LOWER AND UPPER BOUNDS FOR APPROXIMATION OF THE KULLBACK-LEIBLER DIVERGENCE BETWEEN GAUSSIAN MIXTURE MODELS

J.-L. Durrieu, J.-Ph. Thiran

Signal Processing Laboratory (LTS5)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

F. Kelly

Dept. of Electronic and Electrical Engineering
Trinity College Dublin
Ireland

ABSTRACT

Many speech technology systems rely on Gaussian Mixture Models (GMMs). The need for a comparison between two GMMs arises in applications such as speaker verification, model selection or parameter estimation. For this purpose, the Kullback-Leibler (KL) divergence is often used. However, since there is no closed form expression to compute it, it can only be approximated. We propose lower and upper bounds for the KL divergence, which lead to a new approximation and interesting insights into previously proposed approximations. An application to the comparison of speaker models also shows how such approximations can be used to validate assumptions on the models.

Index Terms— Gaussian Mixture Model (GMM), Kullback-Leibler Divergence, speaker comparison, speech processing.

1. INTRODUCTION

Gaussian Mixture Models (GMMs) are widely used to model unknown probability density functions (PDFs). GMMs have many properties that make them particularly useful for parameter estimation. Kullback-Leibler divergences between two PDFs f and g, $D_{\mathrm{KL}}(f||g)$ can be used to compare such distributions. They arise in various (speech processing) applications: to classify speakers [1], as a cost to minimize for parameter estimation [2] or as a Kernel for Support Vector Machines (SVMs) [3, 4].

Let f and g be two PDFs, defined on $\mathbb{R}^d$, where d is the dimension of the observed vectors $\mathbf{x}$. The Kullback-Leibler divergence (KL divergence) between f and g is defined as:

$$D_{\mathrm{KL}}(f||g) = \int_{\mathbb{R}^d} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \tag{1}$$

When f and g are the PDFs of normal random multivariate variables, *i.e.*

$$\log f(\mathbf{x}) = -\frac{1}{2}\log\left((2\pi)^d|\Sigma^f|\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}^f)^T(\Sigma^f)^{-1}(\mathbf{x}-\boldsymbol{\mu}^f)$$
$$f(\mathbf{x}) \triangleq N(\mathbf{x};\boldsymbol{\mu}^f,\Sigma^f) \text{ and } g(\mathbf{x}) \triangleq N(\mathbf{x};\boldsymbol{\mu}^g,\Sigma^g) \tag{2}$$

where $\boldsymbol{\mu}^f$ and Σ^f ($\boldsymbol{\mu}^g$ and Σ^g, respectively) are the mean and covariance matrix of f (resp. g), T is the transpose operator and $|\Sigma^f|$ the determinant of Σ^f, then the KL divergence between f and g has a closed form expression [5]:

$$D_{\mathrm{KL}}(f||g) = \frac{1}{2}\log\frac{|\Sigma^g|}{|\Sigma^f|} + \frac{1}{2}\mathrm{Tr}((\Sigma^g)^{-1}\Sigma^f) + \frac{1}{2}(\boldsymbol{\mu}^f-\boldsymbol{\mu}^g)^T(\Sigma^g)^{-1}(\boldsymbol{\mu}^f-\boldsymbol{\mu}^g) - \frac{d}{2} \tag{3}$$

For GMMs, however, the KL divergence does not have such a closed form expression. Letting f and g now be the PDFs for two GMMs, the expression of f becomes (with an analogous expression for g):

$$f(\mathbf{x}) = \sum_{a=1}^{A} \omega_a^f f_a(\mathbf{x}) = \sum_{a=1}^{A} \omega_a^f N(\mathbf{x};\boldsymbol{\mu}_a^f,\Sigma_a^f) \tag{4}$$

where A and B are the number of components of the GMM for f and g, respectively, and where f_a and g_b, $\forall a, b$, are individual normal PDFs. It is possible to obtain an accurate approximation to the KL divergence between f and g, via Monte-Carlo estimations, but only at a great computational cost. Fast and reliable approximations for the KL divergence are therefore sought after [6, 7]. We propose the calculation of a lower and an upper bound for the KL divergence between two GMMs. The mean of these bounds then provides an approximation comparable to the approximations proposed by Hershey and Olsen [6]. These bounds are essential when one needs to minimize or maximize the KL divergence, since minimizing the upper bounds implies minimizing the divergence.

