VARIATIONAL BHATTACHARYYA DIVERGENCE FOR HIDDEN MARKOV MODELS

John R. Hershey, and Peder A. Olsen

IBM T. J. Watson Research Center

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 simple distributions such as Gaussians, but are intractable for more complex distributions such as hidden Markov models (HMMs) used in speech recognizers. For tasks related to classification error, the Bhattacharyya divergence is of special importance, due to its relationship with the Bayes error. Here we derive novel variational approximations to the Bhattacharyya divergence for HMMs. Remarkably the variational Bhattacharyya divergence can be computed in a simple closed-form expression for a given sequence length. One of the approximations can even be integrated over all possible sequence lengths in a closed-form expression. We apply the variational Bhattacharyya divergence for HMMs to *word confusability*, the problem of estimating the probability of mistaking one spoken word for another.

Index Terms: Bhattacharyya Error, Bhattacharyya divergence, variational methods, Gaussian mixture models (GMMs), hidden Markov models (HMMs)

1. INTRODUCTION

The Bhattacharyya error between two probability density functions $f(x)$ and $g(x)$,[1]

$$B(f,g) \stackrel{\text{def}}{=} \frac{1}{2}\int \sqrt{f(x)g(x)}\,\mathrm{dx}, \tag{1}$$

is commonly used in statistics as a measure of similarity between two probability distributions. The corresponding Bhattacharyya divergence is defined as $D_B(f,g) = -\log 2B(f,g)$.

The Bhattacharyya divergence has previously been used in machine learning as a kernel [2], and in speech recognition for applications such as phoneme clustering for context dependency trees [3], feature selection [4]. The Bhattacharyya divergence cannot be computed analytically for a pair of mixture models. It can, however, be computed analytically for simple distributions such as gaussians. This makes it possible to come up with some reasonable analytical approximations for mixture models [5, 6]. In this paper, we show how some of these approximations can be directly extended to HMMs. We apply Bhattacharyya divergence to the problem of assigning a score indicating the level of confusability between a pair of spoken words, as in [6, 7], where the words are modeled by HMMs.

The Bhattacharyya error satisfies the properties $B(f,g) = B(g,f)$ (*symmetry*), $B(f,g) = 1/2$ if and only if $f = g$ (*identification*), and $0 \le B(f,g) \le 1/2$ almost everywhere. The Bhattacharyya error is closely related to the *Bayes error*, $B_e(f,g) = \frac{1}{2}\int \min(f(x),g(x))\ \mathrm{dx} \le B(f,g)$ via the *power mean inequality*. The Bhattacharyya divergence is also related to the *Kullback-Leibler (KL) divergence* $D_{KL}(f\|g) = \int f(x)\log f(x)/g(x)\ \mathrm{dx} \ge 2D_B(f,g)$, by *Jensen's inequality*.

For two gaussians f and g the Bhattacharyya divergence has a closed–form expression, [8]

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

In fact, the same is true if f and g are any of a wide range of useful distributions known as the exponential family, of which the gaussian is the most famous example. The computation is particularly simple for models with discrete observations. For more complex distributions such as mixture models or hidden Markov models (HMMs), no such closed-form expression exists.

Mixture models: We first consider the case where f and g are mixture models, then derive formulas for hidden Markov models. For the sake of concreteness we use gaussian observation distributions, without loss of generality. The marginal densities of $x \in \mathbb{R}^D$ under f and g are thus

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

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 error between individual components, which we write as $B(f_a,g_b)$. Note that the techniques we introduce apply even if $f_a(x)$ and $g_b(x)$ are not Gaussians, so long as $B(f_a,g_b)$ is known.

