Application

1. Application Number: 1
2. Post Applied for: Associate Professor grade2
3. Name: Shreya B Ballijepalli
4. Category: General
5. Disability: NO
6. Date of Birth: 2019-05-15
7. Nationality: India
8. Gender: Female
9. Marital status: Unmarried
10. Address: 1-19-80/103B, Vijayapuri Colony
11. Email ID: shreyabalijepalli@gmail.com
12. Education

Exam Passed	Board/university	Year of Passing	Specialization	CGPA/Percentage
Bachelors	zsd	2019-05-19	sd	4
Masters	qqw	2019-05-21	wq	3

13. PHD:

University	Year of Graduation	Date of thesis submission	Date of Defence	Specialization	CGPA
hhe	2019-05-10	2019-05-23	2019-05-01	swd	5

14. GATE Year: 1999
15. GATE Score: 1

16. Research Specialization: wea

17. Research Interests: a

18. Post Doc Specialization: a

19. Present Position with Salary Details:

Position	Pay Band	Grade Pay	Consolidated Salary

20. Research/Teaching/Industrial Experience(if any):

Name of the Organization	Start Date	End Date	Full Time (Yes/No)	Designation	Type of Work

21. Projects

Type of Project	project Title	Project Amount	Project Details

22. Referees

Name	Email	Designation	Address
Shreya Ballijepalli	cs15btech11009@iith.ac.in	wer	1-19-80/103B
"Shreya Ballijepalli"	cs15btech11009@iith.ac.in	wert	wert
erty	cs15btech11009@iith.ac.in	1	wert

```
print('Test accuracy:', score[1])
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_7 (Dense)              (None, 100)               30100
_________________________________________________________________
batch_normalization_7 (Batch (None, 100)               400
_________________________________________________________________
activation_7 (Activation)    (None, 100)               0
_________________________________________________________________
dropout_5 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_8 (Dense)              (None, 25)                2525
_________________________________________________________________
batch_normalization_8 (Batch (None, 25)                100
_________________________________________________________________
activation_8 (Activation)    (None, 25)                0
_________________________________________________________________
dropout_6 (Dropout)          (None, 25)                0
_________________________________________________________________
dense_9 (Dense)              (None, 9)                 234
_________________________________________________________________
batch_normalization_9 (Batch (None, 9)                 36
_________________________________________________________________
activation_9 (Activation)    (None, 9)                 0
=================================================================
Total params: 33,395
Trainable params: 33,127
Non-trainable params: 268
_________________________________________________________________
18327 18327
Train on 18327 samples, validate on 6110 samples
Epoch 1/1000
18327/18327 [==============================] - 4s 191us/step - loss: 2.5909 - acc: 0.1030 - val_loss: 2.3216 -
val_acc: 0.0468
Epoch 2/1000
```

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, $D_{\text{KL}}(f||g)$ 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:

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

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]:

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

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:

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

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}:

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

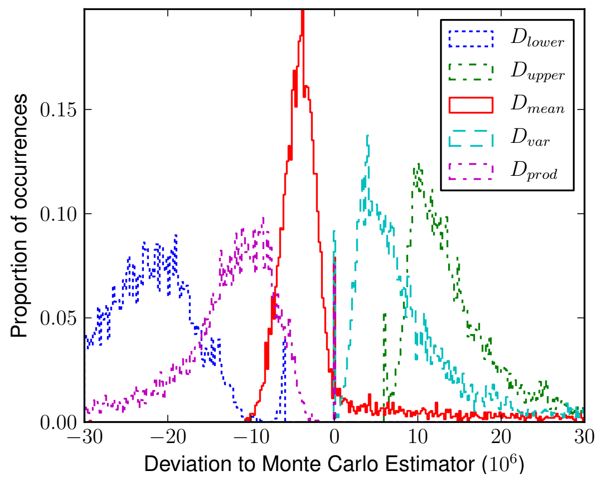


Fig. 1. Histograms of the approximation deviations to the MC estimator, $d = 39$.

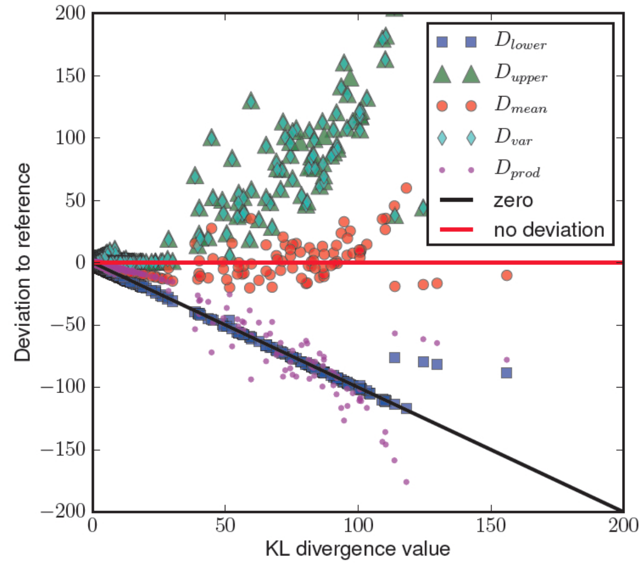


Fig. 2. Deviations from the MC estimator against the reference KL divergence, $d = 3$. In addition to the quantities presented in the article, 2 lines represent the deviation of an "approximation" always equal to 0, and the "no deviation" line.

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

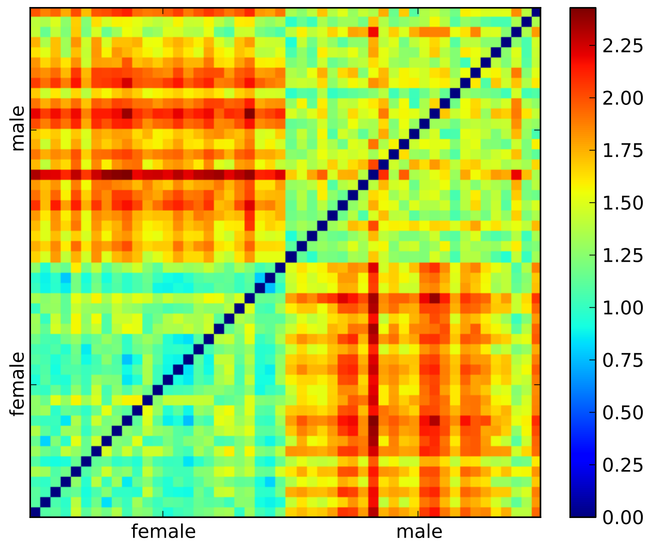


Fig. 3. Confusion matrix for model comparisons with D_{mean}, $d = 36, K = 512$.

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]:

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

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:

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

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.

classify-via-DNN - Google Chrome

Results_OpenKE | Image Resizer to Re | IITH | RP | Python | Convert In | Convert the docx | vasudevram / xtop | python - How to ap | Home | classify-via-DNN

Not secure | 192.168.136.100:8889/notebooks/classify-via-DNN.ipynb

Apps | C Program for... | Deletion in B-... | B-Tree | Set 3 (... | Quora | C++ program t... | Red-Black Tree... | Red-Black Tree... | C++ Multithre... | How to dynam... | Guide to Realt... | sched_setsche...

jupyter classify-via-DNN Last Checkpoint: 19 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run | Raw NBConvert