We first describe previous proposals for approximations of the KL divergence. Then the proposed lower and upper bounds are derived, with discussions about their interpretations. Finally, some numerical results and an application to speaker model comparison are presented.

2. APPROXIMATIONS TO THE KULLBACK-LEIBLER DIVERGENCE

In this section, we recall the approximations presented in [6].

2.1. Monte Carlo Estimation

The KL divergence can be approximated via Monte-Carlo (MC) estimation. It can indeed be expressed as the expectation of the logarithm of the ratio of f over g, under the PDF f. Let X be a (multivariate) random variable, with PDF f. Then, by definition:

$$D_{\mathrm{KL}}(f||g) = E_X\left[\log\left(f(X)/g(X)\right)\right] \tag{5}$$

This work was partly funded by the Swiss CTI agency, project n. 11359.1 PFES-ES, in collaboration with SpeedLingua SA, Lausanne, Switzerland, and partly funded by the Irish Research Council for Science, Engineering and Technology.

The MC methodology can therefore be applied to estimate such expectations, by the following steps:

1. Draw n independent samples $\mathbf{x}_i$ from the PDF f,
2. Compute $D_{\text{MC},n}(f||g) = \frac{1}{n}\sum_i \log\left(f(\mathbf{x}_i)/g(\mathbf{x}_i)\right)$.

By the law of large numbers, $D_{\text{MC},n}(f||g)$ converges to $D_{\text{KL}}(f||g)$ as n tends to infinity. In this work, we chose to consider this MC approximation with $n = 10^6$ as a reference.

2.2. Product of Gaussians Approximation

Hershey and Olsen proposed a decomposition which serves as basis for several of the approximations [6], including the ones proposed here. Let $L_f(g) = E_X[\log g(X)]$, where $X \sim f$. The KL divergence can then be decomposed as:

$$D_{\text{KL}}(f||g) = L_f(f) - L_f(g) \tag{6}$$

The "product of Gaussians" approximation, D_{prod}, is derived thanks to (6) and Jensen's inequality to find upper bounds for $L_f(g)$ and $L_f(f)$:

$$L_f(g) = \sum_a \omega_a^f \int_{\mathbf{x}} f_a(\mathbf{x}) \log(\sum_b \omega_b^g g_b(\mathbf{x}))d\mathbf{x} \tag{7}$$

$$\leq \sum_a \omega_a^f \log\left(\sum_b \omega_b^g \int_{\mathbf{x}} f_a(\mathbf{x}) g_b(\mathbf{x}) d\mathbf{x}\right) \tag{8}$$

$$L_f(g) \leq \sum_a \omega_a^f \log\left(\sum_b \omega_b^g t_{ab}\right) \tag{9}$$

where $t_{ab} \triangleq \int_{\mathbf{x}} f_a(\mathbf{x}) g_b(\mathbf{x}) d\mathbf{x}$ is the normalization constant of the product of the Gaussians. Similarly, we have:

$$L_f(f) \leq \sum_a \omega_a^f \log\left(\sum_\alpha \omega_\alpha^f z_{a\alpha}\right) \tag{10}$$

$$z_{a\alpha} \triangleq \int_{\mathbf{x}} f_a(\mathbf{x}) f_\alpha(\mathbf{x}) d\mathbf{x} \tag{11}$$

Assuming that these upper bounds are close enough to $L_f(g)$ and $L_f(f)$, respectively, these latter quantities can be approximated by their upper bounds, in order to derive D_{prod} [6]:

$$D_{\text{prod}}(f||g) \triangleq \sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f z_{a\alpha}}{\sum_b \omega_b^g t_{ab}} \tag{12}$$

The closed form expression of the normalization constants is given in Appendix A.