Hidden Markov models: A hidden Markov model (HMM) can be considered a special case of a GMM in which each state sequence is considered a mixture component. Hence we can in theory apply any approximation that works for a GMM to an HMM. To formulate the Bhattacharyya divergence for hidden Markov models, we must take care to define them in a way that yields a distribution (integrates to one) over all sequence lengths. For an HMM, f, emitting an observation sequence of length n, we let each state $a_{1:n} = (a_1,\ldots,a_n)$ be a sequence of hidden state discrete random variables, a_t taking values in $\mathcal{E}$, where $\mathcal{E}$ is the set of emitting states. Let $x_{1:n} = (x_1,\ldots,x_n)$ be a sequence of observations, with $x_t \in \mathbb{R}^d$. For the observations we use the shorthand $f_{a_t}(x_t) = \mathcal{N}(x_t;\mu_{a_t},\Sigma_{a_t})$. We also define non-emitting initial and final state values $\mathcal{I}$, and $\mathcal{F}$. The state sequence probabilities are thus formulated as a Markov chain $\pi_{a_{1:n}} = \pi_{a_1|\mathcal{I}}\,\pi_{\mathcal{F}|a_n}\prod_{t=2}^n \pi_{a_t|a_{t-1}}$, where $\pi_{a_1|\mathcal{I}}$ is an initial distribution, $\pi_{a_t|a_{t-1}}$ are transition probabilities, and $\pi_{\mathcal{F}|a_n}$ are the final state transitions. The transition probabilities are normalized such that $\sum_{a_1}\pi_{a_1|\mathcal{I}} = 1$, and $\pi_{\mathcal{F}|a_{t-1}} + \sum_{a_t}\pi_{a_t|a_{t-1}} = 1$, for $2 \le t \le n$. For a given sequence length n, we only consider paths

that reach the final state after exactly n observations. This allows the HMM to describe a distribution over all sequence lengths. The density assigned to a particular sequence length $f(x_{1:n})$ is:

$$f(x_{1:n}) = \sum_{a_{1:n}} \pi_{a_{1:n}} f_{a_{1:n}}(x_{1:n}) \tag{4}$$

$$= \sum_{a_{1:n}} \pi_{a_1|\mathcal{I}} \pi_{\mathcal{F}|a_n} f_{a_1}(x_1) \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} f_{a_t}(x_t), \tag{5}$$

and likewise for $g(x_{1:n})$. Note that we can integrate over all sequences $\mathbf{x} \in \cup_{n=1}^{\infty} \mathbb{R}^{n \times d}$, by summing over sequence lengths; moreover $\int f(\mathbf{x}) = \sum_{n=1}^{\infty} \int f(x_{1:n})\, \mathrm{dx}_{1:n} = 1$, so this is a proper density. Because the number of state sequences increases exponentially with the sequence length, and the observation likelihoods at a given time point are shared among many paths, practical computation must invoke an efficient recursion.

Unfortunately, the recursion does not directly extend to the Bhattacharyya divergence between two HMMs, where we have to consider all combinations of state sequences. With HMMs, as with GMMs, we can reduce the approximation to pair-wise Gaussian Bhattacharyya divergences. Unless there is a jointly recursive structure in the two HMMs, the approximation will not be tractable.

2. VARIATIONAL BOUNDS

Variational bounds for mixture models: We can bound the Bhattacharyya error for mixture models using Jensen's inequality and the concavity of the square root:

$$B(f,g) = \frac{1}{2}\int \sqrt{fg} = \frac{1}{2}\int \sqrt{\sum_{ab} \pi_a \omega_b f_a g_b} \tag{6}$$

$$\geq \sum_{ab} \pi_a \omega_b B(f_a, g_b) = \hat{B}^{\mathrm{jb}}(f,g). \tag{7}$$

