zeta_s}p_{\zeta_s}(x)}\right)$ in I_s can be bounded in a subrange of $[1,k]$, denoted as $[r_{\zeta_s}, R_{\zeta_s}]$. Hence $r_{\zeta_s}w_{\zeta_s}p_{\zeta_s}(x) \leq m(x) \leq R_{\zeta_s}w_{\zeta_s}p_{\zeta_s}(x)$. This forms better bounds than the basic bounds in section 2 because each component in the slab I_s is bounded more accurately. Therefore, we refine the fundamental bounds of $m(x)$ by replacing the boxed eqs. (5) and (6) with

$$\boxed{c_{\nu_s}p_{\nu_s}(x) := r_{\zeta_s}w_{\zeta_s}p_{\zeta_s}(x),} \quad (14)$$

$$\boxed{c_{\delta_s}p_{\delta_s}(x) := R_{\zeta_s}w_{\zeta_s}p_{\zeta_s}(x).} \quad (15)$$

Then, the improved bounds of $I_\alpha(m : m')$ are given by eqs. (9) and (10) after the replacement.

To evaluate r_{ζ_s} and R_{ζ_s} requires iterating through all components in each slab. Therefore the computational complexity is increased to $O(\ell(k+k'))$.

4. VARIANCE-REDUCED BOUNDS

This section further improves the proposed bounds based on the idea of variance reduction [19]. By assumption, $\alpha \geq 1/2$, then $m(x)^\alpha m'(x)^{1-\alpha}$ is more similar to $m(x)$ rather than $m'(x)$, especially when α is close to 1. 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_s. We will therefore bound this ratio term in

$$\int_{I_s} m(x)^\alpha m'(x)^{1-\alpha}\mathrm{d}x = \int_{I_s} m(x)\left(\frac{m(x)^\alpha m'(x)^{1-\alpha}}{m(x)}\right)\mathrm{d}x$$
$$= \sum_{i=1}^{k}\int_{I_s} w_i p_i(x)\left(\frac{m'(x)}{m(x)}\right)^{1-\alpha}\mathrm{d}x. \quad (16)$$

No matter $1/2 \leq \alpha < 1$ or $\alpha > 1$, the function $x^{1-\alpha}$ must be monotonic on $\Re^+$, and we must have that, in each slab I_s, $(m'(x)/m(x))^{1-\alpha}$ ranges between these two functions:

$$\left(\frac{c'_{\nu'_s}p'_{\nu'_s}(x)}{c_{\delta_s}p_{\delta_s}(x)}\right)^{1-\alpha} \quad \text{and} \quad \left(\frac{c'_{\delta'_s}p'_{\delta'_s}(x)}{c_{\nu_s}p_{\nu_s}(x)}\right)^{1-\alpha}, \quad (17)$$

where $c_{\nu_s}p_{\nu_s}(x)$, $c_{\delta_s}p_{\delta_s}(x)$, $c'_{\nu'_s}p'_{\nu'_s}(x)$ and $c'_{\delta'_s}p'_{\delta'_s}(x)$ are defined in eqs. (14) and (15). Similar to the definition of $A^\alpha_{i,j}(I)$ in eq. (8), we define

$$B^\alpha_{i,j,l}(I) = \int_I w_i p_i(x)\left(\frac{c'_l p'_l(x)}{c_j p_j(x)}\right)^{1-\alpha}\mathrm{d}x. \quad (18)$$

Therefore we have

$$L_\alpha(m:m') = \min\mathcal{S}, \quad U_\alpha(m:m') = \max\mathcal{S},$$
$$\mathcal{S} = \left\{\sum_{s=1}^{\ell}\sum_{i=1}^{k} B^\alpha_{i,\delta_s,\nu'_s}(I_s),\ \sum_{s=1}^{\ell}\sum_{i=1}^{k} B^\alpha_{i,\nu_s,\delta'_s}(I_s)\right\}. \quad (19)$$

The remaining problem is to evaluate $B^\alpha_{i,j,l}(I)$ in eq. (18). Similar to section 2, assuming the components are in the same exponential family with natural parameters θ, then

$$B^\alpha_{i,j,l}(I) = w_i\frac{c'^{1-\alpha}_l}{c^{1-\alpha}_j}\exp\Big(F(\bar\theta) - F(\theta_i) - (1-\alpha)F(\theta'_l) + (1-\alpha)F(\theta_j)\Big)\int_I p(x;\bar\theta)\mathrm{d}x \quad (20)$$

can be computed from the CDF of $p(x;\bar\theta)$ if $\bar\theta = \theta_i + (1-\alpha)\theta'_l - (1-\alpha)\theta_j$ is in the natural parameter space; otherwise $B^\alpha_{i,j,l}(I)$ can be numerically integrated by its definition in eq. (18). Despite there is more computation involved, the computational complexity remains the same as the bounds in section 3, i.e., $O(\ell(k+k'))$.