2.3. Variational Approximation

Lower bounds for $L_f(g)$ and $L_f(f)$ can also be derived, using variational parameters as follows [6]:

$$L_f(g) = E_X[\log(\sum_b \omega_b^g g_b(\mathbf{x}))] \tag{13}$$

$$= \sum_a \omega_a^f \int_{\mathbf{x}} f_a(\mathbf{x}) \log\left(\sum_b \omega_b^g \phi_{ba} \frac{g_b(\mathbf{x})}{\phi_{ba}}\right) d\mathbf{x} \tag{14}$$

$$\geq \sum_{ab} \omega_a^f \phi_{ba} \int_{\mathbf{x}} f_a(\mathbf{x}) \log \frac{\omega_b^g g_b(\mathbf{x})}{\phi_{ba}} d\mathbf{x} \tag{15}$$

where $\phi_{ba} \geq 0$, with $\sum_b \phi_{ba} = 1, \forall a, b$. Maximizing the right hand side of the above equation, with respect to ϕ_{ba}, provides a lower bound to $L_f(g)$:

$$L_f(g) \geq \sum_a \omega_a^f \log \sum_b \omega_b^g e^{-D_{\text{KL}}(f_a||g_b)} - \sum_a \omega_a^f H(f_a) \tag{16}$$

where $H(f_a)$ is the entropy of f_a, with a closed form given in Appendix B, and where $D_{\text{KL}}(f_a||g_b)$ also has a closed form expression, as given in Eq. (3). Similarly, $L_f(f)$ has the following variational lower bound:

$$L_f(f) \geq \sum_a \omega_a^f \log \sum_\alpha \omega_\alpha^f e^{-D_{\text{KL}}(f_a||f_\alpha)} - \sum_a \omega_a^f H(f_a) \tag{17}$$

As in the previous section, these lower bounds can be used as approximations for the corresponding quantities in order to derive the "variational" approximation [6]:

$$D_{\text{var}}(f||g) = \sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f e^{-D_{\text{KL}}(f_a||f_\alpha)}}{\sum_b \omega_b^g e^{-D_{\text{KL}}(f_a||g_b)}} \tag{18}$$

These simple closed form expressions make it easy to compute an approximation to D_{KL}, with properties close to that of D_{KL}. However, there does not seem to be a theoretical reason why these quantities should be approximations to D_{KL}, although numerical results have shown their relevance [6]. Since D_{prod} and D_{var} are each the sum of an upper bound with a lower bound, it is difficult to analyze in what sense they approximate the KL divergence.

Based on similar principles, we propose upper and lower bounds that shed a new light on these approximations.

3. UPPER AND LOWER BOUNDS FOR THE KL DIVERGENCE

Strict bounds are mainly useful in the parameter estimation case, and by providing the interval in which we can find the real value of the KL divergence, they provide a well motivated way to design another approximation to the divergence. Using the KL decomposition (6) and the above individual bounds, we propose the following bounds:

Lower bound: Combining Eqs. (9) and (17), we obtain the following lower bound for the KL divergence between GMMs:

$$\underbrace{\sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f e^{-D_{\text{KL}}(f_a||f_\alpha)}}{\sum_b \omega_b^g t_{ab}} - \sum_a \omega_a^f H(f_a)}_{D_{\text{lower}}(f||g)} \leq D_{\text{KL}}(f||g) \tag{19}$$