However, we can improve this bound using variational parameters that express affinities between the states of the two models [6]. Let $\phi_{ab} \geq 0$ satisfy $1 = \sum_{ab} \phi_{ab}$. Then by use of Jensen's inequality we have

$$B(f,g) = \frac{1}{2}\int \sqrt{fg} = \frac{1}{2}\int \sqrt{\sum_{ab} \phi_{ab} \frac{\pi_a f_a \omega_b g_b}{\phi_{ab}}} \tag{8}$$

$$\geq \frac{1}{2} \sum_{ab} \phi_{ab} \int \sqrt{\frac{\pi_a \omega_b}{\phi_{ab}}} \sqrt{f_a g_b} \tag{9}$$

$$= \sum_{ab} \sqrt{\phi_{ab} \pi_a \omega_b} B(f_a, g_b). \tag{10}$$

This bound holds for any ϕ_{ab}. Notice that $\phi_{ab} = \sqrt{\pi_a \omega_b}$, recovers the simple Jensen bound of (6). However, by maximizing with respect to $\phi_{ab} \geq 0$ and the constraint $1 = \sum_{ab} \phi_{ab}$ we get

$$\phi_{ab} = \frac{\pi_a \omega_b B^2(f_a, g_b)}{\sum_{a'b'} \pi_{a'} \omega_{b'} B^2(f_{a'}, g_{b'})}, \tag{11}$$

which upon substitution into (10) gives

$$B(f,g) \geq \sqrt{\sum_{ab} \pi_a \omega_b B^2(f_a, g_b)} = \hat{B}^{\mathrm{vb}}(f,g). \tag{12}$$

One problem with this variational method, as well as the simple Jensen bound, is that they fail to preserve the identification property, that $\hat{B}(f,g) = 1/2$ if and only if $f = g$. This can be enforced by re-normalizing using the geometric mean of $B(f,f)$, and $B(g,g)$: $\hat{B}_{\mathrm{norm}}(f,g) = \hat{B}(f,g)\hat{B}^{-1/2}(f,f)\hat{B}^{-1/2}(g,g)$. The normalized estimate is no longer a bound. Nevertheless it is a better approximation, as shown in Figure 1, relative to Monte Carlo estimates, using the variational importance sampling technique proposed in [6]. Surprisingly, the normalization makes the looser Jensen bound perform better than the variational bound. Empirically it also turns out to work better than other power means. Figure 2 shows that in terms of approximating the Bayes error, the normalized Bhattacharyya approximations are almost as good as the Bhattacharyya divergence itself, as estimated using 1 million samples for each pair of 826 GMMs from a speech recognizer. The iterative variational bound shown here is given in [6], and does not need normalization.

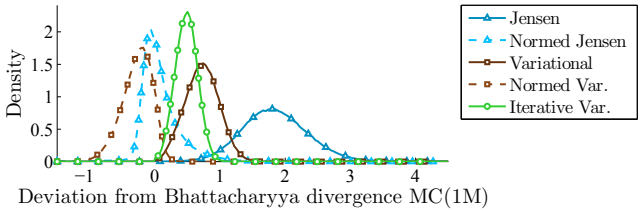

Fig. 1. Distribution of Bhattacharyya approximations relative to MC estimates with 1 million samples, for all pairs of 826 GMMs.

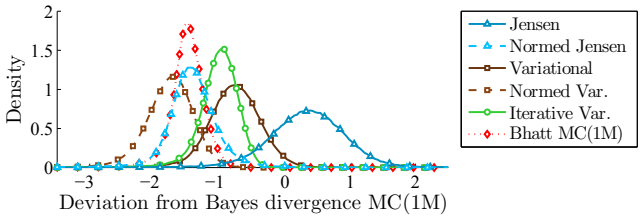

Fig. 2. Distribution of Bhattacharyya approximations relative to the Bayes error estimated with 1 million samples, for all pairs of GMMs.

