Bhattacharyya Error and Divergence using Variational Importance Sampling

Peder A. Olsen[1], John R. Hershey[1]

[1]IBM T. J. Watson Research Center,
{pederao,jrhershe}@us.ibm.com

10.21437/Interspeech.2007-9

Abstract

Many applications require the use of divergence measures between probability distributions. Several of these, such as the Kullback Leibler (KL) divergence and the Bhattacharyya divergence, are tractable for single Gaussians, but intractable for complex distributions such as Gaussian mixture models (GMMs) used in speech recognizers. For tasks related to classification error, the Bhattacharyya divergence is of special importance. Here we derive efficient approximations to the Bhattacharyya divergence for GMMs, using novel variational methods and importance sampling. We introduce a combination of the two, variational importance sampling (VISa), which performs importance sampling using a proposal distribution derived from the variational approximation. VISa achieves the same accuracy as naive importance sampling at a fraction of the computation. Finally we apply the Bhattacharyya divergence to compute word confusability and compare the corresponding estimates using the KL divergence.

Index Terms: Variational importance sampling, Bhattacharyya divergence, variational methods, Gaussian mixture models.

1. Introduction

The Bhattacharyya error [1] between two probability density functions $f(x)$ and $g(x)$, is commonly used in statistics as a measure of similarity between two density distributions. We define the Bhattacharyya measure,

$$B(f,g) \stackrel{\text{def}}{=} \int \sqrt{f(x)g(x)}\,\mathrm{dx}. \tag{1}$$

The Bhattacharyya error is then $\sqrt{p_f p_g}B(f,g)$, where p_f and p_g are priors placed on f and g. Here we assume these priors are equal and deal with $B(f,g)$ directly for simplicity.

The corresponding Bhattacharyya divergence is defined as $D_B(f,g) = -\log B(f,g)$. The Bhattacharyya measure satisfies the properties $0 \le B(f,g) \le 1$, $B(f,g) = B(g,f)$ and $B(f,g) = 1$ if and only if $f = g$. The Bhattacharyya divergence satisfies similar properties.

The Bhattacharyya divergence has been used in machine learning as a kernel, [2], and in speech recognition for applications such as clustering of phonemes, [3], and feature selection, [4]. In this paper we apply the Bhattacharyya divergence to the problem of assigning a score indicating the level of confusability between a pair of words. The KL divergence has previously been used for this purpose, as can be seen in [5, 6, 7].

The use of the Bhattacharyya divergence for this purpose can be motivated by the fact that it closely approximates the *Bayes error*, $B_e(f,g) = 1/2\int \min(f,g)$. Figure 1 shows a scatter plot in which each point represents a pair of GMMs derived from a speech model, for Monte Carlo estimates of the KL divergence and Bhattacharyya divergence, plotted against the *Bayes divergence* $D_{BE} = -log2B_e(f,g)$.

It is clear that the KL divergence makes a poor estimate of the Bayes divergence compared to the Bhattacharyya divergence. The plot of the Bayes divergence using 100K samples shows how estimates of Bayes divergence become unreliable for large divergences, whereas it appears that, in this regime, our estimates of Bhattacharyya divergence more closely match the Bayes divergence than direct estimates of Bayes divergence. The rest of the paper deals with how we can compute accurate estimates of Bhattacharyya divergence.

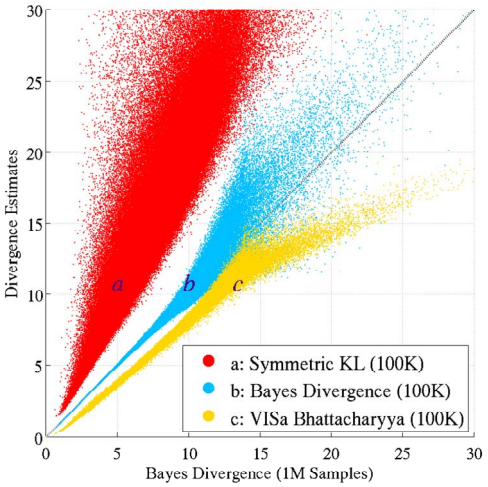


Figure 1: Scatter plot of a) symmetric KL divergence, b) Bayes divergence , and c) Bhattacharyya divergence, estimated via importance sampling with 100K samples, versus the Bayes divergence estimated using 1 million samples, plotted for all pairs of the 826 GMMs.