Upper bound: Similarly, from Eqs. (10) and (16), we obtain:

$$D_{\text{KL}}(f||g) \leq \underbrace{\sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f z_{a\alpha}}{\sum_b \omega_b^g e^{-D_{\text{KL}}(f_a||g_b)}} + \sum_a \omega_a^f H(f_a)}_{D_{\text{upper}}(f||g)} \tag{20}$$

It is worth calculating the mean of D_{lower} and D_{upper}, the "center" of the interval. This is in fact equal to the mean of D_{prod} and D_{var}:

$$\begin{aligned} D_{\text{mean}}(f||g) &\triangleq [D_{\text{upper}}(f||g) + D_{\text{lower}}(f||g)]/2 \\ &= [D_{\text{prod}}(f||g) + D_{\text{var}}(f||g)]/2 \end{aligned} \tag{21}$$

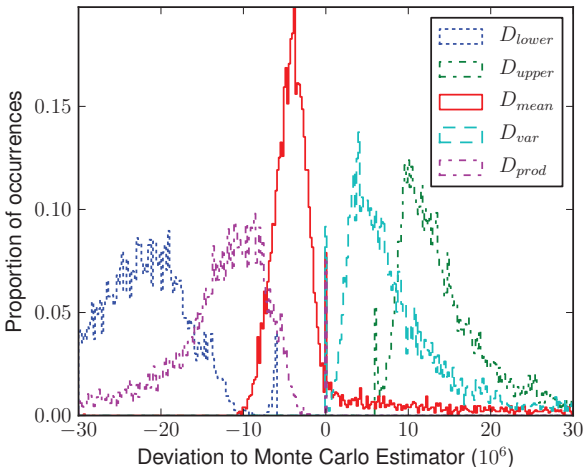

Fig. 1. Histograms of the approximation deviations to the MC estimator, $d = 39$.

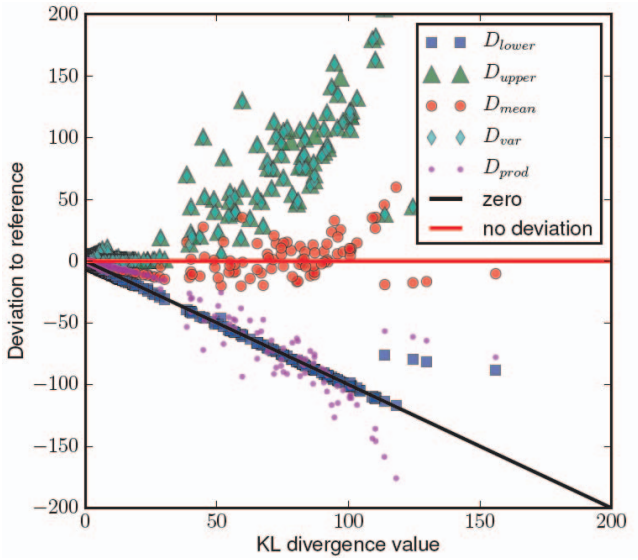

Fig. 2. Deviations from the MC estimator against the reference KL divergence, $d = 3$. In addition to the quantities presented in the article, 2 lines represent the deviation of an "approximation" always equal to 0, and the "no deviation" line.

Since this value is between the lower and upper bounds of the KL divergence, it is a KL approximation as reasonable as D_{prod} or D_{var}. Eq. (21) provides some insight into the results given in [6]: the authors noticed therein that D_{prod} tended to greatly underestimate D_{KL}, while D_{var} was among the best choices as an approximation for D_{KL}. The relation (21) helps us understand why these values can also be considered as approximations, even though their definitions in [6] do not allow much interpretation.

One should also note that for a Gaussian PDF f, $D_{\text{upper}}(f||f) = -D_{\text{lower}}(f||f) = \frac{d}{2}(1-\log 2)$. These "limits", which appear also for GMMs, reveal that the proposed bounds may not be as tight as desired, in spite of the tighter "variational" part of the bound. However, their mean in this case is 0, and D_{mean} is therefore not influenced by these limits. Of the 3 properties of the KL divergence in [6], D_{mean}, like D_{prod} and D_{var}, satisfies the similarity property but not those of identifiablity or positivity.