Variational bounds for hidden Markov models: We extend the bound to HMMs by treating an entire state sequence $a_{1:n}$ as a single state. Thus it is as if we have a variational parameter $\phi_{a_{1:n}b_{1:n}}$ for each pair of sequences. Fortunately the optimized bound of (12) has a tractable form. Since the variational lower bound sums over the product of the two states $\pi_a \omega_b$ we can substitute our HMM into this formula and find the recursion.

$$\hat{B}^{\mathrm{vb}}(f_{1:n}, g_{1:n})^2 = \sum_{a_{1:n} b_{1:n}} \pi_{a_{1:n}} \omega_{b_{1:n}} B^2(f_{a_{1:n}}, g_{b_{1:n}})$$

$$= \sum_{a_{1:n} b_{1:n}} \pi_{a_1|\mathcal{I}} \pi_{\mathcal{F}|a_n} \omega_{b_1|\mathcal{I}} \omega_{\mathcal{F}|b_n} B^2(f_{a_1}, g_{b_1}) \times \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} \omega_{b_t|b_{t-1}} B^2(f_{a_t}, g_{b_t})$$

$$= \sum_{a_1} \pi_{a_1|\mathcal{I}} \sum_{b_1} \omega_{b_1|\mathcal{I}} B^2(f_{a_1}, g_{b_1}) \times \sum_{a_2} \pi_{a_2|a_1} \sum_{b_2} \omega_{b_2|b_1} B^2(f_{a_2}, g_{b_2}) \times \ldots \times \sum_{a_n} \pi_{a_n|a_{n-1}} \pi_{\mathcal{F}|a_n} \sum_{b_n} \omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} B^2(f_{a_n}, g_{b_n}). \tag{13}$$

This can be recursively computed, defining $\tilde{B}_t(a_t, b_t)$ as the contribution from earlier states to the current estimate at state a_t, b_t.

$$\tilde{B}_1(a_1, b_1) = \pi_{a_1|\mathcal{I}}\, \omega_{b_1|\mathcal{I}} \tag{14}$$

$$\tilde{B}_t(a_t, b_t) = \tag{15}$$

$$\sum_{a_{t-1}} \pi_{a_t|a_{t-1}} \sum_{b_{t-1}} \omega_{b_t|b_{t-1}} B^2(f_{a_{t-1}}, g_{b_{t-1}}) \tilde{B}_{t-1}(a_{t-1}, b_{t-1})$$

Handling the end case we get

$$\hat{B}^{\mathrm{vb}}(f_{1:n}, g_{1:n}) = \sqrt{\sum_{a_n} \pi_{\mathcal{F}|a_n} \sum_{b_n} \omega_{\mathcal{F}|b_n} \tilde{B}_n(a_n, b_n) B(a_n, b_n)}.$$

In matrix notation, we write the element-wise product as $A \circ B = \{a_{ij} b_{ij}\}$, the element-wise exponentiation as $A^{\circ n} = \{a_{ij}^n\}$, and the Kronecker product as $A \otimes B = \{a_{ij} B\}$. We define transition matrices $\pi = \{\pi_{a_t|a_{t-1}}\}$ and $\omega = \{\omega_{b_t|b_{t-1}}\}$, and initial and final state probability vectors, $\pi_{\mathcal{I}} = \{\pi_{a_1|\mathcal{I}}\}$, $\omega_{\mathcal{I}} = \{\omega_{b_1|\mathcal{I}}\}$, $\pi_{\mathcal{F}} = \{\pi_{\mathcal{F}|a_n}\}$, $\omega_{\mathcal{F}} = \{\omega_{\mathcal{F}|b_n}\}$, and Bhattacharyya matrices $\mathbf{B} = \{B(f_{a_t}, g_{b_t})\}$, and $\tilde{\mathbf{B}}_t = \{\tilde{B}_t(a_t, b_t)\}$. The recursion is then

$$\tilde{\mathbf{B}}_t = \pi^\top (\tilde{\mathbf{B}}_{t-1} \circ \mathbf{B}^{\circ 2}) \omega \tag{16}$$