For two Gaussians f and g the Bhattacharyya divergence has a closed formed expression,

$$\begin{aligned} D_B(f\,g) &= \frac{1}{8}(\mu_f-\mu_g)^\top\left(\frac{\Sigma_f+\Sigma_g}{2}\right)^{-1}(\mu_f-\mu_g) \\ &\quad +\frac{1}{2}\log\det\left(\frac{\Sigma_f+\Sigma_g}{2}\right)-\frac{1}{4}\log\det(\Sigma_g\Sigma_f) \end{aligned} \tag{2}$$

whereas for two GMMs no such closed form expression exists. In the rest of this paper we consider f and g to be GMMs. The marginal densities of $x \in \mathbb{R}^d$ under f and g are

$$\begin{aligned} f(x) &= \textstyle\sum_a \pi_a \mathcal{N}(x;\mu_a,\Sigma_a) \\ g(x) &= \textstyle\sum_b \omega_b \mathcal{N}(x;\mu_b,\Sigma_b) \end{aligned} \tag{3}$$

where π_a is the prior probability of each state, and $\mathcal{N}(x;\mu_a,\Sigma_a)$ is a Gaussian in x with mean μ_a and covariance Σ_a.

We will frequently use the shorthand notation $f_a(x) = \mathcal{N}(x;\mu_a,\Sigma_a)$ and $g_b(x) = \mathcal{N}(x;\mu_b,\Sigma_b)$. Our estimates of $B(f,g)$ will make use of the Bhattacharyya measure between individual components, which we write as $B(f_a,g_b)$.

2. Monte Carlo sampling

One method that allows us to estimate the Bhattacharyya measure between two mixture models, $B(f,g)$ for large dimension d, with arbitrary accuracy is Monte Carlo simulation. If we draw n samples, $\{x_i\}_{i=1}^n$, from a distribution h and compute $\hat{B}_h(f,g) = \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{f(x_i)g(x_i)}}{h(x_i)}$ the resulting quantity is an unbiased estimate of the Bhattacharyya measure with variance

$$\frac{1}{n}\left(\int \frac{f(x)g(x)}{h(x)}\,\mathrm{dx} - B(f,g)^2\right). \tag{4}$$

When computing the KL divergence with Monte Carlo sampling the traditional choice of sampling distribution is f. Here, this yields an estimator

$$\hat{B}_f(f,g) = \frac{1}{n}\sum_{i=1}^n \sqrt{\frac{g(x_i)}{f(x_i)}} \tag{5}$$

with variance

$$\mathrm{var}[\hat{B}_f(f,g)] = \frac{1}{n}\left(1 - B(f,g)^2\right). \tag{6}$$

From the expression of the variance we see that the estimator has higher accuracy the closer g is to f. For practical purposes this is the most interesting case.

Using importance sampling we can do even better. $B(f,g)$ is symmetric, so a sampling distribution symmetric in f and g, such as, $\frac{f+g}{2}$, is a natural choice. The variance for this choice of sampling distribution is

$$\mathrm{var}[\hat{B}_{\mathrm{avg}}(f,g)] = \frac{1}{n}\left(\int \frac{2fg}{f+g} - B(f,g)^2\right). \tag{7}$$

But $\frac{2fg}{f+g}$, the harmonic mean of f and g, is bounded from above by the arithmetic mean, $\frac{f+g}{2}$, and thus

$$\mathrm{var}[\hat{B}_{\mathrm{avg}}(f,g)] \leq \frac{1}{n}\left(\int \frac{f+g}{2} - B(f,g)^2\right) = \mathrm{var}[\hat{B}_f(f,g)]. \tag{8}$$

We have proved that $\frac{f+g}{2}$ is *uniformly* a better sampling distribution than f. We shall see later in this paper that, by use of variational techniques, we can construct yet better sampling distributions.