5. EXPERIMENTS

This section tests the proposed bounds empirically. Due to space limit, a systematical experimental study is left out. We focus on measuring the α-divergence between two simulated pairs of Gaussian mixture models (GMMs). The details of the first pair $\texttt{GMM}_1$ and $\texttt{GMM}_2$ were given previously [14]. $\texttt{GMM}_3$'s components, in the form $(\text{weight}, \text{mean}, \text{standard deviation})$ are given by $(1/3,-2,1/2)$, $(1/3,0,1)$, $(1/3,2,1/2)$; $\texttt{GMM}_4$ is composed of $(1/3,-2,1)$, $(1/3,0,1/5)$, $(1/3,1,1/2)$. The bounds introduced in sections 2 to 4 are denoted as "Basic", "Adaptive" and "VR", respectively.

Figure 1 visualizes these GMMs and plots the estimations of their α-divergence against α. The red lines mean the upper envelope. The dashed vertical lines mean the elementary intervals. For a clear presentation, only the VR bounds are shown, which tightly surround the true value especially for $\texttt{GMM}_1$ and $\texttt{GMM}_2$. This is because their components are more separated than $\texttt{GMM}_3$ and $\texttt{GMM}_4$, and therefore the mixture signals are better estimated by their envelopes. The VR bounds are not necessarily continuous at $\alpha = 1/2$ because when $\alpha < 1/2$, the analytic form of the VR bounds is based on $D_{1-\alpha}(m':m)$ and is different from eq. (19).

For a more quantitative comparison, table 1 shows the estimated α-divergence by MC, Basic, Adaptive, and VR. As D_α is defined for $\alpha \in \Re\setminus\{0,1\}$, the KL bounds CE(A)LB and CE(A)UB [14] are presented for $\alpha = 0$ or 1. Overall, we have the following order of gap size: Basic > Adaptive > VR. The quality of Basic and Adaptive is degraded when $\alpha \to 0$ or $\alpha \to 1$, because the coefficient $\frac{1}{\alpha(1-\alpha)}$ in eq. (1) turns large. VR is recommended in general for bounding α-divergence.

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 at the cost of additional computation. For further improvement, one can split the fundamental intervals into finer pieces, e.g. based on $I_s = (x_s, \frac{x_s+x_{s+1}}{2}) \cup (\frac{x_s+x_{s+1}}{2}, x_{s+1})$, and then evaluate the proposed bounds.

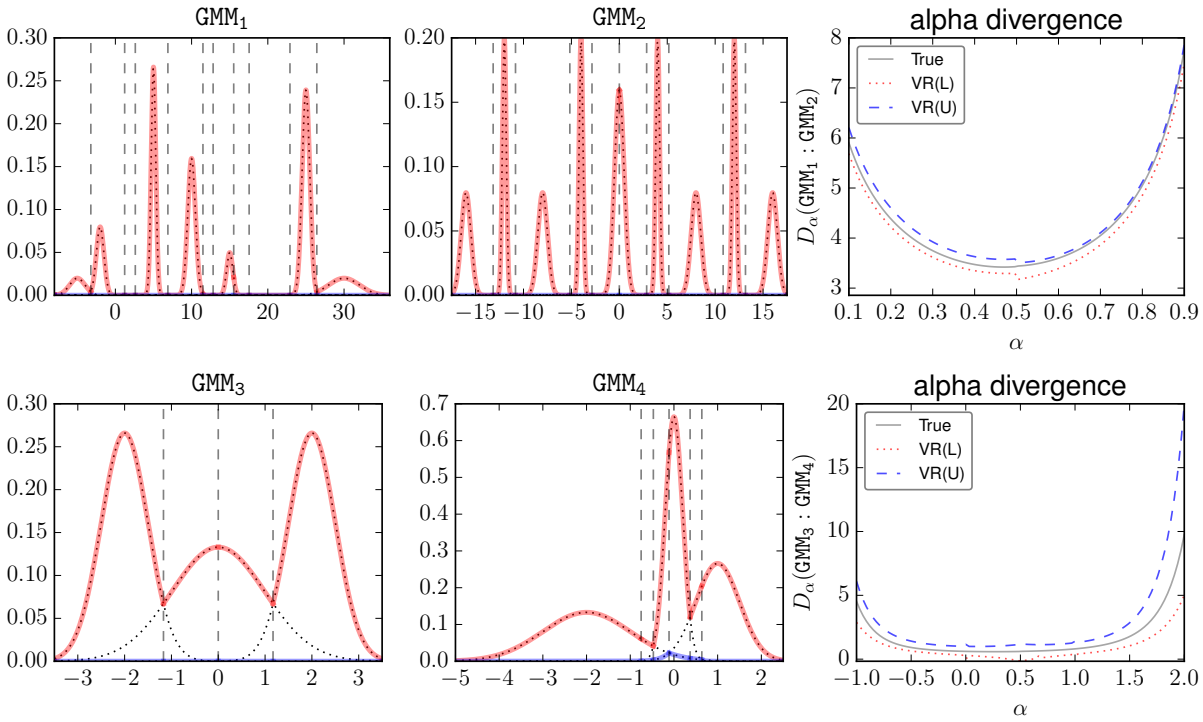

Fig. 1: Two Pairs of GMMs and their α-divergence against different values of α. The "true" value of D_α is estimated by MC using 10^5 random samples. VR(L) and VR(U) denote the variation-reduced lower and upper bounds, respectively. The range of α is selected for each pair for a clear visualization.