$$\mathrm{vec}\, \tilde{\mathbf{B}}_t = \Big((\pi \otimes \omega) \circ (\mathrm{vec}^\top (\mathbf{B}^{\circ 2}) \otimes \mathbf{1}) \Big)\ \mathrm{vec}\, \tilde{\mathbf{B}}_{t-1} \tag{17}$$

$$= A\ \mathrm{vec}\, \tilde{\mathbf{B}}_{t-1}, \tag{18}$$

where $A = (\pi \otimes \omega) \circ (\mathrm{vec}^\top (\mathbf{B}^{\circ 2}) \otimes \mathbf{1})$.
Similarly defining $v_{\mathcal{I}} = \pi_{\mathcal{I}} \otimes \omega_{\mathcal{I}}$ and $v_{\mathcal{F}} = \pi_{\mathcal{F}} \otimes \omega_{\mathcal{F}}$,

$$\hat{B}^{\mathrm{vb}}(f_{1:n}, g_{1:n}) = \sqrt{(v_{\mathcal{F}} \circ \mathrm{vec}\, \mathbf{B}^{\circ 2})^\top A^n\, v_{\mathcal{I}}}. \tag{19}$$

Note that (16) is a more efficient form than (17), taking a factor of K fewer multiplications, where $K = |\mathcal{E}|$ is the number of emitting states. However (19) has the nice property that A^n can be computed using an eigenvalue expansion, which may be much more efficient for longer sequences.
Considering all sequence lengths, the approximation is simply

$$\hat{B}^{\mathrm{vb}}(f, g) = \sum_{n=1}^{\infty} \hat{B}(f_{1:n}, g_{1:n}). \tag{20}$$

Since $\hat{B}(f_{1:n}, g_{1:n}) \to 0$ as $n \to \infty$, in practice the sum can be truncated to the terms that are significantly non-zero.
In the case of the Jensen bound, we can compute the whole sum analytically, at the expense of a looser bound. Defining $A = (\pi \otimes \omega) \circ (\mathrm{vec}^\top (\mathbf{B}) \otimes \mathbf{1})$, we get

$$\hat{B}^{\mathrm{jb}}(f_{1:n}, g_{1:n}) = \Big(v_{\mathcal{F}}^\top \circ \mathrm{vec}\, \mathbf{B}\Big)^\top A^n v_{\mathcal{I}}, \tag{21}$$

Here we can analytically sum over all sequence lengths. Let $A = P^{-1} \Lambda P$ be the eigen-decomposition of non-symmetric non-negative matrix A. Then $\sum_{n=1}^{\infty} A^n = P^{-1}(I - \Lambda)^{-1} P - I = (I - A)^{-1} - I = \mathbf{C}$ if all eigenvalues are less than one in absolute value, which is guaranteed by the Perron-Frobenius theorem [9]. Hence,

$$\hat{B}^{\mathrm{jb}}(f, g) = \sum_{n=1}^{\infty} (v_{\mathcal{F}} \circ \mathrm{vec}\, \mathbf{B})^\top A^n v_{\mathcal{I}} \tag{22}$$

$$= (v_{\mathcal{F}} \circ \mathrm{vec}\, \mathbf{B})^\top \mathbf{C} v_{\mathcal{I}} \tag{23}$$

It is possible to extend the tighter iterative variational bound [6] for mixture models to HMMs, by factorizing the variational parameters into a Markov chain, as was done for the KL Divergence in [10]. However this method is more complicated, and models the distribution over pairs of paths less faithfully.

3. WEIGHTED EDIT DISTANCES

Various types of *weighted edit distances* have been applied to the task of estimating spoken word confusability, as discussed in [11] and [12]. A word is modeled in terms of a left–to–right HMM as in Fig. 3.

Fig. 3. An HMM for **call** with pronunciation `K AO L`. In practice, each phoneme is composed of three states, although here they are shown with one state each.

The confusion between two words can be heuristically modeled in terms of a cartesian product between the two HMMs as seen in Fig. 4. This structure is similar to that used for acoustic perplexity [11] and the average divergence distance [12].