3. The unscented transformation

The unscented transform, [8], is an approach to estimate $E_{f_a}[h(x)]$ in such a way that the approximation is exact *for all* quadratic functions $h(x)$. It is possible to pick $2d$ "sigma" points $\{x_{a,k}\}_{k=1}^{2d}$ such that

$$\int f_a(x)h(x)\,\mathrm{dx} = \frac{1}{2d}\sum_{k=1}^{2d} h(x_{a,k}). \tag{9}$$

One possible choice of the sigma points is

$$x_{a,k} = \mu_a + \sqrt{d\lambda_{a,k}}\;e_{a,k} \tag{10}$$

$$x_{a,d+k} = \mu_a - \sqrt{d\lambda_{a,k}}\;e_{a,k}, \tag{11}$$

for $k = 1,\ldots,d$ where $\lambda_{a,k}$ and $e_{a,k}$ are the eigenvalues and eigenvectors of the covariance Σ_a of the Gaussian f_a. The Bhattacharyya error can be written $B(f,g) = \sum_a \pi_a E_{f_a}[h]$ with $h = \sqrt{g/f}$. Although h is not quadratic, the unscented estimate is then

$$D_{\mathrm{unscented}}(f\|g) = \frac{1}{2d}\sum_a \pi_a \sum_{k=1}^{2d}\sqrt{\frac{g(x_{a,k})}{f(x_{a,k})}}. \tag{12}$$

4. The Gaussian approximation

A commonly used approximation to $B(f,g)$ is to replace f and g with Gaussians, $\hat{f}$ and $\hat{g}$. In one incarnation, one uses Gaussians whose mean and covariance matches that of f and g. The mean and covariance of f are given by

$$\begin{array}{rcl}\mu_{\hat{f}} &=& \sum_a \pi_a\mu_a \\ \Sigma_{\hat{f}} &=& \sum_a \pi_a(\Sigma_a + (\mu_a-\mu_{\hat{f}})(\mu_a-\mu_{\hat{f}})^\top).\end{array} \tag{13}$$

The approximation $B_{\mathrm{gauss}}(f,g)$ is given by the closed-form expression, $B_{\mathrm{gauss}}(f,g) = B(\hat{f},\hat{g})$, using equation (2).

5. First variational bound

Let $\phi_{ab} \geq 0$ satisfy $1 = \sum_{ab}\phi_{ab}$. Then by use of Jensen's inequality and the concavity of the square root we have the following straightforward computation

$$B(f,g) = \int\sqrt{fg} \tag{14}$$

$$= \int\sqrt{\sum_{ab}\phi_{ab}\frac{\pi_a f_a \omega_b g_b}{\phi_{ab}}} \tag{15}$$

$$\geq \sum_{ab}\phi_{ab}\int\sqrt{\frac{\pi_a\omega_b}{\phi_{ab}}}\sqrt{f_a g_b} \tag{16}$$

$$= \sum_{ab}\sqrt{\phi_{ab}\pi_a\omega_b}B(f_a,g_b). \tag{17}$$

This inequality holds in the entire domain of the variational parameters. By maximizing with respect to $\phi_{ab} \geq 0$ and the constraint $1 = \sum_{ab}\phi_{ab}$ we get

$$\phi_{ab} = \frac{\pi_a\omega_b B(f_a,g_b)^2}{\sum_{a'b'}\pi_{a'}\omega_{b'}B(f_{a'},g_{b'})^2} \tag{18}$$

which upon substitution gives

$$B(f,g) \geq \sqrt{\sum_{ab}\pi_a\omega_b B(f_a,g_b)^2}. \tag{19}$$

Every other approximation to the Bhattacharyya measure so far has satisfied the property $B(f,f) = 1$. For this variational estimate it is not the case.