Finally, one should note that the complexities of the different approximations and bounds are roughly equivalent, in $\mathcal{O}(K^2 d)$ for diagonal covariance matrices and equal number of GMM components K. For the MC estimation, the complexity is in $\mathcal{O}(NKd)$. Since obtaining a reliable MC estimation requires $N \gg K$, the use of approximations is clearly advantageous from the computational complexity aspect.

4. NUMERICAL SIMULATIONS AND DISCUSSIONS

4.1. Deviation analysis

In order to compare these bounds and approximations, we created 100 synthetic GMMs, with the number of components K varying from 1 to 10 (10 GMMs for each value of K), for each of the following dimensions d for the vectors: 1, 3, 39. The deviations of the approximations and bounds to the MC estimator of D_{KL}, with $n = 10^6$ as the reference, are analyzed.

The histograms of the deviations for the different approximations and bounds are shown on Fig. 1, for $d = 39$. As expected, D_{lower} and D_{upper} are respectively below and above the reference. They however tend to greatly under- and over-estimate D_{KL}. They are therefore not suitable approximations to the desired divergence, specifically D_{lower} which is actually almost always close to 0, as can be seen on Fig. 2.

D_{var} and D_{prod} are usually closer to D_{KL}, but, as expected, there is no rule as whether they are above or under D_{KL}: for $d = 1$ and $d = 3$, the corresponding histograms even overlap. D_{prod} is generally under D_{KL}, while D_{var} slightly over-estimates it. D_{mean} seems to be closer to the desired value, with deviations more concentrated near 0. According to Fig. 2, the choice of an approximation may also depend on the actual value of the divergence; for small divergences, the approximations appear to be equivalent. For higher values, D_{mean} is a closer fit to the divergence than D_{var}, which tends to overestimate D_{KL}.

4.2. Speaker model comparison

As mentioned, approximations to the KL divergence and its bounds have numerous applications in speech processing. One application is that of speaker comparison, where it can be used as a similarity measure between GMMs representing speakers [1]. We have carried out a speaker comparison using the derived bounds to illustrate this application.

GMMs were trained for 50 speakers (25 male, 25 female) from the YOHO [8] database via adaptation of a gender-independent Universal Background Model (UBM) of 512 mixtures using 5 minutes of data [9]. Pre-processing involved energy-based silence removal and extraction of MFCC vectors of length 12 appended with delta and acceleration coefficients. The 50 models were compared by extracting D_{mean} between each model pair.

A confusion matrix of the comparisons is given in Fig. 3. The clusters of the within-gender and between-gender comparisons are easily identifiable. Between-gender divergence is generally greater than within-gender. This aligns with intuitive expectations about the relationship between male and female speaker models in the acoustic

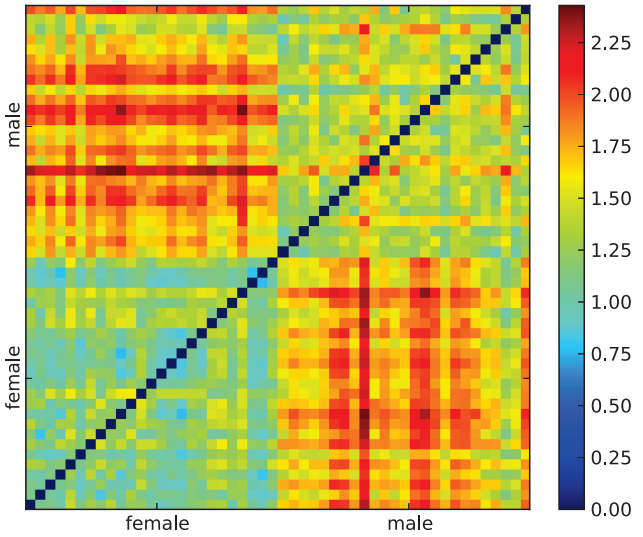

Fig. 3. Confusion matrix for model comparisons with D_{mean}, $d = 36$, $K = 512$.

space *i.e.* that male models are closer to one another than to female models.