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense_4 (Dense)              (None, 80)                24080     
_________________________________________________________________
batch_normalization_4 (Batch (None, 80)                320       
_________________________________________________________________
activation_4 (Activation)    (None, 80)                0         
_________________________________________________________________
dropout_3 (Dropout)          (None, 80)                0         
_________________________________________________________________
dense_5 (Dense)              (None, 80)                6480      
_________________________________________________________________
batch_normalization_5 (Batch (None, 80)                320       
_________________________________________________________________
activation_5 (Activation)    (None, 80)                0         
_________________________________________________________________
dropout_4 (Dropout)          (None, 80)                0         
_________________________________________________________________
dense_6 (Dense)              (None, 9)                 729       
_________________________________________________________________
batch_normalization_6 (Batch (None, 9)                 36        
_________________________________________________________________
activation_6 (Activation)    (None, 9)                 0         
=================================================================
Total params: 31,965
Trainable params: 31,627
Non-trainable params: 338
_________________________________________________________________
18327 18327
Train on 18327 samples, validate on 6110 samples
Epoch 1/1000
18327/18327 [==============================] - 3s 164us/step - loss: 1.9684 - acc: 0.2531 - val_loss: 1.7868 - 
val_acc: 0.3370
Epoch 2/1000
 4608/18327 [======>.......................] - ETA: 1s - loss: 1.8258 - acc: 0.3136
```

application_....html | Screenshot f....png | Show all

Case Study - Walmart

B.Shreya
cs15btech11009

Walmart is an American multinational retail corporation that was started in 1962 by Sam Walton. It operates a chain of hypermarkets, department stores, grocery units and it is the largest private employer in the US as well as the world's largest retailer. Its corporate mission is "**Save people money so they can live better"**. It is known as the "global legislator" because its an important emerging private actor in the transformation of lawmaking in the CSR field.

Effect of Walmart's Expansion:
Walmart was first started in 1962 with its target towards rural towns with a population of lesser than ten thousand people. It then extended the company to large cities and opened international stores across the world.
<u>Affects on businesses of small local merchants:</u>
Whenever Walmart starts a new store in a town, all the small merchants and business vendors feel uncomfortable and fear that they can no longer compete with low prices offered by Walmart. There have also been cases where some merchants have pulled out their businesses when Walmart entered their town. This is also known as "**The Walmart Effect".**

<u>Creation of Urban Sprawl,Traffic Congestion:</u>
Walmart's mega stores are built on vast areas of land. By doing this, Walmart is depressing the economic health of communities and other downtown stores. These stores are also built in areas which are not accessible without driving resulting in a lot of traffic.

<u>Environmental Pollution:</u>
The increased traffic has led to more air pollution, water contamination and call for more roads. The landscape is also affected because of these stores as they have huge areas with many unused parking lots as well.

Walmart's initiative towards Corporate Social Responsibility:
From 2007, Walmart publishes its annual report on its website which is known as the 'Global Responsibility Report'. This report talks about Walmart's constant and progressive efforts towards social responsibility issues. It has made investments in education, health, commitments to fight hunger, support to local farmers.

Walmart's Conflicts:

Gender Discrimination:

There have been several charges of gender and racial discrimination on Walmart. Walmart Stores vs Duke et Al was one such case filed on Walmart in 2001 and it is the largest class action lawsuit in US history. The plaintiff class included 1.6 million women who were lead by Betty Dukes. Dukes was a 54-year-old Walmart worker in California who claimed that despite six years of work and positive performance reviews, she was denied training she needed to advance to a higher position.

Dukes and others claimed that women were discriminated against pay and promotions to top management positions violating the Civil Rights Act of 1964. Walmart appealed to the Ninth Circuit in 2005 that the seven lead plaintiffs were not typical or common of the class. Walmart then turned to the Supreme Court in 2010 after the Ninth Circuit court upheld class certification. In 2011, the Supreme Court reversed class certification saying that the millions of plaintiffs and their claims didn't have enough in common. This case, however, didn't end here as the plaintiffs filed an amended lawsuit in October 2011, limiting the class to female employees in California.

After filling of the lawsuit, Walmart incorporated an Advisory Board on Gender Equality and Diversity, which is aimed at providing equal opportunities for all in top management positions. It has also included a Gender Equality and Diversity gender Policy in its 'Global Responsibility Annual Report'.

Below is a picture from Walmart's 2015 Diversity an Inclusion Report.

<u>Child Labour:</u>

In 2005, a Radio Canada programme Zone Libre reported that Walmart was using child labor at two factories in Bangladesh. Walmart employed children between the ages of 10-15 years for less than $50 a month for manufacturing products and exporting them to Canada.

After this incident, Walmart ceased businesses with the two factories immediately. In a 2005 Ethical Sourcing Report of Walmart, it stated that Walmart ceased to do business with 141 companies because of underage labor violations. The stakeholders affected in this were thousands of poor workers who lost their jobs as a result of this.

Walmart’s 2005 and 2012 COC ‘Standard for Suppliers’ explicitly establish it would not tolerate the use of child labor and it sets 14 as the minimum age for any worker.

<u>Walmart’s Bribery Scandal:</u>

Walmart de Mexico, one of the most successful businesses of Walmart was caught in a massive bribery scandal in April 2012. The bribes which totaled more than $24 million were given to the Mexcian government to win permission to open stores at a much rapid phase. (which wouldn’t have been possible according to the Mexican laws). Walmart’s senior management long knew about the scandal and tried to cover it up.
When this case came into light, it was suggested that Walmart undergo a harsh investigation. However, Walmart opted for an in-house investigation and gave the primary responsibility of the investigation to Walmart de

Mexico itself, again another attempt to conceal the fraud. This was not surprisingly “quickly discontinued.”

Walmart used bribery as a mean to monopolize, neglecting the rules that are set to protect a town and its inhabitants from unsafe commercial development. This action also affected the local businesses to a great extent as consumers were driven towards the “low prices” offered by Walmart.

After this scandal became public, Walmart suffered investor lawsuits, numerous investigations from the Department of Justice and Securities and also brand damage. The stakeholders affected in this scandal were Walmart’s investors, local businesses.

This scandal is still under investigation and it is predicted that criminal charges for some of the Walmart executives are certain.

Conclusion:

Walmart is becoming internationally strong and big day by day. Its low prices have really grabbed customers and have resulted in the shut down of a lot of local businesses. There are several organizations like “Wakeup Walmart”, “Walmart March” which are fighting against the company and the company has also been part of several allegations like poor working conditions, low wages, undertrained workers, etc but it is still going strong day by day.

Resources:

https://www.scribd.com/document/373615247/walmart-corporate-social-responsibility-case-study

https://www.ukessays.com/essays/management/understanding-of-the-case-study-walmart-management-essay.php

https://en.wikipedia.org/wiki/Criticism_of_Walmart

https://www.scribd.com/document/342567422/Case-Study-Over-Csr-Conflicts

https://www.businessinsider.com/walmart-bribery-scandal-2012-4?IR=T

https://cdn.corporate.walmart.com/01/8b/4e0af18a45f3a043fc85196c2cbe/2015-diversity-and-inclusion-report.pdf
https://www.academia.edu/10316637/CASE_STUDY_MEXICO_WALMART_SCANDAL

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}:

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

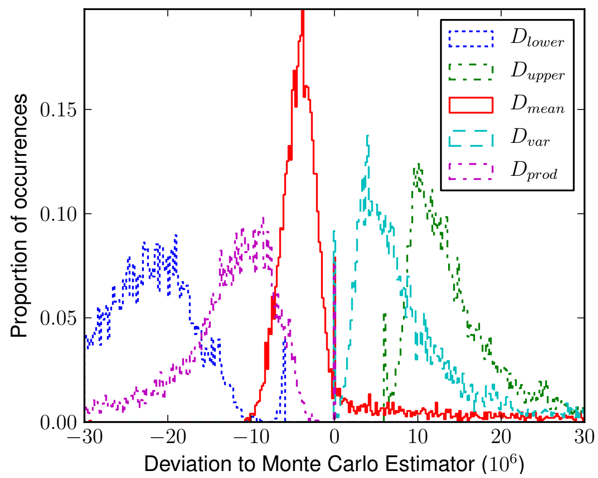


Fig. 1. Histograms of the approximation deviations to the MC estimator, $d = 39$.

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.

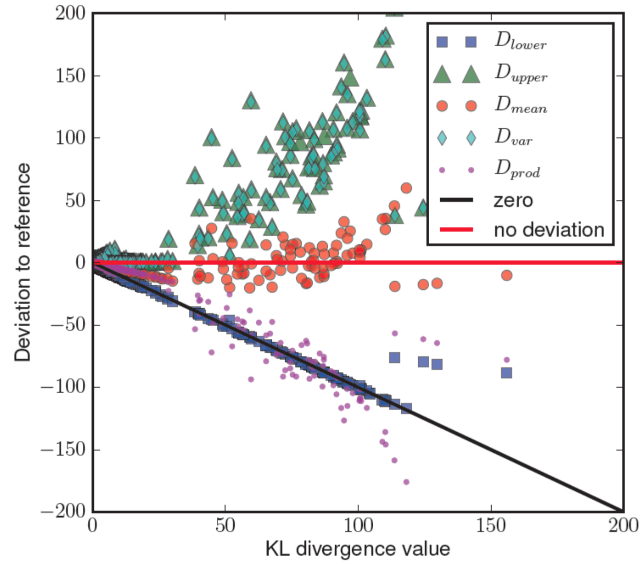


Fig. 2. Deviations from the MC estimator against the reference KL divergence, $d = 3$. In addition to the quantities presented in the article, 2 lines represent the deviation of an "approximation" always equal to 0, and the "no deviation" line.

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

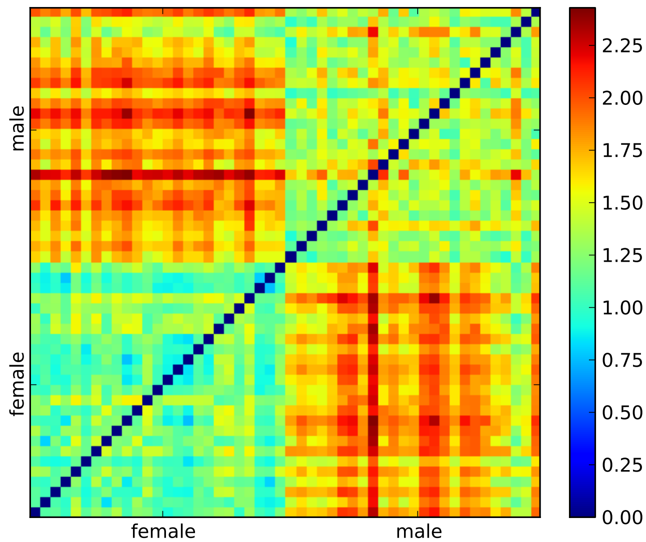


Fig. 3. Confusion matrix for model comparisons with D_{mean}, $d = 36$, $K = 512$.

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]:

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

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:

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

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, Hawaii, 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.