6. Second variational bound

We can follow the approach used in [9], and use the variational principle in yet another way. We introduce the variational parameters $\phi_{b|a} \geq 0$ and $\psi_{a|b} \geq 0$ satisfying the constraints $\sum_a \phi_{a|b} = \sum_b \psi_{b|a} = 1$. Using the variational parameters we may write

$$\begin{array}{rcccl} f &=& \sum_a \pi_a f_a &=& \sum_{ab}\pi_a\psi_{b|a}f_a \\ g &=& \sum_b \omega_b g_b &=& \sum_{ab}\omega_b\phi_{a|b}g_b. \end{array} \tag{20}$$

With this notation we use Jensen's inequality to obtain a lower bound of the Bhattacharyya measure as follows

$$
\begin{aligned}
B(f,g) &= \int f\sqrt{\frac{g}{f}} && (21)\\
&= \int f\sqrt{\sum_{ab}\frac{\pi_a\phi_{b|a}f_a}{f}\,\frac{\omega_b\psi_{a|b}g_b}{\pi_a\phi_{b|a}f_a}}\,\mathrm{dx} && (22)\\
&\geq \sum_{ab}\pi_a\phi_{b|a}\int f_a\sqrt{\frac{\omega_b\psi_{a|b}g_b}{\pi_a\phi_{b|a}f_a}}\mathrm{d}x && (23)\\
&= \sum_{ab}\sqrt{\phi_{b|a}\psi_{a|b}}\sqrt{\pi_a\omega_b}B(f_a,g_b). && (24)
\end{aligned}
$$

This inequality holds for any choice of variational parameters ϕ and ψ. We cannot jointly optimize (24) in ϕ and ψ. For a fixed value of ψ the value of ϕ that maximizes the lower bound is

$$\phi_{b|a} = \frac{\psi_{a|b}\omega_b B(f_a,g_b)^2}{\sum_{b'}\psi_{a|b'}\omega_{b'}B(f_a,g_{b'})^2}. \quad (25)$$

Correspondingly, if we fix ϕ the optimal value for ψ is

$$\psi_{a|b} = \frac{\phi_{b|a}\pi_a B(f_a,g_b)^2}{\sum_{a'}\phi_{a'|b}\pi_{a'}B(f_{a'},g_b)^2}. \quad (26)$$

Each iteration of (25) and (26) increases the value of the lower bound. We use a uniform distribution to initialize ϕ and ψ and iterate until convergence in the lower bound. In this case, it holds that $\hat{B}(f,f)=1$, and equivalently the upper bound $\hat{D}_B(f,f)=0$.

7. Variational importance sampling

Equation (4) gave the variance for a given sampling distribution h. Two sampling distributions can be compared using $\int fg/h$. To choose h we could thus attempt to minimize $\int fg/h$ with respect to $h(x)\geq 0$ and the constraint $\int h = 1$. The optimal choice of h is given by $h=\sqrt{fg}/\int\sqrt{fg}$, which has a variance of 0. However the denominator equals $B(f,g)$, which is the quantity we are trying to compute in the first place. That being said, it nonetheless tells us that we should be searching for a sampling distribution that approximates $\sqrt{fg}$. The variational estimate is such an estimate. We have

$$
\begin{aligned}
B(f,g) &\geq \sum_{ab}\sqrt{\phi_{b|a}\psi_{a|b}}\sqrt{\pi_a\omega_b}B(f_a,g_b) && (27)\\
&= \int\sum_{ab}\sqrt{\phi_{b|a}\psi_{a|b}\pi_a\omega_b}\sqrt{f_ag_b}. && (28)
\end{aligned}
$$

From which we can see that

$$h = \frac{\sum_{ab}\sqrt{\phi_{b|a}\psi_{a|b}\pi_a\omega_b}\sqrt{f_ag_b}}{\int\sum_{ab}\sqrt{\phi_{b|a}\psi_{a|b}\pi_a\omega_b}\sqrt{f_ag_b}} \quad (29)$$

is in some sense an approximation to $\sqrt{fg}/B(f,g)$. Since $\sqrt{f_ag_b}$ is a quadratic exponential, h is a Gaussian mixture distribution, which we know how to sample. After iterating (25) and (26) the variational parameters become sparse, and we can typically prune away enough components, so that the resulting mixture is of comparable size to f and g.