By observation, the KL divergence approximation provides a good estimation of the separation of the real, large GMMs in this test case. However, further work is needed to quantify and directly compare the quality of the estimations in the case of real data.

Finally it is worth noting that the correlation between D_{mean} and D_{var} is very high, meaning that either could be used for comparison purposes.

5. CONCLUSIONS

In this article, a lower and an upper bound for the Kullback-Leibler divergence between two GMM PDFs are proposed. The mean of these bounds provides an approximation to the KL divergence which is shown to be equivalent to previously proposed approximations, with a clearer theoretical motivation.

The closed form expressions of the bounds can be used for model comparisons, model validation, classification, or even to compute gradients whenever KL divergences are involved, for parameter estimation, for instance. Using a similar principle as proposed here, it could also be possible to speed up Monte-Carlo approximations, as shown in [10].

The proposed results could be easily extended to any mixture model, with arbitrary distribution PDFs, provided that closed form expressions for individual PDF divergence exist. The proposed bounds and approximation could at last be extended to the case of hidden Markov models.

A. PRODUCT OF TWO GAUSSIANS

The normalizing constant for the product of two normal PDFs f_a and g_b is given by [11]:

$$\log t_{ab} = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_a^f + \Sigma_b^g| - \frac{1}{2}(\boldsymbol{\mu}_b^g - \boldsymbol{\mu}_a^f)^T(\Sigma_a^f + \Sigma_b^g)^{-1}(\boldsymbol{\mu}_b^g - \boldsymbol{\mu}_a^f) \quad (22)$$

B. ENTROPY OF A MULTIVARIATE NORMAL DISTRIBUTION

Let f be a multivariate normal PDF, $f(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, where $\mathbf{x} \in \mathbb{R}^d$. The entropy $H(f)$ of f is:

$$H(f) \triangleq -\int_{\mathbf{x}} f(\mathbf{x})\log f(\mathbf{x})d\mathbf{x} = \frac{1}{2}\log\left((2\pi e)^d|\Sigma|\right) \quad (23)$$

C. REFERENCES

[1] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *Proc. of International Conference on Spoken Language Processing*, Jeju Island, Korea, October 4-8 2004.

[2] Z. Ghahramani and M.I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2, pp. 245–273, 1997.

[3] P.J. Moreno and P.P. Ho, "A new SVM approach to speaker identification and verification using probabilistic distance kernels," in *Proc. of European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 1-4 2003, vol. 3, pp. 2965–2968.

[4] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[5] S.J. Roberts and W.D. Penny, "Variational Bayes for generalized autoregressive models," Tech. Rep., Oxford University, May 22 2002.

[6] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian Mixture Models," in *Proc. of the International Conference on Audio, Speech and Signal Processing*, Honolulu, Hawai, USA, April 15-20 2007, vol. 4, pp. IV–317.

[7] W. M. Campbell and Z. N. Karam, "Simple and efficient speaker comparison using approximate KL divergence," in *Proc. of Interspeech*, Makuhari, Chiba, Japan, Sept. 26-30 2010, pp. 362 – 365.

[8] J. Campbell and A. Higgins, "YOHO speaker verification," Linguistic Data Consortium, 1994.

[9] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] J.-Y. Chen, J. R. Hershey, P. A. Olsen, and E. Yashchin, "Accelerated Monte Carlo for Kullback-Leibler divergence between Gaussian mixture models," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, March 31-April 4 2008.

[11] P. Ahrendt, "The multivariate Gaussian probability distribution," Tech. Rep., Technical University of Denmark, Jan. 2005.