Learning Diary

B.Shreya
Cs15btech11009

Class 1: 3rd April 2019

This class was about an introduction to ethics where we defined ethics as a normative science.

The discussion then moved on to euthanasia which is the termination of a sick person's life to relieve them from pain. It is also known as mercy killing. We discussed if mercy killing a person who is brain dead is right or wrong. This was followed by a discussion on different values like moral values (concern interpersonal behavior), competence values (concern one's own valuation of one's behavior), personal values (concern the ends that are desirable for the self), social values (concern ends that one should desire for the society) and how a person's decisions are based on his/her values.

We defined the term Machiavellianism: It refers to someone who doesn't really care about other's feelings and does their own thing (to show off their materialism). A set of statements were given and each person had to rate each sentence on a scale of 1 to 5. Based on the total score one can determine if they are a machiavellianist or not.

In the end, we watched a video on The High Price of Materialism by Tim Kasser. In this video, he talks about how consumerism and materialism are affecting the lives of people and making them less happy. He says that people buy into marketing messages that "the good life" is "the goods life" and they not only affect the Earth's resources but are also affecting their own well-being. More material possessions pose a greater risk of anxiety, unhappiness, depression and affect over interpersonal relationships with other people. He calls materialistic and pro-social values as a see-saw, as one goes up, the other comes down. He concludes by talking about solutions like promoting intrinsic values, being close to family and friends, staying away from materialism by using ad-blocks, removing advertising from public spaces for living a healthier, well-being and sustainable life.

Class 2: 6th April 2019
This class began with a discussion on traditionalists, modernists, and post-modernists.
Traditionalists - People who don't want to challenge existing values.
Modernists - People who are ok with questioning values.
Post Modernists - Nothing is really fixed for these people (gray).

We then talked about the quote "The meaning of life is to find a meaning" by Victory Frankle who was a psychotherapist. Here he talks about the importance of meaning as a salve against suffering and the secret to happiness. Meaning brought him through all the hardships that he faced in life (being sent to prison, losing his family) and formed the basis for his entire approach for life.
The discussion then moved on to the theories of ethics in which we first discussed concepts like:

- Personal and business ethics (for example helping a friend in an exam)
- Morality and law (for example LGBTQ community)
- Religion and ethics (Is religion, belief in God necessary to live ethically? Atheists can also be ethical.).

In theories we discussed:

- Utilitarianism: To bring about the greatest possible happiness to all those who are concerned with the actions. If many people benefit from the actions and some suffer it is ok, works as long as the good is more than bad. One example is building a dam, this has a lot of benefits but it may affect the people who are evicted from their homes to build this dam. (in this case, the benefits from the dam are much more than the loss of homes to a few people)

- Kantianism: To always act with dignity, respect and do things from a sense of duty. We can take the example of the job of a police officer, he may not be interested in what he does but he has to do his job.

In the end, we watched Jonathan Haidt's 2008 Ted talk where he talks about liberals and conservatives. He talks about how "being open to changes" is a key distinguisher between these two categories. Liberals crave novelty, new ideas whereas conservatives focus on stability (are low on openness to new experiences). He introduces 5 concepts of morality (harm, justice, purity, authority, ingroup) and what liberals and conservatives think about it. He concludes by saying that one should come out of one's own moral matrix and look through other people's perspectives. This would help us in developing moral humility and changing the world into what we want it to be.

Class 3: 13th April 2019

<u>The Corporation Documentary:</u>

This documentary talks about how corporations were first started by the government for the public interest of people but later became private institutions who bothered only about profits over the interests of people and the environment. It talks about several examples related to this context. One such example is the paper mills in the US that dump toxic waste into the rivers damaging the ecosystem and also putting the lives of people in the surrounding neighborhoods at risk. Another example is that of the company Nike pays lower wages to workers in countries like Indonesia so that they can maximize their profits.

It talks about methods adopted by companies solely to maximize profits and how it affected people, for example, it talks about Monsanto, an American agrochemical and agricultural biotechnology corporation that used Bovine Growth hormone on cows to increase milk productivity and how this caused birth defects and increased risk of cancer in the consumers. We can further see examples of corporate sins made by several famous companies one of them being IBM by giving support in the Worl War II.

The film draws attention towards the mistake made by the Supreme court by granting corporations all rights entitled to a human being thus making it one of the most powerful institutions who adopt methods that cause damage to both the environment and humans for the sole purpose of maximizing profits.

Class 4: 21st April 2019

We first began by describing 4 different situations and what is the right thing to do in each situation.

1. A man uses his company's petrol allowance for his personal car.
This is breaking the social contract with the company because the company trusts you. We can consider other examples for this like faculty using printers of college for personal use.
2. A mobile phone seller is unable to reach his target of selling mobile phones. His friend suggests him to lie so that he can make more money and ensure the safety of his job. - Lying to make profits may initially seem like a good idea but it doesn't really work out in the long run as this changes the brand image and the people may stop buying the company's phones in the future.
3. Lying in a contract: Mr. Shah doesn't have required assets but he overstates his assets to get a contract. He thinks that he can payback after he gets a profit. - It is really dangerous to do this as eventually, the lie will be out and he may have to face legal charges. We can consider the example of Satyam Computers in this context.
4. Bribing: A person gives Rs 20 crore commission to get a contract worth Rs 300 crores. This increases his chances of getting the contract but there

is no guarantee that he will truly get the contract. - Bribing to get something will eventually lead to anarchy in the long run where no one is really following the rules and is bribing to get something. (An example of bribing can be seen in the Case Study that I have written in which Walmart bribes the Mexican Government to establish itself as a monopoly in Mexico.)

This was eventually followed by a discussion about the different types of ethical dilemmas.
Ethical dilemmas:
Loyalty vs Integrity: Loyalty is belonging to a certain organization and integrity is about what someone believes personally, Being Creative: This is an answer which is different from the answers available before us (a third answer), Looking for greater good, Looking for long-term solutions rather than just focusing on the short term like in the case of the mobile seller, he needs to think from the long-term perspective, Analyzing problem from the perspective of different stakeholders (needs to be done by a company before taking any major decisions), Seek professional support.

Corporate Social Responsibility (CSR):
This talks about the importance of the company's actions on society and what it does for the welfare of society. For example, the CSR policies of Walmart (taken up from my Case study) include investments in education, health, commitments to fight hunger, support to local farmers. There have been several arguments against CSR which talk about how CSR is another idea for profit maximization as it improves the image of the company. It is also about long-run self-interest as a better community leads to a better workforce, a better business environment, etc. However, CSR has to keep all stakeholders in mind while making decisions and has to make sure that it is sustainable to all stakeholders in the long run. Next we talked about soft power - where companies need to play a fair game with the competitor (Prisoner’s dilemma).

In the end, we watched a Ted talk by Nick Hanauer where he tells his fellow plutocrats about the increasing economic inequality and how this is about to push our society into conditions resembling pre-revolutionary France.
He argues that increasing wages is one of the solutions to his as it will increase demand and eventually profits. He says that the economy has to be dynamic and needs to come up with new solutions to solve this inequality. He concludes his talk by saying that thriving middle class is the source for prosperity in the economy and the Government must intervene and make sure that capitalism doesn't manipulate this.

Class 5: 27th April 2019

<u>The Big Shot:</u>

This movie is on the 2007 housing marketing clash in the United States. This was followed by a global economic downturn, the Great Recession. The main reason for the "bursting of the bubble" was a high default rate in the United States subprime home mortgage sector.