8. Bhattacharyya divergence experiments

In our experiments we used GMMs from an acoustic model used for speech recognition. The features $x\in\mathbb{R}^d$ are 39 dimensional, $d=39$, and the GMMs all have diagonal covariance. Furthermore the acoustic model consists of a total of 9,998 Gaussians belonging to 826 separate GMMs. The number of Gaussians per GMM varies from 1 to 76 (only 5 mixtures were single Gaussians). The median number of Gaussians per GMM was 9. We used these 826 GMMs to test our various different approximations to the Bhattacharyya divergence. We used all $\binom{826}{2}$ pairs of the GMMs in our tests, and compared each of the methods to the reference approximation, which was the VISa method with one million samples. To justify this reference, for each method we computed an estimate using 100,000 samples, and a reference using one million samples. Then for each estimate we chose the reference that minimized the variance of the deviation. In all cases this best reference was the VISa reference.

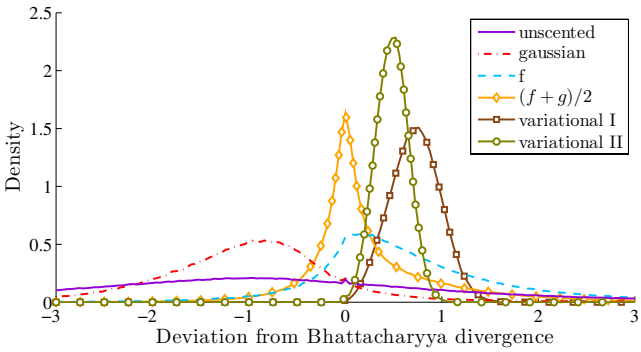


Figure 2: Histograms of deviations from the reference estimate, computed across all pairs of GMMs, for various method.

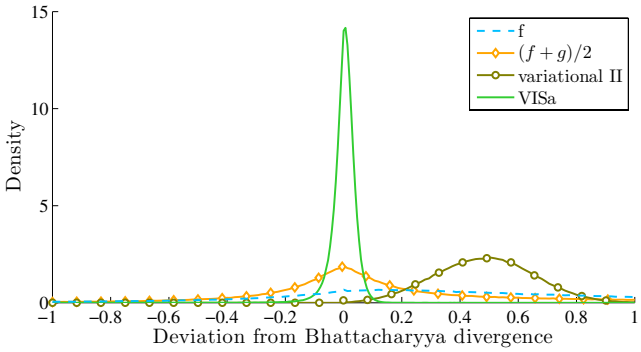


Figure 3: Histogram of various approximations, relative to the reference estimate, computed from all pairs of GMMs. The three Monte Carlo estimates are computed using 1000 samples.

Fig. 2 shows histograms of the two variational bounds, and the Gaussian and unscented approximations. In addition, we have plotted the Monte Carlo methods using the same number of samples, $2dn$, used in the unscented approximation. In our implementation we saw that the computation for the variational lower bound took about 1ms per GMM pair, for the variational approximation 0.6ms, for the Gaussian approximation 9ms and for the unscented approximation 11ms per pair. As seen in Fig. 2 the unscented approximation fails, whereas the Gaussian approximation can compete with sampling from f. The variational

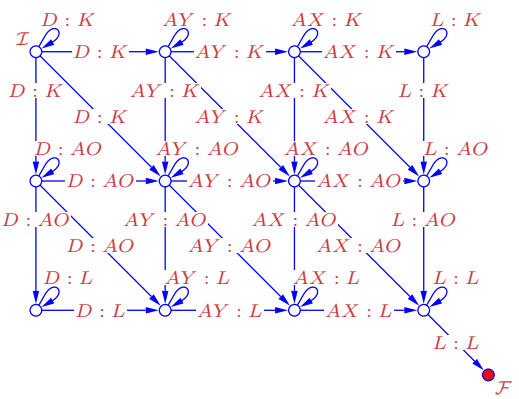


Figure 4: Product HMM for the words *call* and *dial*

methods, which are the cheapest, are also the best of the methods in Fig. 2, in terms of variance, although sampling from $(f+g)/2$ yields less bias. The VISa plot is absent here because it would dwarf the other histograms. The only methods that can give arbitrary accuracy, time permitting, are the Monte Carlo sampling methods. Fig. 3 shows how the various sampling methods compare to each other with 1000 samples. We compared the methods to the best variational estimate, i.e. the lower variational bound. It is clear that the VISa method is far superior to sampling from $\frac{f+g}{2}$, which is again far better than sampling from f. Finally, it is worth noting that we pruned the Gaussians with low priors in the VISa method so as to make the number of Gaussians in the GMM h comparable to that of f, and thus computationally competitive.