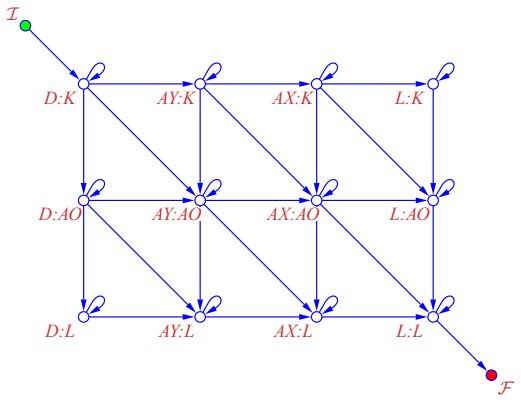

Fig. 4. Product HMM for the words *call* (`K AO L`) and *dial* (`D AY AX L`)

In the *weighted edit distance* (WED), weights are placed on the vertices that assign smaller values when the corresponding phoneme state models are more confusable. The WED is the shortest path (i.e., the Viterbi path) from the initial to the final node in the product graph. [7, 10]

$$D_{\mathrm{WED}}(f, g) = \min_n \min_{a_{1:n}, b_{1:n}} C(a_{1:n}, b_{1:n})$$

where $C(a_{1:n}, b_{1:n}) = \sum_{t=1}^{n} (w_{f_{a_t|a_{t-1}}} + w_{g_{b_t|b_{t-1}}} + w_{f_{a_t}, g_{b_t}})$ is the cost of the path, and the w are costs assigned to each transition. In our experiments we define $w_{f_{a_t|a_{t-1}}} = -\log \pi_{a_t|a_{t-1}}$, and $w_{g_{b_t|b_{t-1}}} = -\log \omega_{b_t|b_{t-1}}$. The weight at each node, $w_{f_{a_t}, g_{b_t}}$, is a dissimilarity measure between the acoustic models for each pair of HMM states. For the KL divergence WED, we define $w_{f_{a_t}, g_{b_t}} = D(f_{a_t} \| g_{b_t})$, and for the Bhattacharyya WED, we define $w_{f_{a_t}, g_{b_t}} = D_B(f_{a_t} \| g_{b_t})$. An interesting variation, which we call the *total weighted edit distance* TWED, is to sum over all paths and sequence lengths:

$$D_{\mathrm{TWED}}(f, g) = -\log \sum_n \sum_{a_{1:n}, b_{1:n}} e^{-C(a_{1:n}, b_{1:n})}. \tag{24}$$

That is, we sum over the probabilities, rather than the costs, which corresponds to the interpretation as a product HMM. The variational Bhattacharyya divergence, $D_{\hat{B}}^{\mathrm{vb}}(f \| g) = -\log \hat{B}^{\mathrm{vb}}(f \| g)$, can be

seen as a special case of the total weighted edit distance, with the pairwise Bhattacharyya weights, $w_{f_{a_t},g_{b_t}} = -2\log B(f_{a_t}\|g_{b_t})$. In addition, the TWED with Bhattacharyya weights, $w_{f_{a_t},g_{b_t}} = -\log B(f_{a_t}\|g_{b_t})$ is identical to $D^{\mathrm{jb}}_{\hat{B}}(f\|g) = -\log \hat{B}^{\mathrm{jb}}(f\|g)$.