The banks gave high mortgage approval and did not check for any minimum security collateral before lending the loans. Even very risky loans were given a really high rating by these banks. Because of the easy availability of loans, many people who couldn't really afford a home also bought properties through mortgage loans eventually leading to the rise of housing prices.
The statistics say that the national median home price ranged from 2.9 to 3.1 times the median household income. As a part of the increase in house prices, financial agreements based on mortgage payments like mortgage-backed securities (MBS) Collateralized debt obligations (CDO) greatly increased. All these eventually led to the financial crisis.

This movie is based on the above scenario and mainly revolves around five individuals namely Michael Burry, one of the first persons to discover the American housing market bubble,Jared Vennet, a Deutsche Bank salesman who understands Burry’s analysis and decides to sell Burry's credit default swaps for his own profit, Mark Baum, FrontPoint hedge fund manager who takes interest in Vennet’s proposal and two young investors Charlie Geller and Jamie Shipley who invest in swaps after discovering Vennet’s strategy. We can see situations in the movie where people have evacuated their homes as they were unable to pay mortgages and situations where owners of houses are not paying mortgages. During their investigation process, they come to know about other huge frauds like synthetic CDOs where chains of increasingly large bets are placed on faulty loans. We can see situations where banks fool Latin-Americans and immigrants into taking up mortgage loans where the people don’t really know what they are getting into. These really show how banks are interested only in getting high profits and are not really thinking about future circumstances. The film eventually ends with the downfall of the housing market with many people losing their jobs and people like Burry and Vennet earning huge amount of profits.

Hierarchical Density Order Embeddings

Ben Athiwaratkun, Andrew Gordon Wilson
Cornell University
Ithaca, NY 14850, USA

Abstract

By representing words with probability densities rather than point vectors, probabilistic word embeddings can capture rich and interpretable semantic information and uncertainty. The uncertainty information can be particularly meaningful in capturing *entailment* relationships – whereby general words such as "entity" correspond to broad distributions that encompass more specific words such as "animal" or "instrument". We introduce *density order embeddings*, which learn hierarchical representations through encapsulation of probability densities. In particular, we propose simple yet effective loss functions and distance metrics, as well as graph-based schemes to select negative samples to better learn hierarchical density representations. Our approach provides state-of-the-art performance on the WordNet hypernym relationship prediction task and the challenging HyperLex lexical entailment dataset – while retaining a rich and interpretable density representation.

1 Introduction

Learning feature representations of natural data such as text and images has become increasingly important for understanding real-world concepts. These representations are useful for many tasks, ranging from semantic understanding of words and sentences (Mikolov et al., 2013; Kiros et al., 2015), image caption generation (Vinyals et al., 2015), textual entailment prediction (Rocktäschel et al., 2015), to language communication with robots (Bisk et al., 2016).

Meaningful representations of text and images capture visual-semantic information, such as hierarchical structure where certain entities are abstractions of others. For instance, an image caption "A dog and a frisbee" is an abstraction of many images with possible lower-level details such as a dog jumping to catch a frisbee or a dog sitting with a frisbee (Figure 1a). A general word such as "object" is also an abstraction of more specific words such as "house" or "pool". Recent work by Vendrov et al. (2016) proposes learning such asymmetric relationships with *order embeddings* – vector representations of non-negative coordinates with partial order structure. These embeddings are shown to be effective for word hypernym classification, image-caption ranking and textual entailment (Vendrov et al., 2016).

Another recent line of work uses probability distributions as rich feature representations that can capture the semantics and uncertainties of concepts, such as Gaussian word embeddings (Vilnis & McCallum, 2015), or extract multiple meanings via multimodal densities (Athiwaratkun & Wilson, 2017). Probability distributions are also natural at capturing orders and are suitable for tasks that involve hierarchical structures. An abstract entity such as "animal" that can represent specific entities such as "insect", "dog", "bird" corresponds to a broad distribution, encapsulating the distributions for these specific entities. For example, in Figure 1c, the distribution for "insect" is more concentrated than for "animal", with a high density occupying a small volume in space.

Such entailment patterns can be observed from density word embeddings through *unsupervised* training based on word contexts (Vilnis & McCallum, 2015; Athiwaratkun & Wilson, 2017). In the unsupervised settings, density embeddings are learned via maximizing the similarity scores between nearby words. In these cases, the density encapsulation behavior arises due to the word occurrence pattern that a general word can often substitute more specific words; for instance, the word "tea" in a sentence "I like iced tea" can be substituted by "beverages", yielding another natural sentence "I like iced beverages". Therefore, the probability density of a general concept such as "beverages" tends to have a larger variance than specific ones such as "tea", reflecting higher uncertainty in meanings since a general word can be used in many contexts. However, the information from word occurrences alone is not sufficient to train meaningful embeddings of some concepts. For instance, it is fairly common to observe sentences “Look at the cat”, or “Look at the dog”, but not “Look at the mammal”. Therefore, due to the way we typically express natural language, it is unlikely that the word “mammal” would be learned as a distribution that encompasses both “cat” and “dog”, since “mammal” rarely occurs in similar contexts.

Rather than relying on the information from word occurrences, one can do *supervised* training of density embeddings on hierarchical data. In this paper, we propose new training methodology to enable effective supervised probabilistic density embeddings. Despite providing rich and intuitive word representations, with a natural ability to represent order relationships, probabilistic embeddings have only been considered in a small number of pioneering works such as Vilnis & McCallum (2015), and these works are almost exclusively focused on *unsupervised embeddings*. Probabilistic Gaussian embeddings trained directly on labeled data have been briefly considered but perform surprisingly poorly compared to other competing models (Vendrov et al., 2016; Vulić et al., 2016).

Our work reaches a very different conclusion: probabilistic Gaussian embeddings can be *highly effective* at capturing ordering and are suitable for modeling hierarchical structures, and can even achieve state-of-the-art results on hypernym prediction and graded lexical entailment tasks, so long as one uses the right training procedures.

In particular, we make the following contributions.

(a) We adopt a new form of loss function for training hierarchical probabilistic order embeddings.

(b) We introduce the notion of soft probabilistic encapsulation orders and a thresholded divergence-based penalty function, which do not over-penalize words with a sufficient encapsulation.

(c) We introduce a new graph-based scheme to select negative samples to contrast the true relationship pairs during training. This approach incorporates hierarchy information to the negative samples that help facilitate training and has added benefits over the hierarchy-agnostic sampling schemes previously used in literature.

(d) We also demonstrate that initializing the right variance scale is highly important for modeling hierarchical data via distributions, allowing the model to exhibit meaningful encapsulation orders.

The outline of our paper is as follows. In Section 2, we introduce the background for Gaussian embeddings (Vilnis & McCallum, 2015) and vector order embeddings (Vendrov et al., 2016). We describe our training methodology in Section 3, where we introduce the notion of soft encapsulation orders (Section 3.2) and explore different divergence measures such as the expected likelihood kernel, KL divergence, and a family of Rényi alpha divergences (Section 3.3). We describe the experiment details in Section 4 and offer a qualitative evaluation of the model in Section 4.3, where we show the visualization of the density encapsulation behavior. We show quantitative results on the WordNet Hypernym prediction task in Section 4.2 and a graded entailment dataset HyperLex in Section 4.4.

In addition, we conduct experiments to show that our proposed changes to learn Gaussian embeddings contribute to the increased performance. We demonstrate (a) the effects of our loss function in Section A.2.3, (b) soft encapsulation in Section A.2.1, (c) negative sample selection in Section 4.4], and (d) initial variance scale in Section A.2.2.

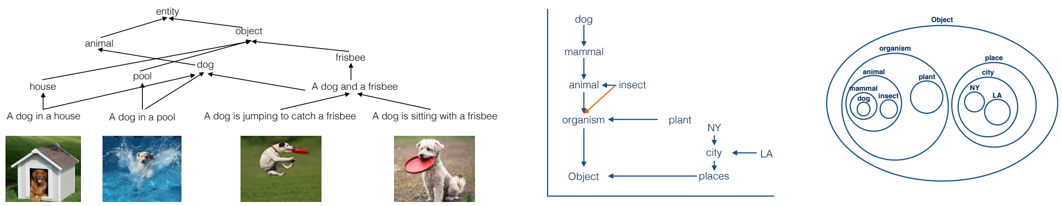
We make our code publicly available.[1]

2 Background and Related Work

2.1 Gaussian Embeddings

Vilnis & McCallum (2015) was the first to propose using probability densities as word embeddings. In particular, each word is modeled as a Gaussian distribution, where the mean vector represents the semantics and the covariance describes the uncertainty or nuances in the meanings. These embeddings are trained on a natural text corpus by maximizing the similarity between words that are in the same local context of sentences. Given a word w with a true context word c_p and a randomly sampled