9. Word confusability experiments

In this section we briefly describe some experimental results where we use the Bhattacharyya divergence in place of KL divergence. The problem of estimating word confusability is discussed in [5] and [6]. A word is modeled in terms of an HMM and so the confusion between the two words can be modeled in terms of a cartesian product between the two HMMs as seen in Fig. 4. This structure is similar to the acoustic perplexity defined in [5] and the average divergence distance, [6], and we draw our methodology from these papers. The edit distance is the shortest path from the initial to the final node in the product graph. We use the edit distance as the indicator for how confusable two words are. In this case the edit distance is equivalent to the Viterbi path.

To measure how well each method can predict recognition errors we used a test suite consisting only of spelling data. There were a total of 38,921 spelling words (a-z) in the test suite with an average word error rate of about 19.3%. A total of 7,500 spelling errors were detected. Given the errors we estimated the probability of correct recognition $\mathrm{P}(w_1|w_2) = C(w_1, w_2)/C(w_2)$. We discarded cases where the error count was low, the total count was low, or the probability was 1.

We take into account the self-loop transition probabilities π_f and π_g corresponding to the HMM nodes associated with f and g, using $D(\pi_f \| \pi_g) = \pi_f \log\left(\frac{\pi_f}{\pi_g}\right) + (1-\pi_f)\log\left(\frac{1-\pi_f}{1-\pi_g}\right)$, as in [10]. To compute the edit distance we then use the following weights in the edit distance computation:

1. KL divergence: $D(f\|g)$.
2. Bhattacharyya divergence: $D_B(f,g)$.
3. Kl divergence with transitions: $\frac{D(f\|g)}{1-\pi_f} + D(\pi_f\|\pi_g)$
4. Bhattacharyya with transitions: $\frac{D_B(f,g)}{1-\pi_f} + D(\pi_f\|\pi_g)$.

Table 1 shows the experimental results using the four different kinds of weights. The Bhattacharyya divergence outperforms the KL divergence, and performs best when combined with the transition probabilities.

method description	squared correlation
KL divergence	0.571
Bhattacharyya	0.610
KL divergence with transitions	0.616
Bhattacharyya with transitions	**0.675**

Table 1: Squared correlation between edit distances and empirical confusability measurements.

10. References

[1] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[2] Tony Jebara and Risi Kondor, "Bhattacharyya and expected likelihood kernels," in *Conference on Learning Theory*, 2003.

[3] Brian Mak and Etienne Barnard, "Phone clustering using the bhattacharyya distance," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 4, pp. 2005–2008.

[4] George Saon and Mukund Padmanabhan, "Minimum bayes error feature selection for continuous speech recognition," in *NIPS*, 2000, pp. 800–806.

[5] Harry Printz and Peder Olsen, "Theory and practice of acoustic confusability," *Computer, Speech and Language*, vol. 16, pp. 131–164, January 2002.

[6] J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 890–906, May 2006.

[7] Qiang Huo and Wei Li, "A DTW-based dissimilarity measure for left-to-right hidden Markov models and its application to word confusability analysis," in *Interspeech 2006 - ICSLP*, Pittsburgh, PA, 2006, pp. 2338–2341.

[8] Simon Julier and Jeffrey K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Tech. Rep. RRG, Dept. of Engineering Science, University of Oxford, 1996.

[9] John Hershey and Peder Olsen, "Approximating the Kullback Leibler divergence between gaussian mixture models," in *ICASSP*, Honolulu, Hawaii, April 2007.

[10] Jia-Yu Chen, Peder Olsen, and John Hershey, "Word confusability - measuring hidden Markov model similarity," in *Proceedings of Interspeech 2007*, August 2007.