	α	MC(10^2)	MC(10^3)	MC(10^4)	Basic		Adaptive		VR	
					L	U	L	U	L	U
$\mathtt{GMM}_1$ & $\mathtt{GMM}_2$	0	15.96 ± 3.9	12.30 ± 1.0	13.63 ± 0.3	11.75	15.89	**13.35**	**14.31**		
	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 ± 0.2	3.47 ± 0.07	-0.60	**3.42**	3.01	**3.42**	**3.17**	3.51
	0.99	40.04 ± 7.7	37.21 ± 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	**93.16**	**94.12**		
$\mathtt{GMM}_3$ & $\mathtt{GMM}_4$	0	0.71 ± 0.2	0.63 ± 0.07	0.62 ± 0.02	-0.44	1.76	**0.30**	**1.00**		
	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	0.92	-0.71	**0.84**	**−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	-0.38	1.81	**0.38**	**1.29**		

Table 1: The estimated D_α and its bounds. The 95% confidence interval is shown for MC. (Run our latest codes [17] to get the best estimation results.)

Note the similarity between KL in eq. (2) and the expression in eq. (16). We give without a formal analysis that: CEAL(U)B [14] is equivalent to VR at the limit $\alpha \to 0$ or $\alpha \to 1$. Experimentally as we slowly set $\alpha \to 0$ or 1, we can observe that VR is consistent with CEAL(U)B.

6. CONCLUSION

This paper introduces 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. They are based on the combinatorial bounds of KL divergence [14], under the general idea to piecewisely bound a mixture by single components. We conclude by emphasizing that the presented methodology can be easily generalized to other divergence [5, 11] relying on Hellinger-type integrals $I_{\alpha,\beta}(p : q) = \int p(x)^\alpha q(x)^\beta \mathrm{d}x$ like the γ-divergence [20] as well as entropy measures [21]. Our Python implementation is available online [17].

7. REFERENCES

[1] Shun-ichi Amari, *Information Geometry and Its Applications*, vol. 194 of *Applied Mathematical Sciences*, Springer, 2016.

[2] Shun-ichi Amari, "α-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.

[3] Andrzej Cichocki and Shun-ichi Amari, "Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[4] Barnabás Póczos and Jeff Schneider, "On the estimation of α-divergences," in *International Conference on Artificial Intelligence and Statistics (AISTATS); JMLR: W&CP 15*, 2011, pp. 609–617.

[5] Frank Nielsen and Richard Nock, "On Rényi and Tsallis entropies and divergences for exponential families," *CoRR*, vol. abs/1105.3259, 2011.

[6] Tom Minka, "Divergence measures and message passing," Tech. Rep. MSR-TR-2005-173, Microsoft Research, 2005.

[7] Carlos Améndola, Mathias Drton, and Bernd Sturmfels, "Maximum likelihood estimates for gaussian mixtures are transcendental," *arXiv*, vol. 1508.06958 [math.ST], 2015.

[8] Ernst Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[9] Tim van Erven and Peter Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[10] Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi, "Kullback information of normal mixture is not an analytic function," *Technical report of IEICE (in Japanese)*, pp. 41–46, 2004.

[11] Frank Nielsen and Richard Nock, "A closed-form expression for the Sharma-Mittal entropy of exponential families," *Journal of Physics A: Mathematical and Theoretical*, vol. 45, no. 3, 2012.

[12] Frank Nielsen and Richard Nock, "On the Chi square and higher-order Chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2014.

[13] Frank Nielsen and Richard Nock, "A series of maximum entropy upper bounds of the differential entropy," *arXiv e-prints*, 2016, 1612.02954 [cs.IT].

[14] Frank Nielsen and Ke Sun, "Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1543–1546, 2016.

[15] Micha Sharir and Pankaj K. Agarwal, *Davenport-Schinzel sequences and their geometric applications*, Cambridge University Press, 1995.

[16] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars, *Computational Geometry: Algorithms and Applications*, Springer-Verlag Berlin Heidelberg, 3rd edition, 2008.

[17] Frank Nielsen and Ke Sun, "PyKLGMM: Python software for computing bounds on the Kullback-Leibler divergence between mixture models," `https://www.lix.polytechnique.fr/~nielsen/KLGMM/`, 2016.

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

[19] Wojciech Jarosz, *Efficient Monte Carlo Methods for Light Transport in Scattering Media*, Ph.D. thesis, UC San Diego, 2008.

[20] Hironori Fujisawa and Shinto Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053–2081, 2008.

[21] Jan Havrda and František Charvát, "Quantification method of classification processes. concept of structural α-entropy," *Kybernetika*, vol. 3, no. 1, pp. 30–35, 1967.