[1] `https://github.com/benathi/density-order-emb`



(a) (b) (c)

Figure 1: (a) Captions and images in the visual-semantic hierarchy. (b) Vector order embedding (Vendrov et al., 2016) where specific entities have higher coordinate values. (c) Density order embedding where specific entities correspond to concentrated distributions encapsulated in broader distributions of general entities.

word c_n (negative context), Gaussian embeddings are learned by minimizing the rank objective in Equation 1, which pushes the similarity of the true context pair $E(w, c_p)$ above that of the negative context pair $E(w, c_n)$ by a margin m.

$$L_m(w, c_p, c_n) = \max(0, m - E(w, c_p) + E(w, c_n)) \tag{1}$$

The similarity score $E(u, v)$ for words u, v can be either $E(u, v) = -\text{KL}(f_u, f_v)$ or $E(u, v) = \log\langle f_u, f_u \rangle_{L_2}$ where f_u, f_v are the distributions of words u and v, respectively. The Gaussian word embeddings contain rich semantic information and performs competitively in many word similarity benchmarks.

The true context word pairs (w, c_p) are obtained from natural sentences in a text corpus such as Wikipedia. In some cases, specific words can be replaced by a general word in a similar context. For instance, “I love cats” or “I love dogs” can be replaced with “I love animals”. Therefore, the trained word embeddings exhibit lexical entailment patterns where specific words such as “dog” and “cat” are concentrated distributions that are encompassed by a more dispersed distribution of “animal”, a word that “cat” and “dog” entail. The broad distribution of a general word agrees with the *distributional informativeness hypothesis* proposed by Santus et al. (2014), which says that a generic word can occur in more general contexts in place of the specific ones that entail it.

However, some word entailment pairs have weak density encapsulation patterns due to the nature of word diction. For instance, even though “dog” and “cat” both entail “mammal”, it is rarely the case that we observe a sentence “I have a mammal” as opposed to “I have a cat” in a natural corpus; therefore, after training density word embeddings on word occurrences, encapsulation of some true entailment instances do not occur.

2.2 Partial Orders and Vector Order Embeddings

We describe the concepts of partial orders and vector order embeddings proposed by Vendrov et al. (2016), which we will later consider in the context of our hierarchical density order embeddings.

A partial order over a set of points X is a binary relation $\preceq$ such that for $a, b, c \in X$, the following properties hold: (1) $a \preceq a$ (reflexivity); (2) if $a \preceq b$ and $b \preceq a$ then $a = b$ (antisymmetry); and (3) if $a \preceq b$ and $b \preceq c$ then $a \preceq c$ (transitivity). An example of a partially ordered set is a set of nodes in a tree where $a \preceq b$ means a is a child node of b. This concept has applications in natural data such as lexical entailment. For words a and b, $a \preceq b$ means that every instance of a is an instance of b, or we can say that a entails b. We also say that (a, b) has a *hypernym* relationship where a is a hyponym of b and b is a hypernym of a. This relationship is asymmetric since $a \preceq b$ does not necessarily imply $(b \preceq a)$. For instance, `aircraft` $\preceq$ `vehicle` but it is not true that `vehicle` $\preceq$ `aircraft`.

An order-embedding is a function $f : (X, \preceq_X) \to (Y, \preceq_Y)$ where $a \preceq_X b$ if and only if $f(a) \preceq_Y f(b)$. Vendrov et al. (2016) proposes to learn the embedding f on $Y = \mathbb{R}_+^N$ where all coordinates are non-negative. Under $\mathbb{R}_+^N$, there exists a partial order relation called the *reversed product order on* $\mathbb{R}_+^N$: $x \preceq y$ if and only if $\forall i, x_i \geq y_i$. That is, a point x entails y if and only if all the coordinate values of x is higher than y's. The origin represents the most general entity at the top of the order hierarchy and the points further away from the origin become more specific. Figure 1b demonstrates the vector order embeddings on $\mathbb{R}_+^N$. We can see that since `insect` $\preceq$ `animal` and `animal` $\preceq$ `organism`, we can infer directly from the embedding that `insect` $\preceq$ `organism` (orange line, diagonal line). To learn the embeddings, Vendrov et al. (2016) proposes a penalty function $E(x, y) = ||\max(0, y - x)||^2$ for a pair $x \preceq y$ which has the property that it is positive if and only if the order is violated.

2.3 Other Related Work

Li et al. (2017) extends Vendrov et al. (2016) for knowledge representation on data such as ConceptNet (Speer et al., 2016). Another related work by Hockenmaier & Lai (2017) embeds words and phrases in a vector space and uses denotational probabilities for textual entailment tasks. Our models offer an improvement on order embeddings and can be applicable to such tasks, which view as a promising direction for future work.

3 Methodology

In Section 3.1, we describe the partial orders that can be induced by density encapsulation. Section 3.2 describes our training approach that softens the notion of strict encapsulation with a viable penalty function.

3.1 Strict Encapsulation Partial Orders

A partial order on probability densities can be obtained by the notion of encapsulation. That is, a density f is more specific than a density g if f is encompassed in g. The degree of encapsulation can vary, which gives rise to multiple order relations. We define an order relation $\preceq_\eta$ for $\eta \geq 0$ where η indicates the degree of encapsulation required for one distribution to entail another. More precisely, for distributions f and g,

$$f \preceq_\eta g \Leftrightarrow \{x : f(x) > \eta\} \subseteq \{x : g(x) > \eta\}. \tag{2}$$

Note that $\{x : f(x) > \eta\}$ is a set where the density f is greater than the threshold η. The relation in Equation 2 says that f entails g if and only if the set of g contains that of f. In Figure 2, we depict two Gaussian distributions with different mean vectors and covariance matrices. Figure 2 (left) shows the density values of distributions f (narrow, blue) and g (broad, orange) and different threshold levels. Figure 2 (right) shows that different η's give rise to different partial orders. For instance, we observe that neither $f \preceq_{\eta_1} g$ nor $g \preceq_{\eta_1} f$ but $f \preceq_{\eta_3} g$.

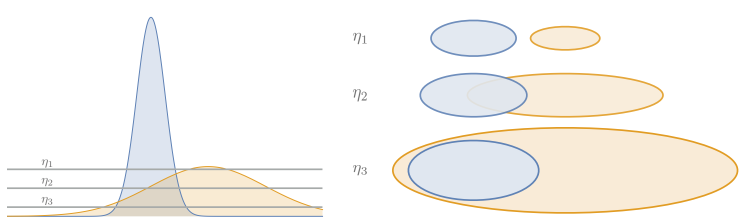


Figure 2: Strict encapsulation orders induced by different η values.

3.2 Soft Encapsulation Orders

A plausible penalty function for the order relation $f \preceq_\eta g$ is a set measure on $\{x : f(x) > \eta\} - \{x : g(x) > \eta\}$. However, this penalty is difficult to evaluate for most distributions, including Gaussians. Instead, we use simple penalty functions based on asymmetric divergence measures between probability densities. Divergence measures $D(\cdot||\cdot)$ have a property that $D(f||g) = 0$ if and only if $f = g$. Using $D(\cdot||\cdot)$ to represent order violation is undesirable since the penalty should be 0 if $f \neq g$ but $f \preceq g$. Therefore, we propose using a thresholded divergence

$$d_\gamma(f, g) = \max(0, D(f||g) - \gamma),$$

which can be zero if f is properly encapsulated in g. We discuss the effectiveness of using divergence thresholds in Section A.2.1.

We note that by using $d_\gamma(\cdot,\cdot)$ as a violation penalty, we no longer have the strict *partial order*. In particular, the notion of transitivity in a partial order is not guaranteed. For instance, if $f \preceq g$ and $g \preceq h$, our density order embeddings would yield $d_\gamma(f,g) = 0$ and $d_\gamma(g,h) = 0$. However, it is not necessarily the case that $d_\gamma(f,h) = 0$ since $D(f||h)$ can be greater than γ. This is not a drawback since a high value of $D(f||h)$ reflects that the hypernym relationship is not direct, requiring many edges from f to h in the hierarchy. The extent of encapsulation contains useful entailment information, as demonstrated in Section 4.4 where our model scores highly correlate with the annotated scores of a challenging lexical entailment dataset and achieves state-of-the-art results.

Another property, antisymmetry, does not strictly hold since if $d_\gamma(f,g) = 0$ and $d_\gamma(g,f) = 0$ does not imply $f = g$. However, in this situation, it is necessary that f and g overlap significantly if γ is small. Due to the fact that the $d_\gamma(\cdot,\cdot)$ does not strictly induce a partial order, we refer to this model as *soft density order embeddings* or simply *density order embeddings*.

3.3 Divergence Measures

3.3.1 Asymmetric Divergence

Kullback-Leibler (KL) Divergence The KL divergence is an asymmetric measure of the difference between probability distributions. For distributions f and g, $\text{KL}(g||f) \equiv \int g(x) \log \frac{g(x)}{f(x)}\, dx$ imposes a high penalty when there is a region of points x such that the density $f(x)$ is low but $g(x)$ is high. An example of such a region is the area on the left of f in Figure 2. This measure penalizes the situation where f is a concentrated distribution relative to g; that is, if the distribution f is encompassed by g, then the KL yields high penalty. For d-dimensional Gaussians $f = \mathcal{N}_d(\mu_f, \Sigma_f)$ and $g = \mathcal{N}_d(\mu_g, \Sigma_g)$,

$$2D_{KL}(f||g) = \log(\det(\Sigma_g)/\det(\Sigma_f)) - d + \text{tr}(\Sigma_g^{-1}\Sigma_f) + (\mu_f - \mu_g)^T \Sigma_g^{-1} (\mu_f - \mu_g) \tag{3}$$

Rényi α-Divergence is a general family of divergence with varying scale of zero-forcing penalty (Renyi, 1961). Equation 4 describes the general form of the α-divergence for $\alpha \neq 0, 1$ (Liese & Vajda, 1987). We note that for $\alpha \to 0$ or 1, we recover the KL divergence and the reverse KL divergence; that is, $\lim_{\alpha\to 1} D_\alpha(f||g) = \text{KL}(f||g)$ and $\lim_{\alpha \to 0} D_\alpha(f||g) = \text{KL}(g||f)$ (Pardo, 2006). The α-divergences are asymmetric for all α's, except for $\alpha = 1/2$.