4. WORD CONFUSABILITY EXPERIMENTS

In this section we describe some experimental results where we use the HMM divergence estimates to approximate spoken word confusability. To measure how well each method can predict recognition errors we used a test suite consisting of spelling data, meaning utterances in which letter sequences are read out, i.e., "J O N" is read as "`jay oh en`." There were a total of 38,921 instances of the spelling words (the letters A-Z) in the test suite with an average letter error rate of about 19.3%. A total of 7,500 recognition errors were detected. Given the errors we estimated the probability of error for each word pair as $\mathrm{E}(w_1, w_2) = \frac{1}{2}\mathrm{P}(w_1|w_2) + \frac{1}{2}\mathrm{P}(w_2|w_1)$, where $\mathrm{P}(w_1|w_2)$ is the fraction of utterances of w_2 that are recognized as w_1. We discarded cases where $w_1 = w_2$, since these dominate the results and exaggerate the performance. We also discarded unreliable cases where the word counts were too low. Continuous speech recognition was used, rather than isolated word recognition, so some recognition errors may have been due to misalignment.

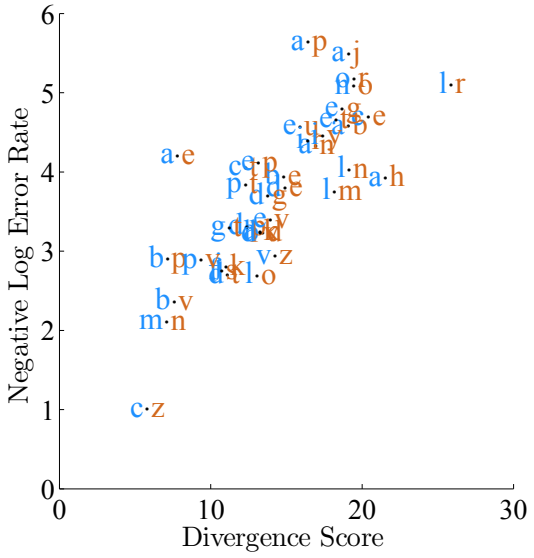

Fig. 5. The negative log error rate for all spelling word pairs compared to the variational HMM Bhattacharyya score.

Figure 5 shows a scatter plot of the variational Bhattacharyya score for each pair of letters, versus the empirical error measurement. Note that similar-sounding combinations of letters appear on the lower left (e.g. "c·z"), and dissimilar combinations appear in the upper right (e.g. "a·p"). We computed the divergences by direct Monte-Carlo sampling of the HMM state sequences. In addition to the Bhattacharyya approximations, we also computed KL divergences and a KL divergence weighted edit distance. Table 1 shows the results using all the different methods. The HMM Bhattacharyya divergence approximations outperform all other methods, even the Monte Carlo Bhattacharyya divergence with 100K samples, much to our surprise. Figure 5 shows a scatter-plot of the variational Bhattacharyya score

Method	Score
MC 100K Min KL Divergence	0.450
MC 100K Bhattacharyya Divergence	0.530
KL Divergence Weighted Edit Distance	0.571
Normalized Bhattacharyya Weighted Edit Distance	0.610
Normalized VB Bhattacharyya Divergence	0.631
Normalized JB Bhattacharyya Divergence	0.646

Table 1. Squared correlation scores between the various model-based divergence measures and the empirical word confusabilities $-\log \mathrm{E}(w_1, w_2)$. VB and JB refer to the variational bound and Jensen bound respectively. Min refers to the $\min(D(f\|g), D(g\|f))$. MC 100K refers to Monte Carlo simulations with 100,000 samples of HMM sequences.

This is natural since the Bhattacharyya divergence is known to yield a tighter bound on the Bayes error than the KL divergence. As with GMMs, the normalized Jensen bound also outperforms the normalized variational bound for HMMs.

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

[6] Peder Olsen and John Hershey, "Bhattacharyya error and divergence using variational importance sampling," in *Proceedings of Interspeech 2007*, August 2007, to appear.

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

[8] Keinosuke Fukunaga, *Statistical Pattern Recognition*, Academic Press, Inc., San Diego, CA, 1990.

[9] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[10] John R. Hershey, Peder A. Olsen, and Steven J. Rennie, "Variational Kullback Leibler divergence for hidden Markov models," in *Proceedings of ASRU*, Kyoto, Japan, December 2007.

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

[12] 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.