$$D_\alpha(f||g) = \frac{1}{\alpha(\alpha-1)} \log\left(\int \frac{f(x)^\alpha}{g(x)^{\alpha-1}}\, dx\right) \tag{4}$$

For two multivariate Gaussians f and g, we can write the Rényi divergence as (Pardo, 2006):

$$2D_\alpha(f||g) = -\frac{1}{\alpha(\alpha-1)} \log \frac{\det\left(\alpha\Sigma_g + (1-\alpha)\Sigma_f\right)}{\left(\det\left(\Sigma_f\right)^{1-\alpha} \cdot \det\left(\Sigma_g\right)^\alpha\right)} + (\mu_f - \mu_g)^T (\alpha\Sigma_g + (1-\alpha)\Sigma_f)^{-1} (\mu_f - \mu_g). \tag{5}$$

The parameter α controls the degree of *zero forcing* where minimizing $D_\alpha(f||g)$ for high α results in f being more concentrated to the region of g with high density. For low α, f tends to be *mass-covering*, encompassing regions of g including the low density regions. Recent work by Li & Turner (2016) demonstrates that different applications can require different degrees of zero-forcing penalty.

3.3.2 Symmetric Divergence

Expected Likelihood Kernel The expected likelihood kernel (ELK) (Jebara et al., 2004) is a symmetric measure of affinity, define as $K(f,g) = \langle f, g\rangle_\mathcal{H}$. For two Gaussians f and g,

$$2\log\langle f, g\rangle_\mathcal{H} = -\log\det(\Sigma_f + \Sigma_g) - d\log(2\pi) - (\mu_f - \mu_g)^T(\Sigma_f + \Sigma_g)^{-1}(\mu_f - \mu_g) \tag{6}$$

Since this kernel is a similarity score, we use its negative as our penalty. That is, $D_{\text{ELK}}(f||g) = -2\log\langle f, g\rangle_\mathcal{H}$. Intuitively, the asymmetric measures should be more successful at training density order embeddings. However, a symmetric measure can result in a correct encapsulation order as well, since a general entity often has to minimize the penalty with many specific elements and consequently ends up having a broad distribution to lower the average loss. The expected likelihood kernel is used to train Gaussian and Gaussian Mixture word embeddings on a large text corpus (Vilnis & McCallum, 2015; Athiwaratkun & Wilson, 2017) where the model performs well on the word entailment dataset (Baroni et al., 2012).

3.4 LOSS FUNCTION

To learn our density embeddings, we use a loss function similar to that of Vendrov et al. (2016). Minimizing this function (Equation 7) is equivalent to minimizing the penalty between a true relationship pair (u, v) where $u \preceq v$, but pushing the penalty to be above a margin m for the negative example (u', v') where $u' \npreceq v'$:

$$\sum_{(u,v)\in\mathcal{D}} d(u, v) + \max\{0, m - d(u', v')\} \tag{7}$$

We note that this loss function is different than the rank-margin loss introduced in the original Gaussian embeddings (Equation 1). Equation 7 aims to reduce the dissimilarity of a true relationship pair $d(u, v)$ with no constraint, unlike in Equation 1, which becomes zero if $d(u, v)$ is above $d(u', v')$ by margin m.

3.5 SELECTING NEGATIVE SAMPLES

In many embedding models such as WORD2VEC (Mikolov et al., 2013) or Gaussian embeddings (Vilnis & McCallum, 2015), negative samples are often used in the training procedure to contrast with true samples from the dataset. For flat data such as words in a text corpus, negative samples are selected randomly from a unigram distribution. We propose new graph-based methods to select negative samples that are suitable for hierarchical data, as demonstrated by the improved performance of our density embeddings. In our experiments, we use various combinations of the following methods.

Method **S1**: A simple negative sampling procedure used by Vendrov et al. (2016) is to replace a true hypernym pair (u, v) with either (u, v') or (u', v) where u', v' are randomly sampled from a uniform distribution of vertices. Method **S2**: We use a negative sample (v, u) if (u, v) is a true relationship pair, to make $D(v||u)$ higher than $D(u||v)$ in order to distinguish the directionality of density encapsulation. Method **S3**: It is important to increase the divergence between neighbor entities that do not entail each other. Let $A(w)$ denote all descendants of w in the training set $\mathcal{D}$, including w itself. We first randomly sample an entity $w \in \mathcal{D}$ that has at least 2 descendants and randomly select a descendant $u \in A(w) - \{w\}$. Then, we randomly select an entity $v \in A(w) - A(u)$ and use the random neighbor pair (v, u) as a negative sample. Note that we can have $u \preceq v$, in which case the pair (v, u) is a reverse relationship. Method **S4**: Same as **S3** except that we sample $v \in A(w) - A(u) - \{w\}$, which excludes the possibility of drawing (w, u).

4 EXPERIMENTS

We have introduced density order embeddings (DOE) to model hierarchical data via encapsulation of probability densities. We propose using a new loss function, graph-based negative sample selections, and a penalty relaxation to induce soft partial orders. In this section, we show the effectiveness of our model on WORDNET hypernym prediction and a challenging graded lexical entailment task, where we achieve state-of-the-art performance.

First, we provide the training details in Section 4.1 and describe the hypernym prediction experiment in 4.2. We offers insights into our model with the qualitative analysis and visualization in Section 4.3. We evaluate our model on HYPERLEX, a lexical entailment dataset in Section 4.4.

4.1 TRAINING DETAILS

We have a similar data setup to the experiment by Vendrov et al. (2016) where we use the transitive closure of WORDNET noun hypernym relationships which contains $82,115$ synsets and $837,888$ hypernym pairs from $84,427$ direct hypernym edges. We obtain the data using the WORDNET API of NLTK version 3.2.1 (Loper & Bird, 2002).

The validation set contains 4000 true hypernym relationships as well as 4000 false hypernym relationships where the false hypernym relationships are constructed from the **S1** negative sampling described in Section 3.5. The same process applies for the test set with another set of 4000 true hypernym relationships and 4000 false hypernym relationships.

We use d-dimensional Gaussian distributions with diagonal covariance matrices. We use $d = 50$ as the default dimension and analyze the results using different d's in Section A.2.4. We initialize the mean vectors to have a unit norm and normalize the mean vectors in the training graph. We initialize the diagonal variance components to be all equal to β and optimize on the unconstrained space of $\log(\Sigma)$. We discuss the important effects of the initial variance scale in Section A.2.2.

We use a minibatch size of 500 true hypernym pairs and use varying number of negative hypernym pairs, depending on the negative sample combination proposed in Section 3.5. We discuss the results for many selection strategies in Section 4.4. We also experiment with multiple divergence measures $D(\cdot||\cdot)$ described in Section 3.3. We use $D(\cdot||\cdot) = D_{KL}(\cdot||\cdot)$ unless stated otherwise. Section A.2.5 considers the results using the α-divergence family with varying degrees of zero-forcing parameter α's. We use the Adam optimizer (Kingma & Ba, 2014) and train our model for at most 20 epochs. For each energy function, we tune the hyperparameters on grids. The hyperparameters are the loss margin m, the initial variance scale β, and the energy threshold γ. We evaluate the results by computing the penalty on the validation set to find the best threshold for binary classification, and use this threshold to perform prediction on the test set. Section A.1 describes the hyperparameters for all our models.

4.2 Hypernym Prediction

We show the prediction accuracy results on the test set of WordNet hypernyms in Table 1. We compare our results with **vector order-embeddings** (VOE) by Vendrov et al. (2016) (VOE model details are in Section 2.2). Another important baseline is the **transitive closure**, which requires no learning and classifies if a held-out edge is a hypernym relationship by determining if it is in the union of the training edges. **word2gauss** and **word2gauss**† are the Gaussian embeddings trained using the loss function in Vilnis & McCallum (2015) (Equation 1) where **word2gauss** is the result reported by Vendrov et al. (2016) and **word2gauss**† is the best performance of our replication (see Section A.2.3 for more details). Our density order embedding (DOE) outperforms the implementation by Vilnis & McCallum (2015) significantly; this result highlights the fact that our different approach for training Gaussian embeddings can be crucial to learning hierarchical representations.

We observe that the symmetric model (ELK) performs quite well for this task despite the fact that the symmetric metric cannot capture directionality. In particular, ELK can accurately detect pairs of concepts with no relationships when they're far away in the density space. In addition, for pairs that are related, ELK can detect pairs that overlap significantly in density space. The lack of directionality has more pronounced effects in the graded lexical entailment task (Section 4.4) where we observe a high degradation in performance if ELK is used instead of KL.

We find that our method outperforms vector order embeddings (VOE) (Vendrov et al., 2016). We also find that DOE is very strong in a 2-dimensional Gaussian embedding example, trained for the purpose of visualization in Section 4.3, despite only having only 4 parameters: 2 from 2-dimensional μ and another 2 from the diagonal Σ. The results of DOE using a symmetric measure also outperforms the baselines on this experiment, but has a slightly lower accuracy than the asymmetric model.

Figure 3 offers an explanation as to why our density order embeddings might be easier to learn, compared to the vector counterpart. In certain cases such as fitting a general concept `entity` to the embedding space, we simply need to adjust the distribution of `entity` to be broad enough to encompass all other concepts. In the vector counterpart, it might be required to shift many points further from the origin to accommodate `entity` to reduce cascading order violations.

Figure 3: (**Left**) Adding a concept `entity` to vector order embedding (**Right**) Adding a concept `entity` to density order embedding

Table 1: Classification accuracy on hypernym relationship test set from WordNet.

Method	Test Accuracy (%)
transitive closure	88.2
word2gauss	86.6
word2gauss†	88.6
VOE (symmetric)	84.2
VOE	90.6
DOE (ELK)	92.1
DOE (KL, reversed)	83.2
DOE (KL)	**92.3**
DOE (KL, $d = 2$)	89.2

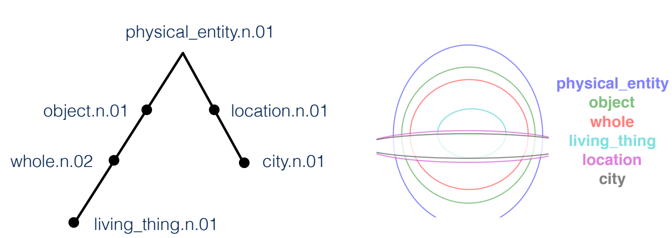
4.3 Qualitative Analysis

For qualitative analysis, we additionally train a 2-dimensional Gaussian model for visualization. Our qualitative analysis shows that the encapsulation behavior can be observed in the trained model. Figure 4 demonstrates the ordering of synsets in the density space. Each ellipse represents a Gaussian distribution where the center is given by the mean vector μ and the major and minor axes are given by the diagonal standard deviations $\sqrt{\Sigma}$, scaled by 300 for the x axis and 30 for the y axis, for visibility.

Most hypernym relationships exhibit encapsulation behavior where the hypernym encompasses the synset that entails it. For instance, the distribution of `whole.n.02` is subsumed in the distribution of `physical_entity.n.01`. Note that `location.n.01` is not entirely encapsulated by `physical_entity.n.01` under this visualization. However, we can still predict which entity should be the hypernym among the two since the KL divergence of one given another would be drastically different. This is because a large part of `physical_entity.n.01` has considerable density at the locations where location.n.01 has very low density. This causes KL(`physical_entity.n.01` || `location.n.01`) to be very high (5103) relative to KL(`location.n.01` || `physical_entity.n.01`) (206). Table 2 shows the KL values for all pairs where we note that the numbers are from the full model ($d = 50$).

Another interesting pair is `city.n.01` $\preceq$ `location.n.01` where we see the two distributions have very similar contours and the encapsulation is not as distinct. In our full model $d = 50$, the distribution of `location.n.01` encompasses `city.n.01`'s, indicated by low KL(`city.n.01`||`location.n.01`) but high KL(`location.n.01`||`city.n.01`).

Figure 4 (**Right**) demonstrates the idea that synsets on the top of the hypernym hierarchy usually have higher "volume". A convenient metric that reflects this quantity is $\log\det(\Sigma)$ for a Gaussian distribution with covariance Σ. We can see that the synset, `physical_entity.n.01`, being the hypernym of all the synsets shown, has the highest $\log\det(\Sigma)$ whereas entities that are more specific such as `object.n.01`, `whole.n.02` and `living_thing` have decreasingly lower volume.



Synset	$\log\det(\Sigma)$
physical_entity.n.01	-219.67
object.n.01	-224.00
whole.n.02	-233.39
living_thing.n.01	-238.33
location.n.01	-261.76
city.n.01	-261.89

Figure 4: [best viewed electronically] (**Left**) Synsets and their hypernym relationships from WordNet. (**Middle**) Visualization of our 2-dimensional Gaussian order embedding. (**Right**) The Gaussian "volume" ($\log\det\Sigma$) of the 50-dimensional Gaussian model.

Table 2: KL(column||row). Cells in boldface indicate true WORDNET hypernym relationships (column $\preceq$ row). Our model predicts a synset pair as a hypernym if the KL less than 1900, where this value is tuned based on the validation set. Most relationship pairs are correctly predicted except for the underlined cells.

	city	location	living_thing	whole	object	physical_entity
city	0	<u>1025</u>	4999	4673	23673	4639
location	**159**	0	4324	4122	26121	5103
living_thing	3623	6798	0	<u>1452</u>	2953	5936
whole	3033	6367	**66**	0	6439	6682
object	<u>138</u>	<u>80</u>	**125**	**77**	0	6618
physical_entity	**232**	**206**	**193**	**166**	**152**	0

4.4 GRADED LEXICAL ENTAILMENT

HYPERLEX is a lexical entailment dataset which has fine-grained human annotated scores between concept pairs, capturing varying degrees of entailment (Vulić et al., 2016). Concept pairs in HYPERLEX reflect many variants of hypernym relationships, such as `no-rel` (no lexical relationship), `ant` (antonyms), `syn` (synonyms), `cohyp` (sharing a hypernym but not a hypernym of each other), `hyp` (hypernym), `rhyp` (reverse hypernym). We use the noun dataset of HYPERLEX for evaluation, which contains 2,163 pairs.

We evaluate our model by comparing our model scores against the annotated scores. Obtaining a high correlation on a fine-grained annotated dataset is a much harder task compared to a binary prediction, since performing well requires meaningful model scores in order to reflect nuances in hypernymy. We use negative divergence as our score for hypernymy scale where large values indicate high degrees of entailment.

We note that the concepts in our trained models are WORDNET synsets, where each synset corresponds to a specific meaning of a word. For instance, `pop.n.03` has a definition "a sharp explosive sound as from a gunshot or drawing a cork" whereas `pop.n.04` corresponds to "music of general appeal to teenagers; ...". For a given pair of words (u, v), we use the score of the synset pair (s'_u, s'_v) that has the lowest KL divergence among all the pairs $S_n \times S_v$ where S_u, S_v are sets of synsets for words u and v, respectively. More precisely, $s(u, v) = -\min_{s_u \in S_u, s_v \in S_v} D(s_u, s_v)$. This pair selection corresponds to choosing the synset pair that has the highest degree of entailment. This approach has been used in word embeddings literature to select most related word pairs (Athiwaratkun & Wilson, 2017). For word pairs that are not in the model, we assign the score equal to the median of all scores. We evaluate our model scores against the human annotated scores using Spearman's rank correlation.

Table 3 shows HYPERLEX results of our models **DOE-A** (asymmetric) and **DOE-S** (symmetric) as well as other competing models. The model **DOE-A** which uses KL divergence and negative sampling approach **S1**, **S2** and **S4** outperforms all other existing models, achieving state-of-the-art performance for the HYPERLEX noun dataset. (See Section A.1 for hyperparameter details) The model **DOE-S** which uses expected likelihood kernel attains a lower score of 0.455 compared to the asymmetric counterpart (**DOE-A**). This result underscores the importance of asymmetric measures which can capture relationship directionality.

We provide a brief summary of competing models: **FR** scores are based on concept word frequency ratio (Weeds et al., 2004). **SLQS** uses entropy-based measure to quantify entailment (Santus et al., 2014). **Vis-ID** calculates scores based on visual generality measures (Kiela et al., 2015). **WN-B** calculates the scores based on the shortest path between concepts in WN taxonomy (Miller, 1995). **w2g** Guassian embeddings trained using the methodology in Vilnis & McCallum (2015). **VOE** Vector order embeddings (Vendrov et al., 2016). **Euc** and **Poin** calculate scores based on the Euclidean distance and Poincaré distance of the trained Poincaré embeddings (Nickel & Kiela, 2017). The models **FR** and **SLQS** are based on word occurrences in text corpus, where **FR** is trained on the British National Corpus and **SLQS** is trained on UKWAC, WACKYPEDIA (Bailey & Thompson, 2006; Baroni et al., 2009) and annotated BLESS dataset (Baroni & Lenci, 2011). Other models **Vis-ID, w2g, VOE, Euc, Poin** and ours are trained on WordNet, with the exception that **Vis-ID** also uses

Table 3: Spearman's correlation for HYPERLEX nouns.

	FR	SLQS	Vis-ID	WN-B	w2g	VOE	Poin	HypV	DOE-S	DOE-A
ρ	0.283	0.229	0.253	0.240	0.192	0.195	0.512	0.540	0.455	**0.590**

Table 4: Spearman's correlation for HYPERLEX nouns for different negative sample schemes.

Negative Samples	ρ
1×**S1**	0.527
2×**S1**	0.529
5×**S1**	0.518
10×**S1**	0.517
1×**S1** + **S2**	0.567
2×**S1** + **S2**	0.567
3×**S1** + **S2**	0.584
5×**S1** + **S2**	0.561
10×**S1** + **S2**	0.550

Negative Samples	ρ
1×**S1** + **S2** + **S4**	**0.590**
2×**S1** + **S2** + **S4**	0.580
5×**S1** + **S2** + **S4**	0.582
1×**S1** + **S2** + **S3**	0.570
2×**S1** + **S2** + **S3**	0.581
S1 + 0.1×**S2** +0.9×**S3**	0.564
S1 + 0.3×**S2** +0.7×**S3**	0.574
S1 + 0.7×**S2** +0.3×**S3**	0.555
S1 + 0.9×**S2** +0.1×**S3**	0.533

Google image search results for visual data. The reported results of **FR**, **SLQS**, **Vis-ID**, **WN-B**, **w2g** and **VOE** are from Vulić et al. (2016).

We note that an implementation of Gaussian embeddings model (**w2g**) reported by Vulić et al. (2016) does not perform well compared to previous benchmarks such as **Vis-ID, FR, SLQS**. Our training approach yields the opposite results and outperforms other highly competitive methods such as Poincaré embeddings and Hypervec. This result highlights the importance of the training approach, even if the concept representation of our work and Vilnis & McCallum (2015) both use Gaussian distributions. In addition, we observe that vector order embeddings (VOE) do not perform well compared to our model, which we hypothesize is due to the "soft" orders induced by the divergence penalty that allows our model scores to more closely reflect hypernymy degrees.

We note another interesting observation that a model trained on a symmetric divergence (ELK) from Section 4.2 can also achieve a high HYPERLEX correlation of 0.532 if KL is used to calculate the model scores. This is because the encapsulation behavior can arise even though the training penalty is symmetric (more explanation in Section 4.2). However, using the symmetric divergence based on ELK results in poor performance on HYPERLEX (0.455), which is expected since it cannot capture the directionality of hypernymy.

We note that another model LEAR obtains an impressive score of 0.686 (Vulić & Mrkšić, 2014). However, LEAR use pre-trained word embeddings such as WORD2VEC or GLOVE as a pre-processing step, leveraging a large vocabulary with rich semantic information. To the best of our knowledge, our model achieves the highest HYPERLEX Spearman's correlation among models without using large-scale pre-trained embeddings.

Table 4 shows the effects of negative sample selection described in Section 3.5. We note again that **S1** is the technique used in literature Socher et al. (2013); Vendrov et al. (2016) and **S2**, **S3**, **S4** are the new techniques we proposed. The notation, for instance, $k \times$ **S1** $+$ **S2** corresponds to using k samples from **S1** and 1 sample from **S2** per each positive sample. We observe that our new selection methods offer strong improvement from the range of $0.51 - 0.52$ (using **S1** alone) to 0.55 or above for most combinations with our new selection schemes.

5 FUTURE WORK

Analogous to recent work by Vulić & Mrkšić (2014) which post-processed word embeddings such as GLOVE or WORD2VEC, our future work includes using the WordNet hierarchy to impose encapsulation orders when training probabilistic embeddings.

In the future, the distribution approach could also be developed for encoder-decoder based models for tasks such as caption generation where the encoder represents the data as a distribution, containing semantic and visual features with uncertainty, and passes this distribution to the decoder which maps to text or images. Such approaches would be reminiscent of variational autoencoders (Kingma & Welling, 2013), which take *samples* from the encoder's distribution.

ACKNOWLEDGEMENTS

We thank NSF IIS-1563887 for support.

REFERENCES

Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. In *ACL*, 2017.

Steve Bailey and Dave Thompson. UKWAC: building the uk's first public web archive. *D-Lib Magazine*, 12(1), 2006.

Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 2011.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 2009.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *EACL*, pp. 23–32, 2012.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. Natural language communication with robots. In *NAACL*, 2016.

Julia Hockenmaier and Alice Lai. Learning to predict denotational probabilities for modeling entailment. In *EACL*, 2017.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *JMLR*, 2004.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. Exploiting image generality for lexical entailment detection. In *ACL*, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, 2015.

Xiang Li, Luke Vilnis, and Andrew McCallum. Improved representation learning for predicting commonsense ontologies. *CoRR*, abs/1708.00549, 2017. URL `http://arxiv.org/abs/1708.00549`.

Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *NIPS*, 2016.

Friedrich Liese and Igor Vajda. *Convex Statistical Distances*. Leipzig : Teubner, 1987.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *ACL workshop*, 2002.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11), November 1995.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NIPS*, 2017.

Leandro Pardo. *Statistical Inference Based on Divergence Measures*, chapter 1, pp. 1–54. Chapman & Hall/CRC, 2006.

Alfred Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., 1961.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664, 2015.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing hypernyms in vector spaces with entropy. In *EACL*, 2014.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 2013.

Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016. URL `http://arxiv.org/abs/1612.03975`.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *ICLR*, 2016.

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *ICLR*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

Ivan Vulić and Nikola Mrkšić. Specialising word vectors for lexical entailment. *CoRR*, abs/1710.06371, 2014.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. Hyperlex: A large-scale evaluation of graded lexical entailment. *CoRR*, 2016.

Julie Weeds, David J. Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *COLING*, 2004.

A Supplementary Materials

A.1 Model Hyperparameters

In Section 4.3, the 2−dimensional Gaussian model is trained with **S-1** method where the number of negative samples is equal to the number of positive samples. The best hyperparameters for $d = 2$ model is $(m, \beta, \gamma) = (100.0, 2 \times 10^{-4}, 3.0)$.

In Section 4.2, the best hyperparameters (m, β, γ) for each of our model are as follows: For Gaussian with KL penalty: $(2000.0, 5 \times 10^{-5}, 500.0)$, , Gaussian with reversed KL penalty: $(1000.0, 1 \times 10^{-4}, 1000.0)$, Gaussian with ELK penalty $(1000, 1 \times 10^{-5}, 10)$.

In Section 4.4, we use the same hyperparameters as in 4.2 with KL penalty, but a different negative sample combination in order to increase the distinguishability of divergence scores. For each positive sample in the training set, we use one sample from each of the methods **S1**, **S2**, **S4**. We note that the model from Section 4.2, using **S1** with the KL penalty obtains a Spearman's correlation of 0.527.

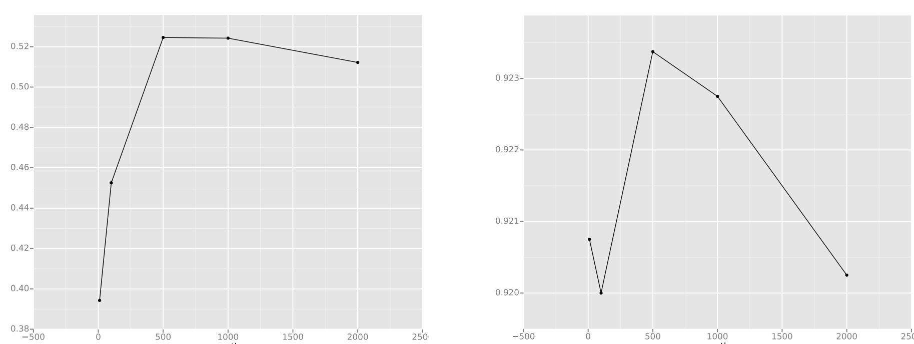
A.2 Analysis of Training Methodology

We emphasize that Gaussian embeddings have been used in the literature, both in the unsupervised settings where word embeddings are trained with local contexts from text corpus, and in supervised settings where concept embeddings are trained to model annotated data such as WordNet . The results in supervised settings such as modeling WordNet have been reported to compare with competing models but often have inferior performance (Vendrov et al., 2016; Vulić et al., 2016). Our paper reaches the opposite conclusion, showing that a different training approach using Gaussian representations can achieve state-of-the-art results.

A.2.1 Divergence Threshold

Consider a relationship $f \preceq g$ where f is a hyponym of g or g is a hypernym of f. Even though the divergence $D(f||g)$ can capture the extent of encapsulation, a density f will have the lowest divergence with respect with g only if $f = g$. In addition, if f is a more concentrated distribution that is encompassed by g, $D(f||g)$ is minimized when f is at the center of g. However, if there any many hyponyms f_1, f_2 of g, the hyponyms can compete to be close to the center, resulting in too much overlapping between f_1 and f_2 if the random sampling to penalize negative pairs is not sufficiently strong. The divergence threshold γ is used such that there is no longer a penalty once the divergence is below a certain level.

We demonstrate empirically that the threshold γ is important for learning meaningful Gaussian distributions. We fix the hyperparameters $m = 2000$ and $\beta = 5 \times 10^{-5}$, with **S1** negative sampling. Figure 5 shows that there is an optimal non-zero threshold and yields the best performance for both WordNet Hypernym prediction and HyperLex Spearman's correlation. We observe that using $\gamma = 0$ is detrimental to the performance, especially on HyperLex results.



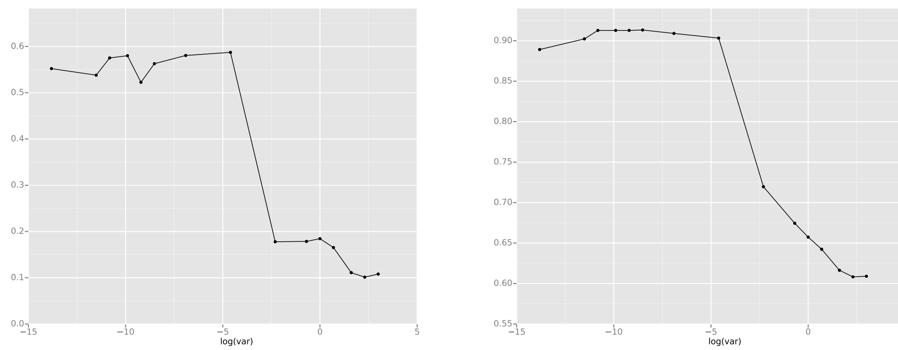
(a) (b)

Figure 5: (a) Spearman's correlation on HyperLex versus γ (b) Test Prediction Accuracy versus γ.

A.2.2 Initial Variance Scale

As opposed to the mean vectors that are randomly initialized, we initialize all diagonal covariance elements to be the same. Even though the variance can adapt during training, we find that different initial scales of variance result in drastically different performance. To demonstrate, in Figure 6, we show the best test accuracy and

HYPERLEX Spearman's correlation for each initial variance scale, with other hyperparameters (margin m and threshold γ) tuned for each variance. We use **S1 + S2 + S4** as a negative sampling method. In general, a low variance scale β increases the scale of the loss and requires higher margin m and threshold γ. We observe that the best prediction accuracy is obtained when $\log(\beta) \approx -10$ or $\beta = 5 \times 10^{-5}$. The best HYPERLEX results are obtained when the scales of β are sufficiently low. The intuition is that low β increases the scale of divergence $D(\cdot||\cdot)$, which increases the ability to capture relationship nuances.



(a) (b)

Figure 6: (a) Spearman's correlation on HYPERLEX versus $\log(\beta)$ (b) Test Prediction Accuracy versus $\log(\beta)$.

A.2.3 LOSS FUNCTION

We verify that for this task, our loss function in Equation 7 in superior to Equation 1 originally proposed by Vilnis & McCallum (2015). We use the exact same setup with new negative sample selections and KL divergence thresholding and compare the two loss functions. Table 5 verifies our claim.

Table 5: Best results for each loss function for two negative sampling setups: **S1** (**Left**) and **S1 + S2 + S4** (**Right**)

	Test Accuracy	HYPERLEX
Eq. 7	0.923	0.527
Eq. 1	0.886	0.524

	Test Accuracy	HYPERLEX
Eq. 7	0.911	0.590
Eq. 1	0.796	0.489

A.2.4 DIMENSIONALITY

Table 6 shows the results for many dimensionalities for two negative sample strategies: **S1** and **S1 + S2 + S4** .

Table 6: Best results for each dimension with negative samples **S1** (**Left**) and **S1 + S2 + S4** (**Right**)

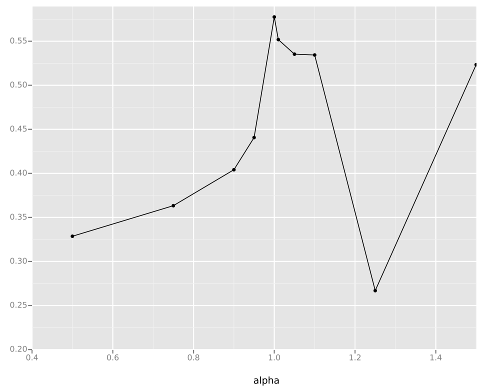
d	Test Accuracy	HYPERLEX
5	0.909	0.437
10	0.919	0.462
20	0.922	0.487
50	0.923	0.527
100	0.924	0.526
200	0.918	0.526

d	Test Accuracy	HYPERLEX
5	0.901	0.483
10	0.909	0.526
20	0.914	0.545
50	0.911	0.590
100	0.913	0.573
200	0.910	0.568

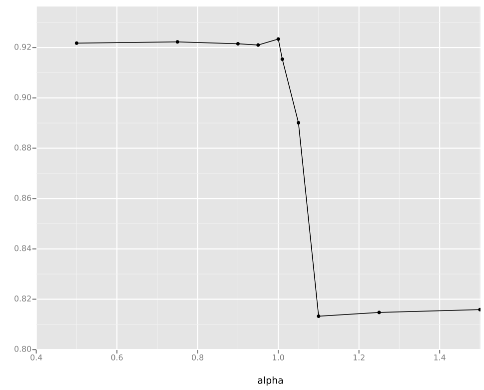
A.2.5 α-DIVERGENCES

Table 7 show the results using models trained and evaluated with $D(\cdot||\cdot) = D_\alpha(\cdot||\cdot)$ with negative sampling approach **S1**. Interestingly, we found that $\alpha \to 1$ (KL) offers the best result for both prediction accuracy and

HYPERLEX . It is possible that $\alpha = 1$ is sufficiently asymmetric enough to distinguish hypernym directionality, but does not have as sharp penalty as in $\alpha > 1$, which can help learning.



(a)



(b)

Figure 7: (a) Spearman's correlation on HYPERLEX versus α (b) Test Prediction Accuracy versus α.